STAT 300: My Introductory Statistics Textbook (Mirzaagha)

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

This text was compiled on 03/18/2025

TABLE OF CONTENTS

3: Regression Analysis

4: Fundamental Principle of Counting and Rules of Probability

9: Hypothesis Testing about Population Mean and Proportion

10: Hypothesis Testing about Two Population Means and Proportions

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing****.*

CHAPTER OVERVIEW

1: Basic Ideas

1.1: Videos

Using stents to prevent strokes - Real case study with a surprising finding

Data Basics - Typical data structures and properties

Data collection principles - Thoughtful data collection is critical to learning from data

Sampling principles and strategies - Different ways to sample from a population

Experiments - Basic principles of experimental design

SECTION OVERVIEW

1.2: Introduction

1.2.1: A Classroom Story and an Inspiration

Several years ago, I was teaching an introductory Statistics course at De Anza College where I had several achieving students who were dedicated to learning the material and who frequently asked me questions during class and office hours. Like many students, they were able to understand the material on descriptive statistics and interpreting graphs. Unlike many introductory Statistics students, they had excellent math and computer skills and went on to master probability, random variables and the Central Limit Theorem.

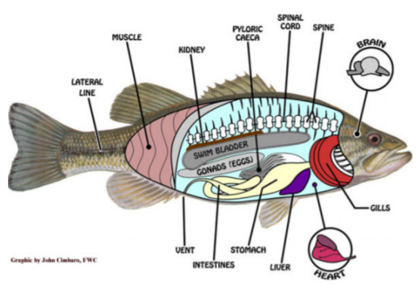
However, when the course turned to inference and hypothesis testing, I watched these students' performance deteriorate. One student asked me after class to again explain the difference between the Null and Alternative Hypotheses. I tried several methods, but it was clear these students never really understood the logic or the reasoning behind the procedure. These students could easily perform the calculations, but they had difficulty choosing the correct model, setting up the test, and stating the conclusion.

These students, (to their credit) continued to work hard; they wanted to understand the material, not simply pass the class. Since these students had excellent math skills, I went deeper into the explanation of Type II error and the statistical power function. Although they could compute power and sample size for different criteria, they still didn't conceptually understand hypothesis testing.

On my long drive home, I was listening to National Public Radio's Talk of the Nation[1] and heard discussion on the difference between the reductionist and holistic approaches to the sciences. The commentator described this as the Western tradition vs. the Eastern tradition. The reductionist or Western method of analyzing a problem, mechanism or phenomenon is to look at the component pieces of the system being studied. For example, a nutritionist breaks a potato down into vitamins, minerals, carbohydrates, fats, calories, fiber and proteins. Reductionist analysis is prevalent in all the sciences, including Inferential Statistics and Hypothesis Testing.

Holistic or Eastern tradition analysis is less concerned with the component parts of a problem, mechanism or phenomenon but rather with how this system operates as a whole, including its surrounding environment. For example, a holistic nutritionist would look at the potato in its environment: when it was eaten, with what other foods it was eaten, how it was grown, or how it was prepared. In holism, the potato is much more than the sum of its parts.

Consider these two renderings of fish:



The first image is a drawing of fish anatomy by John Cimbaro used by the La Crosse Fish Health Center.[2] This drawing tells us a lot about how a fish is constructed, and where its vital organs are located. There is much detail given to the scales, fins, mouth and eyes.

The second image is a watercolor by the Chinese artist Chen Zheng- Long[3] . In this artwork, we learn very little about fish anatomy since we can only see minimalistic eyes, scales and fins. However, the artist shows how fish are social creatures, how their fins move to swim and the type of plants they like. Unlike the first drawing, the drawing teaches us much more about the interaction of the fish in its surrounding environment and much less about how a fish is built.

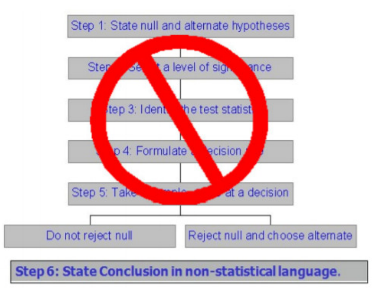
This illustrative example shows the difference between reductionist and holistic analyses. Each rendering teaches something important about the fish: the reductionist drawing of the fish anatomy helps explain how a fish is built and the holistic watercolor helps explain how a fish relates to its environment. Both the reductionist and holistic methods add to knowledge and understanding, and both philosophies are important. Unfortunately, much of Western science has been dominated by the reductionist philosophy, including the backbone of the scientific method, Inferential Statistics.

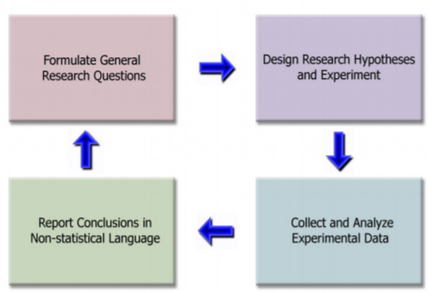
Although science has traditionally been reluctant, often hostile, to embrace or include holistic philosophy in the scientific method, there have been many who now support a multicultural or multi- philosophical approach. In his book Holism and Reductionism in Biology and Ecology[4], Looijen claims that "holism and reductionism should be seen as mutually dependent, and hence co-operating research programs than as conflicting views of nature or of relations between sciences." Holism develops the "macro-laws" that reductionism needs to "delve deeper" into understanding or explaining a concept or phenomena. I believe this claim applies to the study of Statistics as well.

I realize that the problem of my high-achieving students being unable to comprehend hypothesis testing could be cultural – these were international students who may have been schooled under a more holistic philosophy. The Introductory Statistics curriculum and most texts give an incomplete explanation of the logic of Hypothesis Testing, eliminating or barely explaining such topics as Power, the consequence of Type II error or Bayesian alternatives. The problem is how to supplement an Introductory Statistics course with a holistic philosophy without depriving the students of the required reductionist course curriculum – all in one quarter or semester!

I believe it is possible to teach the concept of Inferential Statistics holistically. This course material is a result of that inspiration, and it was designed to supplement, not replace, a traditional course textbook or workbook. This supplemental material includes:

- Examples of deriving research hypotheses from general questions and explanatory conclusions consistent with the general question and test results.
- An in-depth explanation of statistical power and type II error.
- Techniques for checking the validity of model assumptions and identifying potential outliers using graphs and summary statistics.
- Replacement of the traditional step-by-step "cookbook" for hypothesis testing with interrelated procedures.
- De-emphasis of algebraic calculations in favor of a conceptual understanding using computer software to perform tedious calculations.
- Interactive Flash animations to explain the Central Limit Theorem, inference, confidence intervals, and the general hypothesis testing model, which includes Type II error and power.
- PowerPoint Slides of the material for classroom demonstration.
- Excel Data sets for use with computer projects and labs.





This material is limited to one population hypothesis testing but could easily be extended to other models. My experience has been that once students understand the logic of hypothesis testing, the introduction of new models is a minor change in the procedure.

1.2.2: The Blind Man and the Elephant

This old story from China or India was made into the poem The Blind Man and the Elephant by John Godfrey Saxe[5]. Six blind men find excellent empirical evidence from different parts of the elephant and all come to reasoned inferences that match their observations. Their research is flawless and their conclusions are completely wrong, showing the necessity of including holistic analysis in the scientific process.

Here is the poem in its entirety:

> *It was six men of Indostan, to learning much inclined,*
>
> *who went to see the elephant (Though all of them were blind),*
>
> *that each by observation, might satisfy his mind.*
>
> *The first approached the elephant, and, happening to fall,*
>
> *against his broad and sturdy side,*
>
> *at once began to bawl: "God bless me! but the elephant, is nothing but a wall!"*
>
> *The second feeling of the tusk, cried: "Ho! what have we here,*
>
> *so very round and smooth and sharp? To me tis mighty clear,*
>
> *this wonder of an elephant, is very like a spear!"*
>
> *The third approached the animal, and, happening to take,*
>
> *the squirming trunk within his hands, "I see," quoth he,*
>
> *the elephant is very like a snake!"*
>
> *The fourth reached out his eager hand, and felt about the knee:*
>
> *"What most this wondrous beast is like, is mighty plain," quoth he;*
>
> *"Tis clear enough the elephant is very like a tree."*
>
> *The fifth, who chanced to touch the ear, Said; "E'en the blindest man*
>
> *can tell what this resembles most; Deny the fact who can,*
>
> *This marvel of an elephant, is very like a fan!"*
>
> *The sixth no sooner had begun, about the beast to grope,*
>
> *than, seizing on the swinging tail, that fell within his scope,*

"I see," quothe he, "the elephant is very like a rope!"

And so these men of Indostan, disputed loud and long,

each in his own opinion, exceeding stiff and strong,

Though each was partly in the right, and all were in the wrong!

So, oft in theologic wars, the disputants, I ween,

tread on in utter ignorance, of what each other mean,

and prate about the elephant, not one of them has seen!

-John Godfrey Saxe

1.2.3: What can go Wrong in Research - Two Stories

The first story is about a drug that was thought to be effective in research, but was pulled from the market when it was found to be ineffective in practice.

FDA Orders Trimethobenzamide Suppositories Off the market[6]

FDA today ordered makers of unapproved suppositories containing trimethobenzamide hydrochloride to stop manufacturing and distributing those products.

Companies that market the suppositories, according to FDA, are Bio Pharm, Dispensing Solutions, G&W Laboratories, Paddock Laboratories, and Perrigo New York. Bio Pharm also distributes the products, along with Major Pharmaceuticals, PDRX Pharmaceuticals, Physicians Total Care, Qualitest Pharmaceuticals, RedPharm, and Shire U.S. Manufacturing.

FDA had determined in January 1979 that trimethobenzamide suppositories lacked "substantial evidence of effectiveness" and proposed withdrawing approval of any NDA for the products.

"There's a variety of reasons" why it has taken FDA nearly 30 years to finally get the suppositories off the market, Levy said.

At least 21 infant deaths have been associated with unapproved carbinoxamine-containing products, Levy noted.

Many products with unapproved labeling may be included in widely used pharmaceutical reference materials, such as the *Physicians' Desk Reference*, and are sometimes advertised in medical journals, he said.

Regulators urged consumers using suppositories containing trimethobenzamide to contact their health care providers about the products.

The second story is about promising research that was abandoned because the test data showed no significant improvement for patients taking the drug.

Drug Found Ineffective Against Lung Disease[7]

Treatment with interferon gamma-1b (Ifn-g1b) does not improve survival in people with a fatal lung disease called idiopathic pulmonary fibrosis, according to a study that was halted early after no benefit to participants was found.

Previous research had suggested that Ifn-g1b might benefit people with idiopathic pulmonary fibrosis, particularly those with mild to moderate disease.

The new study included 826 people, ages 40 to 79, who lived in Europe and North America. They were given injections of either 200 micrograms of Ifn-g1b (551 people) or a placebo (275) three times a week.

After a median of 64 weeks, 15 percent of those in the Ifn-g1b group and 13 percent in the placebo group had died. Symptoms such as flu-like illness, fatigue, fever and chills were more common among those in the Ifn-g1b group than in the placebo group. The two groups had similar rates of serious side effects, the researchers found.

"We cannot recommend treatment with interferon gamma-1b since the drug did not improve survival for patients with idiopathic pulmonary fibrosis, which refutes previous findings from subgroup analyses of survival in studies of patients with mild-to-moderate physiological impairment of pulmonary function," Dr. Talmadge E. King Jr., of the University of California, San Francisco, and colleagues wrote in the study published online and in an upcoming print issue of *The Lancet*.

The negative findings of this study "should be regarded as definite, [but] they should not discourage patients to participate in one of the several clinical trials currently underway to find effective treatments for this devastating disease," Dr. Demosthenes Bouros, of the Democritus University of Thrace in Greece, wrote in an accompanying editorial.

Bouros added that people deemed suitable "should be enrolled early in the transplantation list, which is today the only mode of treatment that prolongs survival."

Although these are both stories of failures in using drugs to treat diseases, they represent two different aspects of hypothesis testing. In the first story, the suppositories were thought to be effective in treatment from the initial trials, but were later shown to be ineffective in the general population. This is an example of what statisticians call **Type I Error**: supporting a hypothesis (the suppositories are effective) that later turns out to be false.

In the second story, researchers chose to abandon research when the interferon was found to be ineffective in treating lung disease during clinical trials. Now this may have been the correct decision, but what if this treatment was truly effective and the researchers just had an unusual group of test subjects? This would be an example of what statisticians call **Type II Error**: failing to support a hypothesis (the interferon is effective) that later turns out to be true. Unlike the first story, the second story will never result in answer to this question since the treatment will not be released to the general public.

In a traditional Introductory Statistics course, very little time is spent analyzing the potential error shown in the second story. However, both types of error are important and will be explored in this course material.

Preliminary Results – bringing the holistic approach to the entire statistics curriculum.

After writing what are now chapters 8, 9 and 10, I decided to use this holistic approach in several of my courses. I found students were more engaged in the course, were able to understand the logic of hypothesis testing, and would state the appropriate conclusion. I wanted to bring this approach to the entire statistics course and this book is the result.

SECTION OVERVIEW

1.3: Displaying and Analyzing Data with Graphs

1.3.1: Introduction and Examples

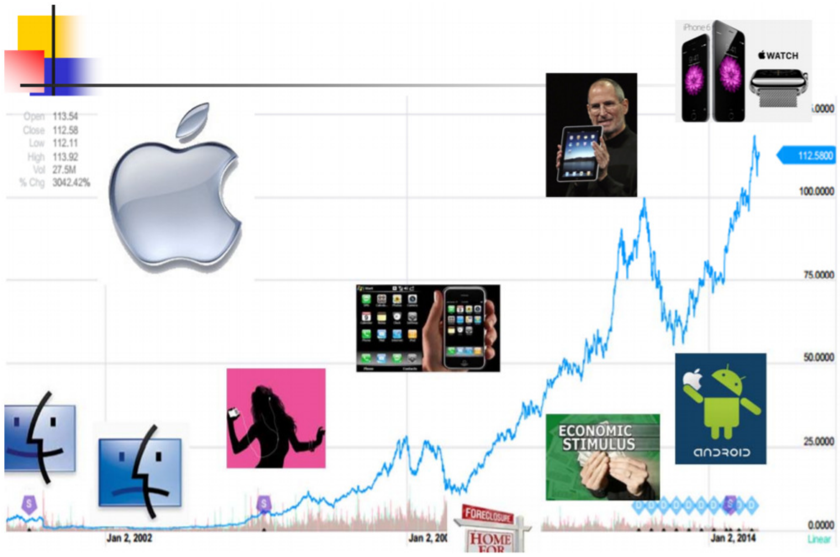
In statistics, we organize data into graphs, which (when properly created) are powerful tools to help us understand, interpret and analyze the phenomena we study.

Here is an example of raw data, the month closing stock price (adjusted for splits) of Apple Inc. from December 1999 to December 2016[9]:

115.82	102.97	106.17	75.50	69.86	52.70	41.97	27.42	11.11	25.77	11.04	9.35	4.19	1.39	0.93	1.42	0.97
110.52	115.73	114.39	74.83	76.83	49.73	40.49	26.01	12.06	23.71	11.93	8.82	4.36	1.36	1.01	1.39	1.07
112.96	116.40	103.43	69.93	77.80	52.67	39.16	24.53	14.00	24.72	10.55	7.49	3.41	1.49	1.05	1.14	1.27
112.47	107.44	96.49	63.79	87.18	49.62	36.92	24.12	14.79	19.97	10.02	6.98	2.52	1.35	0.94	1.01	1.68
105.56	109.84	98.16	65.19	86.93	50.07	31.63	21.89	22.06	18.02	8.83	6.10	2.24	1.47	0.96	1.21	3.96
103.12	117.62	91.10	60.15	79.47	50.81	33.47	21.26	20.68	17.14	8.84	5.55	2.10	1.37	0.99	1.22	3.31
94.60	121.63	88.56	52.71	75.99	43.68	32.73	18.53	21.79	15.88	7.45	4.79	2.12	1.24	1.15	1.51	3.41
98.81	126.33	86.17	59.78	75.17	45.26	33.43	17.67	24.56	15.77	7.78	5.17	1.83	1.17	1.52	1.30	2.73
92.20	120.85	79.89	58.47	75.99	45.56	33.97	16.37	22.63	12.99	9.16	4.69	1.68	0.93	1.58	1.66	4.04
107.20	120.15	72.66	58.45	78.01	45.35	30.58	13.68	18.67	12.09	8.16	5.42	1.76	0.92	1.54	1.44	4.42
95.10	124.05	71.24	58.28	70.58	45.96	26.63	11.62	16.27	11.01	8.91	5.84	1.56	0.98	1.41	1.19	3.73
95.22	112.69	67.37	59.80	59.40	44.15	24.99	11.73	17.61	11.16	9.83	5.00	1.47	0.93	1.61	1.41	3.38

Most people would look at this data and be unable to analyze or interpret what has happened at Apple. However a simple line graph over time is much easier to understand:

The line graph tells the story of Apple, from the dot.com crash in 2000, to the introduction of the first IPod in 2005, the first smart phone in 2007, the economic collapse of 2008, and competition from other operating systems, such as Android:

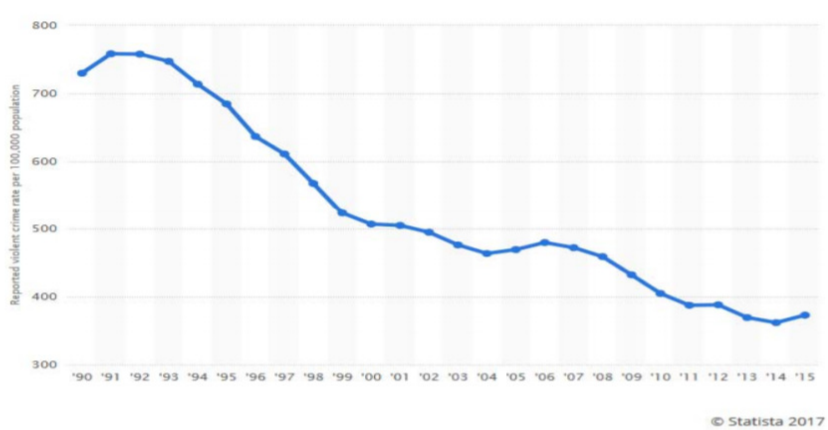


Graphs can help separate perception from reality. The polling organization Gallup has annually asked the question "Is there more crime in the U.S. then there was a year ago, or less?" In virtually every poll done, a large majority has said that crime has gone up.[10]

Is there more crime in the U.S. than there was a year ago, or less?

	More	Less	Same (vol.)	No opinion
	%	%	%	%
2016 Oct 5-9	70	20	6	4
2015 Oct 7-11	70	18	8	4
2014 Oct 12-15	63	21	9	7
2013 Oct 3-6	64	19	9	7
2011 Oct 6-9	68	17	8	8
2010 Oct 7-10	66	17	8	9
2009 Oct 1-4	74	15	6	5
2008 Oct 3-5	67	15	9	9
2007 Oct 4-7	71	14	8	6
2006 Oct 9-12	68	16	8	8
2005 Oct 13-16	67	21	9	3
2004 Oct 11-14	53	28	14	5
2003 Oct 6-8	60	25	11	4
2002 Oct 14-17	62	21	11	6
2001 Oct 11-14	41	43	10	6
2000 Aug 29-Sep 5	47	41	7	5
1998 Oct 23-25	52	35	8	5
1997 Aug 22-25	64	25	6	5
1996 Jul 25-28	71	15	8	6
1993 Oct 13-18	87	4	5	4
1992 Feb 28-Mar 1	89	3	4	4
1990 Sep 10	84	3	7	6

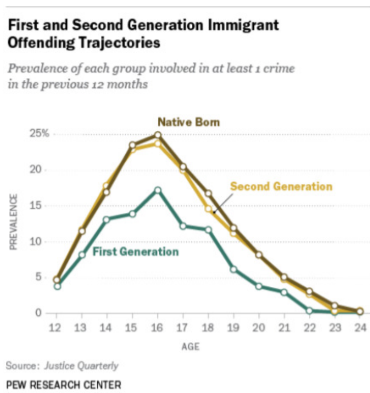
However, actual data from the U.S. Justice Department shows that violent crime rates have actually decreased in almost every since 1990.[11]

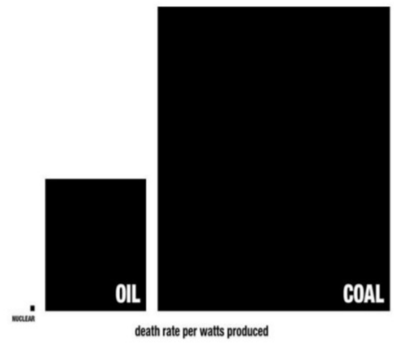


Perhaps people are influenced by stories in the news, which may sensationalize crime, but here is an example of where we can use statistics to challenge these false perceptions.

Here are two other examples of graphs of data. Make your own interpretation:

- Pew Research conducted a study in 2013 on how First Generation immigrant crime rates compares with second generation and native born Americans.[12]
- The Next Big Future conducted a study comparing deaths caused by creating energy from different sources: coal, oil and nuclear.[13]

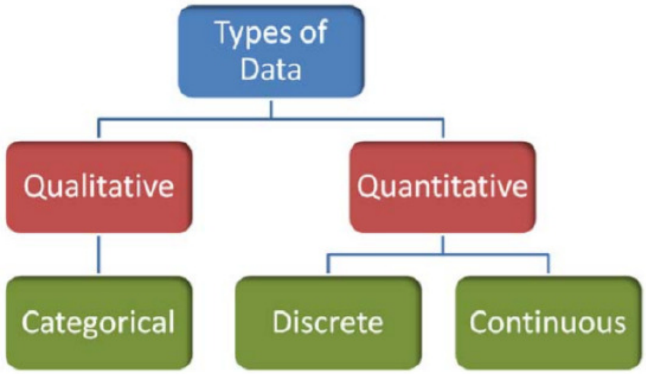




1.3.2: Types of Data

In Statistics, two important concepts are the **population** and the **sample**. If we are collecting data, the population refers to all data for the phenomena that is being studied, while the sample refers to a subset of that data. In statistics, we are almost always analyzing sample data. These concepts will be explored in greater detail in Chapter 3. Until then, we will work with only sample data.

Sample data is a collection of information taken from a population for the purpose of analysis.



Quantitative data are measurements and numeric quantities that can be determined from the data. When describing quantitative data, we can look at the center, spread, shape and unusual features.

Qualitative data are non-numeric values that describe the data. Note that all quantitative data is numeric but some numbers without quantity (such as Zip Code or Social Security Number) are qualitative. When describing categorical data, we are limited to observing counts in each group and comparing the differences in percentages.

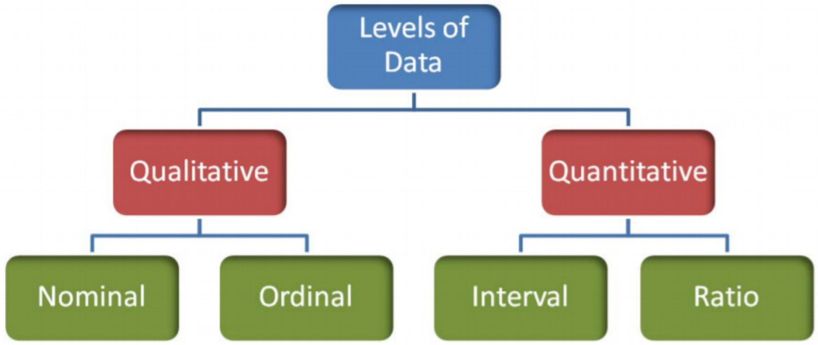
Categorical data are non-numeric values. Some examples of categorical data include eye color, gender, model of computer, and city.

Discrete data are quantitative natural numbers (0, 1, 2, 3, ...). Some examples of discrete data include number of siblings, friends on Facebook, bedrooms in a house. Discrete data are values that are counted, or answers to the question "How many?"

Continuous data are quantitative based on the real numbers. Some examples of continuous data include time to complete an exam, height, and weight. Continuous data are values that are measured, or answers to the question "How much?"

1.3.3: Levels of Data

Data can also be organized into four levels of data, Nominal, Ordinal, Interval and Ratio



Nominal Data are qualitative data that only define attributes, not hierarchal ranking. Examples of nominal data include hair color, ethnicity, gender and any yes/no question.

Ordinal Data are qualitative data that define attributes with a hierarchal ranking. Examples of nominal data include movie rating (G, PG, PG13, R, NC17), T-shirt size (S, M L, XL), or your letter grade on a term paper.

The difference between Nominal and Ordinal data is that Ordinal data can be ranked, while Nominal data are just labels.

Interval Data are quantitative data that have meaningful distance between values, but do not have a "true" zero. Interval data are numeric, but zero is just a place holder. Examples of interval data include temperature in degrees Celsius, and year of birth.

Ratio Data are quantitative data that have meaningful distance between values, and have a "true" zero. Examples of ratio data include time it takes to drive to work, weight, height, and number of children in a family. Most numeric data will be ratio.

One way to tell the difference between Interval and Ratio data is to look if zero has the same value under all possible units. For example zero degrees Celsius is not the same as zero degrees Fahrenheit, so temperature has no true zero. But zero minutes, zero days, zero months all mean the same thing, since for time zero means "no time."

1.3.4: Graphs of Categorical Data

When describing categorical data with graphs, we want to be able to visualize the difference in proportions or percentages within each group. These values are also known as relative frequencies.

Definition: Relative Frequency

n = sample size - The number of observations in your sample size.

Frequency - the number of times a particular value is observed.

Relative frequency - The proportion or percentage of times a particular value is observed.

Relative Frequency = Frequency / n

Example: One categorical variable - marital status

A sample of 500 adults (aged 18 and over) from Santa Clara County, California was taken from the year 2000 United States Census.[14] The results are displayed in the table:

Marital Status	Frequency	Relative Frequency
Married	270	270/500 = 0.540 or 54.0%
Widowed	22	22/500 = 0.044 or 4.4%
Divorced - not remarried	42	42/500 = 0.084 or 8.4%
Separated	10	10/500 = 0.020 or 2.0%
Single - never married	156	156/500 = 0.312 or 31.2%
Total	**500**	**500/500 = 1.000 or 100.0%**

Solution

Analysis - over half of the sampled adults were reported as married. The smallest group was separate which represented only 2% of the sample.

Example: Comparing two categorical variables - presidential approval and gender

Reuters/Ipsos conducts a daily tracking poll of American adults to assess support of the president of the United States. Here are the results of a tracking poll ending August 17, 2017, which includes data from the five days on which Donald Trump made several highly controversial statements regarding violence following a gathering of neo-Nazis and white supremacists in Charlottesville, Virginia. The question is "Overall, do you approve or disapprove of the way Donald Trump is handling his job as president?"[15]

	Female Frequency	Male Frequency	Female Relative Frequency	Male Relative Frequency
Approve	392	404	0.295 or 29.5%	0.400 or 40.0%
Disapprove	846	545	0.634 or 63.4%	0.541 or 54.1%
Unsure/No Opinion	96	59	0.079 or 7.9%	0.059 or 5.9%
Total	**1334**	**1008**	**1.000 or 100%**	**1.000 or 100%**

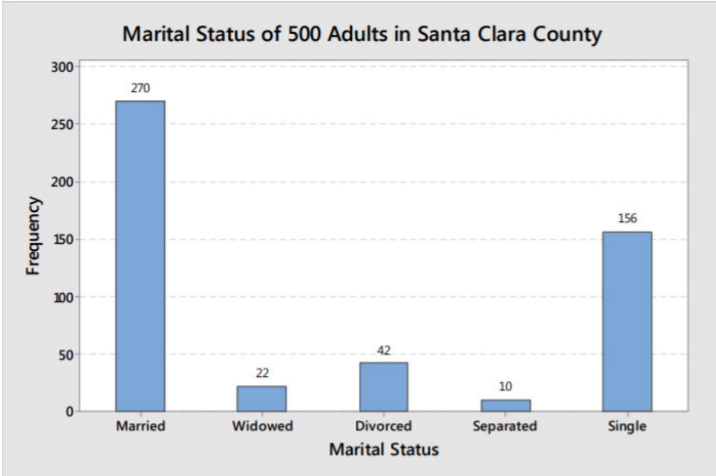
Solution

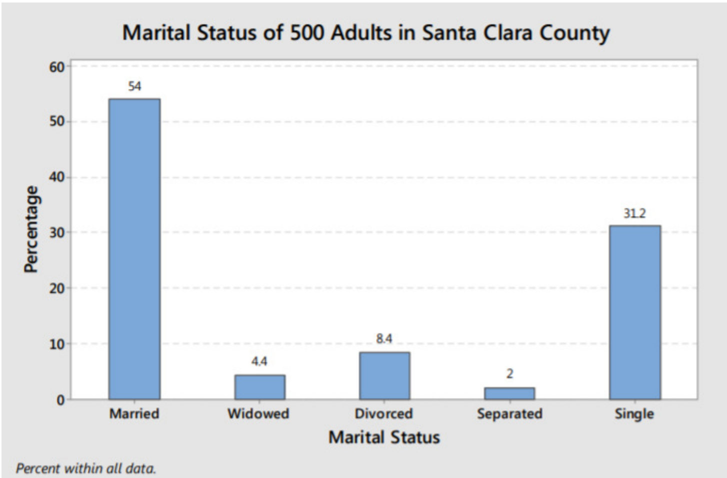
Analysis – Both men and women disapproved of the way Donald Trump was handling his job as president on the date of the poll. Women had a higher disapproval rate than men. In political science, this is called a gender gap.

Bar Graphs

One way to represent categorical data is on a bar graph, where the height of the bar can represent the frequency or relative frequency of each choice.

The graphs below represent the marital status information from the one categorical example. The vertical axis on the first graph shows frequencies for each group, while the second graph shows the relative frequencies (shown here as percentages).

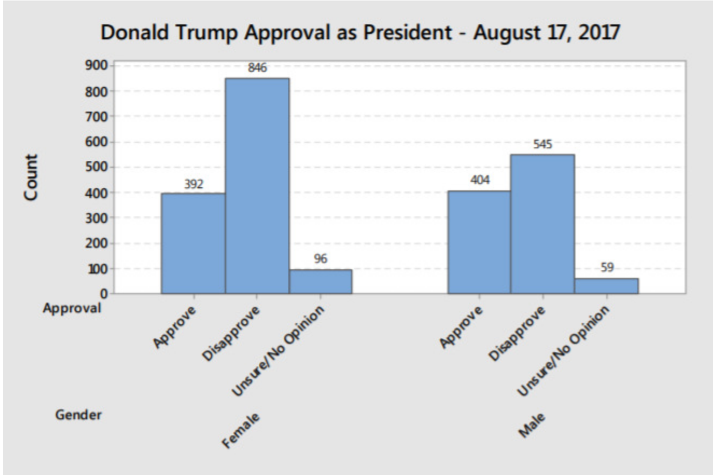




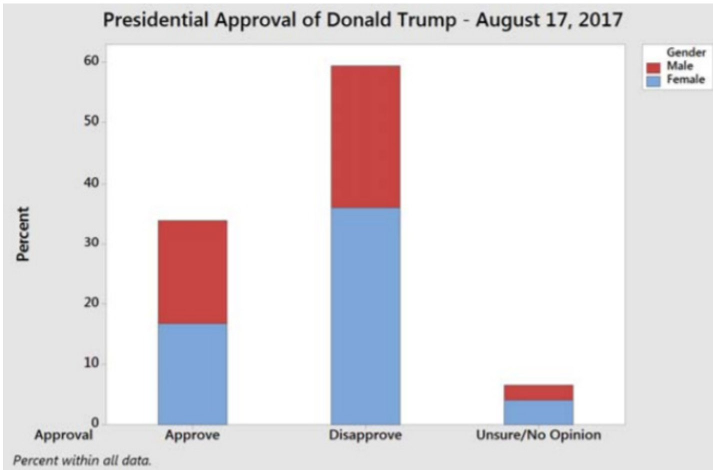
There is no difference in the shape of each graph as the percentage or frequency in each group is directly proportional to the area of each bar.

In either case, we can make the same analysis, that married and single are the most frequently occurring marital statuses.

A **clustered bar** graph can be used to compare categorical variables, such as the presidential approval poll cross- tabulated by gender. You can see in this graph that women have a much stronger disapproval of Trump than men do. In this graph, the vertical axis is frequency, but you could also make the vertical axis relative frequency or percentage.

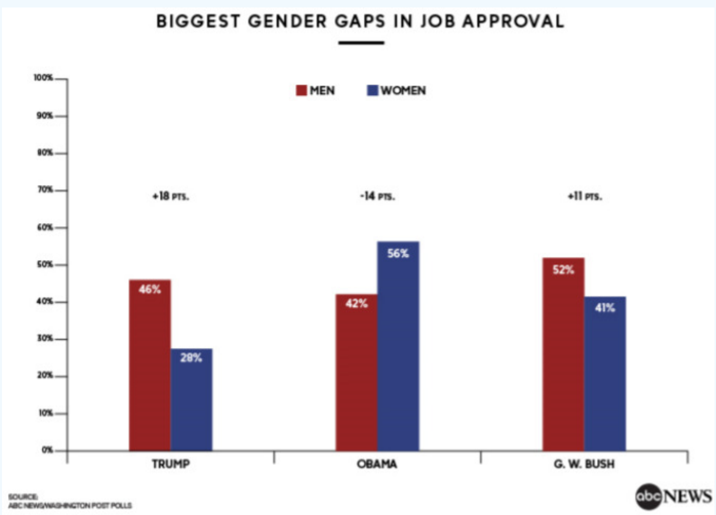


Another way of representing the same data is a **stacked bar graph**, shown here with percentage (relative frequency) as the vertical axis. It is harder to see the difference between men and women, but the total approval/disapproval percentages are easier to read.



Example: Historic gender gaps

Here is another clustered bar graph reported by ABC News, August 21, showing that Trump had a larger gender gap than the two prior presidents, Barack Obama and George W. Bush.[16]

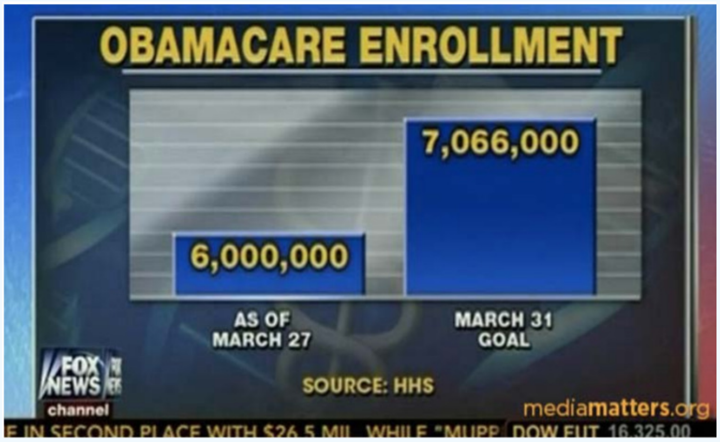


In conclusion, bar graphs are an excellent way to display, analyze and compare categorical data. However, care must be taken to not create misleading graphs.

Example: Misreported Affordable Care Act enrollment

Here is an example of a bar graph reported on the Fox News Channel that distorted the truth about people signing up for the Affordable Care Act (ACA) in 2014, as reported by mediamatters.org[17]

> On March 27 health insurance enrollment through the ACA's exchanges surpassed 6 million, exceeding the revised estimate of enrollees for the program's first year before the March 31 open enrollment deadline. Enrollment appears on track to hit the Congressional Budget Office's initial estimate of 7 million sign-ups, and taking Medicaid enrollees into account, the ACA will have reportedly extended health care coverage to at least 9.5 million previously uninsured individuals.
>
> Fox celebrated the final day of open enrollment by attempting to somehow twist the recent enrollment surge into bad news for the law.
>
> *America's Newsroom* aired an extremely skewed bar chart which made it appear that the 6 million enrollees comprised roughly one-third of the 7 million enrollee goal:



At first look, the graph seemingly shows that the ACA enrollment was well below the projected goal. The graph is misleading for three reasons:

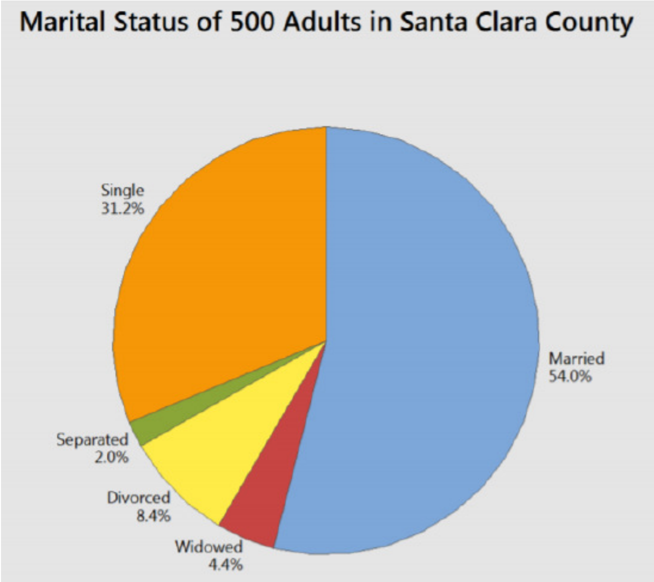
1. The vertical axis doesn't start at zero enrollees, greatly overstating the difference between the two numbers.
2. The graph of the "6,000,000" enrolled failed to include new enrollees in Medicaid, which was part of the "March 31 Goal."
3. The reported enrollment was 4 days before the deadline. Like students doing their homework, many people waited until the last day to enroll.

The actual enrollment numbers far exceeded the goal, the exact opposite of this poorly constructed bar graph.

Pie Charts

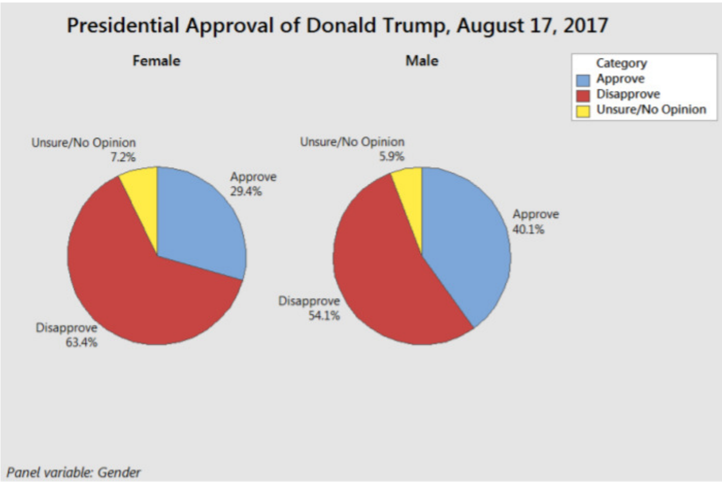
Another way to represent categorical data is a pie chart, in which each slice of the pie represents the relative frequency or percentage of data in each category.

The pie chart shown here represents the marital status of 500 adults in Santa Clara County taken from the 200 census, the same data that was represented by a bar graph in a previous example.



The analysis again shows that most people are married, followed by single.

A **multiple pie chart** can be used to compare the effect of one categorical variable on another.



In the presidential approval poll example, a higher percentage of female adults disapprove of Donald Trump's performance as U.S. President compared to male adults. This is comparable to stacked or clustered bar graphs shown in the prior example.

1.3.5: Graphs of Numeric Data

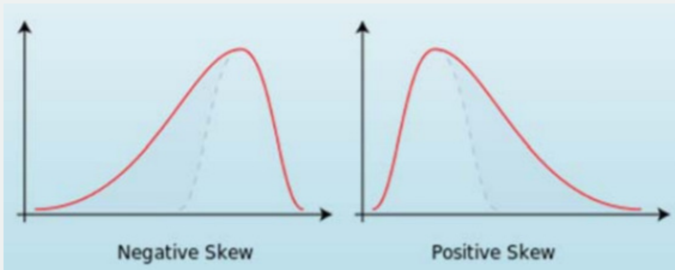
Numeric data is treated differently from categorical data as there exists quantifiable differences in the data values. In analyzing quantitative data, we can describe quantifiable features such as the center, the spread, the shape or skewness[18], and any unusual features (such as outliers).

Interpreting and Describing Numeric Data

Center – Where is the middle of the data, what value would represent the average or typical value?

Spread- How much variability is there in the data? What is range of the data? (range is highest value – lowest value.)

Shape – Are the data values symmetric or is it skewed positive or negative? Are the values clustered toward the center, evenly spread, or clustered towards the extreme values?



Unusual Features – Are there outliers (values that are far removed from the bulk of the data?)

Example: Students browsing the web

This data represents how much time 30 students spent on a web browser (on the Internet) in a 24 hour period.[19]

Data is rounded to the nearest minute.

102	104	85	67	101
71	116	107	99	82
103	97	105	103	95
105	99	86	87	100
109	108	118	87	125
124	112	122	78	92

This data set is continuous, ratio, quantitative data, even though times are rounded to the nearest integer. Sample data presented unsorted in this format are sometimes called **raw data**.

Not much can be understood by looking simply at raw data, so we want to make appropriate graphs to help us conduct preliminary analysis.

1.3.5.1: Stem and Leaf Plots

A stem and leaf plot is a method of tabulating the data to make it easy to interpret. Each data value is split into a "stem" (the first digit or digits) and a "leaf" (the last digit, usually). For example, the stem for 102 minutes would be 10 and the leaf would be 2.

The stems are then written on the left side of the graph and all corresponding leaves are written to the right of each matching stem.

Stem and Leaf Graph

number = stem|leaf
102 = 10|2
71 = 7|1

6	7
7	18
8	25677
9	25799
10	01233455789
11	268
12	245

The stem and leaf plot allows us to do some preliminary analysis of the data. The **center** is around 100 minutes. The **spread** between the highest and lowest numbers is 58 minutes. The **shape** is not symmetric since the data is more spread out towards the lower numbers. In statistics, this is called skewness and we would call this data **negatively skewed**.

Stem and leaf plots can also be used to compare similar data from two groups in a back-to-back format.

In a back-to-back stem and leaf plot, each group would share a common stem and leaves would be written for each group to the left and right of the stem.

Example: Comparing two airlines' passenger loading times

The data shown represents the passenger boarding time (in minutes) for a sample of 16 airplanes each for two different airlines.

Airline A	11, 14, 16, 17, 19, 21, 22, 23, 24, 24, 24, 26, 31, 32, 38, 39
Airline B	8, 11, 13, 14, 15, 16, 16, 18, 19, 19, 21, 21, 22, 24, 26, 31

Airline A will be represented on the left side of the stem, while Airline B will be represented on the right. Instead of using the last digit as the leaf (each row representing 10 minutes), we are instead going to let each row represent 5 minutes. This will allow us to better see the shape of the data.

The center for Airline B is about 5 minutes lower than Airline A. The spread for each airline is about the same. Airline A shape seems slightly skewed towards positive values (skewed positive) while Airline B times are somewhat symmetric.

	0	
	0	8
1 4	1	1 3 4
6 7 9	1	5 6 6 8 9 9
1 2 3 4 4 4	2	1 1 2 4
6	2	6
1 2	3	1
8 9	3	

1.3.5.2: Dot Plots

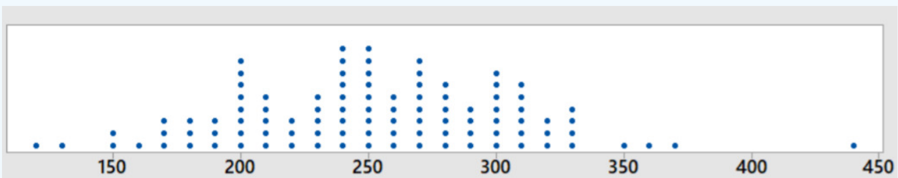
A dot plot represents each value of a data set as a dot on a simple numeric scale. Multiple values are stacked to create a shape for the data. If the data set is large, each dot can represent multiple values of the data.

Example: Weights of apples

A Chilean agricultural researcher collected a sample of 100 Royal Gala apples.[20] The weight of each apple (reported in grams) is shown in the table below:

228	272	196	435	195	242	265	330	298	248
320	189	278	261	203	282	246	203	274	231
282	311	275	297	194	183	308	245	185	260
235	149	312	274	218	307	324	256	203	206
310	182	245	167	297	276	248	262	327	292
287	118	265	235	246	310	200	289	299	230
237	205	164	231	133	222	326	353	252	237
214	274	253	197	244	209	236	290	296	272
315	173	224	202	246	363	299	325	151	242
170	261	270	284	365	213	184	240	302	233

Here is the data organized into a dot plot, in which each dot represents one apple. The scaling of the horizontal axis rounds each apple's weight to the nearest 10 grams.



The center of the data is about 250, meaning that a typical apple would weight about 250 grams. The range of weights is between 110 and 440 grams, although the 440 gram apple is an outlier, an unusually large apple. The next highest weight is only 370 grams. Not counting the outlier, the data is symmetric and clustered towards the center.

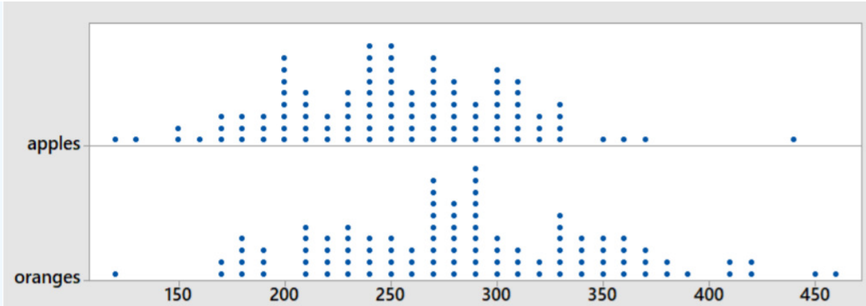
Dot plots can also be used to compare multiple populations.

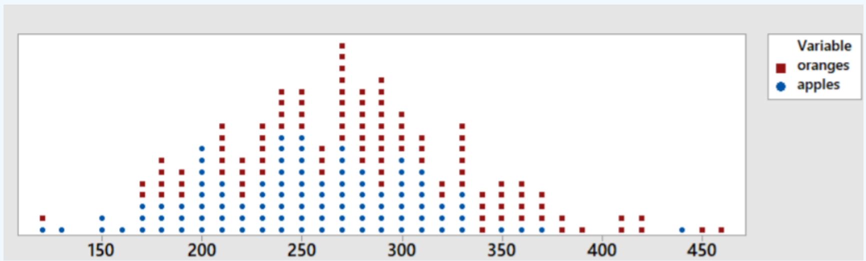
Example: Dot plots can also be used to compare multiple populations

The Chilean agricultural researcher collected a sample of 100 navel oranges[21] and recorded the weight of each orange in grams.

332	298	342	287	392	358	279	165	289	329
265	233	192	214	286	221	381	277	317	285
273	410	419	292	288	283	181	348	356	330
248	245	366	212	458	424	342	208	122	184
285	360	277	363	324	336	230	327	218	237
305	290	249	166	244	273	218	177	277	279
274	194	379	409	286	272	261	306	330	239
350	447	284	304	267	225	193	223	334	264
288	273	229	305	257	342	209	295	238	233
365	348	253	352	304	266	273	372	181	208

We can now add the weights of the oranges to the dot plot of the apple weights made in the prior example. The first chart keeps apples and oranges in separate graphs while the second chart combines data with a different marker for apples and orange. This second chart is called a **stacked dot plot**.





From the graphs, we can see that the typical orange weighs about 30 grams more than the typical apple. The spread of weights for apples and oranges is about the same. The shapes of both graphs are symmetric and clustered towards the center. There is a high outlier for apples at 440 grams and a low outlier for oranges at 120 grams.

1.3.5.3: Grouping Numeric Data

Another way to organize raw data is to group them into **class intervals**, and to then create a **frequency distribution** of these class intervals.

There are many methods of creating class intervals, so we will simply focus on creating intervals of equal width.

How to create class intervals of equal width and a frequency distribution

1. Choose how many intervals you want. Best is between 5 and 15 intervals.
2. Determine the interval width using the formula and rounding UP to a convenient value:

$$\text{IW} = \text{Interval Width} = \frac{\text{Maximum Value - Minimum Value} + 1}{\text{Number of Intervals}}$$

3. Create the **class intervals** starting with the minimum value:

   Min to under Min + IW,

   Min +IW to under Min +2(IW), ...

4. Calculate the **frequency** of each class interval by counting the values in each class interval. Values that are on an endpoint should be put in the lower class interval. This result is called a **frequency distribution**.

Example: Students browsing the web

Let's return to the data that represents how much time 30 students spent on a web browser in a 24 hour period. Data is rounded to the nearest minute.

102	104	85	67	101
71	116	107	99	82
103	97	105	103	95
105	99	86	87	100
109	108	118	87	125
124	112	122	78	92

First we choose how many class intervals. In this example, we will create 5 class intervals.

Next Determine the Class Interval Width and round up to a convenient value.

$$\text{IW} = \frac{125 - 67 + 1}{5} = 11.8 \rightarrow 12$$

Now create class intervals of width 12, starting with the lowest value, 67.

$$(67 \text{ to } 79) \quad (79 \text{ to } 91) \quad (91 \text{ to } 103) \quad (103 \text{ to } 115) \quad (115 \text{ to } 127)$$

Now, create a frequency distribution, by counting how many are in each interval. Values that are on an endpoint should be put in the higher class interval. For example, 103 should be counted in the interval (103 to 115):

Class Interval	Frequency
67 to 79	3
79 to 91	5
91 to 103	8
103 to 115	9
115 to 127	5
Total	30

As we did with categorical data, we can define **Relative Frequency** as the proportion or percentage of values in any Class Interval.

n = sample size - The number of observations in your sample size.

Frequency - the number of times a particular value is observed in a class interval.

Relative frequency - The proportion or percentage of times a particular value is observed in a class interval.

Relative Frequency = Frequency / n

Class Interval	Frequency	Relative Frequency
67 to 79	3	0.100 or 10.0%
79 to 91	5	0.167 or 16.7%
91 to 103	8	0.266 or 26.6%
103 to 115	9	0.300 or 30.0%
115 to 127	5	0.167 or 16.7%
Total	30	**1.000 or 100%**

Note that the value for the (91 to 103) class interval was deliberately rounded down to make the totals add up to exactly 100%

From the frequency distribution, we can see that 30% of the students are on the internet between 103 and 115 minutes per day, while only 10% of students are on the internet between 67 and 79 minutes.

Example: Comparing weights of apples and oranges

A Chilean agricultural researcher collected a sample of 100 Royal Gala apples and 100 navel oranges and measured their weights in grams (see previous example on dot plots).

We will start with a value of 100 and make the interval width equal to 30. Using the tally feature of Minitab, we can create a frequency distribution for the two fruits. Minitab uses “Count” for “Frequency” and reports “Percent” for “Relative Frequency”

Class interval	Apples Count	Apples Percent	Oranges Count	Oranges Percent
100 to 130	1	1.00	1	1.00
130 to 160	3	3.00	0	0.00
160 to 190	9	9.00	6	6.00
190 to 220	15	15.00	10	10.00
220 to 250	23	23.00	14	14.00
250 to 280	18	18.00	18	18.00
280 to 310	16	16.00	19	19.00
310 to 340	11	11.00	9	9.00
340 to 370	3	3.00	13	13.00
370 to 400	0	0.00	4	4.00
400 to 430	0	0.00	4	4.00
430 to 460	1	1.00	2	2.00
Totals	**100**	**100.00**	**100**	**100.00**

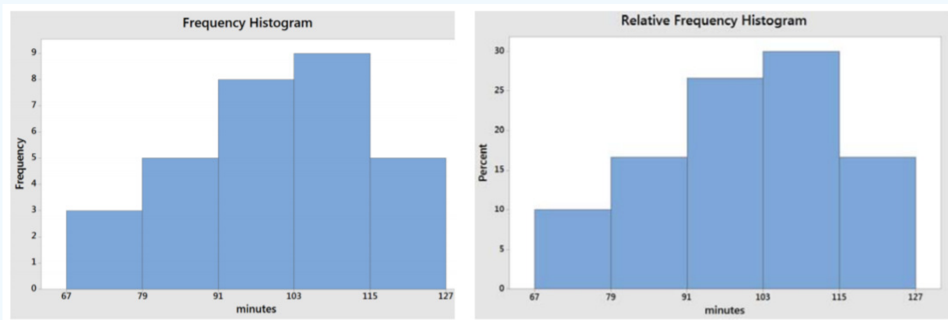
The most frequently occurring interval for apples is 220 to 250 grams while the most frequently occurring interval for oranges is 280 to 310 grams. Notice that there are some intervals with 0 observations, showing a potential high outlier for apples and a low outlier for oranges.

1.3.5.4: Histograms

A histogram is a graph of grouped rectangles where the vertical axis is frequency or relative frequency and the horizontal axis show the endpoints of the class intervals. The area of each rectangle is proportional to the frequency or relative frequency of the class interval represented.

Example: Students browsing the web

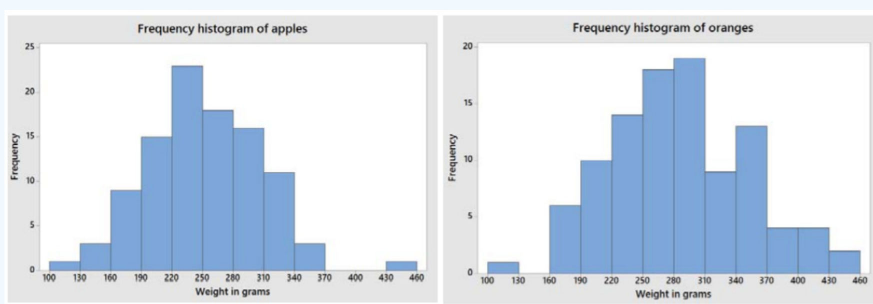
In the earlier example of 30 students browsing the web, we made 5 class intervals of the data. Here histograms represent frequency in the first graph and relative frequency in the second graph. Note that the shape of each graph is identical; all that is different is the scaling of the vertical axis.



Like the stem and leaf diagram, the histogram allows us to interpret and analyze the data. The **center** is around 100 minutes. The **spread** between the highest and lowest numbers is about 60 minutes. The **shape** is slightly **skewed negative**. The data clusters towards the center and there doesn't seem to be any unusual features like outliers.

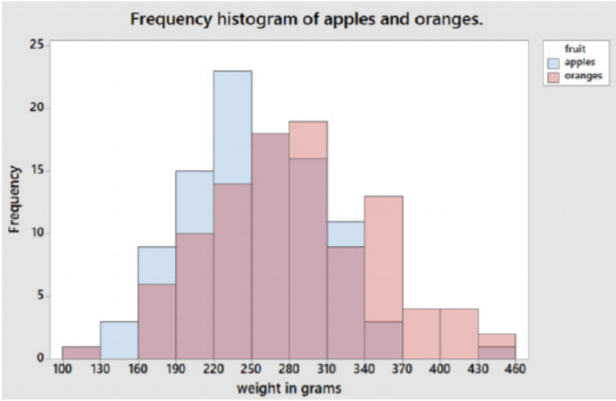
Example: Comparing weights of apples and oranges

First let's make a histogram of apples and oranges separately



For the apples, the center is around 250 grams and for the oranges the center is around 280 grams, meaning the oranges appear slightly heavier. For both apples and oranges, the range is about 360 grams from the minimum to the maximum values. Both graphs seem approximately symmetric. The apples have one value that is unusually high, and the oranges have one value that is unusually low.

Another way of comparing apples and oranges is to combine them into a single graph, also called a **grouped histogram**.



Here, the histograms are laid on top of each other, the light blue and purple match the histogram of apples and the light red and purple match the histogram of the oranges. Here is easier to see that oranges, in general, weigh more than apples.

1.3.5.5: Cumulative Frequency and Relative Frequency

The cumulative frequency of a class interval is the count of all data values less than the right endpoint. The cumulative relative frequency of a class interval is the cumulative frequency divided by the sample size.

Definition: Cumulative Relative Frequency

n = sample size - The number of observations in your sample size.

Cumulative Frequency - the number of times a particular value is observed in a class interval or in any lower class interval.

Cumulative Relative Frequency - The proportion or percentage of times a particular value is observed in a class interval or in any lower class interval.

Cumulative Relative Frequency = Cumulative Frequency / n

Example: Students browsing the web

Let's again return to the data that represents how much time 30 students spent on a web browser in a 24 hour period. Data is rounded to the nearest minute. Earlier we had made a frequency distribution and so we will now add columns for cumulative frequency and cumulative relative frequency.

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
67 to 79	3	0.100 or 10.0%	3	0.100 or 10.0%
79 to 91	5	0.167 or 16.7%	8	0.267 or 26.7%
91 to 103	8	0.266 or 26.6%	16	0.533 or 53.3%
103 to 115	9	0.300 or 30.0%	25	0.833 or 83.3%
115 to 127	5	0.167 or 16.7%	30	1.000 or 100%
Total	30	**1.000 or 100%**		

Note that the last class interval will always have a cumulative relative frequency of 100% of the data.

Some possible ways to interpret cumulative relative frequency: 83.3% of the students are on the internet less that 115 minutes.

The middle value (median) of the data occurs in the interval 91 to 103 minutes since 53.3% of the students are on the internet less than 103 minutes.

Example: Comparing weights of apples and oranges

The tally feature of Minitab can also be used to find cumulative relative frequencies (called cumulative counts and percentages here):

class interval	Apples Count	Apples Percent	Apples CumCnt	Apples CumPct	Oranges Count	Oranges Percent	Oranges CumCnt	Oranges CumPct
100 to 130	1	1.00	1	1.00	1	1.00	1	1.00
130 to 160	3	3.00	4	4.00	0	0.00	1	1.00
160 to 190	9	9.00	13	13.00	6	6.00	7	7.00
190 to 220	15	15.00	28	28.00	10	10.00	17	17.00
220 to 250	23	23.00	51	51.00	14	14.00	31	31.00
250 to 280	18	18.00	69	69.00	18	18.00	49	49.00
280 to 310	16	16.00	85	85.00	19	19.00	68	68.00
310 to 340	11	11.00	96	96.00	9	9.00	77	77.00
340 to 370	3	3.00	99	99.00	13	13.00	90	90.00
370 to 400	0	0.00	99	99.00	4	4.00	94	94.00
400 to 430	0	0.00	99	99.00	4	4.00	98	98.00
430 to 460	1	1.00	100	100.00	2	2.00	100	100.00
Totals	100	100.00			100	100.00		

Cumulative relative frequency can also be used to find percentiles of quantitative data. A **percentile** is the value of the data below which a given percentage of the data fall.

In our example 280 grams would represent the 69th percentile for apples since 69% of apples have weights lower than 280 grams. The 68th percentile for oranges would be 310 grams since 68% of oranges weigh less than 310 grams.

1.3.5.6: Using Ogives to find Percentiles

The table of cumulative relative frequencies can be used to find percentiles for the endpoints. One method of estimating other percentiles of the data is by creating a special graph of cumulative relative frequencies, called an **Ogive**.

An Ogive is a line graph where the vertical axis is cumulative relative frequency and the horizontal axis is the value of the data, specifically the endpoints of the class intervals. The left end point of the first class interval will have a cumulative relative frequency of zero. All other endpoints are given the right endpoint of the corresponding class interval. The points are then connected by line segments.

The graph can then be read to find any percentile desired. For example, the 25th, 50th and 75th percentiles break the data into equal fourths and are called **quartiles**.

Definition: Percentile

Percentile - the value of the data below which a given percentage of the data fall.

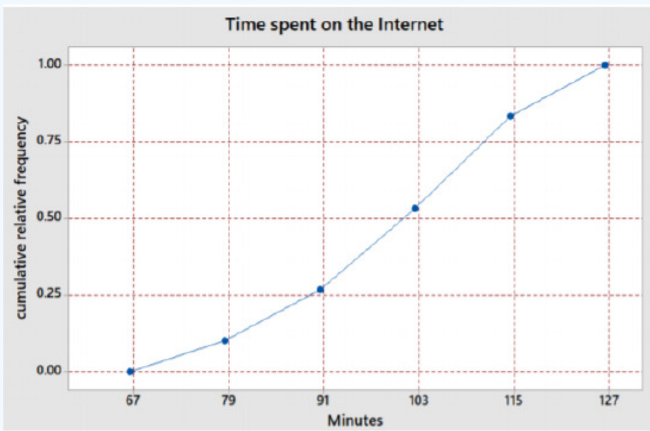
The 25th percentile is also known as the 1st **Quartile**.

The 50th percentile is also known as the 2nd **Quartile** or **median**.

The 75th percentile is also known as the 3rd **Quartile**

Example: Students browsing the web

We can refer to the cumulative relative frequency graph shown in the prior example to make the Ogive shown here.



Using the graph, we can estimate the quartiles of the distributions by where the line graph crosses cumulative relative frequency values of 0.25, 0.50 and 0.75.

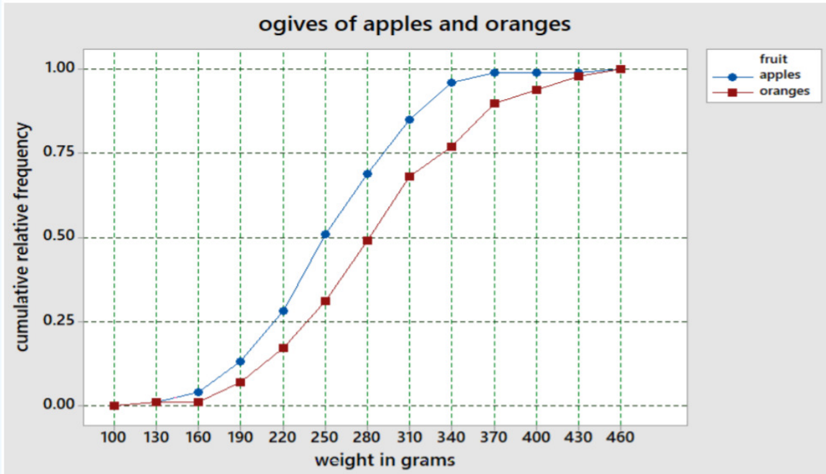
The 1st Quartile is about 87 minutes.

The median is about 100 minutes.

The 3rd Quartile is about 108 minutes.

Example: Comparing weights of apples and oranges

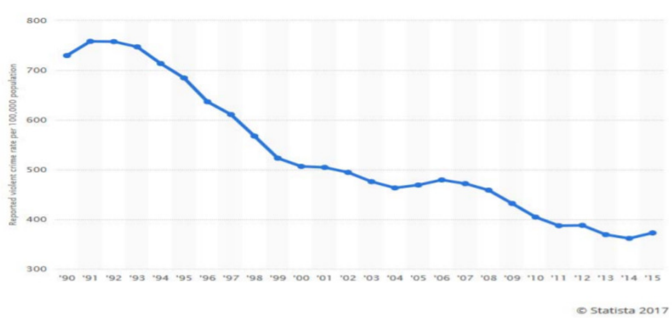
For the cumulative relative frequencies of the weights of apples and oranges, we can put both ogives on a single graph and estimate the quartiles.



Fruit	1st Quartile	Median	3rd Quartile
Apples	210 grams	250 grams	295 grams
Oranges	235 grams	280 grams	335 grams

Line Graphs with time

The ogive is an example of a line graph. A very useful line graph is one in which time is the horizontal axis. An early example from Section 1.1 of this type of line graphs is the historical crime rates. The line graph shows that violent crime has decreased over time.

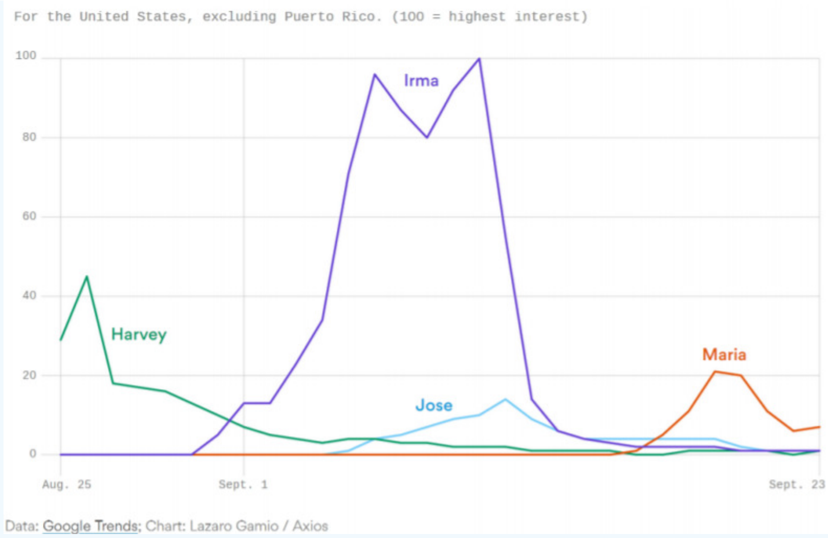


Line graph of Report

Example: Major Hurricanes in the Atlantic Ocean

In a one month period in 2017, four major hurricanes (category 3 or higher) formed in the Atlantic Ocean. Three of these hurricanes did devastating and costly damage to regions of the United States: Hurricane Harvey in Texas, Hurricane Irma in Florida and Hurricane Maria in Puerto Rico and the Virgin Islands. There was also catastrophic damage from these storms in Cuba, Dominica and other Caribbean countries, islands, and territories.

A Google Analytic graph shows that much more attention was paid to Hurricane Irma throughout the days it was threatening Florida.[22]



However, Google Analytics excludes Puerto Rico which took a direct hit from Hurricane Maria. It could also be that after Harvey caused massive flooding in and near Houston, more people became interested in all hurricane activity.

SECTION OVERVIEW

1.4: Introduction to Statistics

This first chapter begins by discussing what statistics are and why the study of statistics is important. Subsequent sections cover a variety of topics all basic to the study of statistics. One theme common to all of these sections is that they cover concepts and ideas important for other chapters in the book.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

1.4.1: What are Statistics?

Learning Objectives

- Describe the range of applications of statistics
- Identify situations in which statistics can be misleading
- Define "Statistics"

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every 8 South Africans is HIV positive.
- By the year 2020, there will be 15 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider the following three scenarios and the interpretations based upon the presented statistics. You will find that the numbers may be right, but the interpretation may be wrong. Try to identify a major flaw with each interpretation before we describe it.

1) A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.

A major flaw is that ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible.

2) The more churches in a city, the more crime there is. Thus, churches lead to crime.

A major flaw is that both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This problem, which we discuss in more detail in the section on Causation in Chapter 6, refers to the third-variable problem. Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

3) 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

A major flaw is that we don't have the information that we need. What is the rate at which marriages are occurring? Suppose only 1% of marriages 25 years ago were interracial and so now 1.75% of marriages are interracial (1.75 is 75% higher than 1). But this latter number is hardly evidence suggesting the acceptability of interracial marriages. In addition, the statistic provided does not rule out the possibility that the number of interracial marriages has seen dramatic fluctuations over the years and this year is not the highest. Again, there is simply not enough information to understand fully the impact of the statistics.

As a whole, these examples show that statistics are not only facts and figures; they are something more than that. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

- Mikki Hebl

1.4.2: Importance of Statistics

Learning Objectives

- Give examples of statistics encountered in everyday life
- Give examples of how statistics can lend credibility to an argument

Like most people, you probably feel that it is important to "take control of your life." But what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard on several occasions. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- Native Americans are significantly more likely to be hit crossing the street than are people of other ethnicities.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are. They come from psychology, health, law, sports, business, etc. Indeed, data and data interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) The present textbook is designed to help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter. The British Prime Minister Benjamin Disraeli is quoted by Mark Twain as having said, "There are three kinds of lies -- lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

We have put the emphasis on defending ourselves against fraudulent claims wrapped up as statistics. We close this section on a more positive note. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

- Mikki Hebl

1.4.3: Descriptive Statistics

Learning Objectives

- Define "descriptive statistics"
- Distinguish between descriptive statistics and inferential statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word "data" refers to the information that has been collected from an experiment, a survey, a historical record, etc. (By the way, "data" is plural. One piece of information is called a "datum.") If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed. Several descriptive statistics are often used at one time, to give a full picture of the data.

Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. Generalizing from our data to another set of cases is the business of inferential statistics, which you'll be studying in another Section. Here we focus on (mere) descriptive statistics. Some descriptive statistics are shown in Table $1.4.3.1$. The table shows the average salaries for various occupations in the United States in 1999. (Click here to see how much individuals with other occupations earn.)

Table $1.4.3.1$: Average salaries for various occupations in 1999.

Salary	**Occupation**
$112,760	pediatricians
$106,130	dentists
$100,090	podiatrists
$ 76,140	physicists
$ 53,410	architects
$ 49,720	school, clinical, and counseling psychologists
$ 47,910	flight attendants
$ 39,560	elementary school teachers
$ 38,710	police officers
$ 18,980	floral designers

Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that we pay the people who educate our children and who protect our citizens a great deal less than we pay people who take care of our feet or our teeth.

For more descriptive statistics, consider Table $1.4.3.2$ which shows the number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. From this table we see that men outnumber women most in Jacksonville, NC, and women outnumber men most in Sarasota, FL. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Table $1.4.3.2$: Number of unmarried men per 100 *unmarried women in U.S. Metro Areas in* 1990.

Cities with mostly men	Men per 100 Women	Cities with mostly women	Men per 100 Women
1. Jacksonville, NC	224	1. Sarasota, FL	66
2. Killeen-Temple, TX	123	2. Bradenton, FL	68
3. Fayetteville, NC	118	3. Altoona, PA	69
4. Brazoria, TX	117	4. Springfield, IL	70
5. Lawton, OK	116	5. Jacksonville, TN	70

Cities with mostly men	Men per 100 Women	Cities with mostly women	Men per 100 Women
6. State College, PA	113	6. Gadsden, AL	70
7. Clarksville-Hopkinsville, TN-KY	113	7. Wheeling, WV	70
8. Anchorage, Alaska	112	8. Charleston, WV	71
9. Salinas-Seaside-Monterey, CA	112	9. St. Joseph, MO	71
10. Bryan-College Station, TX	111	10. Lynchburg, VA	71

NOTE: Unmarried includes never-married, widowed, and divorced persons, 15 years or older.

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the fact that elderly individuals tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

You probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of competition. (The first modern Olympics took place in 1896.) Table 1.4.3.3 shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

Table 1.4.3.3: Winning Olympic marathon times.

Women			
Year	Winner	Country	Time
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40
1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20
Men			
Year	Winner	Country	Time
1896	Spiridon Louis	GRE	2:58:50
1900	Michel Theato	FRA	2:59:45
1904	Thomas Hicks	USA	3:28:53
1906	Billy Sherring	CAN	2:51:23
1908	Johnny Hayes	USA	2:55:18
1912	Kenneth McArthur	S. Afr.	2:36:54
1920	Hannes Kolehmainen	FIN	2:32:35
1924	Albin Stenroos	FIN	2:41:22
1928	Boughra El Ouafi	FRA	2:32:57
1932	Juan Carlos Zabala	ARG	2:31:36

Women			
1936	Sohn Kee-Chung	JPN	2:29:19
1948	Delfo Cabrera	ARG	2:34:51
1952	Emil Ztopek	CZE	2:23:03
1956	Alain Mimoun	FRA	2:25:00
1960	Abebe Bikila	ETH	2:15:16
1964	Abebe Bikila	ETH	2:12:11
1968	Mamo Wolde	ETH	2:20:26
1972	Frank Shorter	USA	2:12:19
1976	Waldemar Cierpinski	E.Ger	2:09:55
1980	Waldemar Cierpinski	E.Ger	2:11:03
1984	Carlos Lopes	POR	2:09:21
1988	Gelindo Bordin	ITA	2:10:32
1992	Hwang Young-Cho	S. Kor	2:13:23
1996	Josia Thugwane	S. Afr.	2:12:36
2000	Gezahenge Abera	ETH	2:10.10
2004	Stefano Baldini	ITA	2:10:55

There are many descriptive statistics that we can compute from the data in the table. To gain insight into the improvement in speed over the years, let us divide the men's times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written $2:44:22$). The mean winning time for the second 13 races is $2:13:18$. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We can't answer this question with descriptive statistics alone. All we can affirm is that the two means are "suggestive."

Examining Table 3 leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how much they will decrease (if at all) in the next century of the Olympics. Might we one day witness a sub-2 hour marathon? The study of statistics can help you make reasonable guesses about the answers to these questions.

- Mikki Hebl

1.4.4: Inferential Statistics

Learning Objectives

- Distinguish between a sample and a population
- Define inferential statistics
- Identify biased samples
- Distinguish between simple random sampling and stratified sampling
- Distinguish between random sampling and random assignment

Populations and samples

In statistics, we often rely on a sample, that is, a small subset of a larger set of data, to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn.

Example $1.4.4.1$

You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Whom will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. The mathematical procedures whereby we convert information about the sample into intelligent guesses about the population fall under the rubric of inferential statistics.

A sample is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferential statistics are based on the assumption that sampling is random. We trust a random sample to represent different segments of society in close to the appropriate proportions (provided the sample is large enough; see below).

Example $1.4.4.2$

We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school. Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls $48,000$ students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors.

To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

Example $1.4.4.3$

A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example $1.4.4.3$, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example $1.4.4.4$

A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In Example $1.4.4.4$, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who cannot do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed). Sampling Bias is Discussed in More Detail Here

Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check your understanding of simple random sampling, consider the following example. What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

Example $1.4.4.5$

A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

In Example $1.4.4.5$, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the $Z's$ may bias the sample. Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the $B's$ is that the “every-other-one” procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For example, if a random sample of 20 subjects were taken from a population with an equal number of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. (To see how to obtain this probability, see the section on the binomial distribution.) Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competing to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional $4,212$ people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

Random Assignment

In experimental research, populations are often hypothetical. For example, in an experiment comparing the effectiveness of a new anti-depressant drug with a placebo, there is no actual population of individuals taking the drug. In this case, a specified population of people with some degree of depression is defined and a random sample is taken from this population. The sample is then randomly divided into two groups; one group is assigned to the treatment condition (drug) and the other group is assigned to the control condition (placebo). This random division of the sample into two groups is called random assignment. Random assignment is critical for the validity of an experiment. For example, consider the bias that could be introduced if the first 20 subjects to show up at the experiment were assigned to the experimental group and the second 20 subjects were assigned to the control group. It is possible that subjects who show up late tend to be more depressed than those who show up early, thus making the experimental group less depressed than the control group even before the treatment was administered.

In experimental research of this kind, failure to assign subjects randomly to groups is generally more serious than having a non-random sample. Failure to randomize (the former error) invalidates the experimental findings. A non-random sample (the latter error) simply restricts the generalizability of the results.

Stratified Sampling

Since simple random sampling often does not ensure a representative sample, a sampling method called stratified random sampling is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let's take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

- Mikki Hebl and David Lane

1.4.5: Sampling Demonstration

Learning Objectives

- Distinguish between simple random sampling and stratified sampling.
- Describe how often random and stratified sampling give exactly the same result.

Review of Sampling

Instructions

The sampling simulation uses a population of 100 animals: 60 lions, 30 turtles, 10 rabbits.

Options

Random Sample: This option allows you to draw a sample of 10 animals at a time with each animal having an equal chance of being selected.

Stratified Sample: This option allows you to draw a sample of 10 animals at a time, with each number of animals from a group being proportional to their group's size of the population.

Simulation Results

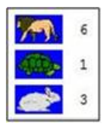


Figure 1.4.5.1: Simulation Results

The number of animals chosen from each group when a sample is drawn is shown next to the picture of the animal.

When you give it a try

Random Sampling

- Begin by leaving the Random Sample option selected.
- Click on the Select 10 Animals button, 10 animals will be selected out of the population.

  Note: The animals become highlighted in blue and a number count of each animal selected will be listed by each animal image.
- Each time you push the Select 10 Animals button, another sample will be drawn and the new tally will be shown on the right side of the previous sample.
- You should get different tally results for each animal as you select Select 10 Animals, however the computer may give you the same number drawn from an animal category every now and then.

Stratified Sample

Note: Your animals should become highlighted in blue and a number count should be listed by each animal image.

- Click on the Reset button, to clear the simulation.
- Select the Stratified Sample option.
- Click on the Select 10 Animals button a few times.
- As you get a new tally for every Select 10 Animals button, notice that the number of animals stays the same, but the animals selected are not always the same animals.

Illustrated Instructions

The opening screen of the sampling simulation displays all 100 animals in the population. You can select between a random sample and a stratified sample directly below the population and then generate a sample of ten animals.

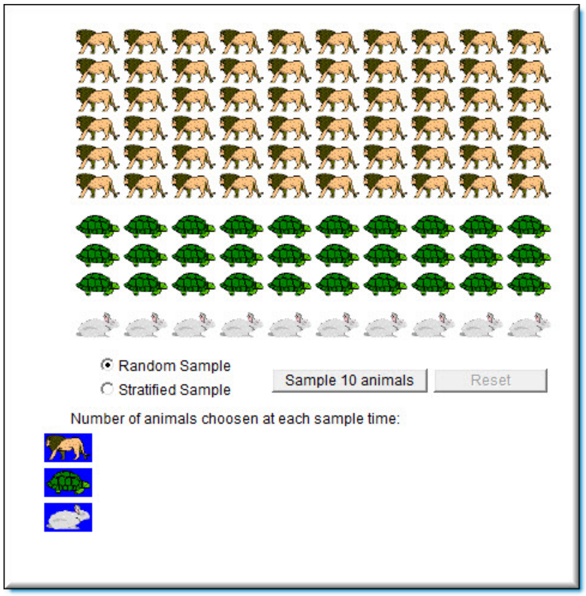


Figure 1.4.5.2: Sample Choices

Below is an example of a random sample. Notice that animals selected are highlighted in the population and the total number of animals selected from each category is listed at the bottom of the simulation.

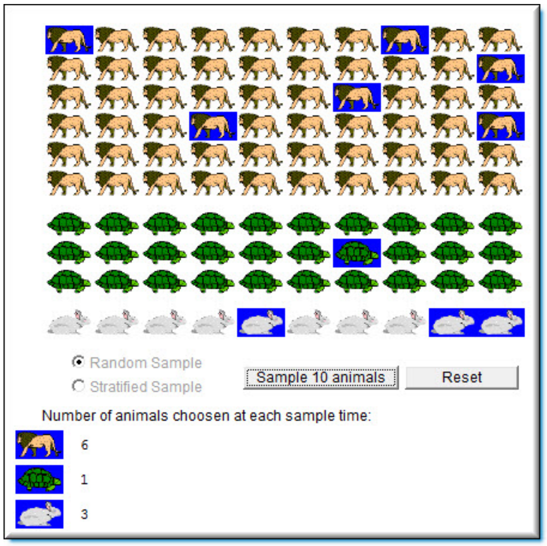


Figure 1.4.5.3: Random Sample

1.4.6: Variables

Learning Objectives

- Define and distinguish between independent and dependent variables
- Define and distinguish between discrete and continuous variables
- Define and distinguish between qualitative and quantitative variables

Independent and Dependent variables

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as π that do not vary). When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant." When a variable is manipulated by an experimenter, it is called an independent variable. The experiment seeks to determine the effect of the *independent variable* on relief from depression. In this example, relief from depression is called a dependent variable. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Example 1.4.6.1

Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month-old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

a. What is the independent variable?
b. What are the dependent variables?

Solution

a. dietary supplement: none, blueberry, strawberry, and spinach
b. memory test and motor skills test

More information on the blueberry study

Example 1.4.6.2

Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the *Journal of the National Cancer Institute* suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

a. What is the independent variable?
b. What is the dependent variable?

Solution

a. supplements: beta-carotene or placebo
b. occurrence of cancer

Example 1.4.6.3

How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

a. What is the independent variable?
b. What is the dependent variable?

Solution

a. brightness of brake lights
b. time to hit brakes

Levels of an Independent Variable: Experiments and Controls

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have five levels. In general, the number of levels of an independent variable is the number of experimental conditions.

Qualitative and Quantitative Variables

An important distinction between variables is between qualitative variables and quantitative variables. Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. The values of a qualitative variable do not imply a numerical ordering. Values of the variable "religion" differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as categorical variables. Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

In the study on the effect of diet discussed in Example $1.4.6.1$, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called discrete variables since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are continuous variables since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Heidi Ziemer

1.4.7: Percentiles

Learning Objectives

- Define percentiles
- Use three formulas for computing percentiles

A test score in and of itself is usually difficult to interpret. For example, if you learned that your score on a measure of shyness was 35 out of a possible 50, you would have little idea how shy you are compared to other people. More relevant is the percentage of people with lower shyness scores than yours. This percentage is called a percentile. If 65% of the scores were below yours, then your score would be the 65^{th} percentile.

Two Simple Definitions of Percentile

There is no universally accepted definition of a percentile. Using the 65^{th} percentile as an example, the 65^{th} percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this "Definition 1." The 65^{th} percentile can also be defined as the smallest score that is greater than or equal to 65% of the scores. This we will call "Definition 2." Unfortunately, these two definitions can lead to dramatically different results, especially when there is relatively little data. Moreover, neither of these definitions is explicit about how to handle rounding. For instance, what rank is required to be higher than 65% of the scores when the total number of scores is 50? This is tricky because 65% of 50 is 32.5. How do we find the lowest number that is higher than 32.5 of the scores? A third way to compute percentiles (presented below) is a weighted average of the percentiles computed according to the first two definitions. This third definition handles rounding more gracefully than the other two and has the advantage that it allows the median to be defined conveniently as the 50^{th} percentile.

Third Definition

Unless otherwise specified, when we refer to "percentile," we will be referring to this third definition of percentiles. Let's begin with an example. Consider the 25^{th} percentile for the 8 numbers in Table $1.4.7.1$ Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

Table $1.4.7.1$: Test Scores.

Number	3	5	7	8	9	11	13	15
Rank	1	2	3	4	5	6	7	8

The first step is to compute the rank (R) of the 25^{th} percentile. This is done using the following formula:

$$R = P/100 \times (N+1) \tag{1.4.7.1}$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = 25/100 \times (8+1) = 9/4 = 2.25 \tag{1.4.7.2}$$

If R is an integer, the P^{th} percentile is the number with rank R. When R is not an integer, we compute the P^{th} percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, $IR = 2$.
2. Define FR as the fractional portion of R. For this example, $FR = 0.25$.
3. Find the scores with Rank IR and with Rank $IR+1$. For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.
4. Interpolate by multiplying the difference between the scores by FR and add the result to the lower score. For these data, this is $(0.25)(7-5)+5 = 5.5$.

Therefore, the 25^{th} percentile is 5.5. If we had used the first definition (the smallest score greater than 25% of the scores), the 25^{th} percentile would have been 7. If we had used the second definition (the smallest score greater than or equal to 25% of the scores), the 25^{th} percentile would have been 5.

For a second example, consider the 20 quiz scores shown in Table $1.4.7.2$

Table 1.4.7.2: 20 *Quiz Scores*.

Number	4	4	4	5	5	5	6	6	7	7	7	8	8	9	9	9	10	10	10
Rank	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	18	19	20

We will compute the 25^{th} and the 85^{th} percentiles. For the 25^{th},

$$R = 25/100 \times (20+1) = 21/4 = 5.25 \tag{1.4.7.3}$$

$$IR = 5 \text{ and } FR = 0.25 \tag{1.4.7.4}$$

Since the score with a rank of IR (which is 5) and the score with a rank of $IR+1$ (which is 6) are both equal to 5, the 25^{th} percentile is 5. In terms of the formula:

$$25^{th} \text{ percentile} = (0.25) \times (5-5) + 5 = 5 \tag{1.4.7.5}$$

For the 85^{th} percentile,

$$R = 85/100 \times (20+1) = 17.85. \tag{1.4.7.6}$$

$$IR = 17 \text{ and } FR = 0.85 \tag{1.4.7.7}$$

Caution: FR does not generally equal the percentile to be computed as it does here.

The score with a rank of 17 is 9 and the score with a rank of 18 is 10. Therefore, the 85^{th} percentile is:

$$(0.85)(10-9) + 9 = 9.85 \tag{1.4.7.8}$$

Consider the 50^{th} percentile of the numbers $2, 3, 5, 9$.

$$R = 50/100 \times (4+1) = 2.5 \tag{1.4.7.9}$$

$$IR = 2 \text{ and } FR = 0.5 \tag{1.4.7.10}$$

The score with a rank of IR is 3 and the score with a rank of $IR+1$ is 5. Therefore, the 50^{th} percentile is:

$$(0.5)(5-3) + 3 = 4 \tag{1.4.7.11}$$

Finally, consider the 50^{th} percentile of the numbers $2, 3, 5, 9, 11$.

$$R = 50/100 \times (5+1) = 3 \tag{1.4.7.12}$$

$$IR = 3 \text{ and } FR = 0 \tag{1.4.7.13}$$

Whenever $FR = 0$, you simply find the number with rank IR. In this case, the third number is equal to 5, so the 50^{th} percentile is 5. You will also get the right answer if you apply the general formula:

$$50^{th} \text{ percentile} = (0.00)(9-5) + 5 = 5 \tag{1.4.7.14}$$

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane

1.4.8: Levels of Measurement

Learning Objectives

- Define and distinguish among nominal, ordinal, interval, and ratio scales
- Identify a scale type
- Discuss the type of scale used in psychological measurement
- Give examples of errors that can be made by failing to understand the proper use of measurement scales

Types of Scales

Before we can conduct a statistical analysis, we need to measure our dependent variable. Exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a dependent variable such as "favorite color," you can simply note the color-word (like "red") that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale types," or just "scales," and are described in this section.

Nominal scales

When measuring using a nominal scale, one simply names or categorizes responses. Gender, handedness, favorite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which green is placed "ahead of" blue. Responses are merely categorized. Nominal scales embody the lowest level of measurement.

Ordinal scales

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person's use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

What if the researcher had measured satisfaction by asking consumers to indicate their level of satisfaction by choosing a number from one to four? Would the difference between the responses of one and two necessarily reflect the same difference in satisfaction as the difference between the responses two and three? The answer is No. Changing the response format to numbers does not change the meaning of the scale. We still are in no position to assert that the mental step from 1 to 2 (for example) is the same as the mental step from 3 to 4.

Interval scales

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label "zero" is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the "zero" label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is "twice as hot" as 40 degrees. Such a claim would depend on an arbitrary decision about where to "start" the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

Ratio scales

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents (or that Bill Gates has a million times more money than you do).

What level of measurement is used for psychological variables?

Rating scales are used frequently in psychological research. For example, experimental subjects may be asked to rate their level of pain, how much they like a consumer product, their attitudes about capital punishment, their confidence in an answer to a test question. Typically these ratings are made on a 5-point or a 7-point scale. These scales are ordinal scales since there is no assurance that a given difference represents the same thing across the range of the scale. For example, there is no way to be sure that a treatment that reduces pain from a rated pain level of 3 to a rated pain level of 2 represents the same level of relief as a treatment that reduces pain from a rated pain level of 7 to a rated pain level of 6.

In memory experiments, the dependent variable is often the number of items correctly recalled. What scale of measurement is this? You could reasonably argue that it is a ratio scale. First, there is a true zero point: some subjects may get no items correct at all. Moreover, a difference of one represents a difference of one item recalled across the entire scale. It is certainly valid to say that someone who recalled 12 items recalled twice as many items as someone who recalled only 6 items.

But number-of-items recalled is a more complicated case than it appears at first. Consider the following example in which subjects are asked to remember as many items as possible from a list of 10. Assume that (a) there are 5 easy items and 5 difficult items, (b) half of the subjects are able to recall all the easy items and different numbers of difficult items, while (c) the other half of the subjects are unable to recall any of the difficult items but they do remember different numbers of easy items. Some sample data are shown below.

Table 1.4.8.1

Subject	Easy Items	Difficult Items	**Score**

Subject	Easy Items					Difficult Items					**Score**
A	0	0	1	1	0	0	0	0	0	0	**2**
B	1	0	1	1	0	0	0	0	0	0	**3**
C	1	1	1	1	1	1	1	0	0	0	**7**
D	1	1	1	1	1	0	1	1	0	1	**8**

Let's compare (1) the difference between Subject $A's$ score of 2 and Subject $B's$ score of 3 with (2) the difference between Subject $C's$ score of 7 and Subject $D's$ score of 8. The former difference is a difference of one easy item; the latter difference is a difference of one difficult item. Do these two differences necessarily signify the same difference in memory? We are inclined to respond "No" to this question since only a little more memory may be needed to retain the additional easy item whereas a lot more memory may be needed to retain the additional hard item. The general point is that it is often inappropriate to consider psychological measurement scales as either interval or ratio.

Consequences of level of measurement

Why are we so interested in the type of scale that measures a dependent variable? The crux of the matter is the relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider a hypothetical study in which 5 children are asked to choose their favorite color from blue, red, yellow, green, and purple. The researcher codes the results as follows:

Table $1.4.8.2$

Color	Code
Blue	1
Red	2
Yellow	3
Green	4
Purple	5

This means that if a child said her favorite color was "Red," then the choice was coded as "2," if the child said her favorite color was "Purple," then the response was coded as 5, and so forth. Consider the following hypothetical data:

Table $1.4.8.3$

Subject	Color	Code
1	Blue	1
2	Blue	1
3	Green	4
4	Green	4
5	Purple	5

Each code is a number, so nothing prevents us from computing the average code assigned to the children. The average happens to be 3, but you can see that it would be senseless to conclude that the average favorite color is yellow (the color with a code of 3). Such nonsense arises because favorite color is a nominal scale, and taking the average of its numerical labels is like counting the number of letters in the name of a snake to see how long the beast is.

Does it make sense to compute the mean of numbers measured on an ordinal scale? This is a difficult question, one that statisticians have debated for decades. You will be able to explore this issue yourself in a simulation shown in the next section and reach your own conclusion. The prevailing (but by no means unanimous) opinion of statisticians is that for almost all practical situations, the mean of an ordinally-measured variable is a meaningful statistic. However, as you will see in the simulation, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Dan Osherson and David M. Lane

1.4.9: Measurements

Learning Objectives

- Understand what it means for a scale to be ordinal and its relationship to interval scales.
- Determine whether an investigator can be misled by computing the means of an ordinal scale.

Instructions

This is a demonstration of a very complex issue. Experts in the field disagree on how to interpret differences on an ordinal scale, so do not be discouraged if it takes you a while to catch on. In this demonstration you will explore the relationship between interval and ordinal scales. The demonstration is based on two brands of baked goods.

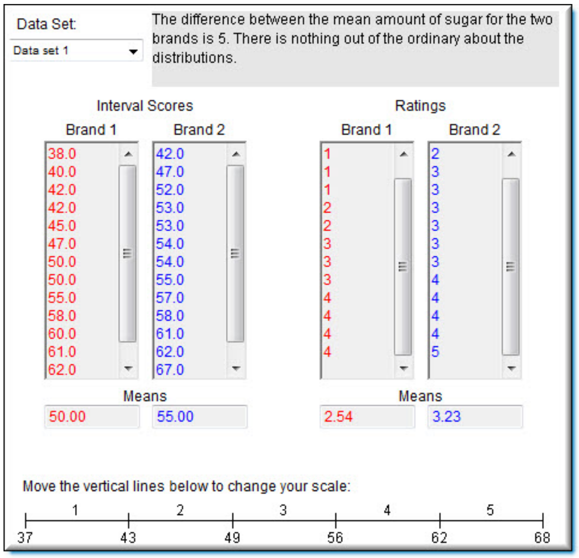


Figure 1.4.9.1: Interval scores and Ratings for two brands

The data on the left side labeled "interval scores" shows the amount of sugar in each of 12 products. The column labeled "$\text{Brand } 1$" contains the sugar content of each of 12 brand-one products. The second column ("$\text{Brand } 1$") shows the sugar content of the brand-two products. The amount of sugar is measured on an interval scale.

A rater tastes each of the products and rates them on a 5-point "sweetness" scale. Rating scales are typically ordinal rather than interval.

The scale at the bottom shows the "mapping" of sugar content onto the ratings. Sugar content between 37 and 43 is rated as 1, between 43 and $49, 2, etc.$ Therefore, the difference between a rating of 1 and a rating of 2 represents, on average a "sugar difference" of 6. A difference between a rating of 2 and a rating of 3 also represents, on average a "sugar difference" of 6. The original ratings are rounded off and displayed are on an interval scale. It is likely that rater's ratings would not be on an interval scale. You can change the cutoff points between ratings by moving the vertical lines with the mouse. As you change these cutoffs, the ratings change automatically. For example, you might see what the ratings would look like if people did not consider something very sweet (rating of 5) unless it was very very sweet.

The mean amount of sugar in $\text{Data Set } 1$ is 50 for the first brand and 55 for the second brand. The obvious conclusion is that, on average, the second brand is sweeter than the first. However, pretend that you only had the ratings to go by and were not aware of the actual amounts of sugar. Would you reach the correct decision if you compared the mean ratings of the two brands. Change the cutoffs for mapping the interval sugar scale onto the ordinal rating scale. Do any mappings lead to incorrect interpretations? Try this with $\text{Data Set } 1$ and with $\text{Data Set } 2$. Try to find a situation where the mean sweetness rating is higher for $\text{Brand } 2$ even though the mean amount of sugar is greater for $\text{Brand } 1$. If you find such a situation, then you have found an instance in which using the means of ordinal data lead to incorrect conclusions. It is possible to find this situation, so look hard.

Keep in mind that in realistic situations, you only know the ratings and not the "true" interval scale that underlies them. If you knew the interval scale, you would use it.

1.4.10: Distributions

Learning Objectives

- Define "distribution"
- Interpret a frequency distribution
- Distinguish between a frequency distribution and a probability distribution
- Construct a grouped frequency distribution for a continuous variable
- Identify the skew of a distribution
- Identify bimodal, leptokurtic, and platykurtic distributions

Distributions of Discrete Variables

A recently purchased a bag of Plain M&M's contained candies of six different colors. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. These counts are shown below in Table $1.4.10.1$

Table 1.4.10.1: Frequencies in the Bag of M&M's

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a frequency table and it describes the distribution of M&M color frequencies. Not surprisingly, this kind of distribution is called a frequency distribution. Often a frequency distribution is shown graphically as in Figure $1.4.10.1$

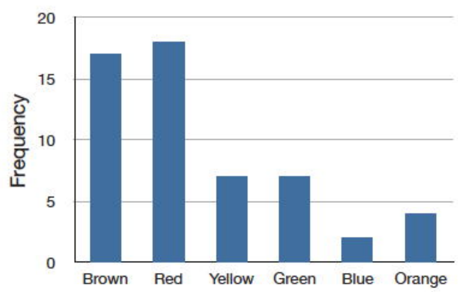


***Figure* $1.4.10.1$:** *Distribution of* 55 *M&M's.*

The distribution shown in Figure $1.4.10.1$ concerns just my one bag of M&M's. You might be wondering about the distribution of colors for all M&M's. The manufacturer of M&M's provides some information about this matter, but they do not tell us exactly how many M&M's of each color they have ever produced. Instead, they report proportions rather than frequencies. Figure $1.4.10.2$ shows these proportions. Since every M&M is one of the six familiar colors, the six proportions shown in the figure add to one. We call Figure $1.4.10.2$ a probability distribution because if you choose an M&M at random, the probability of getting, say, a brown M&M is equal to the proportion of M&M's that are brown (0.30).

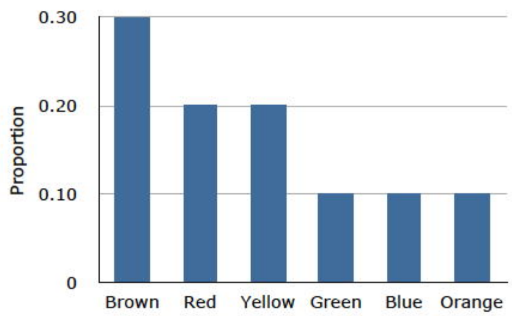


Figure 1.4.10.2: Distribution of all M&M's.

Notice that the distributions in Figures 1.4.10.1 and 1.4.10.2 are not identical. Figure 1.4.10.1 portrays the distribution in a sample of 55 M&M's. Figure 1.4.10.2 shows the proportions for all M&M's. Chance factors involving the machines used by the manufacturer introduce random variation into the different bags produced. Some bags will have a distribution of colors that is close to Figure 1.4.10.2; others will be further away.

Continuous Variables

The variable "color of M&M" used in this example is a discrete variable, and its distribution is also called discrete. Let us now extend the concept of a distribution to continuous variables. The data shown in Table 1.4.10.2 are the times it took one of us (DL) to move the mouse over a small target in a series of 20 trials. The times are sorted from shortest to longest. The variable "time to respond" is a continuous variable. With time measured accurately (to many decimal places), no two response times would be expected to be the same. Measuring time in milliseconds (thousandths of a second) is often precise enough to approximate a continuous variable in Psychology. As you can see in Table 1.4.10.2, measuring DL's responses this way produced times no two of which were the same. As a result, a frequency distribution would be uninformative: it would consist of the 20 times in the experiment, each with a frequency of 1.

Table 1.4.10.2: Response Times

568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

The solution to this problem is to create a grouped frequency distribution. In a grouped frequency distribution, scores falling within various ranges are tabulated. Table 1.4.10.3 shows a grouped frequency distribution for these 20 times.

Table 1.4.10.3: Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5

Range	Frequency
900-1000	0
1000-1100	1

Grouped frequency distributions can be portrayed graphically. Figure $1.4.10.3$ shows a graphical representation of the frequency distribution in Table $1.4.10.3$. This kind of graph is called a histogram. A later chapter contains an entire section devoted to histograms.

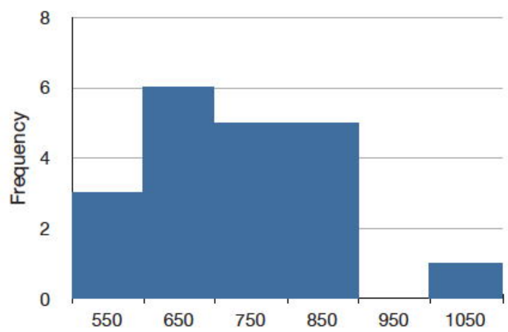


Figure 1.4.10.3: A histogram of the grouped frequency distribution shown in Table 1.4.10.3. The labels on the X*-axis are the middle values of the range they represent.*

Probability Densities

The histogram in Figure $1.4.10.3$ portrays just DL's 20 times in the one experiment he performed. To represent the probability associated with an arbitrary movement (which can take any positive amount of time), we must represent all these potential times at once. For this purpose, we plot the distribution for the continuous variable of time. Distributions for continuous variables are called continuous distributions. They also carry the fancier name probability density. Some probability densities have particular importance in statistics. A very important one is shaped like a bell, and called the normal distribution. Many naturally-occurring phenomena can be approximated surprisingly well by this distribution. It will serve to illustrate some features of all continuous distributions.

An example of a normal distribution is shown in Figure $1.4.10.4$. Do you see the "bell"? The normal distribution doesn't represent a real bell, however, since the left and right tips extend indefinitely (we can't draw them any further so they look like they've stopped in our diagram). The Y-axis in the normal distribution represents the "density of probability." Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis. In Figure $1.4.10.4$, for example, the probability of an observation with value near 40 is about half of the probability of an observation with value near 50. (For more information, please see the chapter on normal distributions.)

Although this text does not discuss the concept of probability density in detail, you should keep the following ideas in mind about the curve that describes a continuous distribution (like the normal distribution). First, the area under the curve equals 1. Second, the probability of any exact value of X is 0. Finally, the area under the curve and bounded between two given points on the X-axis is the probability that a number chosen at random will fall between the two points. Let us illustrate with DL's hand movements. First, the probability that his movement takes some amount of time is one! (We exclude the possibility of him never finishing his gesture.) Second, the probability that his movement takes exactly 598.956432342346576 milliseconds is essentially zero. (We can make the probability as close as we like to zero by making the time measurement more and more precise.) Finally, suppose that the probability of DL's movement taking between 600 and 700 milliseconds is one tenth. Then the continuous distribution for DL's possible times would have a shape that places 10% of the area below the curve in the region bounded by 600 and 700 on the X-axis.

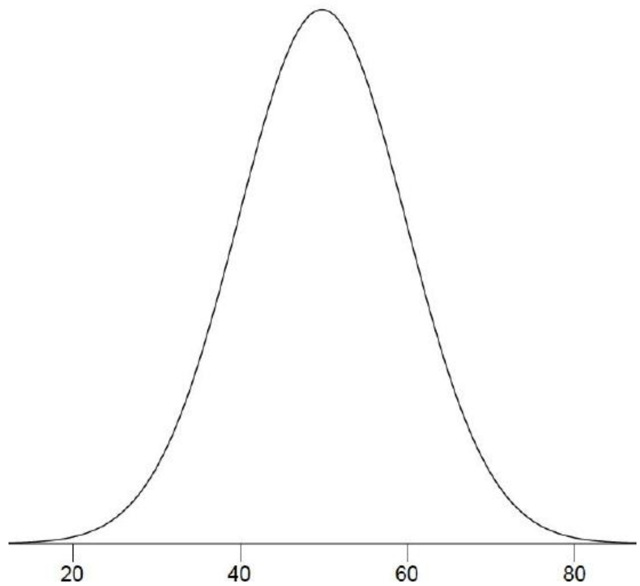


Figure 1.4.10.4: A normal distribution.

Shapes of Distributions

Distributions have different shapes; they don't all look like the normal distribution in Figure $1.4.10.4$ For example, the normal probability density is higher in the middle compared to its two tails. Other distributions need not have this feature. There is even variation among the distributions that we call "normal." For example, some normal distributions are more spread out than the one shown in Figure $1.4.10.4$ (their tails begin to hit the X-axis further from the middle of the curve -- for example, at 10 and 90 if drawn in place of Figure $1.4.10.4$). Others are less spread out (their tails might approach the X-axis at 30 and 70). More information on the normal distribution can be found in a later chapter completely devoted to them.

The distribution shown in Figure $1.4.10.4$ is symmetric; if you folded it in the middle, the two sides would match perfectly. Figure $1.4.10.5$ shows the discrete distribution of scores on a psychology test. This distribution is not symmetric: the tail in the positive direction extends further than the tail in the negative direction. A distribution with the longer tail extending in the positive direction is said to have a positive skew. It is also described as "skewed to the right."

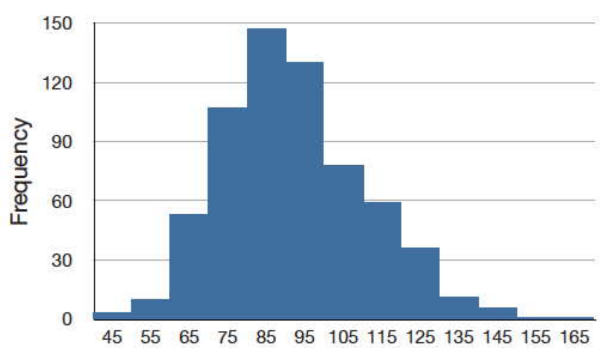


Figure 1.4.10.5: (left) A distribution with a positive skew.

Figure $1.4.10.6$ shows the salaries of major league baseball players in 1974 (in thousands of dollars). This distribution has an extreme positive skew.

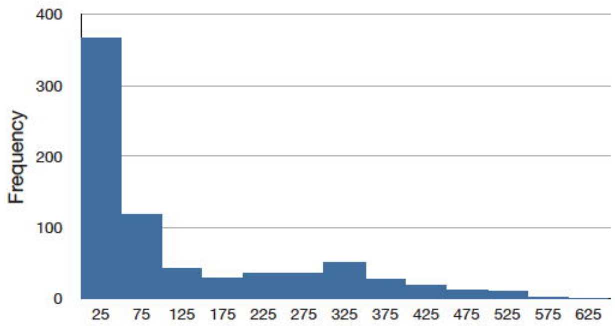


Figure 1.4.10.6: (right) A distribution with a very large positive skew.

A continuous distribution with a positive skew is shown in Figure 1.4.10.7.

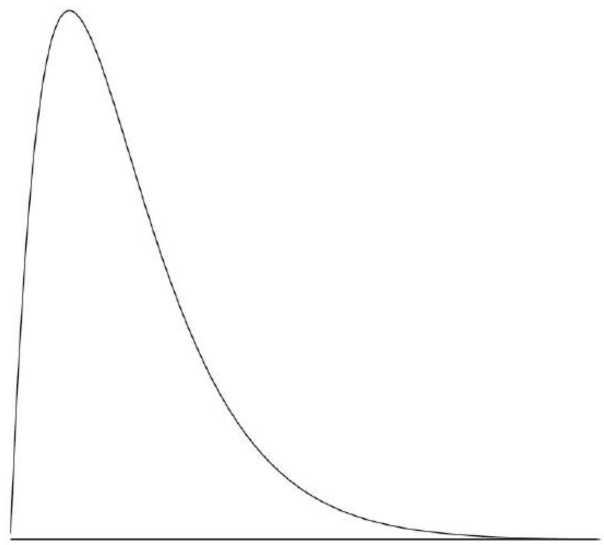


Figure 1.4.10.7: A continuous distribution with a positive skew.

Although less common, some distributions have a negative skew. Figure 1.4.10.8 shows the scores on a 20-point problem on a statistics exam. Since the tail of the distribution extends to the left, this distribution is skewed to the left.

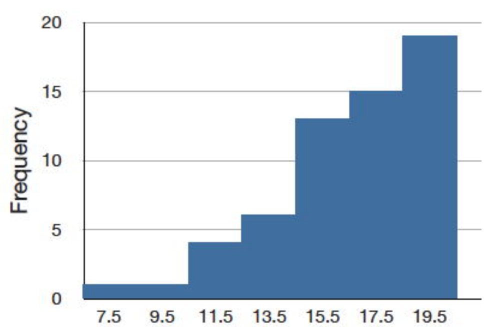


Figure 1.4.10.8: A distribution with negative skew.

The histogram in Figure 1.4.10.8 shows the frequencies of various scores on a 20-point question on a statistics test.

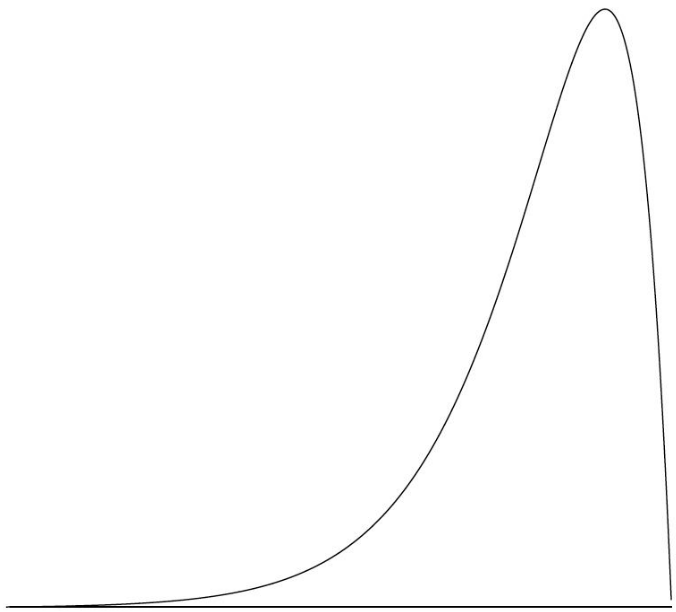


Figure 1.4.10.9: A continuous distribution with a negative skew.

A continuous distribution with a negative skew is shown in Figure 1.4.10.9. The distributions shown so far all have one distinct high point or peak. The distribution in Figure 1.4.10.10 has two distinct peaks. A distribution with two peaks is called a bimodal distribution.

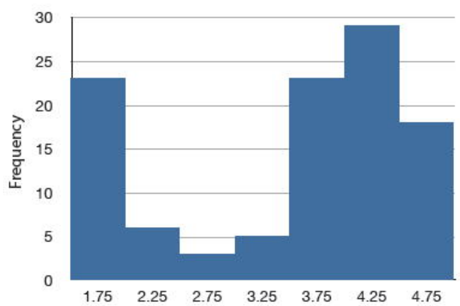


Figure 1.4.10.10. Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

Distributions also differ from each other in terms of how large or "fat" their tails are. Figure 1.4.10.11 shows two distributions that differ in this respect. The upper distribution has relatively more scores in its tails; its shape is called leptokurtic. The lower distribution has relatively fewer scores in its tails; its shape is called platykurtic.

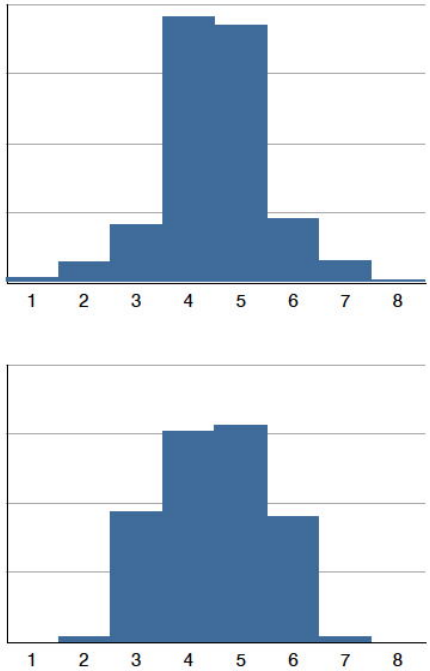


Figure 1.4.10.11. Distributions differing in kurtosis. The top distribution has long tails. It is called "leptokurtic." The bottom distribution has short tails. It is called "platykurtic."

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane and Heidi Ziemer

1.4.11: Summation Notation

Learning Objectives

- Use summation notation to express the sum of all numbers
- Use summation notation to express the sum of a subset of numbers
- Use summation notation to express the sum of squares

Many statistical formulas involve summing numbers. Fortunately there is a convenient notation for expressing summation. This section covers the basics of this summation notation.

Let's say we have a variable X that represents the weights (in grams) of 4 grapes. The data are shown in Table $1.4.11.1$.

Table $1.4.11.1$: *Weights of* 4 *grapes.*

Grape	X
1	4.6
2	5.1
3	4.9
4	4.4

We label Grape $1's$ weight X_1, Grape $2's$ weight X_2, etc. The following formula means to sum up the weights of the four grapes:

$$\sum_{i=1}^{4} X_i \tag{1.4.11.1}$$

The Greek letter capital sigma ($\sum$) indicates summation. The "$i = 1$" at the bottom indicates that the summation is to start with X_1 and the 4 at the top indicates that the summation will end with X_4. The "X_i" indicates that X is the variable to be summed as i goes from 1 to 4. Therefore,

$$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19.0 \tag{1.4.11.2}$$

The symbol

$$\sum_{i=1}^{3} X_i \tag{1.4.11.3}$$

indicates that only the first 3 scores are to be summed. The index variable i goes from 1 to 3.

When all the scores of a variable (such as X) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X \tag{1.4.11.4}$$

Thus, when no values of i are shown, it means to sum all the values of X.

Many formulas involve squaring numbers before they are summed. This is indicated as

$$\sum X^2 = 4.62 + 5.12 + 4.92 + 4.42 = 21.16 + 26.01 + 24.01 + 19.36 = 90.54 \tag{1.4.11.5}$$

Notice that:

$$\left(\sum X\right)^2 \neq \sum X^2 \tag{1.4.11.6}$$

because the expression on the left means to sum up all the values of X and then square the sum ($19^2 = 361$), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products. Table $1.4.11.2$ shows the data for variables X and Y. The cross products (XY) are shown in the third column. The sum of the cross products is $3+4+21=28$.

Table 1.4.11.2*: Cross Products.*

X	Y	XY
1	3	3
2	2	4
3	7	21

In summation notation, this is written as:

$$\sum XY = 28. \tag{1.4.11.7}$$

- David M. Lane

1.4.12: Linear Transformations

Learning Objectives

- Give the formula for a linear transformation
- Determine whether a transformation is linear
- Describe what is linear about a linear transformation

Often it is necessary to transform data from one measurement scale to another. For example, you might want to convert height measured in feet to height measured in inches. Table $1.4.12.1$ shows the heights of four people measured in both feet and inches. To transform feet to inches, you simply multiply by 12. Similarly, to transform inches to feet, you divide by 12.

Table 1.4.12.1: Converting between feet and inches

Feet	Inches
5.00	60
6.25	75
5.50	66
5.75	69

Some conversions require that you multiply by a number and then add a second number. A good example of this is the transformation between degrees Celsius and degrees Fahrenheit. Table $1.4.12.2$ shows the temperatures of five US cities in the early afternoon of $\text{November } 16, 2002$.

Table 1.4.12.2: Temperatures in 5 cities on $11/16/2002$

City	Degrees Fahrenheit	Degrees Celsius
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11

The formula to transform Celsius to Fahrenheit is:

$$F = 1.8C + 32 \tag{1.4.12.1}$$

The formula for converting from Fahrenheit to Celsius is

$$C = 0.5556F - 17.778 \tag{1.4.12.2}$$

The transformation consists of multiplying by a constant and then adding a second constant. For the conversion from Celsius to Fahrenheit, the first constant is 1.8 and the second is 32.

Figure $1.4.12.1$ shows a plot of degrees Celsius as a function of degrees Fahrenheit. Notice that the points form a straight line. This will always be the case if the transformation from one scale to another consists of multiplying by one constant and then adding a second constant. Such transformations are therefore called linear transformations.

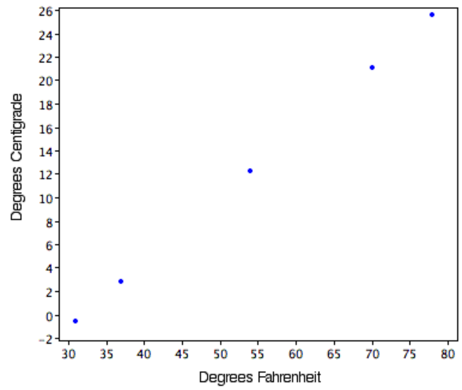


Figure 1.4.12.1*: Degrees Celsius as a function of degrees Fahrenheit*

Many transformations are not linear. With nonlinear transformations, the points in a plot of the transformed variable against the original variable would not fall on a straight line. Examples of nonlinear transformations are: square root, raising to a power, logarithm, and any of the trigonometric functions.

- David M. Lane

1.4.13: Logarithms

Learning Objectives

- Compute logs using different bases
- Perform basic arithmetic operations using logs
- State the relationship between logs and proportional change

The log transformation reduces positive skew. This can be valuable both for making the data more interpretable and for helping to meet the assumptions of inferential statistics.

Basics of Logarithms (Logs)

Logs are, in a sense, the opposite of exponents. Consider the following simple expression:

$$10^2 = 100 \tag{1.4.13.1}$$

Here we can say the base of 10 is raised to the second power. Here is an example of a log:

$$\log_{10}(100) = 2 \tag{1.4.13.2}$$

This can be read as: The log base ten of 100 equals 2. The result is the power that the base of 10 has to be raised to in order to equal the value (100). Similarly,

$$\log_{10}(1000) = 3 \tag{1.4.13.3}$$

since 10 has to be raised to the third power in order to equal $1,000$.

These examples all used base 10, but any base could have been used. There is a base which results in "natural logarithms" and that is called e and equals approximately 2.718. It is beyond the scope here to explain what is "natural" about it. Natural logarithms can be indicated either as: $\ln(x)$ or $\log_e(x)$.

Changing the base of the log changes the result by a multiplicative constant. To convert from $\log_{10}$ to natural logs, you multiply by 2.303. Analogously, to convert in the other direction, you divide by 2.303.

$$\ln X = 2.303 \log_{10} X \tag{1.4.13.4}$$

Taking the antilog of a number undoes the operation of taking the log. Therefore, since $\log_{10}(1000) = 3$, the $antilog_{10}$ of 3 is $10^3 = 1,000$. Taking the antilog of a number simply raises the base of the logarithm in question to that number.

Logs and Proportional Change

A series of numbers that increase proportionally will increase in equal amounts when converted to logs. For example, the numbers in the first column of Table $1.4.13.1$
increase by a factor of 1.5 so that each row is 1.5 times as high as the preceding row. The $\log_{10}$ transformed numbers increase in equal steps of 0.176.

Table 1.4.13.1*: Proportional raw changes are equal in log units*

Raw	Log
4.0	0.602
6.0	0.778
9.0	0.954
13.5	1.130

As another example, if one student increased their score from 100 to 200 while a second student increased theirs from 150 to 300, the percentage change (100%) is the same for both students. The log difference is also the same, as shown below.

$$\begin{aligned} Log_{10}(100) &= 2.000 \\ \log_{10}(200) &= 2.301 \\ Difference &: 0.301 \end{aligned} \tag{1.4.13.5}$$

$$\begin{aligned} \log_{10}(150) &= 2.176 \\ \log_{10}(300) &= 2.477 \\ Difference &: 0.301 \end{aligned}$$

Arithmetic Operations

Rules for logs of products and quotients are shown below.

$$\log(AB) = \log(A) + \log(B) \tag{1.4.13.6}$$

$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B) \tag{1.4.13.7}$$

For example,

$$\log_{10}(10 \times 100) = \log_{10}(10) + \log_{10}(100) = 1 + 2 = 3. \tag{1.4.13.8}$$

Similarly,

$$\log_{10}\left(\frac{100}{10}\right) = \log_{10}(100) - \log_{10}(10) = 2 - 1 = 1. \tag{1.4.13.9}$$

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane

1.4.14: Statistical Literacy

Do Athletes Get Special Treatment?

Prerequisites

Levels of Measurement

Figure 1.4.14.1: Runners

The Board of Trustees at a university commissioned a top management-consulting firm to address the admission processes for academic and athletic programs. The consulting firm wrote a report discussing the trade-off between maintaining academic and athletic excellence. One of their key findings was:

> The standard for an athlete's admission, as reflected in SAT scores alone, is lower than the standard for non-athletes by as much as 20 percent, with the weight of this difference being carried by the so-called "revenue sports" of football and basketball. Athletes are also admitted through a different process than the one used to admit non-athlete students.

What do you think?

Based on what you have learned in this chapter about measurement scales, does it make sense to compare SAT scores using percentages? Why or why not?

> As you may know, the SAT has an arbitrarily-determined lower limit on test scores of 200. Therefore, SAT is measured on either an ordinal scale or, at most, an interval scale. However, it is clearly not measured on a ratio scale. Therefore, it is not meaningful to report SAT score differences in terms of percentages. For example, consider the effect of subtracting 200 from every student's score so that the lowest possible score is 0. How would that affect the difference as expressed in percentages?

Statistical Errors in Politics

Prerequisites

Inferential Statistics

Figure 1.4.14.2: *Survey*

An article about ignorance of statistics in politics quotes a politician commenting on why the "American Community Survey" should be eliminated:

> “We’re spending $70 per person to fill this out. That’s just not cost effective, especially since in the end this is not a scientific survey. It’s a random survey.”

What do you think?

What is wrong with this statement? Despite the error in this statement, what type of sampling could be done so that the sample will be more likely to be representative of the population?

> Randomness is what makes the survey scientific. If the survey were not random, then it would be biased and therefore statistically meaningless, especially since the survey is conducted to make generalizations about the American population. Stratified sampling would likely be more representative of the population.

Reference

Mark C. C., scientopia.org

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Denise Harvey and David Lane

1.4.E: Introduction to Statistics (Exercises)

General Questions

Q1

A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics? (relevant section 1, relevant section 2)

Q2

A cognitive psychologist is interested in comparing two ways of presenting stimuli on subsequent memory. Twelve subjects are presented with each method and a memory test is given. What would be the roles of descriptive and inferential statistics in the analysis of these data? (relevant section 1 & relevant section 2)

Q3

If you are told that you scored in the 80^{th} percentile, from just this information would you know exactly what that means and how it was calculated? Explain. (relevant section)

Q4

A study is conducted to determine whether people learn better with spaced or massed practice. Subjects volunteer from an introductory psychology class. At the beginning of the semester 12 subjects volunteer and are assigned to the massed-practice condition. At the end of the semester 12 subjects volunteer and are assigned to the spaced-practice condition. This experiment involves two kinds of non-random sampling:

1. Subjects are not randomly sampled from some specified population
2. Subjects are not randomly assigned to conditions.

Which of the problems relates to the generality of the results? Which of the problems relates to the validity of the results? Which problem is more serious? (relevant section)

Q5

Give an example of an independent and a dependent variable. (relevant section)

Q6

Categorize the following variables as being qualitative or quantitative: (relevant section)

a. Rating of the quality of a movie on a 7-point scale
b. Age
c. Country you were born in
d. Favorite Color
e. Time to respond to a question

Q7

Specify the level of measurement used for the items in Question 6. (relevant section)

Q8

Which of the following are linear transformations? (relevant section)

a. Converting from meters to kilometers
b. Squaring each side to find the area
c. Converting from ounces to pounds
d. Taking the square root of each person's height.
e. Multiplying all numbers by 2 and then adding 5
f. Converting temperature from Fahrenheit to Centigrade

Q9

The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows:

$$g = 16 + 3s \tag{1.4.E.1}$$

Is this a linear transformation? If a student got a raw score of 20, what is his test grade? (relevant section)

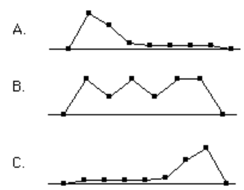
Q10

For the numbers $1, 2, 4, 16$, compute the following: (relevant section)

a. $\sum X$
b. $\sum X^2$
c. $(\sum X)^2$

Q11

Which of the frequency polygons has a large positive skew? Which has a large negative skew? (relevant section)



Q12

What is more likely to have a skewed distribution: time to solve an anagram problem (where the letters of a word or phrase are rearranged into another word or phrase like "dear" and "read" or "funeral" and "real fun") or scores on a vocabulary test? (relevant section)

Questions from Case Studies:

The following questions are from the Angry Moods (AM) case study.

Q13

(AM#1) Which variables are the participant variables? (They act as independent variables in this study.) (relevant section)

Q14

(AM#2) What are the dependent variables? (relevant section)

Q15

(AM#3) Is Anger-Out a quantitative or qualitative variable? (relevant section)

The following question is from the Teacher Ratings (TR) case study.

Q16

(TR#1) What is the independent variable in this study? (relevant section)

The following questions are from the ADHD Treatment (AT) case study.

Q17

(AT#1) What is the independent variable of this experiment? How many levels does it have? (relevant section)

Q18

(AT#2) What is the dependent variable? On what scale (nominal, ordinal, interval, ratio) was it measured? (relevant section)

Select Answers

S9

76

S10

$23,277,529$

1.5: PowerPoints

https://professormo.com/holistic/Powerpoint/ch01.pptx

CHAPTER OVERVIEW

2: Descriptive Statistics

2.1: Videos

Examining numerical data - Mean, standard deviation, histograms, box plots, and more

Considering categorical data - Table proportions, bar graphs, mosaic plots, and more

Case study: gender discrimination - Early inference ideas: testing using randomization

SECTION OVERVIEW

2.2: Graphing Distributions

Graphing data is the first and often most important step in data analysis. In this day of computers, researchers all too often see only the results of complex computer analyses without ever taking a close look at the data themselves. This is all the more unfortunate because computers can create many types of graphs quickly and easily. This chapter covers some classic types of graphs such as bar charts that were invented by William Playfair in the 18th century as well as graphs such as box plots invented by John Tukey in the 20th century.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

2.2.1: Graphing Qualitative Variables

Learning Objectives

- Create a frequency table
- Determine when pie charts are valuable and when they are not
- Create and interpret bar charts
- Identify common graphical mistakes

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table $2.2.1.1$shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" is $85/500 = 0.17$.

Table $2.2.1.1$: *Frequency Table for the iMac Data*

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Pie Charts

The pie chart in Figure $2.2.1.1$shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

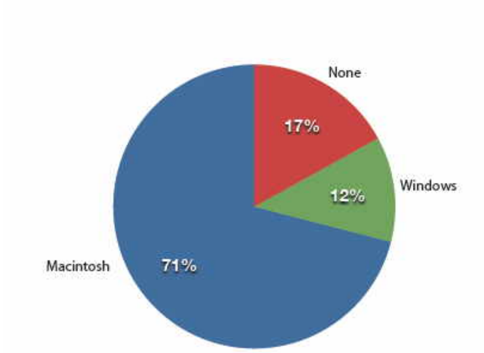


Figure $2.2.1.1$: Pie chart of iMac purchases illustrating frequencies of previous computer ownership

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted, "The only worse design than a pie chart is several of them."

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure $2.2.1.2$ Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations as is typical in pie charts.

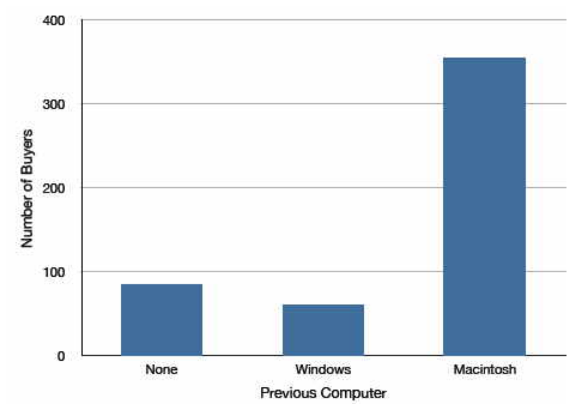


Figure $2.2.1.2$: Bar chart of iMac purchases as a function of previous computer ownership

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure $2.2.1.3$ shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

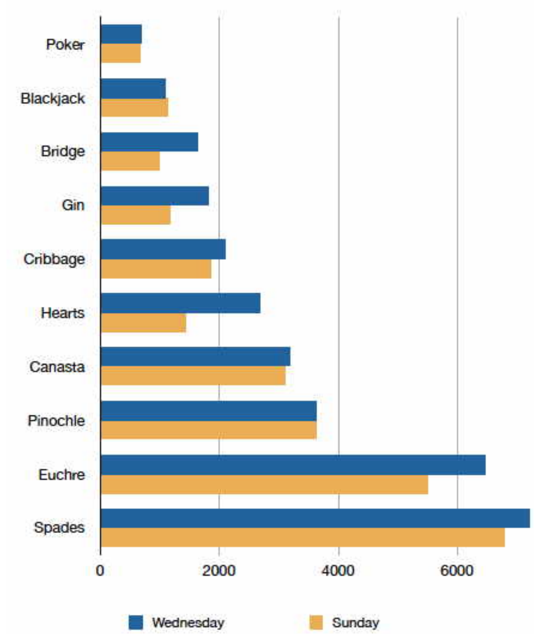


Figure $2.2.1.3$: A bar chart of the number of people playing different card games on Sunday and Wednesday

The bars in Figure $2.2.1.3$ are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in the section Bar Charts.

Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure $2.2.1.4$ are usually not as effective as their two-dimensional counterparts.

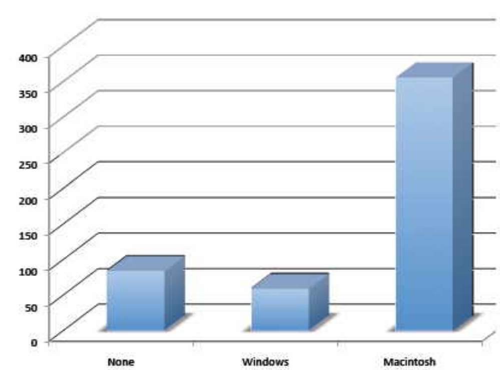


Figure $2.2.1.2$

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure $2.2.1.5$ presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure $2.2.1.5$ is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure $2.2.1.5$ is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure $2.2.1.5$ instead of Figure $2.2.1.2$! Edward Tufte coined the term "lie factor" to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

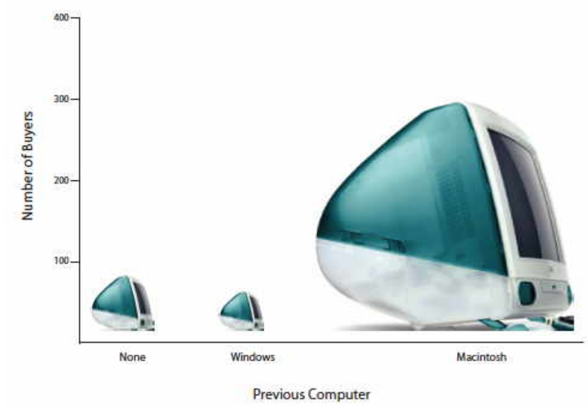


Figure 2.2.1.2 with a lie factor greater than 8

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure $2.2.1.6$ shows the iMac data with a baseline of 50. Once again, the differences in areas suggest a different story than the true differences in percentages. The percentage of Windows-switchers seems minuscule compared to its true value of 12%.

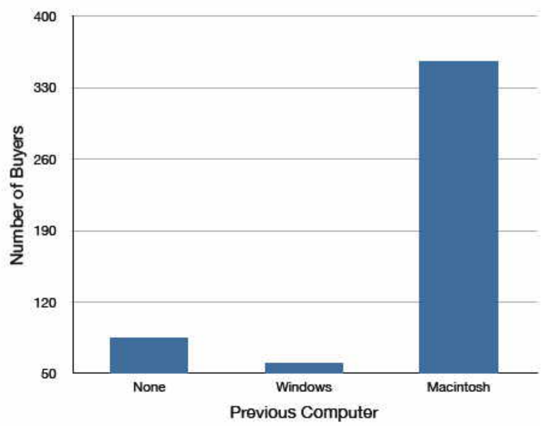


Figure 2.2.1.2 with a baseline of 50

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure $2.2.1.7$ inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure $2.2.1.7$ is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

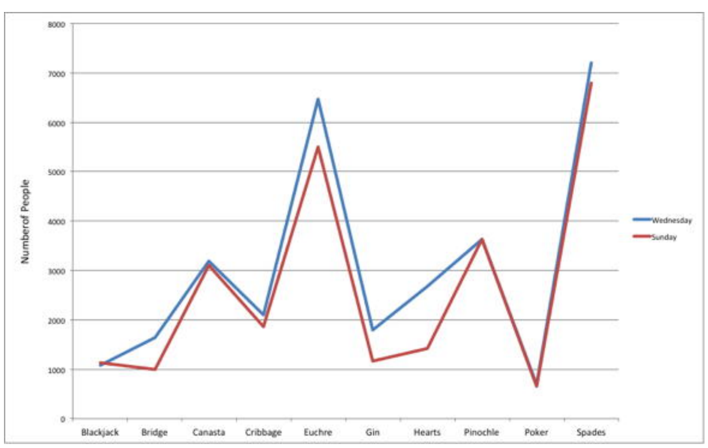


Figure $2.2.1.7$: *A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.*

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

2.2.2: Quantitative Variables

As discussed in the section on variables in Chapter 1, quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, and favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. The upcoming sections cover the following types of graphs:

1. stem and leaf displays
2. histograms
3. frequency polygons
4. box plots
5. bar charts
6. line graphs
7. scatter plots (discussed in a different chapter)
8. dot plots

Some graph types such as stem and leaf displays are best-suited for small to moderate amounts of data, whereas others such as histograms are best-suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

2.2.3: Stem and Leaf Displays

Learning Objectives

- Create and interpret basic stem and leaf displays
- Create and interpret back-to-back stem and leaf displays
- Judge whether a stem and leaf display is appropriate for a given data set

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, an example will get us started. Consider Table $2.2.3.1$ that shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League in the 2000 season.

Table* $2.2.3.1$*:* *Number of touchdown passes

$$\begin{array}{ccccccccc} 37 & 33 & 33 & 32 & 29 & 28 & 28 & 23 & 22 \\ 22 & 22 & 21 & 21 & 21 & 20 & 20 & 19 & 19 \\ 18 & 18 & 18 & 18 & 16 & 15 & 14 & 14 & 14 \\ 12 & 12 & 9 & 6 & & & & & \end{array} \tag{2.2.3.1}$$

A stem and leaf display of the data is shown in Figure $2.2.3.1$. The left portion of Figure $2.2.3.1$ contains the stems. They are the numbers $3, 2, 1, \ and\ 0$, arranged as a column to the left of the bars. Think of these numbers as $10's$ digits. A stem of 3, for example, can be used to represent the $10's$ digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the $1's$ digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

$$\begin{array}{c|cccccccccccc} 3 & 2 & 3 & 3 & 7 \\ 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 8 & 8 & 9 \\ 1 & 2 & 2 & 4 & 4 & 4 & 5 & 6 & 8 & 8 & 8 & 8 & 9 & 9 \\ 0 & 6 & 9 \end{array} \tag{2.2.3.2}$$

Figure* $2.2.3.1$*:* *Stem and leaf display of the number of touchdown passes

To make this clear, let us examine Figure $2.2.3.1$ more closely. In the top row, the four leaves to the right of stem 3 are $2, 3, 3, \ and\ 7$. Combined with the stem, these leaves represent the numbers $32, 33, 33, \ and\ 37$, which are the numbers of TD passes for the first four teams in Table $2.2.3.1$. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table $2.2.3.1$, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure $2.2.3.1$ than in Table $2.2.3.1$. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TDs, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure $2.2.3.2$ shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table $2.2.3.1$. The second row is reserved for the numbers from 30 to 34 and holds the $32, 33, \ and\ 33$ TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

$$\begin{array}{c|ccccccccc} 3 & 7 \\ 3 & 2 & 3 & 3 \\ 2 & 8 & 8 & 9 \\ 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 \\ 1 & 5 & 6 & 8 & 8 & 8 & 8 & 9 & 9 \\ 1 & 2 & 2 & 4 & 4 & 4 \\ 0 & 6 & 9 \end{array} \tag{2.2.3.3}$$

Figure* $2.2.3.2$*:* *Stem and leaf display with the stems split in two

Figure $2.2.3.2$ is more revealing than Figure $1 PageIndex 2$ because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a "back-to-back stem and leaf graph." Figure $2.2.3.3$ shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with $11, 12, \ and\ 13$ TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

$$\begin{array}{r|c|ccccccccc} 1\ 1 & 4 & \\ & 3 & 7 \\ 3\ 3\ 2 & 3 & 2 & 3 & 3 \\ 8\ 8\ 6\ 5 & 2 & 8 & 8 & 9 \\ 4\ 4\ 3\ 3\ 1\ 1\ 1\ 0 & 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 \\ 9\ 8\ 7\ 7\ 7\ 6\ 6\ 6\ 5 & 1 & 5 & 6 & 8 & 8 & 8 & 8 & 9 & 9 \\ 3\ 2\ 1 & 1 & 2 & 2 & 4 & 4 & 4 & 4 \\ 7 & 0 & 6 & 9 \end{array} \tag{2.2.3.4}$$

Figure* $2.2.3.3$*:* *Back-to-back stem and leaf display.

The left side shows the 1998 TD data and the right side shows the 2000 TD data. Figure $2.2.3.3$ helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

Table $2.2.3.2$ shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like “punch”) under two conditions. In one condition, the words were preceded by a non-weapon word such as "bug." In the second condition, the same words were preceded by a weapon word such as "gun" or "knife." The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

Table $2.2.3.2$: *The effects of priming (thousandths of a second)*

$$\begin{array}{cccccccc} 43.2 & 42.9 & 35.6 & 25.6 & 25.4 & 23.6 & & \\ 20.5 & 19.9 & 14.4 & 12.7 & 11.3 & 10.2 & & \\ 10.0 & 9.1 & 7.5 & 5.4 & 4.7 & 3.8 & 2.1 & 1.2 \\ -0.2 & -6.3 & -6.7 & -8.8 & -10.4 & -10.5 & & \\ -14.9 & -14.9 & -15.0 & -18.5 & -27.4 & & & \end{array} \tag{2.2.3.5}$$

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure $2.2.3.4$ Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf), the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number -27. The second-to-last row represents the numbers $-10, -10, -15$, etc. Once again, we have rounded the original values from Table $2.2.3.2$

$$\begin{array}{r|lllllll} 4 & 3 & 3 & & & & & \\ 3 & 6 & & & & & & \\ 2 & 0 & 0 & 4 & 5 & 6 & & \\ 1 & 0 & 0 & 1 & 3 & 4 & & \\ 0 & 1 & 2 & 4 & 5 & 5 & 8 & 9 \\ -0 & 0 & 6 & 7 & 9 & & & \\ -1 & 0 & 0 & 5 & 5 & 5 & 9 & \\ -2 & 7 & & & & & & \end{array} \tag{2.2.3.6}$$

Figure $2.2.3.4$: *Stem and leaf display with negative numbers and rounding*

Observe that the figure contains a row headed by "0" and another headed by "-0". The stem of 0 is for numbers between 0 and 9, whereas the stem of -0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers $1, 2, 4, 5, 5, 8, 9$ and the sixth row holds $0, -6, -7, \ and\ -9$. Values that are exactly 0 before rounding should be split as evenly as possible between the "0" and "-0" rows. In Table $2.2.3.2$, none of the values are 0 before rounding. The "0" that appears in the "-0" row comes from the original value of -0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure $2.2.3.5$ portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between $100,000$ and $500,000$ residents.

$$\begin{array}{r|l} 4 & 8\ 9\ 9 \\ 4 & 6 \\ 4 & 4\ 4\ 5\ 5 \\ 4 & 3\ 3\ 3 \\ 4 & 0\ 1 \\ 3 & 9\ 9 \\ 3 & 6\ 7\ 7\ 7\ 7\ 7 \\ 3 & 5\ 5 \\ 3 & 2\ 2\ 3 \\ 3 & 1\ 1\ 1 \\ 2 & 8\ 8\ 9\ 9 \\ 2 & 6\ 6\ 6\ 6\ 6\ 7 \\ 2 & 4\ 4\ 4\ 4\ 5\ 5 \\ 2 & 2\ 2\ 3\ 3\ 3 \\ 2 & 0\ 0\ 0\ 0\ 0\ 0 \\ 1 & 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 9\ 9\ 9\ 9\ 9\ 9\ 9\ 9\ 9\ 9\ 9 \\ 1 & 6\ 6\ 6\ 6\ 6\ 6\ 7\ 7\ 7\ 7\ 7\ 7 \\ 1 & 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 5\ 5\ 5\ 5\ 5\ 5\ 5\ 5\ 5\ 5\ 5\ 5 \\ 1 & 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 3\ 3\ 3\ 3\ 3\ 3\ 3\ 3 \\ 1 & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1 \end{array} \tag{2.2.3.7}$$

Figure $2.2.3.5$: *Stem and leaf display of populations of* 185 *US cities with populations between* $100,000$ *and* $500,000$ *in* 1998.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest $10,000$. For example, the largest number ($493,559$) was rounded to $490,000$ and then plotted with a stem of 4 and a leaf of 9. The fourth highest number ($463,201$) was rounded to $460,000$ and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of $100,000$ and the leaves represent units of $10,000$. Notice that each stem value is split into five parts: $0-1, 2-3, 4-5, 6-7$, and $8-9$.

Whether your data can be suitably represented by a stem and leaf graph depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure $2.2.3.5$ fit into the $10,000$ and $100,000$ places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

2.2.4: Histograms

Learning Objectives

- Create a grouped frequency distribution
- Create a histogram based on a grouped frequency distribution
- Determine an appropriate bin width

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items, each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table $2.2.4.1$.

Table $2.2.4.1$: *Grouped Frequency Distribution of Psychology Test Scores*

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into intervals, called class intervals. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure $2.2.4.1$.

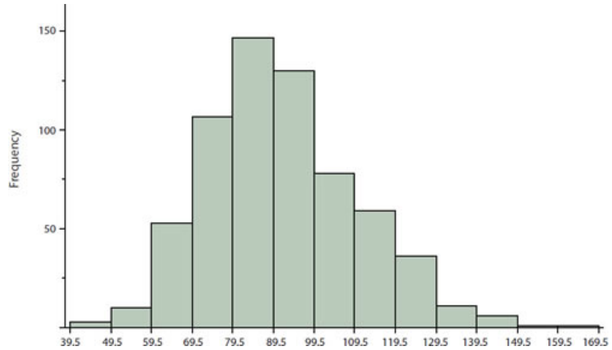


Figure $2.2.4.1$: Histogram of scores on a psychology test.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We'll have more to say about shapes of distributions in the chapter " Summarizing Distributions.")

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence-sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, $4000, 5000$ etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

Sturges' rule

There is more to be said about the widths of the class intervals, sometimes called bin widths. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. There are some "rules of thumb" that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) Sturges' rule is to set the number of intervals as close as possible to $1+\log_2(N)$, where $\log_2(N)$ is the base 2 log of the number of observations. The formula can also be written as $1+3.3\log_2(N)$, where $\log_{10}(N)$ is the log base 10 of the number of observations. According to Sturges' rule, 1000 observations would be graphed with 11 class intervals since 10 is the closest integer to $\log_2(1000)$. We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals instead of the 11 recommended by Sturges' rule. For the psychology test example used above, Sturges' rule recommends 10 intervals while the Rice rule recommends 17. In the end, we compromised and chose 13 intervals for Figure $2.2.4.1$to create a histogram that seemed clearest. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

To provide experience in constructing histograms, we have developed an interactive demonstration. The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

Interactive histogram

2.2.5: Frequency Polygons

Learning Objectives

- Create and interpret frequency polygons
- Create and interpret cumulative frequency polygons
- Create and interpret overlaid frequency polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

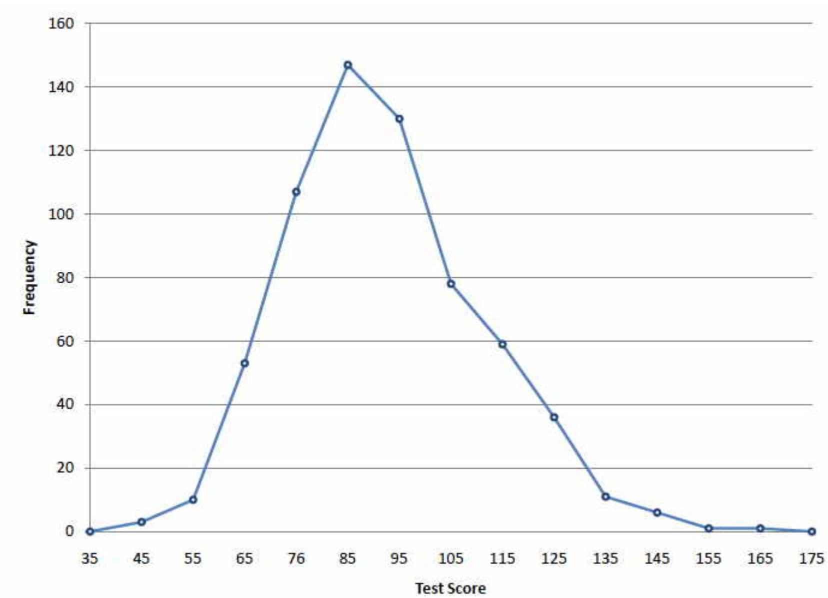


Figure $2.2.5.1$: Frequency polygon for the psychology test scores

A frequency polygon for 642 psychology test scores shown in Figure $2.2.5.1$ was constructed from the frequency table shown in Table $2.2.5.1$.

Table $2.2.5.1$: Frequency Distribution of Psychology Test Scores.

Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173
79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528

Lower Limit	Upper Limit	Count	Cumulative Count
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	179.5	0	642

The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure $2.2.5.1$ Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

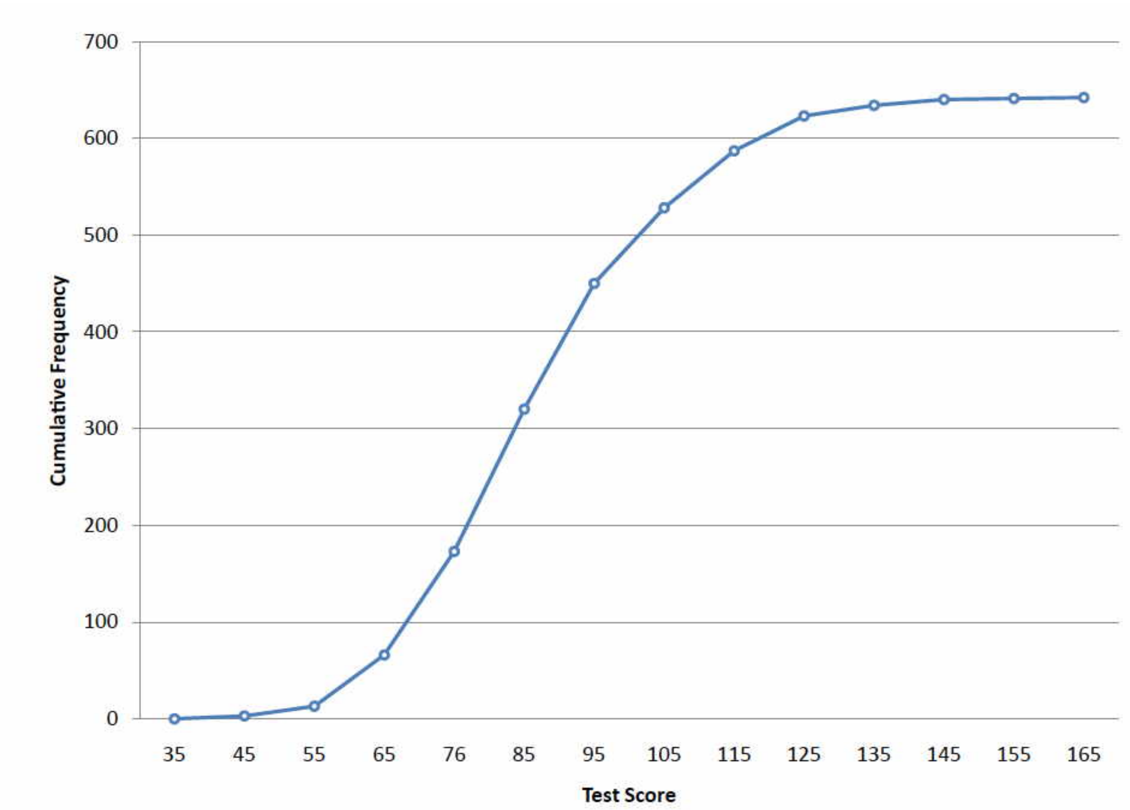


Figure $2.2.5.2$: Cumulative frequency polygon for the psychology test scores

A cumulative frequency polygon for the same test scores is shown in Figure $2.2.5.2$ The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled 35, three in the interval 45, and 10 in the interval 55. Therefore, the Y value corresponding to "55" is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

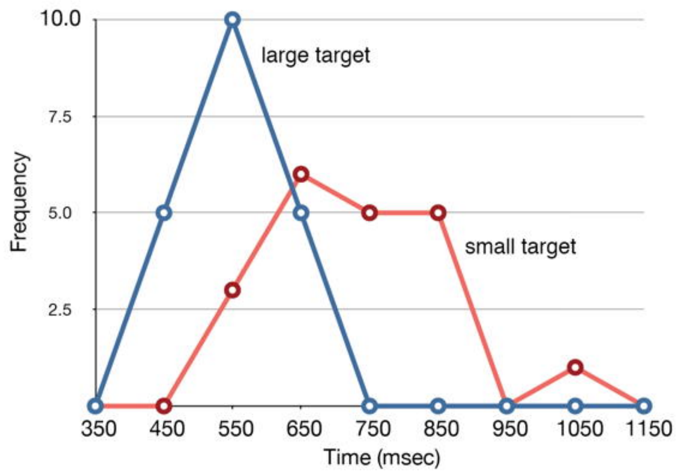


Figure 2.2.5.3: Overlaid frequency polygons

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 2.2.5.3 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 2.2.5.3 The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

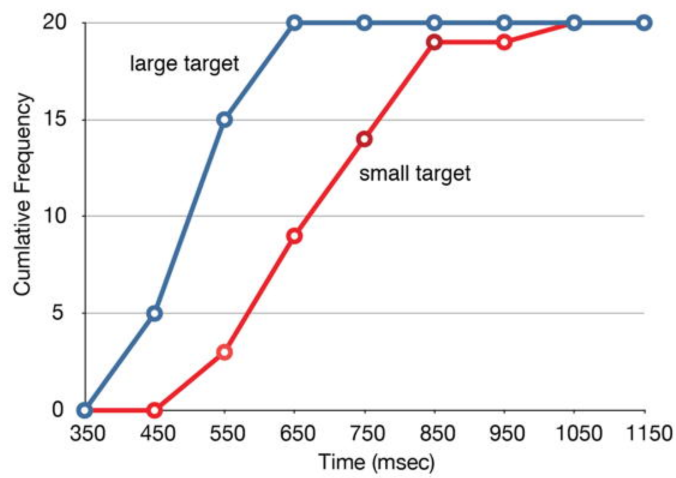


Figure 2.2.5.1: Overlaid cumulative frequency polygons

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 2.2.5.4 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

2.2.6: Box Plots

Learning Objectives

- Define basic terms including hinges, H-spread, step, adjacent value, outside value, and far out value
- Create a box plot
- Create parallel box plots
- Determine whether a box plot is appropriate for a given data set

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section, we present another important graph called a box plot. Box plots are useful for identifying outliers and for comparing distributions. We will explain box plots with the help of data from an in-class experiment. As part of the "Stroop Interference Case Study," students in introductory statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve parallel box plots.

There are several steps in constructing a box plot. The first relies on the 25^{th}, 50^{th}, $and\ 75^{th}$ percentiles in the distribution of scores. Figure $2.2.6.1$shows how these three statistics are used. For each gender, we draw a box extending from the 25^{th} percentile to the 75^{th} percentile. The 50^{th} percentile is drawn inside the box. Therefore,

- the bottom of each box is the 25^{th} percentile,
- the top is the 75^{th} percentile,
- and the line in the middle is the 50^{th} percentile.

The data for the women in our sample are shown in Table $2.2.6.1$.

Table $2.2.6.1$: Women's times

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25^{th} percentile is 17, the 50^{th} percentile is 19, and the 75^{th} percentile is 20. For the men (whose data are not shown), the 25^{th} percentile is 19, the 50^{th} percentile is 22.5, and the 75^{th} percentile is 25.5.

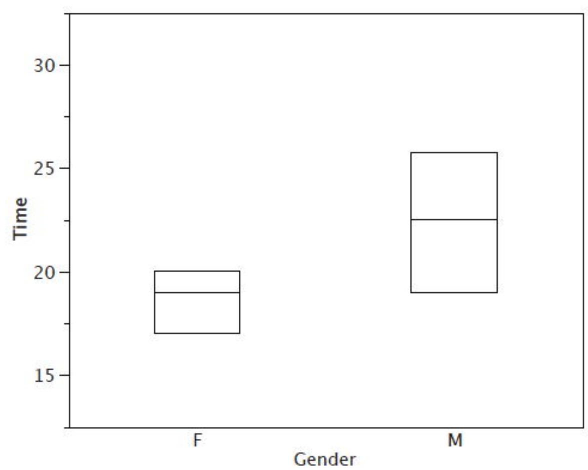


Figure $2.2.6.1$: The first step in creating box plots

Before proceeding, the terminology in Table $2.2.6.2$is helpful.

Table $2.2.6.2$: Box plot terms and values for women's times

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17
H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put "whiskers" above and below each box to give additional information about the spread of the data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women's data).

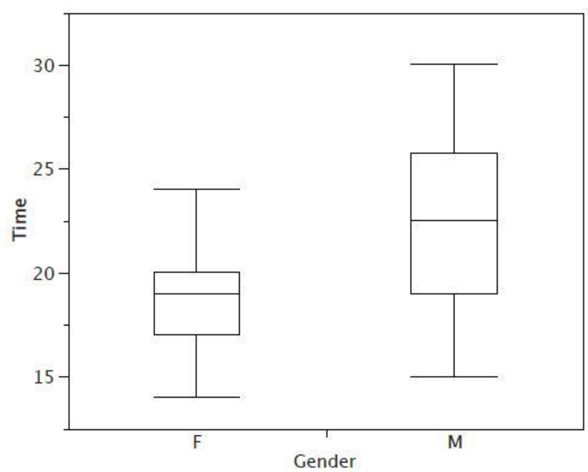


Figure $2.2.6.2$: The box plots with the whiskers drawn

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "$o's$" and far out values are indicated by asterisks ($*$). In our data, there are no far out values and just one outside value. This outside value of 29 is for the women and is shown in Figure $2.2.6.3$.

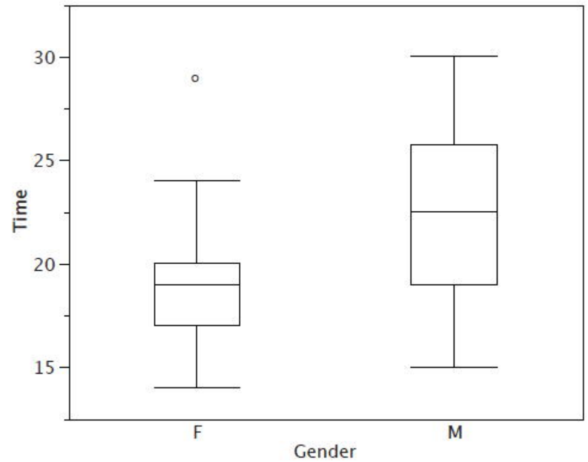


Figure 2.2.6.3: The box plots with the outside value shown

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 2.2.6.4 shows the result of adding means to our box plots.

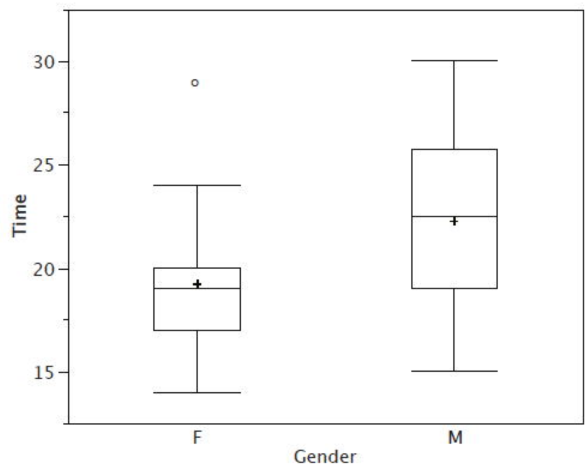


Figure 2.2.6.4: The completed box plots

Figure 2.2.6.4 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25^{th} and 75^{th} percentiles), we see that half the women's times are between 17 and 20 seconds, whereas half the men's times are between 19 and 25.5. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 2.2.6.5 shows the box plot for the women's data with detailed labels.

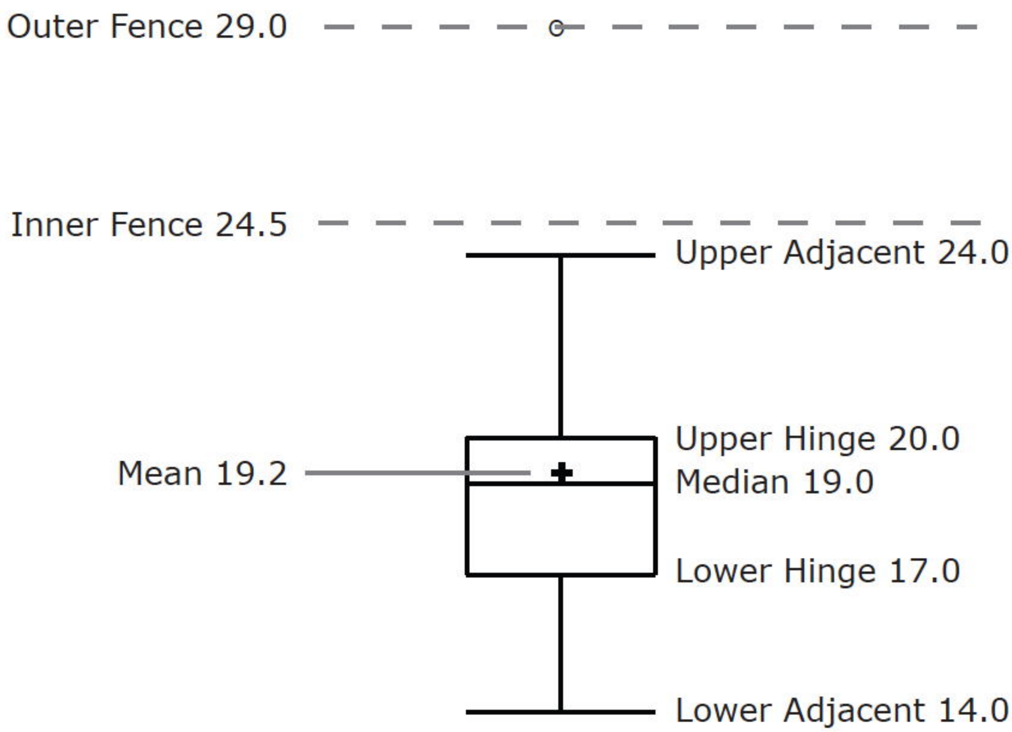


Figure 2.2.6.5: The box plot for the women's data with detailed labels

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot, and to examine these details one should create a histogram and/or a stem and leaf display.

Here are some other examples of box plots:

Example 2.2.6.1: Time to move the mouse over a target

The data come from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The box plots of the two distributions are shown below. You can see that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one.

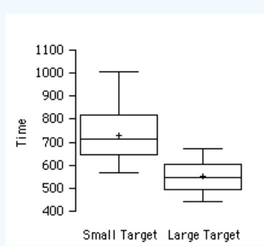


Figure 2.2.6.6: Target Box Plots

Example 2.2.6.2: Draft lottery

In 1969 the war in Vietnam was at its height. An agency called the *Selective Service* was charged with finding a fair procedure to determine which young men would be conscripted ("drafted") into the U.S. military. The procedure was supposed to be fair in the sense of not favoring any culturally or economically defined subgroup of American men. It was decided that choosing "draftees" solely on the basis of a person’s birth date would be fair. A birthday lottery was thus devised. Pieces of paper representing the 366 days of the year (including February 29) were placed in plastic capsules, poured into a rotating drum, and then selected one at a time. The lower the draft number, the sooner the person would be drafted. Men with high enough numbers were not drafted at all.

The first number selected was 258, which meant that someone born on the 258^{th} day of the year ($\text{September } 14$) would be among the first to be drafted. The second number was 115, so someone born on the 115^{th} day ($\text{April } 24$) was among the second group to be drafted. All 366 birth dates were assigned draft numbers in this way.

To crate box plots, we divided the 366 days of the year into thirds. The first third goes from January 1 to May 1, the second from May 2 to August 31, and the last from September 1 to December 31. The three groups of birth dates yield three groups of draft numbers. The draft number for each birthday is the order it was picked in the drawing. The figure below contains box plots of the three sets of draft numbers. As you can see, people born later in the year tended to have lower draft numbers.

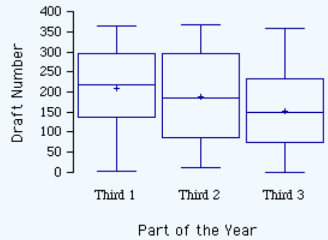


Figure $2.2.6.7$: Draft Lottery Box Plots

Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure $2.2.6.8$ are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.
5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

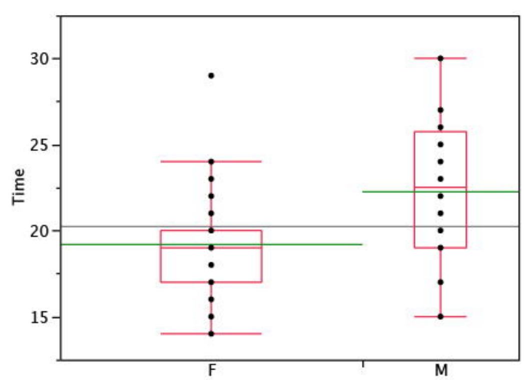


Figure $2.2.6.8$: Box plots showing the individual scores and the means

Each dot in Figure $2.2.6.8$ represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter thepoints. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittererd. Figure $2.2.6.9$shows what jittering looks like.

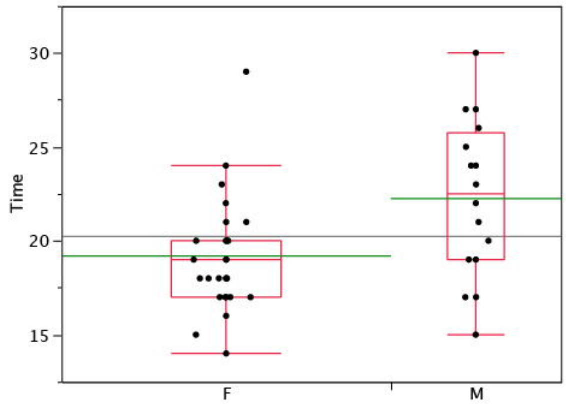


Figure 2.2.6.9: Box plots with the individual scores jittered

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

2.2.7: Box Plot Demo

Learning Objectives

- Understand what it means for a distribution to balance on a fulcrum
- Learn which measure of central tendency will balance a distribution

Instructions

- The box plot in the simulation is based on the data shown to the left of the box plot.
- Notice that various aspects of the box plot such as the mean and median are labeled.
  - These labels can be hidden by unchecking the "show labels on box plot" button.
- Beneath the data is a pair of buttons that let you specify whether you want to be able to enter data or to see statistics based on the data.
- When you use the simulation, try to modify the data in various ways and see how it affects the box plot.
  - Try putting in some extreme values and see if they get labeled as outliers.
  - Outside values are shown as "$o's$" beyond the inner fence.
  - Far out values are shown as $*'s$ outside the outer fence.
- You can delete all the data (by pressing the "Clear All button") and enter your own data.
  - Your data can be typed in or pasted in from another application.
  - When pasting, use keyboard shortcut for pasting (Command-V for Mac, CTRL-V for Windows).
  - When you change the data, the box plot will disappear.

After you have entered new data, click the "Draw box plot" button to redraw the box plot.

Illustrated Instructions

The screenshot below shows the box plot simulation with its default data on the left.

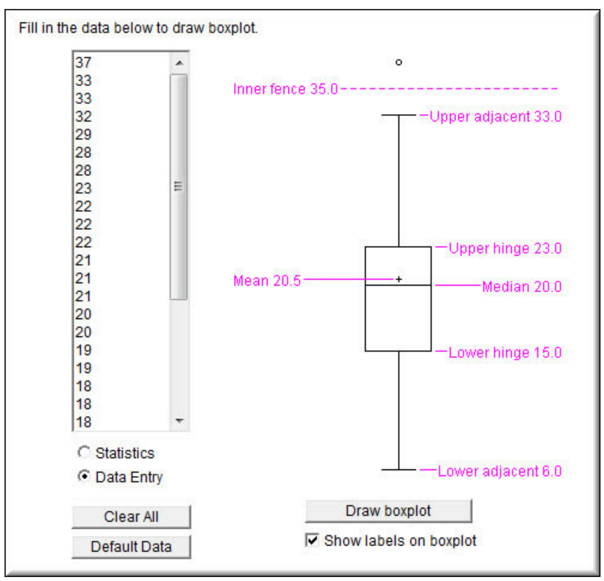


Figure 2.2.7.1: Box Plot simulation

You can change some or all of the data on the left by editing individual numbers or pasting in your own data. When pasting, you must use the keyboard shortcut for pasting (Command-V for Mac, CTRL-V for Windows).

2.2.8: Bar Charts

Learning Objectives

- Create and interpret bar charts
- Judge whether a bar chart or another graph such as a box plot would be more appropriate

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure $2.2.8.1$ shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

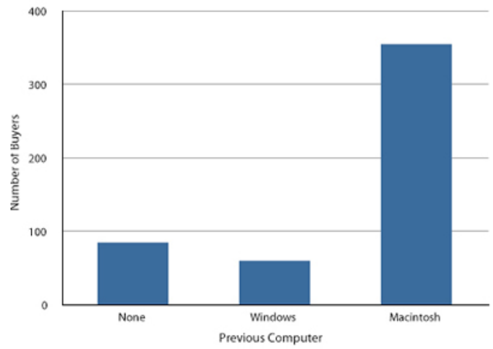


Figure $2.2.8.1$: iMac buyers as a function of previous computer ownership

In this section, we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure $2.2.8.2$ shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from $\text{May } 24\ 2000$ to $\text{May } 24\ 2001$. Notice that both the S & P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity *percentage increase*.

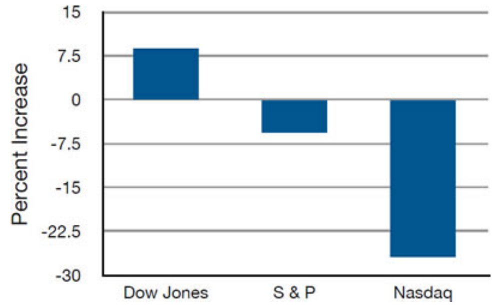


Figure $2.2.8.2$: Percent increase in three stock indexes from May 24th 2000 to May 24th 2001

Bar charts are particularly effective for showing change over time. Figure $2.2.8.3$, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

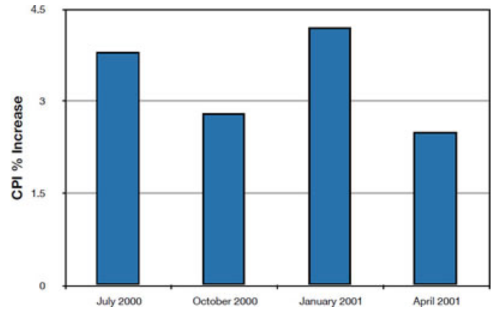


Figure $2.2.8.3$: Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated

Bar charts are often used to compare the means of different experimental conditions. Figure $2.2.8.4$ shows the mean time it took one of us (DL) to move the mouse to either a small target or a large target. On average, more time was required for small targets than for large ones.

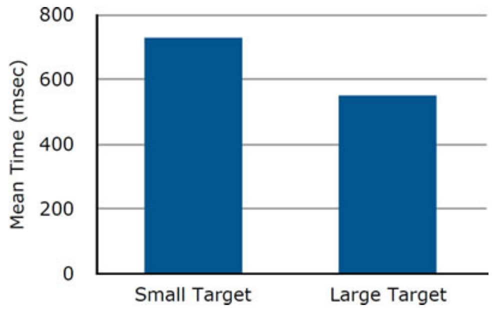


Figure $2.2.8.4$: Bar chart showing the means for the two conditions

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the mouse-movement data is shown in Figure $2.2.8.5$. You can see that Figure $2.2.8.5$ reveals more about the distribution of movement times than does Figure $2.2.8.4$.

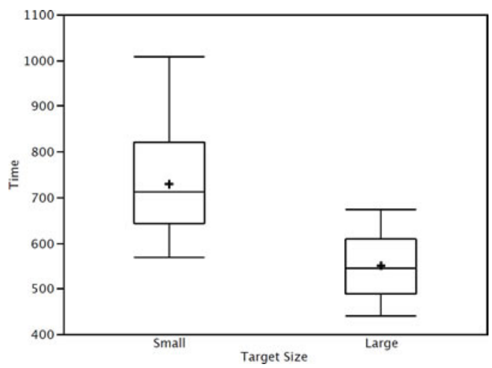


Figure $2.2.8.5$: Box plots of times to move the mouse to the small and large targets

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

2.2.9: Line Graphs

Learning Objectives

- Create and interpret line graphs
- Judge whether a line graph would be appropriate for a given data set

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure $2.2.9.1$ was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

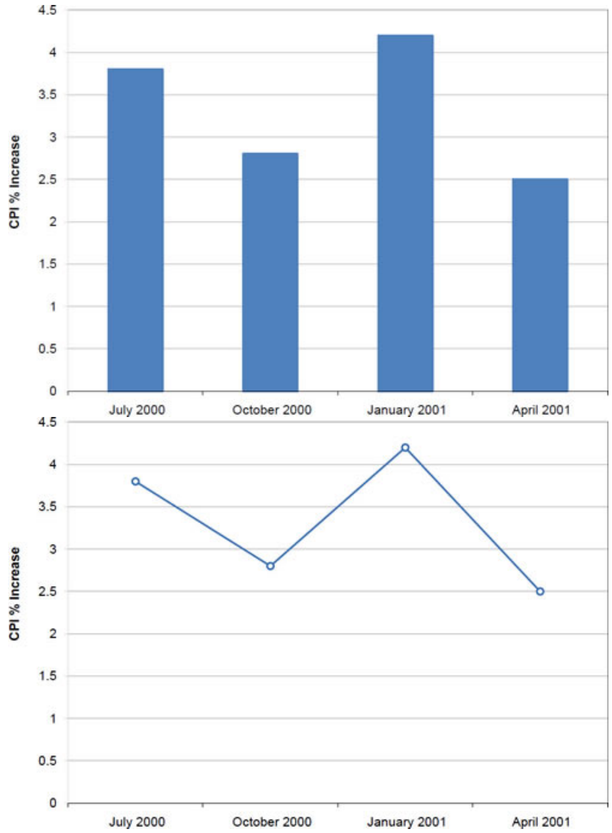


Figure $2.2.9.2$: A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

A line graph of these same data is shown in Figure $2.2.9.2$ Although the figures are similar, the line graph emphasizes the change from period to period.

Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables. Although bar graphs can also be used in this situation, line graphs are generally better at comparing changes over time. Figure $2.2.9.3$, for example, shows percent increases and decreases in five components of the Consumer Price Index (CPI). The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy.

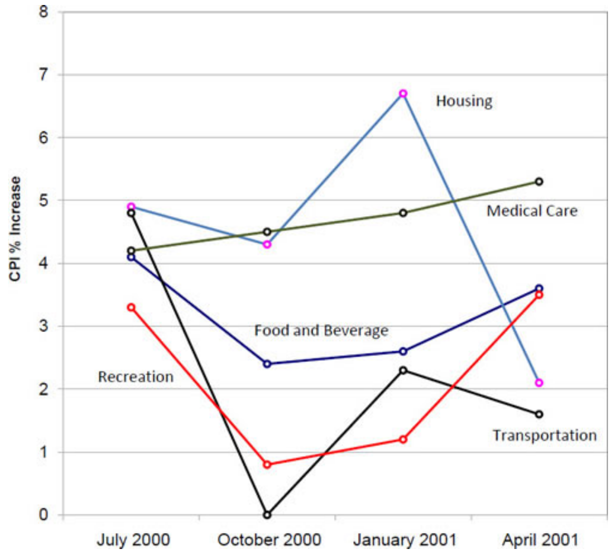


Figure 2.2.9.3: A line graph of the percent change in five components of the CPI over time.

Let us stress that it is misleading to use a line graph when the X-axis contains merely qualitative variables. Figure 2.2.9.4 inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure 2.2.9.4is that it gives the false impression that the games are naturally ordered in a numerical way.

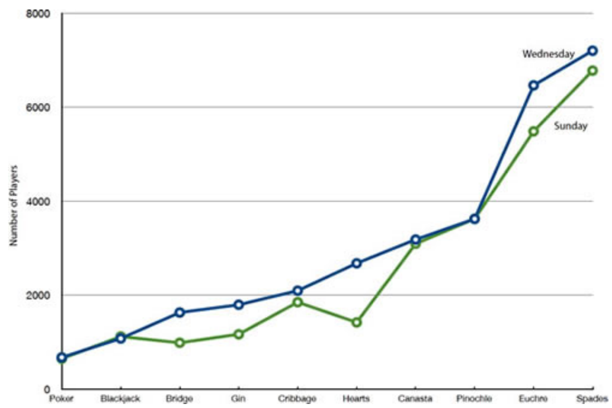


Figure 2.2.9.4: A line graph, inappropriately used, depicting the number of people playing different card games on Sunday and Wednesday.

2.2.10: Dot Plots

Learning Objectives

- Create and interpret dot plots
- Judge whether a dot plot would be appropriate for a given data set

Dot plots can be used to display various types of information. Figure $2.2.10.1$ uses a dot plot to display the number of M & M's of each color found in a bag of M & M's. Each dot represents a single M & M. From the figure, you can see that there were 3 blue M & M's, 19 brown M & M's, etc.

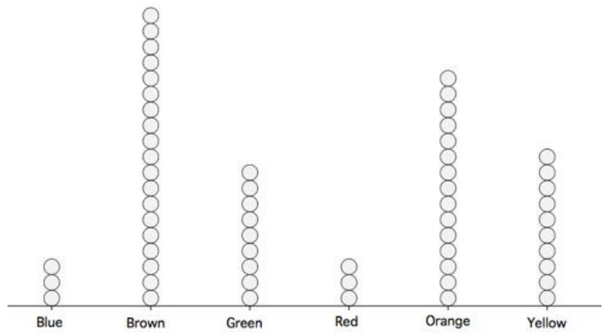


Figure 2.2.10.1: A dot plot showing the number of M & M's of various colors in a bag of M & M's

The dot plot in Figure $2.2.10.2$ shows the number of people playing various card games on the Yahoo website on a Wednesday. Unlike Figure $2.2.10.1$, the location rather than the number of dots represents the frequency.

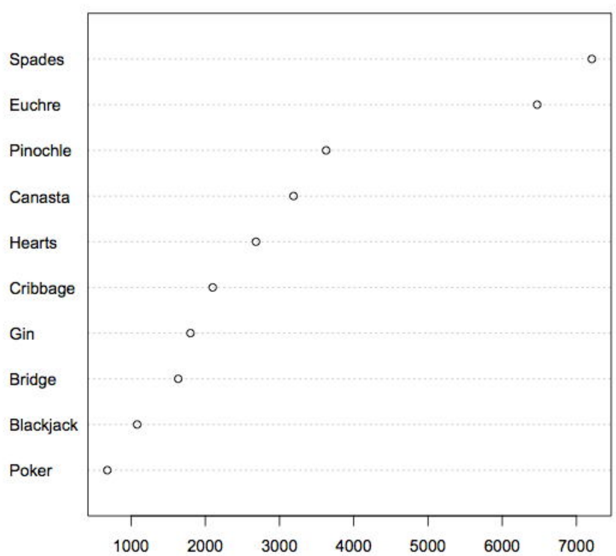


Figure 2.2.10.2: A dot plot showing the number of people playing various card games on a Wednesday

The dot plot in Figure $2.2.10.3$ shows the number of people playing on a Sunday and on a Wednesday. This graph makes it easy to compare the popularity of the games separately for the two days, but does not make it easy to compare the popularity of a given game on the two days.

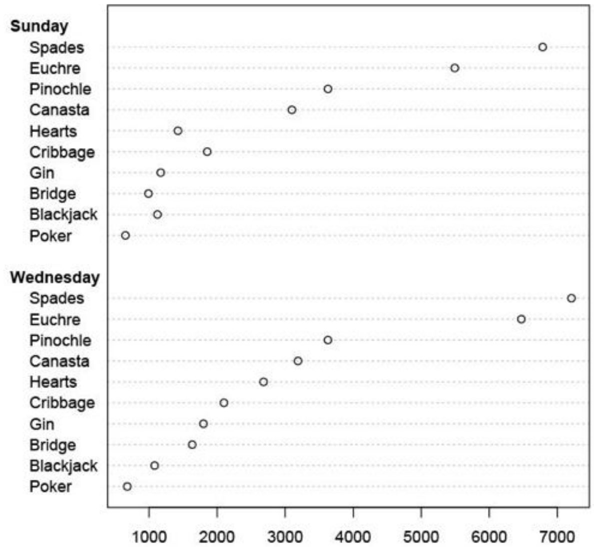


Figure 2.2.10.3: A dot plot showing the number of people playing various card games on a Sunday and on a Wednesday

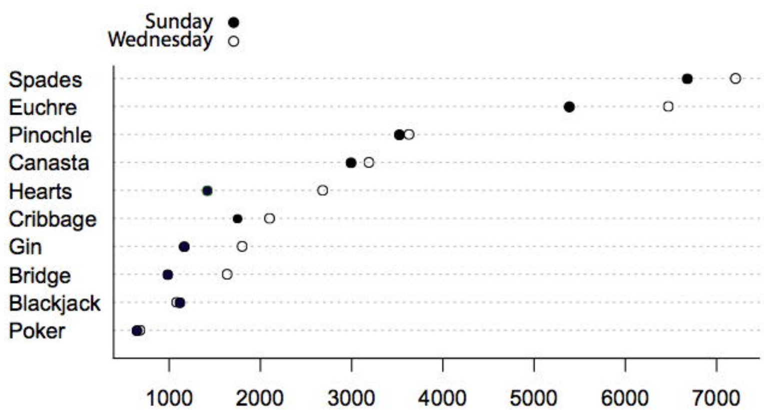


Figure 2.2.10.4: An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday

The dot plot in Figure 2.2.10.4makes it easy to compare the days of the week for specific games while still portraying differences among games.

2.2.11: Statistical Literacy

Are Commercial Vehicles in Texas Unsafe?

Prerequisites

Graphing Distributions

A news report on the safety of commercial vehicles in Texas stated that one out of five commercial vehicles have been pulled off the road in 2012 because they were unsafe. In addition, $12,301$ commercial drivers have been banned from the road for safety violations.

The author presents the bar chart below to provide information about the percentage of fatal crashes involving commercial vehicles in Texas since 2006. The author also quotes DPS director Steven McCraw:

> Commercial vehicles are responsible for approximately 15 percent of the fatalities in Texas crashes. Those who choose to drive unsafe commercial vehicles or drive a commercial vehicle unsafely pose a serious threat to the motoring public.

Example 2.2.11.1

Based on what you have learned in this chapter, does this bar chart below provide enough information to conclude that unsafe or unsafely driven commercial vehicles pose a serious threat to the motoring public? What might you conclude if 30 percent of all the vehicles on the roads of Texas in 2010 were commercial and accounted for 16 percent of fatal crashes?

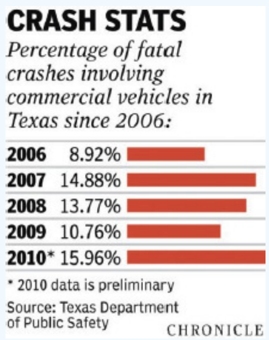


Figure 2.2.11.1: Crash Statistics for commercial vehicles in Texas

Solution

This bar chart does not provide enough information to draw such a conclusion because we don't know, on the average, in a given year what percentage of all vehicles on the road are commercial vehicles. For example, if 30 percent of all the vehicles on the roads of Texas in 2010 are commercial ones and only 16 percent of fatal crashes involved commercial vehicles, then commercial vehicles are safer than non-commercial ones. Note that in this case 70 percent of vehicles are non-commercial and they are responsible for 84 percent of the fatal crashes.

Linear By Design

Example 2.2.11.2

Fox News aired the line graph below showing the number unemployed during four quarters between 2007 and 2010.

Figure 2.2.11.2: Fox news graph showing job loss by quarter

Does Fox News' line graph provide misleading information? Why or Why not?

Solution:

There are major flaws with the Fox News graph. First, the title of the graph is misleading. Although the data show the number unemployed, Fox News' graph is titled "Job **Loss** by Quarter." Second, the intervals on the X-axis are misleading. Although there are 6 months between September 2008 and March 2009 and 15 months between March 2009 and June 2010, the intervals are represented in the graph by very similar lengths. This gives the false impression that unemployment increased steadily.

The graph presented below is corrected so that distances on the X-axis are proportional to the number of days between the dates. This graph shows clearly that the rate of increase in the number unemployed is greater between September 2008 and March 2009 than it is between March 2009 and June 2010.

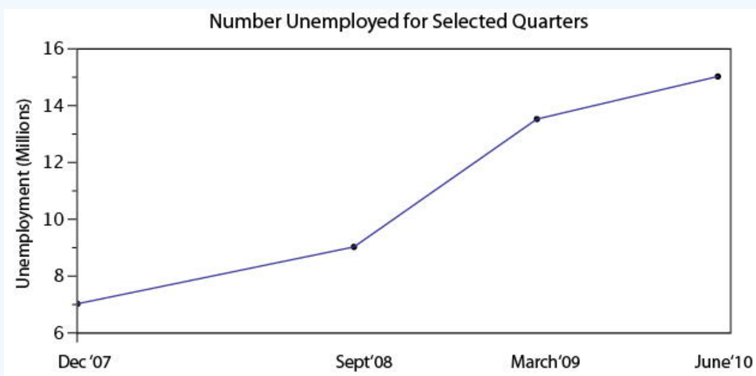


Figure 2.2.11.3: Corrected Fox News graph

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Seyd Ercan and David Lane

2.2.E: Graphing Distributions (Exercises)

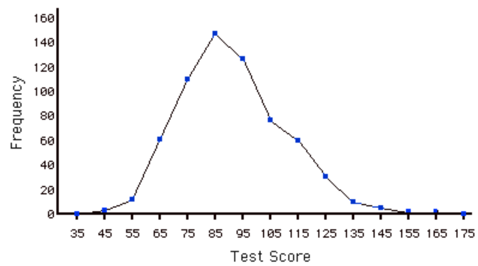
General Questions

Q1

Name some ways to graph quantitative variables and some ways to graph qualitative variables. (relevant section & relevant section)

Q2

Based on the frequency polygon displayed below, the most common test grade was around what score? Explain. (relevant section)



Q3

An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from **three** chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots? (relevant section)

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

Q4

You have to decide between displaying your data with a histogram or with a stem and leaf display. What factor(s) would affect your choice? (relevant section & relevant section)

Q5

In a box plot, what percent of the scores are between the lower and upper hinges? (relevant section)

Q6

A student has decided to display the results of his project on the number of hours people in various countries slept per night. He compared the sleeping patterns of people from the US, Brazil, France, Turkey, China, Egypt, Canada, Norway, and Spain. He was planning on using a line graph to display this data. Is a line graph appropriate? What might be a better choice for a graph? (relevant section & relevant section)

Q7

For the data from the 1977 $\mathrm{Stat.\ and\ Biom.\ 200}$ class for eye color, construct: (relevant section)

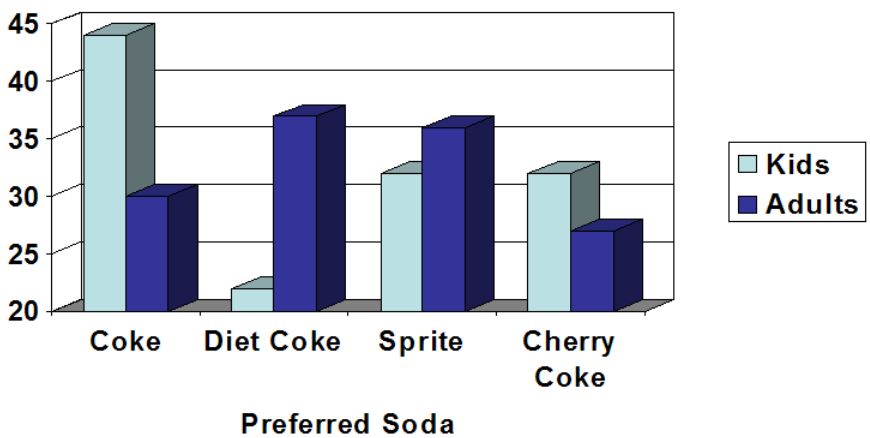
a. pie graph
b. horizontal bar graph
c. vertical bar graph
d. a frequency table with the relative frequency of each eye color

Eye Color	Number of students
Brown	11
Blue	10
Green	4
Gray	1

(Question submitted by J. Warren, UNH)

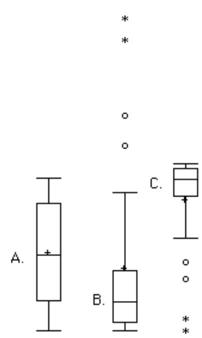
Q8

A graph appears below showing the number of adults and children who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph below could be improved. (relevant section)



Q9

Which of the box plots below has a large positive skew? Which has a large negative skew? (relevant section & relevant section)



Questions from Case Studies

The following questions are from the Angry Moods (AM) case study.

Q10

(AM#6) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the "Anger-Out" scores:

a. Create parallel box plots. (relevant section)
b. Create a back to back stem and leaf displays (You may have trouble finding a computer to do this so you may have to do it by hand.) (relevant section)

Q11

(AM#9) Create parallel box plots for the Anger-In scores by sports participation. (relevant section)

Q12

(AM#11) Plot a histogram of the distribution of the Control-Out scores. (relevant section)

Q13

(AM#14) Create a bar graph comparing the mean Control-In score for the athletes and the non-athletes. What would be a better way to display this data? (relevant section)

Q14

(AM#18) Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger? (relevant section)

The following questions are from the Flatulence (F) case study.

Q15

(F#1) Plot a histogram of the variable "per day." (relevant section)

Q16

(F#7) Create parallel box plots of "how long" as a function gender. Why is the 25^{th} percentile not showing? What can you say about the results? (relevant section)

Q17

(F#9) Create a stem and leaf plot of the variable "how long" What can you say about the shape of the distribution? (relevant section.1)

The following questions are from the Physicians' Reactions (PR) case study.

Q18

(PR#1) Create box plots comparing the time expected to be spent with the average-weight and overweight patients. (relevant section)

Q19

(PR#4) Plot histograms of the time spent with the average-weight and overweight patients. (relevant section)

Q20

(PR#5) To which group does the patient with the highest expected time belong?

The following questions are from the Smiles and Leniency (SL) case study

Q21

(SL#1) Create parallel box plots for the four conditions. (relevant section)

Q22

(SL#3) Create back to back stem and leaf displays for the false smile and neutral conditions. (It may be hard to find a computer program to do this for you, so be prepared to do it by hand). (relevant section)

The following questions are from the ADHD Treatment (AT) case study.

Q23

(AT#3) Create a line graph of the data. Do certain dosages appear to be more effective than others? (relevant section)

Q24

(AT#5) Create a stem and leaf plot of the number of correct responses of the participants after taking the placebo ($d0$ variable). What can you say about the shape of the distribution? (relevant section)

Q25

Create box plots for the four conditions. You may have to rearrange the data to get a computer program to create the box plots.

The following question is from the SAT and College GPA case study.

Q26

Create histograms and stem and leaf displays of both high-school grade point average and university grade point average. In what way(s) do the distributions differ?

Q27

The April 10th issue of the Journal of the American Medical Association reports a study on the effects of anti-depressants. The study involved 340 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort (an herb), Zoloft (Pfizer's cousin of Lilly's Prozac) or placebo for an 8-week period. The following are the mean scores (approximately) for the three groups of subjects over the eight-week experiment. The first column is the baseline. Lower scores mean less depression. Create a graph to display these means.

Placebo	22.5	19.1	17.9	17.1	16.2	15.1	12.1	12.3
Wort	23.0	20.2	18.2	18.0	16.5	16.1	14.2	13.0
Zoloft	22.4	19.2	16.6	15.5	14.2	13.1	11.8	10.5

The following questions are from . Visit the site

Q28

For the graph below, of heights of singers in a large chorus, please write a complete description of the histogram. Be sure to comment on all the important features.

Q29

Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired.

a. What is on the Y-axis? Explain.
b. What is on the X-axis?
c. What would be the probable shape of the salary distribution? Explain why.

Select Answers

S2

85

2.3: PowerPoints

https://professormo.com/holistic/Powerpoint/ch02.pptx

CHAPTER OVERVIEW

3: Regression Analysis

3.1: Videos

Ideas of fitting a line - Also covers residuals and correlation

Fitting a least squares regression line - The notion of a "best fitting" line

Detailed Overview: Fitting a least squares regression line

Types of outliers in regression - Points of high leverage and influential points

Inference for linear regression - Using the t-distribution for inference in regression

3.2: Bivariate Data

In statistics, bivariate data means two variables or measurements per observation. For purposes of this section, we will assume both measurements are numeric data. These variables are usually represented by the letters X and Y.

Example: Sunglasses sales and rainfall

A company selling sunglasses determined the units per 1000 people and the annual rainfall in 5 cities.

X = rainfall in inches

Y = sales of sunglasses per 1000 people.

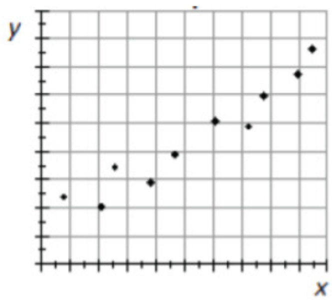
X	Y
10	40
15	35
20	25
30	25
40	15

In this example there are two numeric measurements for each of the five cities.

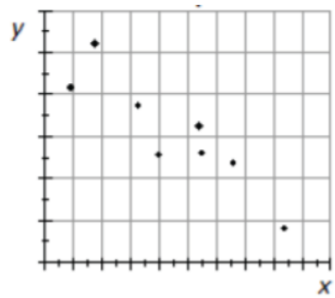
3.2.1: Graphing Bivariate Data with Scatterplots

A scatterplot is a useful graph for looking for relationships between two numeric variables. This relationship is called **correlation**. When performing correlation analysis, ask these questions:

1. What is the **direction** of the correlation?

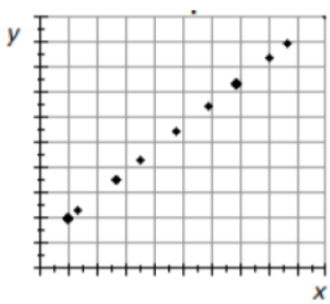


Positive

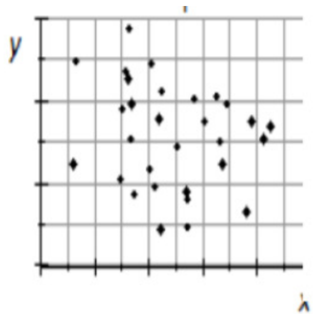


Negative

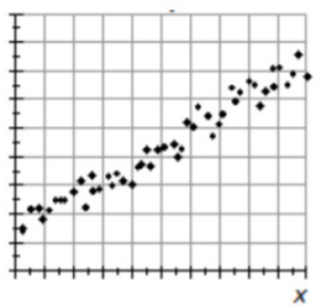
2. What is the **strength** of the correlation?



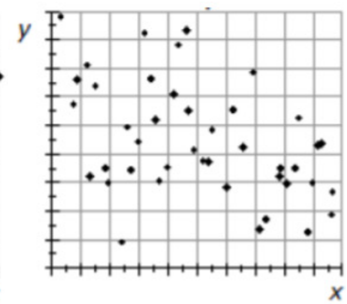
Perfect



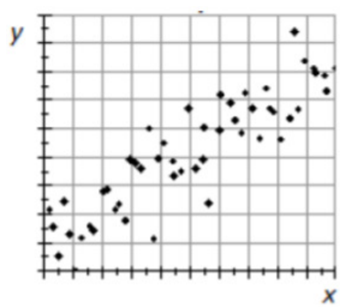
None



Strong

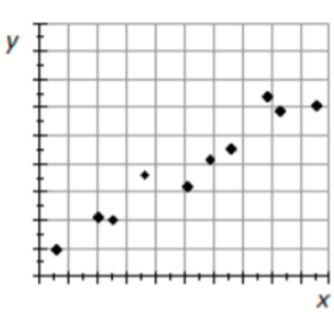


Weak

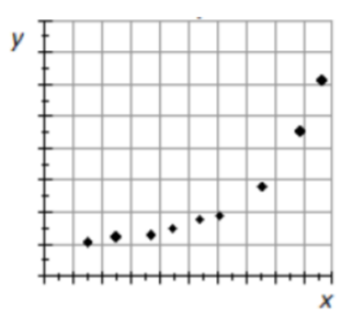


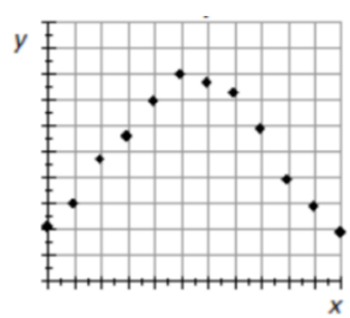
Moderate

3. What is the **shape** of the correlation?



Linear

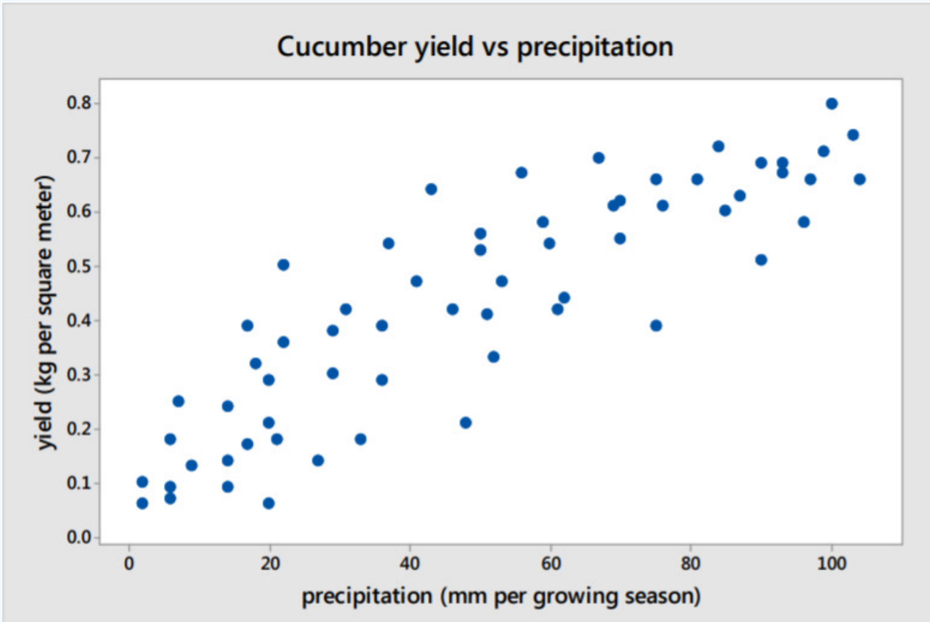




Non-linear

Example: Cucumber yield and rainfall

This scatterplot represents randomly collected data on growing season precipitation and cucumber yield. It is reasonable to suggest that the amount of water received on a field during the growing season will influence the yield of cucumbers growing on it.[32]



Solution

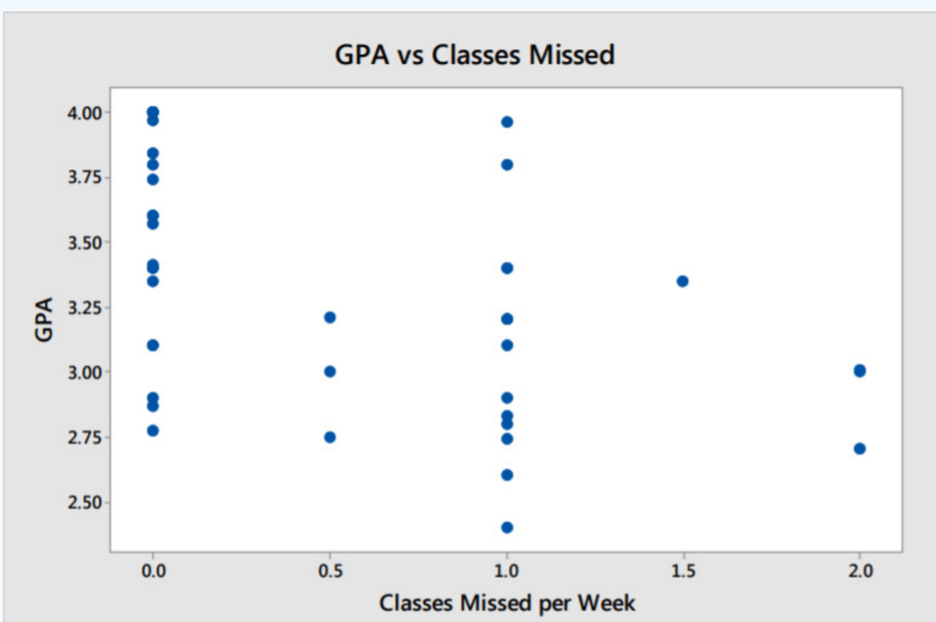
Direction: Correlation is positive, yield increases as precipitation increases.

Strength: There is a moderate to strong correlation.

Shape: Mostly linear, but there may be a slight downward curve in yield as precipitation increases.

Example: GPA and missing class

A group of students at Georgia College conducted a survey asking random students various questions about their academic profile. One part of their study was to see if there is any correlation between various students' GPA and classes missed.[33]



Solution

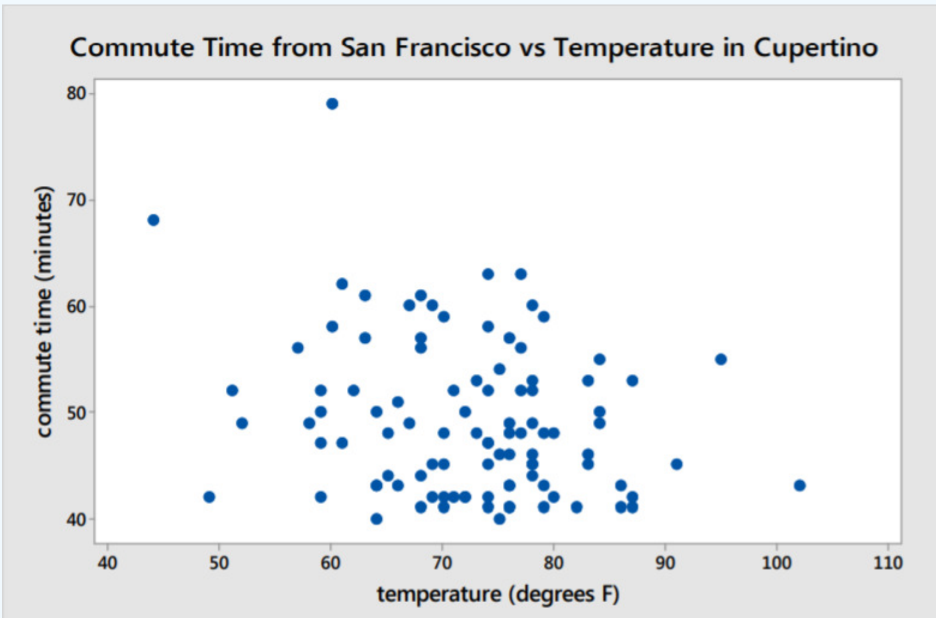
Direction: Correlation, if any, is negative. GPA trends lower for students who miss more classes.

Strength: There is a very weak correlation present.

Shape: Hard to tell, but a linear fit is not unreasonable.

Example: Commute times and temperature

A mathematics instructor commutes by car from his home in San Francisco to De Anza College in Cupertino, California. For 100 randomly selected days during the year, the instructor recorded the commute time and the temperature in Cupertino at time of arrival.



Solution

Direction: There is no obvious direction present.

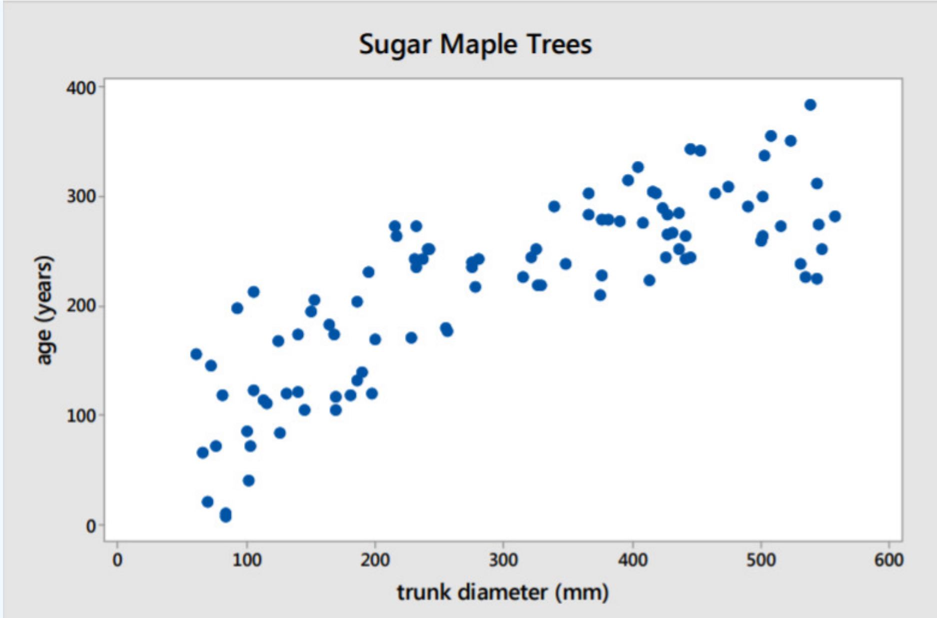
Strength: There is no apparent correlation between commute time and temperature.

Shape: Since there is no apparent correlation, looking for a shape is meaningless.

Other: There are two outliers representing very long commute times.

Example: Age of sugar maple trees

Is it possible to estimate the age of trees by measuring the diameters of the trunks? Data was reconstructed by a comprehensive study by the US Department of Agriculture. The researchers collected data for old growth sugar maple trees in northern US forests.[34]



Solution

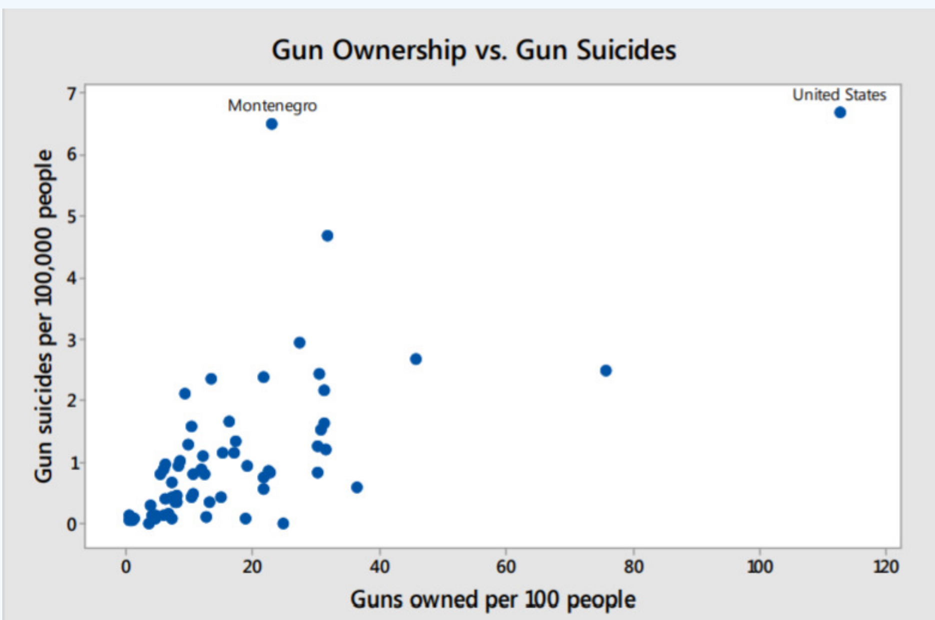
Direction: There is a positive correlation present. Age increases as trunk size increases.

Strength: The correlation is strong.

Shape: The shape of the graph is curved downward meaning the correlation is not linear.

Example: Gun ownership and gun suicides

This scatterplot represents gun ownership and gun suicides for 73 different countries. The data is adjusted to rates per population for comparison purposes.[35]



Solution

Direction: There is a positive correlation present. More gun ownership means more gun suicides.

Strength: The correlation is moderate for most data.

Shape: The shape of the graph is linear for most of the data.

Other: There are a few outliers in which gun ownership is much higher. There is also an outlier with an extremely high suicide rate.

This final example demonstrates that outliers can make it difficult to read graphs. For example, The United States has the highest gun ownership rates and the highest suicide by gun rates among these countries, making the United States stand far away from the bulk of the data in the scatterplot. Montenegro had the second highest suicide by gun rate, but with a much lower gun ownership rate.

3.2.2: Correlation Coefficient

The **correlation coefficient** (represented by the letter r) measures both the direction and strength of a linear relationship or association between two variables. The value r will always take on a value between -1 and 1. Values close to zero indicate a very weak correlation. Values close to 1 or -1 indicate a very strong correlation. The correlation coefficient should not be used for non-linear correlation.

It is important to ignore the sign when determining **strength** of correlation. For example, $r = -0.75$ would indicate a stronger correlation than $r = 0.62$, since -0.75 is farther from zero.

We will use technology to calculate the correlation coefficient, but formulas for manually calculating r are presented at the end of this section.

Interpreting the correlation coefficient (r)

$$-1 \leq r \leq 1$$

$r = 1$ means perfect positive correlation

$r = -1$ means perfect negative correlation

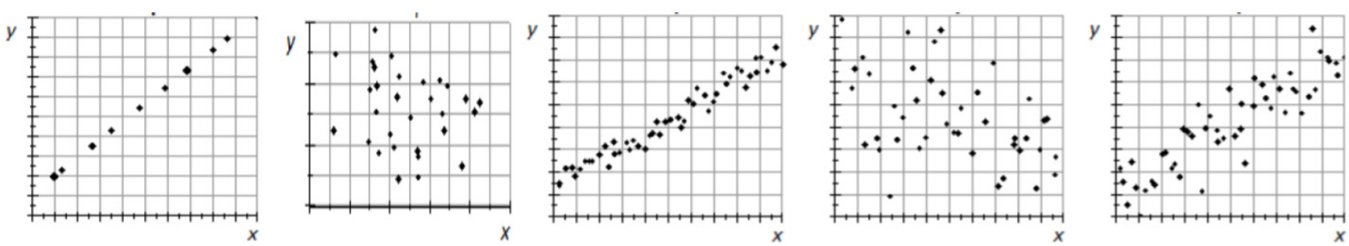
$r = 0$ mean no correlation

The farther r is from zero, the stronger the correlation

$r > 0$ means positive correlation

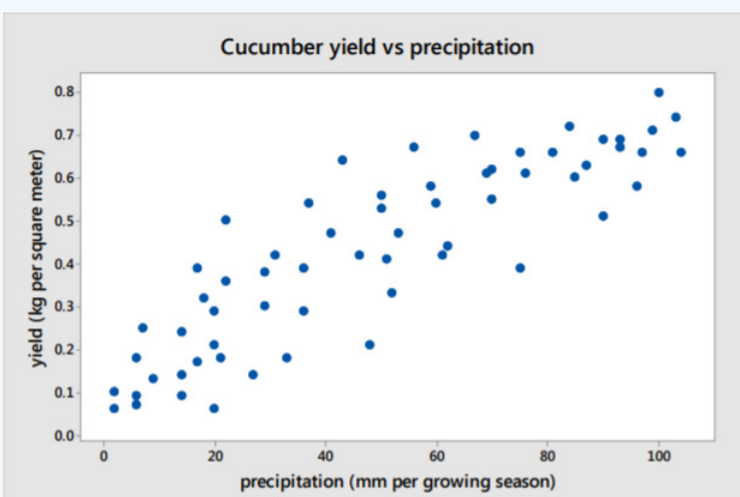
$r < 0$ means negative correlation

Some Examples



r = 1 | r = 0 | r = 0 .95 | r = -0.20 | r = 0.65

Example: Cucumber yield and rainfall

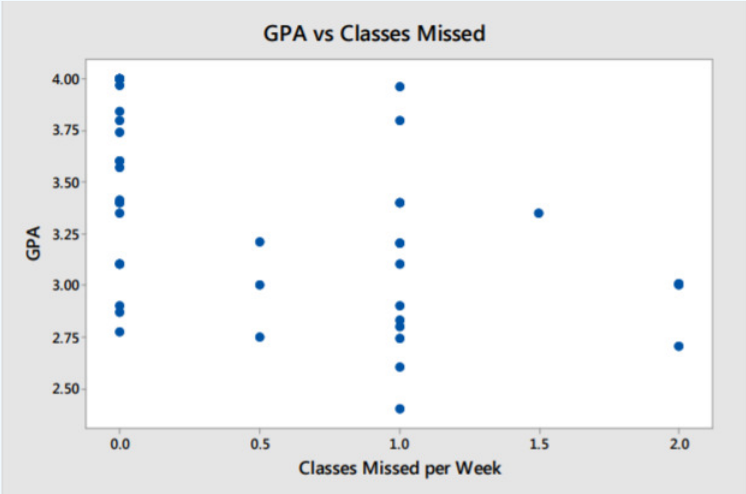


This scatterplot represents randomly collected data on growing season precipitation and cucumber yield.

$r = 0.871$ indicating strong positive correlation.

Example: GPA and missing class

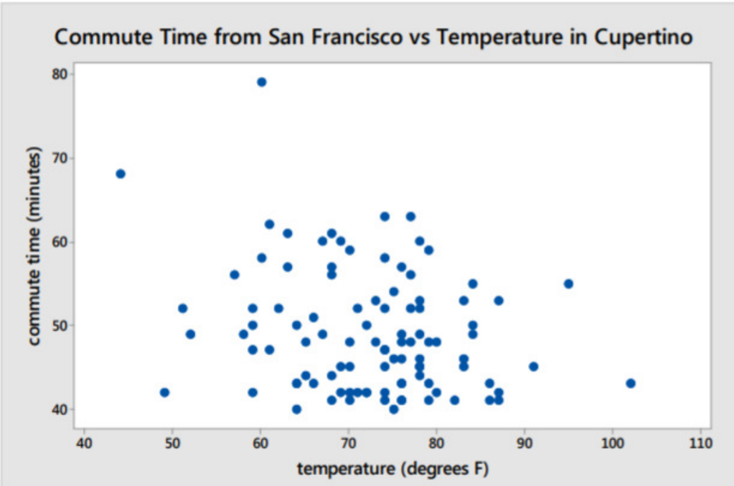
A group of students at Georgia College conducted a survey asking random students various questions about their academic profile. One part of their study was to see if there is any correlation between various students' GPA and classes missed.



$r =$- 0.236 indicating weak negative correlation.

Example: Commute times and temperature

A mathematics instructor commutes by car from his home in San Francisco to De Anza College in Cupertino, California. For 100 randomly selected days during the year, the instructor recorded the commuting time and the temperature in Cupertino at time of arrival.



$r =$- 0.02 indicating no correlation.

Calculating the correlation coefficient

Manually calculating the correlation coefficient is a tedious process, but the needed formulas and one simple example are presented here:

Formulas for calculating the correlation coefficient (r)

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

$$SSX = \Sigma X^2 - \frac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \frac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \frac{1}{n}(\Sigma X \cdot \Sigma Y)$$

Example: Sunglasses sales and rainfall

A company selling sunglasses determined the units sold per 1000 people and the annual rainfall in 5 cities.

X = rainfall in inches

Y = sales of sunglasses per 1000 people.

X	Y
10	40
15	35
20	25
30	25
40	15

Solution

First, find the following sums:

$$\sum X, \sum Y, \sum X^2, \sum Y^2, \sum XY$$

	$X)$	Y	X^2	Y^2	XY
	10	40	100	1600	400
	15	35	225	1225	525
	20	25	400	625	500
	30	25	900	625	750
	40	15	1600	225	600
Σ	**115**	**140**	**3225**	**4300**	**2775**

Then, find SSX, SSY, $SSXY$

$$\begin{aligned} SSX &= 3225 - 115^2/5 &&= 580 \\ SSY &= 4300 - 140^2/5 &&= 380 \\ SSXY &= 2775 - (115)(140)/5 &&= -445 \end{aligned}$$

Finally, calculate r

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} = \frac{-445}{\sqrt{580 \cdot 330}} = -0.9479$$

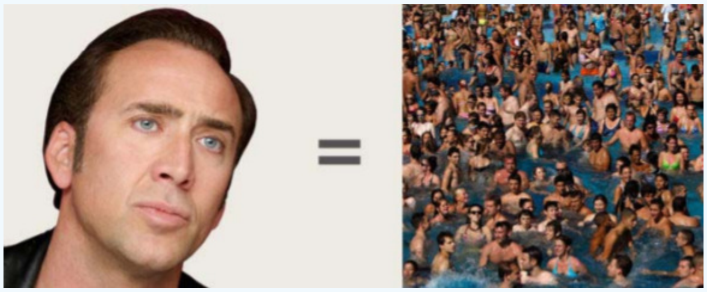
The correlation coefficient is -0.95, indicating a strong, negative correlation between rainfall and sales of sunglasses.

3.2.3: Correlation vs. Causation

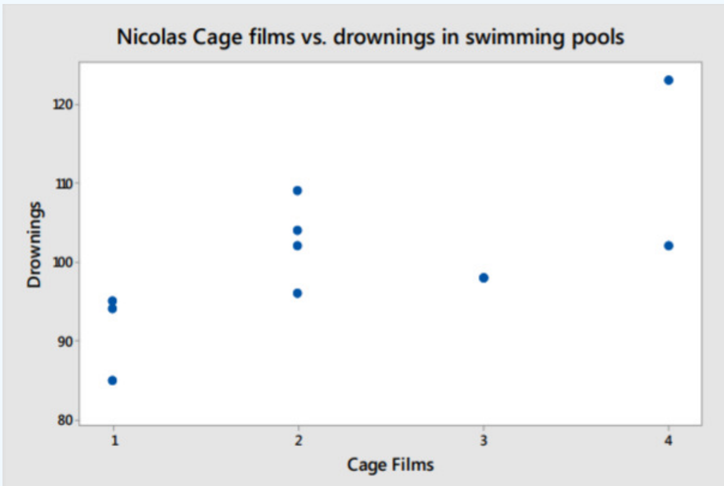
One of the greatest mistakes people make in Statistics is in confusing correlation with causation.

Example: Nicolas Cage movies and drownings

A study done by law student Tyler Vigan showed a moderate to strong correlation between the number of movies Nicolas Cage releases in a year and the number of drownings in swimming pools in the same year.[36]



The scatterplot shows moderate positive correlation, supported by a correlation coefficient of 0.66.

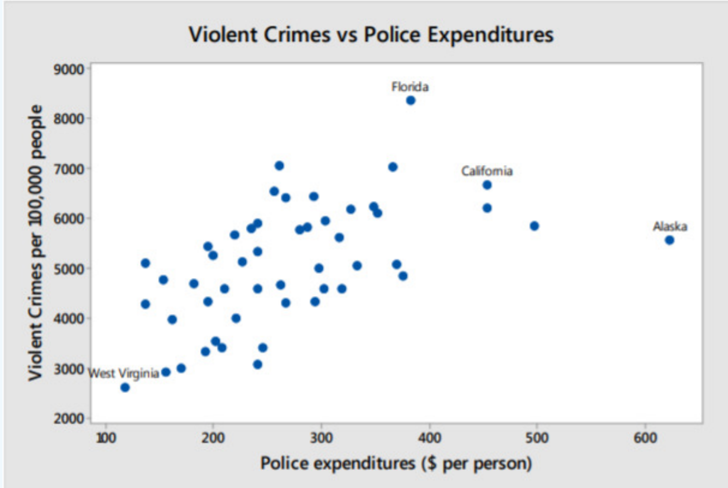


What does this mean? When Nicolas Cage releases a movie, people get excited and go jump in the pool? Or maybe in a year when there are many drownings, Nicolas Cage gets inspired to release a new movie?

This is an example of a **spurious** correlation, a correlation that just happens by chance.

Example: Crime and police expenditures

The scatterplot shows data from all 50 states adjusted for population differences. The horizontal axis is annual police expenditures per person. The vertical axis represents reported violent crimes per 100,000 people per year.



There is a moderate positive correlation present, with a correlation coefficient of 0.547.

What does this mean? Here are possible explanations.

1. **Police cost causes crime**. The more money spent on police, the more crime there is. Eliminate the police to reduce crime.
2. **Crime causes police cost**. The more crime there is, more police get hired. High crime states need to spend more money on the police.
3. **More police means more reported crimes**. The data shows reported crimes, but many crimes go unreported. Having more police means more reported crimes.
4. **Crime and police costs are higher in cities**. States like California, Texas and Florida have major cities where all expenses are higher and there is more crime. So in this example, urbanization is the cause of both variables increasing. (This is an example of a **confounding** variable).

The truth is we can't say why there is a correlation between police expenditures and violent crime. As statisticians, we can only say the variables are correlated, and we cannot support a cause and effect relationship.

In observational studies such as this, **correlation does not equal causation**.

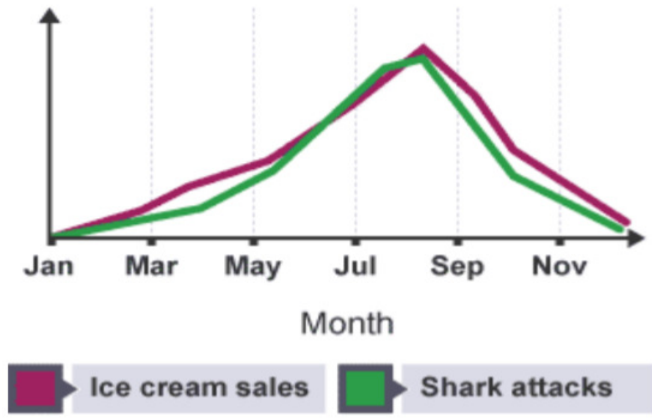
Confounding (lurking) variables

A confounding or lurking variable is a variable that is not known to the researcher, but affects the results of the study.

Research has shown there is a strong, positive correlation between shark attacks and ice cream sales. There is actually a store in New York called Shark's Ice Cream, possibly inspired by this correlation.[37]

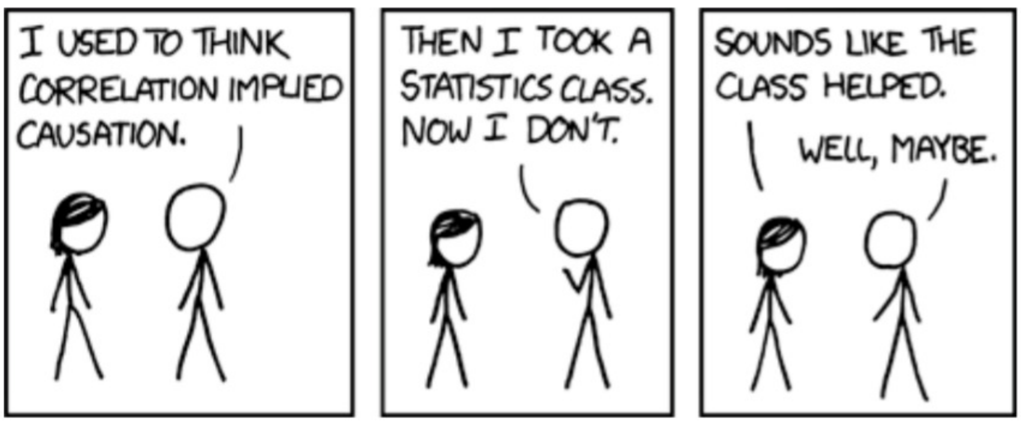
A possible confounding variable might be temperature. On hot days people are more likely to swim in the ocean and are also more likely to buy ice cream.

This graph from the BBC seems to support this claim. [38]Both shark attacks and ice cream sales are highest in the summer months.



In the next section, we will discuss how to design experiments that control for confounding variables.

Hopefully taking this Statistics class will help you avoid making the mistake of confusing correlation and causation. Or, maybe you already knew that, as inspired by this XKCD comic “Correlation.”[39]



SECTION OVERVIEW

3.3: Correlation and Linear Regression

Often in statistical research, we want to discover if there is a relationship between two variables. The **explanatory variable** is the "cause" and the **response variable** is the "effect", although a true cause and effect relationship can only be established in a scientific study that controls for all confounding (lurking) variables.

In Chapter 12, we were interested in determining if a person's gender was a valid explanatory variable of the person's opinion about legalization of marijuana for recreational use. In this case, both the explanatory and response variables are categorical and the appropriate model was the Chi-square Test of Independence.

In Chapter 13, we explored if tofu pizza sales (the response variable) were affected by location of the restaurant (the explanatory variable). In this case, the explanatory variable was categorical but the response was numeric. The appropriate model for this example is One Factor Analysis of Variance (ANOVA).

What if we want to determine if a relationship exists when both the explanatory and response variables are both numeric? For example, does annual rainfall in a city help explain sales of sunglasses? This chapter explores and defines the appropriate model for this type of problem.

3.3.1: Bivariate Data and Scatterplots Review

In Chapter 3, we defined bivariate data as data that have two different numeric variables. In an algebra class, these are also known as ordered pairs. We will let X represent the **independent** (or explanatory) variable and Y represent the **dependent** (or response) variable in this definition. Here is an example of five total pairs in which X represents the annual rainfall in inches in a city and Y represents annual sales of sunglasses per 1000 population.

The best way to graph bivariate data is by using a **Scatterplot** in which X, the independent variable is the vertical axis and Y, the dependent variable is the horizontal axis.

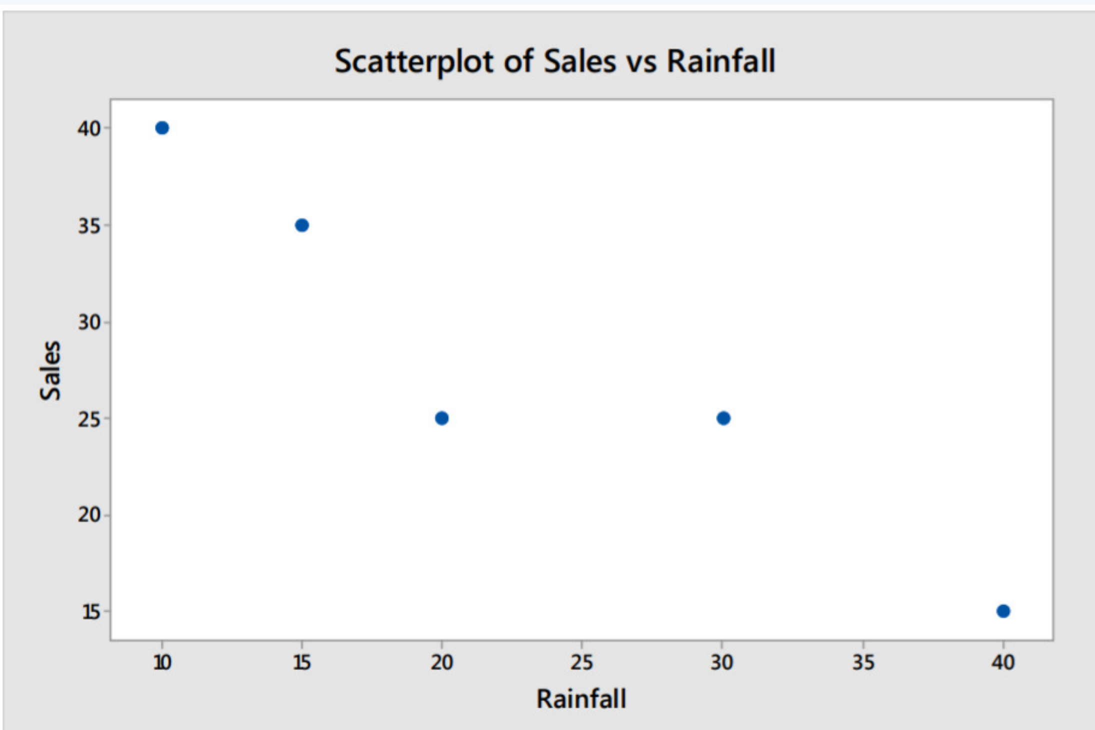
Example: Rainfall and sunglasses sales

Here is an example and scatterplot of five total pairs where X represents the annual rainfall in inches in a city and Y represents annual sales of sunglasses per 1000 population.

X=rainfall	Y=sales
10	40
15	35
20	25
30	25
40	15

Solution

In the scatterplot for this data, it appears that cities with more rainfall have lower sales. It also appears that this relationship is linear, a pattern which can then be exemplified in a statistical model.



3.3.2: The Simple Linear Regression Model

In the scatterplot example shown above, we saw linear correlation between the two dependent variables. We are now going to create a statistical model relating these two variables, but let's start by reviewing a **mathematical linear model** from algebra:

$Y = \beta_0 + \beta_1 X$

Y: Dependent Variable

X: Independent Variable

β_0: Y - intercept

β_1:Slope

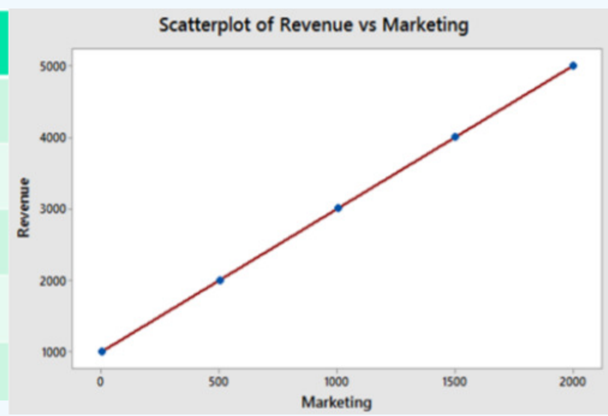
Example

You have a small business producing custom t-shirts. Without marketing, your business has revenue (sales) of $1000 per week. Every dollar you spend marketing will increase revenue by 2 dollars. Let variable X represent the amount spent on marketing and let variable Y represent revenue per week. Write a **mathematical model** that relates X to Y.

Solution

In this example, we are saying that weekly revenue (Y) depends on marketing expense (X). $1000 of weekly revenue represents the vertical intercept, and $2 of weekly revenue per $1 marketing represents the slope, or rate of change of the model. We can choose some value of X and determine Y and then plot the points on a scatterplot to see this linear relationship.

X=marketing	Y=revenue
$0	$1000
$500	$2000
$1000	$3000
$1500	$4000
$2000	$5000

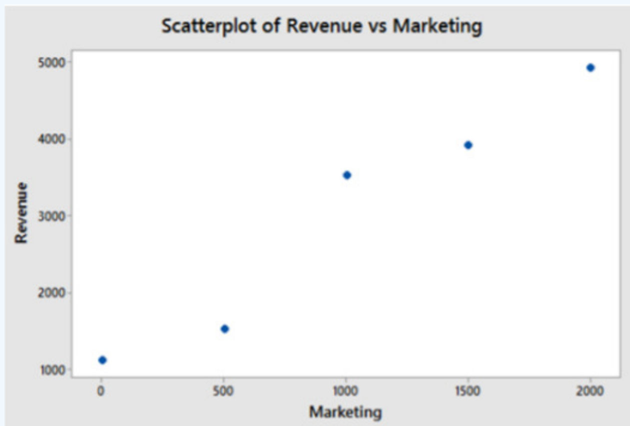


We can then write out the mathematical linear model as an equation:

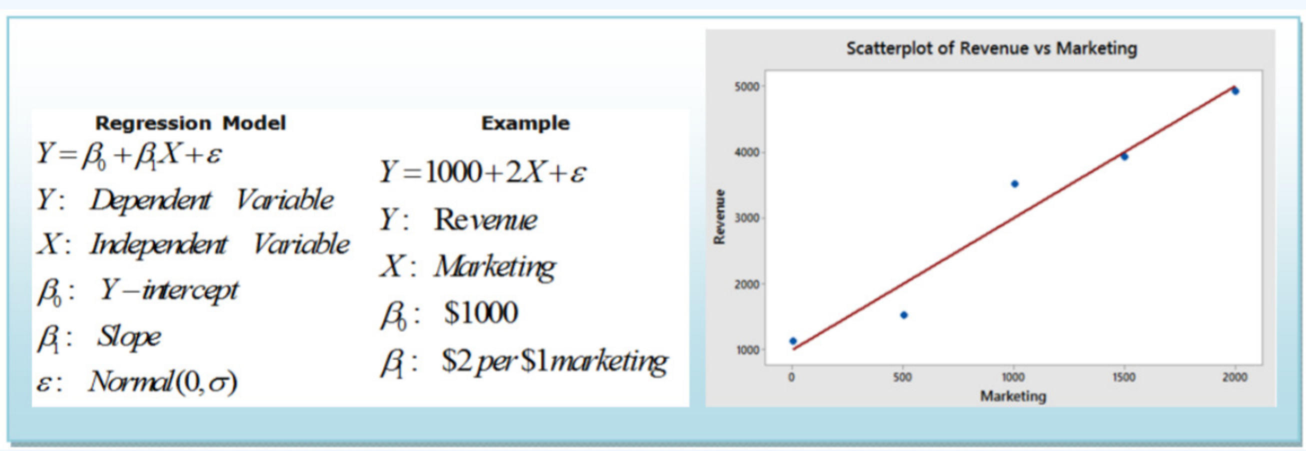
Linear Model	Example
$Y = \beta_0 + \beta_1 X$	$Y = 1000 + 2X$
Y: Dependent Variable	Y: Revenue
X: Independent Variable	X: Marketing
β_0: Y-intercept	β_0: $1000
β_1: Slope	β_1: $2 per $1 marketing

We all learned about these linear models in Algebra classes, but the real world doesn't generally give such perfect results. In particular, we can choose what to spend on marketing, but the actual revenue will have more uncertainty. For example, the true revenue may look more like this:

X=Marketing	Expected Revenue	Y=Actual Revenue	ε=Residual Error
\$0	\$1000	\$1100	+\$100
\$500	\$2000	\$1500	-\$500
\$1000	\$3000	\$3500	+\$500
\$1500	\$4000	\$3900	-\$100
\$2000	\$5000	\$4900	-\$100



The difference between the actual revenue and the expected revenue is called the **residual error**, ε If we assume that the residual error (represented by ε) is a random variable that follows a normal distribution with $\mu = 0$ and σ a constant for all values of X, we have now created a **statistical model** called a **simple linear regression model**.



3.3.3: Estimating the Regression Model with the Least-Square Line

We now return to the case where we know the data and can see the linear correlation in a scatterplot, but we do not know the values of the parameters of the underlying model. The three parameters that are unknown to us are the y-intercept β_0, the slope (β_1) and the standard deviation of the residual error (σ):

Slope parameter: b_1 will be an estimator for β_1

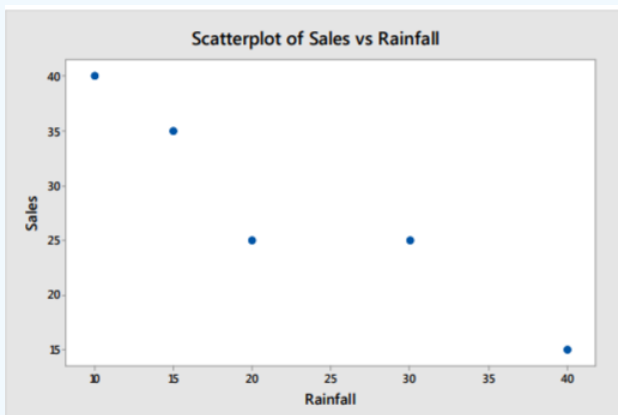
Y-intercept parameter: b_0 will be an estimator for β_0

Standard deviation: s_e will be an estimator for σ

Regression line: $\hat{Y} = b_0 + b_1 X$

Example

Take the example comparing rainfall to sales of sunglasses in which the scatterplot shows a negative correlation. However, there are many lines we could draw. How do we find the line of best fit?



Solution

Minimizing Sum of Squared Residual Errors (SSE)

We are going to define the "best line" as the line that minimizes the Sum of Squared Residual Errors (SSE).

Suppose we try to fit this data with a line that goes through the first and last point. We can then calculate the equation of this line using algebra:

$$\hat{Y} = \frac{145}{3} - \frac{5}{6}X \approx 48.3 - 0.833X$$

The SSE for this line is 47.917:

Rainfall	Sales	Predicted Sales	Residual	Squared Residuals
10	40	40	0	0
15	35	35.833	-0.833	0.694
20	25	31.667	-6.667	44.444
30	25	23.333	1.667	2.778
40	15	15	0	0
		Sum of Squared Residuals =		47.917

Although this line is a good fit, it is not the best line. The slope(b_1) and intercept(b_o) for the line that minimizes SSE is be calculated using the least squares principle formulas:

> **Least squares principle formulas**
>
> $$SSX = \Sigma X^2 - \frac{1}{n}(\Sigma X)^2$$
>
> $$SSY = \Sigma Y^2 - \frac{1}{n}(\Sigma Y)^2$$
>
> $$SSXY = \Sigma XY - \frac{1}{n}(\Sigma X \cdot \Sigma y)$$
>
> $$b_1 = \frac{SSXY}{SSX}$$
>
> $$b_0 = \bar{Y} - b_1 \bar{X}$$

In the Rainfall example where X=Rainfall and Y=Sales of Sunglasses:

	X	Y	X^2	Y^2	XY
	10	40	100	1600	400
	15	35	225	1225	525
	20	25	400	625	500
	30	25	900	625	750
	40	15	1600	225	600
Σ	115	140	3225	4300	2775

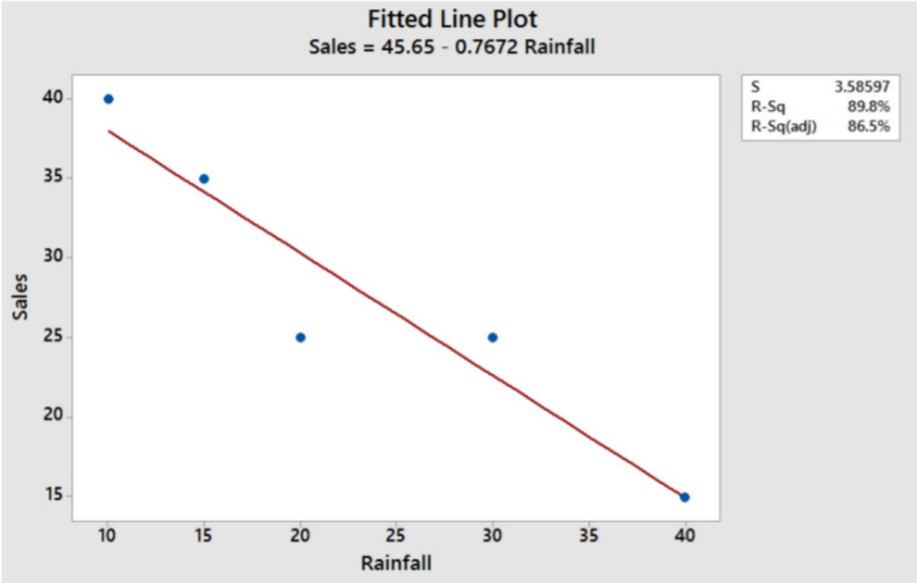
- SSX = 580
- SSY = 380
- SSXY = -445

- b_1 = -.767
- b_0 = 45.647
- $\hat{Y}$ = 45.647 - .767X

The Sum of Squared Residual Errors (SSE) for this line is 38.578, making it the “best line”. (Compare to the value above, in which we picked the line that perfectly fit the two most extreme points).

Rainfall	Sales	Predicted Sales	Residual	Squared Residuals
10	40	37.977	2.023	4.092529
15	35	34.142	0.858	0.736
20	25	30.307	-5.307	28.164
30	25	22.637	2.363	5.584
40	15	14.967	0.033	0.001089
		Sum of Squared Residuals =		38.578

In practice, we will use technology such as Minitab to calculate this line. Here is the example using the Regression Fitted Line Plot option in Minitab, which determines and graphs the regression equation. The point (20,25) has the highest residual error, but the overall Sum of Squared Residual Errors (SSE) is minimized.


Fitted Line Plot
Sales = 45.65 - 0.7672 Rainfall
S 3.58597
R-Sq 89.8%
R-Sq(adj) 86.5%
Sales
Rainfall
40
35
30
25
20
15
10
15
20
25
30
35
40

3.3.4: Hypothesis Test for Simple Linear Regression

We will now describe a hypothesis test to determine if the regression model is meaningful; in other words, does the value of X in any way help predict the expected value of Y?

Simple Linear Regression ANOVA Hypothesis Test

Model Assumptions

- The residual errors are random and are normally distributed.
- The standard deviation of the residual error does not depend on X
- A linear relationship exists between X and Y
- The samples are randomly selected

Test Hypotheses

H_o: X and Y are not correlated

H_a: X and Y are correlated

H_o: β_1 (slope) = 0

H_a: β_1 (slope) ≠ 0

Test Statistic

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$$

$$df_{\text{num}} = 1$$

$$df_{\text{den}} = n - 2$$

Sum of Squares

$$SS_{\text{Total}} = \sum (Y - \bar{Y})^2$$

$$SS_{\text{Error}} = \sum (Y - \hat{Y})^2$$

$$SS_{\text{Regression}} = SS_{\text{Total}} - SS_{\text{Error}}$$

In simple linear regression, this is equivalent to saying "Are X an Y correlated?"

In reviewing the model, $Y = \beta_0 + \beta_1 X + \varepsilon$, as long as the slope (β_1) has any non-zero value, X will add value in helping predict the expected value of Y. However, if there is no correlation between X and Y, the value of the slope (β_1) will be zero. The model we can use is very similar to One Factor ANOVA.

The Results of the test can be summarized in a special ANOVA table:

Source of Variation	Sum of Squares (SS)	[illegible]
Factor (due to X)	$SS_{\text{Regression}}$	[illegible]
Error (Residual)	SS_{Error}	[illegible]
Total	SS_{Total}	n [illegible]

Example: Rainfall and sales of sunglasses

Design: Is there a significant correlation between rainfall and sales of sunglasses?

Research Hypotheses:

H_o: Sales and Rainfall are not correlated H_o: 1 (slope) = 0

H_a: Sales and Rainfall are correlated H_a: 1 (slope) ≠ 0

Type I error would be to reject the Null Hypothesis and t claim that rainfall is correlated with sales of sunglasses, when they are not correlated. The test will be run at a level of significance (α) of 5%.

The test statistic from the table will be $\mathrm{F}=\frac{\mathrm{MSRegression}}{\mathrm{MSError}}$. The degrees of freedom for the numerator will be 1, and the degrees of freedom for denominator will be 5-2=3.

Critical Value for F at $\alpha$of 5% with $df_{num}=1$ and df_{den}=3} is 10.13. Reject \(H_o if $F>10.13$. We will also run this test using the p-value method with statistical software, such as Minitab.

Data/Results

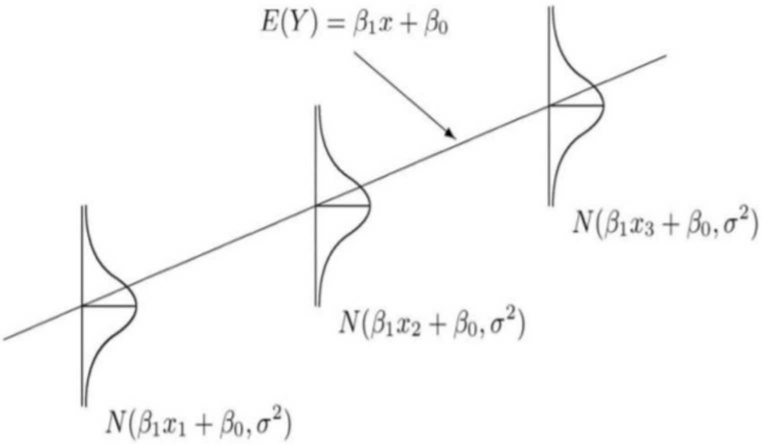
Source	SS	df	MS	F	p-value
Regression	341.422	1	341.422	26.551	0.0142
Error	38.578	3	12.859		
TOTAL	380.000	4			

$F=341.422/12.859=26.551$, which is more than the critical value of 10.13, so Reject H_o. Also, the p-value = 0.0142 < 0.05 which also supports rejecting H_o.

Conclusion

Sales of Sunglasses and Rainfall are negatively correlated.

3.3.5: Estimating σ σ , the standard error of the residuals



The simple linear regression model ($Y = \beta_0 + \beta_1 X + \varepsilon$) includes a random variable ε representing the residual which follows a Normal Distribution with an expected value of 0 and a standard deviation σ which is independent of the value of X. The estimate of σ is called the sample standard error of the residuals and is represented by the symbol s_e. We can use the fact that the Mean Square Error (MSE) from the ANOVA table represents the estimated variance of the residuals errors:

$$S_e = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Example: Rainfall and sales of sunglasses

For the rainfall data, the standard error of the residuals is determined as:

$$s_e = \sqrt{12.859} = 3.586$$

Keep in mind that this is the standard deviation of the residual errors and should not be confused with the standard deviation of Y.

3.3.6: r 2 r2 , The Correlation of Determination

The Regression ANOVA hypothesis test can be used to determine if there is a **significant** correlation between the independent variable (X) and the dependent variable (Y). We now want to investigate the **strength** of correlation.

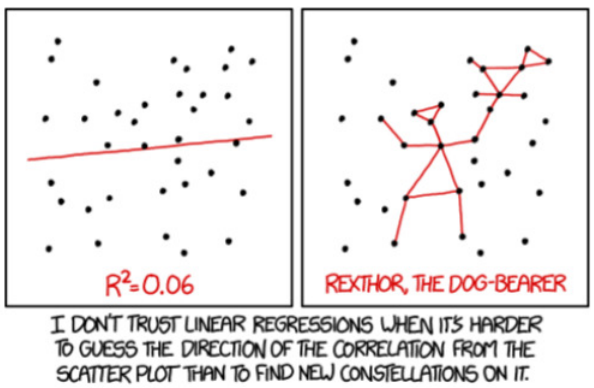
In the earlier chapter on descriptive statistics, we introduced the correlation coefficient (r), a value between -1 and 1. Values of r close to 0 meant there was little correlation between the variables, while values closer to 1 or -1 represented stronger correlations.

In practice, most statisticians and researchers prefer to use r^2, the coefficient of determination as a measure of strength as it represents the proportion or percentage of the variability of Y that is explained by the variability of X. [87]

r^2

\(r^{2}=\dfrac{S S_{\text{regression}}{S S_{\text {Total }}} \dquad 0 \% \leq r^{2} \leq 100 \%\)

r^2 represents the percentage of the variability of Y that is explained by the variability of X.



We can also calculate the correlation coefficient (r) by taking the appropriate square root of r^2, depending on whether the estimate of the slope (b_1) is positive or negative:

$$\text{If } b_1 > 0, r = \sqrt{r^2}$$

$$\text{If } b_1 < 0, r = -\sqrt{r^2}$$

Example: Rainfall and sales of sunglasses

For the rainfall data, the coefficient of determination is:

$$r^2 = \frac{341.422}{380} = 89.85\%$$

89.85% of the variability of sales of sunglasses is explained by rainfall.

We can calculate the correlation coefficient (r) by taking the appropriate square root of r^2:

$$r = -\sqrt{.8985} = -0.9479$$

Here we take the negative square root since the slope of the regression line is negative. This shows that there is a strong, negative correlation between sales of sunglasses and rainfall.

3.3.7: Prediction

One valuable application of the regression model is to make predictions about the value of the dependent variable if the independent variable is known.

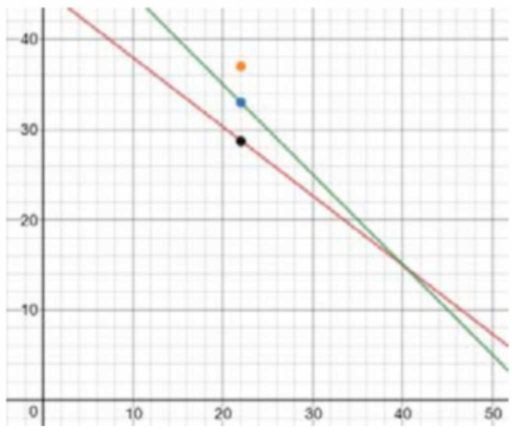
Consider the example about rainfall and sunglasses sales. Suppose we know that a city has 22 inches of rainfall. We can use the regression equation to predict the sales of sunglasses:

$\hat{Y} = 45.647 - .767X$

$\hat{Y}_{22} = 45.647 - .767(22) = 28.7$

For a city with 22 inches of annual rainfall, the model predicts sales of 28.7 per 1000 population.

To measure the **reliability** of this prediction, we can construct confidence intervals. However, we first have to decide what we are estimating. We could (1) be estimating the **expected** sales for a city with 22 inches of rainfall, or we could (2) be predicting the **actual** sales for a city with 22 inches of rainfall.



In the graph shown, the green line represents $Y = \beta_0 + \beta_1 X + \varepsilon$ the actual regression line which is unknown. The red line represents the least square equation, $\hat{Y} = 45.647 - .767X$, which is derived from the data. The black dot represents our prediction $Y_{22} = 28.7$. The green dot represents the correct population **expected** value of Y_{22}, while the yellow dot represents a possible value for the **actual** predicted value of Y_{22}. There is more uncertainty in predicting an actual value of Y_x than the expected value.

Confidence interval and Prediction interval

The **confidence interval** for the **expected** value of Y for a given value of X is given by:

$$\hat{Y}_X \pm t \cdot s_e \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{SSX}}$$

Degrees of freedom for $t = n$- 2

The **prediction interval** for the **actual** value of Y for a given value of X is given by:

$$\hat{Y}_X \pm t \cdot S_e \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SSX}}$$

Degrees of freedom for $t = n$- 2

Example: Rainfall sunglasses sales

1. Find a 95% confidence interval for the expected value of sales for a city with 22 inches of rainfall.
2. Find a 95% prediction interval for the value of sales for a city with 22 inches of rainfall.

Solution

1. **Confidence interval**

$$28.7 \pm 3.182 \cdot 3.586 \sqrt{\frac{1}{5} + \frac{(22-23)^2}{580}} = 28.7 \pm 5.1 \rightarrow (23.6, 33.8)$$

We are 95% confident that the expected annual sales of sunglasses for a city with 22 inches of annual rainfall is between 23.6 and 33.8 sales per 1000 population.

2. **Prediction interval**

$$28.7 \pm 3.182 \cdot 3.586 \sqrt{1 + \frac{1}{5} + \frac{(22-23)^2}{580}} = 28.7 \pm 12.5 \rightarrow (16.2, 41.2)$$

We are 95% confident that the actual annual sales of sunglasses for a city with 22 inches of annual rainfall is between 16.2 and 41.2 sales per 1000 population.

3.3.8: Extrapolation

When using the model to make predictions, care must be taken to only choose values of X that are in the range of X values of the data. In the rainfall/sales example, the values of X range from 10 to 40 inches of rainfall. Choosing a value of X outside this range is called extrapolation and could lead to invalid results. For example, if we use the model to predict sales for a city with 80 inches of rainfall, we get an impossible negative result for sales:

$$\hat{Y} = 45.647 - .767X$$

$$\hat{Y}_{80} = 45.647 - .767(80) = -15.7$$

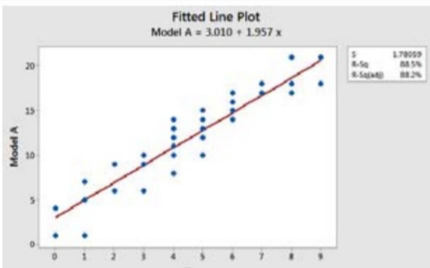
3.3.9: Residual Analysis

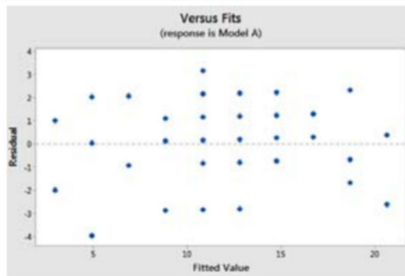
In regression, we assume that the model is linear and that the residual errors ($Y - \hat{Y}$ for each pair) are random and normally distributed. We can analyze the residuals to see if these assumptions are valid and if there are any potential outliers. In particular:

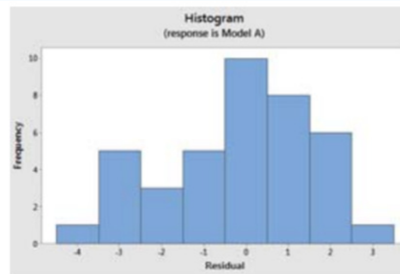
- The residuals should represent a linear model.
- The standard error (standard deviation of the residuals) should not change when the value of X changes.
- The residuals should follow a normal distribution.
- Look for any potential extreme values of X.
- Look for any extreme residual errors.

Example: Model A

Model A is an example of an appropriate linear regression model. We will make three graphs to test the residual; a scatterplot with the regression line, a plot of the residuals, and a histogram of the residuals

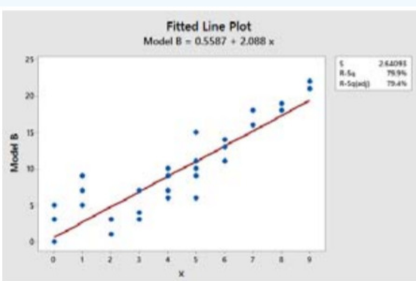


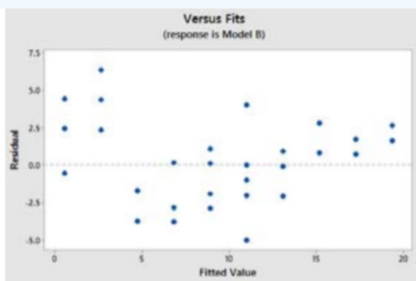


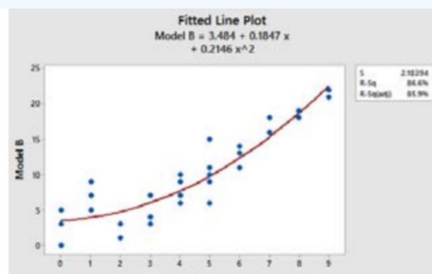


Here we can see the that residuals appear to be random, the fit is linear, and the histogram is approximately bell shaped. In addition, there are no extreme outlier values of X or outlier residuals.

Example: Model B

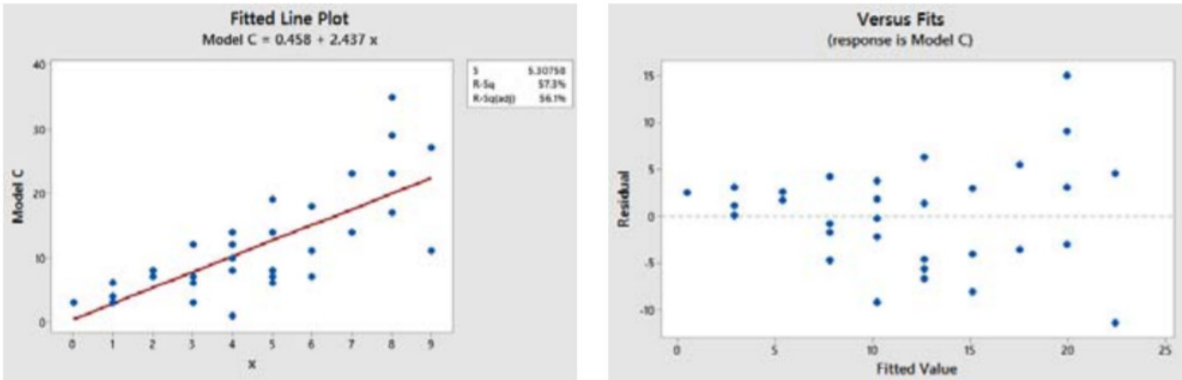






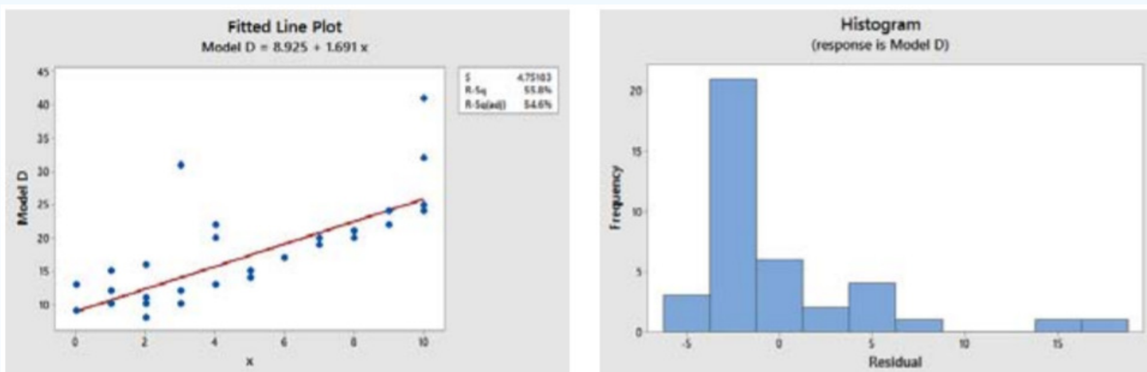
Model B looks like a strong fit, but the residuals are showing a pattern of being positive for low and high values of X and negative for middle values of X. This indicates that the model is not linear and should be fit with a non-linear regression model (for example, the third graph shows a quadratic model).

Example: Model C



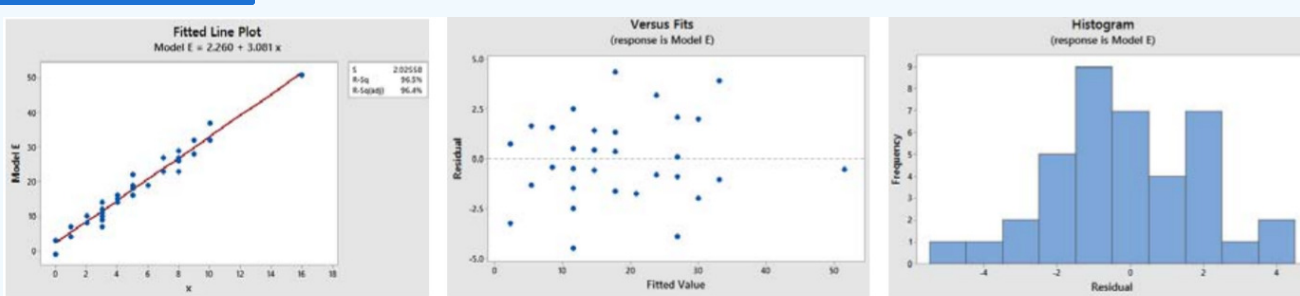
Model C has a linear fit, but the residuals are showing a pattern of being smaller for low values of X and higher for large values of X. This violates the assumption that the standard error should not change when the value of X changes. This phenomena is called **heteroscedasticity** and requires a data transformation to find a more appropriate model.

Example: Model D



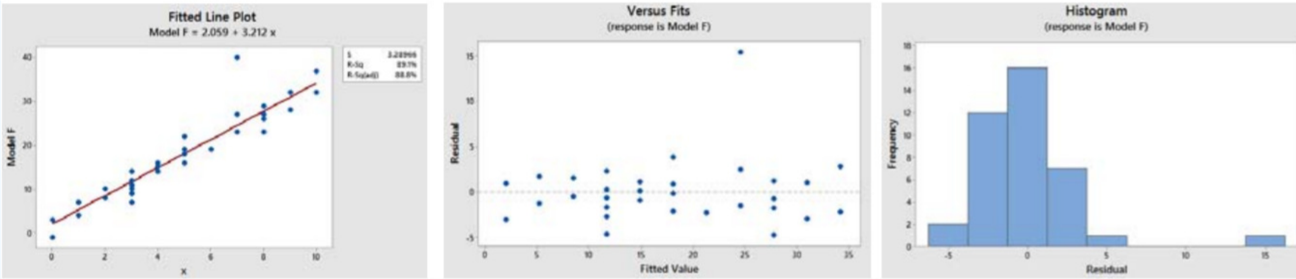
Model D seems to have a linear fit, but the residuals are showing a pattern of being larger when they are positive and smaller when they are negative. This violates the assumption that residuals should follow a normal distribution, as can be seen in the histogram.

Example: Model E



Model E seems to have a linear fit, and the residuals look random and normal. However, the value (16,51) is an extreme outlier value of X and may have an undue influence on the choosing of the regression line.

Example: Model F



Model F seems to have a linear fit, and the residuals look random and normal, except for one outlier at the value (7,40). This outlier is different than the extreme outlier in Model E, but will still have an undue influence on the choosing of the regression line.

SECTION OVERVIEW

3.4: Linear Regression and Correlation

Regression analysis is a statistical process for estimating the relationships among variables and includes many techniques for modeling and analyzing several variables. When the focus is on the relationship between a dependent variable and one or more independent variables.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

3.4.1: Prelude to Linear Regression and Correlation

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it? In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

Figure 3.4.1.1: Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables. In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

3.4.2: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + b\text{x}$$

where a and b are constant numbers. The variable x is the *independent variable*, and y is the *dependent variable*. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example $3.4.2.1$

The following examples are linear equations.

$$y = 3 + 2\text{x}$$

$$y = -0.01 + 1.2\text{x}$$

Exercise $3.4.2.1$

Is the following an example of a linear equation?

$$y = -0.125 - 3.5\text{x}$$

Answer

yes

The graph of a linear equation of the form $y = a + b\text{x}$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example $3.4.2.2$

Graph the equation $y = -1 + 2\text{x}$.

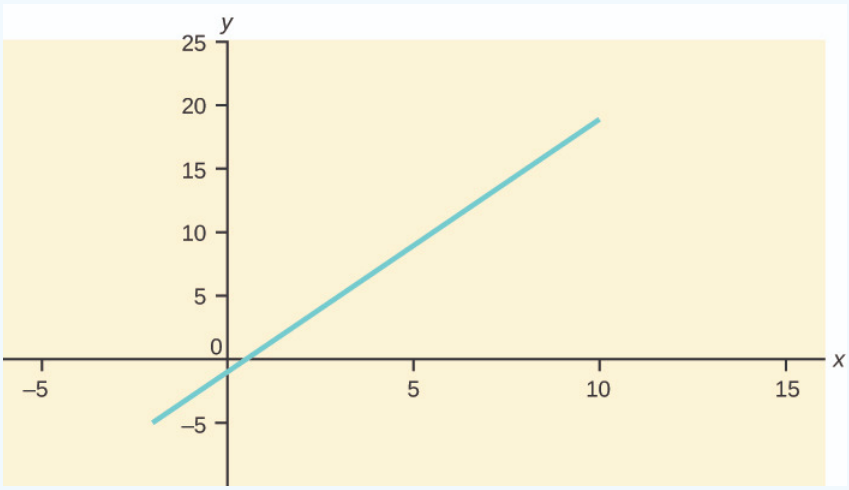


Figure $3.4.2.1$.

Exercise $3.4.2.2$

Is the following an example of a linear equation? Why or why not?

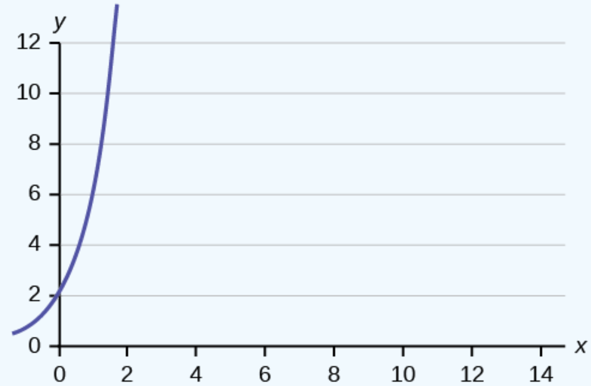


Figure $3.4.2.2$.

Answer

No, the graph is not a straight line; therefore, it is not a linear equation.

Example $3.4.2.3$

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Answer

Let $x =$ the number of hours it takes to get the job done.

Let $y =$ the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32\text{x}$

Exercise $3.4.2.3$

Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Answer

$y = 50 + 20\text{x}$

Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + b\text{x}$, $b =$ slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

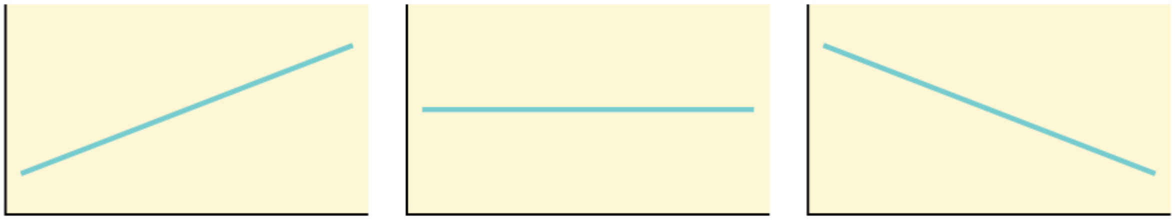


Figure 3.4.2.3:. Three possible graphs of $y = a + b\text{x}$ (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 3.4.2.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of \$25 plus \$15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15\text{x}$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of \$25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns \$15 for each hour she tutors.

Exercise 3.4.2.4

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges \$25 plus \$20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20\text{x}$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of \$25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns \$20 for each hour he works.

Summary

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + b\text{x}$, the constant b that multiplies the x variable (b is called a coefficient) is called the **slope**. The constant a is called the y-intercept.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The **y-intercept** is used to describe the dependent variable when the independent variable equals zero.

Formula Review

$y = a + b\text{x}$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

3.4.2E: Linear Equations (Exercises)

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

Exercise 12.2.5

What are the dependent and independent variables?

Answer

dependent variable: fee amount; independent variable: time

Exercise 12.2.6

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise 12.2.7

Graph the equation from Exercise.

Answer

This is a graph of the equation y = 25 + 12.50x. The x-axis is labeled in intervals of 1 from 0 - 7; the y-axis is labeled in intervals of 25 from 0 - 100. The equation's graph is a line that crosses the y-axis at 25 and is sloped up and to the right, rising 12.50 units for every one unit of run.

Figure $3.4.2E.4$.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise 12.2.8

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise 12.2.9

Graph the equation from Exercise.

Answer

This is a graph of the equation y = 10 + 5x. The x-axis is labeled in intervals of 1 from 0 - 7; the y-axis is labeled in intervals of 10 from 0 - 50. The equation's graph is a line that crosses the y-axis at 10 and is sloped up and to the right, rising 5 units for every one unit of run.

Figure $3.4.2E.5$.

Exercise 12.2.10

Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise 12.2.11

Which of the following equations are linear?

a. $y = 6x + 8$
b. $y + 7 = 3x$
c. $y - x = 8x^2$
d. $4y = 8$

Answer

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

? Exercise 12.2.12

Does the graph show a linear equation? Why or why not?

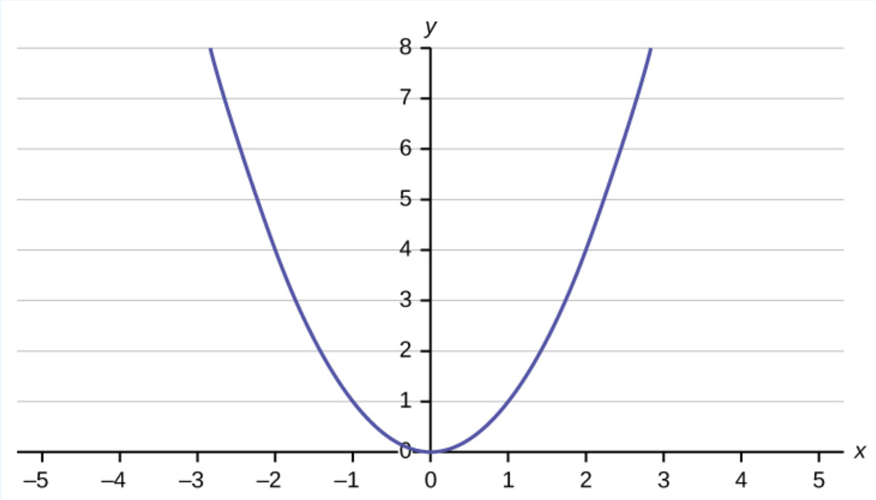


Figure $3.4.2E.6$.

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736

Year	# AIDS cases diagnosed	# AIDS deaths
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

? Exercise 12.2.13

Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Answer

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

? Exercise 12.2.14

What are the independent and dependent variables?

? Exercise 12.2.15

What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of \$50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges \$100 for each hour they clean.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12,000x$.

? Exercise 12.2.16

What are the independent and dependent variables?

? Exercise 12.2.17

How many pounds of soil does the shoreline lose in a year?

Answer

12,000 pounds of soil

Exercise 12.2.18

What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 - 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise 12.2.19

What are the slope and y-intercept? Interpret their meaning.

Answer

The slope is -1.5 ($b = -1.5$). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is $15 ($a = 15$). This means the price of stock before the trading day was $15.

Exercise 12.2.19

If you owned this stock, would you want a positive or negative slope? Why?

3.4.3: Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a *scatter plot*. The following example illustrates a scatter plot.

Example 3.4.3.1

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let $x =$ the year and let $y =$ the number of m-commerce users, in millions.

Table 3.4.3.1: Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

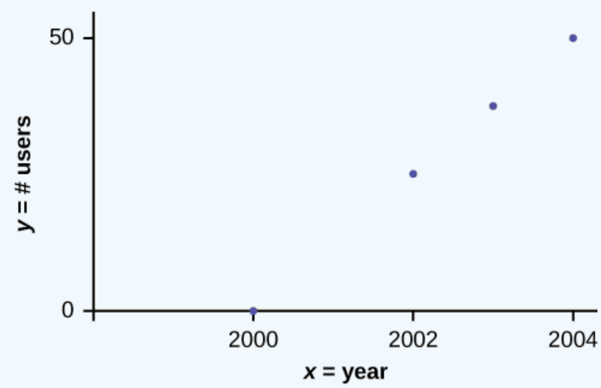


Figure 3.4.3.1: Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot

a. Enter your X data into list L1 and your Y data into list L2.
b. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
c. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
d. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
e. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
f. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
g. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Exercise 3.4.3.1

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Answer

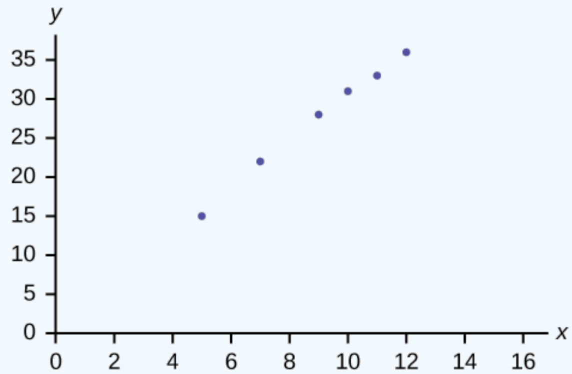


Figure 3.4.3.2

Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the *direction of a relationship* between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the *strength of the relationship* by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatter plot, you want to notice the *overall pattern* and any *deviations* from the pattern. The following scatterplot examples illustrate these concepts.

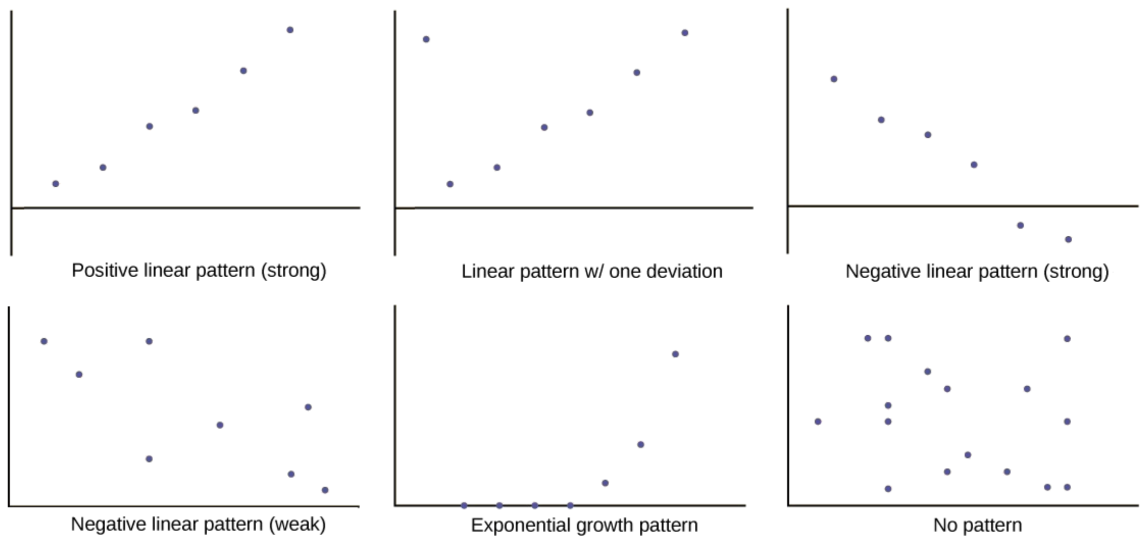


Figure $3.4.3.3$:

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

Summary

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

3.4.3E: Scatter Plots (Exercises)

Exercise $3.4.3E.1$

Does the scatter plot appear linear? Strong or weak? Positive or negative?

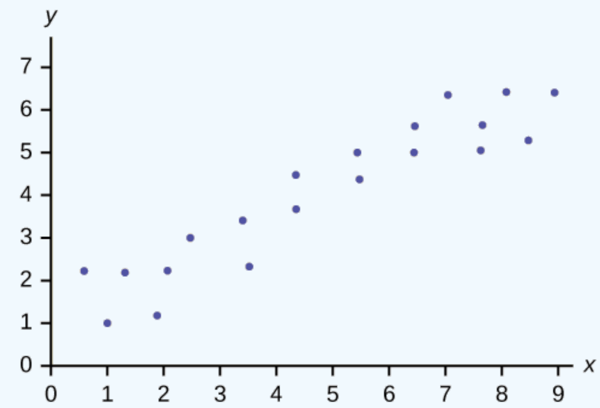


Figure $3.4.3E.4$

Answer

The data appear to be linear with a strong, positive correlation.

Exercise $3.4.3E.3$

Does the scatter plot appear linear? Strong or weak? Positive or negative?

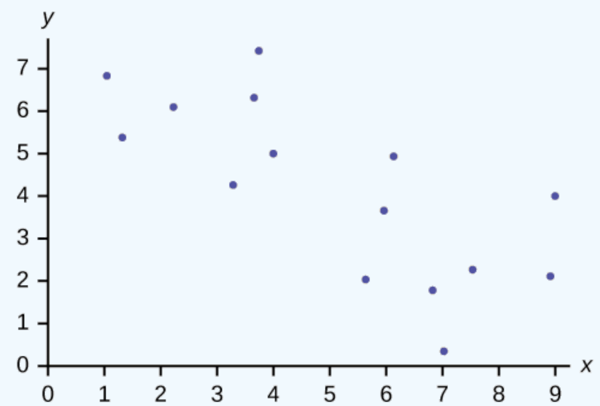


Figure $3.4.3E.5$

Exercise $3.4.3E.4$

Does the scatter plot appear linear? Strong or weak? Positive or negative?

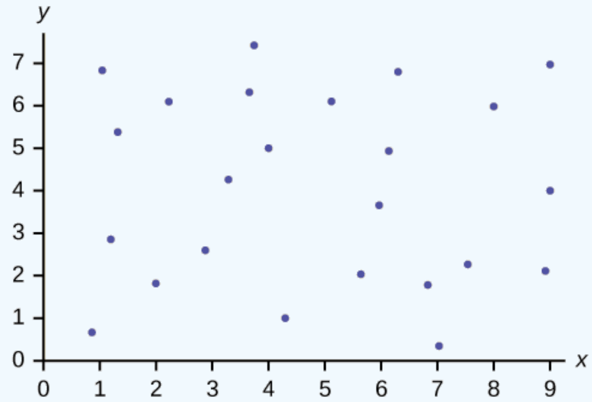


Figure $3.4.3E.6$

Answer

The data appear to have no correlation.

3.5: PowerPoints

https://professormo.com/holistic/Powerpoint/ch13.pptx

CHAPTER OVERVIEW

4: Fundamental Principle of Counting and Rules of Probability

---

4.1: Videos

Defining probability - Core concepts, explained in detail

Probability trees - Useful tool for conditional probability

Would you take this bet? - Thinking through probability and risk

SECTION OVERVIEW

4.2: Probability Topics

Probability theory is concerned with probability, the analysis of random phenomena. The central objects of probability theory are random variables, stochastic processes, and events: mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

4.2.1: Introduction

This is a photo taken of the night sky. A meteor and its tail are shown entering the earth's atmosphere.

Figure $4.2.1.1$. Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Collaborative Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change} | \text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

4.2.2: Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter *S* is used to denote the sample space. For example, if you flip one fair coin, $S = \{\mathrm{H}, \mathrm{T}\}$ where $\mathrm{H} =$ heads and $\mathrm{T} =$ tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(\mathrm{A})$.

Definition: Probability

The *probability* of any outcome is the long-term relative frequency of that outcome. Probabilities are between zero and one, inclusive (that is, zero and one and all numbers between these values).

- $P(\mathrm{A}) = 0$ means the event A can never happen.
- $P(\mathrm{A}) = 1$ means the event A always happens.
- $P(\mathrm{A}) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{\mathrm{HH}, \mathrm{TH}, \mathrm{HT}, \mathrm{TT}\}$ where $\mathrm{T} =$ tails and $\mathrm{H} =$ heads. The sample space has four outcomes. $\mathrm{A} =$ getting one head. There are two outcomes that meet this condition $\{\mathrm{HT}, \mathrm{TH}\}$, so $P(\mathrm{A}) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event $\mathrm{E} =$ rolling a number that is at least five. There are two outcomes {5, 6}. $P(\mathrm{E}) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

Definition: Law of Large Numbers

This important characteristic of probability experiments is known as the law of large numbers which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

The "OR" Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B. For example, let $\text{A} = \{1, 2, 3, 4, 5\}$ and $\text{B} = \{4, 5, 6, 7, 8\}$. $\text{A OR B} = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event

An outcome is in the event A AND B if the outcome is in both A and B at the same time. For example, let A and B be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then $\text{A AND B} = 4, 5$.

The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that

$$P(\text{A}) + P(\text{A}') = 1.$$

For example, let $\text{S} = \{1, 2, 3, 4, 5, 6\}$and let $\text{A} = 1, 2, 3, 4$. Then, $\text{A}' = 5, 6$ and $P(A) = \frac{4}{6}$, $P(\text{A}') = \frac{2}{6}$, and

$$P(\text{A}) + P(\text{A}') = \frac{4}{6} + \frac{2}{6} = 1.$$

The conditional probability of A given B is written $P(\text{A}|\text{B})$. $P(\text{A}|\text{B})$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space**. We calculate the probability of A from the reduced sample space B. The formula to calculate $P(\text{A}|\text{B})$ is

$$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{\text{P(B)}}$$

where $P(\text{B})$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $\text{S} = \{1, 2, 3, 4, 5, 6\}$. Let $\text{A} =$ face is 2 or 3 and $\text{B} =$ face is even (2, 4, 6). To calculate $P(\text{A}|\text{B})$, we count the number of outcomes 2 or 3 in the sample space $\text{B} = \{2, 4, 6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$\begin{aligned} P(\text{A}|\text{B}) &= \frac{\text{P(A AND B)}}{P(\text{B})} \\ &= \frac{\frac{\text{the number of outcomes that are 2 or 3 and even in S}}{6}}{\frac{\text{the number of outcomes that are even in S}}{6}} \\ &= \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3} \end{aligned}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example 4.2.2.1

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ ______________________________

Let event $A =$ the even numbers and event $B =$ numbers greater than 13.

b. $A =$ ______________________, $B =$ ______________________
c. $P(\mathrm{A}) =$ ______________, $P(\mathrm{B}) =$ _________________
d. $\mathrm{A\ AND\ B} =$ _____________________, $\mathrm{A\ OR\ B} =$ _________________
e. $P(\mathrm{A\ AND\ B}) =$ __________, $P(\mathrm{A\ OR\ B}) =$ ______________
f. $\mathrm{A}' =$ ______________, $P(\mathrm{A}') =$ ______________
g. $P(\mathrm{A}) + P(\mathrm{A}') =$ _____________
h. $P(\mathrm{A}|\mathrm{B}) =$ ____________, $P(\mathrm{B}|\mathrm{A}) =$ ______________; are the probabilities equal?

Answer

a. $\mathrm{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $\mathrm{A} = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, \mathrm{B} = \{14, 15, 16, 17, 18, 19\}$
c. $P(\mathrm{A}) = \frac{9}{19}$, $P(\mathrm{B}) = \frac{6}{19}$
d. $\mathrm{A\ AND\ B} = \{14, 16, 18\}$, $\mathrm{A\ OR\ B} = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
e. $P(\mathrm{A\ AND\ B}) = \frac{3}{19}$, $P(\mathrm{A\ OR\ B}) = \frac{12}{19}$
f. $\mathrm{A}' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19 P(\mathrm{A}') = \frac{10}{19}$
g. $P(\mathrm{A}) + P(\mathrm{A}') = 1$ $\left(\left(\frac{9}{19} + \frac{10}{19} = 1\right)\right.$
h. $P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}} = \frac{3}{6}$, $P(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(A)}} = \frac{3}{9}$, No

Exercise 4.2.2.1

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. $S =$ ______________________________
Let event $A =$ the sum is even and event $B =$ the first number is prime.
b. $A =$ ______________________, $B =$ ______________________
c. $P(\mathrm{A}) =$ ______________, $P(\mathrm{B}) =$ _________________
d. $\mathrm{A\ AND\ B} =$ _____________________, $\mathrm{A\ OR\ B} =$ _________________
e. $P(\mathrm{A\ AND\ B}) =$ __________, $P(\mathrm{A\ OR\ B}) =$ ______________
f. $\mathrm{B}' =$ ______________, $P(\mathrm{B}') =$ ______________
g. $P(\mathrm{A}) + P(\mathrm{A}') =$ _____________
h. $P(\mathrm{A}|\mathrm{B}) =$ ____________, $P(\mathrm{B}|\mathrm{A}) =$ ______________; are the probabilities equal?

Answer

a. $\mathrm{S} = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4)\}$
b. $\mathrm{A} = \{(1, 1), (1, 3), (2, 2), (2, 4), (3, 1), (3, 3)\}$
$\mathrm{B} = \{(2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4)\}$
c. $P(\mathrm{A}) = \frac{1}{2}$, $P(\mathrm{B}) = \frac{2}{3}$
d. $\mathrm{A\ AND\ B} = \{(2, 2), (2, 4), (3, 1), (3, 3)\}$
$\mathrm{A\ OR\ B} = \{(1, 1), (1, 3), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4)\}$
e. $P(\mathrm{A\ AND\ B}) = \frac{1}{3}$, $P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
f. $\mathrm{B}' = \{(1, 1), (1, 2), (1, 3), (1, 4)\}$, $P(\mathrm{B}') = \frac{1}{3}$
g. $P(\mathrm{B}) + P(\mathrm{B}') = 1$
h. $P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} = \frac{1}{2}$, $P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} = \frac{2}{3}$, No.

Example $4.2.2.2A$

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

a. Event $\mathrm{T}=$ the outcome is two.
b. Event $\mathrm{A}=$ the outcome is an even number.
c. Event $\mathrm{B}=$ the outcome is less than four.
d. The complement of A.
e. A GIVEN B
f. B GIVEN A
g. A AND B
h. A OR B
i. A OR B'
j. Event $\mathrm{N}=$ the outcome is a prime number.
k. Event $\mathrm{I}=$ the outcome is seven.

Solution

a. $\mathrm{T}=\{2\}, P(\mathrm{T})=\frac{1}{6}$
b. $A=\{2,4,6\}, P(\mathrm{A})=\frac{1}{2}$
c. $\mathrm{B}=\{1,2,3\}, P(\mathrm{B})=\frac{1}{2}$
d. $\mathrm{A}'=\{1,3,5\}, P(\mathrm{A}')=\frac{1}{2}$
e. $\mathrm{A}|\mathrm{B}=\{2\}, P(\mathrm{A}|\mathrm{B})=\frac{1}{3}$
f. $\mathrm{B}|\mathrm{A}=\{2\}, P(\mathrm{B}|\mathrm{A})=\frac{1}{3}$
g. A AND $\mathrm{B}=2, P(\mathrm{A}$ AND $\mathrm{B})=\frac{1}{6}$
h. A OR $\mathrm{B}=\{1,2,3,4,6\}, P(\mathrm{A}$ OR $\mathrm{B})=\frac{5}{6}$
i. A OR $\mathrm{B}'=\{2,4,5,6\}, P(\mathrm{A}$ OR $\mathrm{B}')=\frac{2}{3}$
j. $\mathrm{N}=\{2,3,5\}, P(\mathrm{N})=\frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7)=0$.

Example $4.2.2.2B$

Table describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events $M=$ the subject is male, $F=$ the subject is female, $R=$ the subject is right-handed, $L=$ the subject is left-handed. Compute the following probabilities:

a. $P(\mathrm{M})$
b. $P(\mathrm{F})$
c. $P(\mathrm{R})$
d. $P(\mathrm{L})$
e. $P(\mathrm{M}$ AND $\mathrm{R})$
f. $P(\mathrm{F}$ AND $\mathrm{L})$
g. $P(\mathrm{M}$ OR $\mathrm{F})$
h. $P(\mathrm{M}$ OR $\mathrm{R})$
i. $P(\mathrm{F}$ OR $\mathrm{L})$
j. $P(\mathrm{M'})$

k. $P(\mathrm{R}|\mathrm{M})$
l. $P(\mathrm{F}|\mathrm{L})$
m. $P(\mathrm{L}|\mathrm{F})$

Answer

a. $P(\mathrm{M}) = 0.52$
b. $P(\mathrm{F}) = 0.48$
c. $P(\mathrm{R}) = 0.87$
d. $P(\mathrm{L}) = 0.13$
e. $P(\mathrm{M\ AND\ R}) = 0.43$
f. $P(\mathrm{F\ AND\ L}) = 0.04$
g. $P(\mathrm{M\ OR\ F}) = 1$
h. $P(\mathrm{M\ OR\ R}) = 0.96$
i. $P(\mathrm{F\ OR\ L}) = 0.57$
j. $P(\mathrm{M'}) = 0.48$
k. $P(\mathrm{R}|\mathrm{M}) = 0.8269$ (rounded to four decimal places)
l. $P(\mathrm{F}|\mathrm{L}) = 0.3077$ (rounded to four decimal places)
m. $P(\mathrm{L}|\mathrm{F}) = 0.0833$

References

1. “Countries List by Continent.” Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(\mathrm{S}) = 1$ where S is the sample space

$0 \leq P(\mathrm{A}) \leq 1$

$P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}}$

Glossary

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

Equally Likely

Each outcome of an experiment has the same probability.

Event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

Experiment

a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in *S*. Then:

- $0 \leq P(\mathrm{A}) \leq 1$
- If A and B are any two mutually exclusive events, then $\mathrm{P}(\mathrm{A\ OR\ B}) = P(\mathrm{A}) + P(\mathrm{B})$.
- $P(\mathrm{S}) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event $\mathrm{A\ AND\ B}$ if the outcome is in both $\mathrm{A\ AND\ B}$ at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of *A* GIVEN *B*

$P(\mathrm{A}|\mathrm{B})$ is the probability that event A will occur given that the event B has already occurred.

The Or Event

An outcome is in the event $\mathrm{A\ OR\ B}$ if the outcome is in A or is in B or is in both A and B.

Exercise 3.2.2

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Answer

a. $P(\mathrm{L}') = P(\mathrm{S})$
b. $P(\mathrm{M\ OR\ S})$
c. $P(\mathrm{F\ AND\ L})$
d. $P(\mathrm{M}|\mathrm{L})$
e. $P(\mathrm{L}|\mathrm{M})$
f. $P(\mathrm{S}|\mathrm{F})$
g. $P(\mathrm{F}|\mathrm{L})$

h. $P(\text{F OR L})$
i. $P(\text{M AND S})$
j. $P(\text{F})$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let $H =$ the event of getting a hat.

Let $N =$ the event of getting a noisemaker.

Let $F =$ the event of getting a finger trap.

Let $C =$ the event of getting a bag of confetti.

Exercise 3.2.3

Find $P(\text{H})$.

Exercise 3.2.4

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise 3.2.5

Find $P(\text{F})$.

Exercise 3.2.6

Find $P(\text{C})$.

Answer

$P(\text{C}) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let $B =$ the event of getting a blue jelly bean

Let $G =$ the event of getting a green jelly bean.

Let $O =$ the event of getting an orange jelly bean.

Let $P =$ the event of getting a purple jelly bean.

Let $R =$ the event of getting a red jelly bean.

Let $Y =$ the event of getting a yellow jelly bean.

Exercise 3.2.7

Find $P(\text{B})$.

? Exercise 3.2.8

Find $P(\text{G})$.

Answer

$P(\text{G}) = \frac{20}{150} = \frac{2}{15} = 0.13$

? Exercise 3.2.9

Find $P(\text{P})$.

? Exercise 3.2.10

Find $P(\text{R})$.

Answer

$P(\text{R}) = \frac{22}{150} = \frac{11}{75} = 0.15$

? Exercise 3.2.11

Find $P(\text{Y})$.

? Exercise 3.2.12

Find $P(\text{O})$.

Answer

$P(textO) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

Let $\text{A} =$ the event that a country is in Asia.

Let $\text{E} =$ the event that a country is in Europe.

Let $\text{F} =$ the event that a country is in Africa.

Let $\text{N} =$ the event that a country is in North America.

Let $\text{O} =$ the event that a country is in Oceania.

Let $\text{S} =$ the event that a country is in South America.

? Exercise 3.2.13

Find $P(\text{A})$.

? Exercise 3.2.14

Find $P(\text{E})$.

Answer

$P(\text{E}) = \frac{47}{194} = 0.24$

? Exercise 3.2.15

Find $P(\text{F})$.

? Exercise 3.2.16

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{23}{194} = 0.12$

? Exercise 3.2.17

Find $P(\text{O})$.

? Exercise 3.2.18

Find $P(\text{S})$.

Answer

$P(\text{S}) = \frac{12}{194} = \frac{6}{97} = 0.06$

? Exercise 3.2.19

What is the probability of drawing a red card in a standard deck of 52 cards?

? Exercise 3.2.20

What is the probability of drawing a club in a standard deck of 52 cards?

Answer

$\frac{13}{52} = \frac{1}{4} = 0.25$

? Exercise 3.2.21

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

? Exercise 3.2.22

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Answer

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

CNX_Stats_C03_M01_021.jpg

Figure 4.2.2.1.

Let $\text{B} =$ the event of landing on blue.

Let $\text{R} =$ the event of landing on red.

Let $\text{G} =$ the event of landing on green.

Let $\text{Y} =$ the event of landing on yellow.

? Exercise 3.2.23

If you land on Y, you get the biggest prize. Find $P(\text{Y})$.

? Exercise 3.2.24

If you land on red, you don't get a prize. What is $P(\text{R})$?

Answer

$\text{P}(R) = \frac{4}{8} = 0.5$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

Let $\text{I} =$ the event that a player in an infielder.

Let $\text{O} =$ the event that a player is an outfielder.

Let $\text{H} =$ the event that a player is a great hitter.

Let $\text{N} =$ the event that a player is not a great hitter.

? Exercise 3.2.25

Write the symbols for the probability that a player is not an outfielder.

? Exercise 3.2.26

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Answer

$P(\text{O OR H})$

? Exercise 3.2.27

Write the symbols for the probability that a player is an infielder and is not a great hitter.

? Exercise 3.2.28

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Answer

$P(\text{H}|\text{I})$

? Exercise 3.2.29

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

? Exercise 3.2.30

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Answer

$P(\text{N}|\text{O})$

Exercise 3.2.31

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise 3.2.32

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Answer

$P(\text{I OR N})$

Exercise 3.2.33

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise 3.2.34

Write the symbols for the probability that a player is an infielder.

Answer

$P(\text{I})$

Exercise 3.2.35

What is the word for the set of all possible outcomes?

Exercise 3.2.36

What is conditional probability?

Answer

The likelihood that an event will occur given that another event has already occurred.

Exercise 3.2.37

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

Let $\text{F} =$ event that book is fiction

Let $\text{N} =$ event that book is nonfiction

What is the sample space?

Exercise 3.2.38

What is the sum of the probabilities of an event and its complement?

Answer

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let $\text{E} =$ the event that it lands on an even number. Let $\text{M} =$ the event that it lands on a multiple of three.

Exercise 3.2.39

What does $P(\mathrm{E}|\mathrm{M})$ mean in words?

Exercise 3.2.40

What does $P(\mathrm{E\ OR\ M})$ mean in words?

Answer

the probability of landing on an even number or a multiple of three

4.2.3: Independent and Mutually Exclusive Events

Independent and *mutually exclusive* do not mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A})$
- $P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})$
- $P(\mathrm{A\ AND\ B}) = P(\mathrm{A})P(\mathrm{B})$

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show only one of the above conditions. If two events are NOT independent, then we say that they are dependent.

Sampling a population

Sampling may be done with replacement or without replacement (Figure $4.2.3.1$):

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be *independent*, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be *dependent* or *not independent*.

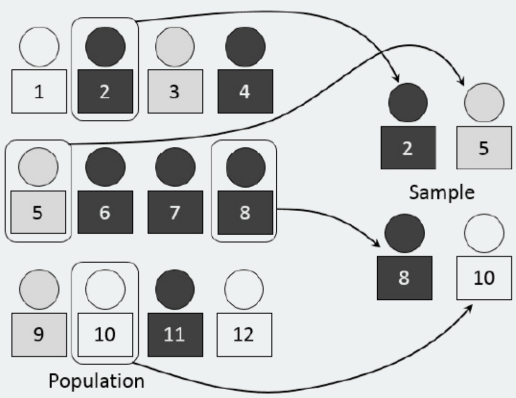


Figure 4.2.3.1: A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". (CC BY-SA 4.0; Dan Kernler).

If it is not known whether A and B are independent or dependent, assume they are dependent until you can show otherwise.

Example $4.2.3.1$: Sampling with and without replacement

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit.

a. Sampling with replacement:

Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are $\{\mathrm{Q}$ of spades, ten of clubs, Q of spades$\}$. You have picked the Q of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:

Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

Exercise 4.2.3.1

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?

Answer a

With replacement

Answer b

No

Example 4.2.3.2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. $S =$ spades, $H =$ Hearts, $D =$ Diamonds, $C =$ Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are $QS, 1D, 1C, QD$.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are $KH, 7D, 6D, KH$.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Answer a

Without replacement

Answer b

With replacement

Exercise 4.2.3.2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. $S =$ spades, $H =$ Hearts, $D =$ Diamonds, $C =$ Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. $QS, 1D, 1C, QD$
b. $KH, 7D, 6D, KH$
c. $QS, 7D, 6D, KS$

Answer - without replacement

a. Possible; b. Impossible, c. Possible

Answer - with replacement

a. Possible; c. Possible, c. Possible

Mutually Exclusive Events

A and B are mutually exclusive events if they **cannot** occur at the same time. This means that A and B do not share any outcomes and $P(\text{A AND B}) = 0$.

For example, suppose the sample space

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Let $\text{A} = \{1, 2, 3, 4, 5\}, \text{B} = \{4, 5, 6, 7, 8\}$ and $\text{C} = \{7, 9\}$. $\text{A AND B} = \{4, 5\}$.

$$P(\text{A AND B}) = \frac{2}{10}$$

and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $P(\text{A AND C}) = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example $4.2.3.3$

Flip two fair coins.

The sample space is $\{HH, HT, TH, TT\}$ where $T =$ tails and $H =$ heads. The outcomes are $HH, HT, TH,$ and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let $\text{A} =$ the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as $\{HH, HT, TH\}$. The outcome HH shows zero tails. HT and TH each show one tail.
- Let $\text{B} =$ the event of getting all tails. B can be written as $\{TT\}$. B is the **complement** of A, so $\text{B} = \text{A}'$. Also, $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$.
- The probabilities for A and for B are $P(\text{A}) = \frac{3}{4}$ and $P(\text{B}) = \frac{1}{4}$.
- Let $\text{C} =$ the event of getting all heads. $\text{C} = \{HH\}$. Since $\text{B} = \{TT\}$, $P(\text{B AND C}) = 0$. B and $\text{C}$are mutually exclusive. B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let $\text{D} =$ event of getting **more than one** tail. $\text{D} = \{TT\}$. $P(\text{D}) = \frac{1}{4}$
- Let $\text{E} =$ event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $\text{E} = \{HT, HH\}$. $P(\text{E}) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let $\text{F} =$ event of getting at least one tail in two flips. $\text{F} = \{HT, TH, TT\}$. $P(\text{F}) = \frac{3}{4}$

Exercise $4.2.3.3$

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Answer

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is $\{BB, BR, RB, RR\}$.

Event $A =$ Getting at least one black card $= \{BB, BR, RB\}$

$P(\text{A}) = \frac{3}{4} = 0.75$

Example 4.2.3.4

Flip two fair coins. Find the probabilities of the events.

a. Let $\text{F} =$ the event of getting at most one tail (zero or one tail).
b. Let $\text{G} =$ the event of getting two faces that are the same.
c. Let $\text{H} =$ the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are F and G mutually exclusive?
e. Let $\text{J} =$ the event of getting all tails. Are J and H mutually exclusive?

Solution

Look at the sample space in Example $4.2.3.3$

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(\text{F}) = \frac{3}{4}$
2. Two faces are the same if HH or TT show up. $P(\text{G}) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(\text{H}) = \frac{2}{4}$
4. F and G share HH so $P(\text{F AND G})$ is not equal to zero (0). F and G are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins (TT). H's outcomes are HH and HT.

J and H have nothing in common so $P(\text{J AND H}) = 0$. J and H are mutually exclusive.

Exercise 4.2.3.4

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let $\text{F} =$ the event of getting the white ball twice.
b. Let $\text{G} =$ the event of getting two balls of different colors.
c. Let $\text{H} =$ the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

Answer

a. $P(\text{F}) = \frac{1}{4}$
b. $P(\text{G}) = \frac{1}{2}$
c. $P(\text{H}) = \frac{1}{2}$
d. Yes
e. No

Example 4.2.3.5

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event $\text{A} =$ a face is odd. Then $\text{A} = \{1, 3, 5\}$. Let event $\text{B} =$ a face is even. Then $\text{B} = \{2, 4, 6\}$.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$. Also, $P(\text{A}) = \frac{3}{6}$ and $P(\text{B}) = \frac{3}{6}$.
- Let event $\text{C} =$ odd faces larger than two. Then $\text{C} = \{3, 5\}$. Let event $\text{D} =$ all even faces smaller than five. Then $\text{D} = \{2, 4\}$. $P(\text{C AND D}) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event $\text{E} =$ all faces less than five. $\text{E} = \{1, 2, 3, 4\}$.

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Answer

No. $\text{C} = \{3, 5\}$ and $\text{E} = \{1, 2, 3, 4\}$. $P(\text{C AND E}) = \frac{1}{6}$. To be mutually exclusive, $P(\text{C AND E})$ must be zero.

- Find $P(\text{C}|\text{A})$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(\text{C}|\text{A})$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(\text{C}|\text{A}) = \frac{2}{3}$.

Exercise 4.2.3.5

Let event $\text{A} =$ learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(\text{A}) = 0.4$ and $P(\text{B}) = 0.2$. $P(\text{A AND B}) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(\text{A}|\text{B}) = P(\text{A})$
- $P(\text{B}|\text{A})$
- $P(\text{A AND B}) = P(\text{A})P(\text{B})$

Answer

$$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{P(\text{B})} = \frac{0.08}{0.2} = 0.4 = P(\text{A}) \tag{4.2.3.1}$$

The events are independent because $P(\text{A}|\text{B}) = P(\text{A})$.

Example 4.2.3.6

Let event $\text{G} =$ taking a math class. Let event $\text{H} =$ taking a science class. Then, $\text{G AND H} =$ taking a math class and a science class. Suppose $P(\text{G}) = 0.6$, $P(\text{H}) = 0.5$, and $P(\text{G AND H}) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(\text{G}|\text{H}) = P(\text{G})$
- $P(\text{H}|\text{G}) = P(\text{H})$
- $P(\text{G AND H}) = P(\text{G})P(\text{H})$

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

a. a. Show that $P(\text{G}|\text{H}) = P(\text{G})$.
b. b. Show $P(\text{G AND H}) = P(\text{G})P(\text{H})$.

Solution

a. $P(\text{G}|\text{H}) = \frac{P(\text{G AND H})}{P(\text{H})} = \frac{0.3}{0.5} = 0.6 = P(\text{G})$
b. $P(\text{G})P(\text{H}) = (0.6)(0.5) = 0.3 = P(\text{G AND H})$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(\text{H}|\text{G}) = P(\text{H})$ to show that G and H are independent events.

Exercise 4.2.3.6

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- $\text{R} =$ a red marble
- $\text{G} =$ a green marble

- $\text{O} =$ an odd-numbered marble
- The sample space is $\text{S} = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(\text{G AND O})$?

Answer

Event G and $\text{O} = \{G1, G3\}$

$P(\text{G and O}) = \dfrac{2}{10} = 0.2$

✓ Example 4.2.3.7

Let event $\text{C} =$ taking an English class. Let event $\text{D} =$ taking a speech class.

Suppose $P(\text{C}) = 0.75$, $P(\text{D}) = 0.3$, $P(\text{C}|\text{D}) = 0.75$ and $P(\text{C AND D}) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(\text{D}|\text{C})$?

Solution

a. Yes, because $P(\text{C}|\text{D}) = P(\text{C})$.
b. No, because $P(\text{C AND D})$ is not equal to zero.
c. $P(\text{D}|\text{C}) = \dfrac{P(\text{C AND D})}{P(\text{C})} = \dfrac{0.225}{0.75} = 0.3$

? Exercise 4.2.3.7

A student goes to the library. Let events $\text{B} =$ the student checks out a book and $\text{D} =$ the student checks out a DVD. Suppose that $P(\text{B}) = 0.40$, $P(\text{D}) = 0.30$ and $P(\text{B AND D}) = 0.20$.

a. Find $P(\text{B}|\text{D})$.
b. Find $P(\text{D}|\text{B})$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Answer

a. $P(\text{B}|\text{D}) = 0.6667$
b. $P(\text{D}|\text{B}) = 0.5$
c. No
d. No

✓ Example 4.2.3.8

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let

- $\text{R} =$ red card is drawn,
- $\text{B} =$ blue card is drawn,
- $\text{E} =$ even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$.

S has eight outcomes.

- $P(\text{R}) = \frac{3}{8}$. $P(\text{B}) = \frac{5}{8}$. $P(\text{R AND B}) = 0$. (You cannot draw one card that is both red and blue.)
- $P(\text{E}) = \frac{3}{8}$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(\text{E}|\text{B}) = \frac{2}{5}$. (There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(\text{B}|\text{E}) = \frac{2}{3}$. (There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(\text{R AND B}) = 0$.
- Let $\text{G} =$ card with a number greater than 3. $\text{G} = \{B4, B5\}$. $P(\text{G}) = \frac{2}{8}$. Let $\text{H} =$ blue card numbered between one and four, inclusive. $\text{H} = \{B1, B2, B3, B4\}$. $P(\text{G}|\text{H}) = \frac{1}{4}$. (The only card in H that has a number greater than three is B4.) Since $\frac{2}{8} = \frac{1}{4}$, $P(\text{G}) = P(\text{G}|\text{H})$, which means that G and H are independent.

Exercise 4.2.3.8

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.

Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Answer

- $P(\text{B}|\text{A}) = 0.67$
- $P(\text{B}) = 0.25$

So $P(\text{B})$ does not equal $P(\text{B}|\text{A})$ which means that $\text{B} and \text{A}$ are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND A}) = 0.20$, not 0.

Example 4.2.3.9

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(\text{F}) = 0.60$; $P(\text{L}) = 0.50$
- $P(\text{F AND L}) = 0.45$
- $P(\text{L}|\text{F}) = 0.75$

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(\text{F}|\text{L})$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(\text{F AND L}) = P(\text{F})P(\text{L})$. We are given that $P(\text{F AND L}) = 0.45$, but $P(\text{F})P(\text{L}) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(\text{F AND L})$ does not equal $P(\text{F})P(\text{L})$.

Solution 2

Check whether $P(\text{L}|\text{F})$ equals $P(\text{L})$. We are given that $P(\text{L}|\text{F}) = 0.75$, but $P(\text{L}) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

? Exercise 4.2.3.9

Mark is deciding which route to take to work. His choices are $\text{I} = \text{the Interstate}$ and $\text{F} = \text{Fifth Street}$

- $P(\text{I}) = 0.44$ and $P(\text{F}) = 0.55$
- $P(\text{I AND F}) = 0$ because Mark will take only one route to work.

What is the probability of $P(\text{I OR F})$?

Answer

Because $P(\text{I AND F}) = 0$,

$P(\text{I OR F}) = P(\text{I}) + P(\text{F}) - P(\text{I AND F}) = 0.44 + 0.56 - 0 = 1$

✓ Example 4.2.3.10

a. Toss one fair coin (the coin has two sides, H and T). The outcomes are ________. Count the outcomes. There are ____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are ________________. Count the outcomes. There are ___ outcomes.
c. Multiply the two numbers of outcomes. The answer is _______.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are $H1$ and $T6$.)
e. Event $\text{A} =$ heads (H) on the coin followed by an even number (2, 4, 6) on the die. A = {_________________}. Find $P(\text{A})$.
f. Event $\text{B} =$ heads on the coin followed by a three on the die. $\text{B} =$ {________}. Find $P(\text{B})$.
g. Are A and B mutually exclusive? (Hint: What is $P(\text{A AND B})$? If $P(\text{A AND B}) = 0$, then A and B are mutually exclusive.)
h. Are A and B independent? (Hint: Is $P(\text{A AND B}) = P(\text{A})P(\text{B})$? If $P(\text{A AND B}) = P(\text{A})P(\text{B})$, then A and B are independent. If not, then they are dependent).

Solution

a. H and T; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. $T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6$
e. $\text{A} = \{H2, H4, H6\}$; $P(\text{A}) = \frac{3}{12}$
f. $\text{B} = \{H3\}$; $P(\text{B}) = \frac{1}{12}$
g. Yes, because $P(\text{A AND B}) = 0$
h. $P(\text{A AND B}) = 0$. $P(\text{A})P(\text{B}) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(\text{A AND B})$ does not equal $P(\text{A})P(\text{B})$, so A and B are dependent.

? Exercise 4.2.3.10

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

a. Compute $P(\mathrm{T})$.
b. Compute $P(\mathrm{T} \mid \mathrm{F})$.
c. Are T and F independent?.
d. Are F and S mutually exclusive?
e. Are F and S independent?

Answer

a. $P(\mathrm{T})=\frac{1}{4}$
b. $P(\mathrm{T} \mid \mathrm{F})=\frac{1}{2}$
c. No
d. No
e. Yes

References

1. Lopez, Shane, Preety Sidhu. "U.S. Teachers Love Their Lives, but Struggle in the Workplace." Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/te...workplace.aspx (accessed May 2, 2013).
2. Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

- If A and B are independent, $P(\mathrm{A \text{ AND } B})=P(\mathrm{A}) P(\mathrm{B}), P(\mathrm{A} \mid \mathrm{B})=P(\mathrm{A}) \quad$ and $P(\mathrm{B} \mid \mathrm{A})=P(\mathrm{B})$.
- If A and B are mutually exclusive, $P(\mathrm{A \text{ OR } B})=P(\mathrm{A})+P(\mathrm{B}) and P(\mathrm{A \text{ AND } B})=0 \quad$.

Exercise 4.2.3.11

E and F are mutually exclusive events. $P(\mathrm{E})=0.4$; $P(\mathrm{F})=0.5$. Find $P(\mathrm{E} \mid \mathrm{F})$.

Exercise 4.2.3.12

J and K are independent events. $P(\mathrm{J} \mid \mathrm{K})=0.3$. Find $P(\mathrm{J})$.

Answer

$P(\mathrm{J})=0.3$

Exercise 4.2.3.13

U and V are mutually exclusive events. $P(\mathrm{U})=0.26$; $P(\mathrm{V})=0.37$. Find:

a. $P(\mathrm{U \text{ AND } V})=$
b. $P(\mathrm{U} \mid \mathrm{V})=$
c. $P(\mathrm{U \text{ OR } V})=$

Exercise 4.2.3.14

Q and R are independent events. $P(\text{Q}) = 0.4$ and $P(\text{Q AND R}) = 0.1$. Find $P(\text{R})$.

Answer

$P(\text{Q AND R}) = P(\text{Q})P(\text{R})$

$0.1 = (0.4)P(\text{R})$

$P(\text{R}) = 0.25$

Bringing It Together

Exercise 4.2.3.16

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into Table.

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\#1}-33| \leq 210 \text{ pounds})$?

Exercise 4.2.3.17

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let $\text{C} =$ a man develops cancer in his lifetime and $\text{P} =$ man has at least one false positive.

a. $P(\text{C}) =$ ______
b. $P(\text{P}|\text{C}) =$ ______
c. $P(\text{P}|\text{C'}) =$ ______
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Answer

a. $P(\text{C}) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise 4.2.3.18

Given events G and H : $P(\text{G}) = 0.43$; $P(\text{H}) = 0.26$; $P(\text{H AND G}) = 0.14$

a. Find $P(\text{H OR G})$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

Exercise 4.2.3.19

Given events J and K : $P(\text{J}) = 0.18$; $P(\text{K}) = 0.37$; $P(\text{J OR K}) = 0.45$

a. Find $P(\text{J AND K})$.
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J AND K).

Answer

a. $P(\text{J OR K}) = P(\text{J}) + P(\text{K}) - P(\text{J AND K}); 0.45 = 0.18 + 0.37 - P(\text{J AND K})$; solve to find $P(\text{J AND K}) = 0.10$
b. $P(\text{NOT (J AND K)}) = 1 - P(\text{J AND K}) = 1 - 0.10 = 0.90$
c. $P(\text{NOT (J OR K)}) = 1 - P(\text{J OR K}) = 1 - 0.45 = 0.55$

Glossary

Dependent Events
If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement
If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement
When sampling is done without replacement, each member of a population may be chosen only once.

The Conditional Probability of One Event Given Another Event
P(*A*|*B*) is the probability that event *A* will occur given that the event *B* has already occurred.

The OR of Two Events
An outcome is in the event *A* OR *B* if the outcome is in *A*, is in *B*, or is in both *A* and *B*.

4.2.4: Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a sample space, then:

$$P(A \text{ AND } B) = P(B)P(A|B) \tag{4.2.4.1}$$

This rule may also be written as:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are *independent*, then

$$P(A|B) = P(A).$$

and Equation 4.2.4.1becomes

$$P(A \text{ AND } B) = P(A)P(B).$$

The Addition Rule

If A and B are defined on a sample space, then:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) \tag{4.2.4.2}$$

If A and B are **mutually exclusive**, then

$$P(A \text{ AND } B) = 0.$$

and Equation 4.2.4.2becomes

$$P(A \text{ OR } B) = P(A) + P(B).$$

Example 4.2.4.1

Klaus is trying to choose where to go on vacation. His two choices are: $\mathrm{A} = \text{New Zealand}$ and $\mathrm{B} = \text{Alaska}$.

- Klaus can only afford one vacation. The probability that he chooses A is $P(\mathrm{A}) = 0.6$ and the probability that he chooses B is $P(\mathrm{B}) = 0.35$.
- $P(\mathrm{A \text{ AND } B}) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(\mathrm{A \text{ OR } B}) = P(\mathrm{A}) + P(\mathrm{B}) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. $\mathrm{A} =$ the event Carlos is successful on his first attempt. $P(\mathrm{A}) = 0.65$. $\mathrm{B} =$ the event Carlos is successful on his second attempt. $P(\mathrm{B}) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are A and B independent?
d. Are A and B mutually exclusive?

Solutions

a. The problem is asking you to find $P(\text{A AND B}) = P(\text{B AND A})$. Since $P(\text{B}|\text{A}) = 0.90 : P(\text{B AND A}) = P(\text{B}|\text{A})P(\text{A}) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(\text{A OR B})$.

$$P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B}) = 0.65 + 0.65 - 0.585 = 0.715 \tag{4.2.4.3}$$

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(\text{B AND A}) = 0.585$.

$$P(\text{B})P(\text{A}) = (0.65)(0.65) = 0.423 \tag{4.2.4.4}$$

$$0.423 \neq 0.585 = P(\text{B AND A}) \tag{4.2.4.5}$$

So, $P(\text{B AND A})$ is **not** equal to $P(\text{B})P(\text{A})$.

d. No, they are not because $P(\text{A and B}) = 0.585$.

To be mutually exclusive, $P(\text{A AND B})$ must equal zero.

Exercise 4.2.4.1

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. $\text{C} =$ the event that Helen makes the first shot. $P(\text{C}) = 0.75$. $\text{D} =$ the event Helen makes the second shot. $P(\text{D}) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Answer

$$P(\text{D}|\text{C}) = 0.85 \tag{4.2.4.6}$$

$$P(\text{C AND D}) = P(\text{D AND C}) \tag{4.2.4.7}$$

$$P(\text{D AND C}) = P(\text{D}|\text{C})P(\text{C}) = (0.85)(0.75) = 0.6375 \tag{4.2.4.8}$$

Helen makes the first and second free throws with probability 0.6375.

Example 4.2.4.2

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?
b. What is the probability that the member practices four times a week?
c. What is the probability that the member is an advanced swimmer and practices four times a week?
d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Answer

a. $\frac{28}{150}$
b. $\frac{80}{150}$
c. $\frac{40}{150}$
d. $P(\text{advanced AND intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

e. No, these are not independent events.

$$P(\text{novice AND practices four times per week}) = 0.0667 \tag{4.2.4.9}$$

$$P(\text{novice})P(\text{practices four times per week}) = 0.0996 \tag{4.2.4.10}$$

$$0.0667 \neq 0.0996 \tag{4.2.4.11}$$

? Exercise 4.2.4.2

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Answer

$$P = \frac{200-140-40}{200} = \frac{20}{200} = 0.1 \tag{4.2.4.12}$$

✓ Example 4.2.4.3

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: $\mathrm{M} =$ math class, $\mathrm{S} =$ speech class, $\mathrm{M}|\mathrm{S} =$ math given speech

a. What is the probability that Felicity enrolls in math and speech?
   Find $P(\mathrm{M\ AND\ S}) = P(\mathrm{M}|\mathrm{S})P(\mathrm{S})$.
b. What is the probability that Felicity enrolls in math or speech classes?
   Find $P(\mathrm{M\ OR\ S}) = P(\mathrm{M}) + P(\mathrm{S}) - P(\mathrm{M\ AND\ S})$.
c. Are M and S independent? Is $P(\mathrm{M}|\mathrm{S}) = P(\mathrm{M})$?
d. Are M and S mutually exclusive? Is $P(\mathrm{M\ AND\ S}) = 0$?

Answer

a. 0.1625, b. 0.6875, c. No, d. No

? Exercise 4.2.4.3

A student goes to the library. Let events $\mathrm{B} =$ the student checks out a book and $\mathrm{D} =$ the student check out a DVD. Suppose that $P(\mathrm{B}) = 0.40$, $P(\mathrm{D}) = 0.30$ and $P(\mathrm{D}|\mathrm{B}) = 0.5$.

a. Find $P(\mathrm{B\ AND\ D})$.
b. Find $P(\mathrm{B\ OR\ D})$.

Answer

a. $P(\mathrm{B\ AND\ D}) = P(\mathrm{D}|\mathrm{B})P(\mathrm{B}) = (0.5)(0.4) = 0.20$.
b. $P(\mathrm{B\ OR\ D}) = P(\mathrm{B}) + P(\mathrm{D}) - P(\mathrm{B\ AND\ D}) = 0.40 + 0.30 - 0.20 = 0.50$

✓ Example 4.2.4.4

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let $\mathrm{B} =$ woman develops breast cancer and let $\mathrm{N} =$ tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?
b. Given that the woman has breast cancer, what is the probability that she tests negative?
c. What is the probability that the woman has breast cancer AND tests negative?

d. What is the probability that the woman has breast cancer or tests negative?
e. Are having breast cancer and testing negative independent events?
f. Are having breast cancer and testing negative mutually exclusive?

Answers

a. $P(\text{B}) = 0.143$; $P(\text{N}) = 0.85$
b. $P(\text{N}|\text{B}) = 0.02$
c. $P(\text{B AND N}) = P(\text{B})P(\text{N}|\text{B}) = (0.143)(0.02) = 0.0029$
d. $P(\text{B OR N}) = P(\text{B}) + P(\text{N}) - P(\text{B AND N}) = 0.143 + 0.85 - 0.0029 = 0.9901$
e. No. $P(\text{N}) = 0.85$; $P(\text{N}|\text{B}) = 0.02$. So, $P(\text{N}|\text{B})$ does not equal $P(\text{N})$.
f. No. $P(\text{B AND N}) = 0.0029$. For B and N to be mutually exclusive, $P(\text{B AND N})$ must be zero

Exercise 4.2.4.4

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Answer

Let $\text{A} =$ student is a senior going to college.

Let $\text{B} =$ student plays sports.

$P(\text{B}) = \dfrac{140}{200}$

$P(\text{B}|\text{A}) = \dfrac{50}{140}$

$P(\text{A AND B}) = P(\text{B}|\text{A})P(\text{A})$

$P(\text{A AND B}) = (\dfrac{140}{200})(\dfrac{50}{140}) = \dfrac{1}{4}$

Example 4.2.4.5

Refer to the information in Example $4.2.4.4$ $\text{P} =$ tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(\text{P}|\text{B}) = 1 - P(\text{N}|\text{B})$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(\text{B AND P}) = P(\text{P}|\text{B})P(\text{B})$.
c. What is the probability that a woman does not develop breast cancer. Find $P(\text{B}') = 1 - P(\text{B})$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(\text{P}) = 1 - P(\text{N})$.

Answer

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Exercise 4.2.4.5

A student goes to the library. Let events $\text{B} =$ the student checks out a book and $\text{D} =$ the student checks out a DVD. Suppose that $P(\text{B}) = 0.40$, $P(\text{D}) = 0.30$ and $P(\text{D}|\text{B}) = 0.5$.

a. Find $P(\text{B}')$.
b. Find $P(\text{D AND B})$.
c. Find $P(\text{B}|\text{D})$.
d. Find $P(\text{D AND B}')$.
e. Find $P(\text{D}|\text{B}')$.

Answer

a. $P(\text{B}') = 0.60$

b. $P(\text{D AND B}) = P(\text{D}|\text{B})P(\text{B}) = 0.20$

c. $P(\text{B}|\text{D}) = \dfrac{P(\text{B AND D})}{P(\text{D})} = \dfrac{(0.20)}{(0.30)} = 0.66$

d. $P(\text{D AND B}') = P(\text{D}) - P(\text{D AND B}) = 0.30 - 0.20 = 0.10$

e. $P(\text{D}|\text{B}') = P(\text{D AND B}')P(\text{B}') = (P(\text{D}) - P(\text{D AND B}))(0.60) = (0.10)(0.60) = 0.06$

References

1. DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at www.field.com/fieldpollonline...rs/Rls2443.pdf (accessed May 2, 2013).
2. Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at www.thestar.com/news/gta/2011..._suggests.html (accessed May 2, 2013).
3. "Mayor's Approval Down." News Release by Forum Research Inc. Available online at www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).
4. "Roulette." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Roulette (accessed May 2, 2013).
5. Shin, Hyon B., Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at www.census.gov/hhes/socdemo/l...acs/ACS-12.pdf (accessed May 2, 2013).
6. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).
7. Data from U.S. Census Bureau.
8. Data from the Wall Street Journal.
9. Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at www.ropercenter.uconn.edu/ (accessed May 2, 2013).
10. Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Review

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(\text{A AND B}) = P(\text{A}|\text{B})P(\text{B})$

The addition rule: $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B})$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- $\text{C} =$ Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- $\text{L} =$ Latino Californians

Suppose that one Californian is randomly selected.

Exercise 4.2.4.5

Find $P(\text{C})$.

? Exercise 4.2.4.6

Find $P(\text{L})$.

Answer

0.376

? Exercise 4.2.4.7

Find $P(\text{C} | \text{L})$.

? Exercise 4.2.4.8

In words, what is $\text{C} | \text{L}$?

Answer

$\text{C} | \text{L}$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

? Exercise 4.2.4.9

Find $P(\text{L AND C})$

? Exercise 4.2.4.10

In words, what is L AND C?

Answer

L AND C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

? Exercise 4.2.4.11

Are L and C independent events? Show why or why not.

? Exercise 4.2.4.12

Find $P(\text{L OR C})$.

Answer

0.6492

? Exercise 4.2.4.13

In words, what is L OR C?

? Exercise 4.2.4.14

Are L and C mutually exclusive events? Show why or why not.

Answer

No, because $P(\text{L AND C})$ does not equal 0.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(\mathrm{A} \mid \mathrm{B})=P(\mathrm{A})$
2. $P(\mathrm{B} \mid \mathrm{A})=P(\mathrm{B})$
3. $P(\mathrm{A} \text{ AND } \mathrm{B})=P(\mathrm{A}) P(\mathrm{B})$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(\mathrm{A} \text{ AND } \mathrm{B})=0$.

4.2.5: Contingency Tables

A *contingency table* provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example 4.2.5.1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

a. Find $P(\text{Person is a cell phone user})$.
b. Find $P(\text{person had no violation in the last year})$.
c. Find $P(\text{Person had no violation in the last year AND was a cell phone user})$.
d. Find $P(\text{Person is a cell phone user OR person had no violation in the last year})$.
e. Find $P(\text{Person is a cell phone user GIVEN person had a violation in the last year})$.
f. Find $P(\text{Person had no violation last year GIVEN person was not a cell phone user})$

Answer

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$
b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$
c. $\frac{280}{755}$
d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$
e. $\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)
f. $\frac{405}{450}$ (The sample space is reduced to the number of persons who were not cell phone users.)

Exercise 4.2.5.1

Table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	**No injury in last year**	**Total**
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is $P(\text{athlete stretches before exercising})$?
b. What is $P(\text{athlete stretches before exercising} \mid \text{no injury in the last year})$?

Answer

a. $P(\text{athlete stretches before exercising}) = \frac{350}{800} = 0.4375$

b. $P(\text{athlete stretches before exercising} | \text{no injury in the last year}) = \frac{295}{514} = 0.5739$

Example 4.2.5.2

Table shows a random sample of 100 hikers and the areas of hiking they prefer.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

a. Complete the table.

b. Are the events "being female" and "preferring the coastline" independent events? Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$
3. Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let $\mathrm{M} =$ being male, and let $\mathrm{L} =$ prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: P(___|___) = ___.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let $\mathrm{F} =$ being female, and let $\mathrm{P} =$ prefers mountain peaks.

1. Find $P(\mathrm{F})$.
2. Find $P(\mathrm{P})$.
3. Find $P(\mathrm{F \text{ AND } P})$.
4. Find $P(\mathrm{F \text{ OR } P})$.

Answers

a.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

b.

$$P(\mathrm{F \text{ AND } C}) = \frac{18}{100} = 0.18$$

$$P(\text{F})P(\text{C}) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(\text{F AND C}) \neq P(\text{F})P(\text{C})$, so the events F and C are not independent.

c.

1. The word 'given' tells you that this is a conditional.
2. $P(\text{M}|\text{L}) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

a. Find $P(\text{F})$.
b. Find $P(\text{P})$.
c. Find $P(\text{F AND P})$.
d. Find $P(\text{F OR P})$.

d.

1. $P(\text{F}) = \frac{45}{100}$
2. $P(\text{P}) = \frac{25}{100}$
3. $P(\text{F AND P}) = \frac{11}{100}$
4. $P(\text{F OR P}) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Exercise 4.2.5.2

Table shows a random sample of 200 cyclists and the routes they prefer. Let $\text{M} =$ males and $\text{H} =$ hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Answer

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$
$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$
$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example 4.2.5.3

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.
b. What is the probability that Alissa does not catch Muddy?
c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. $\frac{41}{60}$

c. $\frac{9}{19}$

✓ Example 4.2.5.4

Table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

a. Find $P(2009 \text{ AND Robbery})$.
b. Find $P(2010 \text{ AND Burglary})$.
c. Find $P(2010 \text{ OR Burglary})$.
d. Find $P(2011|\text{Rape})$
e. Find $P(\text{Vehicle}|2008)$

Answer

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Exercise 4.2.5.3

Table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

References

1. "Blood Types." American Red Cross, 2013. Available online at www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).
2. Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.
3. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).
4. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).
5. "Human Blood Types." Unite Blood Services, 2011. Available online at www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).
6. Samuel, T. M. "Strange Facts about RH Negative Blood." eHow Health, 2013. Available online at www.ehow.com/facts_5552003_st...ive-blood.html (accessed May 2, 2013).
7. "United States: Uniform Crime Report – State Statistics from 1960–2011." The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. Table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Exercise 3.5.4

Find *P*(musician is a female).

Exercise 3.5.5

Find $P(\text{musician is a male AND had private instruction})$.

Answer

$$P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$$

Exercise 3.5.6

Find *P*(musician is a female OR is self taught).

Exercise 3.5.7

Are the events "being a female musician" and "learning music in school" mutually exclusive events?

Answer

The events are not mutually exclusive. It is possible to be a female musician who learned music in school.

Bringing it Together

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

? Exercise 3.5.8

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

? Exercise 3.5.9

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Answer

$$\frac{35,065}{100,450}$$

? Exercise 3.5.10

Find the probability that the person was Latino.

? Exercise 3.5.11

In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

Exercise 3.5.12

In words, explain what it means to pick one person from the study who is “Japanese American **OR** smokes 21 to 30 cigarettes per day.” Also, find the probability.

Exercise 3.5.13

In words, explain what it means to pick one person from the study who is “Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day.” Also, find the probability.

Answer

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4,715}{15,273}$.

Exercise 3.5.14

Prove that smoking level/day and ethnicity are dependent events.

Glossary

contingency table
the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

4.2.6: Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A *tree diagram* is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 4.2.6.1: Probabilities from Sampling with replacement

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, with replacement (remember that "with replacement" means that you put the first ball back in the urn before you select the second ball). The tree diagram using frequencies that show all the possible outcomes follows.

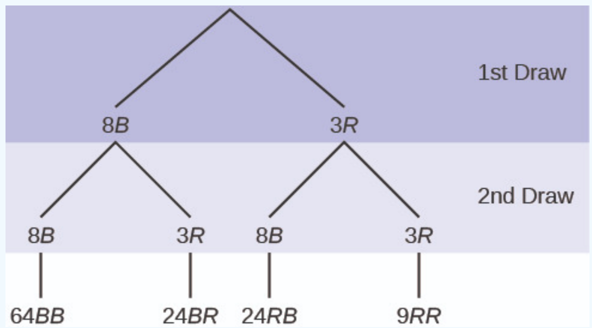


Figure 4.2.6.1: Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as:

*R*1*R*1 *R*1*R*2 *R*1*R*3 *R*2*R*1 *R*2*R*2 *R*2*R*3 *R*3*R*1 *R*3*R*2 *R*3*R*3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the sample space.

Exercise 4.2.6.1

In a standard deck, there are 52 cards. 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(\mathrm{FF})$.

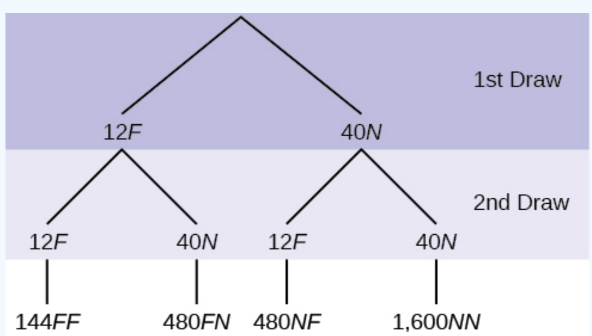


Figure 4.2.6.2:

Answer

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(\text{FF}) = \frac{144}{144+480+480+1,600} = \frac{144}{2,704} = \frac{9}{169} \tag{4.2.6.1}$$

1. a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...
2. b. Using the tree diagram, calculate *P*(*RR*).
3. c. Using the tree diagram, calculate $P(\text{RB OR BR})$.
4. d. Using the tree diagram, calculate $P(\text{R on 1st draw AND B on 2nd draw})$.
5. e. Using the tree diagram, calculate *P*(*R* on 2nd draw GIVEN *B* on 1st draw).
6. Using the tree diagram, calculate $P(\text{BB})$.
7. g. Using the tree diagram, calculate $P(\text{B on the 2nd draw given R on the first draw})$.

Solution

a. *B*1*R*1; *B*1*R*2; *B*1*R*3; *B*2*R*1; *B*2*R*2; *B*2*R*3; *B*3*R*1; *B*3*R*2; *B*3*R*3; *B*4*R*1; *B*4*R*2; *B*4*R*3; *B*5*R*1; *B*5*R*2; *B*5*R*3; *B*6*R*1; *B*6*R*2; *B*6*R*3 *B*7*R*1; *B*7*R*2; *B*7*R*3; *B*8*R*1; *B*8*R*2; *B*8*R*3
b. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$
c. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$
d. $P(\text{R on 1st draw AND B on 2nd draw}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$
e. *P*(*R* on 2nd draw GIVEN *B* on 1st draw) = *P*(*R* on 2nd|*B* on 1st) = $\frac{24}{88}$ = $\frac{3}{11}$ This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $\frac{24}{88}$ = $\frac{3}{11}$
f. $P(\text{BB}) = \frac{64}{121}$
g. $P(\text{B on 2nd draw} | \text{R on 1st draw}) = \frac{8}{11}$. There are 9 + 24 outcomes that have R on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Example 4.2.6.2: Probabilities from Sampling without replacement

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. (remember that "without replacement" means that you do not put the first ball back before you select the second marble). Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

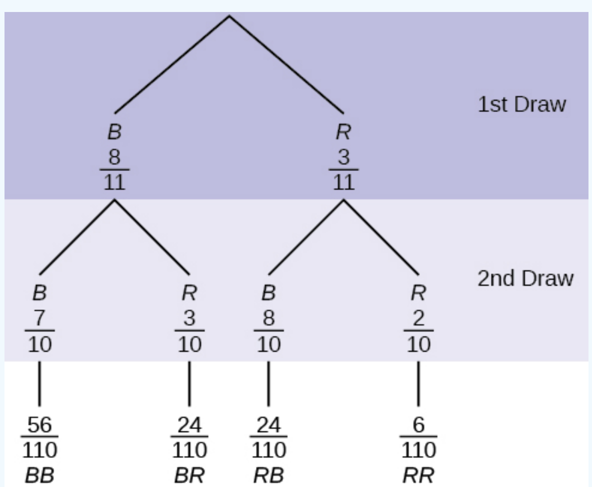


Figure 4.2.6.3: Total $= \frac{56+24+24+6}{110} = \frac{110}{110} = 1$

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw ***without replacement****, so that on the second draw there are ten marbles left in the urn.*

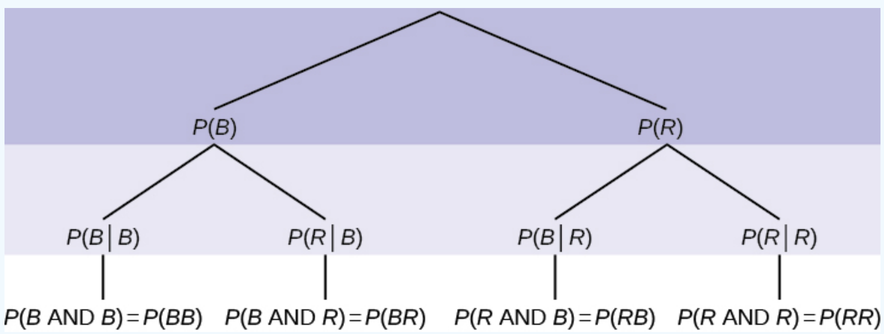
Calculate the following probabilities using the tree diagram.

a. $P(\text{RR}) =$ ________
b. Fill in the blanks: $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) +$ (___)(___) $= \frac{48}{110}$
c. $P(\text{R on 2nd}|\text{B on 1st}) =$
d. Fill in the blanks: $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) =$ (___)(___) $= \frac{24}{110}$
e. Find $P(\text{BB})$.
f. Find $P(\text{B on 2nd}|\text{R on 1st})$.

Answers

a. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$
b. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$
c. $P(\text{R on 2nd}|\text{B on 1st}) = \frac{3}{10}$
d. $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$
e. $P(\text{BB}) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$
f. Using the tree diagram, $P(\text{B on 2nd}|\text{R on 1st}) = P(\text{R}|\text{B}) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.



- $P(\text{R}|\text{R})$ here means $P(\text{R on 2nd}|\text{R on 1st})$
- $P(\text{B}|\text{R})$ here means $P(\text{B on 2nd}|\text{R on 1st})$
- $P(\text{R}|\text{B})$ here means $P(\text{R on 2nd}|\text{B on 1st})$
- $P(\text{B}|\text{B})$ here means $P(\text{B on 2nd}|\text{B on 1st})$

? Exercise 4.2.6.2

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

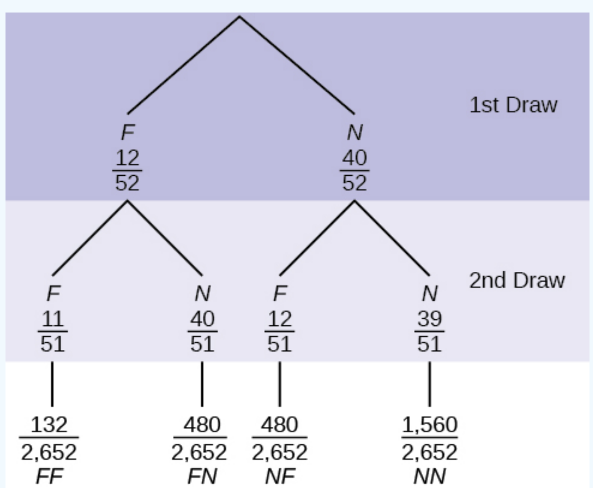


Figure 4.2.6.4:

a. Find $P(\text{FN OR NF})$.

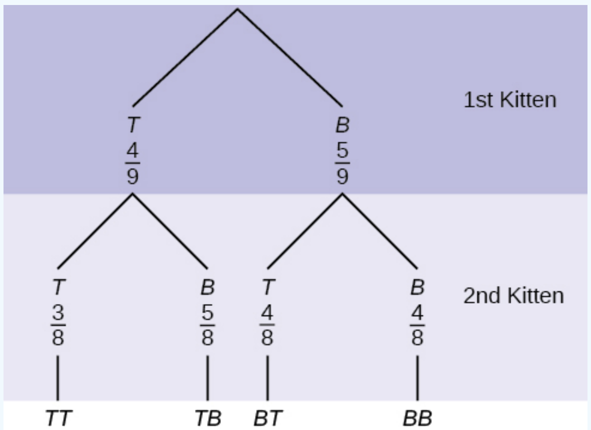
b. Find $P(\text{N}|\text{F})$.
c. Find $P(\text{at most one face card})$.
Hint: "At most one face card" means zero or one face card.
d. Find $P(\text{at least on face card})$.
Hint: "At least one face card" means one or two face cards.

Answer

a. $P(\text{FN OR NF}) = \frac{480}{2,652} + \frac{480}{2,652} = \frac{960}{2,652} = \frac{80}{221}$
b. $P(\text{N}|\text{F}) = \frac{40}{51}$
c. $P(\text{at most one face card}) = \frac{(480+480+1,560)}{2,652} = \frac{2,520}{2,652}$
d. $P(\text{at least one face card}) = \frac{(132+480+480)}{2,652} = \frac{1,092}{2,652}$

Example 4.2.6.3

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



a. What is the probability that both kittens are tabby?
a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
b. What is the probability that one kitten of each coloring is selected?
a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$
c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
d. What is the probability of choosing two kittens of the same color?

Answer

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Exercise 4.2.6.3

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Answer

$\left(\frac{4}{7}\right)\left(\frac{3}{6}\right) + \left(\frac{3}{7}\right)\left(\frac{4}{6}\right)$

Venn Diagram

A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example 4.2.6.4

Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event $\text{A} =$ $\{1, 2, 3, 4, 5, 6\}$ and event $\text{B} = \{6, 7, 8, 9\}$. Then $\text{A AND B} = \{6\}$ and $\text{A OR B} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The Venn diagram is as follows:

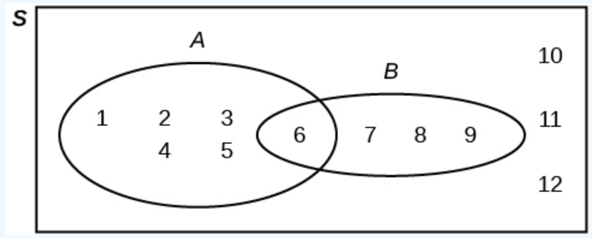


Figure 4.2.6.5:

Exercise 4.2.6.4

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event $\text{C} = \{\text{green, blue, purple}\}$ and event $\text{P} = \{\text{red, yellow, blue}\}$. Then $\text{C AND P} = \{\text{blue}\}$ and $\text{C OR P} = \{\text{green, blue, purple, red, yellow}\}$. Draw a Venn diagram representing this situation.

Answer

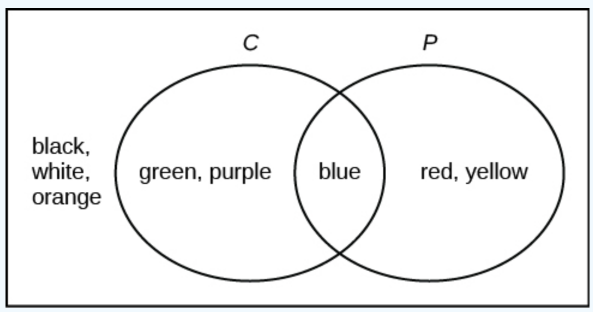


Figure 4.2.6.6:

Example 4.2.6.5

Flip two fair coins. Let $\text{A} =$ tails on the first coin. Let $\text{B} =$ tails on the second coin. Then $\text{A} = \{TT, TH\}$ and $\text{B} = \{TT, HT\}$. Therefore, $\text{A AND B} = \{TT\}$. $\text{A OR B} = \{TH, TT, HT\}$.

The sample space when you flip two fair coins is $X = \{HH, HT, TH, TT\}$. The outcome *HH* is in NEITHER A NOR B. The Venn diagram is as follows:

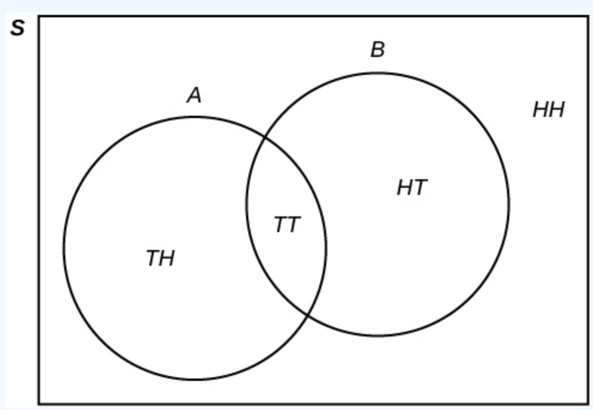
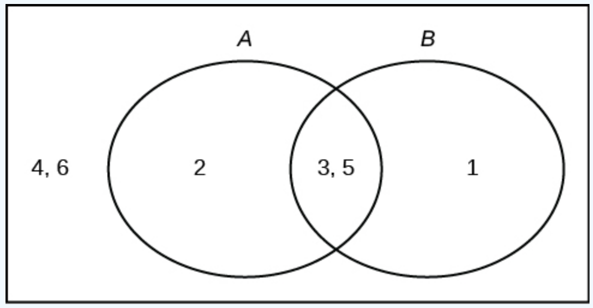


Figure 4.2.6.7:

Exercise $4.2.6.5$

Roll a fair, six-sided die. Let $\mathrm{A} =$ a prime number of dots is rolled. Let $\mathrm{B} =$ an odd number of dots is rolled. Then $\mathrm{A} = \{2, 3, 5\}$ and $\mathrm{B} = \{1, 3, 5\}$. Therefore, $\mathrm{A\ AND\ B} = \{3, 5\}$. $\mathrm{A\ OR\ B} = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $\mathrm{S} = \{1, 2, 3, 4, 5, 6\}$. Draw a Venn diagram representing this situation.

Answer



Example $4.2.6.6$: Probability and Venn Diagrams

Forty percent of the students at a local college belong to a club and 50% work part time. Five percent of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let $\mathrm{C} =$ student belongs to a club and $\mathrm{PT} =$ student works part time.

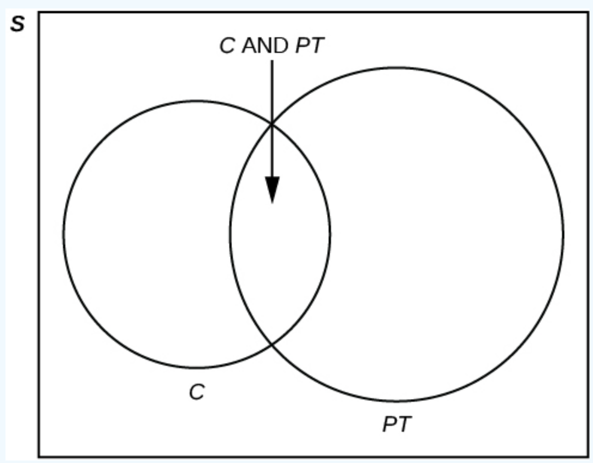


Figure $4.2.6.8$:

If a student is selected at random, find

- the probability that the student belongs to a club. $P(\mathrm{C}) = 0.40$
- the probability that the student works part time. $P(\mathrm{PT}) = 0.50$
- the probability that the student belongs to a club AND works part time. $P(\mathrm{C\ AND\ PT}) = 0.05$
- the probability that the student belongs to a club **given** that the student works part time. $P(\mathrm{C}|\mathrm{PT}) = \frac{P(\mathrm{C\ AND\ PT})}{P(\mathrm{PT})} = \frac{0.05}{0.50} = 0.1$
- the probability that the student belongs to a club **OR** works part time. $P(\mathrm{C\ OR\ PT}) = P(\mathrm{C}) + P(\mathrm{PT}) - P(\mathrm{C\ AND\ PT}) = 0.40 + 0.50 - 0.05 = 0.85$

Exercise $4.2.6.6$

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let $\mathrm{W} =$ works a second job and $\mathrm{S} =$ spouse also works.

Answer

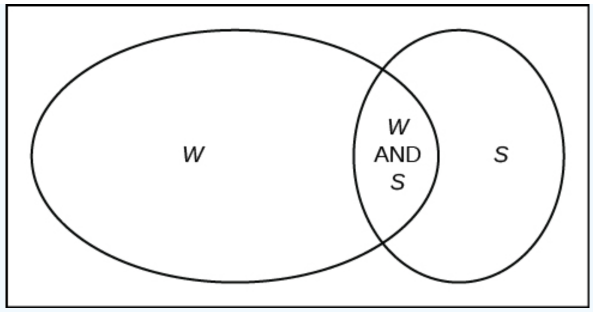


Figure 4.2.6.9:

✔ Example 4.2.6.7

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5−10% of African Americans have the Rh- factor, and 51% have type O blood.

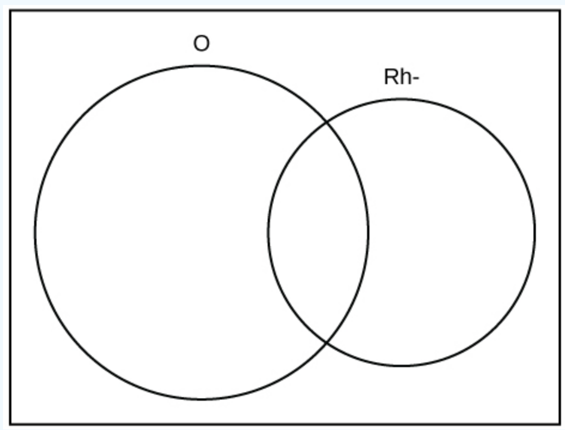


Figure 4.2.6.10:

The "O" circle represents the African Americans with type O blood. The "Rh-" oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let $\mathrm{O} =$ African American with Type O blood and $\mathrm{R} =$ African American with Rh- factor.

a. $P(\mathrm{O}) =$ ___________
b. $P(\mathrm{R}) =$ ___________
c. $P(\mathrm{O\ AND\ R}) =$ ___________
d. $P(\mathrm{O\ OR\ R}) =$ ____________
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Answer

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

? Exercise 4.2.6.7

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or anon-fiction book.
c. In the Venn diagram, describe the overlapping area using a complete sentence.

d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Answer

a. and d. In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.

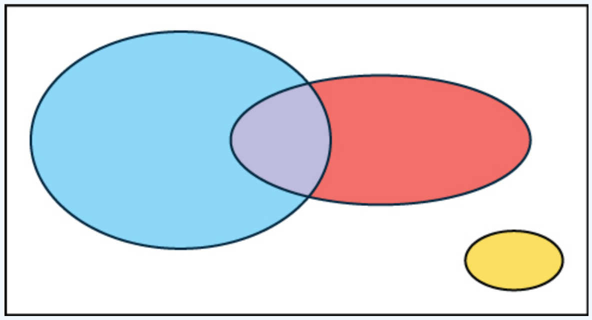


Figure 4.2.6.11:

b. $P(\text{novel or non-fiction}) = P(\text{Blue OR Red}) = P(\text{Blue}) + P(\text{Red}) - P(\text{Blue AND Red}) = 0.6 + 0.4 - 0.2 = 0.8$.

c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

1. Data from Clara County Public H.D.
2. Data from the American Cancer Society.
3. Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).
4. Data from the Federal Highway Administration, part of the United States Department of Transportation.
5. Data from the United States Census Bureau, part of the United States Department of Commerce.
6. Data from USA Today.
7. “Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).
8. “Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

Review

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize. A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space *S* together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

Glossary

Tree Diagram

the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram

the visual representation of a sample space and events in the form of circles or ovals showing their intersections

4.2.7: Probability Topics (Worksheet)

Name: ______________________________

Section: _____________________________

Student ID#:__________________________

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will use theoretical and empirical methods to estimate probabilities.
- The student will appraise the differences between the two estimates.
- The student will demonstrate an understanding of long-term relative frequencies.

Do the Experiment

Count out 40 mixed-color M&Ms® which is approximately one small bag's worth. Record the number of each color in Table. Use the information from this table to complete Table. Next, put the M&Ms in a cup. The experiment is to pick two M&Ms, one at a time. Do **not** look at them as you pick them. The first time through, replace the first M&M before picking the second one. Record the results in the "With Replacement" column of Table. Do this 24 times. The second time through, after picking the first M&M, do **not** replace it before picking the second one. Then, pick the second one. Record the results in the "Without Replacement" column section of Table. After you record the pick, put **both** M&Ms back. Do this a total of 24 times, also. Use the data from Table to calculate the empirical probability questions. Leave your answers in unreduced fractional form. Do **not** multiply out any fractions.

Population

Color	Quantity
Yellow (Y)	
Green (G)	
Blue (BL)	
Brown (B)	
Orange (O)	
Red (R)	

Theoretical Probabilities

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2 \text{ OR } B_1R_2)$		
$P(R_1 \text{ AND } G_2)$		
$P(G_2\|R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

G_2 = green on second pick; R_1 = red on first pick; B_1 = brown on first pick; B_2 = brown on second pick; doubles = both picks are the same colour.

Empirical Results

With Replacement	Without Replacement
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)

Empirical Probabilities

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2 \text{ OR } B_1R_2)$		
$P(R_1 \text{ AND } G_2)$		
$P(G_2 \mid R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Discussion Questions

a. Why are the “With Replacement” and “Without Replacement” probabilities different?
b. Convert P(no yellows) to decimal format for both Theoretical “With Replacement” and for Empirical “With Replacement”. Round to four decimal places.
   1. Theoretical “With Replacement”: P(no yellows) = _______
   2. Empirical “With Replacement”: P(no yellows) = _______
   3. Are the decimal values “close”? Did you expect them to be closer together or farther apart? Why?
c. If you increased the number of times you picked two M&Ms to 240 times, why would empirical probability values change?
d. Would this change (see part 3) cause the empirical probabilities and theoretical probabilities to be closer together or farther apart? How do you know?
e. Explain the differences in what $P(G_1 \text{ AND } R_2)$ and $P(R_1 \mid G_2)$ represent. Hint: Think about the sample space for each probability.

4.2.E: Probability Topics (Exericses)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

3.1: Introduction

3.2: Terminology

Q 3.2.1

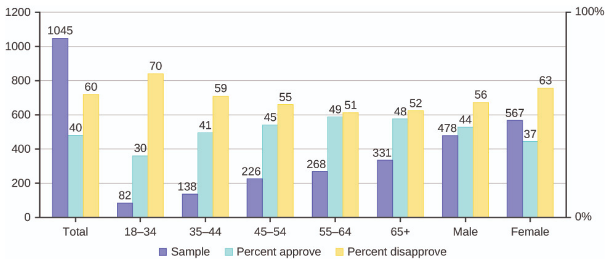


Figure 3.2.3.2.11.

The graph in Figure 3.2.1 displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

Q 3.2.2

Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

S 3.2.2

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

3.3: Independent and Mutually Exclusive Events

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.

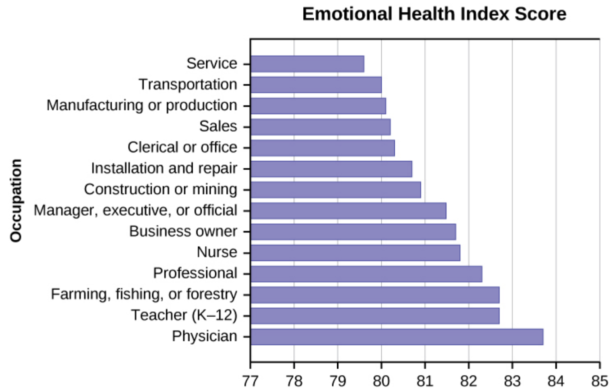


Figure 3.3.1.

Q 3.3.1

Find the probability that an Emotional Health Index Score is 82.7.

Q 3.3.2

Find the probability that an Emotional Health Index Score is 81.0.

S 3.3.2

0

Q 3.3.3

Find the probability that an Emotional Health Index Score is more than 81?

Q 3.3.4

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

S 3.3.4

0.3571

Q 3.3.5

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

Q 3.3.6

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

S 3.3.6

0.2142

Q 3.3.7

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

Q 3.3.8

What occupation has the highest emotional index score?

S 3.3.8

Physician (83.7)

Q 3.3.9

What occupation has the lowest emotional index score?

Q 3.3.10

What is the range of the data?

S 3.3.10

83.7 − 79.6 = 4.1

Q 3.3.11

Compute the average EHIS.

Q 3.3.12

If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

S 3.3.12

$P(\text{Occupation} < 81.3) = 0.5$

3.4: Two Basic Rules of Probability

Q 3.4.1

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- $\text{C} =$ California registered voters who support same-sex marriage.
- $\text{B} =$ California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- $\text{A} =$ California registered voters who are 18 to 39 years old.

a. Find $P(\text{C})$.
b. Find $P(\text{B})$.
c. Find $P(\text{C}|\text{A})$.
d. Find $P(\text{B}|\text{C})$.
e. In words, what is $\text{C}|\text{A}$?
f. In words, what is $\text{B}|\text{C}$?
g. Find $P(\text{C AND B})$.
h. In words, what is C AND B?
i. Find $P(\text{C OR B})$.
j. Are C and B mutually exclusive events? Show why or why not.

Q 3.4.2

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?

e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

S 3.4.2

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.

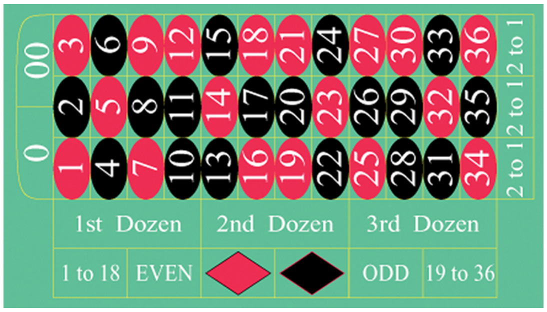


Figure 3.4.1

Q 3.4.3

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find $P(\text{red})$.
c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
d. You bet on an even number. Find $P(\text{even number})$.
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Q 3.4.4

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

S 3.4.4

1. $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$
2. $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$
3. $P(\text{Betting on one number}) = \frac{1}{38}$
4. $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$
5. $P(\text{Betting on two number that touch each other on the table}) = \frac{2}{38}$
6. $P(\text{Betting on 0-00-1-2-3}) = \frac{5}{38}$
7. $P(\text{Betting on 0-1-2; or 0-00-2; or 00-2-3}) = \frac{3}{38}$

Q 3.4.5

Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

Q 3.4.6

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- $\mathrm{G} =$ card drawn is green
- $\mathrm{E} =$ card drawn is even-numbered

  a. List the sample space.
  b. $P(\mathrm{G}) =$ _____
  c. $P(\mathrm{G}|\mathrm{E}) =$ _____
  d. $P(\mathrm{G\ AND\ E}) =$ _____
  e. $P(\mathrm{G\ OR\ E}) =$ _____
  f. Are G and E mutually exclusive? Justify your answer numerically.

S 3.4.6

1. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
2. $\frac{5}{8}$
3. $\frac{2}{3}$
4. $\frac{2}{8}$
5. $\frac{}{6}8$
6. No, because $P(\mathrm{G\ AND\ E})$ does not equal 0.

Q 3.4.7

Roll two fair dice. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(\mathrm{A})$.
c. Let *B* be the event that the sum of the two rolls is at most seven. Find $P(\mathrm{B})$.
d. In words, explain what "$P(\mathrm{A}|\mathrm{B})$" represents. Find $P(\mathrm{A}|\mathrm{B})$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.8

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(\mathrm{A})$.
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

S 3.4.9

The coin toss is independent of the card picked first.

a. $\{(G, H)(G, T)(B, H)(B, T)(R, H)(R, T)\}$
b. $P(\mathrm{A}) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(\mathrm{A\ AND\ B}) = 0$
d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(\mathrm{A\ AND\ C}) = P(\mathrm{A}) = 320$

Q 3.4.10

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(\mathrm{A})$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.11

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(\mathrm{A})$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

S 3.4.12

a. $S = (\mathrm{HHH}), (\mathrm{HHT}), (\mathrm{HTH}), (\mathrm{HTT}), (\mathrm{THH}), (\mathrm{THT}), (\mathrm{TTH}), (\mathrm{TTT})$
b. $\frac{4}{8}$
c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(\mathrm{A\ AND\ B}) = 0$.

Q 3.4.13

Consider the following scenario:

Let $P(\mathrm{C}) = 0.4$.

Let $P(\mathrm{D}) = 0.5$.

Let $P(\mathrm{C}|\mathrm{D}) = 0.6$.

a. Find $P(C AND D)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(\mathrm{C\ OR\ D})$.
e. Find $P(\mathrm{D}|\mathrm{C})$.

Q 3.4.14

Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(\mathrm{Y\ OR\ Z}) = P(\mathrm{Y}) + P(\mathrm{Z}) - P(\mathrm{Y\ AND\ Z})$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(\mathrm{Z})$ if $P(\mathrm{Y\ OR\ Z}) = 0.71$ and $P(\mathrm{Y}) = 0.42$.

S 3.4.14

a. If Y and Z are independent, then $P(\mathrm{Y\ AND\ Z}) = P(\mathrm{Y})P(\mathrm{Z})$, so $P(\mathrm{Y\ OR\ Z}) = P(\mathrm{Y}) + P(\mathrm{Z}) - P(\mathrm{Y})P(\mathrm{Z})$.
b. 0.5

Q 3.4.15

G and H are mutually exclusive events. $P(\mathrm{G})=0.5 P(\mathrm{H})=0.3$

a. Explain why the following statement MUST be false: $P(\mathrm{H} \mid \mathrm{G})=0.4$.
b. Find $P(\mathrm{H} \text{ OR } \mathrm{G})$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Q 3.4.16

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: $\mathrm{E}=$ speaks English at home; $\mathrm{E}'=$ speaks another language at home; $\mathrm{S}=$ speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(\mathrm{E}')=$	i. 0.8043
b. $P(\mathrm{E})=$	ii. 0.623
c. $P(\mathrm{S} \text{ and } \mathrm{E}')=$	iii. 0.1957
d. $P(\mathrm{S} \mid \mathrm{E}')=$	iv. 0.1219

S 3.4.16

a. iii
b. i
c. iv
d. ii

Q 3.4.17

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let $G=$ won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let $\mathrm{F}=$ was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Q 3.4.18

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with \$10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: $\mathrm{R}=$ money returned; $\mathrm{E}=$ economics classes; $\mathrm{O}=$ other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.
c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

S 3.4.18

a. $P(\text{R}) = 0.44$
b. $P(\text{R}|\text{E}) = 0.56$
c. $P(\text{R}|\text{O}) = 0.31$
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(\text{R}|\text{E}) \neq P(\text{R})$.
e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(\text{R}|\text{E}) > P(\text{R})$.

Q 3.4.19

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) = P(\text{hit is a double})$

Q 3.4.29

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

S 3.4.30

a. $P(\text{type O OR Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O AND Rh-})$

   $0.52 = 0.43 + 0.15 - P(\text{type O AND Rh-})$; solve to find $P(\text{type O AND Rh-}) = 0.06$

   6% of people have type O, Rh- blood
b. $P(\text{NOT}(\text{type O AND Rh-})) = 1 - P(\text{type O AND Rh-}) = 1 - 0.06 = 0.94$

   94% of people do not have type O, Rh- blood

Q 3.4.31

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

1. Find the probability that a course has a final exam or a research project.
2. Find the probability that a course has NEITHER of these two requirements.

Q 3.4.32

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

1. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
2. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

S 3.4.32

a. Let $C =$ be the event that the cookie contains chocolate. Let $N =$ the event that the cookie contains nuts.
b. $P(\text{C OR N}) = P(\text{C}) + P(\text{N}) - P(\text{C AND N}) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(\text{C OR N}) = 1 - 0.40 = 0.60$

Q 3.4.33

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let $\text{D} =$ event that a student takes a distance learning class and $\text{E} =$ event that a student is a part time student

a. Find $P(\text{D AND E})$.
b. Find $P(\text{E}|\text{D})$.
c. Find $P(\text{D OR E})$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

3.5: Contingency Tables

Use the information in the Table to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Q 3.5.1

What is the probability that a randomly selected senator has an “Other” affiliation?

S 3.5.1

0

Q 3.5.2

What is the probability that a randomly selected senator is up for reelection in November 2016?

Q 3.5.3

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

S 3.5.3

$\frac{10}{67}$

Q 3.5.4

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Q 3.5.5

Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

S 3.5.5

$\frac{10}{34}$

Q 3.5.6

Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

Q 3.5.7

The events “Republican” and “Up for reelection in 2016” are ________

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

S 3.5.7

d

Q 3.5.8

The events “Other” and “Up for reelection in November 2016” are ________

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Q 3.5.9

This table gives the number of participants in the recent National Health Interview Survey who had been treated for cancer in the previous 12 months. The results are sorted by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex.

Race and Sex	15-24	25-40	41-65	over 65	TOTALS
white, male	1,165	2,036	3,703		8,395
white, female	1,076	2,242	4,060		9,129
black, male	142	194	384		824
black, female	131	290	486		1,061
all others					
TOTALS	2,792	5,279	9,354		21,081

Do not include "all others" for parts f and g.

a. Fill in the column for cancer treatment for individuals over age 65.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a white male.
d. Find the probability that a randomly selected individual was a black female.
e. Find the probability that a randomly selected individual was black
f. Find the probability that a randomly selected individual was a black or white male.
g. Out of the individuals over age 65, find the probability that a randomly selected individual was a black or white male.

S 3.5.9

a.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

b.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

c. $\frac{22{,}050}{29{,}760}$
d. $\frac{330}{29{,}760}$
e. $\frac{2{,}000}{29{,}760}$
f. $\frac{23{,}720}{29{,}760}$
g. $\frac{5{,}010}{6{,}020}$

Use the following information to answer the next two exercises. The table of data obtained from*www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Q 3.5.10

Find $P(\text{hit was made by Babe Ruth})$.

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Q 3.5.11

Find $P(\text{hit was made by Ty Cobb}|\text{The hit was a Home Run})$.

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

S 3.5.11

b

Q 3.5.12

Table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If B is the event of a child having brown hair, find the probability of the complement of B.
g. In words, what does the complement of B represent?

Q 3.5.13

In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data were compiled into the following table.

Shirt#	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

S 3.5.13

a. $\frac{26}{106}$
b. $\frac{33}{106}$
c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right)+\left(\frac{33}{106}\right)-\left(\frac{21}{106}\right)=\left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

Exercise 3.6.8

The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: $\mathrm{C}=$ a man develops cancer in his lifetime; $\mathrm{P}=$ man has at least one false positive. Construct a tree diagram of the situation.

Answer

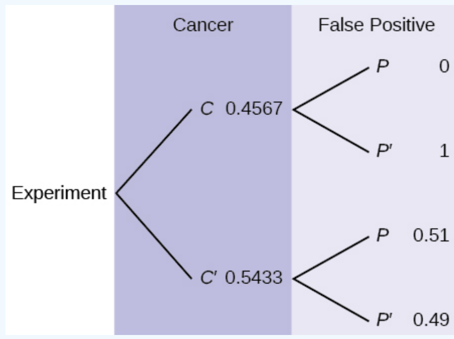


Figure 3.6.12

Bring It Together

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

Exercise 3.6.9

Suppose that you randomly draw two cards, one at a time, **with replacement**.

Let $\mathrm{G}_1=$ first card is green
Let $\mathrm{G}_2=$ second card is green

a. Draw a tree diagram of the situation.
b. Find $P(\mathrm{G}_1 \mathrm{AND} \mathrm{G}_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(\mathrm{G}_2 | \mathrm{G}_1)$.
e. Are G_1 and G_2 independent events? Explain why or why not.

Answer

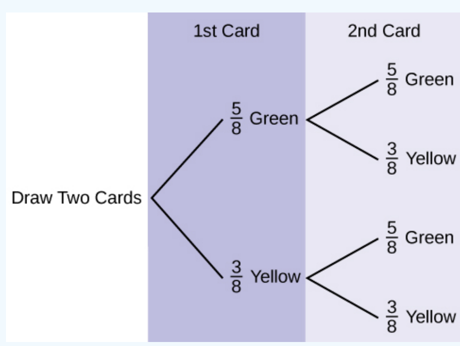
a. 

Figure 3.6.14

b. $P(\mathrm{GG})=\left(\frac{5}{8}\right)\left(\frac{5}{8}\right)=\frac{25}{64}$
c. $P(\text{at least one green})=P(\mathrm{GG})+P(\mathrm{GY})+P(\mathrm{YG})=\frac{25}{64}+\frac{15}{64}+\frac{15}{64}=\frac{55}{64}$
d. $P(\mathrm{G}|\mathrm{G})=\frac{5}{8}$
e. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Exercise 3.6.10

Suppose that you randomly draw two cards, one at a time, **without replacement**.

$\mathrm{G}_1=$ first card is green
$\mathrm{G}_2=$ second card is green

a. Draw a tree diagram of the situation.
b. Find $P(\mathrm{G}_1 \text{AND } \mathrm{G}_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(\mathrm{G}_2|\mathrm{G}_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

Exercise 3.6.11

Complete the following.

a. Construct a table or a tree diagram of the situation.
b. Find $P(\text{driver is female})$.
c. Find $P(\text{driver is age 65 or over}|\text{driver is female})$.
d. Find $P(\text{driver is age 65 or over AND female})$.
e. In words, explain the difference between the probabilities in part c and part d.
f. Find $P(\text{driver is age 65 or over})$.
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Answer

a.

	<20	**20–64**	**>64**	**Totals**
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

b. $P(\mathrm{F})=0.486$
c. $P(>64|\mathrm{F})=0.1361$
d. $P(>64 \text{ and } \mathrm{F})=P(\mathrm{F})P(>64|\mathrm{F})=(0.486)(0.1361)=0.0661$
e. $P(>64|\mathrm{F})$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } \mathrm{F})$ is the percentage of drivers who are female and 65 or older.
f. $P(>64)=P(>64 \text{ and } \mathrm{F})+P(>64 \text{ and } \mathrm{M})=0.1356$
g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } \mathrm{F})=0.0661$.

Exercise 3.6.12

Suppose that 10,000 U.S. licensed drivers are randomly selected.

a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

Exercise 3.6.13

Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Answer

a.

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

b. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$
c. Make the same assumptions as in (b) we have: (0.7708)(1,000) = 771
d. (0.1332)(1,000) = 133

Exercise 3.6.14

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

a. Based on the given data, find $P(\text{H})$ and $P(\text{T})$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

Exercise 3.6.15

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person has a risk factor heterosexual contact})$.
c. Find $P(\text{Person is female OR has a risk factor of IV drug user})$.
d. Find $P(\text{Person is female AND has a risk factor of homosexual/bisexual})$.
e. Find $P(\text{Person is male AND has a risk factor of IV drug user})$.

f. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$.
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

Answer

The completed contingency table is as follows:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	**IV Drug User***	**Heterosexual Contact**	**Other**	**Totals**
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

a. $\frac{255}{2059}$
b. $\frac{196}{3059}$
c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$
g.

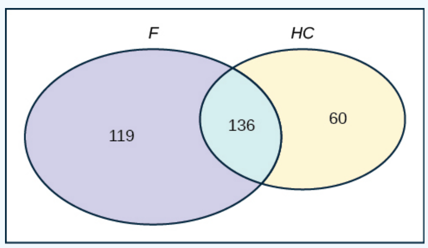


Figure 3.6.15

Exercise 3.6.16

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person obtained the disease through heterosexual contact})$.
c. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (R), four yellow (Y) and five blue (B) beads. For the coin, $P(\text{H}) = \frac{2}{3}$ and $P(\text{T}) = \frac{1}{3}$ where H is heads and T is tails.

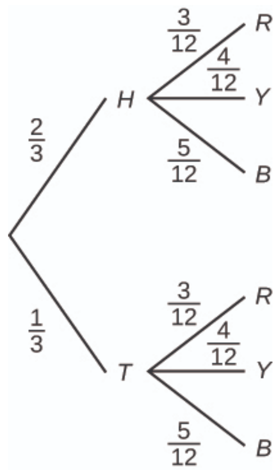


Figure 3.6.1.

Q 3.6.1

Find $P(\text{tossing a Head on the coin AND a Red bead})$

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Q 3.6.2

Find $P(\text{Blue bead})$.

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

S 3.6.2

a

Q 3.6.3

A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find $P(\mathrm{S})$.
e. Let T be the event that the cookies selected were different flavors. Find $P(\mathrm{T})$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find $P(\mathrm{U})$.

3.7: Probability Topics

4.3: PowerPoints

https://professormo.com/holistic/Powerpoint/ch04.pptx

CHAPTER OVERVIEW

5: Discrete Probability

5.1: Videos

Defining probability - Core concepts, explained in detail

Probability trees - Useful tool for conditional probability

Would you take this bet? - Thinking through probability and risk

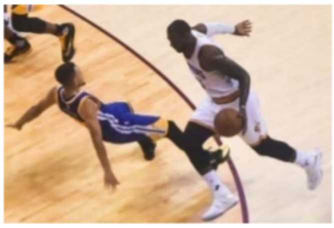
SECTION OVERVIEW

5.2: Probability

In the prior three sections we covered how to obtain and analyze sample data. In the next three sections, we will explore the modeling of populations.

5.2.1: What is Probability?

Rather than defining probability, here are some real life examples:



The Golden State Warriors are trailing the Cleveland Cavaliers by one point late in an important NBA game. Cleveland forward LeBron James fouls Golden State guard Stephen Curry with 1.4 seconds left in the game, meaning Curry will get to shoot 2 free throws. What is the probability the Warriors will win the game?

Thuy is an actress and auditions for a starring role in a Broadway musical. The audition goes extremely well and the director says she did a great job, sings beautifully, and is perfect for the role. He promises to call her back the next day after auditions are completed. What is the probability Thuy will get the role in the musical?

Robert is a student taking a Statistics class for the second time, after dropping the class in the prior quarter. He has a lot of math anxiety, but needs to pass the class to be able to transfer to San Jose State University to continue his dream of becoming a psychologist. What is the probability he will successfully pass the class?

Lupe goes to the doctor after having some pain in her lower back. Her family has a history of kidney problems, so the doctor decides to run some additional tests. What is the probability that Lupe has a kidney disorder that requires treatment?

In all of these examples, it is uncertain or unknown what the actual outcomes will be; however, we can make a guess as to whether each **outcome** is either more likely or less likely. We can quantify this by a value between 0 and 1, or between 0% and 100%. For example, maybe we say The Warriors have a good chance of winning the game since Curry is one of the best free throw shooters in the NBA, say 0.7 or 70% . Maybe Thuy (from her experience in auditioning) is less likely to get the starring role, say 0.2 or 20%. These quantities are called **probabilities**.

Definition: Probability

Probability is the measure of the **likelihood** that an **event A** will occur.

This measure is a quantity between 0 (never) and 1 (always) and will be expressed as **P(A)** (read as “The probability event A occurs.”)

5.2.2: Types of Probability

Classical probability (also called Mathematical Probability) is determined by counting or by using a mathematical formula or model.

Example

The probability of getting a "Heads" when tossing a fair coin is 0.5 or 50%. The probability of rolling a 5 on a fair six-sided die is 1/6, since all numbers are equally likely.

Empirical probability is based on the relative frequencies of historical data, studies or experiments.

Example

The probability that Stephen Curry make a free throw is 90.8% based on the frequency of successes from all prior free throws.

The probability of a random student getting an A in a Statistics class taught by Professor Nguyen is 22.8%, because grade records show that of the 1000 students who took her class in the past, 228 received an A.

In a study of 832 adults with colon cancer, an experimental drug reduced tumors in 131 patients. The probability that the experimental drug reduces colon cancer tumors is 131/832, or 15.7%.

Subjective probability is a “one-shot” educated guess based on anecdotal stories, intuition or a feeling as to whether an event is likely, unlikely or “50-50”. Subjective probability is often inaccurate.

Example

Although Robert is nervous about retaking the Statistics course after dropping the prior quarter, he is 90% sure he will pass the class because the website ratemyprofessor.com gave the instructor very positive reviews.

Jasmine believes that she will probably not like a new movie that is coming out soon because she is not a fan of the actor who is starring in the film. She is about 20% sure she will like the new movie.

No matter how probability is initially derived, the laws and rules of probability will be treated the same.

5.2.3: How to Calculate Classical Probability

We can use counting methods to determine classical probability. However, we need to be careful in our methods to be sure to get the correct answer.

An **Event** is a result of an experiment, usually referred to with a capital letter A, B, C, etc. Consider the experiment of flipping two coins. Then use the letter A to refer to the event of getting exactly one head.

An **Outcome** is a result of the experiment that cannot be broken down into smaller events. Consider event A, getting exactly one head. Note that there are two ways or outcomes to get one head in two tosses, by first getting a head then a tail, or by first getting a tail, then a head. Let's write these distinct outcomes as HT and TH.

The **Sample Space** is the set of all possible outcomes of an experiment. In the experiment of flipping two coins, there are 4 possible outcomes, which can be expressed in set notation.

$$\text{Sample Space } = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$$

We can now redefine an **Event** of an experiment to be a subset of the Sample Space. If event A is getting exactly one head in two coin tosses, then

$$\mathrm{A} = \{\mathrm{HT}, \mathrm{TH}\}$$

After carefully listing the outcomes of the Sample Space and the outcomes of the event, we can then calculate the **probability** the event occurs.

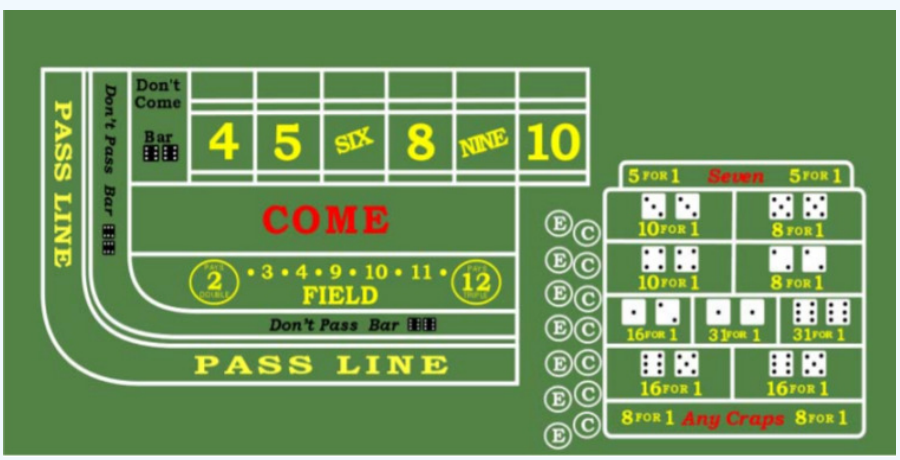
Probability Event Occurs = number of outcomes in Event / number of outcomes in Sample Space

We will use the notation P(A) to mean the probability event A occurs.

In the example, the probability of getting exactly 1 head in two coin tosses is 2 out of 4 or 50%.

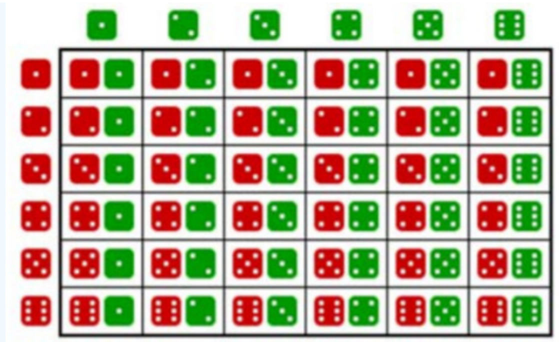
$$P(A) = 2/4 = 0.5 = 50\%$$

Example: Field Bet



In the casino game of craps, two dice are rolled at the same time and then the resulting two numbers are totaled. There are many bets in craps, so let us consider the Field bet. In this bet, the player will win even money if a total of 3, 4, 9, 10 or 11 is rolled. If a total of 2 is rolled, the player will win double the original bet, and if a total of 12 is rolled, the player will win triple the original bet. If a total of 5, 6, 7 or 8 is rolled, the player loses the original bet.

At first glance, this looks like a winning bet for the player since the player wins on 7 different numbers and the casino only wins on 4 different numbers. However, we know that a casino always designs games to give the casino the advantage. Let us carefully use counting methods to calculate the probability of a player winning the Field bet.



Let's first consider the task of listing the sample space of possible outcomes. Since there are two dice rolled, we can consider each outcome to be an ordered pair. There are 6 possible values for the first die and 6 possible values for the second die, meaning that there are 36 ordered pairs or outcomes. In the diagram, the red die is the first roll and the green die is the second roll.

$$\text{Sample Space } = \left\{\begin{array}{l} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{array}\right\}$$

Now define the event W to be the winning pairs of numbers in the Field bet, the pairs that add up to 2, 3, 4, 9, 10, 11 or 12. The winning pairs of numbers are shown in blue and the losing pairs are shown in red.

$$\text{Sample Space } = \left\{\begin{array}{l} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{array}\right\} \quad W = \left\{\begin{array}{l} (1,1),(1,2),(1,3), \\ (2,1),(2,2), \\ (3,1),(3,6), \\ (4,5),(4,6), \\ (5,4),(5,5),(5,6), \\ (6,3),(6,4),(6,5),(6,6) \end{array}\right\}$$

This means that there are 16 outcomes out of 36 in which the player wins. It's now easy to see the that probability of winning is less than 50%, as the casino took the numbers that occur the most frequently.

$$P(W)=\frac{16}{36}=\frac{4}{9} \approx 44.4\%$$

As a final note on this example, you might recall that the casino pays double if the player rolls (1,1) or triple if the player rolls (6,6). Even taking this extra bonus into account, if a player makes 36 $100 bets, the casino will expect to win $2000 (20 numbers x $100), and the player will expect to win $1900 (16 numbers x $100, plus $100 extra for the 2 and $200 extra for the 12), meaning the player loses $100 for every $3600 bet, a house (casino) advantage of 2.78%.

Field Bet – Summary of 36 possible rolls	Amount won on $100 bets
$(1,1)$ (pays double)	+$200
$(6,6)$ (pays triple)	+$300
$(1,2),(1,3),(2,1),(2,2),(3,1),(3,6),(4,5),(4,6),(5,4),(5,5),(5,6),(6,3),(6,4),(6,5)$	+$1400
$(1,4),(1,5),(1,6),(2,3),(2,4),(2,5),(2,6),(3,2),(3,3),(3,4),(3,5),(4,1),(4,2),(4,3),(4,4),(5,1),(5,2),(5,3),(6,1),(6,2)$	-$2000
Overall expected result of 36 rolls ($3600 bet)	-$100

Just remember, in the long run, the casino always wins.

5.3: PowerPoints

https://professormo.com/holistic/Powerpoint/ch05.pptx

CHAPTER OVERVIEW

6: Binomial Probability Distribution

6.1: Videos

Defining probability - Core concepts, explained in detail

Probability trees - Useful tool for conditional probability

Would you take this bet? - Thinking through probability and risk

Normal distribution - Core concepts and several examples

Binomial distribution - Introduction to the binomial distribution

Normal approximation to binomial - A useful technique for some binomial situations

6.2: Mean or Expected Value and Standard Deviation

The expected value is often referred to as the "long-term" average or mean. This means that over the long term of doing an experiment over and over, you would expect this average.

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. As you learned in Chapter 3, probability does not describe the short-term results of an experiment. It gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (the theoretical probability and the relative frequency get closer and closer together). When evaluating the long-term results of statistical experiments, we often want to know the "average" outcome. This "long-term average" is known as the mean or expected value of the experiment and is denoted by the Greek letter μ. In other words, after conducting many trials of an experiment, you would expect this average value.

> To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

Example 6.2.1

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the men's soccer team plays soccer.

Solution

To do the problem, first let the random variable $X =$ the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table adding a column $x * P(x)$. In this column, you will multiply each x value by its probability.

Expected Value Table This table is called an expected value table. The table helps you calculate the expected value or long-term average.

x	$P(x)$	$x * P(x)$
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Add the last column $x * P(x)$ to find the long term average or expected value:

$$(0)(0.2)+(1)(0.5)+(2)(0.3)=0+0.5+0.6=1.1.$$

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu = 1.1$.

Example 6.2.2

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

x	$P(x)$	$x * P(x)$	$(x – \mu)^2 \cdot P(x)$
0	$P(x=0)=\frac{2}{50}$	$(0)\left(\frac{2}{50}\right)=0$	$(0 – 2.1)^2 \cdot 0.04 = 0.1764$

x	$P(x)$	$x * P(x)$	$(x - \mu)^2 \cdot P(x)$
1	$P(x=1)=\frac{11}{50}$	$(1)\left(\frac{11}{50}\right)=\frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x=2)=\frac{23}{50}$	$(2)\left(\frac{23}{50}\right)=\frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x=3)=\frac{9}{50}$	$(3)\left(\frac{9}{50}\right)=\frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x=4)=\frac{4}{50}$	$(4)\left(\frac{4}{50}\right)=\frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x=5)=\frac{1}{50}$	$(5)\left(\frac{1}{50}\right)=\frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

Add the values in the third column of the table to find the expected value of X:

$$\mu = \text{Expected Value} = \frac{105}{50} = 2.1$$

Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x, multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$.

Add the values in the fourth column of the table:

$$0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05$$

The standard deviation of X is the square root of this sum: $\sigma = \sqrt{1.05} \approx 1.0247$

The mean, μ, of a discrete probability function is the expected value.

$$\mu = \sum (x \bullet P(x))$$

The standard deviation, Σ, of the PDF is the square root of the variance.

$$\sigma = \sqrt{\sum [(x - \mu)2 \bullet P(x)]}$$

When all outcomes in the probability distribution are equally likely, these formulas coincide with the mean and standard deviation of the set of possible outcomes.

? Exercise 6.2.2

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	$P(x)$
0	$P(x=0)=\frac{4}{50}$
1	$P(x=1)=\frac{8}{50}$
2	$P(x=2)=\frac{16}{50}$
3	$P(x=3)=\frac{14}{50}$
4	$P(x=4)=\frac{6}{50}$
5	$P(x=5)=\frac{2}{50}$

Answer

The expected value is 2.24

$$(0)\frac{4}{50}+(1)\frac{8}{50}+(2)\frac{16}{50}+(3)\frac{14}{50}+(4)\frac{6}{50}+(5)\frac{2}{50}=0+\frac{8}{50}+\frac{32}{50}+\frac{42}{50}+\frac{24}{50}+\frac{10}{50}=\frac{116}{50}=2.32 \tag{6.2.1}$$

Example 6.2.2

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay $2 to play and could profit $100,000 if you match all five numbers in order (you get your $2 back plus $100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let $X =$ the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and −2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

$$\begin{aligned}\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) &= (1)(10^{-5}) \\ &= 0.00001.\end{aligned}$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

$$1-0.00001=0.99999.1-0.00001=0.99999.$$

The expected value table is as follows:

Add the last column. –1.99998 + 1 = –0.99998

	x	$P(x)$	$xP(x)$
Loss	–2	0.99999	(–2)(0.99999) = –1.99998
Profit	100,000	0.00001	(100000)(0.00001) = 1

Since –0.99998 is about –1, you would, on average, expect to lose approximately $1 for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected LOSS per game after playing this game over and over.

Exercise 6.2.3

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit of playing the game over the long term?

Answer

Let $X =$ the amount of money you profit. The x-values are –$1 and $256.

The probability of guessing the right suit each time is $\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)=\frac{1}{256}=0.0039$

The probability of losing is $1-\frac{1}{256}=\frac{255}{256}=0.9961$

$(0.0039)256+(0.9961)(-1)=0.9984+(-0.9961)=0.0023$ or 0.23 cents.

Example 6.2.4

Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads}) = \frac{2}{3}$ and $P(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

a. Define a random variable X.
b. Complete the following expected value table.
c. What is the expected value, μ? Do you come out ahead?

Solutions

a.

X = amount of profit

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

b.

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

c.

Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Exercise 6.2.4

Suppose you play a game with a spinner. You play each game by spinning the spinner once. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. If you land on red, you pay $10. If you land on blue, you don't pay or win anything. If you land on green, you win $10. Complete the following expected value table.

	x	$P(x)$	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Answer

	x	$P(x)$	$x * P(x)$
Red	–10	$\frac{2}{5}$	$-\frac{20}{5}$
Blue	0	$\frac{2}{5}$	$\frac{0}{5}$

	x	$P(x)$	$x*P(x)$
Green	10	$\frac{1}{5}$	$\frac{1}{5}$

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root. To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x)–\mu^2P(x)$ and take the square root.

x	$P(x)$	$x*P(x)$	$(x–\mu)^2P(x)$
0	0.2	(0)(0.2) = 0	$(0 – 1.1)^2(0.2) = 0.242$
1	0.5	(1)(0.5) = 0.5	$(1 – 1.1)^2(0.5) = 0.005$
2	0.3	(2)(0.3) = 0.6	$(2 – 1.1)^2(0.3) = 0.243$

Add the last column in the table. $0.242+0.005+0.243=0.490$ The standard deviation is the square root of 0.49, or $\sigma=\sqrt{0.49}=0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas for calculating μ and σ.

Example 6.2.5

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like Table and calculate the mean μ and standard deviation σ of X.

Solution

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Use the sample space to complete the following table:

Calculating μ and σ.

x	$P(x)$	$xP(x)$	$(x–\mu)^2\cdot P(x)$
0	$\frac{9}{36}$	0	$(0–1)^2\cdot\frac{9}{36}=\frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1–1)^2\cdot\frac{18}{36}=0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(1–1)^2\cdot\frac{9}{36}=\frac{9}{36}$

Add the values in the third column to find the expected value: $\mu=\frac{36}{36}=1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum:

$$\sigma=\sqrt{\frac{18}{36}} \approx 0.7071. \tag{6.2.2}$$

✔ Example 6.2.6

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win $50. If you lose the bet, you pay $20. Let X = the amount of profit from a bet.

$P(\text{win}) = P(\text{one moderate earthquake will occur}) = 21.42$

$P(\text{loss}) = P(\text{one moderate earthquake will not occur}) = 100$

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X? Construct a table similar to Table and Table to help you answer these questions.

Answer

	x	$P(x)$	$xP(x)$	$(x – \mu^2)P(x)$
win	50	0.2142	10.71	$[50 – (–5.006)]^2(0.2142) = 648.0964$
loss	–20	0.7858	–15.716	$[–20 – (–5.006)]^2(0.7858) = 176.6636$

Mean = Expected Value = 10.71 + (–15.716) = –5.006.

If you make this bet many times under the same conditions, your long term outcome will be an average *loss* of $5.01 per bet.

Standard Deviation $= \sqrt{648.0964+176.6636} \approx 28.7186$

? Exercise 6.2.6

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win $100. If you lose the bet, you pay $10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X.

Answer

	x	$P(x)$	$x \cdot P(x)$	$(x-\mu^2) \cdot P(x)$
win	100	0.0108	1.08	$[100 – (–8.812)]^2 \cdot 0.0108 = 127.8726$
loss	–10	0.9892	–9.892	$[–10 – (–8.812)]^2 \cdot 0.9892 = 1.3961$

Mean = Expected Value $= \mu = 1.08 + (-9.892) = -8.812$

If you make this bet many times under the same conditions, your long term outcome will be an average loss of $8.81 per bet.

Standard Deviation $= \sqrt{127.7826+1.3961} \approx 11.3696$

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Summary

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Formula Review

1. Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$
2. Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x-\mu)^2 P(x)}$

Glossary

Expected Value

expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP(x)$.

Mean

a number that measures the central tendency; a common name for mean is ‘average.’ The term ‘mean’ is a shortened form of ‘arithmetic mean.’ By definition, the mean for a sample (detonated by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$ and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Mean of a Probability Distribution

the long-term average of many trials of a statistical experiment

Standard Deviation of a Probability Distribution

a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

The Law of Large Numbers

As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

References

1. Class Catalogue at the Florida State University. Available online at apps.oti.fsu.edu/RegistrarCo...archFormLegacy (accessed May 15, 2013).
2. “World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. www.world-earthquakes.com/ind...thq_prediction (accessed May 15, 2013).

6.3: PowerPoints

https://professormo.com/holistic/Powerpoint/ch05.pptx

CHAPTER OVERVIEW

7: Continuous Random Variable and Normal Probability Distribution

7.1: Videos

Normal distribution - Core concepts and several examples

Binomial distribution - Introduction to the binomial distribution

Normal approximation to binomial - A useful technique for some binomial situations

7.2: Continuous Random Variable - Introduction

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

> The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Figure $7.2.1$: The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve. The curve is called the probability density function (abbreviated as pdf). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution. Area under the curve is given by a different function called the cumulative distribution function(abbreviated as cdf). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the *x*-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the *x*-axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook. There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter and the next, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

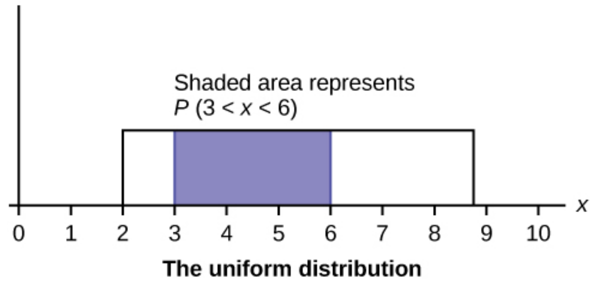


Figure 7.2.2: The graph shows a Uniform Distribution with the area between $x=3$ and $x=6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

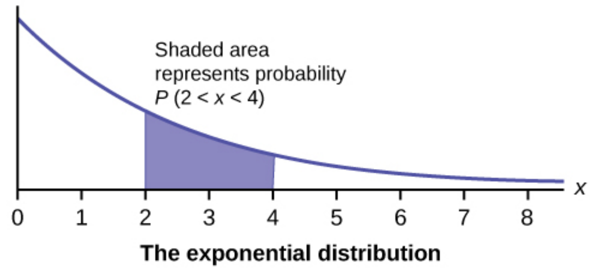


Figure 7.2.3: The graph shows an Exponential Distribution with the area between $x=2$ and $x=4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

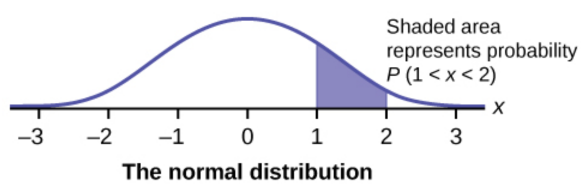


Figure 7.2.4: The graph shows the Standard Normal Distribution with the area between $x=1$ and $x=2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

Glossary

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, $a<x<b$; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a,b)$. The mean is $\mu=\frac{a+b}{2}$ and the standard deviation is $\sigma=\sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x)=\frac{1}{b-a}$ for $a<x<b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x)=\frac{x-a}{b-a}$.

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is $X \sim \operatorname{Exp}(m)$. The mean is $\mu=\frac{1}{m}$ and the standard deviation is $\sigma=\frac{1}{m}$. The probability density function is $f(x)=m e^{-m x}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x)=1-e^{m x}$.

SECTION OVERVIEW

7.3: The Normal Distribution

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ).

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

7.3.1: Prelude to The Normal Distribution

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

Figure 7.3.1.1: If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing

$$f(x)=\frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot\left(\frac{x-\mu}{\sigma}\right)^{2}} \tag{7.3.1.1}$$

The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

The cumulative distribution function is $P(X<x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions.

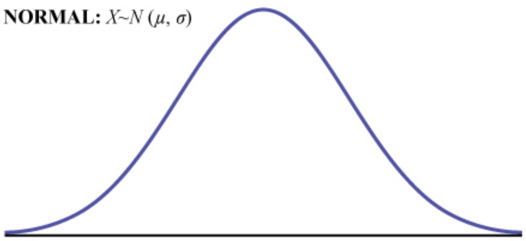


Figure 7.3.1.2: The standard normal distribution

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

COLLABORATIVE CLASSROOM ACTIVITY

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

- $X \sim N(\mu, \sigma)$
- $\mu =$ the mean $\sigma =$ the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x \cdot \mu)^2}{2\sigma^2}} \tag{7.3.1.2}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

7.3.2: The Standard Normal Distribution

Z-Scores

The standard normal distribution is a normal distribution of standardized values called *z-scores*. A z-score is measured in units of the standard deviation.

Definition: Z-Score

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score is:

$$z = \frac{x - \mu}{\sigma} \tag{7.3.2.1}$$

The *z*-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$\begin{aligned} x &= \mu + (z)(\sigma) \\ &= 5 + (3)(2) = 11 \end{aligned}$$

The *z*-score is three.

Since the mean for the standard normal distribution is zero and the standard deviation is one, then the transformation in Equation 7.3.2.1produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ.

> *A z-score is measured in units of the standard deviation.*

Example 7.3.2.1

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then (via Equation 7.3.2.1):

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two** standard deviations (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$)

Now suppose $x = 1$. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$$

(rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: $5 + (-0.67)(6)$ is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.

Exercise 7.3.2.1

What is the z-score of x, when $x = 1$ and $X \sim N(12, 3)$?

Answer

$$z = \frac{1-12}{3} \approx -3.67$$

Example 7.3.2.2

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let $X =$ the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5,2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $x = 2.5$ (verify). This z-score tells you that $x = 10$ is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).
b. Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ __________. This z-score tells you that $x = -3$ is ________ standard deviations to the __________ (right or left) of the mean.

Answers

a. This z-score tells you that $x = 10$ is 2.5 standard deviations to the right of the mean five.

b. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5,6)$ and $Y \sim N(2,1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

$$z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2$$

where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of their own) standard deviations to the right of their respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5,6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2,1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Exercise 7.3.2.2

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is -1.5. This score tells you that $x = 10$ is _____ standard deviations to the ______(right or left) of the mean______(What is the mean?).

Answer

1.5, left, 16

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the *Empirical Rule* says the following:

- About 68% of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for +1σ and –1σ are +1 and –1, respectively.
- The z-scores for +2σ and –2σ are +2 and –2, respectively.
- The z-scores for +3σ and –3σ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.

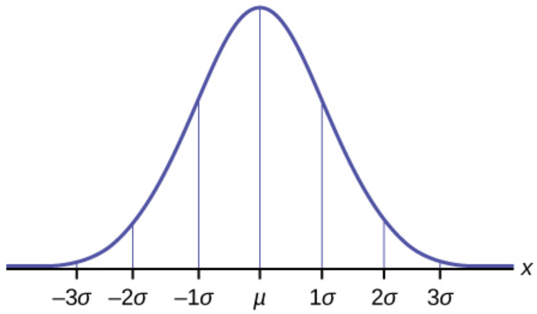


Figure 7.3.2.1

Example 7.3.2.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when $x = 168$ cm is $z =$ _______. This z-score tells you that $x = 168$ is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = 1.27$. What is the male's height? The z-score ($z = 1.27$) tells you that the male's height is ________ standard deviations to the __________ (right or left) of the mean.

Answers

a. –0.32, 0.32, left, 170
b. 177.98, 1.27, right

Exercise 7.3.2.3

Use the information in Example $7.3.2.3$to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when $x = 176$ cm is $z =$ _______. This z-score tells you that $x = 176$ cm is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = -2$. What is the male's height? The z-score ($z = -2$) tells you that the male's height is ________ standard deviations to the __________ (right or left) of the mean.

Answer

Solve the equation $z = \frac{x-\mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$z = \frac{176-170}{6.28}$, This z-score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.

Answer

Solve the equation $z = \frac{x-\mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$X = 157.44$cm, The z-score($z = -2$) tells you that the male's height is two standard deviations to the left of the mean.

Example 7.3.2.4

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the *z*-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Answer

- The z-score (Equation 7.3.2.1) for $x = 160.58$ is $z = -1.5$.
- The z-score for $y = 162.85$ is $z = -1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

Exercise 7.3.2.4

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let $X =$ a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Answer

The *z*-score (Equation 7.3.2.1) for $x_1 = 325$ is $z_1 = -1.15$.

The *z*-score (Equation 7.3.2.1) for $x_2 = 366.21$ is $z_2 = -1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example 7.3.2.5

Suppose *x* has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the *x* values lie within one standard deviation of the mean. Therefore, about 68% of the *x* values lie between –1σ = (–1)(6) = –6 and 1σ = (1)(6) = 6 of the mean 50. The values 50 – 6 = 44 and 50 + 6 = 56 are within one standard deviation from the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between –2σ = (–2)(6) = –12 and 2σ = (2)(6) = 12. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations from the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 99.7% of the x values lie between –3σ = (–3)(6) = –18 and 3σ = (3)(6) = 18 from the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Exercise 7.3.2.5

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Answer

between 20 and 30.

Example 7.3.2.6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

a. About 68% of the y values lie between what two values? These values are ________________. The z-scores are ________________, respectively.
b. About 95% of the y values lie between what two values? These values are ________________. The z-scores are ________________ respectively.
c. About 99.7% of the y values lie between what two values? These values are ________________. The z-scores are ________________, respectively.

Answer

a. About 68% of the values lie between 166.02 and 178.7. The z-scores are –1 and 1.
b. About 95% of the values lie between 159.68 and 185.04. The z-scores are –2 and 2.
c. About 99.7% of the values lie between 153.34 and 191.38. The z-scores are –3 and 3.

Exercise 7.3.2.6

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the y values lie between what two values? These values are ________________. The z-scores are ________________, respectively.
b. About 95% of the y values lie between what two values? These values are ________________. The z-scores are ________________, respectively.
c. About 99.7% of the y values lie between what two values? These values are ________________. The z-scores are ________________, respectively.

Answer a

About 68% of the values lie between the values 41 and 63. The z-scores are –1 and 1, respectively.

Answer b

About 95% of the values lie between the values 30 and 74. The z-scores are –2 and 2, respectively.

Answer c

About 99.7% of the values lie between the values 19 and 85. The z-scores are –3 and 3, respectively.

Summary

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If y is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Formula Review

$Z \sim N(0, 1)$

$z = a$ standardized value (z-score)

mean = 0; standard deviation = 1

To find the K^{th} percentile of X when the z-scores is known:

$k = \mu + (z)\sigma$

z-score: $z = \dfrac{x - \mu}{\sigma}$

$Z =$ the random variable for z-scores

$Z \sim N(0,1)$

Glossary

Standard Normal Distribution

a continuous random variable (RV) $X \sim N(0,1)$; when X follows the standard normal distribution, it is often noted as \(Z \sim N(0, 1)\.

***z*-score**

the linear transformation of the form $z = \frac{x-\mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma$ the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

References

1. "Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewre...reportid=11960 (accessed May 14, 2013).
2. "The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).
3. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at media.collegeboard.com/digita...Group-2012.pdf (accessed May 14, 2013).
4. "Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009." National Center for Education Statistics. Available online at nces.ed.gov/programs/digest/d...s/dt09_147.asp (accessed May 14, 2013).
5. Data from the *San Jose Mercury News*.
6. Data from *The World Almanac and Book of Facts*.
7. "List of stadiums by capacity." Wikipedia. Available online at en.Wikipedia.org/wiki/List_o...ms_by_capacity (accessed May 14, 2013).
8. Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

7.3.2E: The Standard Normal Distribution (Exercises)

Exercise $7.3.2E.7$

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ ____________.

Answer

ounces of water in a bottle

Exercise $7.3.2E.8$

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Exercise $7.3.2E.9$

$X \sim N(1, 2)$

$\sigma =$ _______

Answer

2

Exercise $7.3.2E.10$

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. $X =$ ______________.

Exercise $7.3.2E.11$

$X \sim N(-4, 1)$

What is the median?

Answer

–4

Exercise $7.3.2E.12$

$X \sim N(3, 5)$

$\sigma =$ _______

Exercise $7.3.2E.13$

$X \sim N(-2, 1)$

$\mu =$ _______

Answer

–2

? Exercise 7.3.2E.14

What does a z-score measure?

? Exercise 7.3.2E.15

What does standardizing a normal distribution do to the mean?

Answer

The mean becomes zero.

? Exercise 7.3.2E.16

Is $X \sim N(0,1)$ a standardized normal distribution? Why or why not?

? Exercise 7.3.2E.17

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Answer

$z = 2$

? Exercise 7.3.2E.18

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

? Exercise 7.3.2E.19

What is the z-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

Answer

$z = 2.78$

? Exercise 7.3.2E.20

What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

? Exercise 7.3.2E.21

Suppose $X \sim N(2,6)$. What value of *x* has a *z*-score of three?

Answer

$x = 20$

? Exercise 7.3.2E.22

Suppose $X \sim N(8,1)$. What value of x has a z-score of –2.25?

? Exercise 7.3.2E.23

Suppose $X \sim N(9,5)$. What value of x has a z-score of –0.5?

Answer

$x = 6.5$

Exercise 7.3.2E.24

Suppose $X \sim N(2,3)$. What value of x has a z-score of –0.67?

Exercise 7.3.2E.25

Suppose $X \sim N(4,2)$. What value of x is 1.5 standard deviations to the left of the mean?

Answer

$x = 1$

Exercise 7.3.2E.26

Suppose $X \sim N(4,2)$. What value of x is two standard deviations to the right of the mean?

Exercise 7.3.2E.27

Suppose $X \sim N(8,9)$. What value of x is 0.67 standard deviations to the left of the mean?

Answer

$x = 1.97$

Exercise 7.3.2E.28

Suppose $X \sim N(-1,12)$. What is the z-score of $x = 2$?

Exercise 7.3.2E.29

Suppose $X \sim N(12,6)$. What is the z-score of $x = 2$?

Answer

$z =- 1.67$

Exercise 7.3.2E.30

Suppose $X \sim N(9,3)$. What is the z-score of $x = 9$?

Exercise 7.3.2E.31

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

Answer

$z \approx - 0.33$

Exercise 7.3.2E.32

In a normal distribution, $x = 5$ and $z =- 1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 7.3.2E.33

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

0.67, right

Exercise $7.3.2E.34$

In a normal distribution, $x = -2$ and $z = 6$. This tells you that $z = -2$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise $7.3.2E.35$

In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

3.14, left

Exercise $7.3.2E.36$

In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise $7.3.2E.37$

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Answer

about 68%

Exercise $7.3.2E.38$

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Exercise $7.3.2E.39$

About what percent of x values lie between the second and third standard deviations (both sides)?

Answer

about 4%

Exercise $7.3.2E.40$

Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Exercise $7.3.2E.41$

Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution (i.e., –3).

Answer

between –5 and –1

? Exercise $7.3.2E.42$

Suppose $X \sim N(-3,1)$. Between what x values does 34.14% of the data lie?

? Exercise $7.3.2E.43$

About what percent of x values lie between the mean and three standard deviations?

Answer

about 50%

? Exercise $7.3.2E.44$

About what percent of x values lie between the mean and one standard deviation?

? Exercise $7.3.2E.45$

About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

Answer

about 27%

? Exercise $7.3.2E.46$

About what percent of x values lie between the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

? Exercise $7.3.2E.47$

Define the random variable X in words. $X =$ _______________.

Answer

The lifetime of a Sunshine CD player measured in years.

? Exercise $7.3.2E.48$

$X \sim$ _____(_____,_____)

7.3.3: Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X<x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X<x)$.

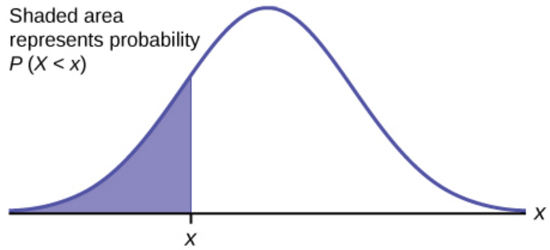


Figure 7.3.3.1.

The area to the right is then $P(X>x)=1-P(X<x)$. Remember, $P(X<x)=$ **Area to the left** of the vertical line through x. $P(X>x)=1-P(X<x)=$ **Area to the right** of the vertical line through x. $P(X<x)$ is the same as $P(X\leq x)$ and $P(X>x)$ is the same as $P(X\geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.To calculate the probability, use the probability tables provided in [link] without the use of technology. The tables include instructions for how to use them.

Example 7.3.3.1

If the area to the left is 0.0228, then the area to the right is $1-0.0228=0.9772$

Exercise 7.3.3.1

If the area to the left of x is 0.012, then what is the area to the right?

Answer

$1-0.012=0.988$

Example 7.3.3.2

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.
c. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).
d. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Answer

a. Let X = a score on the final exam. $X\sim N(63,5)$, where $\mu=63$ and $\sigma=5$

Draw a graph.

Then, find $P(x>65)$.

$$P(x>65)=0.3446$$

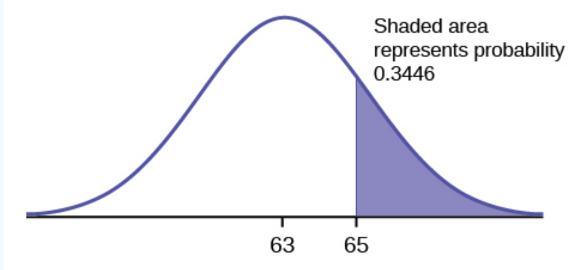


Figure 7.3.3.2.

The probability that any student selected at random scores more than 65 is 0.3446.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go into `2nd DISTR` .

After pressing `2nd DISTR` , press `2:normalcdf` .

The syntax for the instructions are as follows:

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1` , the `EE` key (a 2nd key) and then `99` . Or, you can enter `10^ 99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Historical Note

The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = 65 - 63565 - 635 = 0.4$$

Area to the left is 0.6554.

$$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$$

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Find the percentile for a student scoring 65:

*Press `2nd Distr`
*Press `2:normalcdf` (
*Enter lower bound, upper bound, mean, standard deviation followed by)
*Press `ENTER` .
For this Example, the steps are
`2nd Distr`
`2:normalcdf` (65,1,2nd EE,99,63,5) `ENTER`

The probability that a selected student scored more than 65 is 0.3446.
To find the probability that a selected student scored *more than* 65, subtract the percentile from 1.

Answer

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

$\text{normalcdf}(0, 85, 63, 5) = 1$(rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Answer

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a critical value.

$k = 69.4$

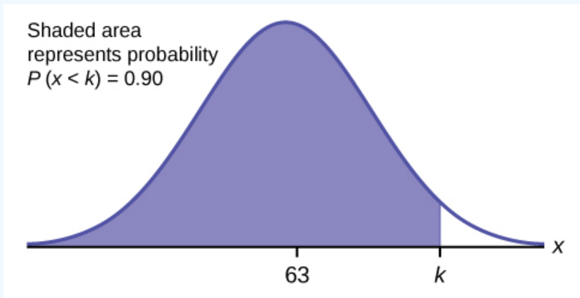


Figure 7.3.3.3.

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

`invNorm` in `2nd DISTR`. invNorm(area to the left, mean, standard deviation)

For this problem, $\text{invNorm}(0.90, 63, 5) = 69.4$

Answer

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.6 and 30% fall at or above.

$$\text{invNorm}(0.70, 63, 5) = 65.6$$

Exercise 7.3.3.2

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a randomly selected golfer scored less than 65.

Answer

$\text{normalcdf}(10^{99}, 65, 68, 3) = 0.1587$

Example 7.3.3.3

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer

a. Let $X =$ the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

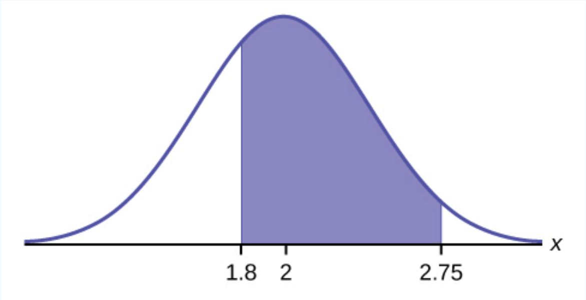


Figure $7.3.3.4$.

$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b.

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25$^{\text{th}}$ percentile,** k, where $P(x < k) = 0.25$.

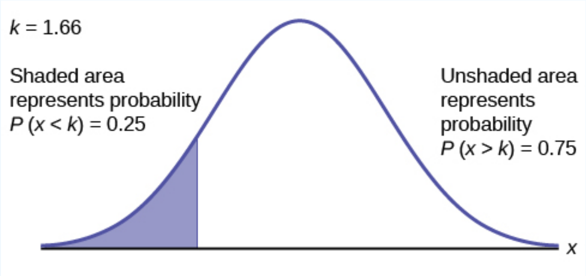


Figure $7.3.3.5$.

$$\text{invNorm}(0.25, 2, 0.5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

? Exercise $7.3.3.3$

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Answer

$$\text{normalcdf}(66, 70, 68, 3) = 0.4950$$

Example 7.3.3.4

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
c. Find the 80$^{\text{th}}$ percentile of this distribution, and interpret it in a complete sentence.

Answer

a. $\text{normalcdf}(23, 64.7, 36.9, 13.9) = 0.8186$
b. $\text{normalcdf}(-10^{99}, 50.8, 36.9, 13.9) = 0.8413$
c. $\text{invNorm}(0.80, 36.9, 13.9) = 48.6$

The 80$^{\text{th}}$ percentile is 48.6 years.

80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Use the information in Example to answer the following questions.

Exercise 7.3.3.4

a. Find the 30$^{\text{th}}$ percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old and at least 0 years old?

70.

Answer

Let $X =$ a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

To find the 30$^{\text{th}}$ percentile, find k such that $P(x < k) = 0.30$.
$\text{invNorm}(0.30, 36.9, 13.9) = 29.6$ years
Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years. Find $P(x < 27)$
(Note that $\text{normalcdf}(-10^{99}, 27, 36.9, 13.9) = 0.2382$. The two answers differ only by 0.0040.)

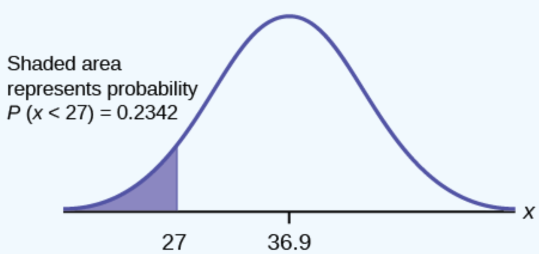


Figure 7.3.3.6.

$$\text{normalcdf}(0, 27, 36.9, 13.9) = 0.2342$$

Example 7.3.3.5

In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information, answer the following questions (round answers to one decimal place).

a. Calculate the interquartile range (IQR).
b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Answer

a.

$$IQR = Q_3 - Q_1$$

Calculate $Q_3 = 75^{\text{th}}$ percentile and $Q_1 = 25^{\text{th}}$ percentile.

$$\text{invNorm}(0.75, 36.9, 13.9) = Q_3 = 46.2754$$

$$\text{invNorm}(0.25, 36.9, 13.9) = Q_1 = 27.5246$$

$$IQR = Q_3 - Q_1 = 18.7508$$

b.

Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.")

$0.40 =$ the area to the right.

Area to the left $= 1 - 0.40 = 0.60$.

The area to the left of $k = 0.60$.

$\text{invNorm}(0.60, 36.9, 13.9) = 40.4215$

$k = 40.42$.

Forty percent of the smartphone users from 13 to 55+ are at least 40.4 years.

Exercise 7.3.3.5

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Answer

a. $Q_1 = 25^{\text{th}}$ percentile $= \text{invNorm}(0.25, 81, 15) = 70.9$
   $Q_3 = 75^{\text{th}}$ percentile $= \text{invNorm}(0.75, 81, 15) = 91.1$
b. The middle 50% of the scores are between 70.9 and 91.1.

Example 7.3.3.6

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
b. The middle 20% of mandarin oranges from this farm have diameters between ______ and ______.
c. Find the 90^{th} percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Answer

a. $\text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$

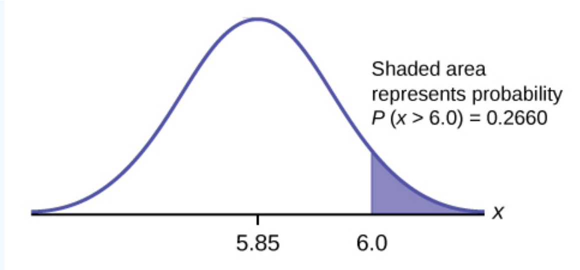


Figure 7.3.3.7.

Answer

b.

$1 - 0.20 = 0.80$

The tails of the graph of the normal distribution each have an area of 0.40.

Find $k1$, the 40$^{\text{th}}$ percentile, and $k2$, the 60$^{\text{th}}$ percentile ($0.40 + 0.20 = 0.60$).

$k1 = \text{invNorm}(0.40, 5.85, 0.24) = 5.79\text{cm}$

$k2 = \text{invNorm}(0.60, 5.85, 0.24) = 5.91\text{cm}$

Answer

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

Exercise 7.3.3.6

Using the information from Example, answer the following:

a. The middle 45% of mandarin oranges from this farm are between ______ and ______.
b. Find the 16$^{\text{th}}$ percentile and interpret it in a complete sentence.

Answer a

The middle area $= 0.40$, so each tail has an area of 0.30.

$-0.40 = 0.60$

The tails of the graph of the normal distribution each have an area of 0.30.

Find $k1$, the 30$^{\text{th}}$ percentile and $k2$, the 70$^{\text{th}}$ percentile ($0.40 + 0.30 = 0.70$).

$k1 = \text{invNorm}(0.30, 5.85, 0.24) = 5.72\text{cm}$

$k2 = \text{invNorm}(0.70, 5.85, 0.24) = 5.98\text{cm}$

Answer b

$\text{normalcdf}(5, 10^{99}, 5.85, 0.24) = 0.9998$

References

1. "Naegele's rule." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).
2. "403: NUMMI." Chicago Public Media & Ira Glass, 2013. Available online at www.thisamericanlife.org/radi...sode/403/nummi (accessed May 14, 2013).
3. "Scratch-Off Lottery Ticket Playing Tips." WinAtTheLottery.com, 2013. Available online at www.winatthelottery.com/publi...partment40.cfm (accessed May 14, 2013).
4. "Smart Phone Users, By The Numbers." Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).
5. "Facebook Statistics." Statistics Brain. Available online at http://www.statisticbrain.com/facebo...tics/(accessed May 14, 2013).

Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of *z*-scores. Its mean is zero, and its standard deviation is one.

Formula Review

- Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.
- Standard Normal Distribution: $Z \sim N(0, 1)$.
- Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)
- Calculator function for the k^{th} percentile: $k = \text{invNorm}$ (area to the left of k, mean, standard deviation)

Exercise 7.3.3.7

How would you represent the area to the left of one in a probability statement?

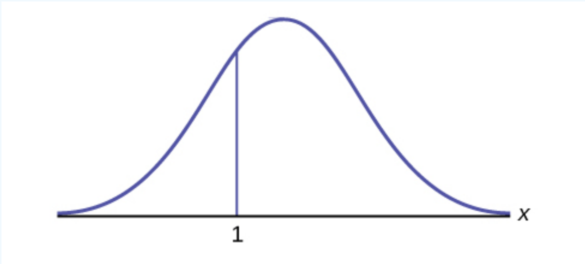


Figure 7.3.3.8.

Answer

$P(x < 1)$

Exercise 7.3.3.8

Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

Answer

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

Exercise 7.3.3.9

How would you represent the area to the left of three in a probability statement?

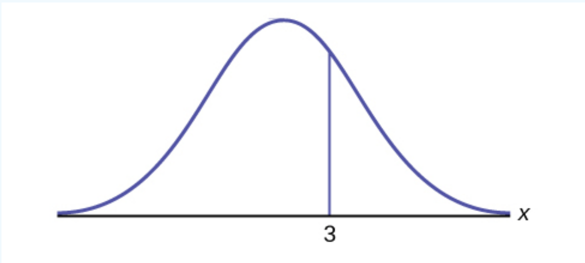


Figure 7.3.3.10.

Exercise 7.3.3.10

What is the area to the right of three?

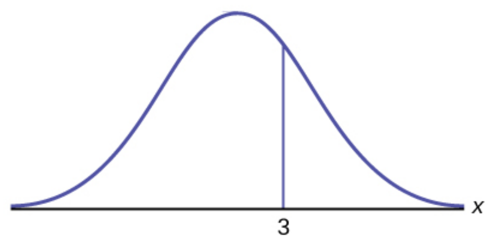


Figure 7.3.3.11.

Answer

$1 - P(x < 3)$ or $P(x > 3)$

Exercise 7.3.3.11

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Exercise 7.3.3.12

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Answer

$1 - 0.543 = 0.457$

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

Exercise 7.3.3.13

Find the probability that $x > 56$.

Exercise 7.3.3.14

Find the probability that $x < 30$.

Answer

0.0013

Exercise 7.3.3.15

Find the 80$^{\text{th}}$ percentile.

Exercise 7.3.3.16

Find the 60$^{\text{th}}$ percentile.

Answer

56.03

Exercise 7.3.3.17

$X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise 7.3.3.18

$X \sim N(-3, 4)$

Find the probability that x is between one and four.

Answer

0.1186

Exercise 7.3.3.19

$X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

Exercise 7.3.3.20

Use the following information to answer the next three exercise: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

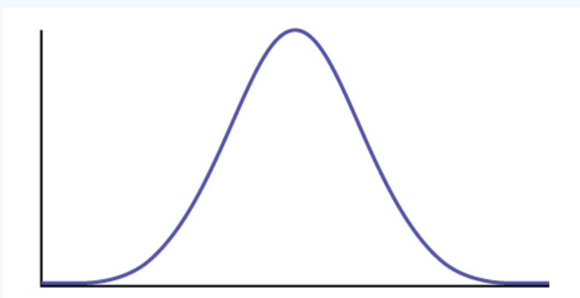


***Figure* 7.3.3.12.**

$P(0 < x <$ ____________$)=$ ___________ (Use zero for the minimum value of x.)

Answer

a. Check student's solution.
b. 3, 0.1979

Exercise 7.3.3.21

Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

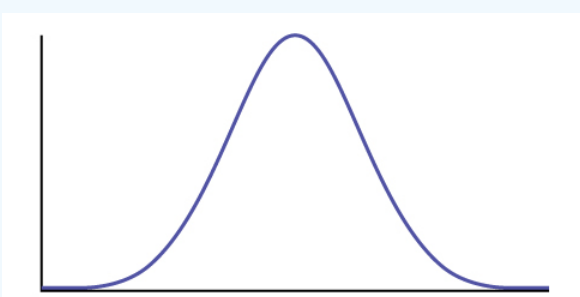


***Figure* 7.3.3.13.**

$P($__________$< x <$__________$) =$__________

Exercise 7.3.3.22

Find the 70th percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.

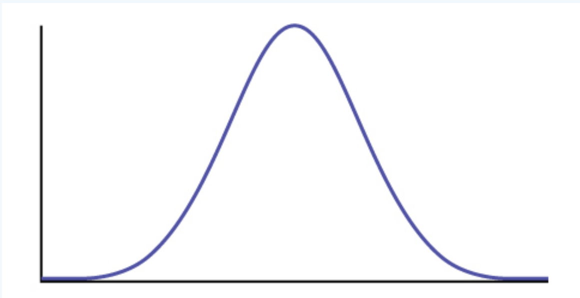


***Figure* $7.3.3.14$.**

$P(x < k) =$ __________ Therefore, $k =$ _________

Answer

a. Check student's solution.
b. 0.70, 4.78 years

SECTION OVERVIEW

7.4: The Central Limit Theorem

In this chapter, you will study means and the **central limit theorem**, which is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

7.4.1: Prelude to the Central Limit Theorem

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**. The **central limit theorem** (clt for short) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means and the sums tend to follow the normal distribution.

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.**

Figure $7.4.1.1$. If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

COLLABORATIVE CLASSROOM ACTIVITY

Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample **mean** of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Since the "mean" when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?

- **Draw the graph for the means using two dice.** Do the sample means show any kind of pattern?
- **Draw the graph for the means using five dice.** Do you see any pattern emerging?
- **Finally, draw the graph for the means using ten dice.** Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt). The central limit theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

Glossary

Sampling Distribution

Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

7.4.2: The Central Limit Theorem for Sums

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution) and suppose:

- μ_x = the mean of X
- σ_x = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\sum X$ consisting of sums tends to be normally distributed and

$$\sum X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x)). \tag{7.4.2.1}$$

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.

The random variable $\sum X$ has the following *z*-score associated with it:

a. $\sum x$ is one sum.
b. $z = \frac{\sum x - (n)(\mu_x)}{(\sqrt{n})(\sigma_x)}$
   i. $(n)(\mu_x)$= the mean of $\sum X$
   ii. $(\sqrt{n})(\sigma_x)$= standard deviation of $\sum X$

Calculator

To find probabilities for sums on the calculator, follow these steps.

2nd `DISTR`

2: `normalcdf`
`normalcdf` (lower value of the area, upper value of the area, (n)(mean), $(\sqrt{n})$(standard deviation))

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* $= n$

Example 7.4.2.1

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7,500.
b. Find the sum that is 1.5 standard deviations above the mean of the sums.

Answer

Let $X =$ one value from the original unknown population. The probability question asks you to find a probability for the sum (or total of) 80 values.

$\sum X =$ the sum or total of 80 values. Since $\mu_x = 90$, $\sigma_x = 15$, and $n = 80$, $\sum X \sim N((80)(90), (\sqrt{80})(15))$

- mean of the sums $= (n)(\mu_x) = (80)(90) = 7,200$
- standard deviation of the sums $= (\sqrt{n})(\sigma_x) = (\sqrt{80})(15) = (80)(15)$
- sum of 80 values $= \sum X = 7,500$

a. Find $P(\sum X > 7,500)$

$P(\sum X > 7,500) = 0.0127$

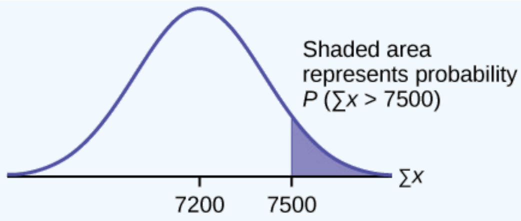


Figure 7.4.2.1.

`normalcdf` (lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated $(lower, upper, (n)(\mu_x, (\sqrt{n}(\sigma_x))$

`normalcdf` $\left(7500, 1E99, (80)(90), (\sqrt{80})(15)\right) = 0.0127$

REMINDER

1E99 = 10^{99}.

Press the `EE` key for E.

b. Find $\sum x$ where $z = 1.5$.

$\sum x = (n)(\nu_x) + (z)(\sqrt{n})(\sigma_x) = (80)(90) + (1.5)(\sqrt{80})(15) = 7,401.2$

Exercise 7.4.2.1

An unknown distribution has a mean of 45 and a standard deviation of eight. A sample size of 50 is drawn randomly from the population. Find the probability that the sum of the 50 values is more than 2,400.

Answer

0.0040

Calculator

To find percentiles for sums on the calculator, follow these steps.

2nd DIStR

3:invNorm

$k = \text{invNorm}(\text{area to the left of} k, (n)(\text{mean}), (\sqrt{n})(\text{standard deviation}))$

where:

- k is the k^{th} **percentile**
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* $= n$

Example 7.4.2.2

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. The sample of size is 50.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
b. Find the probability that the sum of the ages is between 1,500 and 1,800 years.
c. Find the 80th percentile for the sum of the 50 ages.

Answer

a. $\mu_x - n\mu_x = 1,700$ and $\sigma_{\sum X} = \sqrt{n}\sigma_X = (\sqrt{50})(15) = 106.01$
   The distribution is normal for sums by the central limit theorem.

b. $P(1500 < \sum X < 1800) = (1,500, 1,800, (50)(34), (\sqrt{50})(15)) = 0.7974$
c. Let k = the 80$^{\text{th}}$ percentile.
$k = (0.80, (50)(34), (\sqrt{50})(15)) = 1,789.3$

Exercise 7.4.2.2

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 35 years. Suppose the standard deviation is ten years. The sample size is 39.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
b. Find the probability that the sum of the ages is between 1,400 and 1,500 years.
c. Find the 90$^{\text{th}}$ percentile for the sum of the 39 ages.

Answer

a. $\mu_{\sum X} = n\mu_X = 1,365$ and $\sigma_{\sum X} = \sqrt{n}\sigma_x = 62.4$
The distribution is normal for sums by the central limit theorem.
b. $P(1400 < \sum_X < 1500) = \text{normalcdf}(1400, 1500, (39)(35), (\sqrt{39})(10)) = 0.2723$
c. Let k = the 90$^{\text{th}}$ percentile.
$k = \text{invNorm}(0.90, (39)(35), (\sqrt{39})(10)) = 1445.0$

Example 7.4.2.3

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of size 70.

a. What are the mean and standard deviation for the sums?
b. Find the 95$^{\text{th}}$ percentile for the sum of the sample. Interpret this value in a complete sentence.
c. Find the probability that the sum of the sample is at least ten hours.

Answer

a. $\mu_{\sum X} = n\mu_X = 70(8.2) = 574$ minutes and $\sigma_{\sum X}(\sqrt{n})(\sigma_x) = (\sqrt{70})(1) = 8.37$ minutes
b. Let k = the 95$^{\text{th}}$ percentile.
$k = \text{invNorm}(0.95, (70)(8.2), (\sqrt{70})(1)) = 587.76\text{minutes}$
Ninety five percent of the app engagement times are at most 587.76 minutes.
c. ten hours = 600 minutes
$P(\sum X \geq 600) = \text{normalcdf}(600, E99, (70)(8.2), (\sqrt{70})(1)) = 0.0009$

Exercise 7.4.2.3

The mean number of minutes for app engagement by a table use is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample size of 70.

a. What is the probability that the sum of the sample is between seven hours and ten hours? What does this mean in context of the problem?
b. Find the 84$^{\text{th}}$ and 16$^{\text{th}}$ percentiles for the sum of the sample. Interpret these values in context.

Answer

a. 7 hours = 420 minutes
10 hours = 600 minutes
$\text{normalcdf}P(420 \leq \sum X \leq 600) = \text{normalcdf}(420, 600, (70)(8.2), \sqrt{70}(1)) = 0.9991$
This means that for this sample sums there is a 99.9% chance that the sums of usage minutes will be between 420 minutes and 600 minutes.
b. $\text{invNorm}(0.84, (70)(8.2)\sqrt{70}(1)) = 582.32$
$\text{invNorm}(0.16, (70)(8.2), (\sqrt{70}(1)) = 565.68$
Since 84% of the app engagement times are at most 582.32 minutes and 16% of the app engagement times are at most

565.68 minutes, we may state that 68% of the app engagement times are between 565.68 minutes and 582.32 minutes.

References

1. Farago, Peter. “The Truth About Cats and Dogs: Smartphone vs Tablet Usage Differences.” The Flurry Blog, 2013. Posted October 29, 2012. Available online at blog.flurry.com (accessed May 17, 2013).

Review

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution even if the original population is not normally distributed. Additionally, if the original population has a mean of μ_x and a standard deviation of σ_x, the mean of the sums is $n\mu_x$ and the standard deviation is $(\sqrt{n})(\sigma_x)$ where n is the sample size.

Formula Review

- The Central Limit Theorem for Sums: $\sum X \; N[(n)(\mu_x, (\sqrt{n})(\sigma_x))]$
- Mean for Sums $(\sum X) : (n)(\mu_x)$
- The Central Limit Theorem for Sums z-score and standard deviation for sums: $z \text{ for the sample mean} = \frac{\sum x-(n)(\mu_x)}{(\sqrt{n})(\sigma_x)}$
- Standard deviation for Sums $(\sum X) : (\sqrt{n})(\sigma_x)$

7.5: PowerPoints

https://professormo.com/holistic/Powerpoint/ch06.pptx

https://professormo.com/holistic/Powerpoint/ch07.pptx

CHAPTER OVERVIEW

8: Finding Confidence Interval for Population Mean and Proportion

SECTION OVERVIEW

8.1: Inference for Numerical Data

Chapter 4 introduced a framework for statistical inference based on con dence intervals and hypotheses. In this chapter, we encounter several new point estimates and scenarios. In each case, the inference ideas remain the same:

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Apply the ideas from Chapter 4 using the distribution from step 2.

Each section in Chapter 5 explores a new situation: the difference of two means (5.1, 5.2); a single mean or difference of means where we relax the minimum sample size condition (5.3, 5.4); and the comparison of means across multiple groups (5.5). Chapter 6 will introduce scenarios that highlight categorical data.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

8.1.1: One-Sample Means with the t Distribution

The motivation in Chapter 4 for requiring a large sample was two-fold. First, a large sample ensures that the sampling distribution of $\bar{x}$ is nearly normal. We will see in Section 5.3.1 that if the population data are nearly normal, then $\bar{x}$ is also nearly normal regardless of the

[10]*The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances* σ_x^2 *and* σ_y^2*, then the variance of X - Y is* $\sigma_x^2 + \sigma_y^2$ *. Likewise, the variance corresponding to* $\bar{x}_1 - \bar{x}_2$ *is* $\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$ *. Because* $\sigma_{\bar{x}_1}^2$ *and* $\sigma_{\bar{x}_2}^2$ *are just another way of writing* $SE_{\bar{x}_1}^2$ *and* $SE_{\bar{x}_2}^2$*, the variance associated with* $\bar{x}_1 - \bar{x}_2$ *may be written as* $SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2$

sample size. The second motivation for a large sample was that we get a better estimate of the standard error when using a large sample. The standard error estimate will not generally be accurate for smaller sample sizes, and this motivates the introduction of the t distribution, which we introduce in Section 5.3.2.

We will see that the t distribution is a helpful substitute for the normal distribution when we model a sample mean $\bar{x}$ that comes from a small sample. While we emphasize the use of the t distribution for small samples, this distribution may also be used for means from large samples.

The normality condition

We use a special case of the Central Limit Theorem to ensure the distribution of the sample means will be nearly normal, regardless of sample size, provided the data come from a nearly normal distribution.

Central Limit Theorem for normal data

The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

While this seems like a very helpful special case, there is one small problem. It is inherently difficult to verify normality in small data sets.

Caution: Checking the normality condition

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I con dent that outliers are rare?

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

Introducing the t distribution

The second reason we previously required a large sample size was so that we could accurately estimate the standard error using the sample data. In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t distribution. A t distribution, shown as a solid line in Figure 5.10, has a bell shape. However, its tails are thicker than the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[11] These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell shaped t distribution.

[11]*The standard deviation of the t distribution is actually a little more than 1. However, it is useful to always think of the t distribution as having a standard deviation of 1 in all of our applications.*

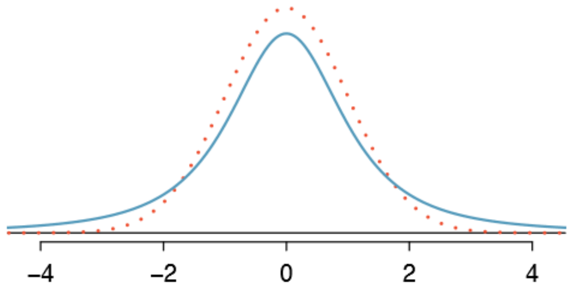


Figure 5.10: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

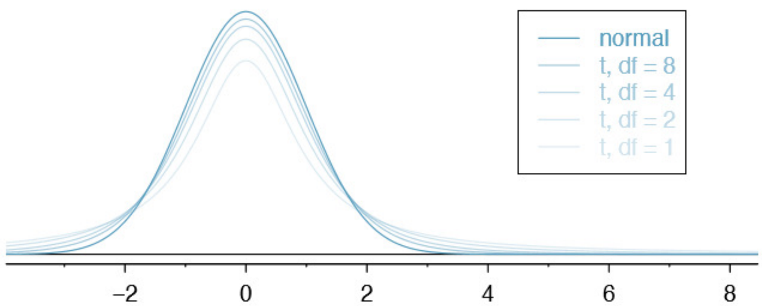


Figure 5.11: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

Several t distributions are shown in Figure 5.11. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution.

Degrees of freedom (df)

The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution. In Section 5.3.3, we relate degrees of freedom to sample size.

We will nd it very useful to become familiar with the t distribution, because it plays a very similar role to the normal distribution during inference for small samples of numerical data. We use a t table, partially shown in Table 5.12, in place of the normal probability table for small sample numerical data. A larger table is presented in Appendix B.2 on page 410.

Each row in the t table represents a t distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t distribution with df = 18, we can examine row 18, which is highlighted in

one tail two tails	0.100 0.200	0.050 0.100	0.025 0.050	0.010 0.020	0.005 0.010

df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	3 5.84
⋮	⋮	⋮	⋮	⋮	⋮
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	⋮
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

Table 5.12: An abbreviated look at the t table. Each row represents a different t distribution. The columns describe the cutoffs for specific tail areas. The row with df = 18 has been highlighted.

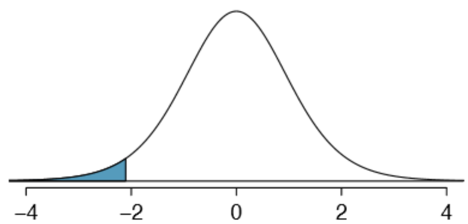


Figure 5.13: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where one tail is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t distributions are symmetric.

Example 5.15 What proportion of the t distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 5.13 and shade the area below -2.10. To nd this area, we identify the appropriate row: df = 18. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

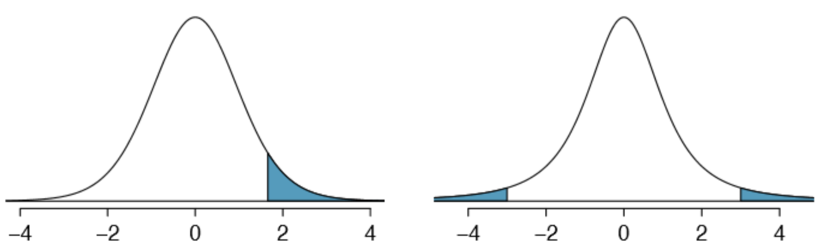


Figure 5.14: Left: The t distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

Example 5.16 A t distribution with 20 degrees of freedom is shown in the left panel of Figure 5.14. Estimate the proportion of the distribution falling above 1.65.

We identify the row in the t table using the degrees of freedom: df = 20. Then we look for 1.65; it is not listed. It falls between the rst and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the rst and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

Example 5.17 A t distribution with 2 degrees of freedom is shown in the right panel of Figure 5.14. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: df = 2. Next, nd the columns that capture 3; because 2:92 < 3 < 4:30, we use the second and third columns. Finally, we nd bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

Exercise 5.18 What proportion of the t distribution with 19 degrees of freedom falls above -1.79 units?[12]

The t distribution as a solution to the standard error problem

When estimating the mean and standard error from a small sample, the t distribution is a more accurate tool than the normal model. This is true for both small and large samples.

TIP: When to use the t distribution

Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

[12]*We find the shaded area above -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.*

To proceed with the t distribution for inference about a single mean, we must check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if it was an experiment or random process, we carefully check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t distribution with n - 1 degrees of freedom (df). For example, if the sample size was 19, then we would use the t distribution with df = 19 - 1 = 18 degrees of freedom and proceed exactly as we did in Chapter 4, except that now we use the t table.

One sample t confidence intervals

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

Figure 5.15: A Risso's dolphin. Photo by Mike Baird (http://www.bairdphotos.com/).

Here we identify a con dence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[13] The data are summarized in Table 5.16. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

[13]Taiji was featured in the movie The Cove, and it is a signi cant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins. Data reference: Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 5.16: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Example 5.19 Are the independence and normality conditions satis ed for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. The summary statistics in Table 5.16 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

In the normal model, we used z* and the standard error to determine the width of a con dence interval. We revise the confidence interval formula slightly when using the t distribution:

$$\bar{x} \pm t*_{df} SE \tag{8.1.1.1}$$

The sample mean and estimated standard error are computed just as before ($\bar{x}$ = 4.4 and $SE = \frac{s}{\sqrt{n}} = 0.528$). The value $t*_{df}$ is a cutoff we obtain based on the con dence level and the t distribution with df degrees of freedom. Before determining this cutoff , we will first need the degrees of freedom.

Degrees of freedom for a single sample

If the sample has n observations and we are examining a single mean, then we use the t distribution with df = n - 1 degrees of freedom.

In our current example, we should use the t distribution with df = 19 -1 = 18 degrees of freedom. Then identifying $t*_{18}$ is similar to how we found z*.

• For a 95% confidence interval, we want to nd the cutoff $t*_{18}$ such that 95% of the t distribution is between $-t*_{18}$ and $t*_{18}$.

• We look in the t table on page 224, nd the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t*_{18} = 2.10$.

Generally the value of $t*_{df}$ is slightly larger than what we would get under the normal model with z*.

Finally, we can substitute all our values into the con dence interval equation to create the 95% con dence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \pm t*_{18} SE \rightarrow 4.4 \pm 2.10 \times 0.528 \rightarrow (3.87, 4.93) \tag{8.1.1.2}$$

We are 95% con dent the average mercury content of muscles in Risso's dolphins is between 3.87 and 4.93 μg/wet gram. This is above the Japanese regulation level of 0.4 μg/wet gram.

Finding a t confidence interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \pm t*_{df} SE \tag{8.1.1.3}$$

where $\bar{x}$ is the sample mean, $t*_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

Exercise 5.20 The FDA's webpage provides some data on mercury content of sh.14 Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[15]

Example 5.21 Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Exercise 5.20. If we are to use the t distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t*_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: df = n - 1 = 14.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row df = 14, we identify $t*_{14} = 1.76$.

Exercise 5.22 Using the results of Exercise 5.20 and Example 5.21, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[16]

One sample t tests

An SAT preparation company claims that its students' scores improve by over 100 points on average after their course. A consumer group would like to evaluate this claim, and they collect data on a random sample of 30 students who took the class. Each of these students took the SAT before and after taking the company's course, and so we have a difference in scores for each student. We will examine these differences $x_1 = 57, x_2 = 133, \dots, x_30 = 140$ as a sample to evaluate the company's claim. (This is paired data, so we analyze the score differences; for a review of the ideas of paired data, see Section 5.1.) The distribution of the differences, shown in Figure 5.17, has mean 135.9 and standard deviation 82.2. Do these data provide convincing evidence to back up the company's claim?

Exercise 5.23 Set up hypotheses to evaluate the company's claim. Use μ_{diff} to represent the true average difference in student scores.[17]

[14]www.fda.gov/Food/FoodSafety/P...bornePathogens

contaminants/Methylmercury/ucm115644.htm

[15]There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red ags for the normal model based on this (limited) information, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

[16]$\bar{x} \pm t*_{14} SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. *We are 90% confident that the average mercury content of croaker white sh (Pacific) is between 0.256 and 0.318 ppm.*

[17]This is a one-sided test. H_0: student scores do not improve by more than 100 after taking the company's course. $\mu_{diff} = 100$ *(we always write the null hypothesis with an equality). H_A: students scores improve by more than 100 points on average after taking the company's course.* $\mu_{diff} > 100$.

Exercise 5.24 Are the conditions to use the t distribution method satisfied?[18]

Just as we did for the normal case, we standardize the sample mean using the Z score to identify the test statistic. However, we will write T instead of Z, because we have a small sample and are basing our inference on the t distribution:

$$T = \frac{\bar{x} - nullvalue}{SE} = \frac{135.9 - 100}{\frac{82.2}{\sqrt{30}}} = 2.39 \tag{8.1.1.4}$$

If the null hypothesis was true, the test statistic T would follow a t distribution with df = n - 1 = 29 degrees of freedom. We can draw a picture of this distribution and mark the observed T, as in Figure 5.18. The shaded right tail represents the p-value: the probability of observing such strong evidence in favor of the SAT company's claim, if the average student improvement is really only 100.

[18]This is a random sample from less than 10% of the company's students (assuming they have more than 300 former students), so the independence condition is reasonable. The normality condition also seems reasonable based on Figure 5.17. We can use the t distribution method. Note that we could use the normal distribution. However, since the sample size (n = 30) just meets the threshold for reasonably estimating the standard error, it is advisable to use the t distribution.

Exercise 5.25 Use the t table in Appendix B.2 on page 410 to identify the p-value. What do you conclude?[19]

Exercise 5.26 Because we rejected the null hypothesis, does this mean that taking the company's class improves student scores by more than 100 points on average?[20]

8.1.2: Paired Data

Are textbooks actually cheaper online? Here we compare the price of textbooks at UCLA's bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses (when a class had multiple books, only the most expensive text was considered). A portion of this data set is shown in Table $8.1.2.1$.

Table 8.1.2.1: Six cases of the textbooks data set.

	dept	course	ucla	amazon	diff
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

Paired Observations and Samples

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

Paired data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the textbook data set, we look at the difference in prices, which is represented as the diff variable in the textbooks data. Here the differences are taken as

$$\text{UCLA price} - \text{Amazon price} \tag{8.1.2.1}$$

for each book. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure $8.1.2.1$. Using differences between paired observations is a common and useful way to analyze paired data.

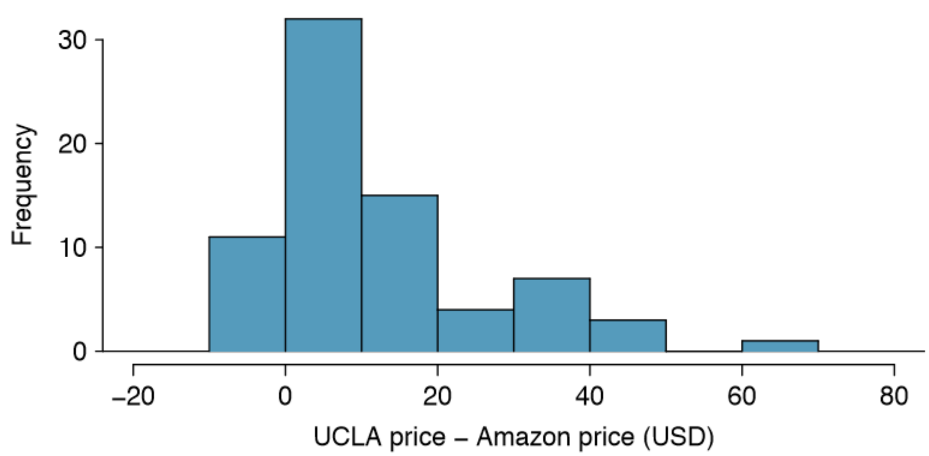


Figure 8.1.2.1: Histogram of the difference in price for each book sampled. These data are strongly skewed.

Exercise 8.1.2.1

The first difference shown in Table 8.1.2.1 is computed as 27.67 - 27.95 = -0.28. Verify the differences are calculated correctly for observations 2 and 3.

Solution

- Observation 2: 40.59 - 31.14 = 9.45.
- Observation 3: 31.68 - 32.00 = -0.32.

Inference for Paired Data

To analyze a paired data set, we use the exact same tools that we developed in Chapter 4. Now we apply them to the differences in the paired observations.

Table 8.1.2.1: Summary statistics for the price differences. There were 73 books, so there are 73 differences.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
73	12.76	14.26

Example 8.1.2.1: UCLA vs. Amazon

Set up and implement a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price.

Solution

There are two scenarios: there is no difference or there is some difference in average prices. The no difference scenario is always the null hypothesis:

- H_0: μ_{diff} = 0. There is no difference in the average textbook price.
- H_A: $\mu_{diff} \neq$ 0. There is a difference in average prices.

Can the normal model be used to describe the sampling distribution of $\bar{x}_{diff}$? We must check that the differences meet the conditions established in Chapter 4. The observations are based on a simple random sample from less than 10% of all books sold at the bookstore, so independence is reasonable; there are more than 30 differences; and the distribution of differences, shown in Figure 8.1.2.1, is strongly skewed, but this amount of skew is reasonable for this sized data set (n = 73). Because all three conditions are reasonably satisfied, we can conclude the sampling distribution of $\bar{x}_{diff}$ nearly normal and our estimate of the standard error will be reasonable.

We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences (s_{diff} = 14.26) and the number of differences (n_{diff} = 73):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{14.26}{\sqrt{73}} = 1.67 \tag{8.1.2.2}$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, which is shown in Figure 8.1.2.1. The p-value is represented by the two (very) small tails.

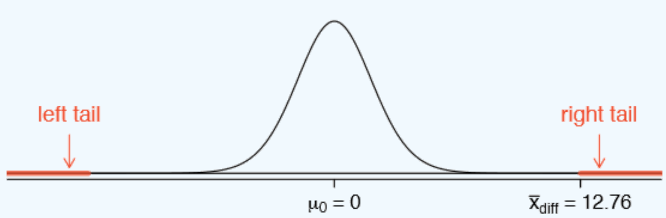


Figure 8.1.2.2: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

To find the tail areas, we compute the test statistic, which is the Z score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$Z = \frac{\bar{x}_{diff} - 0}{SE_{\bar{x}_{diff}}} = \frac{12.76 - 0}{1.67} = 7.59 \tag{8.1.2.3}$$

This Z score is so large it is not even in the table, which ensures the single tail area will be 0.0002 or smaller. Since the p-value corresponds to both tails in this case and the normal distribution is symmetric, the p-value can be estimated as twice the one-tail area:

$$\text{p-value} = 2 \times (\text{one tail area}) \approx 2 \times 0.0002 = 0.0004 \tag{8.1.2.4}$$

Because the p-value is less than 0.05, we reject the null hypothesis. We have found convincing evidence that Amazon is, on average, cheaper than the UCLA bookstore for UCLA course textbooks.

Exercise 8.1.2.1

Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.

Solution

Conditions have already verified and the standard error computed in Example $8.1.2.1$ To find the interval, identify z^* (1.96 for 95% confidence) and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \pm z^* SE \rightarrow 12.76 \pm 1.96 \times 1.67 \rightarrow (9.49, 16.03) \tag{8.1.2.5}$$

We are 95% confident that Amazon is, on average, between \$9.49 and \$16.03 cheaper than the UCLA bookstore for UCLA course books.

8.1.3: Difference of Two Means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. The methods are similar in theory but different in the details. Just as with a single sample, we identify conditions to ensure a point estimate of the difference $\bar{x}_1 - \bar{x}_2$ is nearly normal. Next we introduce a formula for the standard error, which allows us to apply our general tools from Section 4.5.

We apply these methods to two examples: participants in the 2012 Cherry Blossom Run and newborn infants. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

Point Estimates and Standard Errors for Differences of Means

We would like to estimate the average difference in run times for men and women using the run10Samp data set, which was a simple random sample of 45 men and 55 women from all runners in the 2012 Cherry Blossom Run. Table $8.1.3.2$presents relevant summary statistics, and box plots of each sample are shown in Figure 5.6.

Table $8.1.3.2$: Summary statistics for the run time of 100 participants in the 2009 Cherry Blossom Run.

	men	women
$\bar{x}$	87.65	102.13
s	12.5	15.2
n	45	55

The two samples are independent of one-another, so the data are not paired. Instead a point estimate of the difference in average 10 mile times for men and women, $\mu_w - \mu_m$, can be found using the two sample means:

$$\bar{x}_w - \bar{x}_m = 102.13 - 87.65 = 14.48 \tag{8.1.3.1}$$

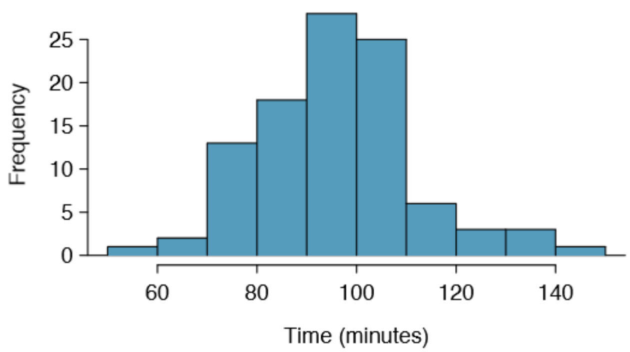


Figure $8.1.3.1$: *A histogram of time for the sample Cherry Blossom Race data.*

Because we are examining two simple random samples from less than 10% of the population, each sample contains at least 30 observations, and neither distribution is strongly skewed, we can safely conclude the sampling distribution of each sample mean is nearly normal. Finally, because each sample is independent of the other (e.g. the data are not paired), we can conclude that the difference in sample means can be modeled using a normal distribution. (Probability theory guarantees that the difference of two independent normal random variables is also normal. Because each sample mean is nearly normal and observations in the samples are independent, we are assured the difference is also nearly normal.)

Conditions for normality of $\bar{x}_1 - \bar{x}_2$

If the sample means, $\bar{x}_1$ and $\bar{x}_2$, each meet the criteria for having nearly normal sampling distributions and the observations in the two samples are independent, then the difference in sample means, $\bar{x}_1 - \bar{x}_2$, will have a sampling distribution that is nearly normal.

We can quantify the variability in the point estimate, $\bar{x}_w - \bar{x}_m$, using the following formula for its standard error:

$$SE_{\bar{x}_w - \bar{x}_m} = \sqrt{\frac{\sigma_w^2}{n_w} + \frac{\sigma_m^2}{n_m}} \tag{8.1.3.2}$$

We usually estimate this standard error using standard deviation estimates based on the samples:

$$SE_{\bar{x}_w - \bar{x}_m} \approx \sqrt{\frac{s_w^2}{n_w} + \frac{s_m^2}{n_m}} \tag{8.1.3.3}$$

$$= \sqrt{\frac{15.2^2}{55} + \frac{12.5^2}{45}} \tag{8.1.3.4}$$

$$= 2.77 \tag{8.1.3.5}$$

Because each sample has at least 30 observations ($n_w = 55$ and $n_m = 45$), this substitution using the sample standard deviation tends to be very good.

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, is nearly normal with mean $\mu_1 - \mu_2$ and estimated standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{8.1.3.6}$$

when each sample mean is nearly normal and all observations are independent.

Confidence Interval for the Difference

When the data indicate that the point estimate $\bar{x}_1 - \bar{x}_2$ comes from a nearly normal distribution, we can construct a confidence interval for the difference in two means from the framework built in Chapter 4. Here a point estimate, $\bar{x}_w - \bar{x}_m = 14.48$, is associated with a normal model with standard error SE = 2.77. Using this information, the general confidence interval formula may be applied in an attempt to capture the true difference in means, in this case using a 95% confidence level:

$$\text{point estimate} \pm z^* SE \rightarrow 14.48 \pm 1.96 \times 2.77 = (9.05, 19.91) \tag{8.1.3.7}$$

Based on the samples, we are 95% confident that men ran, on average, between 9.05 and 19.91 minutes faster than women in the 2012 Cherry Blossom Run.

Exercise 8.1.3.1

What does 95% confidence mean?

Solution

If we were to collected many such samples and create 95% confidence intervals for each, then about 95% of these intervals would contain the population difference, $\mu_w - \mu_m$.

Exercise 8.1.3.2

We may be interested in a different confidence level. Construct the 99% confidence interval for the population difference in average run times based on the sample data.

Solution

The only thing that changes is z*: we use z* = 2:58 for a 99% confidence level. (If the selection of z^* is confusing, see Section 4.2.4 for an explanation.) The 99% confidence interval:

$$14.48 \pm 2.58 \times 2.77 \rightarrow (7.33, 21.63). \tag{8.1.3.8}$$

We are 99% confident that the true difference in the average run times between men and women is between 7.33 and 21.63 minutes.

Hypothesis tests Based on a Difference in Means

A data set called baby smoke represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table $8.1.3.2$ We are particularly interested in two variables: weight and smoke. The weight variable represents the weights of the newborns and the smoke variable describes which mothers smoked during pregnancy. We would like to know if there is convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure $8.1.3.2$

Table $8.1.3.2$: Four cases from the baby smoke data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	⋮
150	45	50	36	9.25	female	nonsmoker

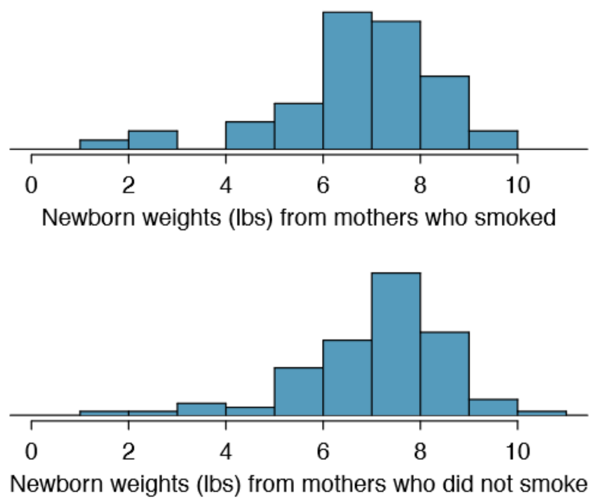


Figure $8.1.3.2$: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. Both distributions exhibit strong skew.

Example $8.1.3.1$

Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

Solution

The null hypothesis represents the case of no difference between the groups.

- H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.
- H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

Summary statistics are shown for each sample in Table $8.1.3.3$ Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, each group's sample size is at least

30 and the skew in each sample distribution is strong (Figure $8.1.3.2$). However, this skew is reasonable for these sample sizes of 50 and 100. Therefore, each sample mean is associated with a nearly normal distribution.

Table $8.1.3.3$: Summary statistics for the baby smoke data set.

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Exercise $8.1.3.3$

a. What is the point estimate of the population difference, $\mu_n - \mu_s$?
b. Can we use a normal distribution to model this difference?
c. Compute the standard error of the point estimate from part (a)

Solution

(a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$.

(b) Because the samples are independent and each sample mean is nearly normal, their difference is also nearly normal.

(c) The standard error of the estimate can be estimated using Equation 8.1.3.6:

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26 \tag{8.1.3.9}$$

The standard error estimate should be sufficiently accurate since the conditions were reasonably satisfied.

Example $8.1.3.2$

If the null hypothesis from Exercise 5.8 was true, what would be the expected value of the point estimate? And the standard deviation associated with this estimate? Draw a picture to represent the p-value.

Solution

If the null hypothesis was true, then we expect to see a difference near 0. The standard error corresponds to the standard deviation of the point estimate: 0.26. To depict the p-value, we draw the distribution of the point estimate as though H_0 was true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.

Example $8.1.3.3$

Compute the p-value of the hypothesis test using the figure in Example 5.9, and evaluate the hypotheses using a signi cance level of $\alpha = 0.05$.

Solution

Since the point estimate is nearly normal, we can nd the upper tail using the Z score and normal probability table:

$$Z = \frac{0.40 - 0}{0.26} = 1.54 \rightarrow \text{upper tail} = 1 - 0.938 = 0.062 \tag{8.1.3.10}$$

Because this is a two-sided test and we want the area of both tails, we double this single tail to get the p-value: 0.124. This p-value is larger than the signi cance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

Exercise 8.1.3.4

Does the conclusion to Example 5.10 mean that smoking and average birth weight are unrelated?

Solution

Absolutely not. It is possible that there is some difference but we did not detect it. If this is the case, we made a Type 2 Error.

Exercise 8.1.3.5

If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect such a difference?

Solution

We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists.

Summary for inference of the difference of two means

When considering the difference of two means, there are two common cases: the two samples are paired or they are independent. (There are instances where the data are neither paired nor independent.) The paired case was treated in Section 5.1, where the one-sample methods were applied to the differences from the paired observations. We examined the second and more complex scenario in this section.

When applying the normal model to the point estimate $\bar{x}_1 - \bar{x}_2$ (corresponding to unpaired data), it is important to verify conditions before applying the inference framework using the normal model. First, each sample mean must meet the conditions for normality; these conditions are described in Chapter 4 on page 168. Secondly, the samples must be collected independently (e.g. not paired data). When these conditions are satisfied, the general inference tools of Chapter 4 may be applied.

For example, a confidence interval may take the following form:

$$\text{point estimate} \pm z^{*} SE \tag{8.1.3.11}$$

When we compute the confidence interval for $\mu_1 - \mu_2$, the point estimate is the difference in sample means, the value z^* corresponds to the confidence level, and the standard error is computed from Equation 8.1.3.6. While the point estimate and standard error formulas change a little, the framework for a confidence interval stays the same. This is also true in hypothesis tests for differences of means.

In a hypothesis test, we apply the standard framework and use the specific formulas for the point estimate and standard error of a difference in two means. The test statistic represented by the Z score may be computed as

$$Z = \frac{\text{point estimate - null value}}{SE} \tag{8.1.3.12}$$

When assessing the difference in two means, the point estimate takes the form $\bar{x}_1 - \bar{x}_2$, and the standard error again takes the form of Equation 8.1.3.6. Finally, the null value is the difference in sample means under the null hypothesis. Just as in Chapter 4, the test statistic Z is used to identify the p-value.

Examining the Standard Error Formula

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}} \tag{8.1.3.13}$$

where s_1 and n_1 represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{8.1.3.14}$$

This special relationship follows from probability theory.

Exercise 8.1.3.6

Prerequisite: Section 2.4. We can rewrite Equation 8.1.3.14in a different way:

$$SE^2_{\bar{x}_1-\bar{x}_2} = SE^2_{\bar{x}_1} + SE^2_{barx_2} \tag{8.1.3.15}$$

Explain where this formula comes from using the ideas of probability theory.[10]

8.1.4: Power Calculations for a Difference of Means (Special Topic)

It is also useful to be able to compare two means for small samples. For instance, a teacher might like to test the notion that two versions of an exam were equally difficult. She could do so by randomly assigning each version to students. If she found that the average scores on the exams were so different that we cannot write it off as chance, then she may want to award extra points to students who took the more difficult exam.

In a medical context, we might investigate whether embryonic stem cells can improve heart pumping capacity in individuals who have suffered a heart attack. We could look for evidence of greater heart health in the stem cell group against a control group.

In this section we use the t distribution for the difference in sample means. We will again drop the minimum sample size condition and instead impose a strong condition on the distribution of the data.

Sampling Distributions for the Difference in two Means

In the example of two exam versions, the teacher would like to evaluate whether there is convincing evidence that the difference in average scores between the two exams is not due to chance.

It will be useful to extend the t distribution method from Section 5.3 to apply to a difference of means:

$$\bar{x}_1 - \bar{x}_2 \tag{8.1.4.1}$$

as a point estimate for

$$\mu_1 - \mu_2 \tag{8.1.4.2}$$

Our procedure for checking conditions mirrors what we did for large samples in Section 5.2. First, we verify the small sample conditions (independence and nearly normal data) for each sample separately, then we verify that the samples are also independent. For instance, if the teacher believes students in her class are independent, the exam scores are nearly normal, and the students taking each version of the exam were independent, then we can use the t distribution for inference on the point estimate $\bar{x}_1 - \bar{x}_2$.

The formula for the standard error of $\bar{x}_1 - \bar{x}_2$., introduced in Section 5.2, also applies to small samples:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.27}$$

[19]*We use the row with 29 degrees of freedom. The value T = 2.39 falls between the third and fourth columns. Because we are looking for a single tail, this corresponds to a p-value between 0.01 and 0.025. The p-value is guaranteed to be less than 0.05 (the default signi cance level), so we reject the null hypothesis. The data provide convincing evidence to support the company's claim that student scores improve by more than 100 points following the class.*

[20]*This is an observational study, so we cannot make this causal conclusion. For instance, maybe SAT test takers tend to improve their score over time even if they don't take a special SAT class, or perhaps only the most motivated students take such SAT courses.*

Because we will use the t distribution, we will need to identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will apply in the examples and exercises.[21]

Using the t distribution for a difference in means

The t distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t distribution and (2) the samples are independent. We estimate the standard error of the difference of two means using Equation ??? .

Two Sample t test

Summary statistics for each exam version are shown in Table 5.19. The teacher would like to evaluate whether this difference is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Table 5.19: Summary statistics of scores for each exam version.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Exercise 8.1.4.1

Construct a two-sided hypothesis test to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, might be due to chance.

Solution

Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. HA: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

Exercise 8.1.4.1

To evaluate the hypotheses in Exercise 5.28 using the t distribution, we must first verify assumptions.

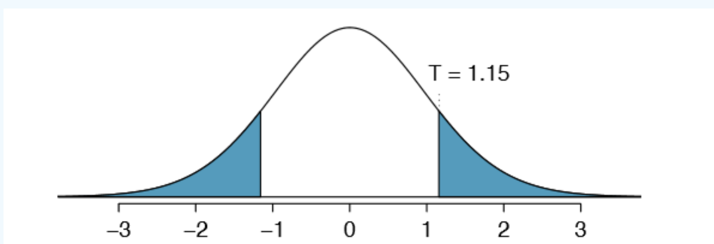
a. Does it seem reasonable that the scores are independent within each group?
b. What about the normality condition for each group?
c. Do you think scores from the two groups would be independent of each other (i.e. the two samples are independent)?[23]

Solution

(a) It is probably reasonable to conclude the scores are independent.

(b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable.

(c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.



After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated using Equation ???:

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62 \tag{8.1.4.3}$$

[21]*This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method.*

Figure 5.20: The t distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate - null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15 \tag{8.1.4.4}$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1 \text{and} n_2 - 1$: df = 26.

Exercise 8.1.4.1

Exercise 5.30 Identify the p-value, shown in Figure 5.20. Use df = 26.

Solution

We examine row df = 26 in the t table. Because this value is smaller than the value in the left column, the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

In Exercise 5.30, we could have used df = 45.97. However, this value is not listed in the table. In such cases, we use the next lower degrees of freedom (unless the computer also provides the p-value). For example, we could have used df = 45 but not df = 46.

Exercise 8.1.4.1

Do embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 5.21 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured. A positive value generally corresponds to increased pumping capacity, which suggests a stronger recovery.

a. Set up hypotheses that will be used to test whether there is convincing evidence that ESCs actually increase the amount of blood the heart pumps.
b. Check conditions for using the t distribution for inference with the point estimate $\bar{x}_1 - \bar{x}_2$. To assist in this assessment, the data are presented in Figure 5.22.[25]

Solution

(a) We first setup the hypotheses:

- H_0: The stem cells do not improve heart pumping function. $\mu_{esc} - \mu_{control} = 0$.
- H_A: The stem cells do improve heart pumping function. $\mu_{esc} - \mu_{control} > 0$.

(b) Because the sheep were randomly assigned their treatment and, presumably, were kept separate from one another, the independence assumption is reasonable for each sample as well as for between samples. The data are very limited, so we can only check for obvious outliers in the raw data in Figure 5.22. Since the distributions are (very) roughly symmetric, we will assume the normality condition is acceptable. Because the conditions are satisfied, we can apply the t distribution.

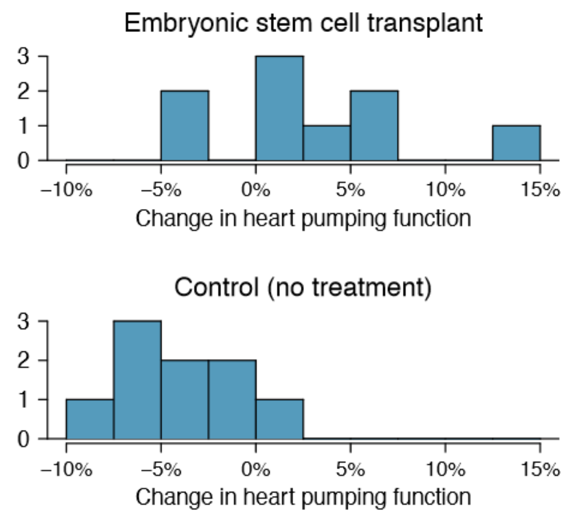


Figure 5.22: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

Table 5.21: Summary statistics of scores, split by exam version.

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Figure 5.23: Distribution of the sample difference of the test statistic if the null hypothesis was true. The shaded area, hardly visible in the right tail, represents the p-value.

Example 8.1.4.1

Use the data from Table 5.21 and df = 8 to evaluate the hypotheses for the ESC experiment described in Exercise 5.31.

Solution

First, we compute the sample difference and the standard error for that point estimate:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.88 \tag{8.1.4.5}$$

$$SE = \frac{\frac{5.17^2}{9} + \frac{2.76^2}{9}}{=} 1.95 \tag{8.1.4.6}$$

The p-value is depicted as the shaded slim right tail in Figure 5.23, and the test statistic is computed as follows:

$$T = \frac{7.88 - 0}{1.95} = 4.03 \tag{8.1.4.7}$$

We use the smaller of $n_1 - 1$ and $n_2 - 1$ (each are the same) for the degrees of freedom: df = 8. Finally, we look for T = 4.03 in the t table; it falls to the right of the last column, so the p-value is smaller than 0.005 (one tail!). Because the p-value is less than 0.005 and therefore also smaller than 0.05, we reject the null hypothesis. The data provide convincing evidence that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack.

Two sample t confidence interval

The results from the previous section provided evidence that ESCs actually help improve the pumping function of the heart. But how large is this improvement? To answer this question, we can use a confidence interval.

Exercise 8.1.4.1

In Exercise 5.31, you found that the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.88$, has a standard error of 1.95. Using df = 8, create a 99% confidence interval for the improvement due to ESCs.

Solution

We know the point estimate, 7.88, and the standard error, 1.95. We also veri ed the conditions for using the t distribution in Exercise 5.31. Thus, we only need identify $t*_8$ to create a 99% con dence interval: $t*_8 = 3.36$. The 99% con dence interval for the improvement from ESCs is given by

$$\text{point estimate} \pm t*_8 SE \rightarrow 7.88 \pm 3.36 \times 1.95 \rightarrow (1.33, 14.43) \tag{8.1.4.8}$$

That is, we are 99% con dent that the true improvement in heart pumping function is somewhere between 1.33% and 14.43%.

Pooled Standard Deviation Estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make our t distribution approach slightly more precise by using a pooled standard deviation. The pooled standard deviation of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s1 and s2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s^2_{pooled} = \frac{s^2_1 \times (n_1 - 1) + s^2_2 \times (n_2 - 1)}{n_1 + n_2 - 2} \tag{8.1.4.9}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} in place of s^2_1 and s^2_2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2 \tag{8.1.4.10}$$

The bene ts of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}^2_1 - \bar{x}^2_2$

Caution: Pooling standard deviations should be done only after careful research

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

8.1.5: Comparing many Means with ANOVA (Special Topic)

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually nd a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

- H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.
- H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

Example $8.1.5.1$

College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences infirstexam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

Solution

The hypotheses may be written in the following form:

- H_0: The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.
- H_A: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

Example $8.1.5.2$

Examine Figure $8.1.5.1$. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

Figure $8.1.5.1$*: Side-by-side dot plot for the outcomes for six groups.*

Solution

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large relative to the variability in the individual observations within each group.

Is Batting Performance Related to Player Position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: out elder (OF), in elder (IF), designated hitter (DH), and catcher (C). We will use a data set called bat10, which includes batting records of 327 Major League Baseball (MLB) players from the 2010 season. Six of the 327 cases represented in bat10 are shown in Table 8.1.5.1, and descriptions for each variable are provided in Table 5.26. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (OBP). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

Table 8.1.5.1: Six cases from the bat10 data matrix.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NTM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

Exercise 8.1.5.1

The null hypothesis under consideration is the following:

$$\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C. \tag{8.1.5.1}$$

Write the null and corresponding alternative hypotheses in plain language.

Solution

- H_0: The average on-base percentage is equal across the four positions.
- H_A: The average on-base

Table 8.1.5.2: Variables and their descriptions for the bat10 data set.

variable	description
name	Player name
team	The abbreviated name of the player's team
position	The player's primary eld position (OF, IF, DH, C)
AB	Number of opportunities at bat
H	Number of hits
HR	Number of home runs
RBI	Number of runs batted in
AVG	Batting average, which is equal to H=AB
OBP	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Example 8.1.5.3

The player positions have been divided into four groups: outfield (OF), infield (IF), designated hitter (DH), and catcher (C). What would be an appropriate point estimate of the batting average by out elders, μ_{OF}?

Solution

A good estimate of the batting average by out elders would be the sample average of AVG for just those players whose position is out field:

$$\bar{x}_{OF} = 0.334. \tag{8.1.5.2}$$

Table 8.1.5.3 provides summary statistics for each group. A side-by-side box plot for the batting average is shown in Figure 8.1.5.1. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

Table 8.1.5.3: Summary statistics of on-base percentage, split by player position.

	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean ($\bar{x}_i$)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045

percentage varies across some (or all) groups.

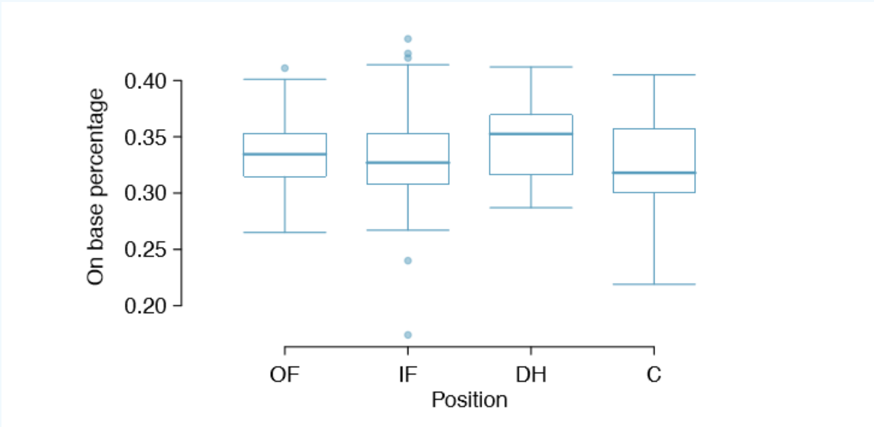


Figure 8.1.5.1: Side-by-side box plot of the on-base percentage for 327 players across four groups. There is one prominent outlier visible in the infield group, but with 154 observations in the in field group, this outlier is not a concern.

Example 8.1.5.1

The largest difference between the sample means is between the designated hitter and the catcher positions. Consider again the original hypotheses:

- H_0: $\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C$
- H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of μ_{DH} and μ_C is statistically significant at a 0.05 significance level?

Solution

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an ination in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional information on the ideas expressed in Example 5.38, we recommend reading about the *prosecutor's fallacy* (*See, for example, www.stat.columbia.edu/~cook/movabletype/archives/2007/05/the prosecutors.html*.)

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in means could have happened just by chance even if there was no difference in the respective population means.

Analysis of variance (ANOVA) and the F test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will simultaneously consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups** (MSG), and it has an associated degrees of freedom, dfG = k - 1 when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote,29 however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the mean square error (MSE), which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE are provided in the footnote[30] for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE} \tag{8.1.5.3}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

Exercise 8.1.5.1

For the baseball data, MSG = 0.00252 and MSE = 0.00127. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 1.994.

Solution

There are k = 4 groups, so $df_G = k - 1 = 3$. There are

$$n = n_1 + n_2 + n_3 + n_4 = 327 \tag{8.1.5.4}$$

total observations, so $df_E = n - k = 323$. Then the F statistic is computed as the ratio of MSG and MSE:

$$F = \frac{MSG}{MSE} = \frac{0.00252}{0.00127} = 1.984 \approx 1.994, \tag{8.1.5.5}$$

(F = 1.994 was computed by using values for MSG and MSE that were not rounded.)

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 3 and 323 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 5.29.

[29]*Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as*

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \tag{8.1.5.6}$$

where SSG is called the sum of squares between groups and n_i is the sample size of group i.

[30]*Let $\bar{x}$ represent the mean of outcomes across all groups. Then the sum of squares total (SST) is computed as*

$$SST = \sum_{i=1}^{n} (y_i - \bar{x})^2 \tag{8.1.5.7}$$

where the sum is over all observations in the data set. Then we compute the sum of squared errors (SSE) in one of two equivalent ways:

$$SSE = SST - SSG \tag{8.1.5.8}$$

$$= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \tag{8.1.5.9}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

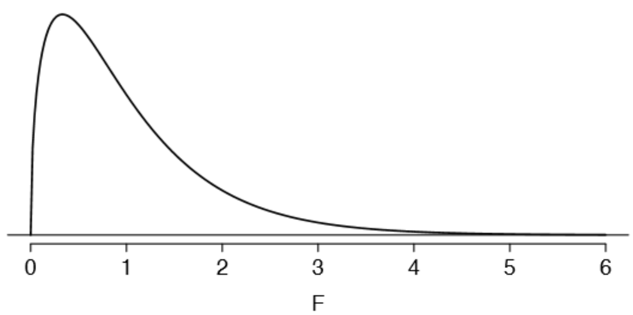


Figure 8.1.5.1*: An F distribution with $df_1 = 3$ and $df_2 = 323$.*

The larger the observed variability in the sample means (MSG) relative to the withingroup observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

The F statistic and the F test

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

Exercise 8.1.5.1

The test statistic for the baseball example is F = 1.994. Shade the area corresponding to the p-value in Figure 5.29. [32]

Example 8.1.5.1

A common method for preparing oxygen is the decomposition

Example 5.42 The p-value corresponding to the shaded area in the solution of Exercise 5.41 is equal to about 0.115. Does this provide strong evidence against the null hypothesis?

The p-value is larger than 0.05, indicating the evidence is not strong enough to reject the null hypothesis at a signi cance level of 0.05. That is, the data do not provide strong evidence that the average on-base percentage varies by player's primary field position.

Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we will see in Chapters 7 and 8. Table 5.30 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F test statistic and p-value can be retrieved from the last columns.

Table 5.30: ANOVA summary for testing whether the average on-base percentage differs across player positions.

	DF	Sum Sq	Mean Sq	F value	Pr(> F)
position	3	0.0076	0.0025	1.9943	0.1147
Residuals	323	0.4080	0.0013		

Graphical Diagnostics for an ANOVA Analysis

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

- **Independence**. If the data are a simple random sample from less than 10% of the population, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). For example, in the MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations.
- **Approximately normal**. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small. The normal probability plots for each group of the MLB data are shown in Figure 5.31; there is some deviation from normality for in elders, but this isn't a substantial concern since there are about 150 observations in that group and the outliers are not extreme. Sometimes in ANOVA there are so many groups or so few observations per group that checking normality for each group isn't reasonable. See the footnote33 for guidance on how to handle such instances.
- **Constant variance**. The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a sideby-side box plot of the outcomes across the groups, as in Figure 5.28 on page 239. In this case, the variability is similar in the four groups but not identical. We see in Table 5.27 on page 238 that the standard deviation varies a bit from one group to the next. Whether these differences are from natural variation is unclear, so we should report this uncertainty with the nal results.

[33]First calculate the residuals of the baseball data, which are calculated by taking the observed values and subtracting the corresponding group means. For example, an out elder with OBP of 0.435 would have a residual of $0.405 - \bar{x}_{OF} = 0.071$. Then to check the normality condition, create a normal probability plot using all the residuals simultaneously.

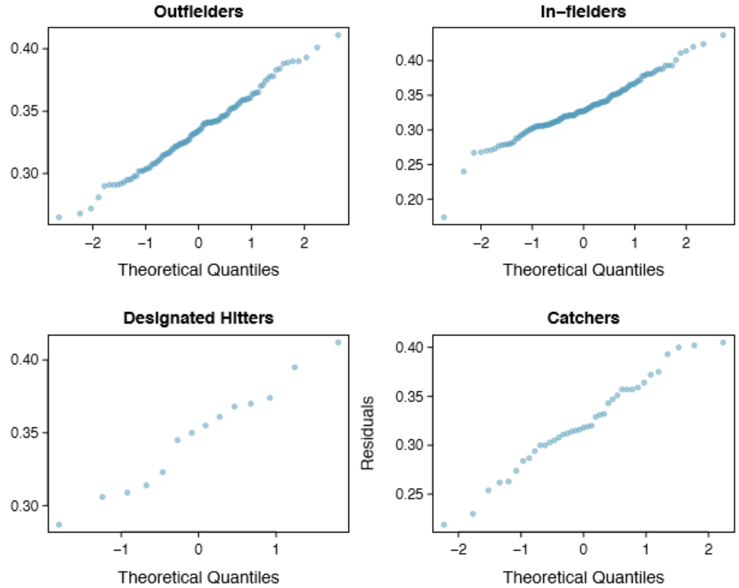


Figure 5.31: Normal probability plot of OBP for each field position.

Caution: Diagnostics for an ANOVA analysis

Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t test, but we use a modi ed signi cance level and a pooled estimate of the standard deviation across groups. Usually this pooled standard deviation can be found in the ANOVA table, e.g. along the bottom of Table 5.30.

Table 5.32: Summary statistics for the first midterm scores in three different lectures of the same course.

Class i	A	B	C
n_i $\bar{x}_i$ s_i	58 75.1 13.9	55 72.0 13.8	51 78.9 13.1

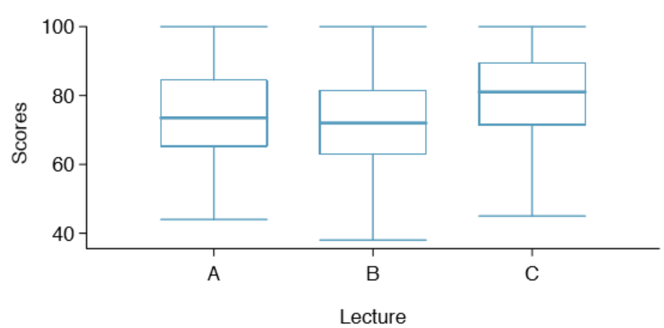


Figure 5.33: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

Example 8.1.5.1

A common method for preparing oxygen is the decomposition

Example 5.43 Example 5.34 on page 236 discussed three statistics lectures, all taught during the same semester. Table 5.32 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 5.33. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. No such situations were evident for these particular data, and we believe that independence is acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be veri ed in Table 5.32, supporting the constant variance assumption.

Exercise 8.1.5.1

A common method for preparing oxygen is the decompositio

Exercise 5.44 An ANOVA was conducted for the midterm data, and summary results are shown in Table 5.34. What should we conclude?[34]

[34]The p-value of the test is 0.0330, less than the default signi cance level of 0.05. Therefore, we reject the null hypothesis and conclude that the difference in the average midterm scores are not due to chance.

Table 5.34: ANOVA summary table for the midterm data.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture Residuals	2 161	1290.11 29810.13	645.06 185.16	3.48	0.0330

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 5.38 on page 238: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modi ed signi cance level.

Multiple comparisons and the Bonferroni correction for

The scenario of testing many pairs of groups is called multiple comparisons. The Bonferroni correction suggests that a more stringent signi cance level is more appropriate for these tests:

$$\alpha^* = \frac{\alpha}{K} \tag{8.1.5.10}$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Example 5.45 In Exercise 5.44, you found strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modi ed signi cance level of $\alpha^* = \frac{0.05}{3} = 0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled} = 13.61$ on df = 161, which is provided in the ANOVA summary table.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{x}_A - \bar{x}_B = 75.1 - 72 = 3.1 SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56 \tag{8.1.5.11}$$

(See Section 5.4.4 on page 235 for additional details.) This results in a T score of 1.21 on df = 161 (we use the df associated with spooled). Statistical software was used to precisely identify the two-tailed p-value since the modi ed signi cance of 0.0167 is not found in the t table. The p-value (0.228) is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on df = 161 and a two-tailed p-value of 0.1462. This p-value is larger than α^*, so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on df = 161 and a two-tailed p-value of 0.0102. This p-value is smaller than α^*. Here we nd strong evidence of a difference in the means of lectures B and C.

We might summarize the ndings of the analysis from Example 5.45 using the following notation:

$$\mu_A \stackrel{?}{=} \mu_B \mu_A \stackrel{?}{=} \mu_C \mu_B \neq \mu_C \tag{8.1.5.12}$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

Caution: Sometimes an ANOVA will reject the null but no groups will have statistically signi cant differences

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, this does not invalidate the ANOVA conclusion. It only means we have not been able to successfully identify which groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high con dence a specific difference as statistically signi cant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

8.1.6: Exercises

Paired data

5.1 Global warming, Part I. Is there strong evidence of global warming? Let's consider a small scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. Then the difference between the two readings (temperature in 2008 - temperature in 1968) was calculated for each of the 51 different locations. The average of these 51 values was 1.1 degrees with a standard deviation of 4.9 degrees. We are interested in determining whether these data provide strong evidence of temperature warming in the continental US.

1. (a) Is there a relationship between the observations collected in 1968 and 2008? Or are the observations in the two groups independent? Explain.
2. (b) Write hypotheses for this research in symbols and in words.
3. (c) Check the conditions required to complete this test.
4. (d) Calculate the test statistic and nd the p-value.
5. (e) What do you conclude? Interpret your conclusion in context.
6. (f) What type of error might we have made? Explain in context what the error means.
7. (g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the temperature measurements from 1968 and 2008 to include 0? Explain your reasoning.

5.2 High School and Beyond, Part I. The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

1. (a) Is there a clear difference in the average reading and writing scores?
2. (b) Are the reading and writing scores of each student independent of each other?
3. (c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?
4. (d) Check the conditions required to complete this test.
5. (e) The average observed difference in scores is $\bar{x}_{\text{read-write}} = -0.545$, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?
6. (f) What type of error might we have made? Explain what the error means in the context of the application.
7. (g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

5.3 Global warming, Part II. We considered the differences between the temperature readings in January 1 of 1968 and 2008 at 51 locations in the continental US in Exercise 5.1. The mean and standard deviation of the reported differences are 1.1 degrees and 4.9 degrees.

1. (a) Calculate a 90% confidence interval for the average difference between the temperature measurements between 1968 and 2008.
2. (b) Interpret this interval in context.
3. (c) Does the confidence interval provide convincing evidence that the temperature was higher in 2008 than in 1968 in the continental US? Explain.

5.4 High school and beyond, Part II. We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey in Exercise 5.3. The mean and standard deviation of the differences are $\bar{x}_{\text{read-write}} = -0.545$ and 8.887 points.

1. (a) Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.
2. (b) Interpret this interval in context.
3. (c) Does the confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

5.5 Gifted children. Researchers collected a simple random sample of 36 children who had been identi ed as gifted in a large city. The following histograms show the distributions of the IQ scores of mothers and fathers of these children. Also provided are some sample statistics.[35]

1. (a) Are the IQs of mothers and the IQs of fathers in this data set related? Explain.
2. (b) Conduct a hypothesis test to evaluate if the scores are equal on average. Make sure to clearly state your hypotheses, check the relevant conditions, and state your conclusion in the context of the data.

5.6 Paired or not? In each of the following scenarios, determine if the data are paired.

1. (a) We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To nd out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock.
2. (b) We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.
3. (c) A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

[35]*F.A. Graybill and H.K. Iyer. Regression Analysis: Concepts and Applications. Duxbury Press, 1994, pp. 511-516.*

Difference of two means

5.7 Math scores of 13 year olds, Part I. The National Assessment of Educational Progress tested a simple random sample of 1,000 thirteen year old students in both 2004 and 2008 (two separate simple random samples). The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively. Calculate a 90% confidence interval for the change in average scores from 2004 to 2008, and interpret this interval in the context of the application. (Reminder: check conditions.)[36]

5.8 Work hours and education, Part I. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. The histograms below display the distributions of hours worked per week for two education groups: those with and without a college degree.37 Suppose we want to estimate the average difference between the number of hours worked per week by all Americans with a college degree and those without a college degree. Summary information for each group is shown in the tables.

1. (a) What is the parameter of interest, and what is the point estimate?
2. (b) Are conditions satisfied for estimating this difference using a confidence interval?
3. (c) Create a 95% confidence interval for the difference in number of hours worked between the two groups, and interpret the interval in context.
4. (d) Can you think of any real world justi cation for your results? (Note: There isn't a single correct answer to this question.)

5.9 Math scores of 13 year olds, Part II. Exercise 5.7 provides data on the average math scores from tests conducted by the National Assessment of Educational Progress in 2004 and 2008. Two separate simple random samples were taken in each of these years. The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively.

1. (a) Do these data provide strong evidence that the average math score for 13 year old students has changed from 2004 to 2008? Use a 10% signi cance level.
2. (b) It is possible that your conclusion in part (a) is incorrect. What type of error is possible for this conclusion? Explain.
3. (c) Based on your hypothesis test, would you expect a 90% confidence interval to contain the null value? Explain.

[36]*National Center for Education Statistics, NAEP Data Explorer.*

[37]*National Opinion Research Center, General Social Survey, 2010.*

5.10 Work hours and education, Part II. The General Social Survey described in Exercise 5.8 included random samples from two groups: US residents with a college degree and US residents without a college degree. For the 505 sampled US residents with a college degree, the average number of hours worked each week was 41.8 hours with a standard deviation of 15.1 hours. For those 667 without a degree, the mean was 39.4 hours with a standard deviation of 15.1 hours. Conduct a hypothesis test to check for a difference in the average number of hours worked for the two groups.

5.11 Does the Paleo diet work? The Paleo diet allows only for foods that humans typically consumed over the last 2.5 million years, excluding those agriculture-type foods that arose during the last 10,000 years or so. Researchers randomly divided 500 volunteers into two equal-sized groups. One group spent 6 months on the Paleo diet. The other group received a pamphlet about controlling portion sizes. Randomized treatment assignment was performed, and at the beginning of the study, the average difference in weights between the two groups was about 0. After the study, the Paleo group had lost on average 7 pounds with a standard deviation of 20 pounds while the control group had lost on average 5 pounds with a standard deviation of 12 pounds.

1. (a) The 95% confidence interval for the difference between the two population parameters (Paleo - control) is given as (-0.891, 4.891). Interpret this interval in the context of the data.
2. (b) Based on this confidence interval, do the data provide convincing evidence that the Paleo diet is more effective for weight loss than the pamphlet (control)? Explain your reasoning.
3. (c) Without explicitly performing the hypothesis test, do you think that if the Paleo group had lost 8 instead of 7 pounds on average, and everything else was the same, the results would then indicate a signi cant difference between the treatment and control groups? Explain your reasoning.

5.12 Weight gain during pregnancy. In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relationship between habits and practices of expectant mothers and the birth of their children. The following histograms show the distributions of weight gain during pregnancy by 867 younger moms (less than 35 years old) and 133 mature moms (35 years old and over) who have been randomly sampled from this large data set. The average weight gain of younger moms is 30.56 pounds, with a standard deviation of 14.35 pounds, and the average weight gain of mature moms is 28.79 pounds, with a standard deviation of 13.48 pounds. Calculate a 95% confidence interval for the difference between the average weight gain of younger and mature moms. Also comment on whether or not this interval provides strong evidence that there is a signi cant difference between the two population means.

5.13 Body fat in women and men. The third National Health and Nutrition Examination Survey collected body fat percentage (BF) data from 13,601 subjects whose ages are 20 to 80. A summary table for these data is given below. Note that BF is given as mean $\pm$ standard error. Construct a 95% confidence interval for the difference in average body fat percentages between men and women, and explain the meaning of this interval.[38]

Gender	n	BF (%)
Men Women	6,580 7,021	23.9 ± 0.07 35.0 ± 0.09

5.14 Child care hours, Part I. The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments. One of the variables collected on the survey is the number of hours parents spend taking care of children in their household under age 6 (feeding, bathing, dressing, holding, or watching them). In 2006, 487 females and 312 males were surveyed for this question. On average, females reported spending 31 hours with a standard deviation of 31 hours, and males reported spending 16 hours with a standard deviation of 21 hours. Calculate a 95% confidence interval for the difference between the average number of hours Chinese males and females spend taking care of their children under age 6. Also comment on whether this interval suggests a significant difference between the two population parameters. You may assume that conditions for inference are satisfied.[39]

One-sample means with the t distribution

5.15 Identify the critical t. An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t value (t?) for the given sample size and confidence level.

1. (a) n = 6, CL = 90%
2. (b) n = 21, CL = 98%
3. (c) n = 29, CL = 95%
4. (d) n = 12, CL = 99%

5.16 Working backwards, Part I. A 90% confidence interval for a population mean is (65,77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

5.17 Working backwards, Part II. A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). This confidence interval is based on a simple random sample of 36 observations. Calculate the sample mean and standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t distribution in any calculations.

5.18 Find the p-value. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given set of hypotheses and T test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

1. (a) $H_A : \mu > \mu_0, n = 11, T = 1.91$
2. (b) $H_A : \mu < \mu_0, n = 17, T = -3.45$
3. (c) $H_A : \mu \neq \mu_0, n = 7, T = 0.83$
4. (d) $H_A : \mu > \mu_0, n = 28, T = 2.13$

[38] *A Romero-Corral et al. "Accuracy of body mass index in diagnosing obesity in the adult general population". In: International Journal of Obesity 32.6 (2008), pp. 959-966.*

[39] *UNC Carolina Population Center, China Health and Nutrition Survey, 2006.*

5.19 Sleep habits of New Yorkers. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. Do these data provide strong evidence that New Yorkers sleep less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

1. (a) Write the hypotheses in symbols and in words.
2. (b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.
3. (c) Find and interpret the p-value in this context. Drawing a picture may be helpful.
4. (d) What is the conclusion of the hypothesis test?
5. (e) If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

5.20 Fuel efficiency of Prius. Fueleconomy.gov, the official US government source for fuel economy information, allows users to share gas mileage information on their vehicles. The histogram below shows the distribution of gas mileage in miles per gallon (MPG) from 14 users who drive a 2012 Toyota Prius. The sample mean is 53.3 MPG and the standard deviation is 5.2 MPG. Note that these data are user estimates and since the source data cannot be veri ed, the accuracy of these estimates are not guaranteed.[40]

1. (a) We would like to use these data to evaluate the average gas mileage of all 2012 Prius drivers. Do you think this is reasonable? Why or why not?
2. (b) The EPA claims that a 2012 Prius gets 50 MPG (city and highway mileage combined). Do these data provide strong evidence against this estimate for drivers who participate on fueleconomy.gov? Note any assumptions you must make as you proceed with the test.
3. (c) Calculate a 95% confidence interval for the average gas mileage of a 2012 Prius by drivers who participate on fueleconomy.gov.

5.21 Find the mean. You are given the following hypotheses:

- $H_0 : \mu = 60$
- $H_A : \mu < 60$

We know that the sample standard deviation is 8 and the sample size is 20. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

5.22 t* vs. z*. For a given confidence level, t* df is larger than z*. Explain how \(t^*_{df} being slightly larger than z* affects the width of the confidence interval.

[40] *Fuelecomy.gov, Shared MPG Estimates: Toyota Prius 2012.*

The t distribution for the difference of two means

5.23 Cleveland vs. Sacramento. Average income varies from one region of the country to another, and it often reects both lifestyles and regional living expenses. Suppose a new graduate is considering a job in two locations, Cleveland, OH and Sacramento, CA, and he wants to see whether the average income in one of these cities is higher than the other. He would like to conduct a t test based on two small samples from the 2000 Census, but he first must consider whether the conditions are met to implement the test. Below are histograms for each city. Should he move forward with the t test? Explain your reasoning.

5.24 Oscar winners. The rst Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2012. Summary statistics for these distributions are also provided. Is a t test appropriate for evaluating whether the difference in the average ages of best actors and actresses might be due to chance? Explain your reasoning.[41]

[41]Oscar winners from 1929 - 2012, data up to 2009 from the Journal of Statistics Education data archive and more current data from Wikipedia.org.

5.25 Friday the 13th, Part I. In the early 1990's, researchers in the UK collected data on traffic ow, number of shoppers, and traffic accident related emergency room admissions on Friday the 13th and the previous Friday, Friday the 6th. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6th and Friday the 13th for many such date pairs. Also given are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th.[42]

	6 th	13 th	Diff.
$\bar{x}$	128,385	126,550	1,835
s	7,259	7,664	1,176
n	10	10	10

1. (a) Are there any underlying structures in these data that should be considered in an analysis? Explain.
2. (b) What are the hypotheses for evaluating whether the number of people out on Friday the 6th is different than the number out on Friday the 13th?
3. (c) Check conditions to carry out the hypothesis test from part (b).
4. (d) Calculate the test statistic and the p-value.
5. (e) What is the conclusion of the hypothesis test?
6. (f) Interpret the p-value in this context.
7. (g) What type of error might have been made in the conclusion of your test? Explain.

5.26 Diamonds, Part I. Prices of diamonds are determined by what is known as the 4 Cs: cut, clarity, color, and carat weight. The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and compare the average prices of the diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. For a 1 carat diamond, we divide the price by 100. The distributions and some sample statistics are shown below.[43]

	0.99 carats	1 carat
Men	$ 44.51	$ 56.81
SD	$ 13.32	$ 16.13
n	23	23

Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.

[42]T.J. Scanlon et al. "Is Friday the 13th Bad For Your Health?" In: BMJ 307 (1993), pp. 1584-1586.

[43]H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

5.27 Friday the 13th, Part II. The Friday the 13th study reported in Exercise 5.25 also provides data on traffic accident related emergency room admissions. The distributions of these counts from Friday the 6th and Friday the 13th are shown below for six such paired dates along with summary statistics. You may assume that conditions for inference are met.

1. (a) Conduct a hypothesis test to evaluate if there is a difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

2. (b) Calculate a 95% confidence interval for the difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.
3. (c) The conclusion of the original study states, "Friday 13th is unlucky for some. The risk of hospital admission as a result of a transport accident may be increased by as much as 52%. Staying at home is recommended." Do you agree with this statement? Explain your reasoning.

5.28 Diamonds, Part II. In Exercise 5.26, we discussed diamond prices (standardized by weight) for diamonds with weights 0.99 carats and 1 carat. See the table for summary statistics, and then construct a 95% confidence interval for the average difference between the standardized prices of 0.99 and 1 carat diamonds. You may assume the conditions for inference are met.

	0.99 carats	1 carat
Men SD n	\$ 44.51 \$ 13.32 23	\$ 56.81 \$ 16.13 23

5.29 Chicken diet and weight, Part I. Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company pro ts, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Below are some summary statistics from this data set along with box plots showing the distribution of weights by feed type.[44]

1. (a) Describe the distributions of weights of chickens that were fed linseed and horsebean.
2. (b) Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different? Use a 5% significance level.
3. (c) What type of error might we have committed? Explain.
4. (d) Would your conclusion change if we used α = 0.01?

5.30 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[45]

5.31 Chicken diet and weight, Part II. Casein is a common weight gain supplement for humans. Does it have an effect on chickens? Using data provided in Exercise 5.29, test the hypothesis that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean. If your hypothesis test yields a statistically significant result, discuss whether or not the higher average weight of chickens can be attributed to the casein diet. Assume that conditions for inference are satisfied.

[44] *Chicken Weights by Feed Type, from the datasets package in R.*

[45] *U.S. Department of Energy, Fuel Economy Data, 2012 Data le.*

5.32 Fuel efficiency of manual and automatic cars, Part II. The table provides summary statistics on highway fuel economy of cars manufactured in 2012 (from Exercise 5.30). Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.[46]

5.33 Gaming and distracted eating, Part I. A group of researchers are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption. To test this hypothesis, they monitored food intake for a group of 44 patients who were randomized into two equal groups. The treatment group ate lunch while playing solitaire, and the control group ate lunch without any added distractions. Patients in the treatment group ate 52.1 grams of biscuits, with a standard deviation of 45.1 grams, and patients in the control group ate 27.1 grams of biscuits, with a standard deviation of 26.4 grams. Do these data provide convincing evidence that the average food intake (measured in amount of biscuits consumed) is different for the patients in the treatment group? Assume that conditions for inference are satisfied.[47]

5.34 Gaming and distracted eating, Part II. The researchers from Exercise 5.33 also investigated the effects of being distracted by a game on how much people eat. The 22 patients in the treatment group who ate their lunch while playing solitaire were asked to do a serial-order recall of the food lunch items they ate. The average number of items recalled by the patients in this group was

4.9, with a standard deviation of 1.8. The average number of items recalled by the patients in the control group (no distraction) was 6.1, with a standard deviation of 1.8. Do these data provide strong evidence that the average number of food items recalled by the patients in the treatment and control groups are different?

5.35 Prison isolation experiment, Part I. Subjects from Central Prison in Raleigh, NC, volunteered for an experiment involving an \isolation" experience. The goal of the experiment was to nd a treatment that reduces subjects' psychopathic deviant T scores. This score measures a person's need for control or their rebellion against control, and it is part of a commonly used mental health test called the Minnesota Multiphasic Personality Inventory (MMPI) test. The experiment had three treatment groups:

1. (1) Four hours of sensory restriction plus a 15 minute "therapeutic" tape advising that professional help is available.
2. (2) Four hours of sensory restriction plus a 15 minute "emotionally neutral" tape on training hunting dogs.
3. (3) Four hours of sensory restriction but no taped message.

Forty-two subjects were randomly assigned to these treatment groups, and an MMPI test was administered before and after the treatment. Distributions of the differences between pre and

[46]U.S. Department of Energy, Fuel Economy Data, 2012 Data file.

[47]R.E. Oldham-Cooper et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake". In: The American Journal of Clinical Nutrition 93.2 (2011), p. 308.

post treatment scores (pre - post) are shown below, along with some sample statistics. Use this information to independently test the effectiveness of each treatment. Make sure to clearly state your hypotheses, check conditions, and interpret results in the context of the data.[48]

5.36 True or false, Part I. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

1. (a) When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal model for the difference in means since $n_2 \geq 30$.
2. (b) As the degrees of freedom increases, the T distribution approaches normality.
3. (c) We use a pooled standard error for calculating the standard error of the difference between means when sample sizes of groups are equal to each other.

Comparing many means with ANOVA

5.37 Chicken diet and weight, Part III. In Exercises 5.29 and 5.31 we compared the effects of two types of feed at a time. A better analysis would rst consider all feed types at once: casein, horsebean, linseed, meat meal, soybean, and sunower. The ANOVA output below can be used to test for differences between the average weights of chicks on different diets.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
feed	5	231,129.16	46,225.83	15.36	0.0000
Residuals	65	195,556.02	3,008.55		

Conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions. Figures and summary statistics are shown below.

5.38 Student performance across discussion sections. A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average nal exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
$\bar{x}_i$	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Section Residuals	7 189	525.01 7584.11	75.00 40.13	1.87	0.0767

Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to proceed with the test.

5.39 Coffee, depression, and physical activity. Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.[49]

	Caffeinated	coffee	consumption			
	$\leq 1 cup/week$	2-6 cus/week	1 cup/day	2-3 cups/day	$\geq 4 cups/day$	Total
Mean SD n	18.7 21.1 12,215	19.6 25.5 6,617	19.3 22.5 17,234	18.9 22.0 12,290	17.5 22.0 2,838	50,739

1. (a) Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.
2. (b) Check conditions and describe any assumptions you must make to proceed with the test.
3. (c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Section Residuals Total	--------------- --------------- ---------------	--------------- 25,564,819 25,575,327	---------------- ----------------	----------------	0.0003

(d) What is the conclusion of the test?

[49]*M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: Archives of internal medicine 171.17 (2011), p. 1571.*

5.40 Work hours and education, Part III. In Exercises 5.8 and 5.10 you worked with data from the General Social Survey in order to compare the average number of hours worked per week by US residents with and without a college degree. However, this analysis didn't take advantage of the original data which contained more accurate information on educational attainment (less than high school, high school, junior college, Bachelor's, and graduate school). Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once instead of re-categorizing them into two groups. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

		Educational	attainment			
	Less than HS	HS	Jr Coll	BAchelor's	Graduate	Total
Mean SD n	38.67 15.81 121	39.6 14.97 546	41.39 18.1 97	42.55 13.62 253	40.85 15.51 155	40.45 15.17 1,172

1. (a) Write hypotheses for evaluating whether the average number of hours worked varies across the ve groups.
2. (b) Check conditions and describe any assumptions you must make to proceed with the test.
3. (c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
degree	---------------	---------------	501.54	----------------	0.0682
Residuals	---------------	267,382	----------------		
Total	---------------	---------------			

(d) What is the conclusion of the test?

5.41 GPA and major. Undergraduate students taking an introductory statistics course at Duke University conducted a survey about GPA and major. The side-by-side box plots show the distribution of GPA among three groups of majors. Also provided is the ANOVA output.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
major	2	0.03	0.02	0.21	0.8068
Residuals	195	15.77	0.08		

1. (a) Write the hypotheses for testing for a difference between average GPA across majors.
2. (b) What is the conclusion of the hypothesis test?
3. (c) How many students answered these questions on the survey, i.e. what is the sample size?

5.42 Child care hours, Part II. Exercise 5.14 introduces the China Health and Nutrition Survey which, among other things, collects information on number of hours Chinese parents spend taking care of their children under age 6. The side by side box plots below show the distribution of this variable by educational attainment of the parent. Also provided below is the ANOVA output for comparing average hours across educational attainment categories.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
education	4	4142.09	1035.52	1.26	0.2846
Residuals	794	653047.83	822.48		

1. (a) Write the hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels.
2. (b) What is the conclusion of the hypothesis test?

5.43 True or false, Part II. Determine if the following statements are true or false in ANOVA, and explain your reasoning for statements you identify as false.

1. (a) As the number of groups increases, the modi ed signi cance level for pairwise tests increases as well.
2. (b) As the total sample size increases, the degrees of freedom for the residuals increases as well.
3. (c) The constant variance condition can be somewhat relaxed when the sample sizes are relatively consistent across groups.
4. (d) The independence assumption can be relaxed when the total sample size is large.

5.44 True or false, Part III. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

If the null hypothesis that the means of four groups are all the same is rejected using ANOVA at a 5% signi cance level, then ...

1. (a) we can then conclude that all the means are different from one another.
2. (b) the standardized variability between groups is higher than the standardized variability within groups.
3. (c) the pairwise analysis will identify at least one pair of means that are signi cantly different.
4. (d) the appropriate to be used in pairwise comparisons is $\frac{0.05}{4} = 0.0125$ since there are four groups.

5.45 Prison isolation experiment, Part II. Exercise 5.35 introduced an experiment that was conducted with the goal of identifying a treatment that reduces subjects' psychopathic deviant T scores, where this score measures a person's need for control or his rebellion against control. In Exercise 5.35 you evaluated the success of each treatment individually. An alternative analysis involves comparing the success of treatments. The relevant ANOVA output is given below.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)

treatment Residuals	2 39	639.48 3740.43	319.74 95.91	3.33	0.0461

1. (a) What are the hypotheses?
2. (b) What is the conclusion of the test? Use a 5% significance level.
3. (c) If in part (b) you determined that the test is signi cant, conduct pairwise tests to determine which groups are different from each other. If you did not reject the null hypothesis in part (b), recheck your solution.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

SECTION OVERVIEW

8.2: Inference for Categorical Data

Chapter 6 introduces inference in the setting of categorical data. We use these methods to answer questions like the following:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

We will find that the methods we learned in previous chapters are very useful in these settings. For example, sample proportions are well characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual confidence interval and hypothesis testing tools. In other instances, such as those with contingency tables or when sample size conditions are not met, we will use a different distribution, though the core ideas remain the same.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

8.2.1: Inference for a Single Proportion

According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[1] This poll included responses of 976 adults.

Identifying when the Sample Proportion is Nearly Normal

A sample proportion can be described as a sample mean. If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0+1+1+\cdots+0}{976} = 0.44 \tag{8.2.1.1}$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes and failures, typically at least 10 of each.

[1]*nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html*

Conditions for the sampling distribution of $\hat{p}$ being nearly normal

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{8.2.1.2}$$

Typically we do not know the true proportion, p, so must substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value - that is, the proportion claimed in the null hypothesis - is used in place of p. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

TIP: Reminder on checking independence of observations

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Alternatively, if the data come from a random process, we must evaluate the independence condition more carefully.

Confidence Intervals for a Proportion

We may want a confidence interval for the proportion of Americans who approve of the job the Supreme Court is doing. Our point estimate, based on a sample of size n = 976 from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use the general confidence interval formula from Section 4.5, we must check the conditions to ensure that the sampling distribution of $\hat{p}$ is nearly normal. We also must determine the standard error of the estimate.

The data are based on a simple random sample and consist of far fewer than 10% of the U.S. population, so independence is confirmed. The sample size must also be sufficiently large, which is checked via the success-failure condition: there were approximately $976 \times \hat{p} = 429$ "successes" and $976 \times (1-\hat{p}) = 547$ "failures" in the sample, both easily greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

Exercise 8.2.1.1

Estimate the standard error of $\hat{p} = 0.44$ using Equation 8.2.1.2. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p.

Answer

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$$

Example 8.2.1.1

Construct a 95% confidence interval for p, the proportion of Americans who trust federal officials most of the time.

Solution

Using the standard error estimate from Exercise 8.2.1.1, the point estimate 0.44, and z* = 1:96 for a 95% confidence interval, the confidence interval may be computed as

$$\text{point estimate} \pm z^* SE \rightarrow 0.44 \pm 1.96 \times 0.016 \rightarrow (0.409, 0.471) \tag{8.2.1.3}$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. If the proportion has not changed since this poll, than we can say with high confidence that the job approval of the Supreme Court is below 50%.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, the sampling distribution of $\hat{p}$ may be well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

Hypothesis Testing for a Proportion

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1-p_0)$ are at least 10, where p_0 is the null value.

Exercise 8.2.1.2

Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Set up a onesided hypothesis test to evaluate this claim.

Answer

Is there convincing evidence that the campaign manager is correct?

- H_0 : p = 0.50,
- H_A : p > 0.50.

Example 8.2.1.2

A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%. Does this provide convincing evidence for the claim of Toohey's manager at the 5% significance level?

Solution

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion,

$$p_0 = 0.5 : np_0 = n(1-p_0) = 500 \times 0.5 = 250 > 10. \tag{8.2.1.4}$$

With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion.

$$\begin{aligned} SE &= \sqrt{\frac{p_0 \times (1-p_0)}{n}} \\ &= \sqrt{\frac{0.5(1-0.5)}{500}} = 0.022 \end{aligned}$$

A picture of the normal model is shown in Figure $8.2.1.1$ with the p-value represented by the shaded region. Based on the normal model, the test statistic can be computed as the Z score of the point estimate:

$$\begin{aligned} Z &= \frac{\text{point estimate - null value}}{SE} \\ &= \frac{0.52-0.50}{0.022} = 0.89 \end{aligned}$$

The upper tail area, representing the p-value, is 0.1867. Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

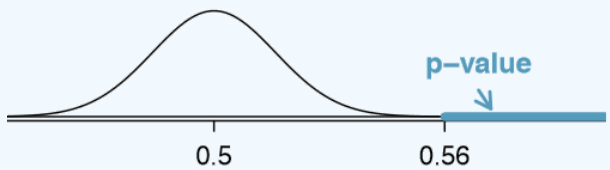


Figure $8.2.1.1$: Sampling distribution of the sample proportion if the null hypothesis is true for Example $8.2.1.2$. The p-value for the test is shaded.

Hypothesis test for a proportion

Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

Choosing a sample size when estimating a proportion

We first encountered sample size computations in Section 4.6, which considered the case of estimating a single mean. We found that these computations were helpful in planning a study to control the size of the standard error of a point estimate. The task was to find a sample size n so that the sample mean would be within some margin of error m of the actual mean with a certain level of confidence. For example, the margin of error for a point estimate using 95% confidence can be written as $1.96 \times SE$. We set up a general equation to represent the problem:

$$ME = z^* SE \leq m \tag{8.2.1.5}$$

where ME represented the actual margin of error and z^* was chosen to correspond to the confidence level. The standard error formula is specified to correspond to the particular setting. For instance, in the case of means, the standard error was given as $\frac{\sigma}{\sqrt{n}}$. In the case of a single proportion, we use $\sqrt{p(1-p)}\, n$ for the standard error.

Planning a sample size before collecting data is equally important when estimating a proportion. For instance, if we are conducting a university survey to determine whether students support a $200 per year increase in fees to pay for a new football stadium, how big of a sample is needed to be sure the margin of error is less than 0.04 using a 95% confidence level?

Example $8.2.1.3$

Find the smallest sample size n so that the margin of error of the point estimate $\hat{p}$ will be no larger than $m = 0.04$ when using a 95% confidence interval.

Solution

For a 95% confidence level, the value z* corresponds to 1.96, and we can write the margin of error expression as follows:

$$ME = z^{*}SE = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04 \tag{8.2.1.6}$$

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could use that value. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this worst case estimate if no other estimate is available:

$$1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.04 \tag{8.2.1.7}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{n} \leq 0.04^2 \tag{8.2.1.8}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} \leq n \tag{8.2.1.9}$$

$$600.25 \leq n \tag{8.2.1.10}$$

We would need at least 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

No estimate of the true proportion is required in sample size computations for a proportion, whereas an estimate of the standard deviation is always needed when computing a sample size for a margin of error for the sample mean. However, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

Example 8.2.1.4

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.

a. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.
b. The sample sizes in (b) vary widely. Which of the three would you suggest using? What would influence your choice?

Solution

(a) For the 1.7% estimate of p, we estimate the appropriate sample size as follows:

$$1.65 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \leq 0.02 \to n \geq 113.7$$

Using the estimate from the first model, we would suggest examining 114 tires (round up!). A similar computation can be accomplished using 0.062 and 0.013 for p: 396 and 88.

(b) We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. (Answers will vary.)

Exercise 8.2.1.4

A recent estimate of Congress' approval rating was 17%.5 What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?

Answer

We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \leq 0.04 \rightarrow n \geq 338.8$$

A sample size of 339 or more would be reasonable.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

8.2.2: Difference of Two Proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, a company weighs whether they should switch to a higher quality parts manufacturer. In the last example, we examine the cancer risk to dogs from the use of yard herbicides.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

Sample Distribution of the Difference of Two Proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when each proportion separately follows a **normal model**, and the samples are **independent**. The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} \tag{8.2.2.1}$$

$$= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{8.2.2.2}$$

where p_1 and p_2 represent the population proportions, and n1 and n2 represent the sample sizes.

For the difference in two means, the standard error formula took the following form:

$$SE_{\hat{x}_1 - \hat{x}_2} = \sqrt{SE^2_{\hat{x}_1} + SE^2_{\hat{x}_2}} \tag{8.2.2.3}$$

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Exercise 5.14.

[5] *www.gallup.com/poll/155144/Congress-Approval-June.aspx*

Table 8.2.2.1: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

	Sample size (ni)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Intervals and tests for $p_1 - p_2$

In the setting of confidence intervals, the sample proportions are used to verify the success/failure condition and also compute standard error, just as was the case with a single proportion.

Example 8.2.2.1

The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[7]

> *As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?*

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed. Table 6.2 shows the results of this experiment. Create and interpret a 90% confidence interval of the difference in approval.

Solution

First the conditions must be verified. Because each group is a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample. Because all conditions are met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13 \tag{8.2.2.4}$$

The standard error may be computed from Equation 8.2.2.2using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025 \tag{8.2.2.5}$$

For a 90% con dence interval, we use z* = 1.65:

$$\text{point estimate} \pm z^* SE \approx 0.13 \pm 1.65 \times 0.025 \rightarrow (0.09, 0.17) \tag{8.2.2.6}$$

We are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still uid on the health insurance mandate.

[7]*www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/.*

Sample sizes for each polling group are approximate.

Exercise 8.2.2.1

A remote control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[8]

Answer

Add texts here. Do not delete this text first.

Example 8.2.2.2

The quality control engineer from Exercise 6.11 collects a sample of gears, examining 1000 gears from each company and nds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Exercise 6.11 using a signi cance level of 5%.

Solution

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the gears are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as

to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, 0.958 - 0.899 = 0.059, can be said to come from a nearly normal distribution.

The standard error can be found using Equation 8.2.2.2:

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114 \tag{8.2.2.7}$$

In this hypothesis test, the sample proportions were used. We will discuss this choice more in Section 6.2.3.

Next, we compute the test statistic and use it to nd the p-value, which is depicted in Figure 8.2.2.1.

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.059-0.03}{0.0114} = 2.54 \tag{8.2.2.8}$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher quality gears actually do pass inspection more than 3% as often as the currently used gears. Based on these results, management will approve the switch to the new supplier.

[8]H_0: The higher quality gears will pass inspection no more than 3% more frequently than the standard quality gears. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher quality gears will pass inspection more than 3% more often than the standard quality gears. $p_{highQ} - p_{standard} > 0.03$.

***Figure 8.2.2.1**: Distribution of the test statistic if the null hypothesis was true.*

The p-value is represented by the shaded area.

Hypothesis testing when H_0: $p_1 = p_2$

Here we use a new example to examine a special estimate of standard error when $H_0: p_1 = p_2$. We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.9 Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table 8.2.2.2

Table 8.2.2.2: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

	cancer	no cancer
2,4 - D	191	304
no 2,4 - D	300	641

Exercise 8.2.2.1

Is this study an experiment or an observational study?

Answer

The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the control group, which is a term usually seen in experiments.

Exercise 8.2.2.1

Exercise 6.14 Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups.[11]

[9]*Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. CaseControl Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.*

Answer

Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.

- H_0: the proportion of dogs with exposure to 2,4-D is the same in "cancer" and \no cancer" dogs, $p_c - p_n = 0$.
- H_A: dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

Example 8.2.2.1: pooled estimate

First are the conditions met to use the normal model and make inference on the results?

(1) It is unclear whether this is a random sample. However, if we believe the dogs in both the cancer and no cancer groups are representative of each respective population and that the dogs in the study do not interact in any way, then we may find it reasonable to assume independence between observations. (2) The success-failure condition holds for each sample.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

In your hypotheses for Exercise $8.2.2.1$, the null is that the proportion of dogs with exposure to 2,4-D is the same in each group. The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. To identify the p-value for this test, we first check conditions (Example 6.15) and compute the standard error of the difference:

$$SE = \sqrt{\frac{p_c(1-p_c)}{n_c} + \frac{p_n(1-p_n)}{n_n}} \tag{8.2.2.9}$$

In a hypothesis test, the distribution of the test statistic is always examined as though the null hypothesis is true, i.e. in this case, $p_c = p_n$. The standard error formula should reflect this equality in the null hypothesis. We will use p to represent the common rate of dogs that are exposed to 2,4-D in the two groups:

$$SE = \sqrt{\frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_n}} \tag{8.2.2.10}$$

We don't know the exposure rate, p, but we can obtain a good estimate of it by pooling the results of both samples:

$$\hat{p} = \frac{\#\text{ of "successes"}}{\#\text{ of cases}} = \frac{191+304}{191+300+304+641} = 0.345 \tag{8.2.2.11}$$

This is called the **pooled estimate** of the sample proportion, and we use it to compute the standard error when the null hypothesis is that $p_1 = p_2$ (e.g. $p_c = p_n$ or $p_c - p_n = 0$) . We also typically use it to verify the success-failure condition.

Pooled estimate of a proportion

When the null hypothesis is $p_1 = p_2$, it is useful to nd the pooled estimate of the shared proportion:

$$\hat{p} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} \tag{8.2.2.12}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1} \tag{8.2.2.13}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

: $p_1 = p_2$

When the null hypothesis suggests the proportions are equal, we use the pooled proportion estimate $(\hat{p})$ to verify the success-failure condition and also to estimate the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_c} + \frac{\hat{p}(1-\hat{p})}{n_n}} \tag{8.2.2.14}$$

Exercise 8.2.2.1

Using Equation 8.2.2.14, $\hat{p} = 0.345, n_1 = 491$, and $n_2 = 945$, verify the estimate for the standard error is SE = 0.026. Next, complete the hypothesis test using a significance level of 0.05. Be certain to draw a picture, compute the p-value, and state your conclusion in both statistical language and plain language.

Answer

Compute the test statistic:

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

8.2.3: Testing for Goodness of Fit using Chi-Square (Special Topic)

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test. In the rst case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Table 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reect the population of eligible jurors, i.e. registered voters.

Table 6.5: Representation by race in a city's juries and population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

[12]*Compute the test statistic:*

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58 \tag{8.2.3.1}$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

A second application, assessing the t of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for the years 1990-2011 are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

Creating a test statistic for one-way tables

Example 8.2.3.1:

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

Solution

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

Exercise 8.2.3.1

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race?

Answer

Answers can be found in Table 6.6.

Table 6.6: Actual and expected make-up of the jurors.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected count	198	19.25	33	24.75	275

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

- H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.
- H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate - null value}}{\text{SE of point estimate}} \tag{8.2.3.2}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to rst compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count - null white count}}{\text{SE of observed white count}} \tag{8.2.3.3}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50 \tag{8.2.3.4}$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

$$\begin{matrix} Black & Hispanic & Other \\ Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 154 & Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39 & Z_4 = \frac{19 - 24:75}{\sqrt{24.75}} = -1.16 \end{matrix} \tag{8.2.3.5}$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they - as a group - tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58 \tag{8.2.3.6}$$

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89 \tag{8.2.3.7}$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual - e.g. a standardized difference of 2.5 - will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{\left(\text{observed count\_1 - null count\_1}\right)^2}{\text{null count\_1}} + \cdots + \frac{\left(\text{observed count\_4 - null count\_4}\right)^2}{\text{null count\_4}} \tag{8.2.3.8}$$

[13]Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown - though not here - that the square root of the count is a better way to standardize the count differences.

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a chi-square distribution. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

The chi-square distribution and finding areas

The chi-square distribution is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters - mean and standard deviation - that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df),** which inuences the shape, center, and spread of the distribution.

Exercise 8.2.3.1

Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

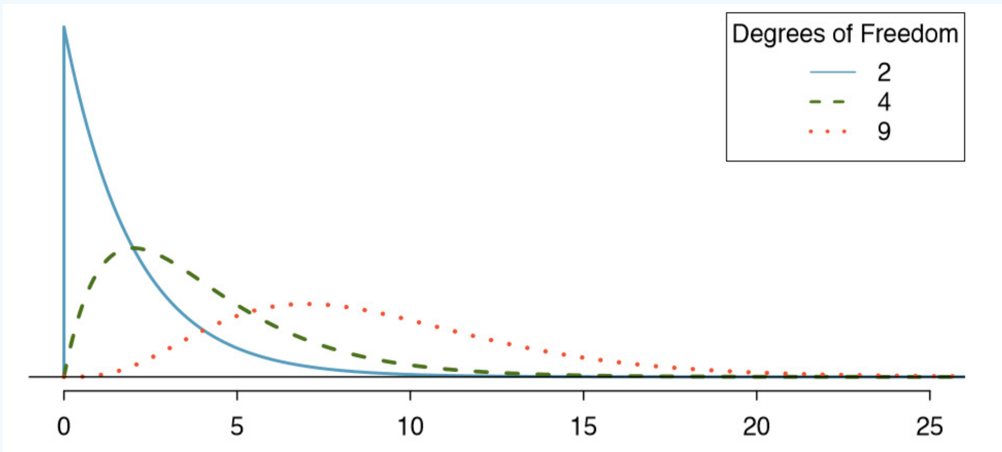


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

Answer

[14](a) The center becomes larger. If we look carefully, we can see that the center of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for df = 2, and then the distributions become more symmetric for the larger degrees of freedom df = 4 and df = 9. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Figure 6.7 and Exercise 6.20 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the **chi-square table**, partially shown in Table 6.8. A more complete table is presented in Appendix B.3 on page 412. This table is very similar to the t table from Sections 5.3 and 5.4: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t table is that the chi-square table only provides upper tail values.

Table 6.8: A section of the chi-square table. A complete table is in Appendix B.3 on page 412.

Upper tail	0.3 0.2 0.1 0.05	0.02 0.01 0.005 0.001
df 1 2 3 4 5	2.41 3.22 4.61 5.99 3.66 4.64 6.25 7.81 4.88 5.99 7.78 9.49 6.06 7.29 9.24 11.07	7.82 9.21 10.60 13.82 9.84 11.34 12.84 16.27 11.67 13.28 14.86 18.47 13.39 15.09 16.75 20.52
6 7	7.23 8.56 10.64 12.59 8.38 9.80 12.02 14.07	15.03 16.81 18.55 22.46 16.62 18.48 20.28 24.32

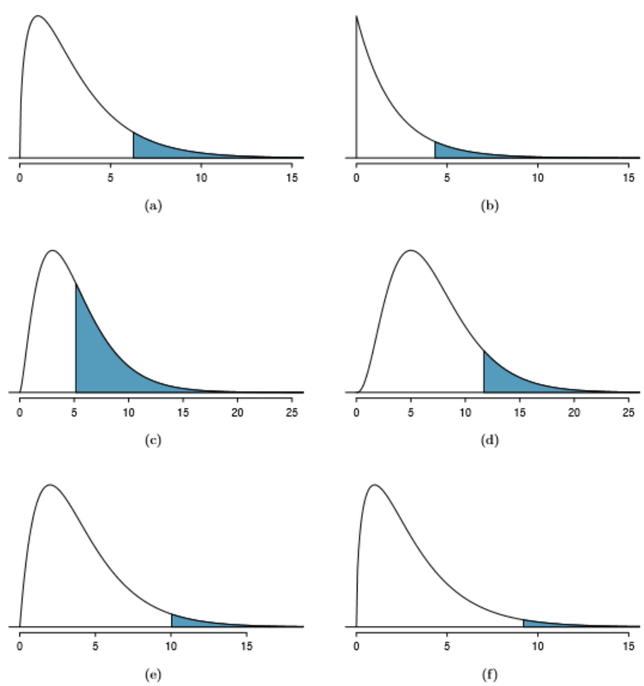


Figure 6.9: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded. (e) 4 degrees of freedom, area above 10 shaded. (f) 3 degrees of freedom, area above 9.21 shaded.

Example 6.21 Figure 6.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 6.8 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value { 6.25 { falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 6.9(a) has area 0.1.

Example 6.22 We rarely observe the exact value in the table. For instance, Figure 6.9(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3, which does not fall in Table 6.8. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 6.9(b) is between 0.1 and 0.2.

Example 6.23 Figure 6.9(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is greater than 0.3.

Exercise 6.24 Figure 6.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

Exercise 6.25 Figure 6.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Exercise 6.26 Figure 6.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

[15]*The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.*

[16]*The area is between 0.02 and 0.05.*

[17]*Between 0.02 and 0.05.*

Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with k - 1 degrees of freedom, where k is the number of bins.

Example 8.2.3.1:

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

Solution

In the jurors example, there were k = 4 categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with k - 1 = 3 degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

Example 8.2.3.1:

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

The chi-square distribution and p-value are shown in Figure 6.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix B.3 or the short table on page 277, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

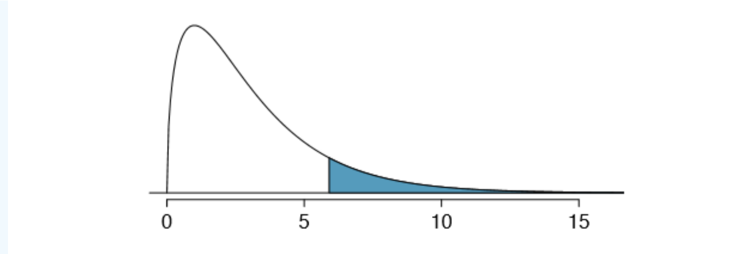


Figure 6.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with df = 3.

Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts $O_1, O_2, \ldots, O_k$ in k categories are unusually different from what might be expected under a null hypothesis. Call the expected counts that are based on the null hypothesis $E_1, E_2, \ldots, E_k$. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with k - 1 degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \tag{8.2.3.9}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

Conditions for the chi-square test

There are three conditions that must be checked before performing a chi-square test:

- **Independence**. Each case that contributes a count to the table must be independent of all the other cases in the table.
- **Sample size / distribution**. Each particular scenario (i.e. cell count) must have at least 5 expected cases.
- **Degrees of freedom** We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one proportion methods introduced in Section 6.1.

Evaluating goodness of fit for a distribution

Section 3.3 would be useful background reading for this example, but it is not a prerequisite. We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model ts a data set. Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as Up or Down (D) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each Up day:

$$\begin{array}{ccccccccccc} \text{Change in price} & 2.52 & -1.46 & 0.51 & -4.07 & 3.36 & 1.10 & -5.46 & -1.03 & -2.99 & 1.71 \\ \text{Outcome} & Up & D & Up & D & Up & Up & D & D & D & Up \\ \text{Days to Up} & 1 & - & 2 & - & 2 & 1 & - & - & - & 4 \end{array} \tag{8.2.3.10}$$

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the kth trial to observe the rst success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next Up trading day, and two more for the third Up day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Table 6.11 shows the number of waiting days for a positive trading day during 1990-2011for the S&P500.

Table 6.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

We consider how many days one must wait until observing an Up day on the S&P500 stock exchange. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a geometric distribution. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an Up day is observed. Under this hypothesis, the number of days until an Up day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an Up day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 from 1990 to 2011 and summarize the waiting times in Table 6.12 and Figure 6.13. The S&P500 was positive on 53.2% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have binned together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the Observed row in Table 6.12, can be compared to the expected counts from the Geometric Model row. The method for computing expected counts is discussed in Table 6.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each

Table 6.12: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model $(P(D) = (1-0.532)^{D-1}(0.532))$ and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 = 11.65\%$ of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

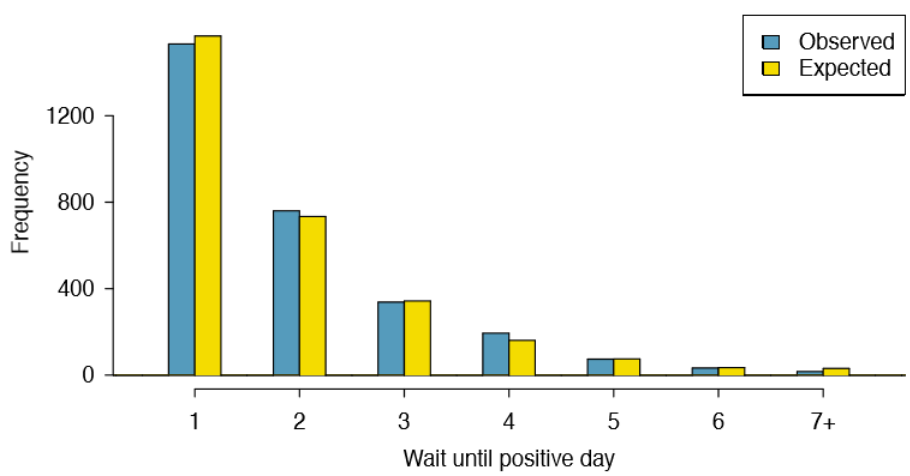


Figure 6.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Example 6.29 Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Table 6.12.

Exercise 6.30 Table 6.12 provides a set of count data for waiting times ($O_1 = 1532, O_2 = 760, \ldots$) and expected counts under the geometric distribution ($E_1 = 1569, E_2 = 734, \ldots$). Compute the chi-square test statistic, X^2.[18]

Exercise 6.31 Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

Example 6.32 If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with df = 6. Using this information, compute a p-value.

Figure 6.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix B.3, we nd that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that

[18] $X^2 = \frac{(1532 - 1569)^2}{1569} + \frac{(760 - 734)^2}{734} + \cdots + \frac{(17 - 31)^2}{31} = 15.08$

[19] *There are k = 7 groups, so we use df = k - 1 = 6.*

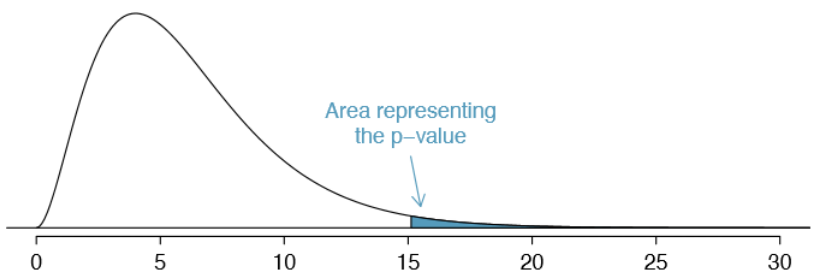


Figure 6.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

Example 6.33 In Example 6.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

8.2.4: Testing for Independence in Two-Way Tables (Special Topic)

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Table 6.15 shows an example of 10,000 queries split into three algorithm groups.[20] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Table 6.15: Google experiment breakdown of test subjects into three search groups.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

[20]Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

Example 8.2.4.1

What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

- H_0: The algorithms each perform equally well.
- H_A: The algorithms do not perform equally well.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. Under scenario (1), we might think that the user was satis ed with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Table 6.16 provides the results from the experiment. These data are very similar to the count data in Section 6.3. However, now the different combinations of two variables are binned in a two-way table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

Table 6.16: Results of the Google search algorithm experiment.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

What is so different about one-way tables and two-way tables?

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for combinations of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the search variable is independent of the algorithm variable.

Expected Counts in Two-way Tables

Example 6.35 From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as $\frac{7078}{10000} = 0.7078$. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satis ed with the initial search:

$$0.7078 \times 5000 = 3539 users \tag{8.2.4.1}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

Exercise 8.2.4.1

Exercise 6.36 Using the same rationale described in Example 6.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

[21]*We would expect 0.7078 * 2500 = 1769.5. It is okay that this is a fraction.*

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 6.35 and Exercise 6.36. These expected counts were used to construct Table 6.17, which is the same as Table 6.16, except now the expected counts have been added in parentheses.

Table 6.17: The observed counts and the (Expected counts)

Search algorithm	current	test 1	test 2	Total
No new search	3511 (3539)	1749 (1769.5)	1818 (1769.5)	7078
New search	1489 (1461)	751 (730.5)	682 (730.5)	2922
Total	5000	2500	2500	10000

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the rst row:

$$0.7078 \times (\text{column 1 total}) = 3539 \tag{8.2.4.2}$$

$$0.7078 \times (\text{column 2 total}) = 1769.5 \tag{8.2.4.3}$$

$$0.7078 \times (\text{column 3 total}) = 1769.5 \tag{8.2.4.4}$$

Looking back to how the fraction 0.7078 was computed - as the fraction of users who did not perform a new search ($\frac{7078}{10000}$) - these three expected counts could have been computed as

$$\frac{\text{row 1 total}}{\text{table total}} (\text{column 1 total}) = 3539 \tag{8.2.4.5}$$

$$\frac{\text{row 1 total}}{\text{table total}} (\text{column 2 total}) = 1769.5 \tag{8.2.4.6}$$

$$\frac{\text{row 1 total}}{\text{table total}} (\text{column 3 total}) = 1769.5 \tag{8.2.4.7}$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

Computing expected counts in a two-way table

To identify the expected count for the ith row and jth column, compute

$$\text{Expected Count}_{\text{row i, col j}} = \frac{(\text{row i total}) \times (\text{column j total})}{\text{table total}} \tag{8.2.4.8}$$

The chi-square Test for Two-way Tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

$$\text{General formula} \frac{(\text{observed count - expected count})^{2}}{\text{expected count}} \tag{8.2.4.9}$$

$$\text{Row 1, Col 1} \frac{(3511-3539)^2}{3539} = 0.222 \tag{8.2.4.10}$$

$$\text{Row 1, Col 2} \frac{(1749-1769.5)^2}{1769.5} = 0.237 \tag{8.2.4.11}$$

$$\vdots \tag{8.2.4.12}$$

$$\text{Row 2, Col 3} \frac{(682-730.5)^2}{730.5} = 3.220 \tag{8.2.4.13}$$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 0.222 + 0.237 + \dots + 3.220 = 6.120 \tag{8.2.4.14}$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus } 1) \times (\text{number of columns minus } 1) \tag{8.2.4.15}$$

In our example, the degrees of freedom parameter is

$$df = (2-1) \times (3-1) = 2 \tag{8.2.4.16}$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic X^2 = 6.12 closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.18.

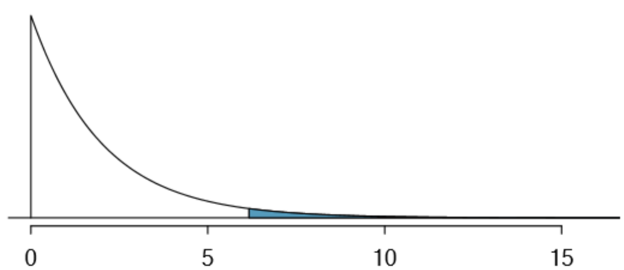


Figure 6.18: Computing the p-value for the Google hypothesis test.

Definition: degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df = (R-1) \times (C-1) \tag{8.2.4.17}$$

where R is the number of rows in the table and C is the number of columns.

[22]*Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.*

Table 6.19: Pew Research poll results of a March 2012 poll.

		Congress		
	Obama	Democrats	Republicans	Total

Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

TIP: Use two-proportion methods for 2-by-2 contingency tables

When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 6.2.

Example 8.2.4.1

Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Solution

Looking in Appendix B.3 on page 412, we examine the row corresponding to 2 degrees of freedom. The test statistic, X^2 = 6.120, falls between the fourth and fth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of α = 0.05 and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Example 8.2.4.1

Table 6.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

Solution

- H_0: There is no difference in approval ratings between the three groups.
- H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23]*See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 6.19 are approximate.*

Exercise 8.2.4.1

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.38. As a rst step, compute the expected values for each of the six table cells.[24]

[24]The expected count for row one / column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119\times1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104\times1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

Exercise 8.2.4.1

Compute the chi-square test statistic.[25]

[25]For each cell, compute $\frac{(\text{obs - exp})^2}{exp}$. For instance, the rst row and rst column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \cdots + 34.0 = 106.4$.

Exercise 8.2.4.1

Because there are 2 rows and 3 columns, the degrees of freedom for the test is df = (2 - 1) (3 - 1) = 2. Use X^2 = 106.4, df = 2, and the chi-square table on page 412 to evaluate whether to reject the null hypothesis.[26]

[26]The test statistic is larger than the right-most column of the df = 2 row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans'

approval has differences among Democrats in Congress, Republicans in Congress, and the president.

8.2.5: Small Sample Hypothesis Testing for a Proportion (Special Topic)

In this section we develop inferential methods for a single proportion that are appropriate when the sample size is too small to apply the normal model to $\hat{p}$. Just like the methods related to the t distribution, these methods can also be applied to large samples.

When the Success-Failure Condition is Not Met

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspect of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients. One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Exercise 8.2.5.1

Exercise 6.42

We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.

Solution

The sample proportion: $\hat{p} = \frac{3}{62} = 0.048$

Example 8.2.5.1

Is it possible to assess the consultant's claim using the data provided?

Solution

No. The claim is that there is a causal connection, but the data are observational. Patients who hire this medical consultant may have lower complication rates for other reasons.

While it is not possible to assess this causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Exercise 8.2.5.1

Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.

Solution

- H_0: There is no association between the consultant's contributions and the clients' complication rate. In statistical language, p = 0.10.
- H_A: Patients who work with the consultant tend to have a complication rate lower than 10%, i.e. p < 0.10.

Example 8.2.5.1

In the examples based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

Solution

The independence assumption may be reasonable if each of the surgeries is from a different surgical team. However, the success-failure condition is not satis ed. Under the null hypothesis, we would anticipate seeing $62 \times 0.10 = 6.2$ complications, not the 10 required for the normal approximation.

The uncertainty associated with the sample proportion should not be modeled using the normal distribution. However, we would still like to assess the hypotheses from Exercise 6.44 in absence of the normal framework. To do so, we need to evaluate the possibility of a sample value (^p) this far below the null value, $p_0 = 0.10$. This possibility is usually measured with a p-value.

The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis is true. Supposing the null hypothesis is true, we can compute the p-value by identifying the chance of observing a test statistic that favors the alternative hypothesis at least as strongly as the observed test statistic. This can be done using simulation.

Generating the null distribution and p-value by simulation

We want to identify the sampling distribution of the test statistic ($\hat{p}$) if the null hypothesis was true. In other words, we want to see how the sample proportion changes due to chance alone. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.

Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients. If this was the case, we could simulate 62 clients to get a sample proportion for the complication rate from the null distribution.

Each client can be simulated using a deck of cards. Take one red card, nine black cards, and mix them up. Then drawing a card is one way of simulating the chance a patient has a complication if the true complication rate is 10% for the data. If we do this 62 times and compute the proportion of patients with complications in the simulation, $\hat{p}_{sim}$, then this sample proportion is exactly a sample from the null distribution.

An undergraduate student was paid $2 to complete this simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e. $\hat{p}_{sim} = \frac{5}{62} = 0.081$.

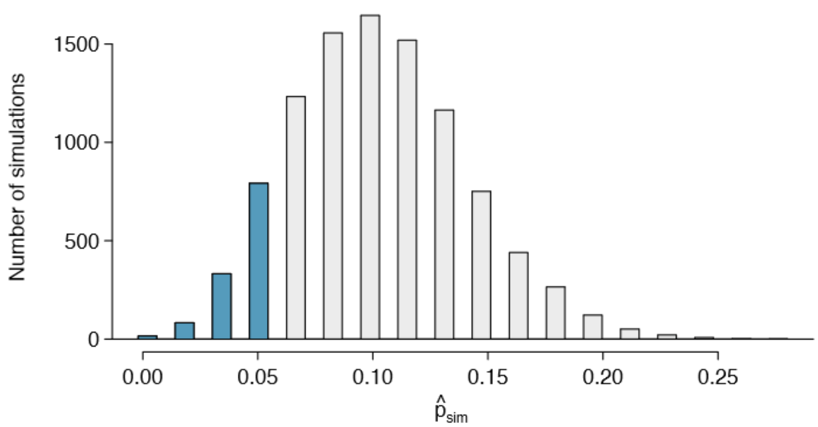


Figure 6.20: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

Example 8.2.5.1

Is this one simulation enough to determine whether or not we should reject the null hypothesis from Exercise 6.44? Explain.

Solution

No. To assess the hypotheses, we need to see a distribution of many $\hat{p}_{sim}$, not just a single draw from this sampling distribution.

One simulation isn't enough to get a sense of the null distribution; many simulation studies are needed. Roughly 10,000 seems sufficient. However, paying someone to simulate 10,000 studies by hand is a waste of time and money. Instead, simulations are typically programmed into a computer, which is much more efficient.

Figure 6.20 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and nd the p-value:

$$\text{left tail} = \frac{\text{Number of observed simulations with} \hat{p}_{sim} \leq 0.048}{10000} \tag{6.47}$$

Of the 10,000 simulated $\hat{p}_{sim}$, 1222 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.1222.

Exercise 8.2.5.1

Because the estimated p-value is 0.1222, which is larger than the signi cance level 0.05, we do not reject the null hypothesis. Explain what this means in plain language in the context of the problem.

Solution

There isn't sufficiently strong evidence to support an association between the consultant's work and fewer surgery complications.

Exercise 8.2.5.1

Does the conclusion in Exercise 6.48 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.

Solution

No. It might be that the consultant's work is associated with a reduction but that there isn't enough data to convincingly show this connection.

One-sided hypothesis test for p with a small sample

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions ($\hat{p}_{sim}$) and use this distribution to compute the tail area, i.e. the p-value.

We continue to use the same rule as before when computing the p-value for a twosided test: double the single tail area, which remains a reasonable approach even when the sampling distribution is asymmetric. However, this can result in p-values larger than 1 when the point estimate is very near the mean in the null distribution; in such cases, we write that the p-value is 1. Also, very large p-values computed in this way (e.g. 0.85), may also be slightly inated.

Exercise 6.48 said the p-value is estimated. It is not exact because the simulated null distribution itself is not exact, only a close approximation. However, we can generate an exact null distribution and p-value using the binomial model from Section 3.4.

Generating the exact null distribution and p-value

The number of successes in n independent cases can be described using the binomial model, which was introduced in Section 3.4. Recall that the probability of observing exactly k successes is given by

$$P(\text{k successes}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{6.50}$$

where p is the true probability of success. The expression $\binom{n}{k}$ is read as n choose k, and the exclamation points represent factorials. For instance, 3! is equal to $3 \times 2 \times 1 = 6, 4!$ is equal to $4 \times 3 \times 2 \times 1 = 24$, and so on (see Section 3.4).

The tail area of the null distribution is computed by adding up the probability in Equation (6.50) for each k that provides at least as strong of evidence favoring the alternative hypothesis as the data. If the hypothesis test is one-sided, then the p-value is represented by a single tail area. If the test is two-sided, compute the single tail area and double it to get the p-value, just as we have done in the past.

Example 8.2.5.1

Compute the exact p-value to check the consultant's claim that her clients' complication rate is below 10%.

Solution

Exactly k = 3 complications were observed in the n = 62 cases cited by the consultant. Since we are testing against the 10% national average, our null hypothesis is p = 0.10. We can compute the p-value by adding up the cases where there are 3 or fewer complications:

$$\text{p-value} = \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \tag{8.2.5.1}$$

$$= \sum_{j=0}^{3} \binom{62}{j} 0.1^j (1-0.1)^{62-j} \tag{8.2.5.2}$$

$$= \binom{62}{0} 0.1^0 (1-0.1)^{62-0} + \binom{62}{1} 0.1^1 (1-0.1)^{62-1} + \binom{62}{2} 0.1^2 (1-0.1)^{62-2} + \binom{62}{3} 0.1^3 (1-0.1)^{62-3} \tag{8.2.5.3}$$

$$= 0.0015 + 0.0100 + 0.0340 + 0.0755 \tag{8.2.5.4}$$

$$= 0.1210 \tag{8.2.5.5}$$

This exact p-value is very close to the p-value based on the simulations (0.1222), and we come to the same conclusion. We do not reject the null hypothesis, and there is not statistically signi cant evidence to support the association.

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 6.20 on page 290.

Using simulation for goodness of fit tests

Simulation methods may also be used to test goodness of t. In short, we simulate a new sample based on the purported bin probabilities, then compute a chi-square test statistic X^2_{sim}. We do this many times (e.g. 10,000 times), and then examine the distribution of these simulated chi-square test statistics. This distribution will be a very precise null distribution for the test statistic X^2 if the probabilities are accurate, and we can nd the upper tail of this null distribution, using a cutoff of the observed test statistic, to calculate the p-value.

Example 8.2.5.1

Section 6.3 introduced an example where we considered whether jurors were racially representative of the population. Would our ndings differ if we used a simulation technique?

Solution

Since the minimum bin count condition was satis ed, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 6.21 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

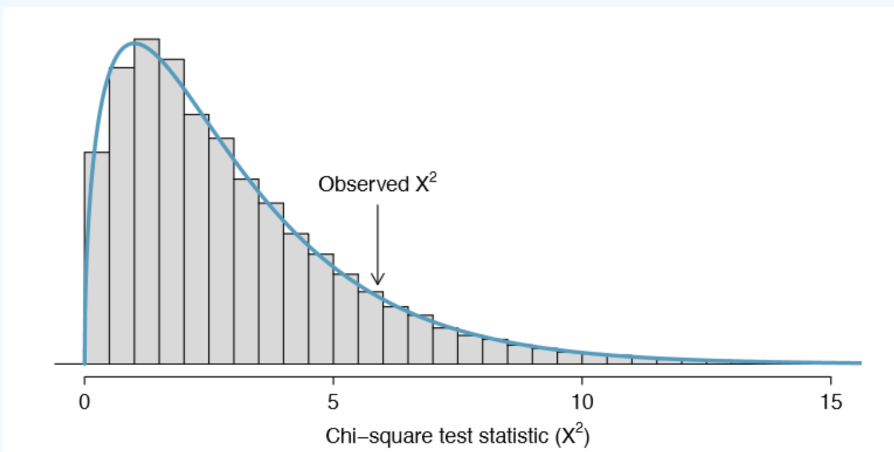


Figure 6.21: The precise null distribution for the juror example from Section 6.3 is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.

8.2.6: Randomization Test (Special Topic)

Cardiopulmonary resuscitation (CPR) is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. This procedure is helpful in maintaining some blood circulation, but the chest compressions involved can also cause internal injuries. Internal bleeding and other injuries complicate additional treatment efforts following arrival at a hospital. For instance, blood thinners may be used to help release a clot that is causing the heart attack. However, the blood thinner would negatively affect an internal injury. Here we consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.(Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial, by Bottiger et al., The Lancet, 2001.) These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive the blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

Example 8.2.6.1

Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival proportion in the control group and p_t represent the survival proportion for the treatment group.

Solution

We are interested in whether the blood thinners are helpful or harmful, so this should be a two-sided test.

- H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group.

$$p_t - p_c = 0. \tag{8.2.6.1}$$

- H_A: Blood thinners do have an impact on survival.

$$p_t - p_c \neq 0. \tag{8.2.6.2}$$

Large Sample Framework for a Difference in Two Proportions

There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did. The study results are shown in Table 6.22.

Table 6.22: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Exercise 8.2.6.1

What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$.

Solution

Observed control survival rate:

$$p_c = \frac{11}{50} = 0.22. \tag{8.2.6.3}$$

Treatment survival rate:

$$p_t = \frac{14}{40} = 0.35. \tag{8.2.6.4}$$

Observed difference:

$$\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13. \tag{8.2.6.5}$$

According to the point estimate, there is a 13% increase in the survival proportion when patients who have undergone CPR outside of the hospital are treated with blood thinners. However, we wonder if this difference could be due to chance. We'd like to investigate this using a large sample framework, but we first need to check the conditions for such an approach.

Example $8.2.6.2$

Can the point estimate of the difference in survival proportions be adequately modeled using a normal distribution?

Solution

We will assume the patients are independent, which is probably reasonable. The success-failure condition is also satisfied. Since the proportions are equal under the null, we can compute the pooled proportion,

$$\hat{p} = \frac{(11+14)}{(50+40)} = 0.278, \tag{8.2.6.6}$$

for checking conditions. We nd the expected number of successes (13.9, 11.1) and failures (36.1, 28.9) are above 10. The normal model is reasonable.

While we can apply a normal framework as an approximation to nd a p-value, we might keep in mind that the expected number of successes is only 13.9 in one group and 11.1 in the other. Below we conduct an analysis relying on the large sample normal theory. We will follow up with a small sample analysis and compare the results.

Example $8.2.6.3$

Assess the hypotheses presented in Example 6.53 using a large sample framework. Use a significance level of $\alpha = 0.05$.

Solution

We suppose the null distribution of the sample difference follows a normal distribution with mean 0 (the null value) and a standard deviation equal to the standard error of the estimate. The null hypothesis in this case would be that the two proportions are the same, so we compute the standard error using the pooled standard error formula from Equation (6.16) on page 273:

$$SE = \sqrt{\frac{p(1-p)}{n_t} + \frac{p(1-p)}{n_c}} \approx \sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}} = 0.095 \tag{8.2.6.7}$$

where we have used the pooled estimate $(\hat{p} = \frac{11+14}{50+40} = 0.278)$ in place of the true proportion, p.

The null distribution with mean zero and standard deviation 0.095 is shown in Figure 6.23. We compute the tail areas to identify the p-value. To do so, we use the Z score of the point estimate:

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - \text{null value}}{SE} = \frac{0.13 - 0}{0.095} = 1.37 \tag{8.2.6.8}$$

If we look this Z score up in Appendix B.1, we see that the right tail has area 0.0853. The p-value is twice the single tail area: 0.176. This p-value does not provide convincing evidence that the blood thinner helps. Thus, there is insufficient evidence to conclude whether or not the blood thinner helps or hurts. (Remember, we never "accept" the null hypothesis - we can only reject or fail to reject.)

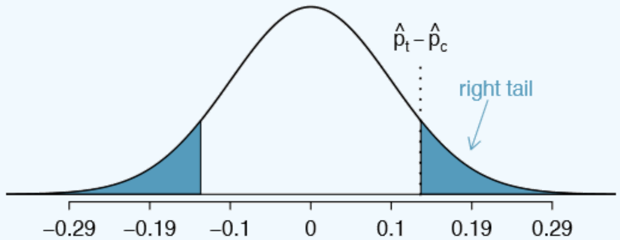


Figure 6.23: The null distribution of the point estimate $\hat{p}_t - \hat{p}_c$ under the large sample framework is a normal distribution with mean 0 and standard deviation equal to the standard error, in this case SE = 0.095. The p-value is represented by the shaded areas.

The p-value 0.176 relies on the normal approximation. We know that when the samples sizes are large, this approximation is quite good. However, when the sample sizes are relatively small as in this example, the approximation may only be adequate. Next we develop a simulation technique, apply it to these data, and compare our results. In general, the small sample method we develop may be used for any size sample, small or large, and should be considered as more accurate than the corresponding large sample technique.

Simulating a Difference under the null Distribution

The ideas in this section were first introduced in the optional Section 1.8. Suppose the null hypothesis is true. Then the blood thinner has no impact on survival and the 13% difference was due to chance. In this case, we can simulate null differences that are due to chance using a randomization technique. (The test procedure we employ in this section is formally called a **permutation test**). By randomly assigning "fake treatment" and "fake control" stickers to the patients' files, we could get a new grouping - one that is completely due to chance. The expected difference between the two proportions under this simulation is zero.

We run this simulation by taking 40 treatment fake and 50 control fake labels and randomly assigning them to the patients. The label counts of 40 and 50 correspond to the number of treatment and control assignments in the actual study. We use a computer program to randomly assign these labels to the patients, and we organize the simulation results into Table 6.24.

Table 6.24: Simulated results for the CPR study under the null hypothesis. The labels were randomly assigned and are independent of the outcome of the patient.

	Survived	Died	Total
Control_fake	15	35	50
Treatment_fake	10	30	40
Total	25	65	90

Exercise 8.2.6.2

What is the difference in survival rates between the two fake groups in Table 6.24? How does this compare to the observed 13% in the real groups?

Solution

The difference is $\hat{p}_{\text{t;fake}} - \hat{p}_{\text{c;fake}} = \dfrac{10}{40} - \dfrac{15}{50} = -0.05$, which is closer to the null value p0 = 0 than what we observed.

The difference computed in Exercise 6.57 represents a draw from the null distribution of the sample differences. Next we generate many more simulated experiments to build up the null distribution, much like we did in Section 6.5.2 to build a null distribution for a one sample proportion.

Caution: Simulation in the two proportion case requires that the null difference is zero

The technique described here to simulate a difference from the null distribution relies on an important condition in the null hypothesis: there is no connection between the two variables considered. In some special cases, the null difference might not be zero, and more advanced methods (or a large sample approximation, if appropriate) would be necessary.

Null distribution for the difference in two proportions

We build up an approximation to the null distribution by repeatedly creating tables like the one shown in Table 6.24 and computing the sample differences. The null distribution from 10,000 simulations is shown in Figure 6.25.

Example $8.2.6.4$

Compare Figures 6.23 and 6.25. How are they similar? How are they different?

Solution

The shapes are similar, but the simulated results show that the continuous approximation of the normal distribution is not very good. We might wonder, how close are the p-values?

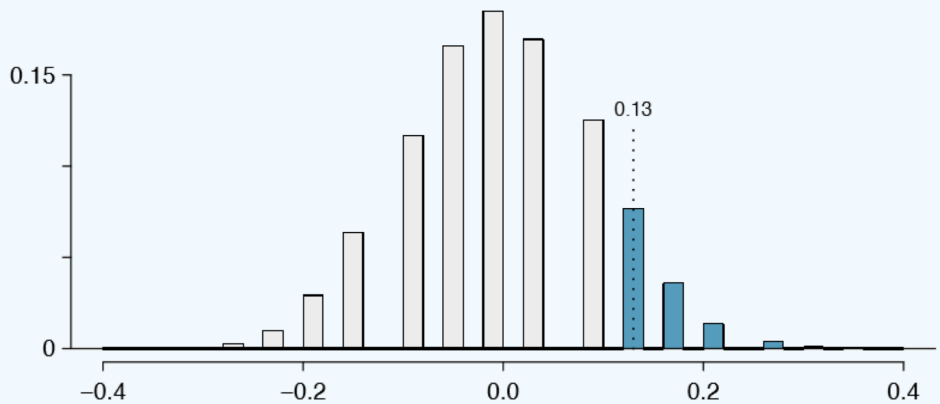


Figure 6.25: An approximation of the null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The p-value is twice the right tail area.

Exercise $8.2.6.3$

The right tail area is about 0.13. (It is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) The p-value is computed by doubling the right tail area: 0.26. How does this value compare with the large sample approximation for the p-value?

Solution

The approximation in this case is fairly poor (p-values: 0.174 vs. 0.26), though we come to the same conclusion. The data do not provide convincing evidence showing the blood thinner helps or hurts patients.

In general, small sample methods produce more accurate results since they rely on fewer assumptions. However, they often require some extra work or simulations. For this reason, many statisticians use small sample methods only when conditions for large sample methods are not satisfied.

Randomization for two-way tables and chi-square

Randomization methods may also be used for the contingency tables. In short, we create a randomized contingency table, then compute a chi-square test statistic X^2_{sim}. We repeat this many times using a computer, and then we examine the distribution of these simulated test statistics. This randomization approach is valid for any sized sample, and it will be more accurate for cases where one or more expected bin counts do not meet the minimum threshold of 5. When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution. As before, we use the upper tail of the null distribution to calculate the p-value.

This page titled 8.2.6: Randomization Test (Special Topic) is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

- **6.6: Randomization Test (Special Topic)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.

8.2.7: Exercises

Inference for a single proportion

6.1 Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

(b) The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

(c) A random sample of 125 college students where 12% are vegetarians would be considered unusual.

(d) A random sample of 250 college students where 12% are vegetarians would be considered unusual.

(e) The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

6.2 Young Americans, Part I. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.[36]

(a) The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.

(b) The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

(c) A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

(d) A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

6.3 Orange tabbies. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

6.4 Young Americans, Part II. About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning.[37]

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

(b) In order for the the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

[36] *A. Vaughn. "Poll finds young adults optimistic, but not about money". In: Los Angeles Times (2011).*

[37] *Demos.org. "The State of Young America: The Poll". In: (2011).*

6.5 Prop 19 in California. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[38]

(a) We are 95% con dent that between 62% and 78% of the California voters in this sample support Prop 19.

(b) We are 95% con dent that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.

(c) If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.

(d) In order to decrease the margin of error to 4%, we would need to quadruple (multiply by 4) the sample size.

(e) Based on this con dence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[39]

(a) We are 95% con dent that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b) We are 95% con dent that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

(d) The margin of error at a 90% con dence level would be higher than 3%.

6.7 Fireworks on July 4th. In late June 2012, Survey USA published results of a survey stating that 56% of the 600 randomly sampled Kansas residents planned to set off reworks on July 4th. Determine the margin of error for the 56% point estimate using a 95% con dence level.40

6.8 Elderly drivers. In January 2011, The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% con dence level.[41]

(a) Verify the margin of error reported by The Marist Poll.

(b) Based on a 95% con dence interval, does the poll provide convincing evidence that more than 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?

[38]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

[39]*Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.*

[40]*Survey USA, News Poll #19333, data collected on June 27, 2012.*

[41]*Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.*

6.9 Life after college. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and nd out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions for constructing a con dence interval based on these data are met.

(c) Calculate a 95% con dence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

(d) What does "95% confidence" mean?

(e) Now calculate a 99% con dence interval for the same parameter and interpret it in the context of the data.

(f) Compare the widths of the 95% and 99% con dence intervals. Which one is wider? Explain.

6.10 Life rating in Greece. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered

"suffering".[42]

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions required for constructing a con dence interval based on these data are met.

(c) Construct a 95% con dence interval for the proportion of Greeks who are "suffering".

(d) Without doing any calculations, describe what would happen to the con dence interval if we decided to use a higher confidence level.

(e) Without doing any calculations, describe what would happen to the con dence interval if we used a larger sample.

6.11 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[43]

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% con dence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

(c) What does "90% confidence" mean?

(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

[42]Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.

[43]studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[44]

(a) Is 48% a sample statistic or a population parameter? Explain.

(b) Construct a 95% con dence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

(c) A critic points out that this 95% con dence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

(d) A news piece on this survey's ndings states, \Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

6.13 Public option, Part I. A Washington Post article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More speci cally, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.[45]

(a) A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

(b) Would you expect a con dence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

6.14 The Civil War. A national survey conducted in 2011 among a simple random sample of 1,507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life.[46]

(a) Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) Interpret the p-value in this context.

(c) Calculate a 90% con dence interval for the proportion of Americans who think the Civil War is still relevant. Interpret the interval in this context, and comment on whether or not the con dence interval agrees with the conclusion of the hypothesis test.

6.15 Browsing on the mobile device. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.[47]

(a) According to an online article, a report from a mobile research company indicates that 38 percent of Chinese mobile web users only access the internet through their cell phones.48 Conduct a hypothesis test to determine if these data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%.

(b) Interpret the p-value in this context.

(c) Calculate a 95% con dence interval for the proportion of Americans who access the internet on their cell phones, and interpret the interval in this context.

[44] *National Opinion Research Center, General Social Survey, 2010.*

[45] *D. Balz and J. Cohen. "Most support public option for health insurance, poll nds". In: The Washington Post (2009).*

[46] *Pew Research Center Publications, Civil War at 150: Still Relevant, Still Divisive, data collected between March 30 - April 3, 2011.*

[47] *Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.*

[48] *S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: M.I.C Gadget (2012).*

6.16 Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.[49]

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a con dence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

6.17 Taste test. Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then lled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identi ed the soda.

(a) Do these data provide strong evidence that these people are able to detect the difference between diet and regular soda, in other words, are the results signi cantly better than just random guessing?

(b) Interpret the p-value in this context.

6.18 Is college worth it? Part II. Exercise 6.16 presents the results of a poll where 48% of 331 Americans who decide to not go to college do so because they cannot afford it.

(a) Calculate a 90% con dence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

(b) Suppose we wanted the margin of error for the 90% con dence level to be about 1.5%. How large of a survey would you recommend?

6.19 College smokers. We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke.

(a) Calculate a 95% con dence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: check conditions)

(b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

6.21 Public option, Part II. Exercise 6.13 presents the results of a poll evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.22 Acetaminophen and liver damage. It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher wants to conduct a study to estimate the proportion of acetaminophen users who have liver damage. For participating in this study, he will pay each subject $20 and provide a free medical consultation if the patient has liver damage.

(a) If he wants to limit the margin of error of his 98% confidence interval to 2%, what is the minimum amount of money he needs to set aside to pay his subjects?

(b) The amount you calculated in part (a) is substantially over his budget so he decides to use fewer subjects. How will this affect the width of his con dence interval?

[49]*Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.*

Difference of two proportions

6.23 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

6.24 Heart transplant success. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might bene t from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.[50]

	Control	treatment
alive dead	4 30	24 45

A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate a con dence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

6.25 Gender and color preference. A 2001 study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.[51]

(a) We are 95% con dent that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% con dent that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% con dence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a signi cant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% con dence interval for ($p_{female} - p_{male}$) cannot be calculated with only the information given in this exercise.

[50]*B. Turnbull et al. "Survivorship of Heart Transplant Data". In: Journal of the American Statistical Association 69 (1974), pp. 74 - 80.*

[51]*L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: Personality and Individual Differences 31.8 (2001), pp. 1375-1379.*

6.26 The Daily Show. A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. A 95% con dence interval for ($p_{collegegrad} - p_{HSorless}$), where p is the proportion of those who watch The Daily Show, is (0.07, 0.15). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[52]

(a) At the 5% significance level, the data provide convincing evidence of a difference between the proportions of college graduates and those with a high school degree or less who watch The Daily Show.

(b) We are 95% con dent that 7% less to 15% more college graduates watch The Daily Show than those with a high school degree or less.

(c) 95% of random samples of 1,099 college graduates and 1,110 people with a high school degree or less will yield differences in sample proportions between 7% and 15%.

(d) A 90% confidence interval for ($p_{collegegrad} - p_{HSorless}$) would be wider.

(e) A 95% confidence interval for ($p_{HSorless} - p_{collegegrad}$) is (-0.15,-0.07).

6.27 Public Option, Part III. Exercise 6.13 presents the results of a poll evaluating support for the health care public option plan in 2009. 70% of 819 Democrats and 42% of 783 Independents support the public option.

(a) Calculate a 95% confidence interval for the difference between ($p_D - p_I$) and interpret it in this context. We have already checked conditions for you.

(b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the public option than the Independent.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% con dence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[53]

6.29 Offshore drilling, Part I. A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[54]

	College Grad	
	Yes	No

Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

[52]*The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.*

[53]*CDC, Perceived Insu�cient Rest or Sleep Among Adults - United States, 2008.*

[54]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

6.30 Sleep deprivation, CA vs. OR, Part II. Exercise 6.28 provides data on sleep deprivation rates of Californians and Oregonians. The proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Conduct a hypothesis test to determine if these data provide strong evidence the rate of sleep deprivation is different for the two states. (Reminder: check conditions)

(b) It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

6.31 Offshore drilling, Part II. Results of a poll evaluating support for drilling for oil and natural gas off the coast of California were introduced in Exercise 6.29.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample support drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who support offshore drilling in California is different than that of noncollege graduates.

6.32 Full body scan, Part I. A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1,137 adults nationwide, interviewed by telephone November 7-10, 2010, where one of the questions on the survey was "Some airports are now using `full-body' digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?" Below is a summary of responses based on party affiliation.[55]

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22

Total	318	369	450

(a) Conduct an appropriate hypothesis test evaluating whether there is a difference in the proportion of Republicans and Democrats who think the full-body scans should be applied in airports. Assume that all relevant conditions are met.

(b) The conclusion of the test in part (a) may be incorrect, meaning a testing error was made. If an error was made, was it a Type I or a Type II error? Explain.

[55] *S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: CBS News (2010).*

6.33 Sleep deprived transportation workers. The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers. The results of the survey are shown below.[56]

		Transportation	Professionals		
	Control	Pilots	Truck Drivers	Train Operators	Bux/Taxi/Limo Drivers
Less than 6 hours of sleep 6 to 8 hours of sleep More than 8 hours	35 193 64	19 132 51	35 117 51	29 119 32	21 131 58
Tota	292	202	203	180	210

Conduct a hypothesis test to evaluate if these data provide evidence of a difference between the proportions of truck drivers and non-transportation workers (the control group) who get less than 6 hours of sleep per day, i.e. are considered sleep deprived.

6.34 Prenatal vitamins and Autism. Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period).[57]

	Autism		
	Autism	Typical development	Total
No vitamin Vitamin	111 143	70 159	181 302
Total	254	229	483

(a) State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.

(b) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

(c) A New York Times article reporting on this study was titled "Prenatal Vitamins May Ward Off Autism". Do you nd the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.[58]

6.35 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called virologic failure). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[59]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

[56]*National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers Sleep, 2012.*

[57]*R.J. Schmidt et al. \Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism". In: Epidemiology 22.4 (2011), p. 476.*

[58]*R.C. Rabin. "Patterns: Prenatal Vitamins May Ward Off Autism". In: New York Times (2011).*

[59]*S. Lockman et al. "Response to antiretroviral therapy after a single, peripartum dose of nevirapine". In: Obstetrical & gynecological survey 62.6 (2007), p. 361.*

6.36 Diabetes and unemployment. A 2012 Gallup poll surveyed Americans about their employment status and whether or not they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.[60]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of incidence of diabetes and employment status.

(c) The sample difference is about 1%. If we completed the hypothesis test, we would nd that the p-value is very small (about 0), meaning the difference is statistically signi cant. Use this result to explain the difference between statistically signi cant and practically significant findings.

Testing for goodness of t using chi-square

6.37 True or false, Part I. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

(c) The chi-square statistic is always positive.

(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

6.38 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found X^2 = 10 with df = 5 you would fail to reject H0 at the 5% signi cance level.

(c) When nding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.39 Open source textbook. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

(a) State the hypotheses for testing if the professor's predictions were inaccurate.

(b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

(c) This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

(d) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

(e) Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

[60]*Gallup Wellbeing, Employed Americans in Better Health Than the Unemployed, data collected Jan. 2, 2011 - May 21, 2012.*

6.40 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[61]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1) Humans evolved, but God had no part in process (2) God created humans in present form (3) Other / No opinion (4)	38% 16% 40% 6%	37% 12% 45% 6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)

Testing for independence in two-way tables

6.41 Quitters. Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

(a) Create a two-way table presenting the results of this study.

(b) Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

i. How many subjects in the "patch + support" group would you expect to quit?

ii. How many subjects in the "only patch" group would you expect to not quit?

6.42 Full body scan, Part II. The table below summarizes a data set we rst encountered in Exercise 6.32 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's party affiliation and his support of full-body scans. It may be useful to rst add on an extra column for row totals before proceeding with the computations.

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22
Total	318	369	450

(a) How many Republicans would you expect to not support the use of full-body scans?

(b) How many Democrats would you expect to support the use of full-body scans?

(c) How many Independents would you expect to not know or not answer?

[61] *Four in 10 Americans Believe in Strict Creationism, December 17, 2010, http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx.*

6.43 Offshore drilling, Part III. The table below summarizes a data set we rst encountered in Exercise 6.29 that examines the responses of a random sample of college graduates and nongraduates on the topic of oil drilling. Complete a chi-square test for

these data to check whether there is a statistically signi cant difference in responses from college graduates and non-graduates.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

6.44 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[62]

		Caffeinated	coffee	consumption		
	$\leq$ 1cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Yes No	670 11,545	373 6,244	905 16,329	564 11,726	95 2,288	2,607 48,132
Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identi ed in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $\frac{(Observed-Expected)^2}{Expected}$.

(e) The test statistic is $X^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.63 Do you agree with this statement? Explain your reasoning.

[62]*M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: Archives of internal medicine 171.17 (2011), p. 1571.*

[63]*A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: New York Times (2011).*

6.45 Privacy on Facebook. A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, "Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?" The responses are cross-tabulated based on gender.[64]

	Gender		
	Male	Female	Total
Yes No Not sure	288 61 10	378 62 7	666 123 17
Total	359	447	806

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

6.46 Shipping holiday gifts. A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age		
	18-34	35-54	55+	Total
USPS	72 (81)	97 (102)	76 (62)	245
UPS	52 (53)	76 (68)	34 (41)	162
FedEx	31 (21)	24 (27)	9 (16)	64
Something else	7 (5)	6 (7)	3 (4)	16
Not sure	3 (5)	6 (5)	4 (3)	13
Total	165	209	126	500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

(b) Are the conditions for inference using a chi-square test satisfied?

Small sample hypothesis testing for a proportion

6.47 Bullying in schools. A 2012 Survey USA poll asked Florida residents how big of a problem they thought bullying was in local schools. 9 out of 191 18-34 year olds responded that bullying is no problem at all. Using these data, is it appropriate to construct a con dence interval using the formula $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for the true proportion of 18-34 year old Floridians who think bullying is no problem at all? If it is appropriate, construct the con dence interval. If it is not, explain why.

[64]*Survey USA, News Poll #17960, data collected February 16-17, 2011.*

6.48 Choose a test. We would like to test the following hypotheses:

H_0 : p = 0.1

$H_A : p \neq 0.1$

The sample size is 120 and the sample proportion is 8.5%. Determine which of the below test(s) is/are appropriate for this situation and explain your reasoning.

I. Z test for a proportion,

i.e. proportion test using normal model

II. Z test for comparing two proportions

III. X^2 test of independence

IV. Simulation test for a proportion

V. t test for a mean

VI. ANOVA

6.49 The Egyptian Revolution. A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the rst couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.[65]

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

(c) Based on large sample theory, we modeled ^p using the normal distribution. Why should we be cautious about this approach for these data?

(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.

(e) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.

[65]*Gallup Politics, Americans' Views of Egypt Sharply More Negative, data collected February 2-5, 2011.*

6.50 Assisted Reproduction. Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g. in vitro fertilization). A 2008 report by the Centers for Disease Control and Prevention estimated that ART has been successful in leading to a live birth in 31% of cases66. A new fertility clinic claims that their success rate is higher than average. A random sample of 30 of their patients yielded a success rate of 40%. A consumer watchdog group would like to determine if this provides strong evidence to support the company's claim.

(a) Write the hypotheses to test if the success rate for ART at this clinic is signi cantly higher than the success rate reported by the CDC.

(b) Based on large sample theory, we modeled ^p using the normal distribution. Why is this not appropriate here?

(c) The normal approximation would be less reliable here, so we should use a simulation strategy. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

(d) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

(e) After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.

[66]*CDC. 2008 Assisted Reproductive Technology Report.*

Hypothesis testing for two proportions

6.51 Social experiment, Part II. Exercise 6.23 introduces a "social experiment" conducted by a TV program that questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

A simulation was conducted to test if people react differently under the two scenarios. 10,000 simulated differences were generated to construct the null distribution shown. The value $\hat{p}_{pr;sim}$ represents the proportion of diners who intervened in the simulation for the provocatively dressed woman, and $\hat{p}_{con;sim}$ is the proportion for the conservatively dressed woman.

(a) What are the hypotheses? For the purposes of this exercise, you may assume that each observed person at the restaurant behaved independently, though we would want to evaluate this assumption more rigorously if we were reporting these results.

(b) Calculate the observed difference between the rates of intervention under the provocative and conservative scenarios: $\hat{p}_{pr} - \hat{p}_{con}$.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

6.52 Is yawning contagious? An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously inuenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.[67]

	Group		
	Treatment	Control	Total
Yawn Not Yawn	10 24	4 12	14 36
Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no inuence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning as $\hat{p}_{trtmt;sim} - \hat{p}_{ctrl;sim}$. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.

(a) What are the hypotheses?

(b) Calculate the observed difference between the yawning rates under the two scenarios.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

[67]*MythBusters, Season 3, Episode 28.*

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

SECTION OVERVIEW

8.3: Confidence Intervals

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

8.3.1: Prelude to Confidence Intervals

> **Learning Objectives**
>
> By the end of this chapter, the student should be able to:
>
> - Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
> - Interpret the Student's t probability distribution as the sample size changes.
> - Discriminate between problems applying the normal and the Student's *t* distributions.
> - Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.

Figure 8.3.1.1. Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\bar{x}$ and s is called a statistic.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \tag{8.3.1.1}$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x}-0.2$ and $\bar{x}+0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x}=2$. Then the unknown population mean μ is between

$$\bar{x}-0.2=2-0.2=1.8 \tag{8.3.1.2}$$

and

$$\bar{x}+0.2=2+0.2=2.2 \tag{8.3.1.3}$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is (1.8, 2.2). This 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, $\bar{x}$. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

> Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Collaborative Exercise

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma=3$ and n = the number of students surveyed.
3. Construct the interval $\left(\bar{x}-2\cdot\frac{\sigma}{\sqrt{n}}, \bar{x}+2\cdot\frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between __________ and ___________.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

---

8.3.2: A Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated *EBM*). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ.

The confidence interval estimate will have the form:

$$(\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The **margin of error** (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

Example $8.3.2.1$

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5: $\bar{x} = 7$ and $EBM = 2.5$

The confidence interval is (7 – 2.5, 7 + 2.5) and calculating the values gives (4.5, 9.5). If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Exercise $8.3.2.1$

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2. What is the confidence interval estimate for the population mean?

Answer

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10$ in both tails, or 5% in each tail, of the normal distribution.

This is a normal distribution curve. The peak of the curve coincides with the point 10 on the horizontal axis. The points 5 and 15 are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 0.90.

Figure 8.3.2.1

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the *z*-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is

$$\frac{\sigma}{\sqrt{n}}$$

This fraction is commonly called the "standard error of the mean" to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N(\mu_x, \frac{\sigma}{\sqrt{n}})$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0,1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95, \alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

using a calculator, computer or a standard normal probability table.

`invNorm` $(0.975, 0, 1) = 1.96$

> Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0,1)$.

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EBM = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

This is a normal distribution curve. The peak of the curve coincides with the point x-bar on the horizontal axis. The points x-bar - EBM and x-bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 1 - a and represents the confidence level. Each unshaded tail has area a/2.

Figure 8.2.2.

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 8.3.2.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3; n = 36$$

The confidence level is 90% (CL = 0.90)

$$CL = 0.90$$

so

$$\alpha = 1 - CL = 1 - 0.90 = 0.10$$

$$\frac{\alpha}{2} = 0.05 z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using $\text{invNorm}(0.95, 0, 1)$ on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225).**

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `7:ZInterval` .
Press `ENTER` .
Arrow to `Stats` and press `ENTER` .
Arrow down and enter three for σ, 68 for $\bar{x}$, 36 for n, and .90 for `C-level` .
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

? Exercise 8.3.2.2

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.1347, 37.8653)

✓ Example 8.3.2.3: Specific Absorption Rate

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\bar{x} = 1.024$$

Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.

This is a normal distribution curve. The point z0.01 is labeled at the right edge of the curve and the region to the right of this point is shaded. The area of this shaded region equals 0.01. The unshaded area equals 0.99.

Figure 8.2.3.

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$\bar{x} - EBM = 1.024 – 0.1431 = 0.8809$

$\bar{x} - EBM = 1.024 – 0.1431 = 1.1671$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to 7:Z Interval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:
  - $\sigma : 0.337$
  - $\bar{x} : 1024$
  - $n : 30$
  - C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

? Exercise $8.3.2.3$

Table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Answer

$$\bar{x} = 0.940$$

$$\frac{\alpha}{2} = \frac{1-CL}{2} = \frac{1-0.93}{2} = 0.035$$

$$z_{0.035} = 1.812$$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$$\bar{x} - EBM = 0.940 - 0.1365 = 0.8035$$

$$\bar{x} + EBM = 0.940 + 0.1365 = 1.0765$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in Example and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

✓ Example $8.3.2.4$

Suppose we change the original problem in Example by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3; n = 36$; The confidence level is 95% (CL = 0.95).

$$CL = 0.95 \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\bar{x} - EBM = 68 - 0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

Part (a) shows a normal distribution curve. A central region with area equal to 0.90 is shaded. Each unshaded tail of the curve has area equal to 0.05. Part (b) shows a normal distribution curve. A central region with area equal to 0.95 is shaded. Each unshaded tail of the curve has area equal to 0.025.

Figure 8.2.4.

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

? Exercise $8.3.2.4$

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Answer

(33.37, 38.63)

Example 8.3.2.5

Suppose we change the original problem in Example to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (*CL*=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Answer

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100 : EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935.$

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25 : EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987.$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Exercise 8.3.2.5

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 8.3.2.6

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68 : $EBM = 68.82 - 68 = 0.82$.
- If we don't know the sample mean: $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

Exercise 8.3.2.6

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Answer

Sample mean is 45, error bound is 2.88

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

$$EBM = \left(z_{\frac{a}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n. In Equation ???, z is $z_{\frac{a}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 8.3.2.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2}$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

? Exercise 8.3.2.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Answer

35 students

References

1. "American Fact Finder." U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...html?refresh=t (accessed July 2, 2013).
2. "Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2, 2013).
3. "Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at research.fhda.edu/factbook/FH...phicTrends.htm (accessed September 30,2013).
4. Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at www.cdc.gov/growthcharts/2000...thchart-us.pdf (accessed July 2, 2013).
5. La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).
6. "Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...prodType=table (accessed July 2, 2013).
7. "Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at www.fec.gov/finance/disclosur...esummary.shtml (accessed July 2, 2013).
8. "National Health and Nutrition Examination Survey." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

Glossary

Confidence Level (CL)

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the $CL = 90$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

8.3.2E: A Single Population Mean using the Normal Distribution (Exercises)

8.3.3: A Single Population Mean using the Student t-Distribution

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t-distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ. If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score

$$t=\frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}, \tag{8.3.3.1}$$

then the t-scores follow a Student's t-distribution with $n-1$ degrees of freedom. The t-score has the same interpretation as the z-score. It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The degrees of freedom, $n-1$, come from the calculation of the sample standard deviation s. Previously, we used n deviations ($x-\bar{x}$ values) to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. We call the number $n-1$ the degrees of freedom (df).

> *For each sample size n, there is a different Student's t-distribution.*

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's t-Distribution.) When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t-table gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t-probabilities.**

The notation for the Student's t-distribution (using T as the random variable) is:

- $T \sim t_{df}$ where $df = n – 1$.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the error bound for a population mean is:

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n – 1$ degrees of freedom, and
- $s =$ sample standard deviation.

The format for the confidence interval is:

$$(\bar{x} - EBM, \bar{x} + EBM). \tag{8.3.3.2}$$

To calculate the confidence interval directly:

Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example 8.3.3.1: Acupuncture

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 8.2267$$

$$s = 1.6722 \; n = 15$$

$$df = 15 – 1 = 14 CL so \alpha = 1 – CL = 1 – 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 t_{\frac{\alpha}{2}} = t_{0.025}$$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$= (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

Now it is just a direct application of Equation 8.3.3.2:

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$
$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).
Arrow to `Data` and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq` .
Arrow down to `C-level` and enter 0.95
Arrow down to `Calculate` and press `ENTER` .
The 95% confidence interval is (7.3006, 9.1527)

When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

? Exercise 8.3.3.1

You do a study of hypnotherapy to determine how effective it is in increasing the number of hourse of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Answer

(8.1634, 9.8032)

✓ Example 8.3.3.2: The Human Toxome Project

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. Table 8.3.3.1 shows how many of the targeted chemicals were found in each infant's cord blood.

Table 8.3.3.1

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution A

From the sample, you can calculate $\bar{x} = 127.45$ and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$, so

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05} \tag{8.3.3.3}$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right) = 1.729 \left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$$\bar{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to Data and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90
Arrow down to `Calculate` and press `ENTER` .

The 90% confidence interval is (117.41, 137.49).

Example $8.3.3.3$

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in Table $8.3.3.2$ Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

Table 8.3.3.2

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution A

- $\bar{x} = 6.133$,
- $s = 5.514$,
- $n = 15$, and
- $df = 15 - 1 = 14$.

$$CL = 0.98, \text{ so } \alpha = 1 - CL = 1 - 0.98 = 0.02$$

$$\frac{\alpha}{2} = 0.01 t_{\frac{\alpha}{2}} = t_{0.01} 2.624$$

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right) = 2.624 \left(\frac{5.514}{\sqrt{15}}\right) - 3.736$$

$$\bar{x} - EBM = 6.133 - 3.736 = 2.397$$

$$\bar{x} + EBM = 6.133 + 3.736 = 9.869$$

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` .
Press `ENTER` .
Arrow to `Data` and press `ENTER` .
Arrow down and enter the name of the list where the data is stored.
Enter `Freq` : 1
Enter `C-Level` : 0.98
Arrow down to `Calculate` and press `Enter` .
The 98% confidence interval is (2.3965, 9,8702).

Reference

1. "America's Best Small Companies." Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).
2. Data from *Microsoft Bookshelf*.
3. Data from http://www.businessweek.com/.
4. Data from http://www.forbes.com/.
5. "Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2,2013).
6. "Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at www.ewg.org/sites/humantoxome...tero%2Fnewborn (accessed July 2, 2013).
7. "Metadata Description of Leadership PAC List." Federal Election Commission. Available online at www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

Glossary

Degrees of Freedom (df)

the number of objects in a sample that are free to vary

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* get larger.
- There is a "family" of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

8.4: PowerPoints

https://professormo.com/holistic/Powerpoint/ch08.pptx

CHAPTER OVERVIEW

9: Hypothesis Testing about Population Mean and Proportion

SECTION OVERVIEW

9.1: Inference for Numerical Data

Chapter 4 introduced a framework for statistical inference based on con dence intervals and hypotheses. In this chapter, we encounter several new point estimates and scenarios. In each case, the inference ideas remain the same:

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Apply the ideas from Chapter 4 using the distribution from step 2.

Each section in Chapter 5 explores a new situation: the difference of two means (5.1, 5.2); a single mean or difference of means where we relax the minimum sample size condition (5.3, 5.4); and the comparison of means across multiple groups (5.5). Chapter 6 will introduce scenarios that highlight categorical data.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

9.1.1: One-Sample Means with the t Distribution

The motivation in Chapter 4 for requiring a large sample was two-fold. First, a large sample ensures that the sampling distribution of $\bar{x}$ is nearly normal. We will see in Section 5.3.1 that if the population data are nearly normal, then $\bar{x}$ is also nearly normal regardless of the

[10] *The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances σ_x^2 and σ_y^2, then the variance of X - Y is $\sigma_x^2 + \sigma_y^2$. Likewise, the variance corresponding to $\bar{x}_1 - \bar{x}_2$ is $\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$. Because $\sigma_{\bar{x}_1}^2$ and $\sigma_{\bar{x}_2}^2$ are just another way of writing $SE_{\bar{x}_1}^2$ and $SE_{\bar{x}_2}^2$, the variance associated with $\bar{x}_1 - \bar{x}_2$ may be written as $SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2$*

sample size. The second motivation for a large sample was that we get a better estimate of the standard error when using a large sample. The standard error estimate will not generally be accurate for smaller sample sizes, and this motivates the introduction of the t distribution, which we introduce in Section 5.3.2.

We will see that the t distribution is a helpful substitute for the normal distribution when we model a sample mean $\bar{x}$ that comes from a small sample. While we emphasize the use of the t distribution for small samples, this distribution may also be used for means from large samples.

The normality condition

We use a special case of the Central Limit Theorem to ensure the distribution of the sample means will be nearly normal, regardless of sample size, provided the data come from a nearly normal distribution.

Central Limit Theorem for normal data

The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

While this seems like a very helpful special case, there is one small problem. It is inherently difficult to verify normality in small data sets.

Caution: Checking the normality condition

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I con dent that outliers are rare?

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

Introducing the t distribution

The second reason we previously required a large sample size was so that we could accurately estimate the standard error using the sample data. In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t distribution. A t distribution, shown as a solid line in Figure 5.10, has a bell shape. However, its tails are thicker than the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[11] These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell shaped t distribution.

[11] *The standard deviation of the t distribution is actually a little more than 1. However, it is useful to always think of the t distribution as having a standard deviation of 1 in all of our applications.*

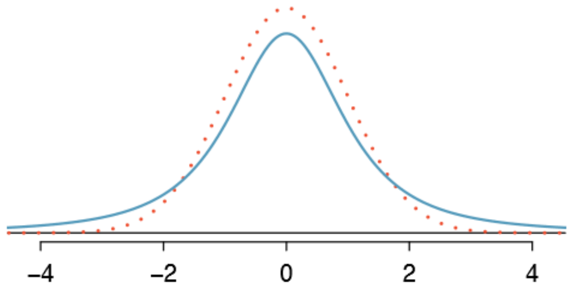


Figure 5.10: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

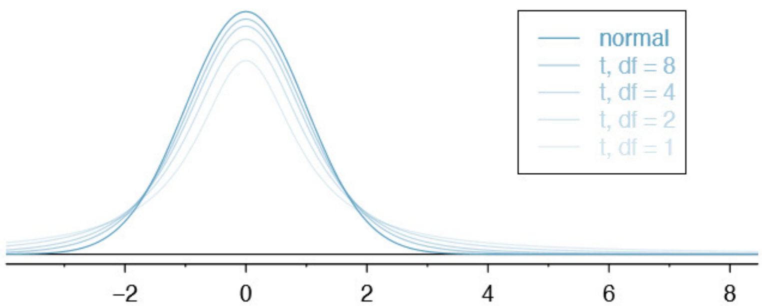


Figure 5.11: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

Several t distributions are shown in Figure 5.11. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution.

Degrees of freedom (df)

The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution. In Section 5.3.3, we relate degrees of freedom to sample size.

We will nd it very useful to become familiar with the t distribution, because it plays a very similar role to the normal distribution during inference for small samples of numerical data. We use a t table, partially shown in Table 5.12, in place of the normal probability table for small sample numerical data. A larger table is presented in Appendix B.2 on page 410.

Each row in the t table represents a t distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t distribution with df = 18, we can examine row 18, which is highlighted in

one tail two tails	0.100 0.200	0.050 0.100	0.025 0.050	0.010 0.020	0.005 0.010

df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	3 5.84
⋮	⋮	⋮	⋮	⋮	⋮
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	⋮
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

Table 5.12: An abbreviated look at the t table. Each row represents a different t distribution. The columns describe the cutoffs for specific tail areas. The row with df = 18 has been highlighted.

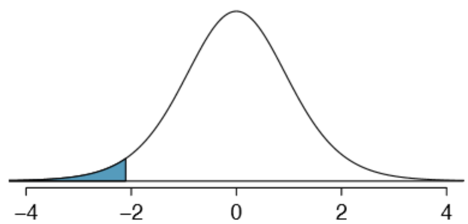


Figure 5.13: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where one tail is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t distributions are symmetric.

Example 5.15 What proportion of the t distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 5.13 and shade the area below -2.10. To nd this area, we identify the appropriate row: df = 18. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

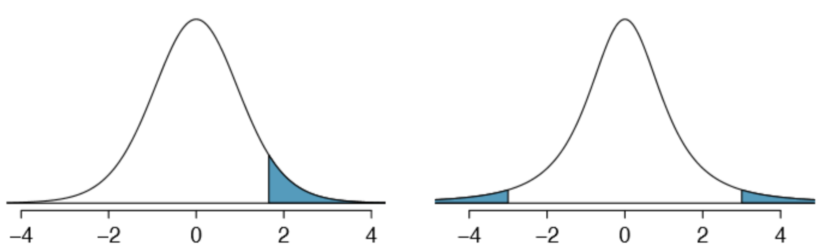


Figure 5.14: Left: The t distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

Example 5.16 A t distribution with 20 degrees of freedom is shown in the left panel of Figure 5.14. Estimate the proportion of the distribution falling above 1.65.

We identify the row in the t table using the degrees of freedom: df = 20. Then we look for 1.65; it is not listed. It falls between the rst and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the rst and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

Example 5.17 A t distribution with 2 degrees of freedom is shown in the right panel of Figure 5.14. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: df = 2. Next, nd the columns that capture 3; because 2:92 < 3 < 4:30, we use the second and third columns. Finally, we nd bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

Exercise 5.18 What proportion of the t distribution with 19 degrees of freedom falls above -1.79 units?[12]

The t distribution as a solution to the standard error problem

When estimating the mean and standard error from a small sample, the t distribution is a more accurate tool than the normal model. This is true for both small and large samples.

TIP: When to use the t distribution

Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

[12] *We find the shaded area above -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.*

To proceed with the t distribution for inference about a single mean, we must check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if it was an experiment or random process, we carefully check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t distribution with n - 1 degrees of freedom (df). For example, if the sample size was 19, then we would use the t distribution with df = 19 - 1 = 18 degrees of freedom and proceed exactly as we did in Chapter 4, except that now we use the t table.

One sample t confidence intervals

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

Figure 5.15: A Risso's dolphin. Photo by Mike Baird (http://www.bairdphotos.com/).

Here we identify a con dence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[13] The data are summarized in Table 5.16. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

[13]Taiji was featured in the movie The Cove, and it is a signi cant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins. Data reference: Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 5.16: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Example 5.19 Are the independence and normality conditions satis ed for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. The summary statistics in Table 5.16 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

In the normal model, we used z* and the standard error to determine the width of a con dence interval. We revise the confidence interval formula slightly when using the t distribution:

$$\bar{x} \pm t*_{df} SE \tag{9.1.1.1}$$

The sample mean and estimated standard error are computed just as before ($\bar{x}$ = 4.4 and $SE = \frac{s}{\sqrt{n}} = 0.528$). The value $t*_{df}$ is a cutoff we obtain based on the con dence level and the t distribution with df degrees of freedom. Before determining this cutoff , we will first need the degrees of freedom.

Degrees of freedom for a single sample

If the sample has n observations and we are examining a single mean, then we use the t distribution with df = n - 1 degrees of freedom.

In our current example, we should use the t distribution with df = 19 -1 = 18 degrees of freedom. Then identifying $t*_{18}$ is similar to how we found z*.

• For a 95% confidence interval, we want to nd the cutoff $t*_{18}$ such that 95% of the t distribution is between $-t*_{18}$ and$t*_{18}$.

• We look in the t table on page 224, nd the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t*_{18} = 2.10$.

Generally the value of $t*_{df}$ is slightly larger than what we would get under the normal model with z*.

Finally, we can substitute all our values into the con dence interval equation to create the 95% con dence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \pm t*_{18} SE \rightarrow 4.4 \pm 2.10 \times 0.528 \rightarrow (3.87, 4.93) \tag{9.1.1.2}$$

We are 95% con dent the average mercury content of muscles in Risso's dolphins is between 3.87 and 4.93 μg/wet gram. This is above the Japanese regulation level of 0.4 μg/wet gram.

Finding a t confidence interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \pm t*_{df} SE \tag{9.1.1.3}$$

where $\bar{x}$ is the sample mean, $t*_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

Exercise 5.20 The FDA's webpage provides some data on mercury content of sh.14 Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[15]

Example 5.21 Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Exercise 5.20. If we are to use the t distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t*_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: df = n - 1 = 14.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row df = 14, we identify $t*_{14} = 1.76$.

Exercise 5.22 Using the results of Exercise 5.20 and Example 5.21, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[16]

One sample t tests

An SAT preparation company claims that its students' scores improve by over 100 points on average after their course. A consumer group would like to evaluate this claim, and they collect data on a random sample of 30 students who took the class. Each of these students took the SAT before and after taking the company's course, and so we have a difference in scores for each student. We will examine these differences $x_1 = 57, x_2 = 133, \ldots, x_30 = 140$ as a sample to evaluate the company's claim. (This is paired data, so we analyze the score differences; for a review of the ideas of paired data, see Section 5.1.) The distribution of the differences, shown in Figure 5.17, has mean 135.9 and standard deviation 82.2. Do these data provide convincing evidence to back up the company's claim?

Exercise 5.23 Set up hypotheses to evaluate the company's claim. Use μ_{diff} to represent the true average difference in student scores.[17]

[14]www.fda.gov/Food/FoodSafety/P...bornePathogens

contaminants/Methylmercury/ucm115644.htm

[15]There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red ags for the normal model based on this (limited) information, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

[16]$\bar{x} \pm t*_{14}SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. *We are 90% confident that the average mercury content of croaker white sh (Pacific) is between 0.256 and 0.318 ppm.*

[17]This is a one-sided test. H_0: student scores do not improve by more than 100 after taking the company's course. $\mu_{diff} = 100$ *(we always write the null hypothesis with an equality). H_A: students scores improve by more than 100 points on average after taking the company's course.* $\mu_{diff} > 100$.

Exercise 5.24 Are the conditions to use the t distribution method satisfied?[18]

Just as we did for the normal case, we standardize the sample mean using the Z score to identify the test statistic. However, we will write T instead of Z, because we have a small sample and are basing our inference on the t distribution:

$$T = \frac{\bar{x} - nullvalue}{SE} = \frac{135.9 - 100}{\frac{82.2}{\sqrt{30}}} = 2.39 \tag{9.1.1.4}$$

If the null hypothesis was true, the test statistic T would follow a t distribution with df = n - 1 = 29 degrees of freedom. We can draw a picture of this distribution and mark the observed T, as in Figure 5.18. The shaded right tail represents the p-value: the probability of observing such strong evidence in favor of the SAT company's claim, if the average student improvement is really only 100.

[18]This is a random sample from less than 10% of the company's students (assuming they have more than 300 former students), so the independence condition is reasonable. The normality condition also seems reasonable based on Figure 5.17. We can use the t distribution method. Note that we could use the normal distribution. However, since the sample size (n = 30) just meets the threshold for reasonably estimating the standard error, it is advisable to use the t distribution.

Exercise 5.25 Use the t table in Appendix B.2 on page 410 to identify the p-value. What do you conclude?[19]

Exercise 5.26 Because we rejected the null hypothesis, does this mean that taking the company's class improves student scores by more than 100 points on average?[20]

9.1.2: Paired Data

Are textbooks actually cheaper online? Here we compare the price of textbooks at UCLA's bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses (when a class had multiple books, only the most expensive text was considered). A portion of this data set is shown in Table $9.1.2.1$.

Table $9.1.2.1$: Six cases of the textbooks data set.

	dept	course	ucla	amazon	diff
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

Paired Observations and Samples

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

Paired data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the textbook data set, we look at the difference in prices, which is represented as the diff variable in the textbooks data. Here the differences are taken as

$$\text{UCLA price} - \text{Amazon price} \tag{9.1.2.1}$$

for each book. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure $9.1.2.1$. Using differences between paired observations is a common and useful way to analyze paired data.

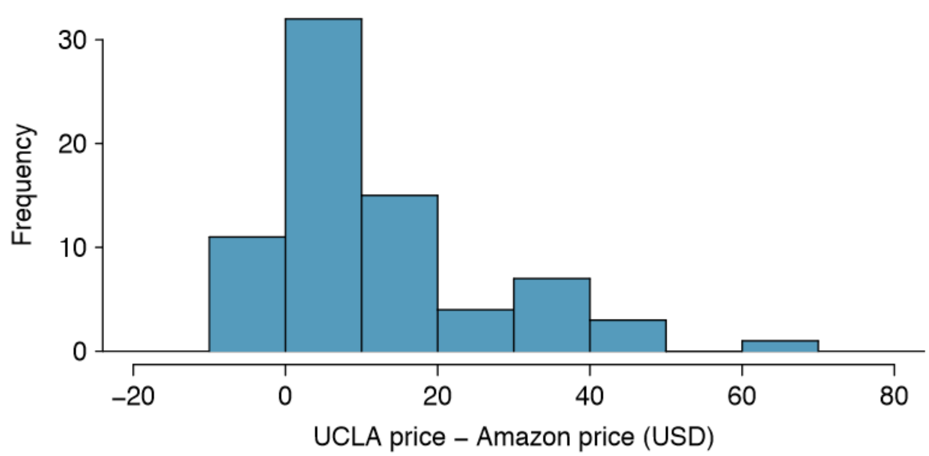


Figure 9.1.2.1: Histogram of the difference in price for each book sampled. These data are strongly skewed.

Exercise 9.1.2.1

The first difference shown in Table 9.1.2.1 is computed as 27.67 - 27.95 = -0.28. Verify the differences are calculated correctly for observations 2 and 3.

Solution

- Observation 2: 40.59 - 31.14 = 9.45.
- Observation 3: 31.68 - 32.00 = -0.32.

Inference for Paired Data

To analyze a paired data set, we use the exact same tools that we developed in Chapter 4. Now we apply them to the differences in the paired observations.

Table 9.1.2.1: Summary statistics for the price differences. There were 73 books, so there are 73 differences.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
73	12.76	14.26

Example 9.1.2.1: UCLA vs. Amazon

Set up and implement a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price.

Solution

There are two scenarios: there is no difference or there is some difference in average prices. The no difference scenario is always the null hypothesis:

- H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.
- H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Can the normal model be used to describe the sampling distribution of $\bar{x}_{diff}$? We must check that the differences meet the conditions established in Chapter 4. The observations are based on a simple random sample from less than 10% of all books sold at the bookstore, so independence is reasonable; there are more than 30 differences; and the distribution of differences, shown in Figure 9.1.2.1, is strongly skewed, but this amount of skew is reasonable for this sized data set (n = 73). Because all three conditions are reasonably satisfied, we can conclude the sampling distribution of $\bar{x}_{diff}$ nearly normal and our estimate of the standard error will be reasonable.

We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences (s_{diff} = 14.26) and the number of differences (n_{diff} = 73):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{14.26}{\sqrt{73}} = 1.67 \tag{9.1.2.2}$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, which is shown in Figure 9.1.2.1. The p-value is represented by the two (very) small tails.

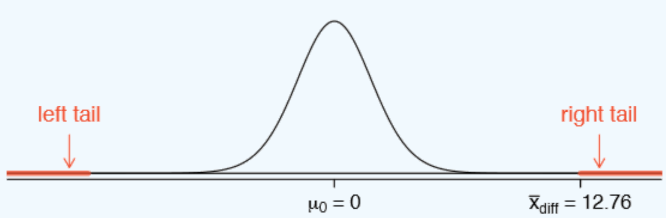


Figure 9.1.2.2: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

To find the tail areas, we compute the test statistic, which is the Z score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$Z=\frac{\bar{x}_{diff}-0}{SE_{\bar{x}_{diff}}}=\frac{12.76-0}{1.67}=7.59 \tag{9.1.2.3}$$

This Z score is so large it is not even in the table, which ensures the single tail area will be 0.0002 or smaller. Since the p-value corresponds to both tails in this case and the normal distribution is symmetric, the p-value can be estimated as twice the one-tail area:

$$\text{p-value}=2\times(\text{one tail area})\approx 2\times 0.0002=0.0004 \tag{9.1.2.4}$$

Because the p-value is less than 0.05, we reject the null hypothesis. We have found convincing evidence that Amazon is, on average, cheaper than the UCLA bookstore for UCLA course textbooks.

Exercise $9.1.2.1$

Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.

Solution

Conditions have already verified and the standard error computed in Example $9.1.2.1$ To find the interval, identify z^* (1.96 for 95% confidence) and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate}\pm z^* SE\rightarrow 12.76\pm 1.96\times 1.67\rightarrow(9.49,16.03) \tag{9.1.2.5}$$

We are 95% confident that Amazon is, on average, between $9.49 and $16.03 cheaper than the UCLA bookstore for UCLA course books.

9.1.3: Difference of Two Means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. The methods are similar in theory but different in the details. Just as with a single sample, we identify conditions to ensure a point estimate of the difference $\bar{x}_1 - \bar{x}_2$ is nearly normal. Next we introduce a formula for the standard error, which allows us to apply our general tools from Section 4.5.

We apply these methods to two examples: participants in the 2012 Cherry Blossom Run and newborn infants. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

Point Estimates and Standard Errors for Differences of Means

We would like to estimate the average difference in run times for men and women using the run10Samp data set, which was a simple random sample of 45 men and 55 women from all runners in the 2012 Cherry Blossom Run. Table $9.1.3.2$presents relevant summary statistics, and box plots of each sample are shown in Figure 5.6.

Table $9.1.3.2$: Summary statistics for the run time of 100 participants in the 2009 Cherry Blossom Run.

	men	women
$\bar{x}$	87.65	102.13
s	12.5	15.2
n	45	55

The two samples are independent of one-another, so the data are not paired. Instead a point estimate of the difference in average 10 mile times for men and women, $\mu_w - \mu_m$, can be found using the two sample means:

$$\bar{x}_w - \bar{x}_m = 102.13 - 87.65 = 14.48 \tag{9.1.3.1}$$

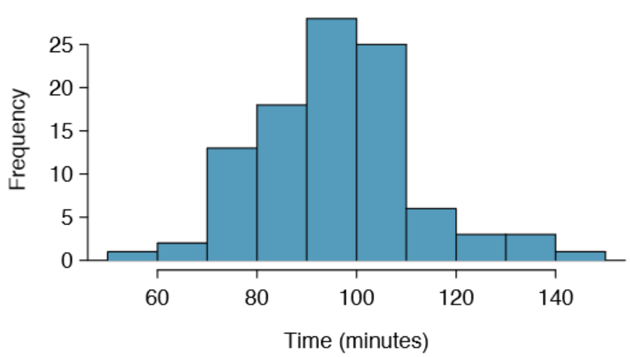


Figure $9.1.3.1$: *A histogram of time for the sample Cherry Blossom Race data.*

Because we are examining two simple random samples from less than 10% of the population, each sample contains at least 30 observations, and neither distribution is strongly skewed, we can safely conclude the sampling distribution of each sample mean is nearly normal. Finally, because each sample is independent of the other (e.g. the data are not paired), we can conclude that the difference in sample means can be modeled using a normal distribution. (Probability theory guarantees that the difference of two independent normal random variables is also normal. Because each sample mean is nearly normal and observations in the samples are independent, we are assured the difference is also nearly normal.)

Conditions for normality of $\bar{x}_1 - \bar{x}_2$

If the sample means, $\bar{x}_1$ and $\bar{x}_2$, each meet the criteria for having nearly normal sampling distributions and the observations in the two samples are independent, then the difference in sample means, $\bar{x}_1 - \bar{x}_2$, will have a sampling distribution that is nearly normal.

We can quantify the variability in the point estimate, $\bar{x}_w - \bar{x}_m$, using the following formula for its standard error:

$$SE_{\bar{x}_w - \bar{x}_m} = \sqrt{\frac{\sigma_w^2}{n_w} + \frac{\sigma_m^2}{n_m}} \tag{9.1.3.2}$$

We usually estimate this standard error using standard deviation estimates based on the samples:

$$SE_{\bar{x}_w - \bar{x}_m} \approx \sqrt{\frac{s_w^2}{n_w} + \frac{s_m^2}{n_m}} \tag{9.1.3.3}$$

$$= \sqrt{\frac{15.2^2}{55} + \frac{12.5^2}{45}} \tag{9.1.3.4}$$

$$= 2.77 \tag{9.1.3.5}$$

Because each sample has at least 30 observations ($n_w = 55$ and $n_m = 45$), this substitution using the sample standard deviation tends to be very good.

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, is nearly normal with mean $\mu_1 - \mu_2$ and estimated standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{9.1.3.6}$$

when each sample mean is nearly normal and all observations are independent.

Confidence Interval for the Difference

When the data indicate that the point estimate $\bar{x}_1 - \bar{x}_2$ comes from a nearly normal distribution, we can construct a confidence interval for the difference in two means from the framework built in Chapter 4. Here a point estimate, $\bar{x}_w - \bar{x}_m = 14.48$, is associated with a normal model with standard error SE = 2.77. Using this information, the general confidence interval formula may be applied in an attempt to capture the true difference in means, in this case using a 95% confidence level:

$$\text{point estimate} \pm z^* SE \rightarrow 14.48 \pm 1.96 \times 2.77 = (9.05, 19.91) \tag{9.1.3.7}$$

Based on the samples, we are 95% confident that men ran, on average, between 9.05 and 19.91 minutes faster than women in the 2012 Cherry Blossom Run.

Exercise 9.1.3.1

What does 95% confidence mean?

Solution

If we were to collected many such samples and create 95% confidence intervals for each, then about 95% of these intervals would contain the population difference, $\mu_w - \mu_m$.

Exercise 9.1.3.2

We may be interested in a different confidence level. Construct the 99% confidence interval for the population difference in average run times based on the sample data.

Solution

The only thing that changes is z*: we use z* = 2:58 for a 99% confidence level. (If the selection of z^* is confusing, see Section 4.2.4 for an explanation.) The 99% confidence interval:

$$14.48 \pm 2.58 \times 2.77 \rightarrow (7.33, 21.63). \tag{9.1.3.8}$$

We are 99% confident that the true difference in the average run times between men and women is between 7.33 and 21.63 minutes.

Hypothesis tests Based on a Difference in Means

A data set called baby smoke represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table 9.1.3.2 We are particularly interested in two variables: weight and smoke. The weight variable represents the weights of the newborns and the smoke variable describes which mothers smoked during pregnancy. We would like to know if there is convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 9.1.3.2

Table 9.1.3.2: Four cases from the baby smoke data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	⋮
150	45	50	36	9.25	female	nonsmoker

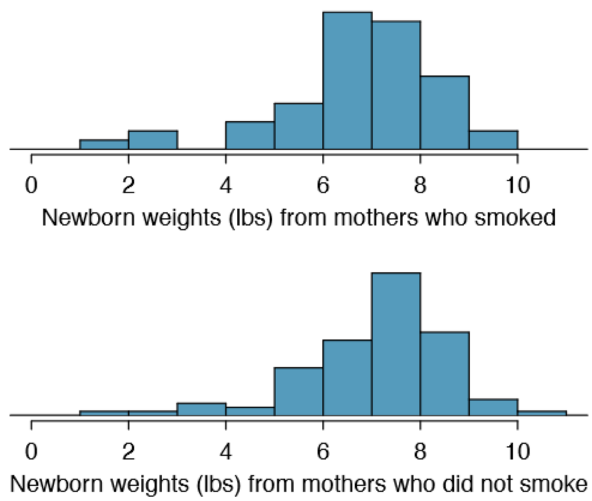


Figure 9.1.3.2: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. Both distributions exhibit strong skew.

Example 9.1.3.1

Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

Solution

The null hypothesis represents the case of no difference between the groups.

- H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.
- H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

Summary statistics are shown for each sample in Table 9.1.3.3 Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, each group's sample size is at least

30 and the skew in each sample distribution is strong (Figure $9.1.3.2$). However, this skew is reasonable for these sample sizes of 50 and 100. Therefore, each sample mean is associated with a nearly normal distribution.

Table $9.1.3.3$: Summary statistics for the baby smoke data set.

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Exercise $9.1.3.3$

a. What is the point estimate of the population difference, $\mu_n - \mu_s$?
b. Can we use a normal distribution to model this difference?
c. Compute the standard error of the point estimate from part (a)

Solution

(a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$.

(b) Because the samples are independent and each sample mean is nearly normal, their difference is also nearly normal.

(c) The standard error of the estimate can be estimated using Equation 9.1.3.6:

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26 \tag{9.1.3.9}$$

The standard error estimate should be sufficiently accurate since the conditions were reasonably satisfied.

Example $9.1.3.2$

If the null hypothesis from Exercise 5.8 was true, what would be the expected value of the point estimate? And the standard deviation associated with this estimate? Draw a picture to represent the p-value.

Solution

If the null hypothesis was true, then we expect to see a difference near 0. The standard error corresponds to the standard deviation of the point estimate: 0.26. To depict the p-value, we draw the distribution of the point estimate as though H_0 was true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.

Example $9.1.3.3$

Compute the p-value of the hypothesis test using the figure in Example 5.9, and evaluate the hypotheses using a signi cance level of $\alpha = 0.05$.

Solution

Since the point estimate is nearly normal, we can nd the upper tail using the Z score and normal probability table:

$$Z = \frac{0.40 - 0}{0.26} = 1.54 \rightarrow \text{upper tail} = 1 - 0.938 = 0.062 \tag{9.1.3.10}$$

Because this is a two-sided test and we want the area of both tails, we double this single tail to get the p-value: 0.124. This p-value is larger than the signi cance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

Exercise 9.1.3.4

Does the conclusion to Example 5.10 mean that smoking and average birth weight are unrelated?

Solution

Absolutely not. It is possible that there is some difference but we did not detect it. If this is the case, we made a Type 2 Error.

Exercise 9.1.3.5

If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect such a difference?

Solution

We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists.

Summary for inference of the difference of two means

When considering the difference of two means, there are two common cases: the two samples are paired or they are independent. (There are instances where the data are neither paired nor independent.) The paired case was treated in Section 5.1, where the one-sample methods were applied to the differences from the paired observations. We examined the second and more complex scenario in this section.

When applying the normal model to the point estimate $\bar{x}_1 - \bar{x}_2$ (corresponding to unpaired data), it is important to verify conditions before applying the inference framework using the normal model. First, each sample mean must meet the conditions for normality; these conditions are described in Chapter 4 on page 168. Secondly, the samples must be collected independently (e.g. not paired data). When these conditions are satisfied, the general inference tools of Chapter 4 may be applied.

For example, a confidence interval may take the following form:

$$\text{point estimate} \pm z^* SE \tag{9.1.3.11}$$

When we compute the confidence interval for $\mu_1 - \mu_2$, the point estimate is the difference in sample means, the value z^* corresponds to the confidence level, and the standard error is computed from Equation 9.1.3.6. While the point estimate and standard error formulas change a little, the framework for a confidence interval stays the same. This is also true in hypothesis tests for differences of means.

In a hypothesis test, we apply the standard framework and use the specific formulas for the point estimate and standard error of a difference in two means. The test statistic represented by the Z score may be computed as

$$Z = \frac{\text{point estimate - null value}}{SE} \tag{9.1.3.12}$$

When assessing the difference in two means, the point estimate takes the form $\bar{x}_1 - \bar{x}_2$, and the standard error again takes the form of Equation 9.1.3.6. Finally, the null value is the difference in sample means under the null hypothesis. Just as in Chapter 4, the test statistic Z is used to identify the p-value.

Examining the Standard Error Formula

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}} \tag{9.1.3.13}$$

where s_1 and n_1 represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{9.1.3.14}$$

This special relationship follows from probability theory.

Exercise 9.1.3.6

Prerequisite: Section 2.4. We can rewrite Equation 9.1.3.14in a different way:

$$SE_{\bar{x}_1-\bar{x}_2}^2 = SE_{\bar{x}_1}^2 + SE_{barx_2}^2 \tag{9.1.3.15}$$

Explain where this formula comes from using the ideas of probability theory.[10]

9.1.4: Power Calculations for a Difference of Means (Special Topic)

It is also useful to be able to compare two means for small samples. For instance, a teacher might like to test the notion that two versions of an exam were equally difficult. She could do so by randomly assigning each version to students. If she found that the average scores on the exams were so different that we cannot write it off as chance, then she may want to award extra points to students who took the more difficult exam.

In a medical context, we might investigate whether embryonic stem cells can improve heart pumping capacity in individuals who have suffered a heart attack. We could look for evidence of greater heart health in the stem cell group against a control group.

In this section we use the t distribution for the difference in sample means. We will again drop the minimum sample size condition and instead impose a strong condition on the distribution of the data.

Sampling Distributions for the Difference in two Means

In the example of two exam versions, the teacher would like to evaluate whether there is convincing evidence that the difference in average scores between the two exams is not due to chance.

It will be useful to extend the t distribution method from Section 5.3 to apply to a difference of means:

$$\bar{x}_1 - \bar{x}_2 \tag{9.1.4.1}$$

as a point estimate for

$$\mu_1 - \mu_2 \tag{9.1.4.2}$$

Our procedure for checking conditions mirrors what we did for large samples in Section 5.2. First, we verify the small sample conditions (independence and nearly normal data) for each sample separately, then we verify that the samples are also independent. For instance, if the teacher believes students in her class are independent, the exam scores are nearly normal, and the students taking each version of the exam were independent, then we can use the t distribution for inference on the point estimate $\bar{x}_1 - \bar{x}_2$.

The formula for the standard error of $\bar{x}_1 - \bar{x}_2$., introduced in Section 5.2, also applies to small samples:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.27}$$

[19]We use the row with 29 degrees of freedom. The value T = 2.39 falls between the third and fourth columns. Because we are looking for a single tail, this corresponds to a p-value between 0.01 and 0.025. The p-value is guaranteed to be less than 0.05 (the default signi cance level), so we reject the null hypothesis. The data provide convincing evidence to support the company's claim that student scores improve by more than 100 points following the class.

[20]This is an observational study, so we cannot make this causal conclusion. For instance, maybe SAT test takers tend to improve their score over time even if they don't take a special SAT class, or perhaps only the most motivated students take such SAT courses.

Because we will use the t distribution, we will need to identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will apply in the examples and exercises.[21]

> **Using the t distribution for a difference in means**
>
> The t distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t distribution and (2) the samples are independent. We estimate the standard error of the difference of two means using Equation ???.

Two Sample t test

Summary statistics for each exam version are shown in Table 5.19. The teacher would like to evaluate whether this difference is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Table 5.19: Summary statistics of scores for each exam version.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Exercise 9.1.4.1

Construct a two-sided hypothesis test to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, might be due to chance.

Solution

Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. HA: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

Exercise 9.1.4.1

To evaluate the hypotheses in Exercise 5.28 using the t distribution, we must first verify assumptions.

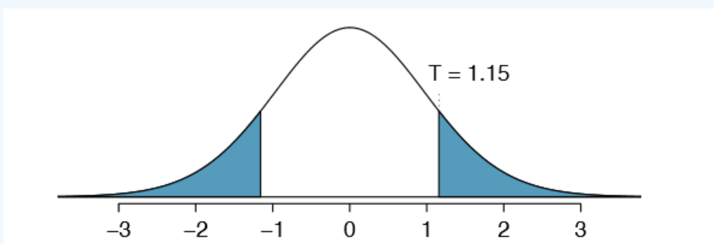
a. Does it seem reasonable that the scores are independent within each group?
b. What about the normality condition for each group?
c. Do you think scores from the two groups would be independent of each other (i.e. the two samples are independent)?[23]

Solution

(a) It is probably reasonable to conclude the scores are independent.

(b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable.

(c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.



After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated using Equation ???:

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62 \tag{9.1.4.3}$$

[21]This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method.

Figure 5.20: The t distribution with 26 degrees of freedom. The shaded right tail represents values with T ≥ 1.15. Because it is a two-sided test, we also shade the corresponding lower tail.

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate - null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15 \tag{9.1.4.4}$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1 \text{and} n_2 - 1$: df = 26.

Exercise 9.1.4.1

Exercise 5.30 Identify the p-value, shown in Figure 5.20. Use df = 26.

Solution

We examine row df = 26 in the t table. Because this value is smaller than the value in the left column, the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

In Exercise 5.30, we could have used df = 45.97. However, this value is not listed in the table. In such cases, we use the next lower degrees of freedom (unless the computer also provides the p-value). For example, we could have used df = 45 but not df = 46.

Exercise 9.1.4.1

Do embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 5.21 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured. A positive value generally corresponds to increased pumping capacity, which suggests a stronger recovery.

a. Set up hypotheses that will be used to test whether there is convincing evidence that ESCs actually increase the amount of blood the heart pumps.
b. Check conditions for using the t distribution for inference with the point estimate $\bar{x}_1 - \bar{x}_2$. To assist in this assessment, the data are presented in Figure 5.22.[25]

Solution

(a) We first setup the hypotheses:

- H_0: The stem cells do not improve heart pumping function. $\mu_{esc} - \mu_{control} = 0$.
- H_A: The stem cells do improve heart pumping function. $\mu_{esc} - \mu_{control} > 0$.

(b) Because the sheep were randomly assigned their treatment and, presumably, were kept separate from one another, the independence assumption is reasonable for each sample as well as for between samples. The data are very limited, so we can only check for obvious outliers in the raw data in Figure 5.22. Since the distributions are (very) roughly symmetric, we will assume the normality condition is acceptable. Because the conditions are satisfied, we can apply the t distribution.

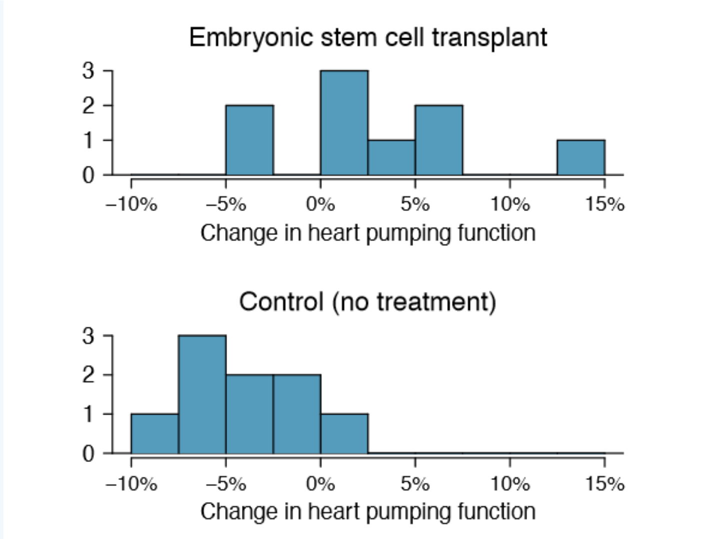


Figure 5.22: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

Table 5.21: Summary statistics of scores, split by exam version.

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Figure 5.23: Distribution of the sample difference of the test statistic if the null hypothesis was true. The shaded area, hardly visible in the right tail, represents the p-value.

Example 9.1.4.1

Use the data from Table 5.21 and df = 8 to evaluate the hypotheses for the ESC experiment described in Exercise 5.31.

Solution

First, we compute the sample difference and the standard error for that point estimate:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.88 \tag{9.1.4.5}$$

$$SE = \frac{\frac{5.17^2}{9} + \frac{2.76^2}{9}}{=} 1.95 \tag{9.1.4.6}$$

The p-value is depicted as the shaded slim right tail in Figure 5.23, and the test statistic is computed as follows:

$$T = \frac{7.88 - 0}{1.95} = 4.03 \tag{9.1.4.7}$$

We use the smaller of $n_1 - 1$ and $n_2 - 1$ (each are the same) for the degrees of freedom: df = 8. Finally, we look for T = 4.03 in the t table; it falls to the right of the last column, so the p-value is smaller than 0.005 (one tail!). Because the p-value is less than 0.005 and therefore also smaller than 0.05, we reject the null hypothesis. The data provide convincing evidence that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack.

Two sample t confidence interval

The results from the previous section provided evidence that ESCs actually help improve the pumping function of the heart. But how large is this improvement? To answer this question, we can use a confidence interval.

Exercise 9.1.4.1

In Exercise 5.31, you found that the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.88$, has a standard error of 1.95. Using df = 8, create a 99% confidence interval for the improvement due to ESCs.

Solution

We know the point estimate, 7.88, and the standard error, 1.95. We also veri ed the conditions for using the t distribution in Exercise 5.31. Thus, we only need identify $t*_8$ to create a 99% con dence interval: $t*_8 = 3.36$. The 99% con dence interval for the improvement from ESCs is given by

$$\text{point estimate} \pm t*_8 SE \rightarrow 7.88 \pm 3.36 \times 1.95 \rightarrow (1.33, 14.43) \tag{9.1.4.8}$$

That is, we are 99% con dent that the true improvement in heart pumping function is somewhere between 1.33% and 14.43%.

Pooled Standard Deviation Estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make our t distribution approach slightly more precise by using a pooled standard deviation. The pooled standard deviation of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s1 and s2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s^2_{pooled} = \frac{s^2_1 \times (n_1 - 1) + s^2_2 \times (n_2 - 1)}{n_1 + n_2 - 2} \tag{9.1.4.9}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} in place of s^2_1 and s^2_2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2 \tag{9.1.4.10}$$

The bene ts of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}^2_1 - \bar{x}^2_2$

Caution: Pooling standard deviations should be done only after careful research

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

9.1.5: Comparing many Means with ANOVA (Special Topic)

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually nd a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

- H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.
- H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

Example 9.1.5.1

College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences infirstexam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

Solution

The hypotheses may be written in the following form:

- H_0: The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.
- H_A: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

Example 9.1.5.2

Examine Figure 9.1.5.1. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

Figure 9.1.5.1: Side-by-side dot plot for the outcomes for six groups.

Solution

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large relative to the variability in the individual observations within each group.

Is Batting Performance Related to Player Position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: out elder (OF), in elder (IF), designated hitter (DH), and catcher (C). We will use a data set called bat10, which includes batting records of 327 Major League Baseball (MLB) players from the 2010 season. Six of the 327 cases represented in bat10 are shown in Table $9.1.5.1$, and descriptions for each variable are provided in Table 5.26. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (OBP). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

Table $9.1.5.1$: Six cases from the bat10 data matrix.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NTM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

Exercise $9.1.5.1$

The null hypothesis under consideration is the following:

$$\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_{C}. \tag{9.1.5.1}$$

Write the null and corresponding alternative hypotheses in plain language.

Solution

- H_0: The average on-base percentage is equal across the four positions.
- H_A: The average on-base

Table $9.1.5.2$: Variables and their descriptions for the bat10 data set.

variable	description
name	Player name
team	The abbreviated name of the player's team
position	The player's primary eld position (OF, IF, DH, C)
AB	Number of opportunities at bat
H	Number of hits
HR	Number of home runs
RBI	Number of runs batted in
AVG	Batting average, which is equal to H=AB
OBP	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Example 9.1.5.3

The player positions have been divided into four groups: outfield (OF), infield (IF), designated hitter (DH), and catcher (C). What would be an appropriate point estimate of the batting average by out elders, μ_{OF}?

Solution

A good estimate of the batting average by out elders would be the sample average of AVG for just those players whose position is out field:

$$\bar{x}_{OF} = 0.334. \tag{9.1.5.2}$$

Table 9.1.5.3 provides summary statistics for each group. A side-by-side box plot for the batting average is shown in Figure 9.1.5.1. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

Table 9.1.5.3: Summary statistics of on-base percentage, split by player position.

	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean ($\bar{x}_i$)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045

percentage varies across some (or all) groups.

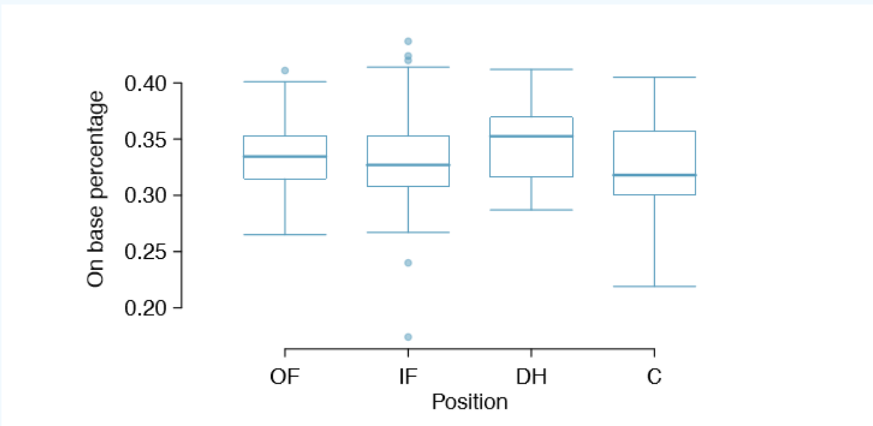


Figure 9.1.5.1: Side-by-side box plot of the on-base percentage for 327 players across four groups. There is one prominent outlier visible in the infield group, but with 154 observations in the in field group, this outlier is not a concern.

Example 9.1.5.1

The largest difference between the sample means is between the designated hitter and the catcher positions. Consider again the original hypotheses:

- H_0: $\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C$
- H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of μ_{DH} and μ_C is statistically significant at a 0.05 significance level?

Solution

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an ination in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional information on the ideas expressed in Example 5.38, we recommend reading about the *prosecutor's fallacy* (*See, for example, www.stat.columbia.edu/~cook/movabletype/archives/2007/05/the prosecutors.html*.)

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in means could have happened just by chance even if there was no difference in the respective population means.

Analysis of variance (ANOVA) and the F test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will simultaneously consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups** (MSG), and it has an associated degrees of freedom, dfG = k - 1 when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote,29 however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the mean square error (MSE), which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE are provided in the footnote[30] for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE} \tag{9.1.5.3}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

Exercise 9.1.5.1

For the baseball data, MSG = 0.00252 and MSE = 0.00127. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 1.994.

Solution

There are k = 4 groups, so $df_G = k - 1 = 3$. There are

$$n = n_1 + n_2 + n_3 + n_4 = 327 \tag{9.1.5.4}$$

total observations, so $df_E = n - k = 323$. Then the F statistic is computed as the ratio of MSG and MSE:

$$F = \frac{MSG}{MSE} = \frac{0.00252}{0.00127} = 1.984 \approx 1.994, \tag{9.1.5.5}$$

(F = 1.994 was computed by using values for MSG and MSE that were not rounded.)

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 3 and 323 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 5.29.

[29]*Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as*

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2 \tag{9.1.5.6}$$

where SSG is called the sum of squares between groups and n_i is the sample size of group i.

[30]*Let $\bar{x}$ represent the mean of outcomes across all groups. Then the sum of squares total (SST) is computed as*

$$SST = \sum_{i=1}^{n} \left(y_i - \bar{x}\right)^2 \tag{9.1.5.7}$$

where the sum is over all observations in the data set. Then we compute the sum of squared errors (SSE) in one of two equivalent ways:

$$SSE = SST - SSG \tag{9.1.5.8}$$

$$= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \tag{9.1.5.9}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

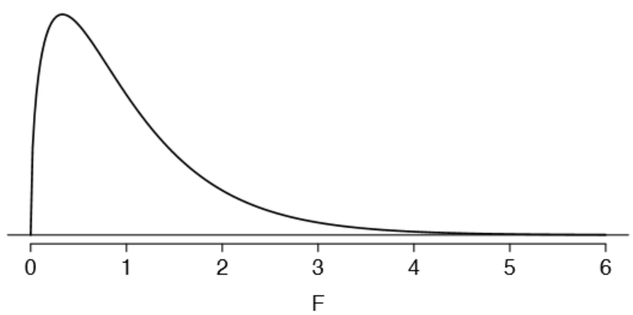


Figure 9.1.5.1*: An F distribution with $df_1 = 3$ and $df_2 = 323$.*

The larger the observed variability in the sample means (MSG) relative to the withingroup observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

The F statistic and the F test

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

Exercise 9.1.5.1

The test statistic for the baseball example is F = 1.994. Shade the area corresponding to the p-value in Figure 5.29. [32]

Example 9.1.5.1

A common method for preparing oxygen is the decomposition

Example 5.42 The p-value corresponding to the shaded area in the solution of Exercise 5.41 is equal to about 0.115. Does this provide strong evidence against the null hypothesis?

The p-value is larger than 0.05, indicating the evidence is not strong enough to reject the null hypothesis at a signi cance level of 0.05. That is, the data do not provide strong evidence that the average on-base percentage varies by player's primary field position.

Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we will see in Chapters 7 and 8. Table 5.30 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F test statistic and p-value can be retrieved from the last columns.

Table 5.30: ANOVA summary for testing whether the average on-base percentage differs across player positions.

	DF	Sum Sq	Mean Sq	F value	Pr(> F)
position	3	0.0076	0.0025	1.9943	0.1147
Residuals	323	0.4080	0.0013		

Graphical Diagnostics for an ANOVA Analysis

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

- **Independence**. If the data are a simple random sample from less than 10% of the population, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). For example, in the MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations.
- **Approximately normal**. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small. The normal probability plots for each group of the MLB data are shown in Figure 5.31; there is some deviation from normality for in elders, but this isn't a substantial concern since there are about 150 observations in that group and the outliers are not extreme. Sometimes in ANOVA there are so many groups or so few observations per group that checking normality for each group isn't reasonable. See the footnote33 for guidance on how to handle such instances.
- **Constant variance**. The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a sideby-side box plot of the outcomes across the groups, as in Figure 5.28 on page 239. In this case, the variability is similar in the four groups but not identical. We see in Table 5.27 on page 238 that the standard deviation varies a bit from one group to the next. Whether these differences are from natural variation is unclear, so we should report this uncertainty with the nal results.

[33]First calculate the residuals of the baseball data, which are calculated by taking the observed values and subtracting the corresponding group means. For example, an out elder with OBP of 0.435 would have a residual of $0.405 - \bar{x}_{OF} = 0.071$. Then to check the normality condition, create a normal probability plot using all the residuals simultaneously.

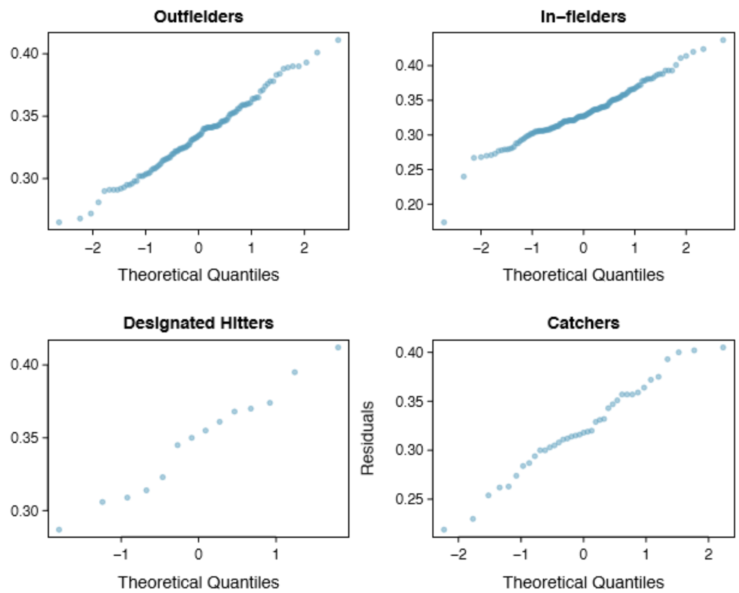


Figure 5.31: Normal probability plot of OBP for each field position.

> **Caution: Diagnostics for an ANOVA analysis**
>
> Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t test, but we use a modi ed signi cance level and a pooled estimate of the standard deviation across groups. Usually this pooled standard deviation can be found in the ANOVA table, e.g. along the bottom of Table 5.30.

Table 5.32: Summary statistics for the first midterm scores in three different lectures of the same course.

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.1	72.0	78.9
s_i	13.9	13.8	13.1

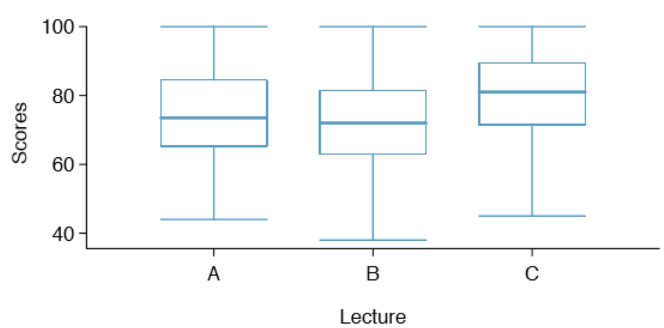


Figure 5.33: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

Example 9.1.5.1

A common method for preparing oxygen is the decomposition

Example 5.43 Example 5.34 on page 236 discussed three statistics lectures, all taught during the same semester. Table 5.32 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 5.33. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. No such situations were evident for these particular data, and we believe that independence is acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be veri ed in Table 5.32, supporting the constant variance assumption.

Exercise 9.1.5.1

A common method for preparing oxygen is the decompositio

Exercise 5.44 An ANOVA was conducted for the midterm data, and summary results are shown in Table 5.34. What should we conclude?[34]

[34]*The p-value of the test is 0.0330, less than the default signi cance level of 0.05. Therefore, we reject the null hypothesis and conclude that the difference in the average midterm scores are not due to chance.*

Table 5.34: ANOVA summary table for the midterm data.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 5.38 on page 238: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modi ed signi cance level.

Multiple comparisons and the Bonferroni correction for

The scenario of testing many pairs of groups is called multiple comparisons. The Bonferroni correction suggests that a more stringent signi cance level is more appropriate for these tests:

$$\alpha^{*}=\frac{\alpha}{K} \tag{9.1.5.10}$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K=\frac{k(k-1)}{2}$.

Example 5.45 In Exercise 5.44, you found strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modi ed signi cance level of $\alpha^{*}=\frac{0.05}{3}=0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled}=13.61$ on df = 161, which is provided in the ANOVA summary table.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{x}_A - \bar{x}_B = 75.1 - 72 = 3.1 SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56 \tag{9.1.5.11}$$

(See Section 5.4.4 on page 235 for additional details.) This results in a T score of 1.21 on df = 161 (we use the df associated with spooled). Statistical software was used to precisely identify the two-tailed p-value since the modi ed signi cance of 0.0167 is not found in the t table. The p-value (0.228) is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on df = 161 and a two-tailed p-value of 0.1462. This p-value is larger than α^*, so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on df = 161 and a two-tailed p-value of 0.0102. This p-value is smaller than α^*. Here we nd strong evidence of a difference in the means of lectures B and C.

We might summarize the ndings of the analysis from Example 5.45 using the following notation:

$$\mu_A \stackrel{?}{=} \mu_B \mu_A \stackrel{?}{=} \mu_C \mu_B \neq \mu_C \tag{9.1.5.12}$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

Caution: Sometimes an ANOVA will reject the null but no groups will have statistically signi cant differences

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, this does not invalidate the ANOVA conclusion. It only means we have not been able to successfully identify which groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high con dence a specific difference as statistically signi cant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

9.1.6: Exercises

Paired data

5.1 Global warming, Part I. Is there strong evidence of global warming? Let's consider a small scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. Then the difference between the two readings (temperature in 2008 - temperature in 1968) was calculated for each of the 51 different locations. The average of these 51 values was 1.1 degrees with a standard deviation of 4.9 degrees. We are interested in determining whether these data provide strong evidence of temperature warming in the continental US.

1. (a) Is there a relationship between the observations collected in 1968 and 2008? Or are the observations in the two groups independent? Explain.
2. (b) Write hypotheses for this research in symbols and in words.
3. (c) Check the conditions required to complete this test.
4. (d) Calculate the test statistic and nd the p-value.
5. (e) What do you conclude? Interpret your conclusion in context.
6. (f) What type of error might we have made? Explain in context what the error means.
7. (g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the temperature measurements from 1968 and 2008 to include 0? Explain your reasoning.

5.2 High School and Beyond, Part I. The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

1. (a) Is there a clear difference in the average reading and writing scores?
2. (b) Are the reading and writing scores of each student independent of each other?
3. (c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?
4. (d) Check the conditions required to complete this test.
5. (e) The average observed difference in scores is $\bar{x}_{\text{read-write}} = -0.545$, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?
6. (f) What type of error might we have made? Explain what the error means in the context of the application.
7. (g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

5.3 Global warming, Part II. We considered the differences between the temperature readings in January 1 of 1968 and 2008 at 51 locations in the continental US in Exercise 5.1. The mean and standard deviation of the reported differences are 1.1 degrees and 4.9 degrees.

1. (a) Calculate a 90% confidence interval for the average difference between the temperature measurements between 1968 and 2008.
2. (b) Interpret this interval in context.
3. (c) Does the confidence interval provide convincing evidence that the temperature was higher in 2008 than in 1968 in the continental US? Explain.

5.4 High school and beyond, Part II. We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey in Exercise 5.3. The mean and standard deviation of the differences are $\bar{x}_{\text{read-write}} = -0.545$ and 8.887 points.

1. (a) Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.
2. (b) Interpret this interval in context.
3. (c) Does the confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

5.5 Gifted children. Researchers collected a simple random sample of 36 children who had been identi ed as gifted in a large city. The following histograms show the distributions of the IQ scores of mothers and fathers of these children. Also provided are some sample statistics.[35]

1. (a) Are the IQs of mothers and the IQs of fathers in this data set related? Explain.
2. (b) Conduct a hypothesis test to evaluate if the scores are equal on average. Make sure to clearly state your hypotheses, check the relevant conditions, and state your conclusion in the context of the data.

5.6 Paired or not? In each of the following scenarios, determine if the data are paired.

1. (a) We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To nd out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock.
2. (b) We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.
3. (c) A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

[35]F.A. Graybill and H.K. Iyer. Regression Analysis: Concepts and Applications. Duxbury Press, 1994, pp. 511-516.

Difference of two means

5.7 Math scores of 13 year olds, Part I. The National Assessment of Educational Progress tested a simple random sample of 1,000 thirteen year old students in both 2004 and 2008 (two separate simple random samples). The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively. Calculate a 90% confidence interval for the change in average scores from 2004 to 2008, and interpret this interval in the context of the application. (Reminder: check conditions.)[36]

5.8 Work hours and education, Part I. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. The histograms below display the distributions of hours worked per week for two education groups: those with and without a college degree.37 Suppose we want to estimate the average difference between the number of hours worked per week by all Americans with a college degree and those without a college degree. Summary information for each group is shown in the tables.

1. (a) What is the parameter of interest, and what is the point estimate?
2. (b) Are conditions satisfied for estimating this difference using a confidence interval?
3. (c) Create a 95% confidence interval for the difference in number of hours worked between the two groups, and interpret the interval in context.
4. (d) Can you think of any real world justi cation for your results? (Note: There isn't a single correct answer to this question.)

5.9 Math scores of 13 year olds, Part II. Exercise 5.7 provides data on the average math scores from tests conducted by the National Assessment of Educational Progress in 2004 and 2008. Two separate simple random samples were taken in each of these years. The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively.

1. (a) Do these data provide strong evidence that the average math score for 13 year old students has changed from 2004 to 2008? Use a 10% signi cance level.
2. (b) It is possible that your conclusion in part (a) is incorrect. What type of error is possible for this conclusion? Explain.
3. (c) Based on your hypothesis test, would you expect a 90% confidence interval to contain the null value? Explain.

[36]National Center for Education Statistics, NAEP Data Explorer.

[37]National Opinion Research Center, General Social Survey, 2010.

5.10 Work hours and education, Part II. The General Social Survey described in Exercise 5.8 included random samples from two groups: US residents with a college degree and US residents without a college degree. For the 505 sampled US residents with a college degree, the average number of hours worked each week was 41.8 hours with a standard deviation of 15.1 hours. For those 667 without a degree, the mean was 39.4 hours with a standard deviation of 15.1 hours. Conduct a hypothesis test to check for a difference in the average number of hours worked for the two groups.

5.11 Does the Paleo diet work? The Paleo diet allows only for foods that humans typically consumed over the last 2.5 million years, excluding those agriculture-type foods that arose during the last 10,000 years or so. Researchers randomly divided 500 volunteers into two equal-sized groups. One group spent 6 months on the Paleo diet. The other group received a pamphlet about controlling portion sizes. Randomized treatment assignment was performed, and at the beginning of the study, the average difference in weights between the two groups was about 0. After the study, the Paleo group had lost on average 7 pounds with a standard deviation of 20 pounds while the control group had lost on average 5 pounds with a standard deviation of 12 pounds.

1. (a) The 95% confidence interval for the difference between the two population parameters (Paleo - control) is given as (-0.891, 4.891). Interpret this interval in the context of the data.
2. (b) Based on this confidence interval, do the data provide convincing evidence that the Paleo diet is more effective for weight loss than the pamphlet (control)? Explain your reasoning.
3. (c) Without explicitly performing the hypothesis test, do you think that if the Paleo group had lost 8 instead of 7 pounds on average, and everything else was the same, the results would then indicate a signi cant difference between the treatment and control groups? Explain your reasoning.

5.12 Weight gain during pregnancy. In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relationship between habits and practices of expectant mothers and the birth of their children. The following histograms show the distributions of weight gain during pregnancy by 867 younger moms (less than 35 years old) and 133 mature moms (35 years old and over) who have been randomly sampled from this large data set. The average weight gain of younger moms is 30.56 pounds, with a standard deviation of 14.35 pounds, and the average weight gain of mature moms is 28.79 pounds, with a standard deviation of 13.48 pounds. Calculate a 95% confidence interval for the difference between the average weight gain of younger and mature moms. Also comment on whether or not this interval provides strong evidence that there is a signi cant difference between the two population means.

5.13 Body fat in women and men. The third National Health and Nutrition Examination Survey collected body fat percentage (BF) data from 13,601 subjects whose ages are 20 to 80. A summary table for these data is given below. Note that BF is given as mean $\pm$ standard error. Construct a 95% confidence interval for the difference in average body fat percentages between men and women, and explain the meaning of this interval.[38]

Gender	n	BF (%)
Men Women	6,580 7,021	23.9 ± 0.07 35.0 ± 0.09

5.14 Child care hours, Part I. The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments. One of the variables collected on the survey is the number of hours parents spend taking care of children in their household under age 6 (feeding, bathing, dressing, holding, or watching them). In 2006, 487 females and 312 males were surveyed for this question. On average, females reported spending 31 hours with a standard deviation of 31 hours, and males reported spending 16 hours with a standard deviation of 21 hours. Calculate a 95% confidence interval for the difference between the average number of hours Chinese males and females spend taking care of their children under age 6. Also comment on whether this interval suggests a significant difference between the two population parameters. You may assume that conditions for inference are satisfied.[39]

One-sample means with the t distribution

5.15 Identify the critical t. An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t value (t?) for the given sample size and confidence level.

1. (a) n = 6, CL = 90%
2. (b) n = 21, CL = 98%
3. (c) n = 29, CL = 95%
4. (d) n = 12, CL = 99%

5.16 Working backwards, Part I. A 90% confidence interval for a population mean is (65,77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

5.17 Working backwards, Part II. A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). This confidence interval is based on a simple random sample of 36 observations. Calculate the sample mean and standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t distribution in any calculations.

5.18 Find the p-value. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given set of hypotheses and T test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

1. (a) $H_A : \mu > \mu_0, n = 11, T = 1.91$
2. (b) $H_A : \mu < \mu_0, n = 17, T = -3.45$
3. (c) $H_A : \mu \neq \mu_0, n = 7, T = 0.83$
4. (d) $H_A : \mu > \mu_0, n = 28, T = 2.13$

[38]*A Romero-Corral et al. "Accuracy of body mass index in diagnosing obesity in the adult general population". In: International Journal of Obesity 32.6 (2008), pp. 959-966.*

[39]*UNC Carolina Population Center, China Health and Nutrition Survey, 2006.*

5.19 Sleep habits of New Yorkers. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. Do these data provide strong evidence that New Yorkers sleep less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

1. (a) Write the hypotheses in symbols and in words.
2. (b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.
3. (c) Find and interpret the p-value in this context. Drawing a picture may be helpful.
4. (d) What is the conclusion of the hypothesis test?
5. (e) If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

5.20 Fuel efficiency of Prius. Fueleconomy.gov, the official US government source for fuel economy information, allows users to share gas mileage information on their vehicles. The histogram below shows the distribution of gas mileage in miles per gallon (MPG) from 14 users who drive a 2012 Toyota Prius. The sample mean is 53.3 MPG and the standard deviation is 5.2 MPG. Note that these data are user estimates and since the source data cannot be veri ed, the accuracy of these estimates are not guaranteed.[40]

1. (a) We would like to use these data to evaluate the average gas mileage of all 2012 Prius drivers. Do you think this is reasonable? Why or why not?
2. (b) The EPA claims that a 2012 Prius gets 50 MPG (city and highway mileage combined). Do these data provide strong evidence against this estimate for drivers who participate on fueleconomy.gov? Note any assumptions you must make as you proceed with the test.
3. (c) Calculate a 95% confidence interval for the average gas mileage of a 2012 Prius by drivers who participate on fueleconomy.gov.

5.21 Find the mean. You are given the following hypotheses:

- $H_0 : \mu = 60$
- $H_A : \mu < 60$

We know that the sample standard deviation is 8 and the sample size is 20. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

5.22 t* vs. z*. For a given confidence level, t* df is larger than z*. Explain how \(t^*_{df} being slightly larger than z* affects the width of the confidence interval.

[40]*Fuelecomy.gov, Shared MPG Estimates: Toyota Prius 2012.*

The t distribution for the difference of two means

5.23 Cleveland vs. Sacramento. Average income varies from one region of the country to another, and it often reects both lifestyles and regional living expenses. Suppose a new graduate is considering a job in two locations, Cleveland, OH and Sacramento, CA, and he wants to see whether the average income in one of these cities is higher than the other. He would like to conduct a t test based on two small samples from the 2000 Census, but he first must consider whether the conditions are met to implement the test. Below are histograms for each city. Should he move forward with the t test? Explain your reasoning.

5.24 Oscar winners. The rst Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2012. Summary statistics for these distributions are also provided. Is a t test appropriate for evaluating whether the difference in the average ages of best actors and actresses might be due to chance? Explain your reasoning.[41]

[41]*Oscar winners from 1929 - 2012, data up to 2009 from the Journal of Statistics Education data archive and more current data from Wikipedia.org.*

5.25 Friday the 13th, Part I. In the early 1990's, researchers in the UK collected data on traffic ow, number of shoppers, and traffic accident related emergency room admissions on Friday the 13th and the previous Friday, Friday the 6th. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6th and Friday the 13th for many such date pairs. Also given are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th.[42]

	6 th	13 th	Diff.
$\bar{x}$	128,385	126,550	1,835
s	7,259	7,664	1,176
n	10	10	10

1. (a) Are there any underlying structures in these data that should be considered in an analysis? Explain.
2. (b) What are the hypotheses for evaluating whether the number of people out on Friday the 6th is different than the number out on Friday the 13th?
3. (c) Check conditions to carry out the hypothesis test from part (b).
4. (d) Calculate the test statistic and the p-value.
5. (e) What is the conclusion of the hypothesis test?
6. (f) Interpret the p-value in this context.
7. (g) What type of error might have been made in the conclusion of your test? Explain.

5.26 Diamonds, Part I. Prices of diamonds are determined by what is known as the 4 Cs: cut, clarity, color, and carat weight. The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and compare the average prices of the diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. For a 1 carat diamond, we divide the price by 100. The distributions and some sample statistics are shown below.[43]

	0.99 carats	1 carat
Men	\$ 44.51	\$ 56.81
SD	\$ 13.32	\$ 16.13
n	23	23

Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.

[42]*T.J. Scanlon et al. "Is Friday the 13th Bad For Your Health?" In: BMJ 307 (1993), pp. 1584-1586.*

[43]*H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.*

5.27 Friday the 13th, Part II. The Friday the 13th study reported in Exercise 5.25 also provides data on traffic accident related emergency room admissions. The distributions of these counts from Friday the 6th and Friday the 13th are shown below for six such paired dates along with summary statistics. You may assume that conditions for inference are met.

1. (a) Conduct a hypothesis test to evaluate if there is a difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

2. (b) Calculate a 95% confidence interval for the difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.
3. (c) The conclusion of the original study states, "Friday 13th is unlucky for some. The risk of hospital admission as a result of a transport accident may be increased by as much as 52%. Staying at home is recommended." Do you agree with this statement? Explain your reasoning.

5.28 Diamonds, Part II. In Exercise 5.26, we discussed diamond prices (standardized by weight) for diamonds with weights 0.99 carats and 1 carat. See the table for summary statistics, and then construct a 95% confidence interval for the average difference between the standardized prices of 0.99 and 1 carat diamonds. You may assume the conditions for inference are met.

	0.99 carats	1 carat
Men SD n	\$ 44.51 \$ 13.32 23	\$ 56.81 \$ 16.13 23

5.29 Chicken diet and weight, Part I. Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company pro ts, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Below are some summary statistics from this data set along with box plots showing the distribution of weights by feed type.[44]

1. (a) Describe the distributions of weights of chickens that were fed linseed and horsebean.
2. (b) Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different? Use a 5% significance level.
3. (c) What type of error might we have committed? Explain.
4. (d) Would your conclusion change if we used $\alpha = 0.01$?

5.30 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[45]

5.31 Chicken diet and weight, Part II. Casein is a common weight gain supplement for humans. Does it have an effect on chickens? Using data provided in Exercise 5.29, test the hypothesis that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean. If your hypothesis test yields a statistically significant result, discuss whether or not the higher average weight of chickens can be attributed to the casein diet. Assume that conditions for inference are satisfied.

[44]*Chicken Weights by Feed Type, from the datasets package in R.*

[45]*U.S. Department of Energy, Fuel Economy Data, 2012 Data le.*

5.32 Fuel efficiency of manual and automatic cars, Part II. The table provides summary statistics on highway fuel economy of cars manufactured in 2012 (from Exercise 5.30). Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.[46]

5.33 Gaming and distracted eating, Part I. A group of researchers are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption. To test this hypothesis, they monitored food intake for a group of 44 patients who were randomized into two equal groups. The treatment group ate lunch while playing solitaire, and the control group ate lunch without any added distractions. Patients in the treatment group ate 52.1 grams of biscuits, with a standard deviation of 45.1 grams, and patients in the control group ate 27.1 grams of biscuits, with a standard deviation of 26.4 grams. Do these data provide convincing evidence that the average food intake (measured in amount of biscuits consumed) is different for the patients in the treatment group? Assume that conditions for inference are satisfied.[47]

5.34 Gaming and distracted eating, Part II. The researchers from Exercise 5.33 also investigated the effects of being distracted by a game on how much people eat. The 22 patients in the treatment group who ate their lunch while playing solitaire were asked to do a serial-order recall of the food lunch items they ate. The average number of items recalled by the patients in this group was

4.9, with a standard deviation of 1.8. The average number of items recalled by the patients in the control group (no distraction) was 6.1, with a standard deviation of 1.8. Do these data provide strong evidence that the average number of food items recalled by the patients in the treatment and control groups are different?

5.35 Prison isolation experiment, Part I. Subjects from Central Prison in Raleigh, NC, volunteered for an experiment involving an \isolation" experience. The goal of the experiment was to nd a treatment that reduces subjects' psychopathic deviant T scores. This score measures a person's need for control or their rebellion against control, and it is part of a commonly used mental health test called the Minnesota Multiphasic Personality Inventory (MMPI) test. The experiment had three treatment groups:

1. (1) Four hours of sensory restriction plus a 15 minute "therapeutic" tape advising that professional help is available.
2. (2) Four hours of sensory restriction plus a 15 minute "emotionally neutral" tape on training hunting dogs.
3. (3) Four hours of sensory restriction but no taped message.

Forty-two subjects were randomly assigned to these treatment groups, and an MMPI test was administered before and after the treatment. Distributions of the differences between pre and

[46]*U.S. Department of Energy, Fuel Economy Data, 2012 Data file.*

[47]*R.E. Oldham-Cooper et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake". In: The American Journal of Clinical Nutrition 93.2 (2011), p. 308.*

post treatment scores (pre - post) are shown below, along with some sample statistics. Use this information to independently test the effectiveness of each treatment. Make sure to clearly state your hypotheses, check conditions, and interpret results in the context of the data.[48]

5.36 True or false, Part I. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

1. (a) When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal model for the difference in means since $n_2 \geq 30$.
2. (b) As the degrees of freedom increases, the T distribution approaches normality.
3. (c) We use a pooled standard error for calculating the standard error of the difference between means when sample sizes of groups are equal to each other.

Comparing many means with ANOVA

5.37 Chicken diet and weight, Part III. In Exercises 5.29 and 5.31 we compared the effects of two types of feed at a time. A better analysis would rst consider all feed types at once: casein, horsebean, linseed, meat meal, soybean, and sunower. The ANOVA output below can be used to test for differences between the average weights of chicks on different diets.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
feed Residuals	5 65	231,129.16 195,556.02	46,225.83 3,008.55	15.36	0.0000

Conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions. Figures and summary statistics are shown below.

5.38 Student performance across discussion sections. A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average nal exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
$\bar{x}_i$	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Section Residuals	7 189	525.01 7584.11	75.00 40.13	1.87	0.0767

Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to proceed with the test.

5.39 Coffee, depression, and physical activity. Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.[49]

	Caffeinated	coffee	consumption			
	$\leq 1 cup/week$	2-6 cus/week	1 cup/day	2-3 cups/day	$\geq 4 cups/day$	Total
Mean SD n	18.7 21.1 12,215	19.6 25.5 6,617	19.3 22.5 17,234	18.9 22.0 12,290	17.5 22.0 2,838	50,739

1. (a) Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.
2. (b) Check conditions and describe any assumptions you must make to proceed with the test.
3. (c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Section Residuals Total	--------------- --------------- ---------------	--------------- 25,564,819 25,575,327	---------------- ----------------	----------------	0.0003

(d) What is the conclusion of the test?

[49] *M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: Archives of internal medicine 171.17 (2011), p. 1571.*

5.40 Work hours and education, Part III. In Exercises 5.8 and 5.10 you worked with data from the General Social Survey in order to compare the average number of hours worked per week by US residents with and without a college degree. However, this analysis didn't take advantage of the original data which contained more accurate information on educational attainment (less than high school, high school, junior college, Bachelor's, and graduate school). Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once instead of re-categorizing them into two groups. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

		Educational	attainment			
	Less than HS	HS	Jr Coll	BAchelor's	Graduate	Total
Mean SD n	38.67 15.81 121	39.6 14.97 546	41.39 18.1 97	42.55 13.62 253	40.85 15.51 155	40.45 15.17 1,172

1. (a) Write hypotheses for evaluating whether the average number of hours worked varies across the ve groups.
2. (b) Check conditions and describe any assumptions you must make to proceed with the test.
3. (c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
degree Residuals Total	--------------- --------------- ---------------	--------------- 267,382 ---------------	501.54 ----------------	----------------	0.0682

(d) What is the conclusion of the test?

5.41 GPA and major. Undergraduate students taking an introductory statistics course at Duke University conducted a survey about GPA and major. The side-by-side box plots show the distribution of GPA among three groups of majors. Also provided is the ANOVA output.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
major Residuals	2 195	0.03 15.77	0.02 0.08	0.21	0.8068

1. (a) Write the hypotheses for testing for a difference between average GPA across majors.
2. (b) What is the conclusion of the hypothesis test?
3. (c) How many students answered these questions on the survey, i.e. what is the sample size?

5.42 Child care hours, Part II. Exercise 5.14 introduces the China Health and Nutrition Survey which, among other things, collects information on number of hours Chinese parents spend taking care of their children under age 6. The side by side box plots below show the distribution of this variable by educational attainment of the parent. Also provided below is the ANOVA output for comparing average hours across educational attainment categories.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
education Residuals	4 794	4142.09 653047.83	1035.52 822.48	1.26	0.2846

1. (a) Write the hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels.
2. (b) What is the conclusion of the hypothesis test?

5.43 True or false, Part II. Determine if the following statements are true or false in ANOVA, and explain your reasoning for statements you identify as false.

1. (a) As the number of groups increases, the modi ed signi cance level for pairwise tests increases as well.
2. (b) As the total sample size increases, the degrees of freedom for the residuals increases as well.
3. (c) The constant variance condition can be somewhat relaxed when the sample sizes are relatively consistent across groups.
4. (d) The independence assumption can be relaxed when the total sample size is large.

5.44 True or false, Part III. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

If the null hypothesis that the means of four groups are all the same is rejected using ANOVA at a 5% signi cance level, then ...

1. (a) we can then conclude that all the means are different from one another.
2. (b) the standardized variability between groups is higher than the standardized variability within groups.
3. (c) the pairwise analysis will identify at least one pair of means that are signi cantly different.
4. (d) the appropriate to be used in pairwise comparisons is $\frac{0.05}{4} = 0.0125$ since there are four groups.

5.45 Prison isolation experiment, Part II. Exercise 5.35 introduced an experiment that was conducted with the goal of identifying a treatment that reduces subjects' psychopathic deviant T scores, where this score measures a person's need for control or his rebellion against control. In Exercise 5.35 you evaluated the success of each treatment individually. An alternative analysis involves comparing the success of treatments. The relevant ANOVA output is given below.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)

treatment Residuals	2 39	639.48 3740.43	319.74 95.91	3.33	0.0461

1. (a) What are the hypotheses?
2. (b) What is the conclusion of the test? Use a 5% significance level.
3. (c) If in part (b) you determined that the test is signi cant, conduct pairwise tests to determine which groups are different from each other. If you did not reject the null hypothesis in part (b), recheck your solution.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

SECTION OVERVIEW

9.2: Inference for Categorical Data

Chapter 6 introduces inference in the setting of categorical data. We use these methods to answer questions like the following:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

We will find that the methods we learned in previous chapters are very useful in these settings. For example, sample proportions are well characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual confidence interval and hypothesis testing tools. In other instances, such as those with contingency tables or when sample size conditions are not met, we will use a different distribution, though the core ideas remain the same.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

9.2.1: Inference for a Single Proportion

According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[1] This poll included responses of 976 adults.

Identifying when the Sample Proportion is Nearly Normal

A sample proportion can be described as a sample mean. If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0+1+1+\cdots+0}{976} = 0.44 \tag{9.2.1.1}$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes and failures, typically at least 10 of each.

[1]*nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html*

Conditions for the sampling distribution of $\hat{p}$ being nearly normal

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{9.2.1.2}$$

Typically we do not know the true proportion, p, so must substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value - that is, the proportion claimed in the null hypothesis - is used in place of p. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

TIP: Reminder on checking independence of observations

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Alternatively, if the data come from a random process, we must evaluate the independence condition more carefully.

Confidence Intervals for a Proportion

We may want a confidence interval for the proportion of Americans who approve of the job the Supreme Court is doing. Our point estimate, based on a sample of size n = 976 from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use the general confidence interval formula from Section 4.5, we must check the conditions to ensure that the sampling distribution of $\hat{p}$ is nearly normal. We also must determine the standard error of the estimate.

The data are based on a simple random sample and consist of far fewer than 10% of the U.S. population, so independence is confirmed. The sample size must also be sufficiently large, which is checked via the success-failure condition: there were approximately $976 \times \hat{p} = 429$ "successes" and $976 \times (1-\hat{p}) = 547$ "failures" in the sample, both easily greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

Exercise 9.2.1.1

Estimate the standard error of $\hat{p} = 0.44$ using Equation 9.2.1.2. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p.

Answer

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$$

Example 9.2.1.1

Construct a 95% confidence interval for p, the proportion of Americans who trust federal officials most of the time.

Solution

Using the standard error estimate from Exercise 9.2.1.1, the point estimate 0.44, and z* = 1:96 for a 95% confidence interval, the confidence interval may be computed as

$$\text{point estimate} \pm z^{*} SE \rightarrow 0.44 \pm 1.96 \times 0.016 \rightarrow (0.409, 0.471) \tag{9.2.1.3}$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. If the proportion has not changed since this poll, than we can say with high confidence that the job approval of the Supreme Court is below 50%.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, the sampling distribution of $\hat{p}$ may be well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

Hypothesis Testing for a Proportion

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1-p_0)$ are at least 10, where p_0 is the null value.

Exercise 9.2.1.2

Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Set up a onesided hypothesis test to evaluate this claim.

Answer

Is there convincing evidence that the campaign manager is correct?

- H_0 : p = 0.50,
- H_A : p > 0.50.

Example 9.2.1.2

A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%. Does this provide convincing evidence for the claim of Toohey's manager at the 5% significance level?

Solution

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion,

$$p_0 = 0.5 : np_0 = n(1-p_0) = 500 \times 0.5 = 250 > 10. \tag{9.2.1.4}$$

With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion.

$$\begin{aligned} SE &= \sqrt{\frac{p_0 \times (1-p_0)}{n}} \\ &= \sqrt{\frac{0.5(1-0.5)}{500}} = 0.022 \end{aligned}$$

A picture of the normal model is shown in Figure $9.2.1.1$ with the p-value represented by the shaded region. Based on the normal model, the test statistic can be computed as the Z score of the point estimate:

$$\begin{aligned} Z &= \frac{\text{point estimate - null value}}{SE} \\ &= \frac{0.52-0.50}{0.022} = 0.89 \end{aligned}$$

The upper tail area, representing the p-value, is 0.1867. Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

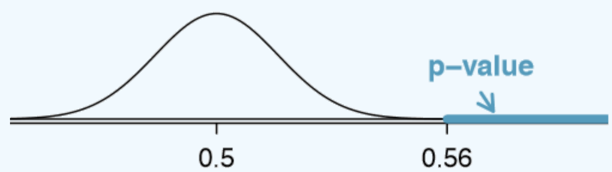


Figure $9.2.1.1$: Sampling distribution of the sample proportion if the null hypothesis is true for Example $9.2.1.2$. The p-value for the test is shaded.

Hypothesis test for a proportion

Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

Choosing a sample size when estimating a proportion

We first encountered sample size computations in Section 4.6, which considered the case of estimating a single mean. We found that these computations were helpful in planning a study to control the size of the standard error of a point estimate. The task was to find a sample size n so that the sample mean would be within some margin of error m of the actual mean with a certain level of confidence. For example, the margin of error for a point estimate using 95% confidence can be written as $1.96 \times SE$. We set up a general equation to represent the problem:

$$ME = z^{*}SE \leq m \tag{9.2.1.5}$$

where ME represented the actual margin of error and z^{*} was chosen to correspond to the confidence level. The standard error formula is specified to correspond to the particular setting. For instance, in the case of means, the standard error was given as $\frac{\sigma}{\sqrt{n}}$.

In the case of a single proportion, we use $\sqrt{p(1-p)}\,n$ for the standard error.

Planning a sample size before collecting data is equally important when estimating a proportion. For instance, if we are conducting a university survey to determine whether students support a $200 per year increase in fees to pay for a new football stadium, how big of a sample is needed to be sure the margin of error is less than 0.04 using a 95% confidence level?

Example $9.2.1.3$

Find the smallest sample size n so that the margin of error of the point estimate $\hat{p}$ will be no larger than $m = 0.04$ when using a 95% confidence interval.

Solution

For a 95% confidence level, the value z* corresponds to 1.96, and we can write the margin of error expression as follows:

$$ME = z^{*} SE = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04 \tag{9.2.1.6}$$

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could use that value. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this worst case estimate if no other estimate is available:

$$1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.04 \tag{9.2.1.7}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{n} \leq 0.04^2 \tag{9.2.1.8}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} \leq n \tag{9.2.1.9}$$

$$600.25 \leq n \tag{9.2.1.10}$$

We would need at least 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

No estimate of the true proportion is required in sample size computations for a proportion, whereas an estimate of the standard deviation is always needed when computing a sample size for a margin of error for the sample mean. However, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

Example 9.2.1.4

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.

a. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.
b. The sample sizes in (b) vary widely. Which of the three would you suggest using? What would influence your choice?

Solution

(a) For the 1.7% estimate of p, we estimate the appropriate sample size as follows:

$$1.65 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \leq 0.02 \rightarrow n \geq 113.7$$

Using the estimate from the first model, we would suggest examining 114 tires (round up!). A similar computation can be accomplished using 0.062 and 0.013 for p: 396 and 88.

(b) We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. (Answers will vary.)

Exercise 9.2.1.4

A recent estimate of Congress' approval rating was 17%.5 What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?

Answer

We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \leq 0.04 \rightarrow n \geq 338.8$$

A sample size of 339 or more would be reasonable.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

9.2.2: Difference of Two Proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, a company weighs whether they should switch to a higher quality parts manufacturer. In the last example, we examine the cancer risk to dogs from the use of yard herbicides.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

Sample Distribution of the Difference of Two Proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when each proportion separately follows a **normal model**, and the samples are **independent**. The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} \tag{9.2.2.1}$$

$$= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{9.2.2.2}$$

where p_1 and p_2 represent the population proportions, and n1 and n2 represent the sample sizes.

For the difference in two means, the standard error formula took the following form:

$$SE_{\hat{x}_1 - \hat{x}_2} = \sqrt{SE^2_{\hat{x}_1} + SE^2_{\hat{x}_2}} \tag{9.2.2.3}$$

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Exercise 5.14.

[5]*www.gallup.com/poll/155144/Congress-Approval-June.aspx*

Table 9.2.2.1: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

	Sample size (ni)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Intervals and tests for $p_1 - p_2$

In the setting of confidence intervals, the sample proportions are used to verify the success/failure condition and also compute standard error, just as was the case with a single proportion.

Example 9.2.2.1

The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[7]

> *As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?*

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed. Table 6.2 shows the results of this experiment. Create and interpret a 90% confidence interval of the difference in approval.

Solution

First the conditions must be verified. Because each group is a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample. Because all conditions are met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13 \tag{9.2.2.4}$$

The standard error may be computed from Equation 9.2.2.2using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025 \tag{9.2.2.5}$$

For a 90% con dence interval, we use z* = 1.65:

$$\text{point estimate} \pm z^{*} SE \approx 0.13 \pm 1.65 \times 0.025 \to (0.09, 0.17) \tag{9.2.2.6}$$

We are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still uid on the health insurance mandate.

[7] *www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/.*

Sample sizes for each polling group are approximate.

Exercise 9.2.2.1

A remote control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[8]

Answer

Add texts here. Do not delete this text first.

Example 9.2.2.2

The quality control engineer from Exercise 6.11 collects a sample of gears, examining 1000 gears from each company and nds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Exercise 6.11 using a signi cance level of 5%.

Solution

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the gears are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as

to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, 0.958 - 0.899 = 0.059, can be said to come from a nearly normal distribution.

The standard error can be found using Equation 9.2.2.2:

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114 \tag{9.2.2.7}$$

In this hypothesis test, the sample proportions were used. We will discuss this choice more in Section 6.2.3.

Next, we compute the test statistic and use it to nd the p-value, which is depicted in Figure 9.2.2.1.

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54 \tag{9.2.2.8}$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher quality gears actually do pass inspection more than 3% as often as the currently used gears. Based on these results, management will approve the switch to the new supplier.

[8]H_0: The higher quality gears will pass inspection no more than 3% more frequently than the standard quality gears. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher quality gears will pass inspection more than 3% more often than the standard quality gears. $p_{highQ} - p_{standard} > 0.03$.

***Figure* 9.2.2.1*:** *Distribution of the test statistic if the null hypothesis was true.*

The p-value is represented by the shaded area.

Hypothesis testing when H_0: $p_1 = p_2$

Here we use a new example to examine a special estimate of standard error when $H_0: p_1 = p_2$. We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.9 Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table 9.2.2.2

Table 9.2.2.2: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

	cancer	no cancer
2,4 - D	191	304
no 2,4 - D	300	641

Exercise 9.2.2.1

Is this study an experiment or an observational study?

Answer

The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the control group, which is a term usually seen in experiments.

Exercise 9.2.2.1

Exercise 6.14 Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups.[11]

[9]*Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. CaseControl Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.*

Answer

Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.

- H_0: the proportion of dogs with exposure to 2,4-D is the same in "cancer" and \no cancer" dogs, $p_c - p_n = 0$.
- H_A: dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

Example 9.2.2.1: pooled estimate

First are the conditions met to use the normal model and make inference on the results?

(1) It is unclear whether this is a random sample. However, if we believe the dogs in both the cancer and no cancer groups are representative of each respective population and that the dogs in the study do not interact in any way, then we may find it reasonable to assume independence between observations. (2) The success-failure condition holds for each sample.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

In your hypotheses for Exercise 9.2.2.1, the null is that the proportion of dogs with exposure to 2,4-D is the same in each group. The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. To identify the p-value for this test, we first check conditions (Example 6.15) and compute the standard error of the difference:

$$SE = \sqrt{\frac{p_c(1-p_c)}{n_c} + \frac{p_n(1-p_n)}{n_n}} \tag{9.2.2.9}$$

In a hypothesis test, the distribution of the test statistic is always examined as though the null hypothesis is true, i.e. in this case, $p_c = p_n$. The standard error formula should reflect this equality in the null hypothesis. We will use p to represent the common rate of dogs that are exposed to 2,4-D in the two groups:

$$SE = \sqrt{\frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_n}} \tag{9.2.2.10}$$

We don't know the exposure rate, p, but we can obtain a good estimate of it by pooling the results of both samples:

$$\hat{p} = \frac{\#\text{ of ”successes”}}{\#\text{ of cases}} = \frac{191+304}{191+300+304+641} = 0.345 \tag{9.2.2.11}$$

This is called the **pooled estimate** of the sample proportion, and we use it to compute the standard error when the null hypothesis is that $p_1 = p_2$ (e.g. $p_c = p_n$ or $p_c - p_n = 0$) . We also typically use it to verify the success-failure condition.

Pooled estimate of a proportion

When the null hypothesis is $p_1 = p_2$, it is useful to nd the pooled estimate of the shared proportion:

$$\hat{p} = \frac{\text{number of ”successes”}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} \tag{9.2.2.12}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1} \tag{9.2.2.13}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

: $p_1 = p_2$

When the null hypothesis suggests the proportions are equal, we use the pooled proportion estimate $(\hat{p})$ to verify the success-failure condition and also to estimate the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_c} + \frac{\hat{p}(1-\hat{p})}{n_n}} \tag{9.2.2.14}$$

Exercise 9.2.2.1

Using Equation 9.2.2.14, $\hat{p} = 0.345, n_1 = 491$, and $n_2 = 945$, verify the estimate for the standard error is SE = 0.026. Next, complete the hypothesis test using a significance level of 0.05. Be certain to draw a picture, compute the p-value, and state your conclusion in both statistical language and plain language.

Answer

Compute the test statistic:

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

9.2.3: Testing for Goodness of Fit using Chi-Square (Special Topic)

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test. In the rst case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Table 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reect the population of eligible jurors, i.e. registered voters.

Table 6.5: Representation by race in a city's juries and population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

[12]*Compute the test statistic:*

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58 \tag{9.2.3.1}$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

A second application, assessing the t of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for the years 1990-2011 are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

Creating a test statistic for one-way tables

Example 9.2.3.1:

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

Solution

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

Exercise 9.2.3.1

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race?

Answer

Answers can be found in Table 6.6.

Table 6.6: Actual and expected make-up of the jurors.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected count	198	19.25	33	24.75	275

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

- H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.
- H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate - null value}}{\text{SE of point estimate}} \tag{9.2.3.2}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to rst compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count - null white count}}{\text{SE of observed white count}} \tag{9.2.3.3}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205-198}{\sqrt{198}} = 0.50 \tag{9.2.3.4}$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

$$\begin{matrix} Black & Hispanic & Other \\ Z_2 = \frac{26-19.25}{\sqrt{19.25}} = 154 & Z_3 = \frac{25-33}{\sqrt{33}} = -1.39 & Z_4 = \frac{19-24:75}{\sqrt{24.75}} = -1.16 \end{matrix} \tag{9.2.3.5}$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they - as a group - tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58 \tag{9.2.3.6}$$

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89 \tag{9.2.3.7}$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual - e.g. a standardized difference of 2.5 - will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{\left(\text{observed count\_1 - null count\_1}\right)^2}{\text{null count\_1}} + \dots + \frac{\left(\text{observed count\_4 - null count\_4}\right)^2}{\text{null count\_4}} \tag{9.2.3.8}$$

[13]Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown - though not here - that the square root of the count is a better way to standardize the count differences.

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a chi-square distribution. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

The chi-square distribution and finding areas

The chi-square distribution is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters - mean and standard deviation - that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df),** which inuences the shape, center, and spread of the distribution.

Exercise 9.2.3.1

Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

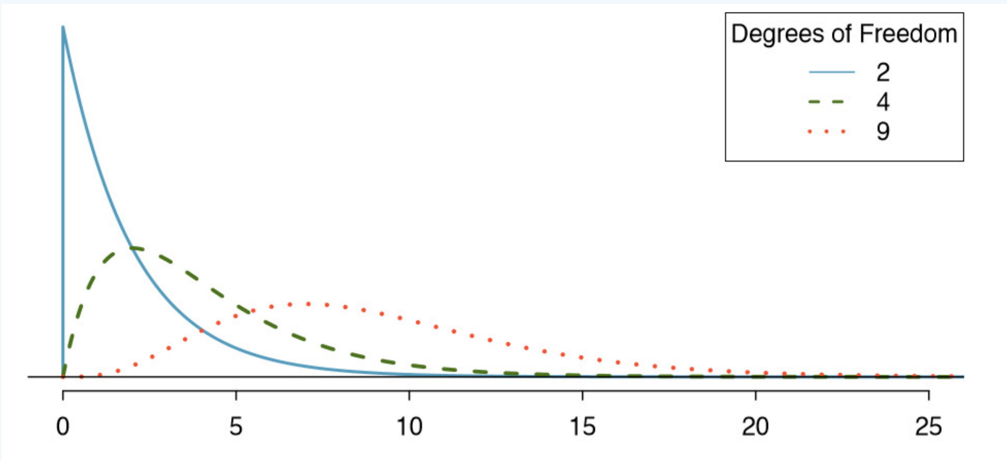


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

Answer

[14](a) The center becomes larger. If we look carefully, we can see that the center of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for df = 2, and then the distributions become more symmetric for the larger degrees of freedom df = 4 and df = 9. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Figure 6.7 and Exercise 6.20 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the **chi-square table**, partially shown in Table 6.8. A more complete table is presented in Appendix B.3 on page 412. This table is very similar to the t table from Sections 5.3 and 5.4: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t table is that the chi-square table only provides upper tail values.

Table 6.8: A section of the chi-square table. A complete table is in Appendix B.3 on page 412.

Upper tail	0.3 0.2 0.1 0.05	0.02 0.01 0.005 0.001
df 1 2 3 4 5	2.41 3.22 4.61 5.99 3.66 4.64 6.25 7.81 4.88 5.99 7.78 9.49 6.06 7.29 9.24 11.07	7.82 9.21 10.60 13.82 9.84 11.34 12.84 16.27 11.67 13.28 14.86 18.47 13.39 15.09 16.75 20.52
6 7	7.23 8.56 10.64 12.59 8.38 9.80 12.02 14.07	15.03 16.81 18.55 22.46 16.62 18.48 20.28 24.32

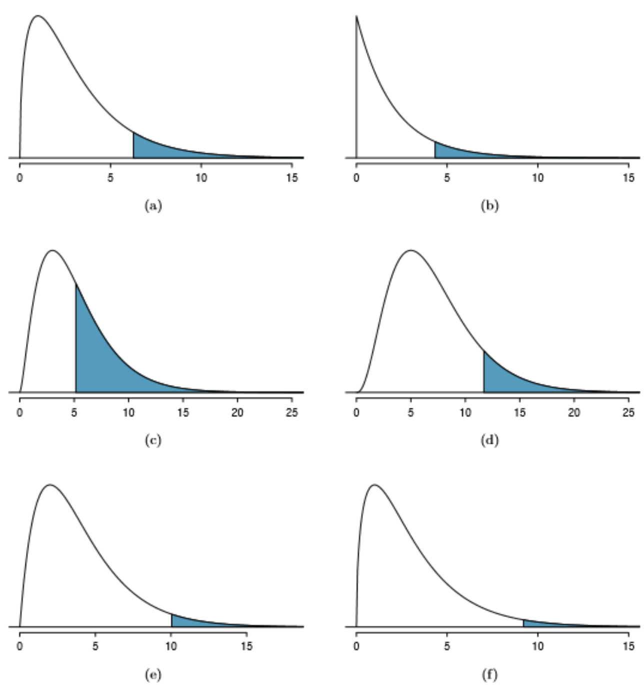


Figure 6.9: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded. (e) 4 degrees of freedom, area above 10 shaded. (f) 3 degrees of freedom, area above 9.21 shaded.

Example 6.21 Figure 6.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 6.8 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value { 6.25 { falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 6.9(a) has area 0.1.

Example 6.22 We rarely observe the exact value in the table. For instance, Figure 6.9(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3, which does not fall in Table 6.8. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 6.9(b) is between 0.1 and 0.2.

Example 6.23 Figure 6.9(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is greater than 0.3.

Exercise 6.24 Figure 6.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

Exercise 6.25 Figure 6.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Exercise 6.26 Figure 6.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

[15]*The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.*

[16]*The area is between 0.02 and 0.05.*

[17]*Between 0.02 and 0.05.*

Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with k - 1 degrees of freedom, where k is the number of bins.

Example $9.2.3.1$:

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

Solution

In the jurors example, there were k = 4 categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with k - 1 = 3 degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

Example $9.2.3.1$:

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

The chi-square distribution and p-value are shown in Figure 6.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix B.3 or the short table on page 277, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

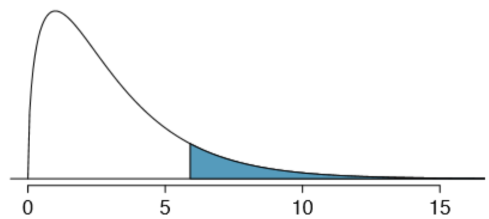


Figure 6.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with df = 3.

Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts $O_1, O_2, \ldots, O_k$ in k categories are unusually different from what might be expected under a null hypothesis. Call the expected counts that are based on the null hypothesis $E_1, E_2, \ldots, E_k$. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with k - 1 degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \tag{9.2.3.9}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

Conditions for the chi-square test

There are three conditions that must be checked before performing a chi-square test:

- **Independence**. Each case that contributes a count to the table must be independent of all the other cases in the table.
- **Sample size / distribution**. Each particular scenario (i.e. cell count) must have at least 5 expected cases.
- **Degrees of freedom** We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one proportion methods introduced in Section 6.1.

Evaluating goodness of fit for a distribution

Section 3.3 would be useful background reading for this example, but it is not a prerequisite. We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model ts a data set. Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as Up or Down (D) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each Up day:

$$\begin{array}{rcccccccccc} \text{Change in price} & 2.52 & -1.46 & 0.51 & -4.07 & 3.36 & 1.10 & -5.46 & -1.03 & -2.99 & 1.71 \\ \text{Outcome} & Up & D & Up & D & Up & Up & D & D & D & Up \\ \text{Days to Up} & 1 & - & 2 & - & 2 & 1 & - & - & - & 4 \end{array} \tag{9.2.3.10}$$

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the kth trial to observe the rst success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next Up trading day, and two more for the third Up day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Table 6.11 shows the number of waiting days for a positive trading day during 1990-2011for the S&P500.

Table 6.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

We consider how many days one must wait until observing an Up day on the S&P500 stock exchange. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a geometric distribution. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an Up day is observed. Under this hypothesis, the number of days until an Up day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an Up day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 from 1990 to 2011 and summarize the waiting times in Table 6.12 and Figure 6.13. The S&P500 was positive on 53.2% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have binned together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the Observed row in Table 6.12, can be compared to the expected counts from the Geometric Model row. The method for computing expected counts is discussed in Table 6.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each

Table 6.12: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model $(P(D) = (1 - 0.532)^{D-1}(0.532))$ and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 = 11.65\%$ of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

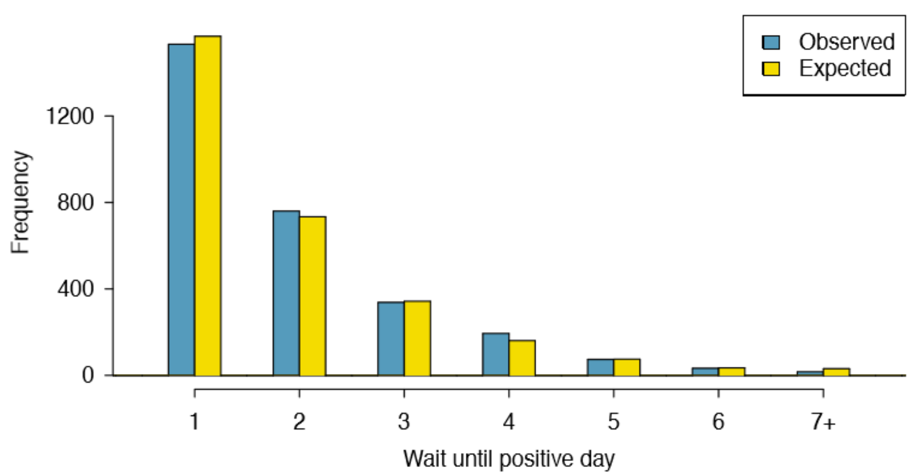


Figure 6.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Example 6.29 Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Table 6.12.

Exercise 6.30 Table 6.12 provides a set of count data for waiting times ($O_1 = 1532, O_2 = 760, \ldots$) and expected counts under the geometric distribution ($E_1 = 1569, E_2 = 734, \ldots$). Compute the chi-square test statistic, X^2.[18]

Exercise 6.31 Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

Example 6.32 If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with df = 6. Using this information, compute a p-value.

Figure 6.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix B.3, we nd that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that

[18] $X^2 = \frac{(1532-1569)^2}{1569} + \frac{(760-734)^2}{734} + \cdots + \frac{(17-31)^2}{31} = 15.08$

[19] *There are k = 7 groups, so we use df = k - 1 = 6.*

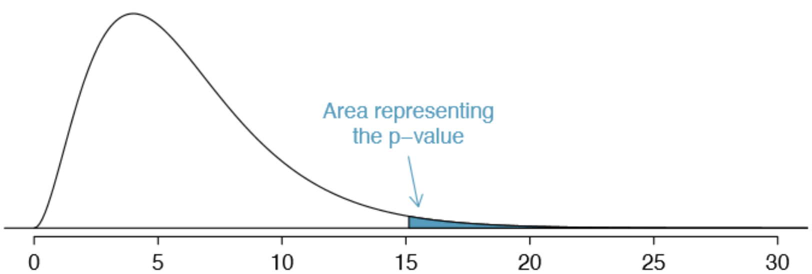


Figure 6.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

Example 6.33 In Example 6.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

9.2.4: Testing for Independence in Two-Way Tables (Special Topic)

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Table 6.15 shows an example of 10,000 queries split into three algorithm groups.[20] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Table 6.15: Google experiment breakdown of test subjects into three search groups.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

[20]Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

Example 9.2.4.1

What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

- H_0: The algorithms each perform equally well.
- H_A: The algorithms do not perform equally well.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. Under scenario (1), we might think that the user was satis ed with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Table 6.16 provides the results from the experiment. These data are very similar to the count data in Section 6.3. However, now the different combinations of two variables are binned in a two-way table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

Table 6.16: Results of the Google search algorithm experiment.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

What is so different about one-way tables and two-way tables?

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for combinations of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the search variable is independent of the algorithm variable.

Expected Counts in Two-way Tables

Example 6.35 From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as $\frac{7078}{10000} = 0.7078$. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satis ed with the initial search:

$$0.7078 \times 5000 = 3539 users \tag{9.2.4.1}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

Exercise 9.2.4.1

Exercise 6.36 Using the same rationale described in Example 6.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

[21] *We would expect 0.7078 * 2500 = 1769.5. It is okay that this is a fraction.*

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 6.35 and Exercise 6.36. These expected counts were used to construct Table 6.17, which is the same as Table 6.16, except now the expected counts have been added in parentheses.

Table 6.17: The observed counts and the (Expected counts)

Search algorithm	current	test 1	test 2	Total
No new search	3511 (3539)	1749 (1769.5)	1818 (1769.5)	7078
New search	1489 (1461)	751 (730.5)	682 (730.5)	2922
Total	5000	2500	2500	10000

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the rst row:

$$0.7078 \times (\text{column 1 total}) = 3539 \tag{9.2.4.2}$$

$$0.7078 \times (\text{column 2 total}) = 1769.5 \tag{9.2.4.3}$$

$$0.7078 \times (\text{column 3 total}) = 1769.5 \tag{9.2.4.4}$$

Looking back to how the fraction 0.7078 was computed - as the fraction of users who did not perform a new search ($\frac{7078}{10000}$) - these three expected counts could have been computed as

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 1 total}) = 3539 \tag{9.2.4.5}$$

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 2 total}) = 1769.5 \tag{9.2.4.6}$$

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 3 total}) = 1769.5 \tag{9.2.4.7}$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

Computing expected counts in a two-way table

To identify the expected count for the ith row and jth column, compute

$$\text{Expected Count}_{\text{row i, col j}} = \frac{(\text{row i total}) \times (\text{column j total})}{\text{table total}} \tag{9.2.4.8}$$

The chi-square Test for Two-way Tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

$$\text{General formula} \frac{(\text{observed count - expected count})^{2}}{\text{expected count}} \tag{9.2.4.9}$$

$$\text{Row 1, Col 1} \frac{(3511-3539)^2}{3539} = 0.222 \tag{9.2.4.10}$$

$$\text{Row 1, Col 2} \frac{(1749-1769.5)^2}{1769.5} = 0.237 \tag{9.2.4.11}$$

$$\vdots \tag{9.2.4.12}$$

$$\text{Row 2, Col 3} \frac{(682-730.5)^2}{730.5} = 3.220 \tag{9.2.4.13}$$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120 \tag{9.2.4.14}$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1}) \tag{9.2.4.15}$$

In our example, the degrees of freedom parameter is

$$df = (2-1) \times (3-1) = 2 \tag{9.2.4.16}$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic X^2 = 6.12 closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.18.

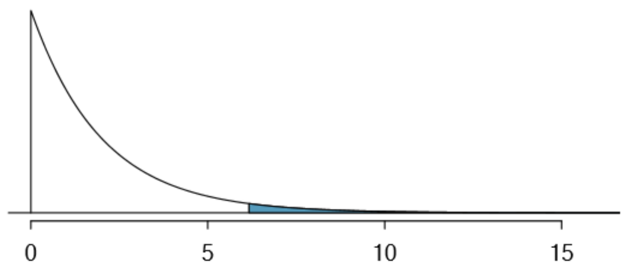


Figure 6.18: Computing the p-value for the Google hypothesis test.

Definition: degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df = (R-1) \times (C-1) \tag{9.2.4.17}$$

where R is the number of rows in the table and C is the number of columns.

[22]*Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.*

Table 6.19: Pew Research poll results of a March 2012 poll.

		Congress		
	Obama	Democrats	Republicans	Total

Approve Disapprove	842 616	736 646	541 842	2119 2104
Total	1458	1382	1383	4223

TIP: Use two-proportion methods for 2-by-2 contingency tables

When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 6.2.

Example 9.2.4.1

Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Solution

Looking in Appendix B.3 on page 412, we examine the row corresponding to 2 degrees of freedom. The test statistic, X^2 = 6.120, falls between the fourth and fth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of α = 0.05 and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Example 9.2.4.1

Table 6.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

Solution

- H_0: There is no difference in approval ratings between the three groups.
- H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23]See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 6.19 are approximate.

Exercise 9.2.4.1

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.38. As a rst step, compute the expected values for each of the six table cells.[24]

[24]The expected count for row one / column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119\times1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104\times1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

Exercise 9.2.4.1

Compute the chi-square test statistic.[25]

[25]For each cell, compute $\frac{(\text{obs - exp})^2}{exp}$. For instance, the rst row and rst column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \cdots + 34.0 = 106.4$.

Exercise 9.2.4.1

Because there are 2 rows and 3 columns, the degrees of freedom for the test is df = (2 - 1) (3 - 1) = 2. Use X^2 = 106.4, df = 2, and the chi-square table on page 412 to evaluate whether to reject the null hypothesis.[26]

[26]The test statistic is larger than the right-most column of the df = 2 row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans'

approval has differences among Democrats in Congress, Republicans in Congress, and the president.

9.2.5: Small Sample Hypothesis Testing for a Proportion (Special Topic)

In this section we develop inferential methods for a single proportion that are appropriate when the sample size is too small to apply the normal model to $\hat{p}$. Just like the methods related to the t distribution, these methods can also be applied to large samples.

When the Success-Failure Condition is Not Met

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspect of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients. One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Exercise 9.2.5.1

Exercise 6.42

We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.

Solution

The sample proportion: $\hat{p} = \frac{3}{62} = 0.048$

Example 9.2.5.1

Is it possible to assess the consultant's claim using the data provided?

Solution

No. The claim is that there is a causal connection, but the data are observational. Patients who hire this medical consultant may have lower complication rates for other reasons.

While it is not possible to assess this causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Exercise 9.2.5.1

Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.

Solution

- H_0: There is no association between the consultant's contributions and the clients' complication rate. In statistical language, p = 0.10.
- H_A: Patients who work with the consultant tend to have a complication rate lower than 10%, i.e. p < 0.10.

Example 9.2.5.1

In the examples based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

Solution

The independence assumption may be reasonable if each of the surgeries is from a different surgical team. However, the success-failure condition is not satis ed. Under the null hypothesis, we would anticipate seeing $62 \times 0.10 = 6.2$ complications, not the 10 required for the normal approximation.

The uncertainty associated with the sample proportion should not be modeled using the normal distribution. However, we would still like to assess the hypotheses from Exercise 6.44 in absence of the normal framework. To do so, we need to evaluate the possibility of a sample value (^p) this far below the null value, $p_0 = 0.10$. This possibility is usually measured with a p-value.

The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis is true. Supposing the null hypothesis is true, we can compute the p-value by identifying the chance of observing a test statistic that favors the alternative hypothesis at least as strongly as the observed test statistic. This can be done using simulation.

Generating the null distribution and p-value by simulation

We want to identify the sampling distribution of the test statistic ($\hat{p}$) if the null hypothesis was true. In other words, we want to see how the sample proportion changes due to chance alone. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.

Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients. If this was the case, we could simulate 62 clients to get a sample proportion for the complication rate from the null distribution.

Each client can be simulated using a deck of cards. Take one red card, nine black cards, and mix them up. Then drawing a card is one way of simulating the chance a patient has a complication if the true complication rate is 10% for the data. If we do this 62 times and compute the proportion of patients with complications in the simulation, $\hat{p}_{sim}$, then this sample proportion is exactly a sample from the null distribution.

An undergraduate student was paid $2 to complete this simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e. $\hat{p}_{sim} = \frac{5}{62} = 0.081$.

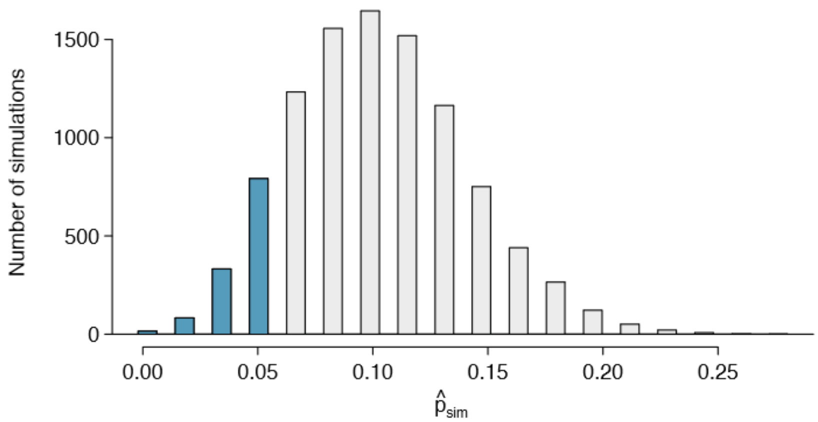


Figure 6.20: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

Example 9.2.5.1

Is this one simulation enough to determine whether or not we should reject the null hypothesis from Exercise 6.44? Explain.

Solution

No. To assess the hypotheses, we need to see a distribution of many $\hat{p}_{sim}$, not just a single draw from this sampling distribution.

One simulation isn't enough to get a sense of the null distribution; many simulation studies are needed. Roughly 10,000 seems sufficient. However, paying someone to simulate 10,000 studies by hand is a waste of time and money. Instead, simulations are typically programmed into a computer, which is much more efficient.

Figure 6.20 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and nd the p-value:

$$\text{left tail} = \frac{\text{Number of observed simulations with} \hat{p}_{sim} \leq 0.048}{10000} \tag{6.47}$$

Of the 10,000 simulated $\hat{p}_{sim}$, 1222 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.1222.

Exercise 9.2.5.1

Because the estimated p-value is 0.1222, which is larger than the signi cance level 0.05, we do not reject the null hypothesis. Explain what this means in plain language in the context of the problem.

Solution

There isn't sufficiently strong evidence to support an association between the consultant's work and fewer surgery complications.

Exercise 9.2.5.1

Does the conclusion in Exercise 6.48 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.

Solution

No. It might be that the consultant's work is associated with a reduction but that there isn't enough data to convincingly show this connection.

One-sided hypothesis test for p with a small sample

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions ($\hat{p}_{sim}$) and use this distribution to compute the tail area, i.e. the p-value.

We continue to use the same rule as before when computing the p-value for a twosided test: double the single tail area, which remains a reasonable approach even when the sampling distribution is asymmetric. However, this can result in p-values larger than 1 when the point estimate is very near the mean in the null distribution; in such cases, we write that the p-value is 1. Also, very large p-values computed in this way (e.g. 0.85), may also be slightly inated.

Exercise 6.48 said the p-value is estimated. It is not exact because the simulated null distribution itself is not exact, only a close approximation. However, we can generate an exact null distribution and p-value using the binomial model from Section 3.4.

Generating the exact null distribution and p-value

The number of successes in n independent cases can be described using the binomial model, which was introduced in Section 3.4. Recall that the probability of observing exactly k successes is given by

$$P(\text{k successes}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{6.50}$$

where p is the true probability of success. The expression $\binom{n}{k}$ is read as n choose k, and the exclamation points represent factorials. For instance, 3! is equal to $3 \times 2 \times 1 = 6, 4!$ is equal to $4 \times 3 \times 2 \times 1 = 24$, and so on (see Section 3.4).

The tail area of the null distribution is computed by adding up the probability in Equation (6.50) for each k that provides at least as strong of evidence favoring the alternative hypothesis as the data. If the hypothesis test is one-sided, then the p-value is represented by a single tail area. If the test is two-sided, compute the single tail area and double it to get the p-value, just as we have done in the past.

Example 9.2.5.1

Compute the exact p-value to check the consultant's claim that her clients' complication rate is below 10%.

Solution

Exactly k = 3 complications were observed in the n = 62 cases cited by the consultant. Since we are testing against the 10% national average, our null hypothesis is p = 0.10. We can compute the p-value by adding up the cases where there are 3 or fewer complications:

$$\text{p-value} = \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \tag{9.2.5.1}$$

$$= \sum_{j=0}^{3} \binom{62}{j} 0.1^{j} (1-0.1)^{62-j} \tag{9.2.5.2}$$

$$= \binom{62}{0} 0.1^{0} (1-0.1)^{62-0} + \binom{62}{1} 0.1^{1} (1-0.1)^{62-1} + \binom{62}{2} 0.1^{2} (1-0.1)^{62-2} + \binom{62}{3} 0.1^{3} (1-0.1)^{62-3} \tag{9.2.5.3}$$

$$= 0.0015 + 0.0100 + 0.0340 + 0.0755 \tag{9.2.5.4}$$

$$= 0.1210 \tag{9.2.5.5}$$

This exact p-value is very close to the p-value based on the simulations (0.1222), and we come to the same conclusion. We do not reject the null hypothesis, and there is not statistically signi cant evidence to support the association.

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 6.20 on page 290.

Using simulation for goodness of fit tests

Simulation methods may also be used to test goodness of t. In short, we simulate a new sample based on the purported bin probabilities, then compute a chi-square test statistic X^2_{sim}. We do this many times (e.g. 10,000 times), and then examine the distribution of these simulated chi-square test statistics. This distribution will be a very precise null distribution for the test statistic X^2 if the probabilities are accurate, and we can nd the upper tail of this null distribution, using a cutoff of the observed test statistic, to calculate the p-value.

Example 9.2.5.1

Section 6.3 introduced an example where we considered whether jurors were racially representative of the population. Would our ndings differ if we used a simulation technique?

Solution

Since the minimum bin count condition was satis ed, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 6.21 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

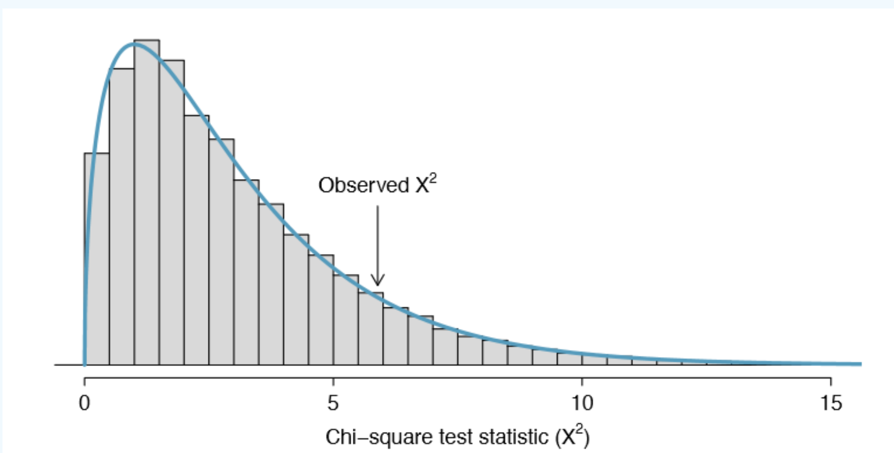


Figure 6.21: The precise null distribution for the juror example from Section 6.3 is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.

9.2.6: Randomization Test (Special Topic)

Cardiopulmonary resuscitation (CPR) is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. This procedure is helpful in maintaining some blood circulation, but the chest compressions involved can also cause internal injuries. Internal bleeding and other injuries complicate additional treatment efforts following arrival at a hospital. For instance, blood thinners may be used to help release a clot that is causing the heart attack. However, the blood thinner would negatively affect an internal injury. Here we consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.(Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial, by Bottiger et al., The Lancet, 2001.) These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive the blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

Example 9.2.6.1

Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival proportion in the control group and p_t represent the survival proportion for the treatment group.

Solution

We are interested in whether the blood thinners are helpful or harmful, so this should be a two-sided test.

- H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group.

$$p_t - p_c = 0. \tag{9.2.6.1}$$

- H_A: Blood thinners do have an impact on survival.

$$p_t - p_c \neq 0. \tag{9.2.6.2}$$

Large Sample Framework for a Difference in Two Proportions

There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did. The study results are shown in Table 6.22.

Table 6.22: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Exercise 9.2.6.1

What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$.

Solution

Observed control survival rate:

$$p_c = \frac{11}{50} = 0.22. \tag{9.2.6.3}$$

Treatment survival rate:

$$p_t = \frac{14}{40} = 0.35. \tag{9.2.6.4}$$

Observed difference:

$$\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13. \tag{9.2.6.5}$$

According to the point estimate, there is a 13% increase in the survival proportion when patients who have undergone CPR outside of the hospital are treated with blood thinners. However, we wonder if this difference could be due to chance. We'd like to investigate this using a large sample framework, but we first need to check the conditions for such an approach.

Example $9.2.6.2$

Can the point estimate of the difference in survival proportions be adequately modeled using a normal distribution?

Solution

We will assume the patients are independent, which is probably reasonable. The success-failure condition is also satisfied. Since the proportions are equal under the null, we can compute the pooled proportion,

$$\hat{p} = \frac{(11+14)}{(50+40)} = 0.278, \tag{9.2.6.6}$$

for checking conditions. We nd the expected number of successes (13.9, 11.1) and failures (36.1, 28.9) are above 10. The normal model is reasonable.

While we can apply a normal framework as an approximation to nd a p-value, we might keep in mind that the expected number of successes is only 13.9 in one group and 11.1 in the other. Below we conduct an analysis relying on the large sample normal theory. We will follow up with a small sample analysis and compare the results.

Example $9.2.6.3$

Assess the hypotheses presented in Example 6.53 using a large sample framework. Use a significance level of $\alpha = 0.05$.

Solution

We suppose the null distribution of the sample difference follows a normal distribution with mean 0 (the null value) and a standard deviation equal to the standard error of the estimate. The null hypothesis in this case would be that the two proportions are the same, so we compute the standard error using the pooled standard error formula from Equation (6.16) on page 273:

$$SE = \sqrt{\frac{p(1-p)}{n_t} + \frac{p(1-p)}{n_c}} \approx \sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}} = 0.095 \tag{9.2.6.7}$$

where we have used the pooled estimate $\left(\hat{p} = \frac{11+14}{50+40} = 0.278\right)$ in place of the true proportion, p.

The null distribution with mean zero and standard deviation 0.095 is shown in Figure 6.23. We compute the tail areas to identify the p-value. To do so, we use the Z score of the point estimate:

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - \text{null value}}{SE} = \frac{0.13 - 0}{0.095} = 1.37 \tag{9.2.6.8}$$

If we look this Z score up in Appendix B.1, we see that the right tail has area 0.0853. The p-value is twice the single tail area: 0.176. This p-value does not provide convincing evidence that the blood thinner helps. Thus, there is insufficient evidence to conclude whether or not the blood thinner helps or hurts. (Remember, we never "accept" the null hypothesis - we can only reject or fail to reject.)

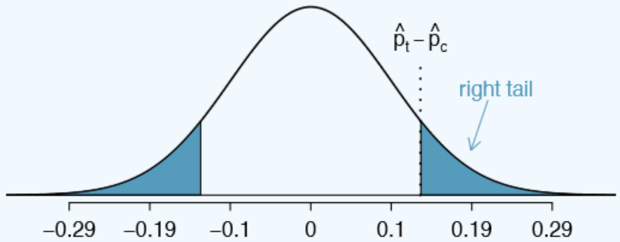


Figure 6.23: The null distribution of the point estimate $\hat{p}_t - \hat{p}_c$ under the large sample framework is a normal distribution with mean 0 and standard deviation equal to the standard error, in this case SE = 0.095. The p-value is represented by the shaded areas.

The p-value 0.176 relies on the normal approximation. We know that when the samples sizes are large, this approximation is quite good. However, when the sample sizes are relatively small as in this example, the approximation may only be adequate. Next we develop a simulation technique, apply it to these data, and compare our results. In general, the small sample method we develop may be used for any size sample, small or large, and should be considered as more accurate than the corresponding large sample technique.

Simulating a Difference under the null Distribution

The ideas in this section were first introduced in the optional Section 1.8. Suppose the null hypothesis is true. Then the blood thinner has no impact on survival and the 13% difference was due to chance. In this case, we can simulate null differences that are due to chance using a randomization technique. (The test procedure we employ in this section is formally called a **permutation test**). By randomly assigning "fake treatment" and "fake control" stickers to the patients' files, we could get a new grouping - one that is completely due to chance. The expected difference between the two proportions under this simulation is zero.

We run this simulation by taking 40 treatment fake and 50 control fake labels and randomly assigning them to the patients. The label counts of 40 and 50 correspond to the number of treatment and control assignments in the actual study. We use a computer program to randomly assign these labels to the patients, and we organize the simulation results into Table 6.24.

Table 6.24: Simulated results for the CPR study under the null hypothesis. The labels were randomly assigned and are independent of the outcome of the patient.

	Survived	Died	Total
Control_fake	15	35	50
Treatment_fake	10	30	40
Total	25	65	90

Exercise 9.2.6.2

What is the difference in survival rates between the two fake groups in Table 6.24? How does this compare to the observed 13% in the real groups?

Solution

The difference is $\hat{p}_{\text{t;fake}} - \hat{p}_{\text{c;fake}} = \frac{10}{40} - \frac{15}{50} = -0.05$, which is closer to the null value p0 = 0 than what we observed.

The difference computed in Exercise 6.57 represents a draw from the null distribution of the sample differences. Next we generate many more simulated experiments to build up the null distribution, much like we did in Section 6.5.2 to build a null distribution for a one sample proportion.

Caution: Simulation in the two proportion case requires that the null difference is zero

The technique described here to simulate a difference from the null distribution relies on an important condition in the null hypothesis: there is no connection between the two variables considered. In some special cases, the null difference might not be zero, and more advanced methods (or a large sample approximation, if appropriate) would be necessary.

Null distribution for the difference in two proportions

We build up an approximation to the null distribution by repeatedly creating tables like the one shown in Table 6.24 and computing the sample differences. The null distribution from 10,000 simulations is shown in Figure 6.25.

Example $9.2.6.4$

Compare Figures 6.23 and 6.25. How are they similar? How are they different?

Solution

The shapes are similar, but the simulated results show that the continuous approximation of the normal distribution is not very good. We might wonder, how close are the p-values?

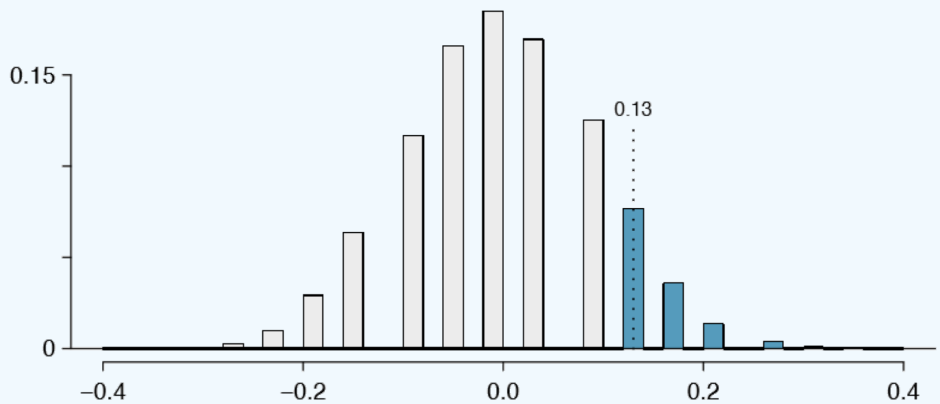


Figure 6.25: An approximation of the null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The p-value is twice the right tail area.

Exercise $9.2.6.3$

The right tail area is about 0.13. (It is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) The p-value is computed by doubling the right tail area: 0.26. How does this value compare with the large sample approximation for the p-value?

Solution

The approximation in this case is fairly poor (p-values: 0.174 vs. 0.26), though we come to the same conclusion. The data do not provide convincing evidence showing the blood thinner helps or hurts patients.

In general, small sample methods produce more accurate results since they rely on fewer assumptions. However, they often require some extra work or simulations. For this reason, many statisticians use small sample methods only when conditions for large sample methods are not satisfied.

Randomization for two-way tables and chi-square

Randomization methods may also be used for the contingency tables. In short, we create a randomized contingency table, then compute a chi-square test statistic X^2_{sim}. We repeat this many times using a computer, and then we examine the distribution of these simulated test statistics. This randomization approach is valid for any sized sample, and it will be more accurate for cases where one or more expected bin counts do not meet the minimum threshold of 5. When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution. As before, we use the upper tail of the null distribution to calculate the p-value.

9.2.7: Exercises

Inference for a single proportion

6.1 Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

(b) The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

(c) A random sample of 125 college students where 12% are vegetarians would be considered unusual.

(d) A random sample of 250 college students where 12% are vegetarians would be considered unusual.

(e) The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

6.2 Young Americans, Part I. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.[36]

(a) The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.

(b) The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

(c) A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

(d) A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

6.3 Orange tabbies. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

6.4 Young Americans, Part II. About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning.[37]

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

(b) In order for the the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

[36]*A. Vaughn. "Poll finds young adults optimistic, but not about money". In: Los Angeles Times (2011).*

[37]*Demos.org. "The State of Young America: The Poll". In: (2011).*

6.5 Prop 19 in California. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[38]

(a) We are 95% con dent that between 62% and 78% of the California voters in this sample support Prop 19.

(b) We are 95% con dent that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.

(c) If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.

(d) In order to decrease the margin of error to 4%, we would need to quadruple (multiply by 4) the sample size.

(e) Based on this con dence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[39]

(a) We are 95% con dent that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b) We are 95% con dent that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

(d) The margin of error at a 90% con dence level would be higher than 3%.

6.7 Fireworks on July 4th. In late June 2012, Survey USA published results of a survey stating that 56% of the 600 randomly sampled Kansas residents planned to set off reworks on July 4th. Determine the margin of error for the 56% point estimate using a 95% con dence level.40

6.8 Elderly drivers. In January 2011, The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% con dence level.[41]

(a) Verify the margin of error reported by The Marist Poll.

(b) Based on a 95% con dence interval, does the poll provide convincing evidence that more than 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?

[38]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

[39]*Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.*

[40]*Survey USA, News Poll #19333, data collected on June 27, 2012.*

[41]*Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.*

6.9 Life after college. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and nd out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions for constructing a con dence interval based on these data are met.

(c) Calculate a 95% con dence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

(d) What does "95% confidence" mean?

(e) Now calculate a 99% con dence interval for the same parameter and interpret it in the context of the data.

(f) Compare the widths of the 95% and 99% con dence intervals. Which one is wider? Explain.

6.10 Life rating in Greece. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered

"suffering".[42]

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions required for constructing a con dence interval based on these data are met.

(c) Construct a 95% con dence interval for the proportion of Greeks who are "suffering".

(d) Without doing any calculations, describe what would happen to the con dence interval if we decided to use a higher confidence level.

(e) Without doing any calculations, describe what would happen to the con dence interval if we used a larger sample.

6.11 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[43]

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% con dence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

(c) What does "90% confidence" mean?

(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

[42]Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.

[43]studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[44]

(a) Is 48% a sample statistic or a population parameter? Explain.

(b) Construct a 95% con dence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

(c) A critic points out that this 95% con dence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

(d) A news piece on this survey's ndings states, \Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

6.13 Public option, Part I. A Washington Post article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More speci cally, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.[45]

(a) A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

(b) Would you expect a con dence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

6.14 The Civil War. A national survey conducted in 2011 among a simple random sample of 1,507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life.[46]

(a) Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) Interpret the p-value in this context.

(c) Calculate a 90% con dence interval for the proportion of Americans who think the Civil War is still relevant. Interpret the interval in this context, and comment on whether or not the con dence interval agrees with the conclusion of the hypothesis test.

6.15 Browsing on the mobile device. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.[47]

(a) According to an online article, a report from a mobile research company indicates that 38 percent of Chinese mobile web users only access the internet through their cell phones.48 Conduct a hypothesis test to determine if these data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%.

(b) Interpret the p-value in this context.

(c) Calculate a 95% con dence interval for the proportion of Americans who access the internet on their cell phones, and interpret the interval in this context.

[44]*National Opinion Research Center, General Social Survey, 2010.*

[45]*D. Balz and J. Cohen. "Most support public option for health insurance, poll nds". In: The Washington Post (2009).*

[46]*Pew Research Center Publications, Civil War at 150: Still Relevant, Still Divisive, data collected between March 30 - April 3, 2011.*

[47]*Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.*

[48]*S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: M.I.C Gadget (2012).*

6.16 Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.[49]

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a con dence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

6.17 Taste test. Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then lled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identi ed the soda.

(a) Do these data provide strong evidence that these people are able to detect the difference between diet and regular soda, in other words, are the results signi cantly better than just random guessing?

(b) Interpret the p-value in this context.

6.18 Is college worth it? Part II. Exercise 6.16 presents the results of a poll where 48% of 331 Americans who decide to not go to college do so because they cannot afford it.

(a) Calculate a 90% con dence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

(b) Suppose we wanted the margin of error for the 90% con dence level to be about 1.5%. How large of a survey would you recommend?

6.19 College smokers. We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke.

(a) Calculate a 95% con dence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: check conditions)

(b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

6.21 Public option, Part II. Exercise 6.13 presents the results of a poll evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.22 Acetaminophen and liver damage. It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher wants to conduct a study to estimate the proportion of acetaminophen users who have liver damage. For participating in this study, he will pay each subject $20 and provide a free medical consultation if the patient has liver damage.

(a) If he wants to limit the margin of error of his 98% confidence interval to 2%, what is the minimum amount of money he needs to set aside to pay his subjects?

(b) The amount you calculated in part (a) is substantially over his budget so he decides to use fewer subjects. How will this affect the width of his con dence interval?

[49]*Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.*

Difference of two proportions

6.23 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

6.24 Heart transplant success. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might bene t from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.[50]

	Control	treatment
alive dead	4 30	24 45

A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate a con dence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

6.25 Gender and color preference. A 2001 study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.[51]

(a) We are 95% con dent that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% con dent that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% con dence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a signi cant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% con dence interval for ($p_{female} - p_{male}$) cannot be calculated with only the information given in this exercise.

[50]*B. Turnbull et al. "Survivorship of Heart Transplant Data". In: Journal of the American Statistical Association 69 (1974), pp. 74 - 80.*

[51]*L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: Personality and Individual Differences 31.8 (2001), pp. 1375-1379.*

6.26 The Daily Show. A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. A 95% con dence interval for ($p_{collegegrad} - p_{HSorless}$), where p is the proportion of those who watch The Daily Show, is (0.07, 0.15). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[52]

(a) At the 5% significance level, the data provide convincing evidence of a difference between the proportions of college graduates and those with a high school degree or less who watch The Daily Show.

(b) We are 95% con dent that 7% less to 15% more college graduates watch The Daily Show than those with a high school degree or less.

(c) 95% of random samples of 1,099 college graduates and 1,110 people with a high school degree or less will yield differences in sample proportions between 7% and 15%.

(d) A 90% confidence interval for ($p_{collegegrad} - p_{HSorless}$) would be wider.

(e) A 95% confidence interval for ($p_{HSorless} - p_{collegegrad}$) is (-0.15,-0.07).

6.27 Public Option, Part III. Exercise 6.13 presents the results of a poll evaluating support for the health care public option plan in 2009. 70% of 819 Democrats and 42% of 783 Independents support the public option.

(a) Calculate a 95% confidence interval for the difference between ($p_D - p_I$) and interpret it in this context. We have already checked conditions for you.

(b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the public option than the Independent.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% con dence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[53]

6.29 Offshore drilling, Part I. A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[54]

	College Grad	
	Yes	No

Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

[52]*The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.*

[53]*CDC, Perceived Insu�cient Rest or Sleep Among Adults - United States, 2008.*

[54]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

6.30 Sleep deprivation, CA vs. OR, Part II. Exercise 6.28 provides data on sleep deprivation rates of Californians and Oregonians. The proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Conduct a hypothesis test to determine if these data provide strong evidence the rate of sleep deprivation is different for the two states. (Reminder: check conditions)

(b) It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

6.31 Offshore drilling, Part II. Results of a poll evaluating support for drilling for oil and natural gas off the coast of California were introduced in Exercise 6.29.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample support drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who support offshore drilling in California is different than that of noncollege graduates.

6.32 Full body scan, Part I. A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1,137 adults nationwide, interviewed by telephone November 7-10, 2010, where one of the questions on the survey was "Some airports are now using `full-body' digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?" Below is a summary of responses based on party affiliation.[55]

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22

Total	318	369	450

(a) Conduct an appropriate hypothesis test evaluating whether there is a difference in the proportion of Republicans and Democrats who think the full-body scans should be applied in airports. Assume that all relevant conditions are met.

(b) The conclusion of the test in part (a) may be incorrect, meaning a testing error was made. If an error was made, was it a Type I or a Type II error? Explain.

[55]*S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: CBS News (2010).*

6.33 Sleep deprived transportation workers. The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers. The results of the survey are shown below.[56]

		Transportation	Professionals		
	Control	Pilots	Truck Drivers	Train Operators	Bux/Taxi/Limo Drivers
Less than 6 hours of sleep 6 to 8 hours of sleep More than 8 hours	35 193 64	19 132 51	35 117 51	29 119 32	21 131 58
Tota	292	202	203	180	210

Conduct a hypothesis test to evaluate if these data provide evidence of a difference between the proportions of truck drivers and non-transportation workers (the control group) who get less than 6 hours of sleep per day, i.e. are considered sleep deprived.

6.34 Prenatal vitamins and Autism. Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period).[57]

	Autism		
	Autism	Typical development	Total
No vitamin Vitamin	111 143	70 159	181 302
Total	254	229	483

(a) State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.

(b) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

(c) A New York Times article reporting on this study was titled "Prenatal Vitamins May Ward Off Autism". Do you nd the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.[58]

6.35 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called virologic failure). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[59]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

[56]*National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers Sleep, 2012.*

[57]*R.J. Schmidt et al. \Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism". In: Epidemiology 22.4 (2011), p. 476.*

[58]*R.C. Rabin. "Patterns: Prenatal Vitamins May Ward Off Autism". In: New York Times (2011).*

[59]*S. Lockman et al. "Response to antiretroviral therapy after a single, peripartum dose of nevirapine". In: Obstetrical & gynecological survey 62.6 (2007), p. 361.*

6.36 Diabetes and unemployment. A 2012 Gallup poll surveyed Americans about their employment status and whether or not they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.[60]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of incidence of diabetes and employment status.

(c) The sample difference is about 1%. If we completed the hypothesis test, we would nd that the p-value is very small (about 0), meaning the difference is statistically signi cant. Use this result to explain the difference between statistically signi cant and practically significant findings.

Testing for goodness of t using chi-square

6.37 True or false, Part I. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

(c) The chi-square statistic is always positive.

(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

6.38 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found X^2 = 10 with df = 5 you would fail to reject H0 at the 5% signi cance level.

(c) When nding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.39 Open source textbook. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

(a) State the hypotheses for testing if the professor's predictions were inaccurate.

(b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

(c) This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

(d) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

(e) Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

[60]*Gallup Wellbeing, Employed Americans in Better Health Than the Unemployed, data collected Jan. 2, 2011 - May 21, 2012.*

6.40 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[61]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1) Humans evolved, but God had no part in process (2) God created humans in present form (3) Other / No opinion (4)	38% 16% 40% 6%	37% 12% 45% 6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)

Testing for independence in two-way tables

6.41 Quitters. Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

(a) Create a two-way table presenting the results of this study.

(b) Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

i. How many subjects in the "patch + support" group would you expect to quit?

ii. How many subjects in the "only patch" group would you expect to not quit?

6.42 Full body scan, Part II. The table below summarizes a data set we rst encountered in Exercise 6.32 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's party affiliation and his support of full-body scans. It may be useful to rst add on an extra column for row totals before proceeding with the computations.

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22
Total	318	369	450

(a) How many Republicans would you expect to not support the use of full-body scans?

(b) How many Democrats would you expect to support the use of full-body scans?

(c) How many Independents would you expect to not know or not answer?

[61] *Four in 10 Americans Believe in Strict Creationism, December 17, 2010, http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx*.

6.43 Offshore drilling, Part III. The table below summarizes a data set we rst encountered in Exercise 6.29 that examines the responses of a random sample of college graduates and nongraduates on the topic of oil drilling. Complete a chi-square test for

these data to check whether there is a statistically signi cant difference in responses from college graduates and non-graduates.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

6.44 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[62]

		Caffeinated	coffee	consumption		
	$\leq$ 1cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Yes No	670 11,545	373 6,244	905 16,329	564 11,726	95 2,288	2,607 48,132
Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identi ed in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $\frac{(Observed-Expected)^2}{Expected}$.

(e) The test statistic is $X^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.63 Do you agree with this statement? Explain your reasoning.

[62]*M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: Archives of internal medicine 171.17 (2011), p. 1571.*

[63]*A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: New York Times (2011).*

6.45 Privacy on Facebook. A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, "Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?" The responses are cross-tabulated based on gender.[64]

	Gender		
	Male	Female	Total
Yes No Not sure	288 61 10	378 62 7	666 123 17
Total	359	447	806

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

6.46 Shipping holiday gifts. A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age		
	18-34	35-54	55+	Total
USPS	72 (81)	97 (102)	76 (62)	245
UPS	52 (53)	76 (68)	34 (41)	162
FedEx	31 (21)	24 (27)	9 (16)	64
Something else	7 (5)	6 (7)	3 (4)	16
Not sure	3 (5)	6 (5)	4 (3)	13
Total	165	209	126	500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

(b) Are the conditions for inference using a chi-square test satisfied?

Small sample hypothesis testing for a proportion

6.47 Bullying in schools. A 2012 Survey USA poll asked Florida residents how big of a problem they thought bullying was in local schools. 9 out of 191 18-34 year olds responded that bullying is no problem at all. Using these data, is it appropriate to construct a con dence interval using the formula $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for the true proportion of 18-34 year old Floridians who think bullying is no problem at all? If it is appropriate, construct the con dence interval. If it is not, explain why.

[64]*Survey USA, News Poll #17960, data collected February 16-17, 2011.*

6.48 Choose a test. We would like to test the following hypotheses:

$H_0 : p = 0.1$

$H_A : p \neq 0.1$

The sample size is 120 and the sample proportion is 8.5%. Determine which of the below test(s) is/are appropriate for this situation and explain your reasoning.

I. Z test for a proportion,

i.e. proportion test using normal model

II. Z test for comparing two proportions

III. X^2 test of independence

IV. Simulation test for a proportion

V. t test for a mean

VI. ANOVA

6.49 The Egyptian Revolution. A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the rst couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.[65]

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

(c) Based on large sample theory, we modeled ^p using the normal distribution. Why should we be cautious about this approach for these data?

(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.

(e) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.

[65]*Gallup Politics, Americans' Views of Egypt Sharply More Negative, data collected February 2-5, 2011.*

6.50 Assisted Reproduction. Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g. in vitro fertilization). A 2008 report by the Centers for Disease Control and Prevention estimated that ART has been successful in leading to a live birth in 31% of cases66. A new fertility clinic claims that their success rate is higher than average. A random sample of 30 of their patients yielded a success rate of 40%. A consumer watchdog group would like to determine if this provides strong evidence to support the company's claim.

(a) Write the hypotheses to test if the success rate for ART at this clinic is signi cantly higher than the success rate reported by the CDC.

(b) Based on large sample theory, we modeled ^p using the normal distribution. Why is this not appropriate here?

(c) The normal approximation would be less reliable here, so we should use a simulation strategy. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

(d) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

(e) After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.

[66]*CDC. 2008 Assisted Reproductive Technology Report.*

Hypothesis testing for two proportions

6.51 Social experiment, Part II. Exercise 6.23 introduces a "social experiment" conducted by a TV program that questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

A simulation was conducted to test if people react differently under the two scenarios. 10,000 simulated differences were generated to construct the null distribution shown. The value $\hat{p}_{pr;sim}$ represents the proportion of diners who intervened in the simulation for the provocatively dressed woman, and $\hat{p}_{con;sim}$ is the proportion for the conservatively dressed woman.

(a) What are the hypotheses? For the purposes of this exercise, you may assume that each observed person at the restaurant behaved independently, though we would want to evaluate this assumption more rigorously if we were reporting these results.

(b) Calculate the observed difference between the rates of intervention under the provocative and conservative scenarios: $\hat{p}_{pr} - \hat{p}_{con}$.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

6.52 Is yawning contagious? An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously inuenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.[67]

	Group		
	Treatment	Control	Total
Yawn Not Yawn	10 24	4 12	14 36
Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no inuence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning as $\hat{p}_{trtmt;sim} - \hat{p}_{ctrl;sim}$. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.

(a) What are the hypotheses?

(b) Calculate the observed difference between the yawning rates under the two scenarios.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

[67]*MythBusters, Season 3, Episode 28.*

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

SECTION OVERVIEW

9.3: Hypothesis Testing with One Sample

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

9.3.1: Prelude to Hypothesis Testing

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

Figure 9.3.1.1: You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

A statistician will make a decision about these claims. This process is called "hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analysis of the data, to reject the null hypothesis. In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

- Set up two contradictory hypotheses.
- Collect sample data (in homework problems, the data or summary statistics will be given to you).
- Determine the correct distribution to perform the hypothesis test.
- Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
- Make a decision and write a meaningful conclusion.

> To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

9.3.2: Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis:** It is a statement of no difference between the variables—they are not related. This can often be considered the status quo and as a result if you cannot accept the null it requires some action.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0. This is usually what the researcher is trying to prove.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Table 9.3.2.1: Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal $(\neq)$ **or** greater than (>) **or** less than (<)
greater than or equal to $(\geq)$	less than (<)
less than or equal to $(\geq)$	more than (>)

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

Example 9.3.2.1

- H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
- H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Exercise 9.3.2.1

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Answer

- H_0: The drug reduces cholesterol by 25%. $p = 0.25$
- H_a: The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Example 9.3.2.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

- $H_0 : \mu = 2.0$
- $H_a : \mu \neq 2.0$

Exercise 9.3.2.2

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

- $H_0 : \mu\_66$
- $H_a : \mu\_66$

Answer

- $H_0 : \mu = 66$
- $H_a : \mu \neq 66$

Example 9.3.2.3

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

- $H_0 : \mu \geq 5$
- $H_a : \mu < 5$

Exercise 9.3.2.3

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. $H_0 : \mu\_45$
b. $H_a : \mu\_45$

Answer

a. $H_0 : \mu \geq 45$
b. $H_a : \mu < 45$

Example 9.3.2.4

In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

- $H_0 : p \leq 0.066$
- $H_a : p > 0.066$

Exercise 9.3.2.4

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

a. $H_0 : p\_0.40$
b. $H_a : p\_0.40$

Answer

a. $H_0 : p = 0.40$
b. $H_a : p > 0.40$

COLLABORATIVE EXERCISE

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality $(=, \leq \text{ or } \geq)$
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., $(\neq, >, \text{or } <)$.
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_a has:	equal $(=)$	greater than or equal to $(\geq)$	less than or equal to $(\leq)$
then H_a has:	not equal $(\neq)$ **or** greater than $(>)$ **or** less than $(<)$	less than $(<)$	greater than $(>)$

- If $\alpha \leq p$-value, then do not reject H_0.
- If$\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

9.3.2E: Null and Alternative Hypotheses (Exercises)

Exercise $9.3.2E.5$

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

Answer

The random variable is the mean Internet speed in Megabits per second.

Exercise $9.3.2E.1$

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

Exercise $9.3.2E.1$

The American family has an average of two children. What is the random variable? Describe in words.

Answer

The random variable is the mean number of children an American family has.

Exercise $9.3.2E.8$

The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

Exercise $9.3.2E.9$

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

Answer

The random variable is the proportion of people picked at random in Times Square visiting the city.

Exercise $9.3.2E.10$

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

Exercise $9.3.2E.11$

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

Answer

a. $H_0 : p = 0.42$
b. $H_a : p < 0.42$

Exercise $9.3.2E.12$

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years.

Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0:_______
b. H_a:_______

Exercise $9.3.2E.13$

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0:__________
b. H_a:__________

Answer

a. $H_0 : \mu = 15$
b. $H_a : \mu \neq 15$

Exercise

The National Institute 9.2.14 of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

1. H_0:_______
2. H_a:_______

9.3.3: Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 is Actually True	H_0 is Actually False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject** H_0 when H_0 **is true (correct decision).**
2. The decision is to **reject** H_0 when H_0 **is true** (incorrect decision known as aType I error).
3. The decision is **not to reject** H_0 when, in fact, H_0 **is false** (incorrect decision known as a Type II error).
4. The decision is to **reject** H_0 when H_0 **is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

- $\alpha =$ probability of a Type I error $= P(\text{Type I error}) =$ probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta =$ probability of a Type II error $= P(\text{Type II error}) =$ probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The *Power of the Test* is $1-\beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test. The following are examples of Type I and Type II errors.

Example 9.3.3.1: Type I vs. Type II errors

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

- **Type I error**: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
- **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

$\alpha =$ **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.

$\beta =$ **probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Exercise 9.3.3.1

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Answer

- **Type I error**: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.
- **Type II error**: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Example 9.3.3.2

Suppose the null hypothesis, H_0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

- **Type I error**: The emergency crew thinks that the victim is dead when, in fact, the victim is alive.
- **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

$\alpha =$ **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive $= P(\text{Type I error})$.

$\beta =$ **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead $= P(\text{Type II error})$.

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Exercise 9.3.3.2

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Answer

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Example 9.3.3.3

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome.

- **Type I error**: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.
- **Type II error**: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Exercise 9.3.3.3

"Red tide" is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Answer

In this scenario, an appropriate null hypothesis would be H_0: the mean level of toxins is at most $800\mu g$, $H_0 : \mu_0 \leq 800\mu g$.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most $800\mu g$. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than $800\mu g$). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example 9.3.3.4

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

- **Type I**: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
- **Type II**: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

? Exercise 9.3.3.4

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Answer

Type I error: c

Type II error: b

Summary

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected. The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1-\beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

- $\alpha =$ probability of a Type I error $= P(\text{Type I error}) =$ probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta =$ probability of a Type II error $= P(\text{Type II error}) =$ probability of not rejecting the null hypothesis when the null hypothesis is false.

Glossary

Type 1 Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

9.3.3E: Outcomes and the Type I and Type II Errors (Exercises)

Exercise $9.3.3E.5$

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Answer

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

Exercise $9.3.3E.6$

A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

Exercise $9.3.3E.7$

For Exercise 9.12, what are α and β in words?

Answer

$\alpha =$ the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

$\beta =$ the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

Exercise $9.3.3E.8$

In words, describe $1-\beta$ For Exercise $9.3.3E$.

Exercise $9.3.3E.9$

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

Answer

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

Exercise $9.3.3E.10$

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

Exercise $9.3.3E.11$

The power of a test is 0.981. What is the probability of a Type II error?

Answer

0.019

Exercise $9.3.3E.12$

A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

Exercise $9.3.3E.13$

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

Answer

0.998

Exercise $9.3.3E.14$

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

9.3.4: Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a single population mean, the distribution for the test is for *means*:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{9.3.4.1}$$

or

$$t_{df} \tag{9.3.4.2}$$

The population parameter is μ. The estimated value (point estimate) for μ is $\bar{x}$, the sample mean.

If you are testing a single population proportion, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p-q}{n}}\right) \tag{9.3.4.3}$$

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean** μ using a Student's t-distribution (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a binomial distribution which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five $(np > 5$ and $nq > 5)$. Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Summary

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of successes and the mean number of failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α, then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's t-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\bar{x}}$ and $sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Student's *t*-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t-distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

9.3.4E: Distribution Needed for Hypothesis Testing (Exercises)

Exercise $9.3.4E.1$

Which two distributions can you use for hypothesis testing for this chapter?

Answer

A normal distribution or a Student's t-distribution

Exercise $9.3.4E.2$

Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.

Exercise $9.3.4E.3$

Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

Answer

Use a Student's t-distribution

Exercise $9.3.4E.4$

A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

Exercise $9.3.4E.5$

A population has a mean of 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

Answer

a normal distribution for a single population mean

Exercise $9.3.4E.6$

It is thought that 42% of respondents in a taste test would prefer Brand A. In a particular test of 100 people, 39% preferred Brand A. What distribution should you use to perform a hypothesis test?

Exercise $9.3.4E.7$

You are performing a hypothesis test of a single population mean using a Student's t-distribution. What must you assume about the distribution of the data?

Answer

It must be approximately normally distributed.

Exercise $9.3.4E.8$

You are performing a hypothesis test of a single population mean using a Student's t-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

Exercise $9.3.4E.9$

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

Answer

They must both be greater than five.

Exercise $9.3.4E.10$

You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

Exercise $9.3.4E.11$

You are performing a hypothesis test of a single population proportion. The data come from which distribution?

Answer

binomial distribution

9.3.5: Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very *unlikely* to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the p-value. The p-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p-value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example $9.3.5.1$

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.

- The null hypothesis could be $H_0 : \mu \leq 15$
- The alternate hypothesis is $H_a : \mu > 15$

The words **"is more than"** translates as a "$>$" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ **is known** ($\sigma = 0.5 cm.$), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16.$$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

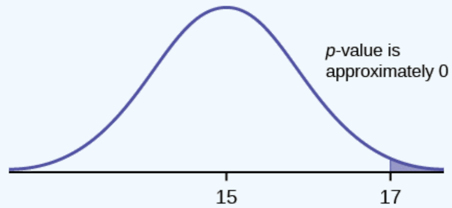


Figure 9.3.5.1

$p\text{-value} = P(\bar{x} > 17)$ which is approximately zero.

A p-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Exercise 9.3.5.1

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

- $H_0 : \mu \leq 12$
- $H_a : \mu > 12$

The p-value is 0.0013

Draw a graph that shows the p-value.

Answer

$p\text{-value} = 0.0013$

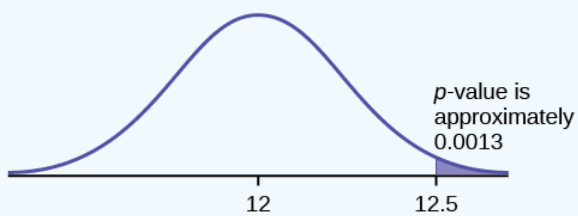


Figure 9.3.5.2

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p-value and a preset or preconceived α (also called a "**significance level**"). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a decision to reject or not reject H_0, do as follows:

- If $\alpha > p\text{-value}$, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a, may be correct.
- If $\alpha \leq p\text{-value}$, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis,H_a, may be correct.

When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Conclusion: After you make your decision, write a thoughtful conclusion about the hypotheses in terms of the given problem.

Example 9.3.5.2

When using the p-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

- If the p-value is low, the null must go.
- If the p-value is high, the null must fly.

This memory aid relates a p-value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p-value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when ____.

The results of the sample data _______________________________________.

Do not reject the null when hypothesis when ____.

The results of the sample data ____.

Answer

Reject the null hypothesis when **the p-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the p-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Exercise 9.3.5.2

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- $H_0 : p = 0.50, H_a : p > 0.50$
- $\alpha = 0.01$
- p-value $= 0.025$

Interpret the results and state a conclusion in simple, non-technical terms.

Answer

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support It's a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

- $\alpha > p - value$, reject the null hypothesis
- $\alpha \leq p - value$, do not reject the null hypothesis

Glossary

Level of Significance of the Test

probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

***p*-value**

the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

9.3.5E: Rare Events, the Sample, Decision and Conclusion (Exercises)

Exercise $9.3.5E.3$

When do you reject the null hypothesis?

Exercise $9.3.5E.4$

The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

Answer

The outcome of winning is very unlikely.

Exercise $9.3.5E.5$

The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

Exercise $9.3.5E.6$

It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p-value is almost zero. State the null and alternative hypotheses and interpret the p-value.

Answer

$H_0 : \mu \geq 73$

$H_a : \mu \leq 73$

The p-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

Exercise $9.3.5E.7$

The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

Exercise $9.3.5E.8$

Does the shaded region represent a low or a high p-value compared to a level of significance of 1%?

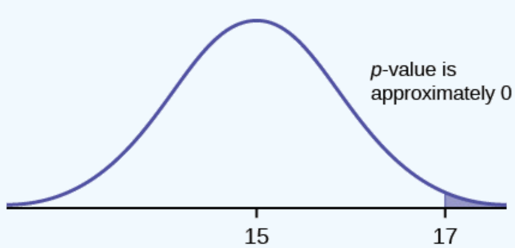


Figure $9.3.5E.3$.

Answer

The shaded region shows a low p-value.

Exercise $9.3.5E.9$

What should you do when $\alpha > p$-value?

Exercise $9.3.5E.10$

What should you do if $\alpha = p$-value?

Answer

Do not reject H_0.

Exercise $9.3.5E.11$

If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Exercise $9.3.5E.12$

Is this a test of means or proportions?

Answer

means

Exercise $9.3.5E.13$

What symbol represents the random variable for this test?

Exercise $9.3.5E.14$

In words, define the random variable for this test.

Answer

the mean time spent in jail for 26 first time convicted burglars

Exercise $9.3.5E.15$

Is the population standard deviation known and, if so, what is it?

Exercise $9.3.5E.16$

Calculate the following:

a. $\bar{x}$ ________
b. σ ________
c. s_x ________
d. n ________

Answer

a. 3
b. 1.5
c. 1.8
d. 26

? Exercise $9.3.5E.17$

Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

? Exercise $9.3.5E.18$

State the distribution to use for the hypothesis test.

Answer

$\bar{X} - N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

? Exercise $9.3.5E.19$

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   $H0$: ____________________ H_a: ____________________
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Calculate the following:
   i. $\bar{x}$ = _____________
   ii. s = ____________
   iii. n = ____________
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.3.6: Additional Information and Full Hypothesis Test Examples

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use before collecting the sample data.
- If no level of significance is given, a common standard to use is $\alpha = 0.05$.
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The alternative hypothesis, H_a, tells you if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- Thinking about the meaning of the p-value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value such as 0.4, as opposed to a p-value of 0.056 ($\alpha = 0.05$ is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 9.3.6.1

$H_0 : \mu = 5, H_a : \mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:

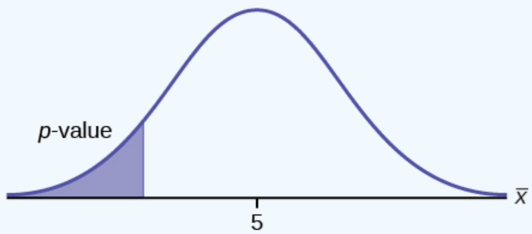


Figure 9.3.6.1

Exercise 9.3.6.1

$H_0 : \mu = 10, H_a : \mu < 10$

Assume the p-value is 0.0935. What type of test is this? Draw the picture of the p-value.

Answer

left-tailed test

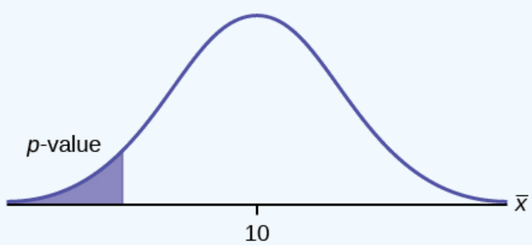


Figure 9.3.6.2

Example 9.3.6.2

$H_0 : \mu \leq 0.2, H_a : \mu > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:

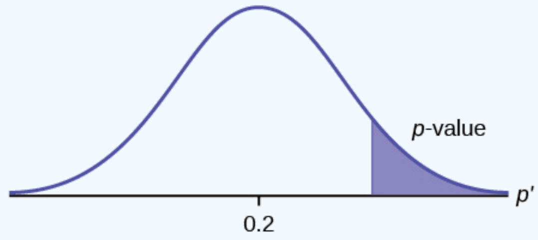


Figure 9.3.6.3

Exercise 9.3.6.2

$H_0 : \mu \leq 1, H_a : \mu > 1$

Assume the p-value is 0.1243. What type of test is this? Draw the picture of the p-value.

Answer

right-tailed test

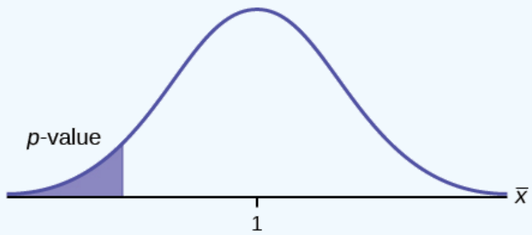


Figure 9.3.6.4

Example 9.3.6.3

$H_0 : \mu = 50, H_a : \mu \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.

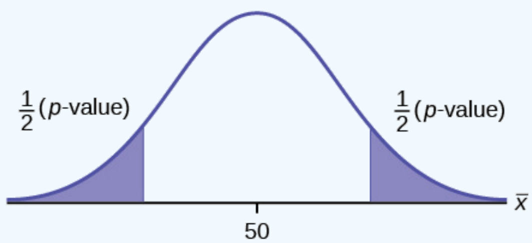


Figure 9.3.6.5

Exercise 9.3.6.3

$H_0 : \mu = 0.5, H_a : \mu \neq 0.5$

Assume the p-value is 0.2564. What type of test is this? Draw the picture of the p-value.

Answer

two-tailed test

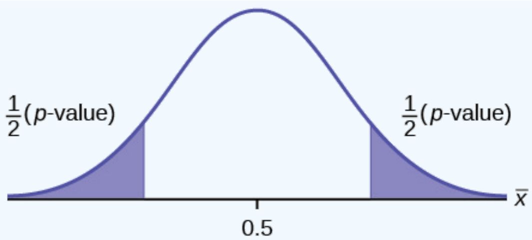


Figure 9.3.6.6

Full Hypothesis Test Examples

Example 9.3.6.4

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.**Conduct a hypothesis test using a preset α = 0.05. Assume that the swim times for the 25-yard freestyle are normal.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

$H_0 : \mu = 16.43, H_a : \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X} =$ the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population standard deviation is known: $\sigma = 0.8$)

$\bar{X} - N\left(\mu, \frac{\sigma_x}{\sqrt{n}}\right)$ Therefore, $\bar{X} - N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the $p-\text{value}$ using the normal distribution for a mean:

$p\text{-value} = P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

$p\text{-value} = 0.0187$ (This is called the **actual level of significance**.) The $p-\text{value}$ is the area to the left of the sample mean is given as 16.

Graph:

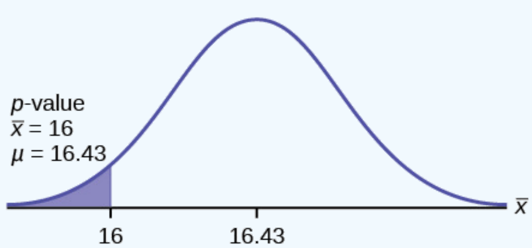


Figure 9.3.6.7

$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the $p-\text{value}$: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the $p-\text{value}$:

$\alpha = 0.05 p\text{-value} = 0.0187 \alpha > p\text{-value}$

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The *p*-value can easily be calculated.

Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the *p*-value ($p = 0.0187$) but it also calculates the test statistic (*z*-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ ($p-\text{value}$). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the `Z-Test` function finds the *p*-value by doing a normal probability calculation using the central limit theorem:

$P(\bar{X} < 16)$ `2nd DISTR normcdf` $((-10^{99}, 16, 16.43, \frac{0.8}{\sqrt{15}})$

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

? Exercise 9.3.6.4

The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco’s mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Press STAT and arrow over to TESTS. Press 1: Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as $<$, $\neq$, or $>$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value but it also calculates the test statistic (*z*-score) for the sample mean. Select $<$, $\neq$, or $>$ for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and *p*-value. Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Answer

Since the problem is about a mean, this is a test of a single population mean.

- $H_0 : \mu = 40$
- $H_a : \mu > 40$
- $p = 0.0062$

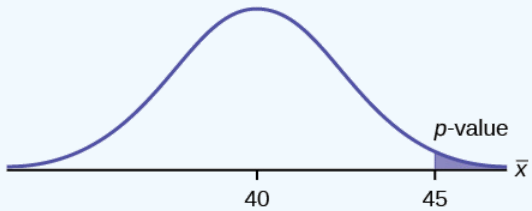


Figure 9.3.6.8

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Historical Note

The traditional way to compare the two probabilities, α and the $p-\text{value}$, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{\bar{x}-\mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$. For this problem, $\bar{x} = 16$, $\mu_x = 16.43$ from the null hypotheses is, $\sigma_x = 0.8$, and $n = 15$.)

You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p-\text{value}$), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the $p-\text{value}$, 0.0187 is considerably smaller than $\alpha = 0.05$. You can be confident about your decision to reject. The graph shows α, the $p-\text{value}$, and the test statistics and the critical value.

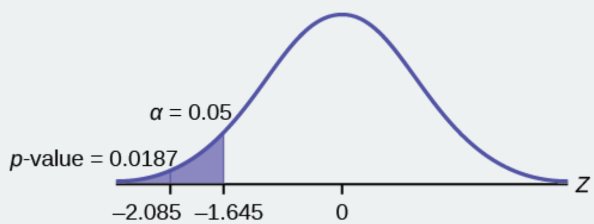


Figure 9.3.6.9

Example 9.3.6.5

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is more than 275 pounds.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a test of a single population mean.

- $H_0 : \mu = 275$
- $H_a : \mu > 275$

This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X} =$ the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

- $\bar{X} - N\left(275, \frac{55}{\sqrt{30}}\right)$
- $\bar{x} = 286.2$ pounds (from the data).
- $\sigma = 55$ pounds **(Always use σ if you know it.)** We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the p-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$$p\text{-value} = P(\bar{x} > 286.2) = 0.1323.$$

Interpretation of the p-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

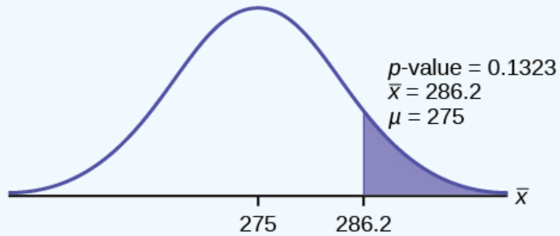


Figure 9.3.6.10

Compare α and the $p-\text{value}$:

$\alpha = 0.025 p-value = 0.1323$

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The $p-\text{value}$ can easily be calculated.

Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the $p-\text{value}$ ($p = 0.1331$), a little different from the previous calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ ($p-\text{value}$). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 9.3.6.6

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Answer

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$H_0 : \mu = 65 \qquad H_a : \mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X} =$ average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p-value using the Student's t-distribution:

$p\text{-value} = P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the *p*-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

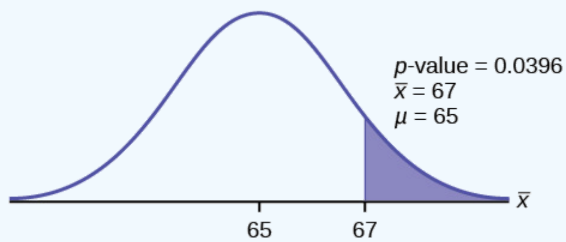


Figure 9.3.6.11

Compare α and the $p-$value:

Since $\alpha = 0.05$ and $p\text{-value} = 0.0396$. $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated.

Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

? Exercise 9.3.6.6

It is believed that a stock price for a particular company will grow at a rate of $5 per week with a standard deviation of $1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: $4, $3, $2, $3, $1, $7, $2, $1, $1, $2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : \mu = 5$
- $H_a : \mu < 5$
- $p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than $5 a week.

- Type I Error: To conclude that the stock price is growing slower than $5 a week when, in fact, the stock price is growing at $5 a week (reject the null hypothesis when the null hypothesis is true).

- Type II Error: To conclude that the stock price is growing at a rate of $5 a week when, in fact, the stock price is growing slower than $5 a week (do not reject the null hypothesis when the null hypothesis is false).

Example 9.3.6.7

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Answer

Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

$H_0 : p = 0.50 \qquad H_a : p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

$$P' - N\left(p, \sqrt{\frac{p-q}{n}}\right)$$

Therefore,

$$P' - N\left(0.5, \sqrt{\frac{0.5-0.5}{100}}\right)$$

where $p = 0.50, q = 1 - p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = P(p' < 0.47 \text{ or } p' > 0.53) = 0.5485$$

where

$$x = 53, p' = \frac{x}{n} = \frac{53}{100} = 0.53$$

.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in Figure).

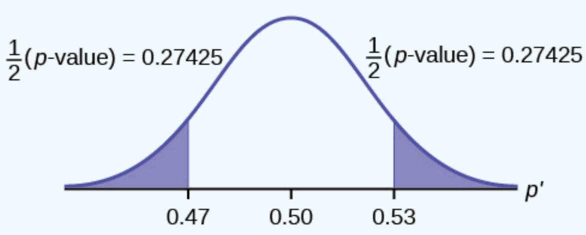


Figure 9.3.6.12

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and $p\text{-value} = 0.5485$. $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated.

Press `STAT` and arrow over to `TESTS`. Press `5:1-PropZTest`. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 9.3.6.7

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Answer

Since the problem is about percentages, this is a test of single population proportions.

- $H_0 : p = 0.85$
- $H_a : p \neq 0.85$
- $p = 0.7554$

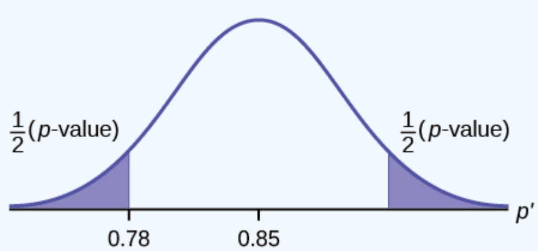


Figure 9.3.6.13

Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example 9.3.6.8

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Answer

Set up the Hypothesis Test:

$H_0: p = 0.30, H_a: p \neq 0.30$

Determine the distribution needed:

The **random variable** is $P' =$ proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' - N\left(0.30, \sqrt{\frac{(0.30 \cdot 0.70)}{150}}\right)$

? Exercise 9.3.6.8.2

a. The value that helps determine the p-value is p'. Calculate p'.

Answer

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43, n = 150$

$p' = 43150$

? Exercise 9.3.6.8.3

b. What is a **success** for this problem?

Answer

b. A success is having three cell phones in a household.

? Exercise 9.3.6.8.4

c. What is the level of significance?

Answer

c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

? Exercise 9.3.6.8.5

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Calculate the p-value.

Answer

d. $p\text{-value} = 0.7216$

? Exercise 9.3.6.8.6

e. Make a decision. _____________(Reject/Do not reject) H_0 because____________.

Answer

e. Assuming that $\alpha = 0.05, \alpha < p$-value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

? Exercise 9.3.6.8

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : p = 0.92$
- $H_a : p < 0.92$
- $p\text{-value} = 0.0046$

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

- Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example $9.3.6.9$

My dog has so many fleas,

They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!
Before his shampoo
I counted 42.

At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out

Use the new shampoo or go without?

Answer

Set up the hypothesis test:

$H_0 : p \leq 0.25 \qquad H_a : p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\bar{X}$ or P' represents.

$P' =$ The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal:

$$N\left(0.25, \sqrt{\frac{(0.25)1-0.25}{42}}\right)$$

Test Statistic: $z = 2.3163$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = 0.0103$$

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

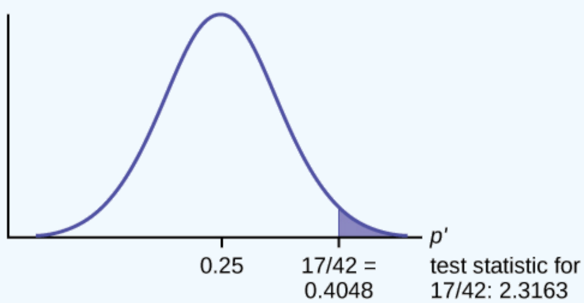


Figure 9.3.6.14

Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha < p$-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

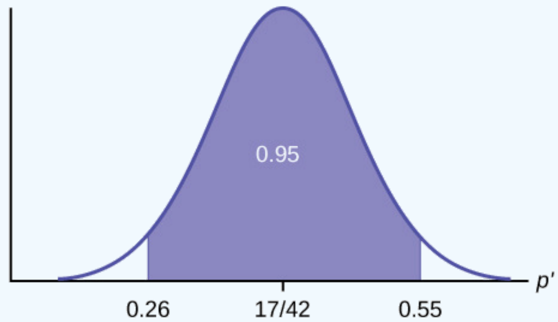


Figure 9.3.6.15

Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion *p* of fleas that are killed by the new shampoo is between 26% and 55%.

This test result is not very definitive since the p-value *is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.*

Example 9.3.6.10

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Answer

Let's follow a four-step process to answer this statistical question.

1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
   a. $H_0 : \mu \leq 1$
   b. $H_a : \mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

alt

Figure 9.3.6.7.

Figure 9.3.6.8.

alt

Figure 9.3.6.9.

Figure 9.3.6.10.

4. **State the Conclusions**: Since the p-value$(p = 0.036)$ is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example 9.3.6.11

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Answer

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
   a. $H_0 : p \leq 0.00034$
   b. $H_a : p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420,019$. The sample is sufficiently large because we have $np = 420,019(0.00034) = 142.8$, $nq = 420,019(0.99966) = 419,876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are

   alt

   Figure $\mathbf{9.3.6.11.}$

   Figure $\mathbf{9.3.6.12.}$
4. Since the p-value $= 0.0073$ is greater than our alpha value $= 0.005$, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example $9.3.6.12$

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Answer

We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion *z*-test. The hypotheses for the test will be
   a. $H_0 : p = 0.00078$
   b. $H_a : p \neq 0.00078$
3. The following screen shots display the summary statistics from the hypothesis test.

   alt

   Figure $\mathbf{9.3.6.13.}$

   Figure $\mathbf{9.3.6.14.}$
4. Since the p-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.

4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

Exercise 9.3.6.8

Assume $H_0: \mu = 9$ and $H_a: \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 9.3.6.9

Assume $H_0: \mu \leq 6$ and $H_a: \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 9.3.6.10

Assume $H_0: p = 0.25$ and $H_a: p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

Exercise 9.3.6.11

Draw the general graph of a left-tailed test.

Exercise 9.3.6.12

Draw the graph of a two-tailed test.

Answer

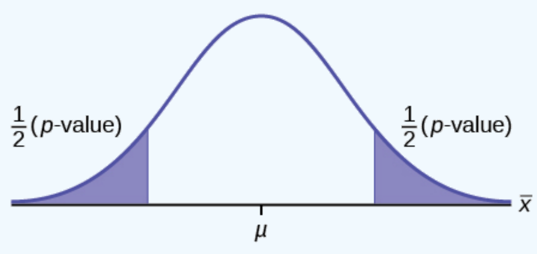


Figure 9.3.6.16

Exercise 9.3.6.13

A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Exercise 9.3.6.14

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

Answer

a right-tailed test

Exercise 9.3.6.15

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

Exercise 9.3.6.16

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

Answer

a left-tailed test

Exercise 9.3.6.17

If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

Exercise 9.3.6.18

Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 9.3.6.19

Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 9.3.6.20

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)
11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. "Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005." Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).
16. "Foothill-De Anza Community College District." De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. "Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark." Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. "How often does sexual assault occur?" RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N(n\mu, \sqrt{n}\sigma)$. If the size *n* of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

9.3.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)

Name: ______________________________

Section: _____________________________

Student ID#:__________________________

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_0: _____________
2. H_a: _____________
3. In words, define the random variable. __________ = ______________________
4. The distribution to use for the test is ______________________.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately.Shade the actual level of significance.
   a. Graph:

Figure 9.7.1.

   b. Determine the $p\text{-value}$.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Language Survey

About 42.3% of Californians and 19.6% of all Americans over age five speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percent of the students at your school who speak a language other than English at home is different from 42.3%.

1. H_0: ___________
2. H_a: ___________
3. In words, define the random variable. __________ = _______________
4. The distribution to use for the test is ________________
5. Determine the test statistic using your data.

6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

Figure 9.7.2.

    b. Determine the p-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey

Suppose that young adults own an average of three pairs of jeans. Survey eight people from your class to determine if the average is higher than three. Assume the population is normal.

1. H_0: ______________
2. H_a: ______________
3. In words, define the random variable. ___________ = ________________________
4. The distribution to use for the test is _________________________.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

Figure 9.7.3.

    b. Determine the p-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

9.3.E: Hypothesis Testing with One Sample (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

9.1: Introduction

9.2: Null and Alternative Hypotheses

Q 9.2.1

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter $(\mu \text{or} p)$.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

S 9.2.1

a. $H_0: \mu = 34; H_a: \mu \neq 34$
b. $H_0: p \leq 0.60; H_a: p > 0.60$
c. $H_0: \mu \geq 100,000; H_a: \mu < 100,000$
d. $H_0: p = 0.29; H_a: p \neq 0.29$
e. $H_0: p = 0.05; H_a: p < 0.05$
f. $H_0: \mu \leq 10; H_a: \mu > 10$
g. $H_0: p = 0.50; H_a: p \neq 0.50$
h. $H_0: \mu = 6; H_a: \mu \neq 6$
i. $H_0: p \geq 0.11; H_a: p < 0.11$
j. $H_0: \mu \leq 20,000; H_a: \mu > 20,000$

Q 9.2.2

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

Q 9.2.3

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

S 9.2.3

c

Q 9.2.4

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. $H_0 : \bar{x} = 4.5, H_a : \bar{x} > 4.5$
b. $H_0 : \mu \geq 4.5, H_a : \mu < 4.5$
c. $H_0 : \mu = 4.75, H_a : \mu > 4.75$
d. $H_0 : \mu = 4.5, H_a : \mu > 4.5$

9.3: Outcomes and the Type I and Type II Errors

Q 9.3.1

State the Type I and Type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

S 9.3.1

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.
f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.

i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
j. Type I error: We conclude that the average tuition cost at private universities is more than \$20,000, though in reality it is at most \$20,000. Type II error: We conclude that the average tuition cost at private universities is at most \$20,000 when, in fact, it is more than \$20,000.

Q 9.3.2

For statements a-j in Exercise 9.109, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

Q 9.3.3

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is "the drug is unsafe." What is the Type II Error?

a. To conclude the drug is safe when in, fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

S 9.3.3

b

Q 9.3.4

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is ________.

a. at least 20%, when in fact, it is less than 20%.
b. 20%, when in fact, it is 20%.
c. less than 20%, when in fact, it is at least 20%.
d. less than 20%, when in fact, it is less than 20%.

Q 9.3.4

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

S 9.3.4

d

Q 9.3.5

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

9.4: Distribution Needed for Hypothesis Testing

Q 9.4.1

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\bar{X} \sim$ ________________

a. $N\left(7.24, \frac{1.93}{\sqrt{22}}\right)$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

S 9.4.1

d

9.5: Rare Events, the Sample, Decision and Conclusion

Q 9.5.1

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0 : ____________________ H_a : ____________________
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:
   i. $x =$ ________________
   ii. $n =$ ________________
   iii. $p' =$ _____________
g. Calculate $\sigma_x =$ __________. Show the formula set-up.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.6: Additional Information and Full Hypothesis Test Examples

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's t-distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

Q 9.6.1.

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using $\alpha = 0.05$, is the data highly inconsistent with the claim?

S 9.6.1

a. $H_0 : \mu \geq 50,000$
b. $H_a : \mu < 50,000$
c. Let $\bar{X} =$ the average lifespan of a brand of tires.
d. normal distribution
e. $z = -2.315$
f. $p\text{-value} = 0.0103$
g. Check student's solution.
h. i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.
i. $(43,537, 49,463)$

Q 9.6.2

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

Q 9.6.3

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

S 9.6.3

a. $H_0 : \mu = \$1.00$
b. $H_a : \mu \neq \$1.00$
c. Let $\bar{X} =$ the average cost of a daily newspaper.
d. normal distribution
e. $z =-0.866$
f. $p\text{-value} = 0.3865$
g. Check student's solution.
h. i. $\alpha : 0.01$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.01.
   iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $1. The mean cost could be $1.
i. $(\$0.84, \$1.06)$

Q 9.6.4

An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

Q 9.6.5

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let $x =$ the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?

S 9.6.5

a. $H_0 : \mu = 10$
b. $H_a : \mu \neq 10$
c. Let $\bar{X}$ the mean number of sick days an employee takes per year.
d. Student's *t*-distribution
e. $t = -1.12$
f. $p\text{-value} = 0.300$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The $p\text{-value}$ is greater than 0.05.
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.
i. $(4.9443, 11.806)$

Q 9.6.6

In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

Q 9.6.7

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

S 9.6.7

a. $H_0 : p \geq 0.6$
b. $H_a : p < 0.6$
c. Let $P' =$ the proportion of students who feel more enriched as a result of taking Elementary Statistics.
d. normal for a single proportion
e. 1.12
f. $p\text{-value} = 0.1308$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The $p\text{-value}$ is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.
i. Confidence Interval: $(0.409, 0.654)$

   The "plus-4s" confidence interval is $(0.411, 0.648)$

Q 9.6.8

A Nissan Motor Corporation advertisement read, "The average man's I.Q. is 107. The average brown trout's I.Q. is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

Q 9.6.9

Refer to Exercise 9.119. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout's mean I.Q. is **not** four.

S 9.6.9

a. $H_0 : \mu = 4$
b. $H_a : \mu \neq 4$
c. Let $\bar{X}$ the average I.Q. of a set of brown trout.
d. two-tailed Student's t-test
e. $t = 1.95$
f. $p\text{-value} = 0.076$
g. Check student's solution.
h. i. $\alpha : 0.05$
  ii. Decision: Reject the null hypothesis.
  iii. Reason for decision: The $p\text{-value}$ is greater than 0.05
  iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.
i. $(3.8865, 5.9468)$

Q 9.6.10

According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

Q 9.6.11

A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

S 9.6.11

a. $H_0 : p \geq 0.13$
b. $H_a : p < 0.13$
c. Let $P' =$ the proportion of Americans who have seen or sensed angels
d. normal for a single proportion
e. –2.688
f. $p\text{-value} = 0.0036$
g. Check student's solution.
h. i. alpha: 0.05
  ii. Decision: Reject the null hypothesis.
  iii. Reason for decision: The p-valuee is less than 0.05.
  iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.
i. $(0, 0.0623)$.

  The"plus-4s" confidence interval is (0.0022, 0.0978)

Q 9.6.12

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Q 9.6.13

Use the “Lap time” data for Lap 4 (see [link]) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

S 9.6.13

a. $H_0 : \mu \geq 129$
b. $H_a : \mu < 129$
c. Let $\bar{X} =$ the average time in seconds that Terri finishes Lap 4.
d. Student's t-distribution
e. $t = 1.209$
f. 0.8792
g. Check student’s solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The $p\text{-value}$ is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that Terri’s mean lap time is less than 129 seconds.
i. $(128.63, 130.37)$

Q 9.6.14

Use the “Initial Public Offering” data (see [link]) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note

The following questions were written by past students. They are excellent problems!

Q 9.6.15

"Asian Family Reunion," by Chau Nguyen

Every two years it comes around.

We all get together from different towns.

In my honest opinion,

It's not a typical family reunion.

Not forty, or fifty, or sixty,

But how about seventy companions!

The kids would play, scream, and shout

One minute they're happy, another they'll pout.

The teenagers would look, stare, and compare

From how they look to what they wear.

The men would chat about their business

That they make more, but never less.

Money is always their subject

And there's always talk of more new projects.

The women get tired from all of the chats

They head to the kitchen to set out the mats.

Some would sit and some would stand

Eating and talking with plates in their hands.

Then come the games and the songs

And suddenly, everyone gets along!

With all that laughter, it's sad to say

That it always ends in the same old way.

They hug and kiss and say "good-bye"

And then they all begin to cry!

I say that 60 percent shed their tears

But my mom counted 35 people this year.

She said that boys and men will always have their pride,

So we won't ever see them cry.

I myself don't think she's correct,

So could you please try this problem to see if you object?

S 9.6.15

a. $H_0 : p = 0.60$
b. $H_a : p < 0.60$
c. Let $P' =$ the proportion of family members who shed tears at a reunion.
d. normal for a single proportion
e. –1.71
f. 0.0438
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $< \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the p-value and alpha are quite close, so other tests should be done.
i. We are 95% confident that between 38.29% and 61.71% of family members will shed tears at a family reunion. $(0.3829, 0.6171)$ The"plus-4s" confidence interval (see chapter 8) is $(0.3861, 0.6139)$

Note that here the "large-sample" $1 - \text{PropZTest}$ provides the approximate p-value of 0.0438. Whenever a p-value based on a normal approximation is close to the level of significance, the exact p-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

Q 9.6.16

"The Problem with Angels," by Cyndy Dowling

Although this problem is wholly mine,

The catalyst came from the magazine, Time.

On the magazine cover I did find

The realm of angels tickling my mind.

Inside, 69% I found to be

In angels, Americans do believe.

Then, it was time to rise to the task,

Ninety-five high school and college students I did ask.

Viewing all as one group,

Random sampling to get the scoop.

So, I asked each to be true,

"Do you believe in angels?" Tell me, do!

Hypothesizing at the start,

Totally believing in my heart

That the proportion who said yes

Would be equal on this test.

Lo and behold, seventy-three did arrive,

Out of the sample of ninety-five.

Now your job has just begun,

Solve this problem and have some fun.

Q 9.6.17

"Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,

I had to find some sane defense.

Some light and lifting simple play

To float my math anxiety away.

Blowing bubbles lifts me high

Takes my troubles to the sky.

POIK! They're gone, with all my stress

Bubble therapy is the best.

The label said each time I blew

The average number of bubbles would be at least 22.

I blew and blew and this I found

From 64 blows, they all are round!

But the number of bubbles in 64 blows

Varied widely, this I know.

20 per blow became the mean

They deviated by 6, and not 16.

From counting bubbles, I sure did relax

But now I give to you your task.

Was 22 a reasonable guess?

Find the answer and pass this test!

S 9.6.17

a. $H_0 : \mu \geq 22$
b. $H_a : \mu < 22$
c. Let $\bar{X} =$ the mean number of bubbles per blow.
d. Student's t-distribution

e. –2.667
f. $p\text{-value} = 0.00486$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The $p\text{-value}$ is less than 0.05.
iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.
i. $(18.501, 21.499)$

Q 9.6.18

"Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles

Bred puppies with numerous freckles

The Dalmatians he sought

Possessed spot upon spot

The more spots, he thought, the more shekels.

His competitors did not agree

That freckles would increase the fee.

They said, "Spots are quite nice

But they don't affect price;

One should breed for improved pedigree."

The breeders decided to prove

This strategy was a wrong move.

Breeding only for spots

Would wreak havoc, they thought.

His theory they want to disprove.

They proposed a contest to Spreckles

Comparing dog prices to freckles.

In records they looked up

One hundred one pups:

Dalmatians that fetched the most shekels.

They asked Mr. Spreckles to name

An average spot count he'd claim

To bring in big bucks.

Said Spreckles, "Well, shucks,

It's for one hundred one that I aim."

Said an amateur statistician

Who wanted to help with this mission.

"Twenty-one for the sample

Standard deviation's ample:

They examined one hundred and one

Dalmatians that fetched a good sum.

They counted each spot,

Mark, freckle and dot

And tallied up every one.

Instead of one hundred one spots

They averaged ninety six dots

Can they muzzle Spreckles'

Obsession with freckles

Based on all the dog data they've got?

Q 9.6.19

"Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:

Price per box of Mac and Cheese:

- 5 stores @ $2.02
- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ 0.75.

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

S 9.6.19

a. $H_0 : \mu \leq 1$
b. $H_a : \mu > 1$
c. Let $\bar{X} =$ the mean cost in dollars of macaroni and cheese in a certain town.
d. Student's t-distribution
e. $t = 0.340$
f. $p\text{-value} = 0.36756$
g. Check student's solution.
h.
    i. $\alpha : 0.05$
    ii. Decision: Do not reject the null hypothesis.
    iii. Reason for decision: The p-value is greater than 0.05
    iv. Conclusion: The mean cost could be $1, or less. At the 5% significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $1.

i. $(0.8291, 1.241)$

Q 9.6.20

"William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi

THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet's tutor
- HOROTIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how cans't thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What piece of work is man, how noble in reason, how infinite in faculties? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

Q 9.6.21

"Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

S 9.6.21

a. $H_0 : p = 0.01$
b. $H_a : p > 0.01$
c. Let $P' =$ the proportion of errors generated
d. Normal for a single proportion
e. 2.13
f. 0.0165
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The $p\text{-value}$ is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.
i. Confidence interval: $(0, 0.094)$.

   The"plus-4s" confidence interval is $(0.004, 0.144)$

Q 9.6.22

"Japanese Girls' Names"

by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

Q 9.6.23

"Phillip's Wish," by Suzanne Osorio

My nephew likes to play

Chasing the girls makes his day.

He asked his mother

If it is okay

To get his ear pierced.

She said, “No way!”

To poke a hole through your ear,

Is not what I want for you, dear.

He argued his point quite well,

Says even my macho pal, Mel,

Has gotten this done.

It’s all just for fun.

C’mon please, mom, please, what the hell.

Again Phillip complained to his mother,

Saying half his friends (including their brothers)

Are piercing their ears

And they have no fears

He wants to be like the others.

She said, “I think it’s much less.

We must do a hypothesis test.

And if you are right,

I won’t put up a fight.

But, if not, then my case will rest.”

We proceeded to call fifty guys

To see whose prediction would fly.

Nineteen of the fifty

Said piercing was nifty

And earrings they’d occasionally buy.

Then there’s the other thirty-one,

Who said they’d never have this done.

So now this poem’s finished.

Will his hopes be diminished,

Or will my nephew have his fun?

S 9.6.23

a. $H_0 : p = 0.50$
b. $H_a : p < 0.50$
c. Let $P' =$ the proportion of friends that has a pierced ear.
d. normal for a single proportion
e. –1.70
f. $p\text{-value} = 0.0448$
g. Check student’s solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The $p\text{-value}$ is less than 0.05. (However, they are very close.)
   iv. Conclusion: There is sufficient evidence to support the claim that less than 50% of his friends have pierced ears.
i. Confidence Interval: $(0.245, 0.515)$ The “plus-4s” confidence interval is $(0.259, 0.519)$

Q 9.6.24

"The Craven," by Mark Salangsang

Once upon a morning dreary

In stats class I was weak and weary.

Pondering over last night's homework

Whose answers were now on the board

This I did and nothing more.

While I nodded nearly napping

Suddenly, there came a tapping.

As someone gently rapping,

Rapping my head as I snore.

Quoth the teacher, "Sleep no more."

"In every class you fall asleep,"

The teacher said, his voice was deep.

"So a tally I've begun to keep

Of every class you nap and snore.

The percentage being forty-four."

"My dear teacher I must confess,

While sleeping is what I do best.

The percentage, I think, must be less,

A percentage less than forty-four."

This I said and nothing more.

"We'll see," he said and walked away,

And fifty classes from that day

He counted till the month of May

The classes in which I napped and snored.

The number he found was twenty-four.

At a significance level of 0.05,

Please tell me am I still alive?

Or did my grade just take a dive

Plunging down beneath the floor?

Upon thee I hereby implore.

Q 9.6.25

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

S 9.6.25

a. $H_0 : p = 0.40$
b. $H_a : p < 0.40$

c. Let P' = the proportion of schoolmates who fear public speaking.
d. normal for a single proportion
e. –1.01
f. p-value $= 0.1563$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.
i. Confidence Interval: $(0.3241, 0.4240)$ The "plus-4s" confidence interval is $(0.3257, 0.4250)$

Q 9.6.26

Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

Q 9.6.27

According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

S 9.6.27

a. $H_0 : p = 0.14$
b. $H_a : p < 0.14$
c. Let P' = the proportion of NYC residents that smoke.
d. normal for a single proportion
e. –0.2756
f. p-value $= 0.3914$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.
i. Confidence Interval: $(0.0502, 0.2070)$ The "plus-4s" confidence interval (see chapter 8) is $(0.0676, 0.2297)$

Q 9.6.28

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

Q 9.6.29

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

S 9.6.29

a. $H_0 : \mu = 69,110$
b. $H_0 : \mu > 69,110$
c. Let $\bar{X}$ = the mean salary in dollars for California registered nurses.

d. Student's t-distribution
e. $t = 1.719$
f. p-value : 0.0466
g. Check student’s solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.
i. $(\$68,757, \$73,485)$

Q 9.6.30

La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

Q 9.6.31

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?

After conducting the test, your decision and conclusion are

a. Reject H_0: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject H_0: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject H_0: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
d. Reject H_0: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

S 9.6.31

c

Q 9.6.32

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.

At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

Q 9.6.33

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

S 9.6.33

c

Instructions: For the following ten exercises,

Hypothesis testing: For the following ten exercises, answer each question.

State the null and alternate hypothesis.

State the p-value.

State α.

What is your decision?

Write a conclusion.

Answer any other questions asked in the problem.

Q 9.6.34

According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

Q 9.6.35

A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

S 9.6.35

a. $H_0: p = 0.488$ $H_a: p \neq 0.488$
b. $p\text{-value} = 0.0114$
c. $\alpha = 0.05$
d. Reject the null hypothesis.
e. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.
f. The survey does not appear to be accurate.

Q 9.6.36

Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

Q 9.6.37

The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

S 9.6.37

a. $H_0: p = 0.517$ $H_0: p \neq 0.517$
b. $p\text{-value} = 0.9203.$
c. $\alpha = 0.05.$
d. Do not reject the null hypothesis.

e. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

Q 9.6.38

For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

Q 9.6.39

The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05 level$, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

S 9.6.39

a. $H_0 : \mu \geq 11.52$ $H_a : \mu < 11.52$
b. $p\text{-value} = 0.000002$which is almost 0.
c. $\alpha = 0.05$.
d. Reject the null hypothesis.
e. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1% because the $p\text{-value}$ is almost 0.

Q 9.6.40

A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

Q 9.6.41

A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3; 2; 1; 3; 7; 2; 9; 4; 6; 6; 8; 0; 5; 6; 4; 2; 1; 3; 4; 1

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?

S 9.6.42

1. $H_0 : \mu \leq 5.8$ $H_a : \mu > 5.8$
2. $p\text{-value} = 0.9987$
3. $\alpha = 0.05$
4. Do not reject the null hypothesis.
5. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

Q 9.6.42

According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:

5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2

At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

Q 9.6.43

The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At α = 0.01 level, is the student academic group's claim correct?

S 9.6.43

a. $H_0 : \mu \geq 150$ $H_0 : \mu < 150$
b. p-value $= 0.0622$
c. $\alpha = 0.01$
d. Do not reject the null hypothesis.
e. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

9.7: Hypothesis Testing of a Single Mean and Single Proportion

9.4: PowerPoints

https://professormo.com/holistic/Powerpoint/ch09.pptx

CHAPTER OVERVIEW

10: Hypothesis Testing about Two Population Means and Proportions

SECTION OVERVIEW

10.1: Inference for Categorical Data

Chapter 6 introduces inference in the setting of categorical data. We use these methods to answer questions like the following:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

We will find that the methods we learned in previous chapters are very useful in these settings. For example, sample proportions are well characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual confidence interval and hypothesis testing tools. In other instances, such as those with contingency tables or when sample size conditions are not met, we will use a different distribution, though the core ideas remain the same.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

10.1.1: Inference for a Single Proportion

According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[1] This poll included responses of 976 adults.

Identifying when the Sample Proportion is Nearly Normal

A sample proportion can be described as a sample mean. If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0+1+1+\cdots+0}{976} = 0.44 \tag{10.1.1.1}$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes and failures, typically at least 10 of each.

[1]*nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html*

Conditions for the sampling distribution of $\hat{p}$ being nearly normal

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{10.1.1.2}$$

Typically we do not know the true proportion, p, so must substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value - that is, the proportion claimed in the null hypothesis - is used in place of p. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

TIP: Reminder on checking independence of observations

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Alternatively, if the data come from a random process, we must evaluate the independence condition more carefully.

Confidence Intervals for a Proportion

We may want a confidence interval for the proportion of Americans who approve of the job the Supreme Court is doing. Our point estimate, based on a sample of size n = 976 from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use the general confidence interval formula from Section 4.5, we must check the conditions to ensure that the sampling distribution of $\hat{p}$ is nearly normal. We also must determine the standard error of the estimate.

The data are based on a simple random sample and consist of far fewer than 10% of the U.S. population, so independence is confirmed. The sample size must also be sufficiently large, which is checked via the success-failure condition: there were approximately $976 \times \hat{p} = 429$ "successes" and $976 \times (1-\hat{p}) = 547$ "failures" in the sample, both easily greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

Exercise 10.1.1.1

Estimate the standard error of $\hat{p} = 0.44$ using Equation 10.1.1.2. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p.

Answer

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$$

Example 10.1.1.1

Construct a 95% confidence interval for p, the proportion of Americans who trust federal officials most of the time.

Solution

Using the standard error estimate from Exercise 10.1.1.1, the point estimate 0.44, and z* = 1:96 for a 95% confidence interval, the confidence interval may be computed as

$$\text{point estimate} \pm z^{*} SE \rightarrow 0.44 \pm 1.96 \times 0.016 \rightarrow (0.409, 0.471) \tag{10.1.1.3}$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. If the proportion has not changed since this poll, than we can say with high confidence that the job approval of the Supreme Court is below 50%.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, the sampling distribution of $\hat{p}$ may be well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

Hypothesis Testing for a Proportion

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1-p_0)$ are at least 10, where p_0 is the null value.

Exercise 10.1.1.2

Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Set up a onesided hypothesis test to evaluate this claim.

Answer

Is there convincing evidence that the campaign manager is correct?

- H_0 : p = 0.50,
- H_A : p > 0.50.

Example 10.1.1.2

A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%. Does this provide convincing evidence for the claim of Toohey's manager at the 5% significance level?

Solution

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion,

$$p_0 = 0.5 : np_0 = n(1-p_0) = 500 \times 0.5 = 250 > 10. \tag{10.1.1.4}$$

With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion.

$$\begin{aligned} SE &= \sqrt{\frac{p_0 \times (1-p_0)}{n}} \\ &= \sqrt{\frac{0.5(1-0.5)}{500}} = 0.022 \end{aligned}$$

A picture of the normal model is shown in Figure 10.1.1.1 with the p-value represented by the shaded region. Based on the normal model, the test statistic can be computed as the Z score of the point estimate:

$$\begin{aligned} Z &= \frac{\text{point estimate - null value}}{SE} \\ &= \frac{0.52-0.50}{0.022} = 0.89 \end{aligned}$$

The upper tail area, representing the p-value, is 0.1867. Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

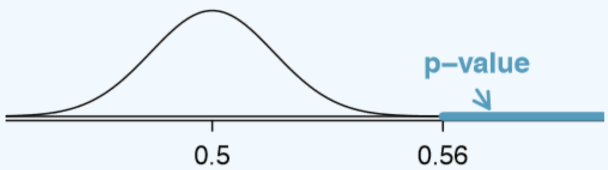


Figure 10.1.1.1: Sampling distribution of the sample proportion if the null hypothesis is true for Example 10.1.1.2. The p-value for the test is shaded.

Hypothesis test for a proportion

Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

Choosing a sample size when estimating a proportion

We first encountered sample size computations in Section 4.6, which considered the case of estimating a single mean. We found that these computations were helpful in planning a study to control the size of the standard error of a point estimate. The task was to find a sample size n so that the sample mean would be within some margin of error m of the actual mean with a certain level of confidence. For example, the margin of error for a point estimate using 95% confidence can be written as $1.96 \times SE$. We set up a general equation to represent the problem:

$$ME = z^* SE \leq m \tag{10.1.1.5}$$

where ME represented the actual margin of error and z^* was chosen to correspond to the confidence level. The standard error formula is specified to correspond to the particular setting. For instance, in the case of means, the standard error was given as $\frac{\sigma}{\sqrt{n}}$.

In the case of a single proportion, we use $\sqrt{p(1-p)}\,n$ for the standard error.

Planning a sample size before collecting data is equally important when estimating a proportion. For instance, if we are conducting a university survey to determine whether students support a $200 per year increase in fees to pay for a new football stadium, how big of a sample is needed to be sure the margin of error is less than 0.04 using a 95% confidence level?

Example 10.1.1.3

Find the smallest sample size n so that the margin of error of the point estimate $\hat{p}$ will be no larger than $m = 0.04$ when using a 95% confidence interval.

Solution

For a 95% confidence level, the value z* corresponds to 1.96, and we can write the margin of error expression as follows:

$$ME = z^{*}SE = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04 \tag{10.1.1.6}$$

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could use that value. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this worst case estimate if no other estimate is available:

$$1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.04 \tag{10.1.1.7}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{n} \leq 0.04^2 \tag{10.1.1.8}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} \leq n \tag{10.1.1.9}$$

$$600.25 \leq n \tag{10.1.1.10}$$

We would need at least 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

No estimate of the true proportion is required in sample size computations for a proportion, whereas an estimate of the standard deviation is always needed when computing a sample size for a margin of error for the sample mean. However, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

Example 10.1.1.4

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.

a. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.
b. The sample sizes in (b) vary widely. Which of the three would you suggest using? What would influence your choice?

Solution

(a) For the 1.7% estimate of p, we estimate the appropriate sample size as follows:

$$1.65 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \leq 0.02 \rightarrow n \geq 113.7$$

Using the estimate from the first model, we would suggest examining 114 tires (round up!). A similar computation can be accomplished using 0.062 and 0.013 for p: 396 and 88.

(b) We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. (Answers will vary.)

Exercise 10.1.1.4

A recent estimate of Congress' approval rating was 17%.5 What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?

Answer

We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \leq 0.04 \rightarrow n \geq 338.8$$

A sample size of 339 or more would be reasonable.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

10.1.2: Difference of Two Proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, a company weighs whether they should switch to a higher quality parts manufacturer. In the last example, we examine the cancer risk to dogs from the use of yard herbicides.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

Sample Distribution of the Difference of Two Proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when each proportion separately follows a **normal model**, and the samples are **independent**. The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} \tag{10.1.2.1}$$

$$= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{10.1.2.2}$$

where p_1 and p_2 represent the population proportions, and n1 and n2 represent the sample sizes.

For the difference in two means, the standard error formula took the following form:

$$SE_{\hat{x}_1 - \hat{x}_2} = \sqrt{SE^2_{\hat{x}_1} + SE^2_{\hat{x}_2}} \tag{10.1.2.3}$$

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Exercise 5.14.

[5]*www.gallup.com/poll/155144/Congress-Approval-June.aspx*

Table 10.1.2.1: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

	Sample size (ni)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Intervals and tests for $p_1 - p_2$

In the setting of confidence intervals, the sample proportions are used to verify the success/failure condition and also compute standard error, just as was the case with a single proportion.

Example 10.1.2.1

The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[7]

> *As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?*

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed. Table 6.2 shows the results of this experiment. Create and interpret a 90% confidence interval of the difference in approval.

Solution

First the conditions must be verified. Because each group is a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample. Because all conditions are met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13 \tag{10.1.2.4}$$

The standard error may be computed from Equation 10.1.2.2using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025 \tag{10.1.2.5}$$

For a 90% con dence interval, we use z* = 1.65:

$$\text{point estimate} \pm z^* SE \approx 0.13 \pm 1.65 \times 0.025 \rightarrow (0.09, 0.17) \tag{10.1.2.6}$$

We are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still uid on the health insurance mandate.

[7] *www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/.*

Sample sizes for each polling group are approximate.

Exercise 10.1.2.1

A remote control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[8]

Answer

Add texts here. Do not delete this text first.

Example 10.1.2.2

The quality control engineer from Exercise 6.11 collects a sample of gears, examining 1000 gears from each company and nds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Exercise 6.11 using a signi cance level of 5%.

Solution

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the gears are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as

to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, 0.958 - 0.899 = 0.059, can be said to come from a nearly normal distribution.

The standard error can be found using Equation 10.1.2.2:

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114 \tag{10.1.2.7}$$

In this hypothesis test, the sample proportions were used. We will discuss this choice more in Section 6.2.3.

Next, we compute the test statistic and use it to nd the p-value, which is depicted in Figure $10.1.2.1$.

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.059-0.03}{0.0114} = 2.54 \tag{10.1.2.8}$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher quality gears actually do pass inspection more than 3% as often as the currently used gears. Based on these results, management will approve the switch to the new supplier.

[8] H_0: The higher quality gears will pass inspection no more than 3% more frequently than the standard quality gears. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher quality gears will pass inspection more than 3% more often than the standard quality gears. $p_{highQ} - p_{standard} > 0.03$.

Figure* $10.1.2.1$*: Distribution of the test statistic if the null hypothesis was true.

The p-value is represented by the shaded area.

Hypothesis testing when H_0: $p_1 = p_2$

Here we use a new example to examine a special estimate of standard error when $H_0: p_1 = p_2$. We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.9 Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table $10.1.2.2$

Table 10.1.2.2: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

	cancer	no cancer
2,4 - D	191	304
no 2,4 - D	300	641

Exercise 10.1.2.1

Is this study an experiment or an observational study?

Answer

The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the control group, which is a term usually seen in experiments.

Exercise 10.1.2.1

Exercise 6.14 Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups.[11]

[9] *Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. CaseControl Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.*

Answer

Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.

- H_0: the proportion of dogs with exposure to 2,4-D is the same in "cancer" and \no cancer" dogs, $p_c - p_n = 0$.
- H_A: dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

Example 10.1.2.1: pooled estimate

First are the conditions met to use the normal model and make inference on the results?

(1) It is unclear whether this is a random sample. However, if we believe the dogs in both the cancer and no cancer groups are representative of each respective population and that the dogs in the study do not interact in any way, then we may find it reasonable to assume independence between observations. (2) The success-failure condition holds for each sample.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

In your hypotheses for Exercise $10.1.2.1$, the null is that the proportion of dogs with exposure to 2,4-D is the same in each group. The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. To identify the p-value for this test, we first check conditions (Example 6.15) and compute the standard error of the difference:

$$SE = \sqrt{\frac{p_c(1-p_c)}{n_c} + \frac{p_n(1-p_n)}{n_n}} \tag{10.1.2.9}$$

In a hypothesis test, the distribution of the test statistic is always examined as though the null hypothesis is true, i.e. in this case, $p_c = p_n$. The standard error formula should reflect this equality in the null hypothesis. We will use p to represent the common rate of dogs that are exposed to 2,4-D in the two groups:

$$SE = \sqrt{\frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_n}} \tag{10.1.2.10}$$

We don't know the exposure rate, p, but we can obtain a good estimate of it by pooling the results of both samples:

$$\hat{p} = \frac{\#\text{ of "successes"}}{\#\text{ of cases}} = \frac{191+304}{191+300+304+641} = 0.345 \tag{10.1.2.11}$$

This is called the **pooled estimate** of the sample proportion, and we use it to compute the standard error when the null hypothesis is that $p_1 = p_2$ (e.g. $p_c = p_n$ or $p_c - p_n = 0$) . We also typically use it to verify the success-failure condition.

Pooled estimate of a proportion

When the null hypothesis is $p_1 = p_2$, it is useful to nd the pooled estimate of the shared proportion:

$$\hat{p} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} \tag{10.1.2.12}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1} \tag{10.1.2.13}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

: $p_1 = p_2$

When the null hypothesis suggests the proportions are equal, we use the pooled proportion estimate $(\hat{p})$ to verify the success-failure condition and also to estimate the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_c} + \frac{\hat{p}(1-\hat{p})}{n_n}} \tag{10.1.2.14}$$

Exercise 10.1.2.1

Using Equation 10.1.2.14, $\hat{p} = 0.345, n_1 = 491$, and $n_2 = 945$, verify the estimate for the standard error is SE = 0.026. Next, complete the hypothesis test using a significance level of 0.05. Be certain to draw a picture, compute the p-value, and state your conclusion in both statistical language and plain language.

Answer

Compute the test statistic:

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

10.1.3: Testing for Goodness of Fit using Chi-Square (Special Topic)

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test. In the rst case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Table 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reect the population of eligible jurors, i.e. registered voters.

Table 6.5: Representation by race in a city's juries and population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

[12]*Compute the test statistic:*

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58 \tag{10.1.3.1}$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

A second application, assessing the t of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for the years 1990-2011 are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

Creating a test statistic for one-way tables

Example 10.1.3.1:

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

Solution

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

Exercise 10.1.3.1

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race?

Answer

Answers can be found in Table 6.6.

Table 6.6: Actual and expected make-up of the jurors.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected count	198	19.25	33	24.75	275

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

- H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.
- H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate - null value}}{\text{SE of point estimate}} \tag{10.1.3.2}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to rst compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count - null white count}}{\text{SE of observed white count}} \tag{10.1.3.3}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205-198}{\sqrt{198}} = 0.50 \tag{10.1.3.4}$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

$$\begin{array}{ccc} Black & Hispanic & Other \\ Z_2 = \frac{26-19.25}{\sqrt{19.25}} = 154 & Z_3 = \frac{25-33}{\sqrt{33}} = -1.39 & Z_4 = \frac{19-24:75}{\sqrt{24.75}} = -1.16 \end{array} \tag{10.1.3.5}$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they - as a group - tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58 \tag{10.1.3.6}$$

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89 \tag{10.1.3.7}$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual - e.g. a standardized difference of 2.5 - will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{(\text{observed count\_1 - null count\_1})^2}{\text{null count\_1}} + \cdots + \frac{(\text{observed count\_4 - null count\_4})^2}{\text{null count\_4}} \tag{10.1.3.8}$$

[13]Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown - though not here - that the square root of the count is a better way to standardize the count differences.

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a chi-square distribution. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

The chi-square distribution and finding areas

The chi-square distribution is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters - mean and standard deviation - that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df),** which inuences the shape, center, and spread of the distribution.

Exercise 10.1.3.1

Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

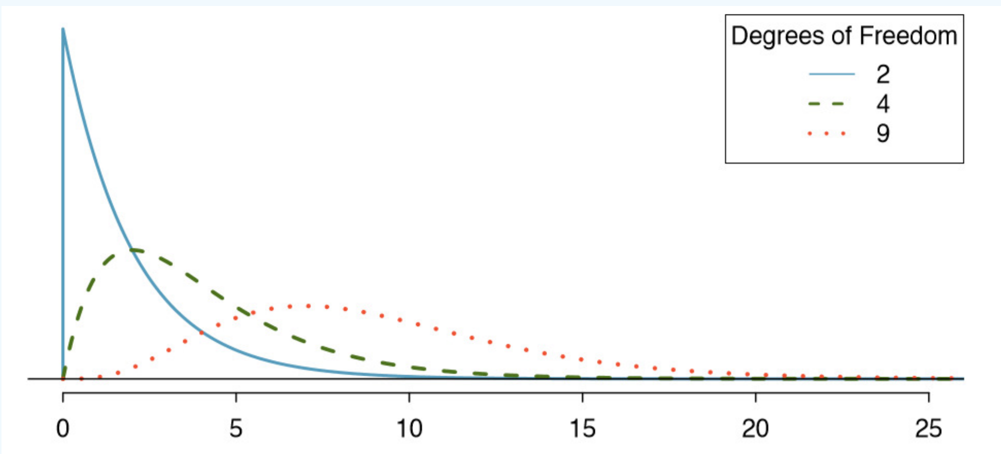


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

Answer

[14](a) The center becomes larger. If we look carefully, we can see that the center of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for df = 2, and then the distributions become more symmetric for the larger degrees of freedom df = 4 and df = 9. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Figure 6.7 and Exercise 6.20 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the **chi-square table**, partially shown in Table 6.8. A more complete table is presented in Appendix B.3 on page 412. This table is very similar to the t table from Sections 5.3 and 5.4: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t table is that the chi-square table only provides upper tail values.

Table 6.8: A section of the chi-square table. A complete table is in Appendix B.3 on page 412.

Upper tail	0.3 0.2 0.1 0.05	0.02 0.01 0.005 0.001
df 1 2 3 4 5	2.41 3.22 4.61 5.99 3.66 4.64 6.25 7.81 4.88 5.99 7.78 9.49 6.06 7.29 9.24 11.07	7.82 9.21 10.60 13.82 9.84 11.34 12.84 16.27 11.67 13.28 14.86 18.47 13.39 15.09 16.75 20.52
6 7	7.23 8.56 10.64 12.59 8.38 9.80 12.02 14.07	15.03 16.81 18.55 22.46 16.62 18.48 20.28 24.32

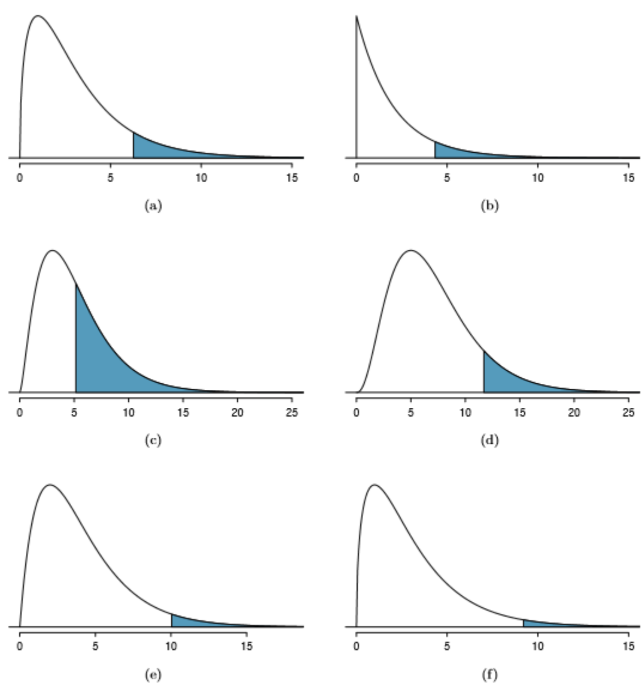


Figure 6.9: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded. (e) 4 degrees of freedom, area above 10 shaded. (f) 3 degrees of freedom, area above 9.21 shaded.

Example 6.21 Figure 6.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 6.8 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value { 6.25 { falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 6.9(a) has area 0.1.

Example 6.22 We rarely observe the exact value in the table. For instance, Figure 6.9(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3, which does not fall in Table 6.8. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 6.9(b) is between 0.1 and 0.2.

Example 6.23 Figure 6.9(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is greater than 0.3.

Exercise 6.24 Figure 6.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

Exercise 6.25 Figure 6.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Exercise 6.26 Figure 6.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

[15]*The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.*

[16]*The area is between 0.02 and 0.05.*

[17]*Between 0.02 and 0.05.*

Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with k - 1 degrees of freedom, where k is the number of bins.

Example 10.1.3.1:

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

Solution

In the jurors example, there were k = 4 categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with k - 1 = 3 degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

Example 10.1.3.1:

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

The chi-square distribution and p-value are shown in Figure 6.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix B.3 or the short table on page 277, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

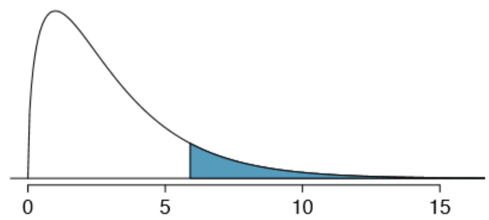


Figure 6.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with df = 3.

Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts $O_1, O_2, \ldots, O_k$ in k categories are unusually different from what might be expected under a null hypothesis. Call the expected counts that are based on the null hypothesis $E_1, E_2, \ldots, E_k$. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with k - 1 degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \tag{10.1.3.9}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

Conditions for the chi-square test

There are three conditions that must be checked before performing a chi-square test:

- **Independence**. Each case that contributes a count to the table must be independent of all the other cases in the table.
- **Sample size / distribution**. Each particular scenario (i.e. cell count) must have at least 5 expected cases.
- **Degrees of freedom** We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one proportion methods introduced in Section 6.1.

Evaluating goodness of fit for a distribution

Section 3.3 would be useful background reading for this example, but it is not a prerequisite. We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model ts a data set. Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as Up or Down (D) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each Up day:

$$\begin{array}{ccccccccccc} \text{Change in price} & 2.52 & -1.46 & 0.51 & -4.07 & 3.36 & 1.10 & -5.46 & -1.03 & -2.99 & 1.71 \\ \text{Outcome} & Up & D & Up & D & Up & Up & D & D & D & Up \\ \text{Days to Up} & 1 & - & 2 & - & 2 & 1 & - & - & - & 4 \end{array} \tag{10.1.3.10}$$

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the kth trial to observe the rst success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next Up trading day, and two more for the third Up day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Table 6.11 shows the number of waiting days for a positive trading day during 1990-2011for the S&P500.

Table 6.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

We consider how many days one must wait until observing an Up day on the S&P500 stock exchange. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a geometric distribution. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an Up day is observed. Under this hypothesis, the number of days until an Up day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an Up day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 from 1990 to 2011 and summarize the waiting times in Table 6.12 and Figure 6.13. The S&P500 was positive on 53.2% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have binned together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the Observed row in Table 6.12, can be compared to the expected counts from the Geometric Model row. The method for computing expected counts is discussed in Table 6.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each

Table 6.12: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model $(P(D) = (1-0.532)^{D-1}(0.532))$ and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 =$ 11.65% of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

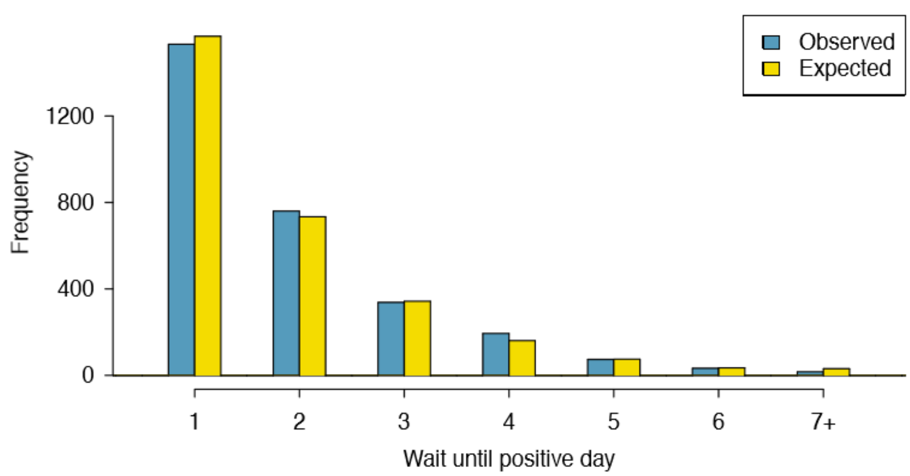


Figure 6.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Example 6.29 Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Table 6.12.

Exercise 6.30 Table 6.12 provides a set of count data for waiting times ($O_1 = 1532, O_2 = 760, \ldots$) and expected counts under the geometric distribution ($E_1 = 1569, E_2 = 734, \ldots$). Compute the chi-square test statistic, X^2.[18]

Exercise 6.31 Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

Example 6.32 If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with df = 6. Using this information, compute a p-value.

Figure 6.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix B.3, we nd that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that

[18] $X^2 = \frac{(1532 - 1569)^2}{1569} + \frac{(760 - 734)^2}{734} + \cdots + \frac{(17 - 31)^2}{31} = 15.08$

[19] *There are k = 7 groups, so we use df = k - 1 = 6.*

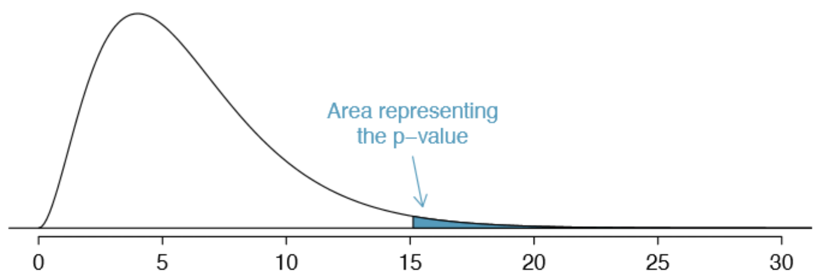


Figure 6.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

Example 6.33 In Example 6.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

10.1.4: Testing for Independence in Two-Way Tables (Special Topic)

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Table 6.15 shows an example of 10,000 queries split into three algorithm groups.[20] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Table 6.15: Google experiment breakdown of test subjects into three search groups.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

[20]Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

Example 10.1.4.1

What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

- H_0: The algorithms each perform equally well.
- H_A: The algorithms do not perform equally well.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. Under scenario (1), we might think that the user was satis ed with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Table 6.16 provides the results from the experiment. These data are very similar to the count data in Section 6.3. However, now the different combinations of two variables are binned in a two-way table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

Table 6.16: Results of the Google search algorithm experiment.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

What is so different about one-way tables and two-way tables?

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for combinations of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the search variable is independent of the algorithm variable.

Expected Counts in Two-way Tables

Example 6.35 From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as $\frac{7078}{10000} = 0.7078$. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satis ed with the initial search:

$$0.7078 \times 5000 = 3539 users \tag{10.1.4.1}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

Exercise 10.1.4.1

Exercise 6.36 Using the same rationale described in Example 6.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

[21]*We would expect 0.7078 * 2500 = 1769.5. It is okay that this is a fraction.*

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 6.35 and Exercise 6.36. These expected counts were used to construct Table 6.17, which is the same as Table 6.16, except now the expected counts have been added in parentheses.

Table 6.17: The observed counts and the (Expected counts)

Search algorithm	current	test 1	test 2	Total
No new search	3511 (3539)	1749 (1769.5)	1818 (1769.5)	7078
New search	1489 (1461)	751 (730.5)	682 (730.5)	2922
Total	5000	2500	2500	10000

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the rst row:

$$0.7078 \times (\text{column 1 total}) = 3539 \tag{10.1.4.2}$$

$$0.7078 \times (\text{column 2 total}) = 1769.5 \tag{10.1.4.3}$$

$$0.7078 \times (\text{column 3 total}) = 1769.5 \tag{10.1.4.4}$$

Looking back to how the fraction 0.7078 was computed - as the fraction of users who did not perform a new search ($\frac{7078}{10000}$) - these three expected counts could have been computed as

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 1 total}) = 3539 \tag{10.1.4.5}$$

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 2 total}) = 1769.5 \tag{10.1.4.6}$$

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 3 total}) = 1769.5 \tag{10.1.4.7}$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

Computing expected counts in a two-way table

To identify the expected count for the ith row and jth column, compute

$$\text{Expected Count}_{\text{row i, col j}} = \frac{(\text{row i total}) \times (\text{column j total})}{\text{table total}} \tag{10.1.4.8}$$

The chi-square Test for Two-way Tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

$$\text{General formula} \frac{(\text{observed count - expected count})^{2}}{\text{expected count}} \tag{10.1.4.9}$$

$$\text{Row 1, Col 1} \frac{(3511-3539)^{2}}{3539}=0.222 \tag{10.1.4.10}$$

$$\text{Row 1, Col 2} \frac{(1749-1769.5)^{2}}{1769.5}=0.237 \tag{10.1.4.11}$$

$$\vdots \tag{10.1.4.12}$$

$$\text{Row 2, Col 3} \frac{(682-730.5)^{2}}{730.5}=3.220 \tag{10.1.4.13}$$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^{2}=0.222+0.237+\cdots+3.220=6.120 \tag{10.1.4.14}$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df=(\text{number of rows minus } 1)\times(\text{number of columns minus } 1) \tag{10.1.4.15}$$

In our example, the degrees of freedom parameter is

$$df=(2-1)\times(3-1)=2 \tag{10.1.4.16}$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic X^2 = 6.12 closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.18.

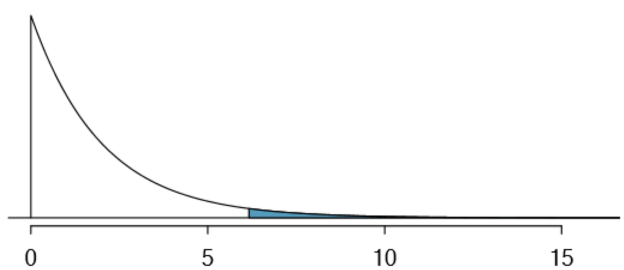


Figure 6.18: Computing the p-value for the Google hypothesis test.

Definition: degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df=(R-1)\times(C-1) \tag{10.1.4.17}$$

where R is the number of rows in the table and C is the number of columns.

[22] *Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.*

Table 6.19: Pew Research poll results of a March 2012 poll.

		Congress		
	Obama	Democrats	Republicans	Total

Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

TIP: Use two-proportion methods for 2-by-2 contingency tables

When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 6.2.

Example 10.1.4.1

Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Solution

Looking in Appendix B.3 on page 412, we examine the row corresponding to 2 degrees of freedom. The test statistic, X^2 = 6.120, falls between the fourth and fth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of α = 0.05 and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Example 10.1.4.1

Table 6.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

Solution

- H_0: There is no difference in approval ratings between the three groups.
- H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23]See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 6.19 are approximate.

Exercise 10.1.4.1

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.38. As a rst step, compute the expected values for each of the six table cells.[24]

[24]The expected count for row one / column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119\times1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104\times1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

Exercise 10.1.4.1

Compute the chi-square test statistic.[25]

[25]For each cell, compute $\frac{(\text{obs - exp})^2}{exp}$. For instance, the rst row and rst column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \cdots + 34.0 = 106.4$.

Exercise 10.1.4.1

Because there are 2 rows and 3 columns, the degrees of freedom for the test is df = (2 - 1) (3 - 1) = 2. Use X^2 = 106.4, df = 2, and the chi-square table on page 412 to evaluate whether to reject the null hypothesis.[26]

[26]The test statistic is larger than the right-most column of the df = 2 row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans'

approval has differences among Democrats in Congress, Republicans in Congress, and the president.

10.1.5: Small Sample Hypothesis Testing for a Proportion (Special Topic)

In this section we develop inferential methods for a single proportion that are appropriate when the sample size is too small to apply the normal model to $\hat{p}$. Just like the methods related to the t distribution, these methods can also be applied to large samples.

When the Success-Failure Condition is Not Met

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspect of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients. One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Exercise 10.1.5.1

Exercise 6.42

We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.

Solution

The sample proportion: $\hat{p} = \frac{3}{62} = 0.048$

Example 10.1.5.1

Is it possible to assess the consultant's claim using the data provided?

Solution

No. The claim is that there is a causal connection, but the data are observational. Patients who hire this medical consultant may have lower complication rates for other reasons.

While it is not possible to assess this causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Exercise 10.1.5.1

Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.

Solution

- H_0: There is no association between the consultant's contributions and the clients' complication rate. In statistical language, p = 0.10.
- H_A: Patients who work with the consultant tend to have a complication rate lower than 10%, i.e. p < 0.10.

Example 10.1.5.1

In the examples based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

Solution

The independence assumption may be reasonable if each of the surgeries is from a different surgical team. However, the success-failure condition is not satis ed. Under the null hypothesis, we would anticipate seeing $62 \times 0.10 = 6.2$ complications, not the 10 required for the normal approximation.

The uncertainty associated with the sample proportion should not be modeled using the normal distribution. However, we would still like to assess the hypotheses from Exercise 6.44 in absence of the normal framework. To do so, we need to evaluate the possibility of a sample value (^p) this far below the null value, $p_0 = 0.10$. This possibility is usually measured with a p-value.

The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis is true. Supposing the null hypothesis is true, we can compute the p-value by identifying the chance of observing a test statistic that favors the alternative hypothesis at least as strongly as the observed test statistic. This can be done using simulation.

Generating the null distribution and p-value by simulation

We want to identify the sampling distribution of the test statistic ($\hat{p}$) if the null hypothesis was true. In other words, we want to see how the sample proportion changes due to chance alone. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.

Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients. If this was the case, we could simulate 62 clients to get a sample proportion for the complication rate from the null distribution.

Each client can be simulated using a deck of cards. Take one red card, nine black cards, and mix them up. Then drawing a card is one way of simulating the chance a patient has a complication if the true complication rate is 10% for the data. If we do this 62 times and compute the proportion of patients with complications in the simulation, $\hat{p}_{sim}$, then this sample proportion is exactly a sample from the null distribution.

An undergraduate student was paid $2 to complete this simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e. $\hat{p}_{sim} = \frac{5}{62} = 0.081$.

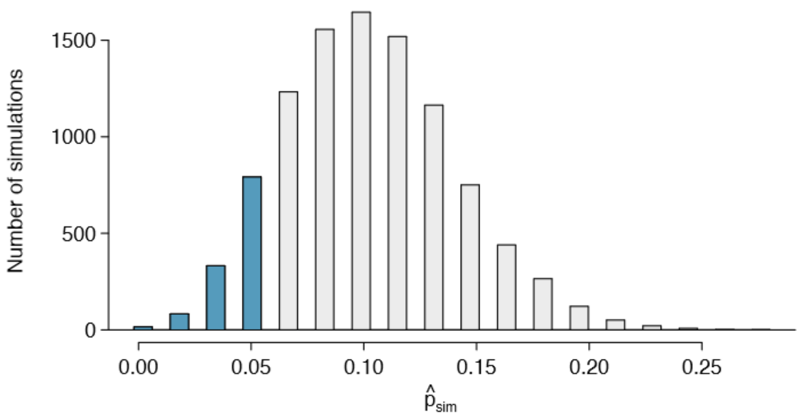


Figure 6.20: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

Example 10.1.5.1

Is this one simulation enough to determine whether or not we should reject the null hypothesis from Exercise 6.44? Explain.

Solution

No. To assess the hypotheses, we need to see a distribution of many $\hat{p}_{sim}$, not just a single draw from this sampling distribution.

One simulation isn't enough to get a sense of the null distribution; many simulation studies are needed. Roughly 10,000 seems sufficient. However, paying someone to simulate 10,000 studies by hand is a waste of time and money. Instead, simulations are typically programmed into a computer, which is much more efficient.

Figure 6.20 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and nd the p-value:

$$\text{left tail} = \frac{\text{Number of observed simulations with} \hat{p}_{sim} \leq 0.048}{10000} \tag{6.47}$$

Of the 10,000 simulated $\hat{p}_{sim}$, 1222 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.1222.

Exercise 10.1.5.1

Because the estimated p-value is 0.1222, which is larger than the signi cance level 0.05, we do not reject the null hypothesis. Explain what this means in plain language in the context of the problem.

Solution

There isn't sufficiently strong evidence to support an association between the consultant's work and fewer surgery complications.

Exercise 10.1.5.1

Does the conclusion in Exercise 6.48 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.

Solution

No. It might be that the consultant's work is associated with a reduction but that there isn't enough data to convincingly show this connection.

One-sided hypothesis test for p with a small sample

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions ($\hat{p}_{sim}$) and use this distribution to compute the tail area, i.e. the p-value.

We continue to use the same rule as before when computing the p-value for a twosided test: double the single tail area, which remains a reasonable approach even when the sampling distribution is asymmetric. However, this can result in p-values larger than 1 when the point estimate is very near the mean in the null distribution; in such cases, we write that the p-value is 1. Also, very large p-values computed in this way (e.g. 0.85), may also be slightly inated.

Exercise 6.48 said the p-value is estimated. It is not exact because the simulated null distribution itself is not exact, only a close approximation. However, we can generate an exact null distribution and p-value using the binomial model from Section 3.4.

Generating the exact null distribution and p-value

The number of successes in n independent cases can be described using the binomial model, which was introduced in Section 3.4. Recall that the probability of observing exactly k successes is given by

$$P(\text{k successes}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{6.50}$$

where p is the true probability of success. The expression $\binom{n}{k}$ is read as n choose k, and the exclamation points represent factorials. For instance, 3! is equal to $3 \times 2 \times 1 = 6$, 4! is equal to $4 \times 3 \times 2 \times 1 = 24$, and so on (see Section 3.4).

The tail area of the null distribution is computed by adding up the probability in Equation (6.50) for each k that provides at least as strong of evidence favoring the alternative hypothesis as the data. If the hypothesis test is one-sided, then the p-value is represented by a single tail area. If the test is two-sided, compute the single tail area and double it to get the p-value, just as we have done in the past.

Example 10.1.5.1

Compute the exact p-value to check the consultant's claim that her clients' complication rate is below 10%.

Solution

Exactly k = 3 complications were observed in the n = 62 cases cited by the consultant. Since we are testing against the 10% national average, our null hypothesis is p = 0.10. We can compute the p-value by adding up the cases where there are 3 or fewer complications:

$$\text{p-value} = \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \tag{10.1.5.1}$$

$$= \sum_{j=0}^{3} \binom{62}{j} 0.1^j (1-0.1)^{62-j} \tag{10.1.5.2}$$

$$= \binom{62}{0} 0.1^0 (1-0.1)^{62-0} + \binom{62}{1} 0.1^1 (1-0.1)^{62-1} + \binom{62}{2} 0.1^2 (1-0.1)^{62-2} + \binom{62}{3} 0.1^3 (1-0.1)^{62-3} \quad (10.1.5.3)$$

$$= 0.0015 + 0.0100 + 0.0340 + 0.0755 \quad (10.1.5.4)$$

$$= 0.1210 \quad (10.1.5.5)$$

This exact p-value is very close to the p-value based on the simulations (0.1222), and we come to the same conclusion. We do not reject the null hypothesis, and there is not statistically signi cant evidence to support the association.

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 6.20 on page 290.

Using simulation for goodness of fit tests

Simulation methods may also be used to test goodness of t. In short, we simulate a new sample based on the purported bin probabilities, then compute a chi-square test statistic X^2_{sim}. We do this many times (e.g. 10,000 times), and then examine the distribution of these simulated chi-square test statistics. This distribution will be a very precise null distribution for the test statistic X^2 if the probabilities are accurate, and we can nd the upper tail of this null distribution, using a cutoff of the observed test statistic, to calculate the p-value.

Example 10.1.5.1

Section 6.3 introduced an example where we considered whether jurors were racially representative of the population. Would our ndings differ if we used a simulation technique?

Solution

Since the minimum bin count condition was satis ed, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 6.21 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

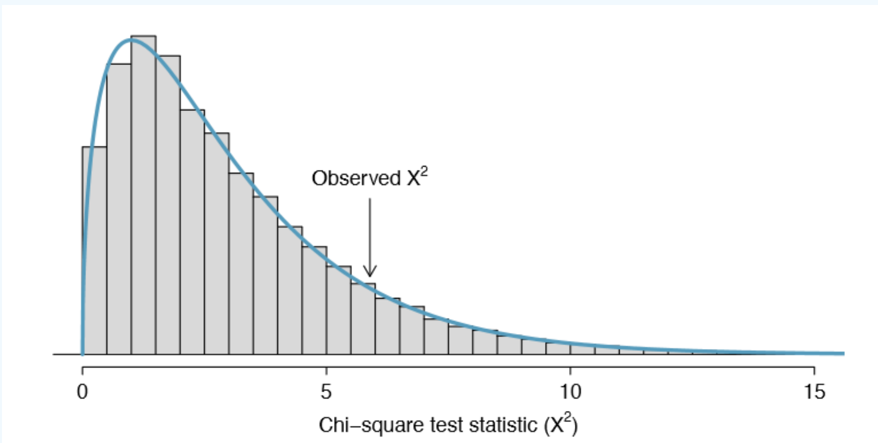


Figure 6.21: The precise null distribution for the juror example from Section 6.3 is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.

10.1.6: Randomization Test (Special Topic)

Cardiopulmonary resuscitation (CPR) is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. This procedure is helpful in maintaining some blood circulation, but the chest compressions involved can also cause internal injuries. Internal bleeding and other injuries complicate additional treatment efforts following arrival at a hospital. For instance, blood thinners may be used to help release a clot that is causing the heart attack. However, the blood thinner would negatively affect an internal injury. Here we consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.(Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial, by Bottiger et al., The Lancet, 2001.) These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive the blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

Example 10.1.6.1

Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival proportion in the control group and p_t represent the survival proportion for the treatment group.

Solution

We are interested in whether the blood thinners are helpful or harmful, so this should be a two-sided test.

- H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group.

$$p_t - p_c = 0. \tag{10.1.6.1}$$

- H_A: Blood thinners do have an impact on survival.

$$p_t - p_c \neq 0. \tag{10.1.6.2}$$

Large Sample Framework for a Difference in Two Proportions

There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did. The study results are shown in Table 6.22.

Table 6.22: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Exercise 10.1.6.1

What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$.

Solution

Observed control survival rate:

$$p_c = \frac{11}{50} = 0.22. \tag{10.1.6.3}$$

Treatment survival rate:

$$p_t = \frac{14}{40} = 0.35. \tag{10.1.6.4}$$

Observed difference:

$$\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13. \tag{10.1.6.5}$$

According to the point estimate, there is a 13% increase in the survival proportion when patients who have undergone CPR outside of the hospital are treated with blood thinners. However, we wonder if this difference could be due to chance. We'd like to investigate this using a large sample framework, but we first need to check the conditions for such an approach.

Example 10.1.6.2

Can the point estimate of the difference in survival proportions be adequately modeled using a normal distribution?

Solution

We will assume the patients are independent, which is probably reasonable. The success-failure condition is also satisfied. Since the proportions are equal under the null, we can compute the pooled proportion,

$$\hat{p} = \frac{(11+14)}{(50+40)} = 0.278, \tag{10.1.6.6}$$

for checking conditions. We nd the expected number of successes (13.9, 11.1) and failures (36.1, 28.9) are above 10. The normal model is reasonable.

While we can apply a normal framework as an approximation to nd a p-value, we might keep in mind that the expected number of successes is only 13.9 in one group and 11.1 in the other. Below we conduct an analysis relying on the large sample normal theory. We will follow up with a small sample analysis and compare the results.

Example 10.1.6.3

Assess the hypotheses presented in Example 6.53 using a large sample framework. Use a significance level of $\alpha = 0.05$.

Solution

We suppose the null distribution of the sample difference follows a normal distribution with mean 0 (the null value) and a standard deviation equal to the standard error of the estimate. The null hypothesis in this case would be that the two proportions are the same, so we compute the standard error using the pooled standard error formula from Equation (6.16) on page 273:

$$SE = \sqrt{\frac{p(1-p)}{n_t} + \frac{p(1-p)}{n_c}} \approx \sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}} = 0.095 \tag{10.1.6.7}$$

where we have used the pooled estimate $\left(\hat{p} = \frac{11+14}{50+40} = 0.278\right)$ in place of the true proportion, p.

The null distribution with mean zero and standard deviation 0.095 is shown in Figure 6.23. We compute the tail areas to identify the p-value. To do so, we use the Z score of the point estimate:

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - \text{null value}}{SE} = \frac{0.13-0}{0.095} = 1.37 \tag{10.1.6.8}$$

If we look this Z score up in Appendix B.1, we see that the right tail has area 0.0853. The p-value is twice the single tail area: 0.176. This p-value does not provide convincing evidence that the blood thinner helps. Thus, there is insufficient evidence to conclude whether or not the blood thinner helps or hurts. (Remember, we never "accept" the null hypothesis - we can only reject or fail to reject.)

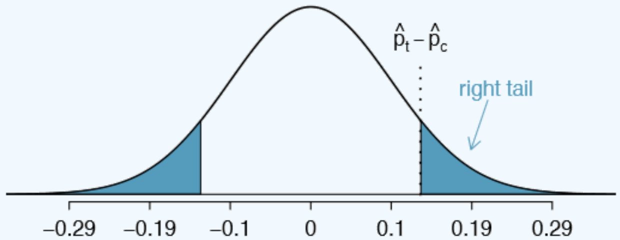


Figure 6.23: The null distribution of the point estimate $\hat{p}_t - \hat{p}_c$ under the large sample framework is a normal distribution with mean 0 and standard deviation equal to the standard error, in this case SE = 0.095. The p-value is represented by the shaded areas.

The p-value 0.176 relies on the normal approximation. We know that when the samples sizes are large, this approximation is quite good. However, when the sample sizes are relatively small as in this example, the approximation may only be adequate. Next we develop a simulation technique, apply it to these data, and compare our results. In general, the small sample method we develop may be used for any size sample, small or large, and should be considered as more accurate than the corresponding large sample technique.

Simulating a Difference under the null Distribution

The ideas in this section were first introduced in the optional Section 1.8. Suppose the null hypothesis is true. Then the blood thinner has no impact on survival and the 13% difference was due to chance. In this case, we can simulate null differences that are due to chance using a randomization technique. (The test procedure we employ in this section is formally called a **permutation test**). By randomly assigning "fake treatment" and "fake control" stickers to the patients' files, we could get a new grouping - one that is completely due to chance. The expected difference between the two proportions under this simulation is zero.

We run this simulation by taking 40 treatment fake and 50 control fake labels and randomly assigning them to the patients. The label counts of 40 and 50 correspond to the number of treatment and control assignments in the actual study. We use a computer program to randomly assign these labels to the patients, and we organize the simulation results into Table 6.24.

Table 6.24: Simulated results for the CPR study under the null hypothesis. The labels were randomly assigned and are independent of the outcome of the patient.

	Survived	Died	Total
Control_fake	15	35	50
Treatment_fake	10	30	40
Total	25	65	90

Exercise 10.1.6.2

What is the difference in survival rates between the two fake groups in Table 6.24? How does this compare to the observed 13% in the real groups?

Solution

The difference is $\hat{p}_{\text{t;fake}} - \hat{p}_{\text{c;fake}} = \frac{10}{40} - \frac{15}{50} = -0.05$, which is closer to the null value p0 = 0 than what we observed.

The difference computed in Exercise 6.57 represents a draw from the null distribution of the sample differences. Next we generate many more simulated experiments to build up the null distribution, much like we did in Section 6.5.2 to build a null distribution for a one sample proportion.

Caution: Simulation in the two proportion case requires that the null difference is zero

The technique described here to simulate a difference from the null distribution relies on an important condition in the null hypothesis: there is no connection between the two variables considered. In some special cases, the null difference might not be zero, and more advanced methods (or a large sample approximation, if appropriate) would be necessary.

Null distribution for the difference in two proportions

We build up an approximation to the null distribution by repeatedly creating tables like the one shown in Table 6.24 and computing the sample differences. The null distribution from 10,000 simulations is shown in Figure 6.25.

Example 10.1.6.4

Compare Figures 6.23 and 6.25. How are they similar? How are they different?

Solution

The shapes are similar, but the simulated results show that the continuous approximation of the normal distribution is not very good. We might wonder, how close are the p-values?

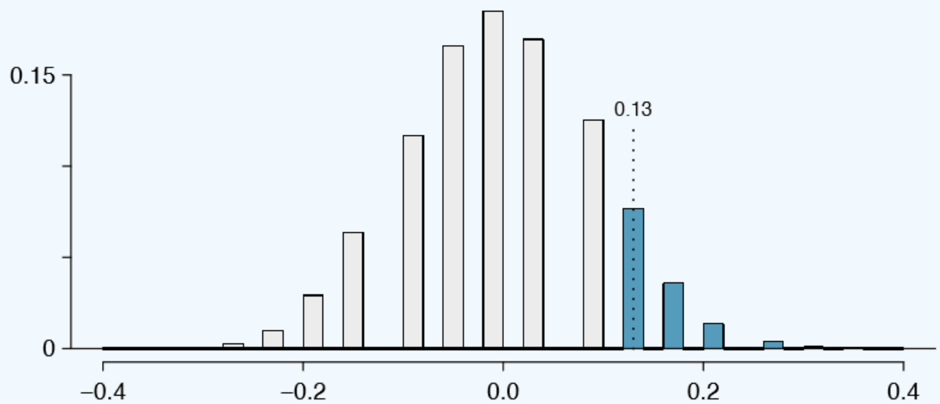


Figure 6.25: An approximation of the null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The p-value is twice the right tail area.

Exercise 10.1.6.3

The right tail area is about 0.13. (It is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) The p-value is computed by doubling the right tail area: 0.26. How does this value compare with the large sample approximation for the p-value?

Solution

The approximation in this case is fairly poor (p-values: 0.174 vs. 0.26), though we come to the same conclusion. The data do not provide convincing evidence showing the blood thinner helps or hurts patients.

In general, small sample methods produce more accurate results since they rely on fewer assumptions. However, they often require some extra work or simulations. For this reason, many statisticians use small sample methods only when conditions for large sample methods are not satisfied.

Randomization for two-way tables and chi-square

Randomization methods may also be used for the contingency tables. In short, we create a randomized contingency table, then compute a chi-square test statistic X^2_{sim}. We repeat this many times using a computer, and then we examine the distribution of these simulated test statistics. This randomization approach is valid for any sized sample, and it will be more accurate for cases where one or more expected bin counts do not meet the minimum threshold of 5. When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution. As before, we use the upper tail of the null distribution to calculate the p-value.

10.1.7: Exercises

Inference for a single proportion

6.1 Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

(b) The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

(c) A random sample of 125 college students where 12% are vegetarians would be considered unusual.

(d) A random sample of 250 college students where 12% are vegetarians would be considered unusual.

(e) The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

6.2 Young Americans, Part I. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.[36]

(a) The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.

(b) The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

(c) A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

(d) A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

6.3 Orange tabbies. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

6.4 Young Americans, Part II. About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning.[37]

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

(b) In order for the the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

[36] *A. Vaughn. "Poll finds young adults optimistic, but not about money". In: Los Angeles Times (2011).*

[37] *Demos.org. "The State of Young America: The Poll". In: (2011).*

6.5 Prop 19 in California. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[38]

(a) We are 95% con dent that between 62% and 78% of the California voters in this sample support Prop 19.

(b) We are 95% con dent that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.

(c) If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.

(d) In order to decrease the margin of error to 4%, we would need to quadruple (multiply by 4) the sample size.

(e) Based on this con dence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[39]

(a) We are 95% con dent that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b) We are 95% con dent that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

(d) The margin of error at a 90% con dence level would be higher than 3%.

6.7 Fireworks on July 4th. In late June 2012, Survey USA published results of a survey stating that 56% of the 600 randomly sampled Kansas residents planned to set off reworks on July 4th. Determine the margin of error for the 56% point estimate using a 95% con dence level.40

6.8 Elderly drivers. In January 2011, The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% con dence level.[41]

(a) Verify the margin of error reported by The Marist Poll.

(b) Based on a 95% con dence interval, does the poll provide convincing evidence that more than 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?

[38] *Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

[39] *Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.*

[40] *Survey USA, News Poll #19333, data collected on June 27, 2012.*

[41] *Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.*

6.9 Life after college. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and nd out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions for constructing a con dence interval based on these data are met.

(c) Calculate a 95% con dence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

(d) What does "95% confidence" mean?

(e) Now calculate a 99% con dence interval for the same parameter and interpret it in the context of the data.

(f) Compare the widths of the 95% and 99% con dence intervals. Which one is wider? Explain.

6.10 Life rating in Greece. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered

"suffering".[42]

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions required for constructing a con dence interval based on these data are met.

(c) Construct a 95% con dence interval for the proportion of Greeks who are "suffering".

(d) Without doing any calculations, describe what would happen to the con dence interval if we decided to use a higher confidence level.

(e) Without doing any calculations, describe what would happen to the con dence interval if we used a larger sample.

6.11 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[43]

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% con dence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

(c) What does "90% confidence" mean?

(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

[42]Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.

[43]studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[44]

(a) Is 48% a sample statistic or a population parameter? Explain.

(b) Construct a 95% con dence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

(c) A critic points out that this 95% con dence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

(d) A news piece on this survey's ndings states, \Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

6.13 Public option, Part I. A Washington Post article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More speci cally, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.[45]

(a) A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

(b) Would you expect a con dence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

6.14 The Civil War. A national survey conducted in 2011 among a simple random sample of 1,507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life.[46]

(a) Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) Interpret the p-value in this context.

(c) Calculate a 90% con dence interval for the proportion of Americans who think the Civil War is still relevant. Interpret the interval in this context, and comment on whether or not the con dence interval agrees with the conclusion of the hypothesis test.

6.15 Browsing on the mobile device. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.[47]

(a) According to an online article, a report from a mobile research company indicates that 38 percent of Chinese mobile web users only access the internet through their cell phones.48 Conduct a hypothesis test to determine if these data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%.

(b) Interpret the p-value in this context.

(c) Calculate a 95% con dence interval for the proportion of Americans who access the internet on their cell phones, and interpret the interval in this context.

[44]National Opinion Research Center, General Social Survey, 2010.

[45]D. Balz and J. Cohen. "Most support public option for health insurance, poll nds". In: The Washington Post (2009).

[46]Pew Research Center Publications, Civil War at 150: Still Relevant, Still Divisive, data collected between March 30 - April 3, 2011.

[47]Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.

[48]S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: M.I.C Gadget (2012).

6.16 Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.[49]

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a con dence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

6.17 Taste test. Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then lled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identi ed the soda.

(a) Do these data provide strong evidence that these people are able to detect the difference between diet and regular soda, in other words, are the results signi cantly better than just random guessing?

(b) Interpret the p-value in this context.

6.18 Is college worth it? Part II. Exercise 6.16 presents the results of a poll where 48% of 331 Americans who decide to not go to college do so because they cannot afford it.

(a) Calculate a 90% con dence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

(b) Suppose we wanted the margin of error for the 90% con dence level to be about 1.5%. How large of a survey would you recommend?

6.19 College smokers. We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke.

(a) Calculate a 95% con dence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: check conditions)

(b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

6.21 Public option, Part II. Exercise 6.13 presents the results of a poll evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.22 Acetaminophen and liver damage. It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher wants to conduct a study to estimate the proportion of acetaminophen users who have liver damage. For participating in this study, he will pay each subject \$20 and provide a free medical consultation if the patient has liver damage.

(a) If he wants to limit the margin of error of his 98% confidence interval to 2%, what is the minimum amount of money he needs to set aside to pay his subjects?

(b) The amount you calculated in part (a) is substantially over his budget so he decides to use fewer subjects. How will this affect the width of his con dence interval?

[49]*Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.*

Difference of two proportions

6.23 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

6.24 Heart transplant success. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might bene t from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.[50]

	Control	treatment
alive dead	4 30	24 45

A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate a con dence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

6.25 Gender and color preference. A 2001 study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.[51]

(a) We are 95% con dent that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% con dent that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% con dence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a signi cant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% con dence interval for ($p_{female} - p_{male}$) cannot be calculated with only the information given in this exercise.

[50]*B. Turnbull et al. "Survivorship of Heart Transplant Data". In: Journal of the American Statistical Association 69 (1974), pp. 74 - 80.*

[51]*L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: Personality and Individual Differences 31.8 (2001), pp. 1375-1379.*

6.26 The Daily Show. A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. A 95% con dence interval for ($p_{collegegrad} - p_{HSorless}$), where p is the proportion of those who watch The Daily Show, is (0.07, 0.15). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[52]

(a) At the 5% significance level, the data provide convincing evidence of a difference between the proportions of college graduates and those with a high school degree or less who watch The Daily Show.

(b) We are 95% con dent that 7% less to 15% more college graduates watch The Daily Show than those with a high school degree or less.

(c) 95% of random samples of 1,099 college graduates and 1,110 people with a high school degree or less will yield differences in sample proportions between 7% and 15%.

(d) A 90% confidence interval for ($p_{collegegrad} - p_{HSorless}$) would be wider.

(e) A 95% confidence interval for ($p_{HSorless} - p_{collegegrad}$) is (-0.15,-0.07).

6.27 Public Option, Part III. Exercise 6.13 presents the results of a poll evaluating support for the health care public option plan in 2009. 70% of 819 Democrats and 42% of 783 Independents support the public option.

(a) Calculate a 95% confidence interval for the difference between ($p_D - p_I$) and interpret it in this context. We have already checked conditions for you.

(b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the public option than the Independent.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% con dence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[53]

6.29 Offshore drilling, Part I. A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[54]

	College Grad	
	Yes	No

Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

[52]*The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.*

[53]*CDC, Perceived Insu�cient Rest or Sleep Among Adults - United States, 2008.*

[54]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

6.30 Sleep deprivation, CA vs. OR, Part II. Exercise 6.28 provides data on sleep deprivation rates of Californians and Oregonians. The proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Conduct a hypothesis test to determine if these data provide strong evidence the rate of sleep deprivation is different for the two states. (Reminder: check conditions)

(b) It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

6.31 Offshore drilling, Part II. Results of a poll evaluating support for drilling for oil and natural gas off the coast of California were introduced in Exercise 6.29.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample support drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who support offshore drilling in California is different than that of noncollege graduates.

6.32 Full body scan, Part I. A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1,137 adults nationwide, interviewed by telephone November 7-10, 2010, where one of the questions on the survey was "Some airports are now using `full-body' digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?" Below is a summary of responses based on party affiliation.[55]

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22

Total	318	369	450

(a) Conduct an appropriate hypothesis test evaluating whether there is a difference in the proportion of Republicans and Democrats who think the full-body scans should be applied in airports. Assume that all relevant conditions are met.

(b) The conclusion of the test in part (a) may be incorrect, meaning a testing error was made. If an error was made, was it a Type I or a Type II error? Explain.

[55] *S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: CBS News (2010).*

6.33 Sleep deprived transportation workers. The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers. The results of the survey are shown below.[56]

		Transportation	Professionals		
	Control	Pilots	Truck Drivers	Train Operators	Bux/Taxi/Limo Drivers
Less than 6 hours of sleep 6 to 8 hours of sleep More than 8 hours	35 193 64	19 132 51	35 117 51	29 119 32	21 131 58
Tota	292	202	203	180	210

Conduct a hypothesis test to evaluate if these data provide evidence of a difference between the proportions of truck drivers and non-transportation workers (the control group) who get less than 6 hours of sleep per day, i.e. are considered sleep deprived.

6.34 Prenatal vitamins and Autism. Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period).[57]

	Autism		
	Autism	Typical development	Total
No vitamin Vitamin	111 143	70 159	181 302
Total	254	229	483

(a) State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.

(b) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

(c) A New York Times article reporting on this study was titled "Prenatal Vitamins May Ward Off Autism". Do you nd the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.[58]

6.35 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called virologic failure). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[59]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

[56]*National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers Sleep, 2012.*

[57]*R.J. Schmidt et al. \Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism". In: Epidemiology 22.4 (2011), p. 476.*

[58]*R.C. Rabin. "Patterns: Prenatal Vitamins May Ward Off Autism". In: New York Times (2011).*

[59]*S. Lockman et al. "Response to antiretroviral therapy after a single, peripartum dose of nevirapine". In: Obstetrical & gynecological survey 62.6 (2007), p. 361.*

6.36 Diabetes and unemployment. A 2012 Gallup poll surveyed Americans about their employment status and whether or not they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.[60]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of incidence of diabetes and employment status.

(c) The sample difference is about 1%. If we completed the hypothesis test, we would nd that the p-value is very small (about 0), meaning the difference is statistically signi cant. Use this result to explain the difference between statistically signi cant and practically significant findings.

Testing for goodness of t using chi-square

6.37 True or false, Part I. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

(c) The chi-square statistic is always positive.

(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

6.38 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found X^2 = 10 with df = 5 you would fail to reject H0 at the 5% signi cance level.

(c) When nding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.39 Open source textbook. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

(a) State the hypotheses for testing if the professor's predictions were inaccurate.

(b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

(c) This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

(d) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

(e) Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

[60]*Gallup Wellbeing, Employed Americans in Better Health Than the Unemployed, data collected Jan. 2, 2011 - May 21, 2012.*

6.40 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[61]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1) Humans evolved, but God had no part in process (2) God created humans in present form (3) Other / No opinion (4)	38% 16% 40% 6%	37% 12% 45% 6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)

Testing for independence in two-way tables

6.41 Quitters. Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

(a) Create a two-way table presenting the results of this study.

(b) Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

i. How many subjects in the "patch + support" group would you expect to quit?

ii. How many subjects in the "only patch" group would you expect to not quit?

6.42 Full body scan, Part II. The table below summarizes a data set we rst encountered in Exercise 6.32 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's party affiliation and his support of full-body scans. It may be useful to rst add on an extra column for row totals before proceeding with the computations.

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22
Total	318	369	450

(a) How many Republicans would you expect to not support the use of full-body scans?

(b) How many Democrats would you expect to support the use of full-body scans?

(c) How many Independents would you expect to not know or not answer?

[61]*Four in 10 Americans Believe in Strict Creationism, December 17, 2010, http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx*.

6.43 Offshore drilling, Part III. The table below summarizes a data set we rst encountered in Exercise 6.29 that examines the responses of a random sample of college graduates and nongraduates on the topic of oil drilling. Complete a chi-square test for

these data to check whether there is a statistically signi cant difference in responses from college graduates and non-graduates.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

6.44 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[62]

		Caffeinated	coffee	consumption		
	$\leq$ 1cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Yes No	670 11,545	373 6,244	905 16,329	564 11,726	95 2,288	2,607 48,132
Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identi ed in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $\frac{(Observed-Expected)^2}{Expected}$.

(e) The test statistic is $X^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.63 Do you agree with this statement? Explain your reasoning.

[62]*M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: Archives of internal medicine 171.17 (2011), p. 1571.*

[63]*A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: New York Times (2011).*

6.45 Privacy on Facebook. A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, "Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?" The responses are cross-tabulated based on gender.[64]

	Gender		
	Male	Female	Total
Yes No Not sure	288 61 10	378 62 7	666 123 17
Total	359	447	806

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

6.46 Shipping holiday gifts. A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age		
	18-34	35-54	55+	Total
USPS	72 (81)	97 (102)	76 (62)	245
UPS	52 (53)	76 (68)	34 (41)	162
FedEx	31 (21)	24 (27)	9 (16)	64
Something else	7 (5)	6 (7)	3 (4)	16
Not sure	3 (5)	6 (5)	4 (3)	13
Total	165	209	126	500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

(b) Are the conditions for inference using a chi-square test satisfied?

Small sample hypothesis testing for a proportion

6.47 Bullying in schools. A 2012 Survey USA poll asked Florida residents how big of a problem they thought bullying was in local schools. 9 out of 191 18-34 year olds responded that bullying is no problem at all. Using these data, is it appropriate to construct a con dence interval using the formula $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for the true proportion of 18-34 year old Floridians who think bullying is no problem at all? If it is appropriate, construct the con dence interval. If it is not, explain why.

[64]*Survey USA, News Poll #17960, data collected February 16-17, 2011.*

6.48 Choose a test. We would like to test the following hypotheses:

H_0 : p = 0.1

$H_A : p \neq 0.1$

The sample size is 120 and the sample proportion is 8.5%. Determine which of the below test(s) is/are appropriate for this situation and explain your reasoning.

I. Z test for a proportion,

i.e. proportion test using normal model

II. Z test for comparing two proportions

III. X^2 test of independence

IV. Simulation test for a proportion

V. t test for a mean

VI. ANOVA

6.49 The Egyptian Revolution. A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the rst couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.[65]

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

(c) Based on large sample theory, we modeled ^p using the normal distribution. Why should we be cautious about this approach for these data?

(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.

(e) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.

[65]*Gallup Politics, Americans' Views of Egypt Sharply More Negative, data collected February 2-5, 2011.*

6.50 Assisted Reproduction. Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g. in vitro fertilization). A 2008 report by the Centers for Disease Control and Prevention estimated that ART has been successful in leading to a live birth in 31% of cases66. A new fertility clinic claims that their success rate is higher than average. A random sample of 30 of their patients yielded a success rate of 40%. A consumer watchdog group would like to determine if this provides strong evidence to support the company's claim.

(a) Write the hypotheses to test if the success rate for ART at this clinic is signi cantly higher than the success rate reported by the CDC.

(b) Based on large sample theory, we modeled ^p using the normal distribution. Why is this not appropriate here?

(c) The normal approximation would be less reliable here, so we should use a simulation strategy. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

(d) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

(e) After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.

[66]*CDC. 2008 Assisted Reproductive Technology Report.*

Hypothesis testing for two proportions

6.51 Social experiment, Part II. Exercise 6.23 introduces a "social experiment" conducted by a TV program that questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

A simulation was conducted to test if people react differently under the two scenarios. 10,000 simulated differences were generated to construct the null distribution shown. The value $\hat{p}_{pr;sim}$ represents the proportion of diners who intervened in the simulation for the provocatively dressed woman, and $\hat{p}_{con;sim}$ is the proportion for the conservatively dressed woman.

(a) What are the hypotheses? For the purposes of this exercise, you may assume that each observed person at the restaurant behaved independently, though we would want to evaluate this assumption more rigorously if we were reporting these results.

(b) Calculate the observed difference between the rates of intervention under the provocative and conservative scenarios: $\hat{p}_{pr} - \hat{p}_{con}$.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

6.52 Is yawning contagious? An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously inuenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.[67]

	Group		
	Treatment	Control	Total
Yawn Not Yawn	10 24	4 12	14 36
Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no inuence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning as $\hat{p}_{trtmt;sim} - \hat{p}_{ctrl;sim}$. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.

(a) What are the hypotheses?

(b) Calculate the observed difference between the yawning rates under the two scenarios.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

[67]*MythBusters, Season 3, Episode 28.*

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

SECTION OVERVIEW

10.2: Hypothesis Testing with Two Samples

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded. To compare two means or two proportions, you work with two groups. The groups are classified either as independent or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

Contributors and Attributions

10.2.1: Two Population Means with Unknown Standard Deviations

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations:
   - if the sample sizes are small, the distributions are important (should be normal)
   - if the sample sizes are large, the distributions are not important (need not be normal)

> The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\bar{X}_1 - \bar{X}_2$, and divide by the standard error in order to standardize the difference. The result is a t-score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\bar{X}_1 - \bar{X}_2$.

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \tag{10.2.1.1}$$

The test statistic (*t*-score) is calculated as follows:

$$\frac{(\bar{x} - \bar{x}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{10.2.1.2}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_1, respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means. μ_1 and μ_2 are the population means.

The number of *degrees of freedom* (df) requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The df are not always a whole number. The test statistic calculated previously is approximated by the Student's *t*-distribution with df as follows:

> **Degrees of freedom**
>
> $$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{10.2.1.3}$$

When both sample sizes n_1 and n_2 are five or larger, the Student's *t* approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

> It is not necessary to compute the degrees of freedom by hand. A calculator or computer easily computes it.

Example 10.2.1.1: Independent groups

The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in Table 10.2.1.1. Each populations has a normal distribution.

Table 10.2.1.1

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2	0.8660.866
Boys	16	3.2	1.00

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

Answer

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two independent groups, two population means.

Random variable: $\bar{X}_g - \bar{X}_b =$ difference in the sample mean amount of time girls and boys play sports each day.

- $H_0 : \mu_g = \mu_b$
- $H_0 : \mu_g - \mu_b = 0$
- $H_a : \mu_g \neq \mu_b$
- $H_a : \mu_g - \mu_b \neq 0$

The words **"the same"** tell you H_0 has an "=". Since there are no other words to indicate H_a, assume it says **"is different."** This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. **Do not pool the variances.**

Calculate the *p*-value using a Student's *t*-distribution: p-value $= 0.0054$

Graph:

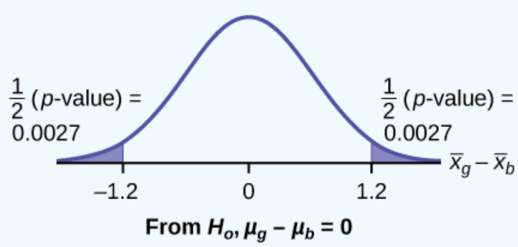


Figure 10.2.1.1: Normal distribution curve representing the difference in the average amount of time girls and boys play sports all day

$$s_g = 0.866 \tag{10.2.1.4}$$

$$s_b = 1 \tag{10.2.1.5}$$

So,

$$\bar{x}_g - \bar{x}_b = 2 - 3.2 = -1.2 \tag{10.2.1.6}$$

Half the p-value is below –1.2 and half is above 1.2.

Make a decision: Since $\alpha > p$-value, reject H_0. This means you reject $\mu_g = \mu_b$. The means are different.

Press `STAT`. Arrow over to `TESTS` and press `4:2-SampTTest`. Arrow over to Stats and press `ENTER`. Arrow down and enter `2` for the first sample mean, $\sqrt{0.866}$ for Sx1, `9` for n1, `3.2` for the second sample mean, `1` for Sx2, and `16` for n2. Arrow down to μ1: and arrow to `does not equal` μ2. Press `ENTER`. Arrow down to Pooled: and `No`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is $p = 0.0054$, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged seven to 11 play sports per day is different (mean number of hours boys aged seven to 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged seven to 11 play sports per day is greater than the mean number of hours played by boys).

? Exercise 10.2.1.1

Two samples are shown in Table. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

Table 10.2.1.2

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

Answer

The p-value is 0.4125, which is much higher than 0.05, so we decline to reject the null hypothesis. There is not sufficient evidence to conclude that the means of the two populations are not the same.

When the sum of the sample sizes is larger than $30 (n_1 + n_2 > 30)$ you can use the normal distribution to approximate the Student's t.

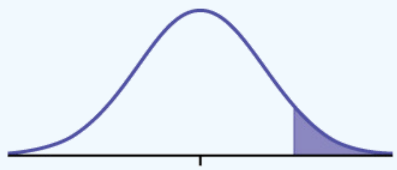
✓ Example 10.2.1.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes,** on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

a. Is this a test of two means or two proportions?
b. Are the populations standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?

Solutions

a. two means
b. unknown
c. Student's t
d. $\bar{X}_A - \bar{X}_B$
e. $H_0 : \mu_A \leq \mu_B$ and $H_a : \mu_A > \mu_B$



f. right

g. g. 0.1928
h. h. Do not reject.
i. i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

? Exercise 10.2.1.2

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

a. Are the population standard deviations known?
b. Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

Answer

a. They are unknown.
b. The $p\text{-value} = 0.0878$. At the 5% level of significance, there is insufficient evidence to conclude that the workers of Company A stay longer with the company.

✓ Example 10.2.1.3

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in Table $10.2.1.3$and Table $10.2.1.4$

Table 10.2.1.3: Online Class

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Table 10.2.1.4: Face-to-face Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?

c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right, left, or two tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?
i. At the ___ level of significance, from the sample data, there ______ (is/is not) sufficient evidence to conclude that ______.

(See the conclusion in Example, and write yours in a similar fashion)

Be careful not to mix up the information for Group 1 and Group 2!

Answer

a. two means
b. unknown
c. Student's t
d. $\bar{X}_1 - \bar{X}_2$
e.
   i. $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the final exam scores are equal for the online and face-to-face statistics classes.
   ii. $H_a : \mu_1 < \mu_2$ Alternative hypothesis: the mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.
f. left-tailed
g. $p\text{-value} = 0.0011$

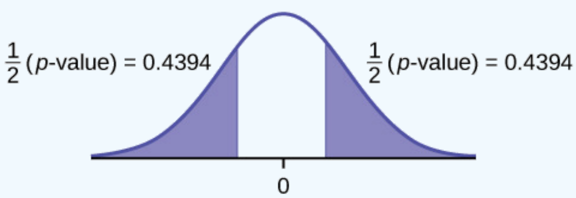


Figure **10.2.1.3.**

h. Reject the null hypothesis
i. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.

   At the 5% level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ_1: and arrow to $\neq \mu_1$ (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's d is a measure of effect size based on the differences between two means. Cohen’s d, named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen’s standards of small, medium, and large effect sizes.

Table 10.2.1.5: Cohen's Standard Effect Sizes

Size of effect	d
Small	0.2
medium	0.5

Size of effect	d
Large	0.8

Cohen's d is the measure of the difference between two means divided by the pooled standard deviation: $d=\frac{\bar{x}_2-\bar{x}_2}{s_{\text{pooled}}}$ where $s_{pooled}=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

Example 10.2.1.4

Calculate Cohen’s d for Example. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Answer

$$\mu_1=4 s_1=1.5 n_1=11$$

$$\mu_2=3.5 s_2=1 n_2=9$$

$$d=0.384$$

The effect is small because 0.384 is between Cohen’s value of 0.2 for small effect size and 0.5 for medium effect size. The size of the differences of the means for the two colleges is small indicating that there is not a significant difference between them.

Example 10.2.1.5

Calculate Cohen’s d for Example. Is the size of the effect small, medium or large? Explain what the size of the effect means for this problem.

Answer

$d=0.834$; Large, because 0.834 is greater than Cohen’s 0.8 for a large effect size. The size of the differences between the means of the Final Exam scores of online students and students in a face-to-face class is large indicating a significant difference.

Example 10.2.6

Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the northeast and in the west as identified by Nasdaq on May 24, 2013 are listed in Table and Table, respectively.

Northeast

94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0

West

126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5

Is there a difference in the weighted alpha of the top 30 stocks of banks in the northeast and in the west? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right, left, or two tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?
i. At the ___ level of significance, from the sample data, there ______ (is/is not) sufficient evidence to conclude that ______.
j. Calculate Cohen's d and interpret it.

Answer

a. two means
b. unknown
c. Student's-t
d. $\bar{X}_1 - \bar{X}_2$
e. i. $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the weighted alphas are equal.
   ii. $H_a : \mu_1 \neq \mu_2$ Alternative hypothesis : the means of the weighted alphas are not equal.
f. two-tailed
g. $p\text{-value} = 0.8787$
h. Do not reject the null hypothesis
i. This indicates that the trends in stocks are about the same in the top 30 banks in each region.

This is a normal distribution curve with mean equal to zero. Both the right and left tails of the curve are shaded. Each tail represents 1/2(p-value) = 0.4394.

Figure **10.2.1.4.**

5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean weighted alphas for the banks in the northeast and the west are different

j. $d = 0.040$, Very small, because 0.040 is less than Cohen's value of 0.2 for small effect size. The size of the difference of the means of the weighted alphas for the two regions of banks is small indicating that there is not a significant difference between their trends in stocks.

References

1. Data from Graduating Engineer + Computer Careers. Available online at www.graduatingengineer.com
2. Data from *Microsoft Bookshelf*.
3. Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).
4. "List of current United States Senators by Age." Wikipedia. Available online at en.Wikipedia.org/wiki/List_of...enators_by_age (accessed June 17, 2013).
5. "Sectoring by Industry Groups." Nasdaq. Available online at www.nasdaq.com/markets/barcha...&base=industry (accessed June 17, 2013).
6. "Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).
7. "World Series History." Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

Review

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $\bar{X}_1 - \bar{X}_2 =$ the difference of the sampling means
- Distribution: Student's t-distribution with degrees of freedom (variances not pooled)

Formula Review

Standard error:

$$SE = \sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}} \tag{10.2.1.7}$$

Test statistic (t-score):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{10.2.1.8}$$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2} + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2 \tag{10.2.1.9}$$

where:

- s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.
- x_1 and x_2 are the sample means.

Cohen's d is the measure of effect size:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \tag{10.2.1.10}$$

where

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{10.2.1.11}$$

Glossary

Degrees of Freedom (df)

the number of objects in a sample that are free to vary.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Variable (Random Variable)

a characteristic of interest in a population being studied. Common notation for variables are upper-case Latin letters X, Y, Z,... Common notation for a specific value from the domain (set of all possible values of a variable) are lower-case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if $X =$ hair color, then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the random variable X takes only after performing the experiment.

10.2.2: Two Population Means with Known Standard Deviations

Even though this situation is not likely (knowing the population standard deviations is not likely), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal and both populations must be normal. The random variable is $\bar{X}_1 - \bar{X}_2$. The normal distribution has the following format:

Normal distribution is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{10.2.2.1}$$

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}} \tag{10.2.2.2}$$

The test statistic (z-score) is:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \tag{10.2.2.3}$$

Example 10.2.2.1

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. Twenty floors are randomly assigned to test each wax. Both populations have a normal distributions. The data are recorded in Table.

Wax	Sample Mean Number of Months Floor Wax Lasts	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Does the data indicate that **wax 1 is more effective than wax 2**? Test at a 5% level of significance.

Answer

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\bar{X}_1 - \bar{X}_2 =$ difference in the mean number of months the competing floor waxes last.

- $H_0 : \mu_1 \leq \mu_2$
- $H_a : \mu_1 > \mu_2$

The words **"is more effective"** says that **wax 1 lasts longer than wax 2**, on average. "Longer" is a “>” symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The population standard deviations are known so the distribution is normal. Using Equation 10.2.2.1, the distribution is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right) \tag{10.2.2.4}$$

Since $\mu_1 \leq \mu_2$ then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

Calculate the p-value **using the normal distribution:** $p\text{-value} = 0.1799$

Graph:

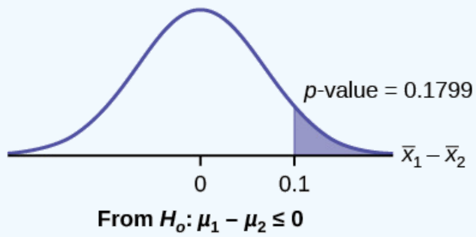


Figure 10.3.1.

$\bar{X}_1 - \bar{X}_2 = 3 - 2.9 = 0.1$

Compare α and the p-value: $\alpha = 0.05$ and $p\text{-value} = 0.1799$. Therefore, $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

Press `STAT` . Arrow over to `TESTS` and press `3:2-SampZTest` . Arrow over to `Stats` and press `ENTER` . Arrow down and enter `.33` for sigma1, `.36` for sigma2, `3` for the first sample mean, `20` for n1, `2.9` for the second sample mean, and `20` for n2. Arrow down to $\mu 1$: and arrow to $> \mu_2$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The p-value is $p = 0.1799$ and the test statistic is 0.9157. Do the procedure again, but instead of `Calculate` do `Dra` .

? Exercise 10.2.2.1

The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. Table shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5% level of significance.

Engine	Sample Mean Number of RPM	Population Standard Deviation
1	1,500	50
2	1,600	60

Answer

The p-value is almost 0, so we reject the null hypothesis. There is sufficient evidence to conclude that Engine 2 runs at a higher RPM than Engine 1.

✓ Example 10.2.2.2: Age of Senators

An interested citizen wanted to know if Democratic U. S. senators are older than Republican U.S. senators, on average. On May 26 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years.

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance.

Answer

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is 30 + 30 = 60, which is greater than 30, so we can use the normal approximation to the Student's-t distribution. Subscripts: 1: Democratic senators 2: Republican senators

Random variable: $\bar{X}_1 - \bar{X}_2 =$ difference in the mean age of Democratic and Republican U.S. senators.

- $H_0 : \mu_1 \le \mu_2 \quad H_0 : \mu_1 - \mu_2 \le 0$

- $H_a: \mu_1 > \mu_2 \quad H_a: \mu_1 - \mu_2 > 0$

The words "older than" translates as a “>” symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The distribution is the normal approximation to the Student’s t for means, independent groups. Using the formula, the distribution is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}}\right] \tag{10.2.2.5}$$

Since $\mu_1 \leq \mu_2, \mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

(Calculating the *p*-value using the normal distribution gives $p\text{-value} = 0.4040$)

Graph:

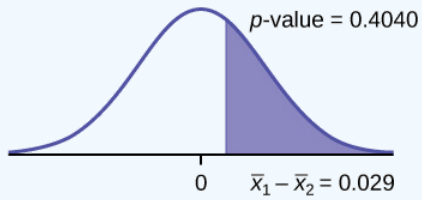


Figure 10.3.2.

Compare α and the p-value: $\alpha = 0.05$ and $p\text{-value} = 0.4040$. Therefore, $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

References

1. Data from the United States Census Bureau. Available online at www.census.gov/prod/cen2010/b...c2010br-02.pdf
2. Hinduja, Sameer. “Sexting Research and Gender Differences.” Cyberbulling Research Center, 2013. Available online at cyberbullying.us/blog/sexting...r-differences/ (accessed June 17, 2013).
3. “Smart Phone Users, By the Numbers.” Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).
4. Smith, Aaron. “35% of American adults own a Smartphone.” Pew Internet, 2013. Available online at www.pewinternet.org/~/media/F...martphones.pdf (accessed June 17, 2013).
5. “State-Specific Prevalence of Obesity AmongAduls—Unites States, 2007.” MMWR, CDC. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm (accessed June 17, 2013).
6. “Texas Crime Rates 1960–1012.” FBI, Uniform Crime Reports, 2013. Available online at: http://www.disastercenter.com/crime/txcrime.htm (accessed June 17, 2013).

Review

A hypothesis test of two population means from independent samples where the population standard deviations are known will have these characteristics:

- Random variable: $\bar{X}_1 - \bar{X}_2 =$ the difference of the means
- Distribution: normal distribution

Formula Review

Normal Distribution:

$$\bar{X}_1 - \bar{X}_2 = \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{10.2.2.6}$$

Generally $\mu_1 - \mu_2 = 0$.

Test Statistic (*z*-score):

$$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1}+\frac{(\sigma_2)^2}{n_2}}} \tag{10.2.2.7}$$

Generally $\mu_1-\mu_2=0$.

where:

σ_1 and σ_2 are the known population standard deviations. n_1 and n_1 are the sample sizes. $\bar{x}_1$ and $\bar{x}_2$ are the sample means. μ_1 and μ_2 are the population means

---

10.2.3: Comparing Two Independent Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_0 : p_A = p_B$. To conduct the test, we use a pooled proportion, p_c.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{10.2.3.1}$$

The distribution for the differences is:

$$p_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{10.2.3.2}$$

The test statistic (*z*-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{10.2.3.3}$$

Example 10.2.3.1

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Answer

The problem asks for a difference in proportions, making it a test of two proportions.

Let A and B be the subscripts for medication A and medication B, respectively. Then p_A and p_B are the desired population proportions.

Random Variable:$P'_A – P'_B =$ difference in the proportions of adult patients who did not react after 30 minutes to medication A and to medication B.

$H_0 : p_A = p_B$

$p_A - p_B = 0$

$H_a : p_A \neq p_B$

$p_A - p_B \neq 0$

The words **"is a difference"** tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 8001 - p_c = 0.92 \tag{10.2.3.4}$$

$$p'_A - p'_B \sim N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200}+\frac{1}{200}\right)}\right] \tag{10.2.3.5}$$

$p'_A - p'_B$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution: $p\text{-value} = 0.1404$.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

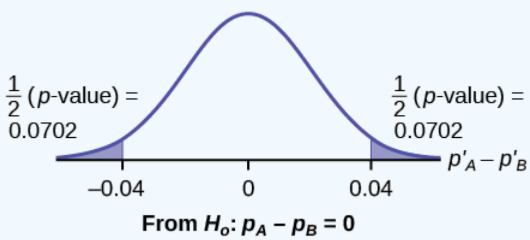


Figure 10.4.1.

$p'_A - p'_B = 0.1 - 0.06 = 0.04$.

Half the p-value is below -0.04, and half is above 0.04.

Compare α and the p-value : $\alpha = 0.01$ and the p-value $= 0.1404$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication *A* and medication *B*.

Press `STAT` . Arrow over to `TESTS` and press `6:2-PropZTest` . Arrow down and enter `20` for x1, `200` for n1, `12` for x2, and `200` for n2. Arrow down to `p1` : and arrow to `not equal p2` . Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The p-value is p = 0.1404 and the test statistic is 1.47. Do the procedure again, but instead of `Calculate` do `Draw` .

? Exercise 10.2.3.1

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

Answer

The p-value is 0.0379, so we can reject the null hypothesis. At the 5% significance level, the data support that there is a difference in the pressure tolerances between the two valves.

✓ Example 10.2.3.2: Sexting

A research study was conducted about gender differences in "sexting." The researcher believed that the proportion of girls involved in "sexting" is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in Table. Is the proportion of girls sending sexts less than the proportion of boys "sexting?" Test at a 1% level of significance.

	Males	Females
Sent "sexts"	183	156

	Males	Females
Total number surveyed	2231	2169

Answer

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then p_M and p_F are the desired population proportions.

Random variable: $p'_F - p'_M =$ difference in the proportions of males and females who sent "sexts."

$H_a : p_F = p_m \quad H_0 : p_F - p_M = 0$

$H_a : p_F < p_m \quad H_a : p_F - p_M < 0$

The words **"less than"** tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_C = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077 \tag{10.2.3.6}$$

$$1 - p_C = 0.923 \tag{10.2.3.7}$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right) \tag{10.2.3.8}$$

$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the p-value **using the normal distribution:**

$p\text{-value} = 0.1045$

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:

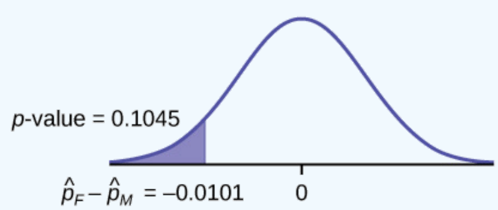


Figure 10.4.2.

Decision: Since $\alpha < p\text{-value}$, Do not reject H_0

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending "sexts" is less than the proportion of boys sending "sexts."

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press `ENTER` . Arrow down to Calculate and press ENTER. The p-value is $P = 0.1045$ and the test statistic is $z = -1.256$.

Example 10.2.3.3

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

Answer

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then p_W and p_A are the desired population proportions.

Random variable: $p'_W - p'_A =$ difference in the proportions of Android and iPhone users.

$H_0 : p_W = p_A \quad H_0 : p_W - p_A = 0$

$H_a : p_W > p_A \quad H_a : p_W - p_A < 0$

The words "more popular" indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$$p_C = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927 \tag{10.2.3.9}$$

$$1 - p_C = 0.9073 \tag{10.2.3.10}$$

Therefore,

$$p'_W - p'_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right) \tag{10.2.3.11}$$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:

$p\text{-value} = 0.0077$

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05

Graph:

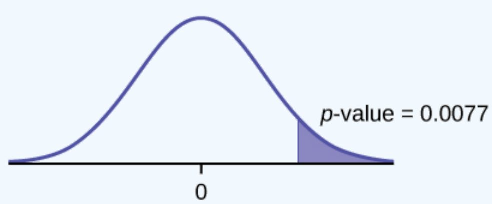


Figure 10.4.3.

Decision: Since $\alpha > p\text{-value}$, reject the H_0.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to p1: and arrow to greater than p2. Press ENTER. Arrow down to Calculate and press ENTER. The P-value is P = 0.0092 and the test statistic is Z = 2.33.

Example 10.2.3.3

A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category. Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Which distribution do you use to perform the test?

c. What is the random variable?
d. What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.
e. Is this test right-, left-, or two-tailed?
f. What is the p-value?
g. Do you reject or not reject the null hypothesis?
h. At the ___ level of significance, from the sample data, there ______ (is/is not) sufficient evidence to conclude that ____________.

Solutions

a. two proportions

b. normal for two proportions

c. Subscripts: 1 = 2010, 2 = 2011 $P'_1 - P'_2$

d. Subscripts: 1 = 2010, 2 = 2011 $H_0 : p_1 = p_2 H_0 : p_1 - p_2 = 0 \quad H_0 : p_1 \neq p_2 H_0 : p_1 - p_2 \neq 0$

e. two-tailed

f. p-value $= 0.00086$

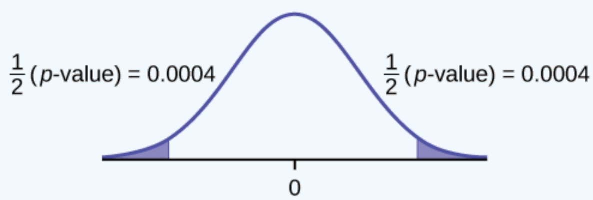


Figure 10.4.4.

g. Reject the H_0.

h. At the 5% significance level, from the sample data, there is sufficient evidence to conclude that there is a difference between the proportion of forcible rapes in 2011 and 2010.

References

1. Data from *Educational Resources*, December catalog.
2. Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013).
3. Data from Hyatt Hotels. Available online at hyatt.com (accessed June 17, 2013).
4. Data from Statistics, United States Department of Health and Human Services.
5. Data from Whitney Exhibit on loan to San Jose Museum of Art.
6. Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).
7. Data from the Chancellor's Office, California Community Colleges, November 1994.
8. "State of the States." Gallup, 2013. Available online at www.gallup.com/poll/125066/St...ef=interactive (accessed June 17, 2013).
9. "West Nile Virus." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).

Review

- Test of two population proportions from independent samples.
- Random variable: $\hat{p}_A - \hat{p}_B =$ difference between the two estimated proportions
- Distribution: normal distribution

Formula Review

Pooled Proportion:

$$p_c = \frac{x_F + x_M}{n_F + n_M} \tag{10.2.3.12}$$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}\right] \tag{10.2.3.13}$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

Test Statistic (*z*-score):

$$z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \tag{10.2.3.14}$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

and

- p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,
- P_c is the pooled proportion, and n_A and n_B are the sample sizes.

Glossary

Pooled Proportion

estimate of the common value of p_1 and p_2.

10.2.4: Matched or Paired Samples

When using a hypothesis test for matched or paired samples, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student's t-test for a single population mean with $n-1$ degrees of freedom, where n is the number of differences.

The test statistic (t-score) is:

$$t=\frac{\bar{x}_d-\mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{10.2.4.1}$$

Example 10.2.4.1

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in Table. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Answer

Corresponding "before" and "after" values form matched pairs. (Calculate "after" – "before.")

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

The data for the test are the differences: $\{0.2, -4.1, -1.6, -1.8, -3.2, -2, -2.9, -9.6\}$

The sample mean and sample standard deviation of the differences are: $\bar{x}_d=-3.13$ and $s_d=2.91$ Verify these values.

Let μ_d be the population mean for the differences. We use the subscript dd to denote "differences."

Random variable:

$\bar{X}_d =$ the mean difference of the sensory measurements

$$H_0 : \mu_d \geq 0 \tag{10.2.4.2}$$

The null hypothesis is zero or positive, meaning that there is the same or more pain felt after hypnotism. That means the subject shows no improvement. μ_d is the population mean of the differences.

$$H_a : \mu_d < 0 \tag{10.2.4.3}$$

The alternative hypothesis is negative, meaning there is less pain felt after hypnotism. That means the subject shows improvement. The score should be lower after hypnotism, so the difference ought to be negative to indicate improvement.

Distribution for the test:

The distribution is a Student's *t* with $df = n - 1 = 8 - 1 = 7$. Use t_7. *(Notice that the test is for a single population mean.)*

Calculate the *p*-value using the Student's-t distribution:

$$p\text{-value} = 0.0095 \tag{10.2.4.4}$$

Graph:

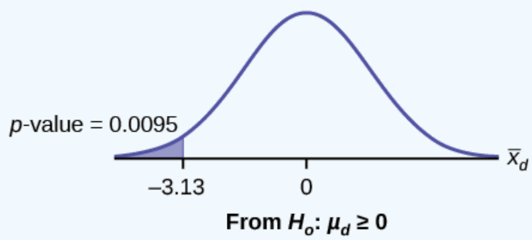


Figure 10.5.1.

$\bar{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$\bar{x}_d = -3.13$

$s_d = 2.91$

Compare α and the *p*-value

$\alpha = 0.05$ and $p\text{-value} = 0.0095$. $\alpha > p\text{-value}$

Make a decision

Since $\alpha > p\text{-value}$, reject H_0. This means that $\mu_d < 0$ and there is improvement.

Conclusion

At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

> For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after - before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction, and you will have the differences in the third list.

Use your list of differences as the data. Press `STAT` and arrow over to `TESTS` . Press `2:T-Test` . Arrow over to `Data` and press `ENTER` . Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to μ: and arrow over to `<` μ_0. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The *p*-value is 0.0094, and the test statistic is -3.04. Do these instructions again except, arrow to `Draw` (instead of `Calculate`). Press `ENTER` .

Exercise 10.2.4.1

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Answer

The p-value is 0.0130, so we can reject the null hypothesis. There is enough evidence to suggest that the diet lowers cholesterol.

Example 10.2.4.2

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.

Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: $\{90, 11, -8, -8\}$ Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$$\bar{x}_d = 21.3 \tag{10.2.4.5}$$

and

$$s_d = 46.7 \tag{10.2.4.6}$$

> The data given here would indicate that the distribution is actually right-skewed. The difference 90 may be an extreme outlier? It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are actually negative.

Using the difference data, this becomes a test of a single __________ (fill in the blank).

Define the random variable: $\bar{X}$ mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

- $H_0 : \mu_d \leq 0$,
- $H_a : \mu_d > 0$

Graph:

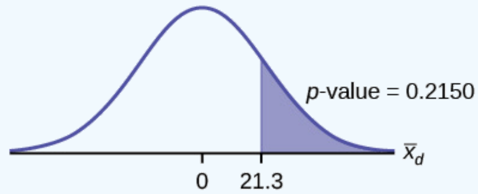


Figure 10.5.2.

Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is not to reject the null hypothesis, because $\alpha < p\text{-value}$.

What is the conclusion?

At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Exercise 10.2.4.2

A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data recorded in Table. Are the scores, on average, higher after the class? Test at a 5% level.

SAT Scores	Student 1	Student 2	Student 3	Student 4
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Answer

The p-value is 0.0874, so we decline to reject the null hypothesis. The data do not support that the class improves SAT scores significantly.

Example 10.2.4.3

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The data were collected and recorded in Table.

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Record the **differences** data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: $\{2, 12, 7, -1, 2, 0, 4\}$The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation. $\bar{x} = 3.71$, $s_d = 4.5$.

Random variable: $\bar{X} =$ mean difference in the distances between the hands.

Distribution for the hypothesis test: t_6

$H_0 : \mu_d = 0 \quad H_a : \mu_d \neq 0$

Graph:

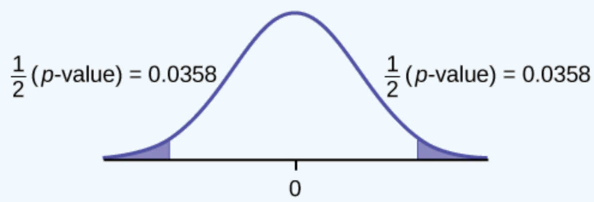


Figure 10.5.3.

Calculate the *p*-value: The p-value is 0.0716 (using the data directly).

(test statistic = 2.18. $p\text{-value} = 0.0719$ using $(\bar{x}_d = 3.71$, $s_d = 4.5$.

Decision: Assume $\alpha = 0.05$. Since $\alpha < p\text{-value}$, Do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the children's weaker and dominant hands to push the shot-put.

? Exercise 10.2.4.3

Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded in Table. Conduct a hypothesis test to determine whether the mean difference in distances between the dominant and off-hand is significant. Test at the 5% level.

	Player 1	**Player 2**	**Player 3**	**Player 4**	**Player 5**
Dominant Hand	120	111	135	140	125
Off-hand	105	109	98	111	99

Answer

The p-level is 0.0230, so we can reject the null hypothesis. The data show that the players do not throw the same distance with their off-hands as they do with their dominant hands.

Review

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random Variable: $x_d =$ mean of the differences
- Distribution: Student's t-distribution with $n-1$ degrees of freedom
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

Formula Review

Test Statistic (*t*-score):

$$t = \frac{\bar{x}_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{10.2.4.7}$$

where:

x_d is the mean of the sample differences. μ_d is the mean of the population differences. s_d is the sample standard deviation of the differences. n is the sample size.

10.3: PowerPoints

https://professormo.com/holistic/Powerpoint/ch10.pptx

CHAPTER OVERVIEW

11: Hypothesis Testing about Goodness of Fit (Multinomial)

SECTION OVERVIEW

11.1: Inference for Categorical Data

Chapter 6 introduces inference in the setting of categorical data. We use these methods to answer questions like the following:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

We will find that the methods we learned in previous chapters are very useful in these settings. For example, sample proportions are well characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual confidence interval and hypothesis testing tools. In other instances, such as those with contingency tables or when sample size conditions are not met, we will use a different distribution, though the core ideas remain the same.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

11.1.1: Inference for a Single Proportion

According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[1] This poll included responses of 976 adults.

Identifying when the Sample Proportion is Nearly Normal

A sample proportion can be described as a sample mean. If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0+1+1+\cdots+0}{976} = 0.44 \tag{11.1.1.1}$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes and failures, typically at least 10 of each.

[1] *nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html*

Conditions for the sampling distribution of $\hat{p}$ being nearly normal

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{11.1.1.2}$$

Typically we do not know the true proportion, p, so must substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value - that is, the proportion claimed in the null hypothesis - is used in place of p. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

TIP: Reminder on checking independence of observations

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Alternatively, if the data come from a random process, we must evaluate the independence condition more carefully.

Confidence Intervals for a Proportion

We may want a confidence interval for the proportion of Americans who approve of the job the Supreme Court is doing. Our point estimate, based on a sample of size n = 976 from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use the general confidence interval formula from Section 4.5, we must check the conditions to ensure that the sampling distribution of $\hat{p}$ is nearly normal. We also must determine the standard error of the estimate.

The data are based on a simple random sample and consist of far fewer than 10% of the U.S. population, so independence is confirmed. The sample size must also be sufficiently large, which is checked via the success-failure condition: there were approximately $976 \times \hat{p} = 429$ "successes" and $976 \times (1-\hat{p}) = 547$ "failures" in the sample, both easily greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

Exercise 11.1.1.1

Estimate the standard error of $\hat{p} = 0.44$ using Equation 11.1.1.2. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p.

Answer

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$$

Example 11.1.1.1

Construct a 95% confidence interval for p, the proportion of Americans who trust federal officials most of the time.

Solution

Using the standard error estimate from Exercise 11.1.1.1, the point estimate 0.44, and z* = 1:96 for a 95% confidence interval, the confidence interval may be computed as

$$\text{point estimate} \pm z^* SE \rightarrow 0.44 \pm 1.96 \times 0.016 \rightarrow (0.409, 0.471) \tag{11.1.1.3}$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. If the proportion has not changed since this poll, than we can say with high confidence that the job approval of the Supreme Court is below 50%.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, the sampling distribution of $\hat{p}$ may be well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

Hypothesis Testing for a Proportion

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1-p_0)$ are at least 10, where p_0 is the null value.

Exercise 11.1.1.2

Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Set up a onesided hypothesis test to evaluate this claim.

Answer

Is there convincing evidence that the campaign manager is correct?

- H_0 : p = 0.50,
- H_A : p > 0.50.

Example 11.1.1.2

A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%. Does this provide convincing evidence for the claim of Toohey's manager at the 5% significance level?

Solution

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion,

$$p_0 = 0.5 : np_0 = n(1-p_0) = 500 \times 0.5 = 250 > 10. \tag{11.1.1.4}$$

With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion.

$$SE = \sqrt{\frac{p_0 \times (1-p_0)}{n}}$$
$$= \sqrt{\frac{0.5(1-0.5)}{500}} = 0.022$$

A picture of the normal model is shown in Figure 11.1.1.1 with the p-value represented by the shaded region. Based on the normal model, the test statistic can be computed as the Z score of the point estimate:

$$Z = \frac{\text{point estimate - null value}}{SE}$$
$$= \frac{0.52-0.50}{0.022} = 0.89$$

The upper tail area, representing the p-value, is 0.1867. Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

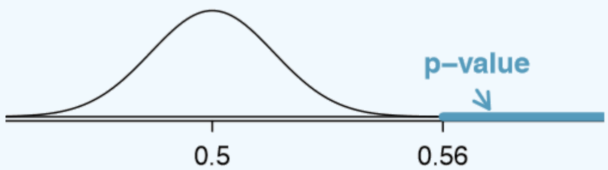


Figure 11.1.1.1: Sampling distribution of the sample proportion if the null hypothesis is true for Example 11.1.1.2. The p-value for the test is shaded.

Hypothesis test for a proportion

Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

Choosing a sample size when estimating a proportion

We first encountered sample size computations in Section 4.6, which considered the case of estimating a single mean. We found that these computations were helpful in planning a study to control the size of the standard error of a point estimate. The task was to find a sample size n so that the sample mean would be within some margin of error m of the actual mean with a certain level of confidence. For example, the margin of error for a point estimate using 95% confidence can be written as $1.96 \times SE$. We set up a general equation to represent the problem:

$$ME = z^* SE \leq m \tag{11.1.1.5}$$

where ME represented the actual margin of error and z^* was chosen to correspond to the confidence level. The standard error formula is specified to correspond to the particular setting. For instance, in the case of means, the standard error was given as $\frac{\sigma}{\sqrt{n}}$.

In the case of a single proportion, we use $\sqrt{p(1-p)}n$ for the standard error.

Planning a sample size before collecting data is equally important when estimating a proportion. For instance, if we are conducting a university survey to determine whether students support a $200 per year increase in fees to pay for a new football stadium, how big of a sample is needed to be sure the margin of error is less than 0.04 using a 95% confidence level?

Example 11.1.1.3

Find the smallest sample size n so that the margin of error of the point estimate $\hat{p}$ will be no larger than $m = 0.04$ when using a 95% confidence interval.

Solution

For a 95% confidence level, the value z* corresponds to 1.96, and we can write the margin of error expression as follows:

$$ME = z^* SE = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04 \tag{11.1.1.6}$$

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could use that value. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this worst case estimate if no other estimate is available:

$$1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.04 \tag{11.1.1.7}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{n} \leq 0.04^2 \tag{11.1.1.8}$$

$$1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} \leq n \tag{11.1.1.9}$$

$$600.25 \leq n \tag{11.1.1.10}$$

We would need at least 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

No estimate of the true proportion is required in sample size computations for a proportion, whereas an estimate of the standard deviation is always needed when computing a sample size for a margin of error for the sample mean. However, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

Example 11.1.1.4

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.

a. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.
b. The sample sizes in (b) vary widely. Which of the three would you suggest using? What would influence your choice?

Solution

(a) For the 1.7% estimate of p, we estimate the appropriate sample size as follows:

$$1.65 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \leq 0.02 \rightarrow n \geq 113.7$$

Using the estimate from the first model, we would suggest examining 114 tires (round up!). A similar computation can be accomplished using 0.062 and 0.013 for p: 396 and 88.

(b) We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. (Answers will vary.)

Exercise 11.1.1.4

A recent estimate of Congress' approval rating was 17%.5 What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?

Answer

We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \leq 0.04 \rightarrow n \geq 338.8$$

A sample size of 339 or more would be reasonable.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

11.1.2: Difference of Two Proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, a company weighs whether they should switch to a higher quality parts manufacturer. In the last example, we examine the cancer risk to dogs from the use of yard herbicides.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

Sample Distribution of the Difference of Two Proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when each proportion separately follows a **normal model**, and the samples are **independent**. The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} \tag{11.1.2.1}$$

$$= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{11.1.2.2}$$

where p_1 and p_2 represent the population proportions, and n1 and n2 represent the sample sizes.

For the difference in two means, the standard error formula took the following form:

$$SE_{\hat{x}_1 - \hat{x}_2} = \sqrt{SE^2_{\hat{x}_1} + SE^2_{\hat{x}_2}} \tag{11.1.2.3}$$

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Exercise 5.14.

[5]*www.gallup.com/poll/155144/Congress-Approval-June.aspx*

Table 11.1.2.1: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

	Sample size (ni)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Intervals and tests for $p_1 - p_2$

In the setting of confidence intervals, the sample proportions are used to verify the success/failure condition and also compute standard error, just as was the case with a single proportion.

Example 11.1.2.1

The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[7]

> *As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?*

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed. Table 6.2 shows the results of this experiment. Create and interpret a 90% confidence interval of the difference in approval.

Solution

First the conditions must be verified. Because each group is a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample. Because all conditions are met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13 \tag{11.1.2.4}$$

The standard error may be computed from Equation 11.1.2.2using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025 \tag{11.1.2.5}$$

For a 90% con dence interval, we use z* = 1.65:

$$\text{point estimate} \pm z^{*} SE \approx 0.13 \pm 1.65 \times 0.025 \rightarrow (0.09, 0.17) \tag{11.1.2.6}$$

We are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still uid on the health insurance mandate.

[7] www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/.

Sample sizes for each polling group are approximate.

Exercise 11.1.2.1

A remote control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[8]

Answer

Add texts here. Do not delete this text first.

Example 11.1.2.2

The quality control engineer from Exercise 6.11 collects a sample of gears, examining 1000 gears from each company and nds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Exercise 6.11 using a signi cance level of 5%.

Solution

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the gears are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as

to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, 0.958 - 0.899 = 0.059, can be said to come from a nearly normal distribution.

The standard error can be found using Equation 11.1.2.2:

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114 \tag{11.1.2.7}$$

In this hypothesis test, the sample proportions were used. We will discuss this choice more in Section 6.2.3.

Next, we compute the test statistic and use it to nd the p-value, which is depicted in Figure 11.1.2.1.

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54 \tag{11.1.2.8}$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher quality gears actually do pass inspection more than 3% as often as the currently used gears. Based on these results, management will approve the switch to the new supplier.

[8]H_0: The higher quality gears will pass inspection no more than 3% more frequently than the standard quality gears. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher quality gears will pass inspection more than 3% more often than the standard quality gears. $p_{highQ} - p_{standard} > 0.03$.

***Figure 11.1.2.1**: Distribution of the test statistic if the null hypothesis was true.*

The p-value is represented by the shaded area.

Hypothesis testing when H_0: $p_1 = p_2$

Here we use a new example to examine a special estimate of standard error when $H_0 : p_1 = p_2$. We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.9 Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table 11.1.2.2

Table 11.1.2.2: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

	cancer	no cancer
2,4 - D	191	304
no 2,4 - D	300	641

Exercise 11.1.2.1

Is this study an experiment or an observational study?

Answer

The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the control group, which is a term usually seen in experiments.

Exercise 11.1.2.1

Exercise 6.14 Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups.[11]

[9]Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. CaseControl Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.

Answer

Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.

- H_0: the proportion of dogs with exposure to 2,4-D is the same in "cancer" and \no cancer" dogs, $p_c - p_n = 0$.
- H_A: dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

Example 11.1.2.1: pooled estimate

First are the conditions met to use the normal model and make inference on the results?

(1) It is unclear whether this is a random sample. However, if we believe the dogs in both the cancer and no cancer groups are representative of each respective population and that the dogs in the study do not interact in any way, then we may find it reasonable to assume independence between observations. (2) The success-failure condition holds for each sample.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

In your hypotheses for Exercise 11.1.2.1, the null is that the proportion of dogs with exposure to 2,4-D is the same in each group. The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. To identify the p-value for this test, we first check conditions (Example 6.15) and compute the standard error of the difference:

$$SE = \sqrt{\frac{p_c(1-p_c)}{n_c} + \frac{p_n(1-p_n)}{n_n}} \tag{11.1.2.9}$$

In a hypothesis test, the distribution of the test statistic is always examined as though the null hypothesis is true, i.e. in this case, $p_c = p_n$. The standard error formula should reflect this equality in the null hypothesis. We will use p to represent the common rate of dogs that are exposed to 2,4-D in the two groups:

$$SE = \sqrt{\frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_n}} \tag{11.1.2.10}$$

We don't know the exposure rate, p, but we can obtain a good estimate of it by pooling the results of both samples:

$$\hat{p} = \frac{\#\text{ of ”successes”}}{\#\text{ of cases}} = \frac{191+304}{191+300+304+641} = 0.345 \tag{11.1.2.11}$$

This is called the **pooled estimate** of the sample proportion, and we use it to compute the standard error when the null hypothesis is that $p_1 = p_2$ (e.g. $p_c = p_n$ or $p_c - p_n = 0$) . We also typically use it to verify the success-failure condition.

Pooled estimate of a proportion

When the null hypothesis is $p_1 = p_2$, it is useful to nd the pooled estimate of the shared proportion:

$$\hat{p} = \frac{\text{number of ”successes”}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} \tag{11.1.2.12}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1} \tag{11.1.2.13}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

: $p_1 = p_2$

When the null hypothesis suggests the proportions are equal, we use the pooled proportion estimate $(\hat{p})$ to verify the success-failure condition and also to estimate the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_c} + \frac{\hat{p}(1-\hat{p})}{n_n}} \tag{11.1.2.14}$$

Exercise 11.1.2.1

Using Equation 11.1.2.14, $\hat{p} = 0.345$, $n_1 = 491$, and $n_2 = 945$, verify the estimate for the standard error is SE = 0.026. Next, complete the hypothesis test using a significance level of 0.05. Be certain to draw a picture, compute the p-value, and state your conclusion in both statistical language and plain language.

Answer

Compute the test statistic:

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

Contributors

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

11.1.3: Testing for Goodness of Fit using Chi-Square (Special Topic)

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test. In the rst case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Table 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reect the population of eligible jurors, i.e. registered voters.

Table 6.5: Representation by race in a city's juries and population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

[12]*Compute the test statistic:*

$$Z = \frac{\text{point estimate - null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58 \tag{11.1.3.1}$$

We leave the picture to you. Looking up Z = 2.58 in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: 1- 0.9951 = 0.0049. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

A second application, assessing the t of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for the years 1990-2011 are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

Creating a test statistic for one-way tables

Example 11.1.3.1:

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

Solution

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

Exercise 11.1.3.1

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race?

Answer

Answers can be found in Table 6.6.

Table 6.6: Actual and expected make-up of the jurors.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected count	198	19.25	33	24.75	275

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

- H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.
- H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate - null value}}{\text{SE of point estimate}} \tag{11.1.3.2}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to rst compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count - null white count}}{\text{SE of observed white count}} \tag{11.1.3.3}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50 \tag{11.1.3.4}$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

$$\begin{array}{ccc} Black & Hispanic & Other \\ Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 154 & Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39 & Z_4 = \frac{19 - 24:75}{\sqrt{24.75}} = -1.16 \end{array} \tag{11.1.3.5}$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they - as a group - tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58 \tag{11.1.3.6}$$

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89 \tag{11.1.3.7}$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual - e.g. a standardized difference of 2.5 - will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{(\text{observed count\_1 - null count\_1})^{\,2}}{\text{null count\_1}} + \cdots + \frac{(\text{observed count\_4 - null count\_4})^{\,2}}{\text{null count\_4}} \tag{11.1.3.8}$$

[13]Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown - though not here - that the square root of the count is a better way to standardize the count differences.

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a chi-square distribution. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

The chi-square distribution and finding areas

The chi-square distribution is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters - mean and standard deviation - that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df),** which inuences the shape, center, and spread of the distribution.

Exercise 11.1.3.1

Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

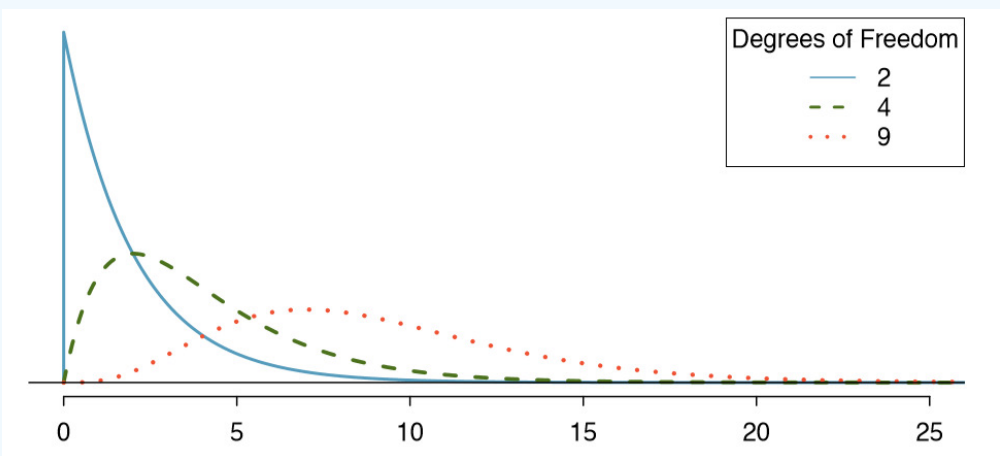


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

Answer

[14](a) The center becomes larger. If we look carefully, we can see that the center of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for df = 2, and then the distributions become more symmetric for the larger degrees of freedom df = 4 and df = 9. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Figure 6.7 and Exercise 6.20 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the **chi-square table**, partially shown in Table 6.8. A more complete table is presented in Appendix B.3 on page 412. This table is very similar to the t table from Sections 5.3 and 5.4: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t table is that the chi-square table only provides upper tail values.

Table 6.8: A section of the chi-square table. A complete table is in Appendix B.3 on page 412.

Upper tail	0.3 0.2 0.1 0.05	0.02 0.01 0.005 0.001
df 1 2 3 4 5	2.41 3.22 4.61 5.99 3.66 4.64 6.25 7.81 4.88 5.99 7.78 9.49 6.06 7.29 9.24 11.07	7.82 9.21 10.60 13.82 9.84 11.34 12.84 16.27 11.67 13.28 14.86 18.47 13.39 15.09 16.75 20.52
6 7	7.23 8.56 10.64 12.59 8.38 9.80 12.02 14.07	15.03 16.81 18.55 22.46 16.62 18.48 20.28 24.32

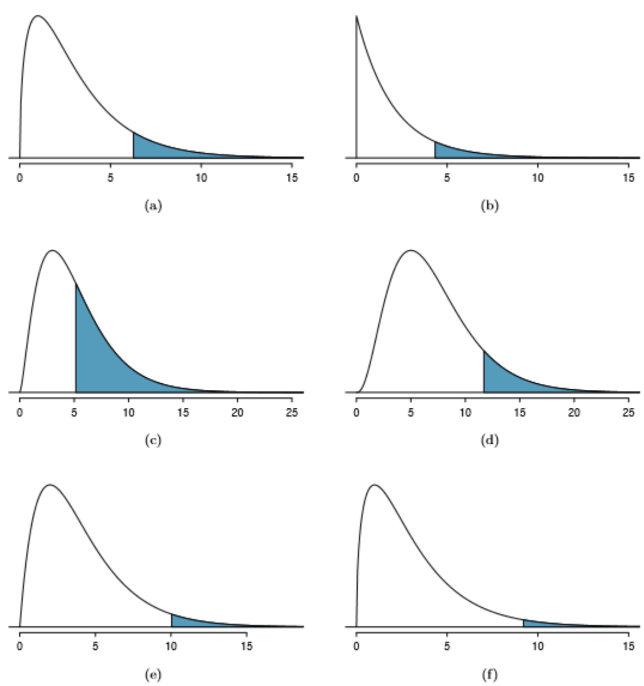


Figure 6.9: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded. (e) 4 degrees of freedom, area above 10 shaded. (f) 3 degrees of freedom, area above 9.21 shaded.

Example 6.21 Figure 6.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 6.8 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value { 6.25 { falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 6.9(a) has area 0.1.

Example 6.22 We rarely observe the exact value in the table. For instance, Figure 6.9(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3, which does not fall in Table 6.8. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 6.9(b) is between 0.1 and 0.2.

Example 6.23 Figure 6.9(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is greater than 0.3.

Exercise 6.24 Figure 6.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

Exercise 6.25 Figure 6.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Exercise 6.26 Figure 6.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

[15] *The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.*

[16] *The area is between 0.02 and 0.05.*

[17] *Between 0.02 and 0.05.*

Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with k - 1 degrees of freedom, where k is the number of bins.

Example 11.1.3.1:

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

Solution

In the jurors example, there were k = 4 categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with k - 1 = 3 degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

Example 11.1.3.1:

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

The chi-square distribution and p-value are shown in Figure 6.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix B.3 or the short table on page 277, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

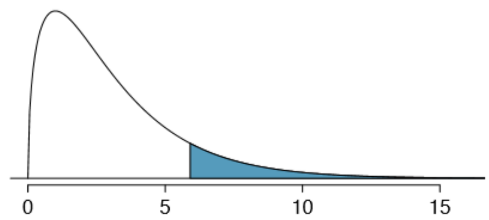


Figure 6.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with df = 3.

Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts $O_1, O_2, \ldots, O_k$ in k categories are unusually different from what might be expected under a null hypothesis. Call the expected counts that are based on the null hypothesis $E_1, E_2, \ldots, E_k$. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with k - 1 degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \tag{11.1.3.9}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

Conditions for the chi-square test

There are three conditions that must be checked before performing a chi-square test:

- **Independence**. Each case that contributes a count to the table must be independent of all the other cases in the table.
- **Sample size / distribution**. Each particular scenario (i.e. cell count) must have at least 5 expected cases.
- **Degrees of freedom** We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one proportion methods introduced in Section 6.1.

Evaluating goodness of fit for a distribution

Section 3.3 would be useful background reading for this example, but it is not a prerequisite. We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model ts a data set. Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as Up or Down (D) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each Up day:

$$\begin{array}{rcccccccccc} \text{Change in price} & 2.52 & -1.46 & 0.51 & -4.07 & 3.36 & 1.10 & -5.46 & -1.03 & -2.99 & 1.71 \\ \text{Outcome} & Up & D & Up & D & Up & Up & D & D & D & Up \\ \text{Days to Up} & 1 & - & 2 & - & 2 & 1 & - & - & - & 4 \end{array} \tag{11.1.3.10}$$

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the kth trial to observe the rst success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next Up trading day, and two more for the third Up day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Table 6.11 shows the number of waiting days for a positive trading day during 1990-2011for the S&P500.

Table 6.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

We consider how many days one must wait until observing an Up day on the S&P500 stock exchange. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a geometric distribution. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an Up day is observed. Under this hypothesis, the number of days until an Up day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an Up day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 from 1990 to 2011 and summarize the waiting times in Table 6.12 and Figure 6.13. The S&P500 was positive on 53.2% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have binned together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the Observed row in Table 6.12, can be compared to the expected counts from the Geometric Model row. The method for computing expected counts is discussed in Table 6.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each

Table 6.12: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model $(P(D) = (1 - 0.532)^{D-1}(0.532))$ and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 =$ 11.65% of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

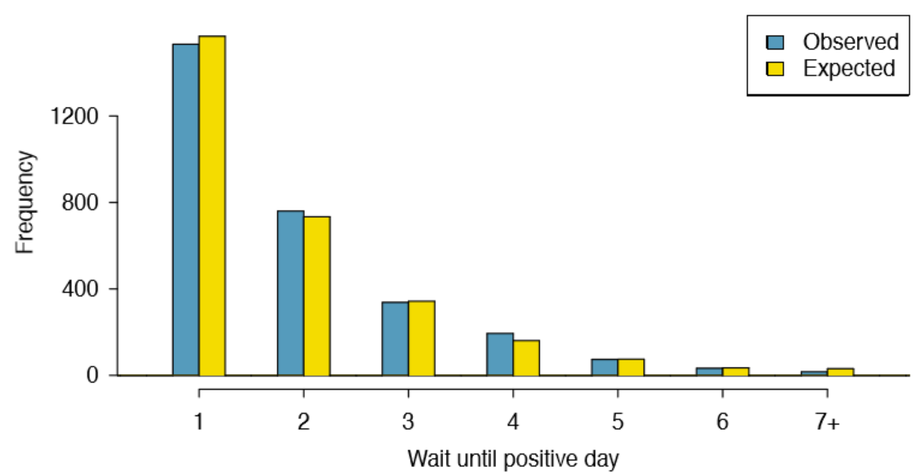


Figure 6.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Example 6.29 Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Table 6.12.

Exercise 6.30 Table 6.12 provides a set of count data for waiting times ($O_1 = 1532, O_2 = 760, \ldots$) and expected counts under the geometric distribution ($E_1 = 1569, E_2 = 734, \ldots$). Compute the chi-square test statistic, X^2.[18]

Exercise 6.31 Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

Example 6.32 If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with df = 6. Using this information, compute a p-value.

Figure 6.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix B.3, we nd that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that

[18] $X^2 = \dfrac{(1532-1569)^2}{1569} + \dfrac{(760-734)^2}{734} + \cdots + \dfrac{(17-31)^2}{31} = 15.08$

[19] *There are k = 7 groups, so we use df = k - 1 = 6.*

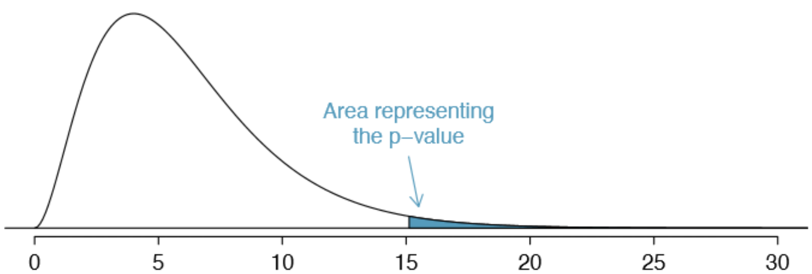


Figure 6.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

Example 6.33 In Example 6.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

11.1.4: Testing for Independence in Two-Way Tables (Special Topic)

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Table 6.15 shows an example of 10,000 queries split into three algorithm groups.[20] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Table 6.15: Google experiment breakdown of test subjects into three search groups.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

[20]Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

Example 11.1.4.1

What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

- H_0: The algorithms each perform equally well.
- H_A: The algorithms do not perform equally well.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. Under scenario (1), we might think that the user was satis ed with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Table 6.16 provides the results from the experiment. These data are very similar to the count data in Section 6.3. However, now the different combinations of two variables are binned in a two-way table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

Table 6.16: Results of the Google search algorithm experiment.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

What is so different about one-way tables and two-way tables?

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for combinations of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the search variable is independent of the algorithm variable.

Expected Counts in Two-way Tables

Example 6.35 From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as $\frac{7078}{10000} = 0.7078$. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satis ed with the initial search:

$$0.7078 \times 5000 = 3539 users \tag{11.1.4.1}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

Exercise 11.1.4.1

Exercise 6.36 Using the same rationale described in Example 6.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

[21] *We would expect 0.7078 * 2500 = 1769.5. It is okay that this is a fraction.*

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 6.35 and Exercise 6.36. These expected counts were used to construct Table 6.17, which is the same as Table 6.16, except now the expected counts have been added in parentheses.

Table 6.17: The observed counts and the (Expected counts)

Search algorithm	current	test 1	test 2	Total
No new search	3511 (3539)	1749 (1769.5)	1818 (1769.5)	7078
New search	1489 (1461)	751 (730.5)	682 (730.5)	2922
Total	5000	2500	2500	10000

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the rst row:

$$0.7078 \times (\text{column 1 total}) = 3539 \tag{11.1.4.2}$$

$$0.7078 \times (\text{column 2 total}) = 1769.5 \tag{11.1.4.3}$$

$$0.7078 \times (\text{column 3 total}) = 1769.5 \tag{11.1.4.4}$$

Looking back to how the fraction 0.7078 was computed - as the fraction of users who did not perform a new search ($\frac{7078}{10000}$) - these three expected counts could have been computed as

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 1 total}) = 3539 \tag{11.1.4.5}$$

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 2 total}) = 1769.5 \tag{11.1.4.6}$$

$$\frac{\text{row 1 total}}{\text{table total}}(\text{column 3 total}) = 1769.5 \tag{11.1.4.7}$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

Computing expected counts in a two-way table

To identify the expected count for the ith row and jth column, compute

$$\text{Expected Count}_{\text{row i, col j}} = \frac{(\text{row i total}) \times (\text{column j total})}{\text{table total}} \tag{11.1.4.8}$$

The chi-square Test for Two-way Tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

$$\text{General formula} \frac{(\text{observed count - expected count})^{2}}{\text{expected count}} \tag{11.1.4.9}$$

$$\text{Row 1, Col 1} \frac{(3511 - 3539)^2}{3539} = 0.222 \tag{11.1.4.10}$$

$$\text{Row 1, Col 2} \frac{(1749 - 1769.5)^2}{1769.5} = 0.237 \tag{11.1.4.11}$$

$$\vdots\vdots \tag{11.1.4.12}$$

$$\text{Row 2, Col 3} \frac{(682 - 730.5)^2}{730.5} = 3.220 \tag{11.1.4.13}$$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120 \tag{11.1.4.14}$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1}) \tag{11.1.4.15}$$

In our example, the degrees of freedom parameter is

$$df = (2 - 1) \times (3 - 1) = 2 \tag{11.1.4.16}$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic X^2 = 6.12 closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.18.

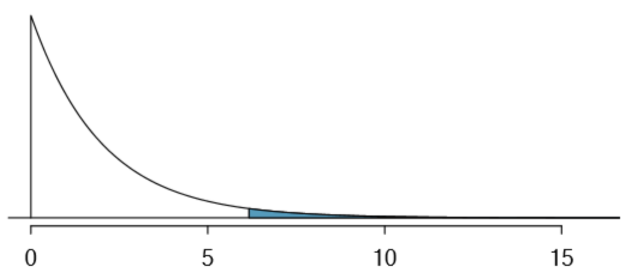


Figure 6.18: Computing the p-value for the Google hypothesis test.

Definition: degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df = (R - 1) \times (C - 1) \tag{11.1.4.17}$$

where R is the number of rows in the table and C is the number of columns.

[22]*Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.*

Table 6.19: Pew Research poll results of a March 2012 poll.

		Congress		
	Obama	Democrats	Republicans	Total

Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

TIP: Use two-proportion methods for 2-by-2 contingency tables

When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 6.2.

Example 11.1.4.1

Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Solution

Looking in Appendix B.3 on page 412, we examine the row corresponding to 2 degrees of freedom. The test statistic, X^2 = 6.120, falls between the fourth and fth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of α = 0.05 and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Example 11.1.4.1

Table 6.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

Solution

- H_0: There is no difference in approval ratings between the three groups.
- H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23]See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 6.19 are approximate.

Exercise 11.1.4.1

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.38. As a rst step, compute the expected values for each of the six table cells.[24]

[24]The expected count for row one / column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119 \times 1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104 \times 1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

Exercise 11.1.4.1

Compute the chi-square test statistic.[25]

[25]For each cell, compute $\frac{(\text{obs - exp})^2}{exp}$. For instance, the rst row and rst column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \dots + 34.0 = 106.4$.

Exercise 11.1.4.1

Because there are 2 rows and 3 columns, the degrees of freedom for the test is df = (2 - 1) (3 - 1) = 2. Use X^2 = 106.4, df = 2, and the chi-square table on page 412 to evaluate whether to reject the null hypothesis.[26]

[26]The test statistic is larger than the right-most column of the df = 2 row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans'

approval has differences among Democrats in Congress, Republicans in Congress, and the president.

11.1.5: Small Sample Hypothesis Testing for a Proportion (Special Topic)

In this section we develop inferential methods for a single proportion that are appropriate when the sample size is too small to apply the normal model to $\hat{p}$. Just like the methods related to the t distribution, these methods can also be applied to large samples.

When the Success-Failure Condition is Not Met

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspect of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients. One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Exercise 11.1.5.1

Exercise 6.42

We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.

Solution

The sample proportion: $\hat{p} = \frac{3}{62} = 0.048$

Example 11.1.5.1

Is it possible to assess the consultant's claim using the data provided?

Solution

No. The claim is that there is a causal connection, but the data are observational. Patients who hire this medical consultant may have lower complication rates for other reasons.

While it is not possible to assess this causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Exercise 11.1.5.1

Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.

Solution

- H_0: There is no association between the consultant's contributions and the clients' complication rate. In statistical language, p = 0.10.
- H_A: Patients who work with the consultant tend to have a complication rate lower than 10%, i.e. p < 0.10.

Example 11.1.5.1

In the examples based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

Solution

The independence assumption may be reasonable if each of the surgeries is from a different surgical team. However, the success-failure condition is not satis ed. Under the null hypothesis, we would anticipate seeing $62 \times 0.10 = 6.2$ complications, not the 10 required for the normal approximation.

The uncertainty associated with the sample proportion should not be modeled using the normal distribution. However, we would still like to assess the hypotheses from Exercise 6.44 in absence of the normal framework. To do so, we need to evaluate the possibility of a sample value (^p) this far below the null value, $p_0 = 0.10$. This possibility is usually measured with a p-value.

The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis is true. Supposing the null hypothesis is true, we can compute the p-value by identifying the chance of observing a test statistic that favors the alternative hypothesis at least as strongly as the observed test statistic. This can be done using simulation.

Generating the null distribution and p-value by simulation

We want to identify the sampling distribution of the test statistic ($\hat{p}$) if the null hypothesis was true. In other words, we want to see how the sample proportion changes due to chance alone. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.

Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients. If this was the case, we could simulate 62 clients to get a sample proportion for the complication rate from the null distribution.

Each client can be simulated using a deck of cards. Take one red card, nine black cards, and mix them up. Then drawing a card is one way of simulating the chance a patient has a complication if the true complication rate is 10% for the data. If we do this 62 times and compute the proportion of patients with complications in the simulation, $\hat{p}_{sim}$, then this sample proportion is exactly a sample from the null distribution.

An undergraduate student was paid $2 to complete this simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e. $\hat{p}_{sim} = \frac{5}{62} = 0.081$.

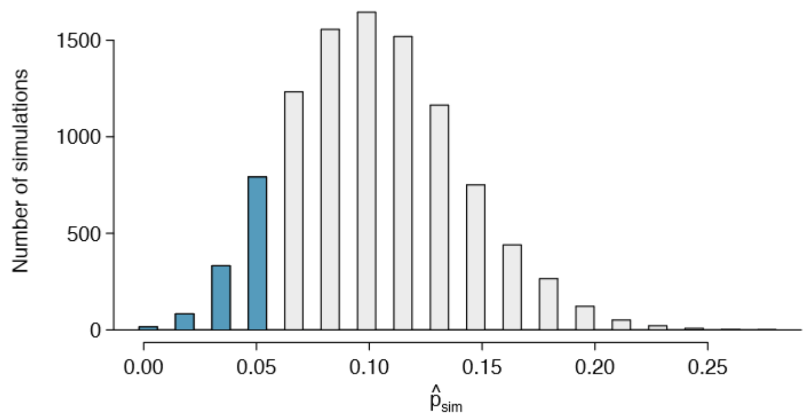


Figure 6.20: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

Example 11.1.5.1

Is this one simulation enough to determine whether or not we should reject the null hypothesis from Exercise 6.44? Explain.

Solution

No. To assess the hypotheses, we need to see a distribution of many $\hat{p}_{sim}$, not just a single draw from this sampling distribution.

One simulation isn't enough to get a sense of the null distribution; many simulation studies are needed. Roughly 10,000 seems sufficient. However, paying someone to simulate 10,000 studies by hand is a waste of time and money. Instead, simulations are typically programmed into a computer, which is much more efficient.

Figure 6.20 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and nd the p-value:

$$\text{left tail} = \frac{\text{Number of observed simulations with} \hat{p}_{sim} \leq 0.048}{10000} \tag{6.47}$$

Of the 10,000 simulated $\hat{p}_{sim}$, 1222 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.1222.

Exercise 11.1.5.1

Because the estimated p-value is 0.1222, which is larger than the signi cance level 0.05, we do not reject the null hypothesis. Explain what this means in plain language in the context of the problem.

Solution

There isn't sufficiently strong evidence to support an association between the consultant's work and fewer surgery complications.

Exercise 11.1.5.1

Does the conclusion in Exercise 6.48 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.

Solution

No. It might be that the consultant's work is associated with a reduction but that there isn't enough data to convincingly show this connection.

One-sided hypothesis test for p with a small sample

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions ($\hat{p}_{sim}$) and use this distribution to compute the tail area, i.e. the p-value.

We continue to use the same rule as before when computing the p-value for a twosided test: double the single tail area, which remains a reasonable approach even when the sampling distribution is asymmetric. However, this can result in p-values larger than 1 when the point estimate is very near the mean in the null distribution; in such cases, we write that the p-value is 1. Also, very large p-values computed in this way (e.g. 0.85), may also be slightly inated.

Exercise 6.48 said the p-value is estimated. It is not exact because the simulated null distribution itself is not exact, only a close approximation. However, we can generate an exact null distribution and p-value using the binomial model from Section 3.4.

Generating the exact null distribution and p-value

The number of successes in n independent cases can be described using the binomial model, which was introduced in Section 3.4. Recall that the probability of observing exactly k successes is given by

$$P(\text{k successes}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{6.50}$$

where p is the true probability of success. The expression $\binom{n}{k}$ is read as n choose k, and the exclamation points represent factorials. For instance, 3! is equal to $3 \times 2 \times 1 = 6$, $4!$ is equal to $4 \times 3 \times 2 \times 1 = 24$, and so on (see Section 3.4).

The tail area of the null distribution is computed by adding up the probability in Equation (6.50) for each k that provides at least as strong of evidence favoring the alternative hypothesis as the data. If the hypothesis test is one-sided, then the p-value is represented by a single tail area. If the test is two-sided, compute the single tail area and double it to get the p-value, just as we have done in the past.

Example 11.1.5.1

Compute the exact p-value to check the consultant's claim that her clients' complication rate is below 10%.

Solution

Exactly k = 3 complications were observed in the n = 62 cases cited by the consultant. Since we are testing against the 10% national average, our null hypothesis is p = 0.10. We can compute the p-value by adding up the cases where there are 3 or fewer complications:

$$\text{p-value} = \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \tag{11.1.5.1}$$

$$= \sum_{j=0}^{3} \binom{62}{j} 0.1^j (1-0.1)^{62-j} \tag{11.1.5.2}$$

$$= \binom{62}{0} 0.1^0 (1-0.1)^{62-0} + \binom{62}{1} 0.1^1 (1-0.1)^{62-1} + \binom{62}{2} 0.1^2 (1-0.1)^{62-2} + \binom{62}{3} 0.1^3 (1-0.1)^{62-3} \tag{11.1.5.3}$$

$$= 0.0015 + 0.0100 + 0.0340 + 0.0755 \tag{11.1.5.4}$$

$$= 0.1210 \tag{11.1.5.5}$$

This exact p-value is very close to the p-value based on the simulations (0.1222), and we come to the same conclusion. We do not reject the null hypothesis, and there is not statistically signi cant evidence to support the association.

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 6.20 on page 290.

Using simulation for goodness of fit tests

Simulation methods may also be used to test goodness of t. In short, we simulate a new sample based on the purported bin probabilities, then compute a chi-square test statistic X^2_{sim}. We do this many times (e.g. 10,000 times), and then examine the distribution of these simulated chi-square test statistics. This distribution will be a very precise null distribution for the test statistic X^2 if the probabilities are accurate, and we can nd the upper tail of this null distribution, using a cutoff of the observed test statistic, to calculate the p-value.

Example 11.1.5.1

Section 6.3 introduced an example where we considered whether jurors were racially representative of the population. Would our ndings differ if we used a simulation technique?

Solution

Since the minimum bin count condition was satis ed, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 6.21 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

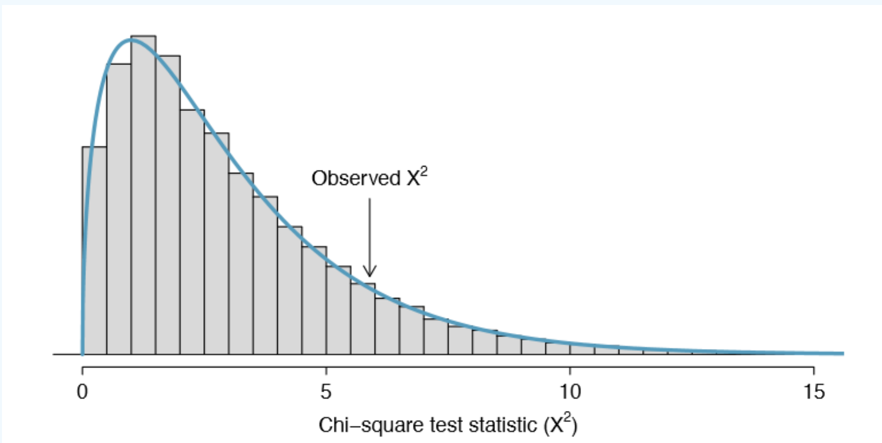


Figure 6.21: The precise null distribution for the juror example from Section 6.3 is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.

11.1.6: Randomization Test (Special Topic)

Cardiopulmonary resuscitation (CPR) is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. This procedure is helpful in maintaining some blood circulation, but the chest compressions involved can also cause internal injuries. Internal bleeding and other injuries complicate additional treatment efforts following arrival at a hospital. For instance, blood thinners may be used to help release a clot that is causing the heart attack. However, the blood thinner would negatively affect an internal injury. Here we consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.(Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial, by Bottiger et al., The Lancet, 2001.) These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive the blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

Example 11.1.6.1

Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival proportion in the control group and p_t represent the survival proportion for the treatment group.

Solution

We are interested in whether the blood thinners are helpful or harmful, so this should be a two-sided test.

- H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group.

$$p_t - p_c = 0. \tag{11.1.6.1}$$

- H_A: Blood thinners do have an impact on survival.

$$p_t - p_c \neq 0. \tag{11.1.6.2}$$

Large Sample Framework for a Difference in Two Proportions

There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did. The study results are shown in Table 6.22.

Table 6.22: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Exercise 11.1.6.1

What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$.

Solution

Observed control survival rate:

$$p_c = \frac{11}{50} = 0.22. \tag{11.1.6.3}$$

Treatment survival rate:

$$p_t = \frac{14}{40} = 0.35. \tag{11.1.6.4}$$

Observed difference:

$$\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13. \tag{11.1.6.5}$$

According to the point estimate, there is a 13% increase in the survival proportion when patients who have undergone CPR outside of the hospital are treated with blood thinners. However, we wonder if this difference could be due to chance. We'd like to investigate this using a large sample framework, but we first need to check the conditions for such an approach.

Example 11.1.6.2

Can the point estimate of the difference in survival proportions be adequately modeled using a normal distribution?

Solution

We will assume the patients are independent, which is probably reasonable. The success-failure condition is also satisfied. Since the proportions are equal under the null, we can compute the pooled proportion,

$$\hat{p} = \frac{(11+14)}{(50+40)} = 0.278, \tag{11.1.6.6}$$

for checking conditions. We nd the expected number of successes (13.9, 11.1) and failures (36.1, 28.9) are above 10. The normal model is reasonable.

While we can apply a normal framework as an approximation to nd a p-value, we might keep in mind that the expected number of successes is only 13.9 in one group and 11.1 in the other. Below we conduct an analysis relying on the large sample normal theory. We will follow up with a small sample analysis and compare the results.

Example 11.1.6.3

Assess the hypotheses presented in Example 6.53 using a large sample framework. Use a significance level of $\alpha = 0.05$.

Solution

We suppose the null distribution of the sample difference follows a normal distribution with mean 0 (the null value) and a standard deviation equal to the standard error of the estimate. The null hypothesis in this case would be that the two proportions are the same, so we compute the standard error using the pooled standard error formula from Equation (6.16) on page 273:

$$SE = \sqrt{\frac{p(1-p)}{n_t} + \frac{p(1-p)}{n_c}} \approx \sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}} = 0.095 \tag{11.1.6.7}$$

where we have used the pooled estimate $\left(\hat{p} = \frac{11+14}{50+40} = 0.278\right)$ in place of the true proportion, p.

The null distribution with mean zero and standard deviation 0.095 is shown in Figure 6.23. We compute the tail areas to identify the p-value. To do so, we use the Z score of the point estimate:

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - \text{null value}}{SE} = \frac{0.13-0}{0.095} = 1.37 \tag{11.1.6.8}$$

If we look this Z score up in Appendix B.1, we see that the right tail has area 0.0853. The p-value is twice the single tail area: 0.176. This p-value does not provide convincing evidence that the blood thinner helps. Thus, there is insufficient evidence to conclude whether or not the blood thinner helps or hurts. (Remember, we never "accept" the null hypothesis - we can only reject or fail to reject.)

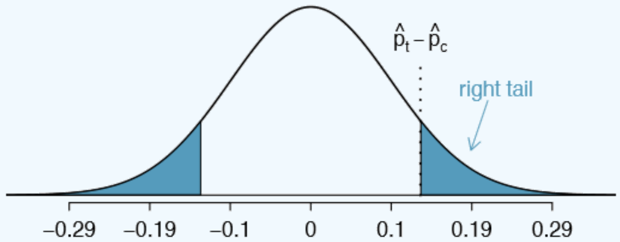


Figure 6.23: The null distribution of the point estimate $\hat{p}_t - \hat{p}_c$ under the large sample framework is a normal distribution with mean 0 and standard deviation equal to the standard error, in this case SE = 0.095. The p-value is represented by the shaded areas.

The p-value 0.176 relies on the normal approximation. We know that when the samples sizes are large, this approximation is quite good. However, when the sample sizes are relatively small as in this example, the approximation may only be adequate. Next we develop a simulation technique, apply it to these data, and compare our results. In general, the small sample method we develop may be used for any size sample, small or large, and should be considered as more accurate than the corresponding large sample technique.

Simulating a Difference under the null Distribution

The ideas in this section were first introduced in the optional Section 1.8. Suppose the null hypothesis is true. Then the blood thinner has no impact on survival and the 13% difference was due to chance. In this case, we can simulate null differences that are due to chance using a randomization technique. (The test procedure we employ in this section is formally called a **permutation test**). By randomly assigning "fake treatment" and "fake control" stickers to the patients' files, we could get a new grouping - one that is completely due to chance. The expected difference between the two proportions under this simulation is zero.

We run this simulation by taking 40 treatment fake and 50 control fake labels and randomly assigning them to the patients. The label counts of 40 and 50 correspond to the number of treatment and control assignments in the actual study. We use a computer program to randomly assign these labels to the patients, and we organize the simulation results into Table 6.24.

Table 6.24: Simulated results for the CPR study under the null hypothesis. The labels were randomly assigned and are independent of the outcome of the patient.

	Survived	Died	Total
Control_fake	15	35	50
Treatment_fake	10	30	40
Total	25	65	90

Exercise 11.1.6.2

What is the difference in survival rates between the two fake groups in Table 6.24? How does this compare to the observed 13% in the real groups?

Solution

The difference is $\hat{p}_{\text{t;fake}} - \hat{p}_{\text{c;fake}} = \frac{10}{40} - \frac{15}{50} = -0.05$, which is closer to the null value p0 = 0 than what we observed.

The difference computed in Exercise 6.57 represents a draw from the null distribution of the sample differences. Next we generate many more simulated experiments to build up the null distribution, much like we did in Section 6.5.2 to build a null distribution for a one sample proportion.

Caution: Simulation in the two proportion case requires that the null difference is zero

The technique described here to simulate a difference from the null distribution relies on an important condition in the null hypothesis: there is no connection between the two variables considered. In some special cases, the null difference might not be zero, and more advanced methods (or a large sample approximation, if appropriate) would be necessary.

Null distribution for the difference in two proportions

We build up an approximation to the null distribution by repeatedly creating tables like the one shown in Table 6.24 and computing the sample differences. The null distribution from 10,000 simulations is shown in Figure 6.25.

Example 11.1.6.4

Compare Figures 6.23 and 6.25. How are they similar? How are they different?

Solution

The shapes are similar, but the simulated results show that the continuous approximation of the normal distribution is not very good. We might wonder, how close are the p-values?

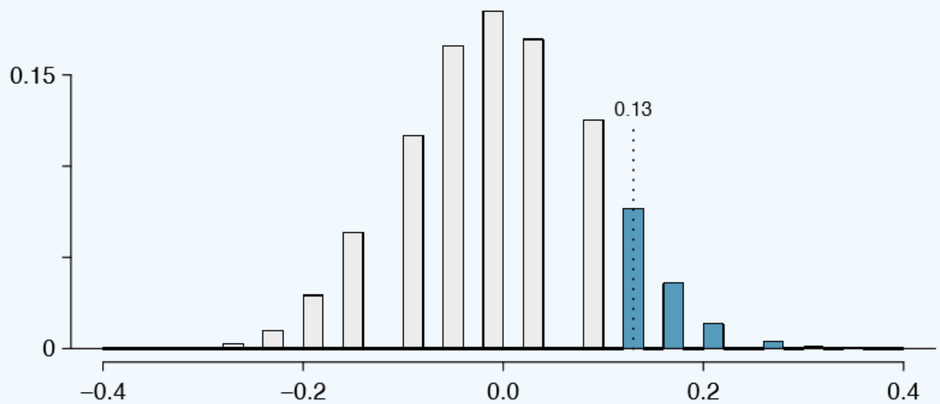


Figure 6.25: An approximation of the null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The p-value is twice the right tail area.

Exercise 11.1.6.3

The right tail area is about 0.13. (It is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) The p-value is computed by doubling the right tail area: 0.26. How does this value compare with the large sample approximation for the p-value?

Solution

The approximation in this case is fairly poor (p-values: 0.174 vs. 0.26), though we come to the same conclusion. The data do not provide convincing evidence showing the blood thinner helps or hurts patients.

In general, small sample methods produce more accurate results since they rely on fewer assumptions. However, they often require some extra work or simulations. For this reason, many statisticians use small sample methods only when conditions for large sample methods are not satisfied.

Randomization for two-way tables and chi-square

Randomization methods may also be used for the contingency tables. In short, we create a randomized contingency table, then compute a chi-square test statistic X^2_{sim}. We repeat this many times using a computer, and then we examine the distribution of these simulated test statistics. This randomization approach is valid for any sized sample, and it will be more accurate for cases where one or more expected bin counts do not meet the minimum threshold of 5. When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution. As before, we use the upper tail of the null distribution to calculate the p-value.

11.1.7: Exercises

Inference for a single proportion

6.1 Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

(b) The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

(c) A random sample of 125 college students where 12% are vegetarians would be considered unusual.

(d) A random sample of 250 college students where 12% are vegetarians would be considered unusual.

(e) The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

6.2 Young Americans, Part I. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.[36]

(a) The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.

(b) The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

(c) A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

(d) A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

6.3 Orange tabbies. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

6.4 Young Americans, Part II. About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning.[37]

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

(b) In order for the the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

[36]*A. Vaughn. "Poll finds young adults optimistic, but not about money". In: Los Angeles Times (2011).*

[37]*Demos.org. "The State of Young America: The Poll". In: (2011).*

6.5 Prop 19 in California. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[38]

(a) We are 95% con dent that between 62% and 78% of the California voters in this sample support Prop 19.

(b) We are 95% con dent that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.

(c) If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.

(d) In order to decrease the margin of error to 4%, we would need to quadruple (multiply by 4) the sample size.

(e) Based on this con dence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[39]

(a) We are 95% con dent that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b) We are 95% con dent that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

(d) The margin of error at a 90% con dence level would be higher than 3%.

6.7 Fireworks on July 4th. In late June 2012, Survey USA published results of a survey stating that 56% of the 600 randomly sampled Kansas residents planned to set off reworks on July 4th. Determine the margin of error for the 56% point estimate using a 95% con dence level.40

6.8 Elderly drivers. In January 2011, The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% con dence level.[41]

(a) Verify the margin of error reported by The Marist Poll.

(b) Based on a 95% con dence interval, does the poll provide convincing evidence that more than 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?

[38]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

[39]*Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.*

[40]*Survey USA, News Poll #19333, data collected on June 27, 2012.*

[41]*Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.*

6.9 Life after college. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and nd out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions for constructing a con dence interval based on these data are met.

(c) Calculate a 95% con dence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

(d) What does "95% confidence" mean?

(e) Now calculate a 99% con dence interval for the same parameter and interpret it in the context of the data.

(f) Compare the widths of the 95% and 99% con dence intervals. Which one is wider? Explain.

6.10 Life rating in Greece. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered

"suffering".[42]

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions required for constructing a con dence interval based on these data are met.

(c) Construct a 95% con dence interval for the proportion of Greeks who are "suffering".

(d) Without doing any calculations, describe what would happen to the con dence interval if we decided to use a higher confidence level.

(e) Without doing any calculations, describe what would happen to the con dence interval if we used a larger sample.

6.11 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[43]

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% con dence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

(c) What does "90% confidence" mean?

(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

[42]Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.

[43]studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[44]

(a) Is 48% a sample statistic or a population parameter? Explain.

(b) Construct a 95% con dence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

(c) A critic points out that this 95% con dence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

(d) A news piece on this survey's ndings states, \Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

6.13 Public option, Part I. A Washington Post article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More speci cally, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.[45]

(a) A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

(b) Would you expect a con dence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

6.14 The Civil War. A national survey conducted in 2011 among a simple random sample of 1,507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life.[46]

(a) Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) Interpret the p-value in this context.

(c) Calculate a 90% con dence interval for the proportion of Americans who think the Civil War is still relevant. Interpret the interval in this context, and comment on whether or not the con dence interval agrees with the conclusion of the hypothesis test.

6.15 Browsing on the mobile device. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.[47]

(a) According to an online article, a report from a mobile research company indicates that 38 percent of Chinese mobile web users only access the internet through their cell phones.48 Conduct a hypothesis test to determine if these data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%.

(b) Interpret the p-value in this context.

(c) Calculate a 95% con dence interval for the proportion of Americans who access the internet on their cell phones, and interpret the interval in this context.

[44] *National Opinion Research Center, General Social Survey, 2010.*

[45] *D. Balz and J. Cohen. "Most support public option for health insurance, poll nds". In: The Washington Post (2009).*

[46] *Pew Research Center Publications, Civil War at 150: Still Relevant, Still Divisive, data collected between March 30 - April 3, 2011.*

[47] *Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.*

[48] *S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: M.I.C Gadget (2012).*

6.16 Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.[49]

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a con dence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

6.17 Taste test. Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then lled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identi ed the soda.

(a) Do these data provide strong evidence that these people are able to detect the difference between diet and regular soda, in other words, are the results signi cantly better than just random guessing?

(b) Interpret the p-value in this context.

6.18 Is college worth it? Part II. Exercise 6.16 presents the results of a poll where 48% of 331 Americans who decide to not go to college do so because they cannot afford it.

(a) Calculate a 90% con dence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

(b) Suppose we wanted the margin of error for the 90% con dence level to be about 1.5%. How large of a survey would you recommend?

6.19 College smokers. We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke.

(a) Calculate a 95% con dence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: check conditions)

(b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

6.21 Public option, Part II. Exercise 6.13 presents the results of a poll evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.22 Acetaminophen and liver damage. It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher wants to conduct a study to estimate the proportion of acetaminophen users who have liver damage. For participating in this study, he will pay each subject $20 and provide a free medical consultation if the patient has liver damage.

(a) If he wants to limit the margin of error of his 98% confidence interval to 2%, what is the minimum amount of money he needs to set aside to pay his subjects?

(b) The amount you calculated in part (a) is substantially over his budget so he decides to use fewer subjects. How will this affect the width of his con dence interval?

[49] *Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.*

Difference of two proportions

6.23 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

6.24 Heart transplant success. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might bene t from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.[50]

	Control	treatment
alive dead	4 30	24 45

A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate a con dence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

6.25 Gender and color preference. A 2001 study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.[51]

(a) We are 95% con dent that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% con dent that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% con dence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a signi cant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% con dence interval for $(p_{female} - p_{male})$ cannot be calculated with only the information given in this exercise.

[50]*B. Turnbull et al. "Survivorship of Heart Transplant Data". In: Journal of the American Statistical Association 69 (1974), pp. 74 - 80.*

[51]*L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: Personality and Individual Differences 31.8 (2001), pp. 1375-1379.*

6.26 The Daily Show. A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. A 95% con dence interval for $(p_{collegegrad} - p_{HSorless})$, where p is the proportion of those who watch The Daily Show, is (0.07, 0.15). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[52]

(a) At the 5% significance level, the data provide convincing evidence of a difference between the proportions of college graduates and those with a high school degree or less who watch The Daily Show.

(b) We are 95% con dent that 7% less to 15% more college graduates watch The Daily Show than those with a high school degree or less.

(c) 95% of random samples of 1,099 college graduates and 1,110 people with a high school degree or less will yield differences in sample proportions between 7% and 15%.

(d) A 90% confidence interval for $(p_{collegegrad} - p_{HSorless})$ would be wider.

(e) A 95% confidence interval for $(p_{HSorless} - p_{collegegrad})$ is (-0.15,-0.07).

6.27 Public Option, Part III. Exercise 6.13 presents the results of a poll evaluating support for the health care public option plan in 2009. 70% of 819 Democrats and 42% of 783 Independents support the public option.

(a) Calculate a 95% confidence interval for the difference between $(p_D - p_I)$ and interpret it in this context. We have already checked conditions for you.

(b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the public option than the Independent.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% con dence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[53]

6.29 Offshore drilling, Part I. A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[54]

	College Grad	
	Yes	No

Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

[52]*The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.*

[53]*CDC, Perceived Insu�cient Rest or Sleep Among Adults - United States, 2008.*

[54]*Survey USA, Election Poll #16804, data collected July 8-11, 2010.*

6.30 Sleep deprivation, CA vs. OR, Part II. Exercise 6.28 provides data on sleep deprivation rates of Californians and Oregonians. The proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Conduct a hypothesis test to determine if these data provide strong evidence the rate of sleep deprivation is different for the two states. (Reminder: check conditions)

(b) It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

6.31 Offshore drilling, Part II. Results of a poll evaluating support for drilling for oil and natural gas off the coast of California were introduced in Exercise 6.29.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample support drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who support offshore drilling in California is different than that of noncollege graduates.

6.32 Full body scan, Part I. A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1,137 adults nationwide, interviewed by telephone November 7-10, 2010, where one of the questions on the survey was "Some airports are now using `full-body' digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?" Below is a summary of responses based on party affiliation.[55]

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22

Total	318	369	450

(a) Conduct an appropriate hypothesis test evaluating whether there is a difference in the proportion of Republicans and Democrats who think the full-body scans should be applied in airports. Assume that all relevant conditions are met.

(b) The conclusion of the test in part (a) may be incorrect, meaning a testing error was made. If an error was made, was it a Type I or a Type II error? Explain.

[55] *S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: CBS News (2010).*

6.33 Sleep deprived transportation workers. The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers. The results of the survey are shown below.[56]

		Transportation	Professionals		
	Control	Pilots	Truck Drivers	Train Operators	Bux/Taxi/Limo Drivers
Less than 6 hours of sleep 6 to 8 hours of sleep More than 8 hours	35 193 64	19 132 51	35 117 51	29 119 32	21 131 58
Tota	292	202	203	180	210

Conduct a hypothesis test to evaluate if these data provide evidence of a difference between the proportions of truck drivers and non-transportation workers (the control group) who get less than 6 hours of sleep per day, i.e. are considered sleep deprived.

6.34 Prenatal vitamins and Autism. Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period).[57]

	Autism		
	Autism	Typical development	Total
No vitamin Vitamin	111 143	70 159	181 302
Total	254	229	483

(a) State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.

(b) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

(c) A New York Times article reporting on this study was titled "Prenatal Vitamins May Ward Off Autism". Do you nd the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.[58]

6.35 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called virologic failure). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[59]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

[56]*National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers Sleep, 2012.*

[57]*R.J. Schmidt et al. \Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism". In: Epidemiology 22.4 (2011), p. 476.*

[58]*R.C. Rabin. "Patterns: Prenatal Vitamins May Ward Off Autism". In: New York Times (2011).*

[59]*S. Lockman et al. "Response to antiretroviral therapy after a single, peripartum dose of nevirapine". In: Obstetrical & gynecological survey 62.6 (2007), p. 361.*

6.36 Diabetes and unemployment. A 2012 Gallup poll surveyed Americans about their employment status and whether or not they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.[60]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of incidence of diabetes and employment status.

(c) The sample difference is about 1%. If we completed the hypothesis test, we would nd that the p-value is very small (about 0), meaning the difference is statistically signi cant. Use this result to explain the difference between statistically signi cant and practically significant findings.

Testing for goodness of t using chi-square

6.37 True or false, Part I. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

(c) The chi-square statistic is always positive.

(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

6.38 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found X^2 = 10 with df = 5 you would fail to reject H0 at the 5% signi cance level.

(c) When nding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.39 Open source textbook. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

(a) State the hypotheses for testing if the professor's predictions were inaccurate.

(b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

(c) This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

(d) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

(e) Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

[60]*Gallup Wellbeing, Employed Americans in Better Health Than the Unemployed, data collected Jan. 2, 2011 - May 21, 2012.*

6.40 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[61]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1) Humans evolved, but God had no part in process (2) God created humans in present form (3) Other / No opinion (4)	38% 16% 40% 6%	37% 12% 45% 6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)

Testing for independence in two-way tables

6.41 Quitters. Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

(a) Create a two-way table presenting the results of this study.

(b) Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

i. How many subjects in the "patch + support" group would you expect to quit?

ii. How many subjects in the "only patch" group would you expect to not quit?

6.42 Full body scan, Part II. The table below summarizes a data set we rst encountered in Exercise 6.32 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's party affiliation and his support of full-body scans. It may be useful to rst add on an extra column for row totals before proceeding with the computations.

	Party Affiliation		
	Republican	Democrat	Independent
Should Should not Don't know/No answer	264 38 16	299 55 15	351 77 22
Total	318	369	450

(a) How many Republicans would you expect to not support the use of full-body scans?

(b) How many Democrats would you expect to support the use of full-body scans?

(c) How many Independents would you expect to not know or not answer?

[61] *Four in 10 Americans Believe in Strict Creationism, December 17, 2010, http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx.*

6.43 Offshore drilling, Part III. The table below summarizes a data set we rst encountered in Exercise 6.29 that examines the responses of a random sample of college graduates and nongraduates on the topic of oil drilling. Complete a chi-square test for

these data to check whether there is a statistically signi cant difference in responses from college graduates and non-graduates.

	College Grad	
	Yes	No
Support Oppose Do not know	154 180 104	132 126 131
Total	438	389

6.44 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[62]

		Caffeinated	coffee	consumption		
	$\leq$ 1cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Yes No	670 11,545	373 6,244	905 16,329	564 11,726	95 2,288	2,607 48,132
Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identi ed in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $\frac{(Observed-Expected)^2}{Expected}$.

(e) The test statistic is $X^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.63 Do you agree with this statement? Explain your reasoning.

[62]*M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: Archives of internal medicine 171.17 (2011), p. 1571.*

[63]*A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: New York Times (2011).*

6.45 Privacy on Facebook. A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, "Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?" The responses are cross-tabulated based on gender.[64]

	Gender		
	Male	Female	Total
Yes No Not sure	288 61 10	378 62 7	666 123 17
Total	359	447	806

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

6.46 Shipping holiday gifts. A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age		
	18-34	35-54	55+	Total
USPS	72 (81)	97 (102)	76 (62)	245
UPS	52 (53)	76 (68)	34 (41)	162
FedEx	31 (21)	24 (27)	9 (16)	64
Something else	7 (5)	6 (7)	3 (4)	16
Not sure	3 (5)	6 (5)	4 (3)	13
Total	165	209	126	500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

(b) Are the conditions for inference using a chi-square test satisfied?

Small sample hypothesis testing for a proportion

6.47 Bullying in schools. A 2012 Survey USA poll asked Florida residents how big of a problem they thought bullying was in local schools. 9 out of 191 18-34 year olds responded that bullying is no problem at all. Using these data, is it appropriate to construct a con dence interval using the formula $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for the true proportion of 18-34 year old Floridians who think bullying is no problem at all? If it is appropriate, construct the con dence interval. If it is not, explain why.

[64]*Survey USA, News Poll #17960, data collected February 16-17, 2011.*

6.48 Choose a test. We would like to test the following hypotheses:

H_0 : p = 0.1

$H_A : p \neq 0.1$

The sample size is 120 and the sample proportion is 8.5%. Determine which of the below test(s) is/are appropriate for this situation and explain your reasoning.

I. Z test for a proportion,

i.e. proportion test using normal model

II. Z test for comparing two proportions

III. X^2 test of independence

IV. Simulation test for a proportion

V. t test for a mean

VI. ANOVA

6.49 The Egyptian Revolution. A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the rst couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.[65]

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

(c) Based on large sample theory, we modeled ^p using the normal distribution. Why should we be cautious about this approach for these data?

(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.

(e) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.

[65]*Gallup Politics, Americans' Views of Egypt Sharply More Negative, data collected February 2-5, 2011.*

6.50 Assisted Reproduction. Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g. in vitro fertilization). A 2008 report by the Centers for Disease Control and Prevention estimated that ART has been successful in leading to a live birth in 31% of cases66. A new fertility clinic claims that their success rate is higher than average. A random sample of 30 of their patients yielded a success rate of 40%. A consumer watchdog group would like to determine if this provides strong evidence to support the company's claim.

(a) Write the hypotheses to test if the success rate for ART at this clinic is signi cantly higher than the success rate reported by the CDC.

(b) Based on large sample theory, we modeled ^p using the normal distribution. Why is this not appropriate here?

(c) The normal approximation would be less reliable here, so we should use a simulation strategy. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

(d) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

(e) After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.

[66]*CDC. 2008 Assisted Reproductive Technology Report.*

Hypothesis testing for two proportions

6.51 Social experiment, Part II. Exercise 6.23 introduces a "social experiment" conducted by a TV program that questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

	Scenario		
	Provocative	Conservative	Total
Yes No	5 15	15 10	20 25
Total	20	25	45

A simulation was conducted to test if people react differently under the two scenarios. 10,000 simulated differences were generated to construct the null distribution shown. The value $\hat{p}_{pr;sim}$ represents the proportion of diners who intervened in the simulation for the provocatively dressed woman, and $\hat{p}_{con;sim}$ is the proportion for the conservatively dressed woman.

(a) What are the hypotheses? For the purposes of this exercise, you may assume that each observed person at the restaurant behaved independently, though we would want to evaluate this assumption more rigorously if we were reporting these results.

(b) Calculate the observed difference between the rates of intervention under the provocative and conservative scenarios: $\hat{p}_{pr} - \hat{p}_{con}$.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

6.52 Is yawning contagious? An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously inuenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.[67]

	Group		
	Treatment	Control	Total
Yawn Not Yawn	10 24	4 12	14 36
Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no inuence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning as $\hat{p}_{trtmt;sim} - \hat{p}_{ctrl;sim}$. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.

(a) What are the hypotheses?

(b) Calculate the observed difference between the yawning rates under the two scenarios.

(c) Estimate the p-value using the gure above and determine the conclusion of the hypothesis test.

[67]*MythBusters, Season 3, Episode 28.*

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

SECTION OVERVIEW

11.2: The Chi-Square Distribution

A chi-squared test is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true.

11.2.1: Facts About the Chi-Square Distribution

The notation for the chi-square distribution is:

$$\chi \sim \chi^2_{df} \tag{11.2.1.1}$$

where $df =$ degrees of freedom which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is

$$\sigma = \sqrt{2(df)}. \tag{11.2.1.2}$$

The random variable is shown as χ^2, but may be any upper case letter. The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + \ldots + (Z_k)^2 \tag{11.2.1.3}$$

a. The curve is nonsymmetrical and skewed to the right.
b. There is a different chi-square curve for each df.

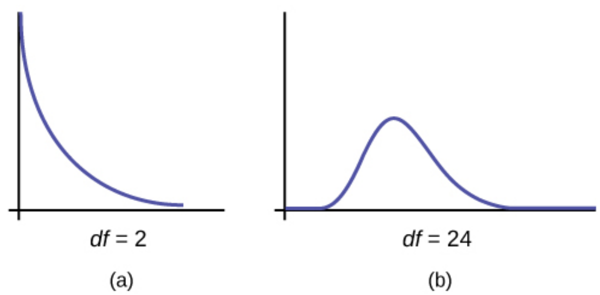


Figure 11.2.1.1

c. The test statistic for any test is always greater than or equal to zero.
d. When $df > 90$, the chi-square curve approximates the normal distribution. For $\chi \sim \chi^2_{1,000}$ the mean, $\mu = df = 1,000$ and the standard deviation, $\mu = \sqrt{2(1,000)}$. Therefore, $X \sim N(1,000, 44.7)$, approximately.
e. The mean, μ, is located just to the right of the peak.

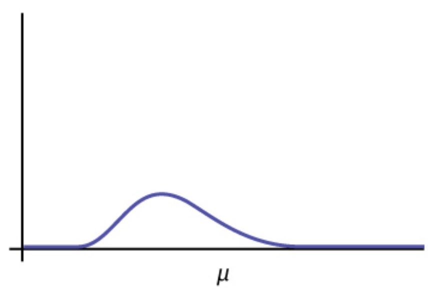


Figure 11.2.1.2

References

1. Data from *Parade Magazine*.
2. “HIV/AIDS Epidemiology Santa Clara County.”Santa Clara County Public Health Department, May 2011.

Review

The chi-square distribution is a useful tool for assessment in a series of problem categories. These problem categories include primarily (i) whether a data set fits a particular distribution, (ii) whether the distributions of two populations are the same, (iii) whether two events might be independent, and (iv) whether there is a different variability than expected within a population.

An important parameter in a chi-square distribution is the degrees of freedom df in a given problem. The random variable in the chi-square distribution is the sum of squares of *df* standard normal variables, which must be independent. The key characteristics of the chi-square distribution also depend directly on the degrees of freedom.

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom df. For $df > 90$, the curve approximates the normal distribution. Test statistics based on the chi-square distribution are always greater than or equal to zero. Such application tests are almost always right-tailed tests.

Formula Review

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_{df})^2 \tag{11.2.1.4}$$

chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2(df)}$ Chi-Square distribution population standard deviation

? Exercise 11.2.1.1

If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

Answer

mean $= 25$ and standard deviation $= 7.0711$

? Exercise 11.2.1.2

If $df > 90$, the distribution is _____________. If $df = 15$, the distribution is ________________.

? Exercise 11.2.1.3

When does the chi-square curve approximate a normal distribution?

Answer

when the number of degrees of freedom is greater than 90

? Exercise 11.2.1.4

Where is μ located on a chi-square curve?

? Exercise 11.2.1.5

Is it more likely the *df* is 90, 20, or two in the graph?

Figure 11.2.1.3.

Answer

$df = 2$

11.2.2: Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data "fit" a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.

The test statistic for a goodness-of-fit test is:

$$\sum_{k} \frac{(O-E)^2}{E} \tag{11.2.2.1}$$

where:

- $O =$ observed values (data)
- $E =$ expected values (from theory)
- $k =$ the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is $df = (\text{number of categories} - 1)$.

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

> The expected value for each cell needs to be at least five in order for you to use this test.

Example 11.3.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to the table below.

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9–11	6
12+	2

A random survey across all mathematics courses was then done to determine the actual number **(observed)** of absences in a course. The chart in the table below displays the results of that survey.

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

- H_0: Student absenteeism **fits** faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

- H_a: Student absenteeism **does not fit** faculty perception.

Exercise 11.2.2.1.1

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Answer

a. **No.** Notice that the expected number of absences for the "12+" entry is less than five (it is two). Combine that group with the "9–11" group to create new tables where the number of students for each entry are at least five. The new results are in the table below.

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

Exercise 11.2.2.1.2

b. What is the number of degrees of freedom (df)?

Answer

b. There are four "cells" or categories in each of the new tables.

$df = \text{number of cells} - 1 = 4 - 1 = 3$

Exercise 11.2.2.1

A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in the table below.

Number produced	Number defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

A random sample was taken to determine the actual number of defects. The table below shows the results of the survey.

Number produced	Number defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

Answer

H_0:The number of defects fits expectations.

H_a:The number of defects does not fit expectations.

$df = 4$

Example 11.3.2

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the table below. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

Day of the Week Employees were Most Absent

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

Answer

The null and alternative hypotheses are:

- H_0: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: $15+12+9+9+15=60$), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values $(12, 12, 12, 12, 12)$
- Observed (O) values $(15, 12, 9, 9, 15)$
- $(O\text{–}E)$
- $(O\text{–}E)^2$
- $\frac{(O-E)^2}{E}$

Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

To find the p-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. (Use a computer or calculator to find the p-value. You should get p-value $= 0.5578$.)

The dfs are the number of cells $-1 = 5-1 = 4$

Press `2nd DISTR` . Arrow down to χ^2cdf. Press `ENTER` . Enter `(3,10^99,4)` . Rounded to four decimal places, you should see 0.5578, which is the p-value.

Next, complete a graph like the following one with the proper labeling and shading. (You should shade the right tail.)

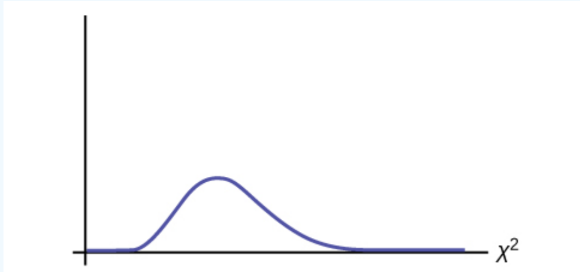


Figure 11.2.2.1.

The decision is not to reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example Example has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF` . To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF` . Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw` . Make sure you clear any lists before you start. **To Clear Lists in the calculators:** Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Alternatively, you can press `STAT` and press 4 (for `ClrList`). Enter the list name and press `ENTER` .

Exercise 11.2.2.2

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 49 students were asked on which night of the week they did the most homework. The results were distributed as in the table below.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Answer

$df = 6$

$p\text{-value} = 0.6093$

We decline to reject the null hypothesis. There is not enough evidence to support that students do not do the majority of their homework equally throughout the week.

Example 11.3.3

One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as in the table below.

Number of Televisions	Percent

Number of Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the data in the table below.

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (O) frequency values.

Exercise 11.2.2.3.1

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Answer

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are shown in the table below.

Number of Televisions	Percent	Expected Frequency
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
over 3	8	(0.08)(600) = 48

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter $0.10 * 600$.

H_0: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where $df = (\text{the number of cells}) - 1 = 5 - 1 = 4$.

Note 11.3.3.1

$df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:

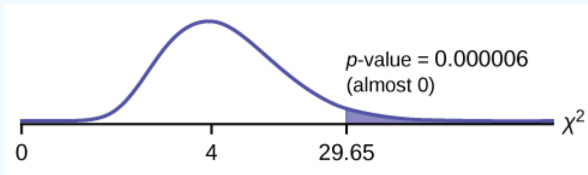


Figure 11.2.2.2.

Probability statement: $p\text{-value} = P(\chi^2 > 29.65) = 0.000006$

Compare *α* and the *p*-value:

$\alpha = 0.01$

$p\text{-value} = 0.000006$

So, $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of Example). Into `L1`, put the observed frequencies `66`, `119`, `349`, `60`, `15`. Into `L2`, put the expected frequencies `.10*600,` `.16*600`, `.55*600`, `.11*600`, `.08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"` `(Enter L3)`. Rounded to 2 decimal places, you should see `29.65`. Press `2nd DISTR`. Press `7` or Arrow down to `7:χ2cdf` and press `ENTER`. Enter `(29.65,1E99,4)`. Rounded to four places, you should see `5.77E-6 = .000006` (rounded to six decimal places), which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

Exercise 11.2.2.3

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in the table below.

Number of Pets	Percent
0	18
1	25
2	30

Number of Pets	Percent
3	18
4+	9

A random sample of 1,000 students from the Eastern United States resulted in the data in the table below.

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

At the 1% significance level, does it appear that the distribution "number of pets" of students in the Eastern United States is different from the distribution for the United States student population as a whole? What is the p-value?

Answer

$p\text{-value} = 0.0036$

We reject the null hypothesis that the distributions are the same. There is sufficient evidence to conclude that the distribution "number of pets" of students in the Eastern United States is different from the distribution for the United States student population as a whole.

✓ Example 11.3.4

Suppose you flip two coins 100 times. The results are 20 *HH*, 27 *HT*, 30 *TH*, and 23 *TT*. Are the coins fair? Test at a 5% significance level.

Answer

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is HH, HT, TH, TT. Out of 100 flips, you would expect 25 *HH*, 25 *HT*, 25 *TH*, and 25 *TT*. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins $(20HH, 27HT, 30TH, 23TT)$ fit the expected distribution?"

Random Variable: Let $X =$ the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the **number of cells is three**. Since $X =$ the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

H_0: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:

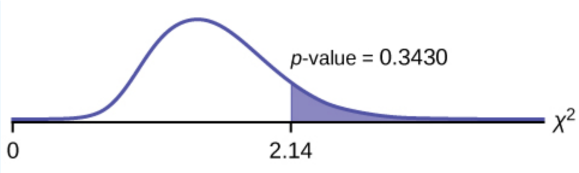


Figure 11.2.2.3.

Probability statement: $p\text{-value} = P(\chi^2 > 2.14) = 0.3430$

Compare α and the p-value:

$\alpha = 0.05$

$p\text{-value} = 0.3430$

$\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Press `STAT` and `ENTER`. Make sure you clear lists `L1`, `L2`, and `L3` if they have data in them. Into `L1`, put the observed frequencies `20`, `57`, `23`. Into `L2`, put the expected frequencies `25`, `50`, `25`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`. `Enter L3`. Rounded to two decimal places, you should see `2.14`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press `7`). Press `ENTER`. Enter `2.14,1E99,2)`. Rounded to four places, you should see `.3430`, which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

Exercise 11.2.2.4

Students in a social studies class hypothesize that the literacy rates across the world for every region are 82%. The table below shows the actual literacy rates across the world broken down by region. What are the test statistic and the degrees of freedom?

MDG Region	Adult Literacy Rate (%)
Developed Regions	99.0
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91.0
Eastern Asia	93.8
Southern Asia	61.9
South-Eastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Answer

$df = 9$

χ^2 test statistic $= 26.38$

This is a nonsymmetric chi-square curve with df = 9. The values 0, 9, and 26.38 are labeled on the horizontal axis. The value 9 coincides with the peak of the curve. A vertical upward line extends from 26.38 to the curve, and the region to the right of this line is shaded. The shaded area is equal to the p-value.

Figure 11.2.2.4.

Press `STAT` and `ENTER`. Make sure you clear lists `L1, L2,` and `L3` if they have data in them. Into L1, put the observed frequencies `99, 99.5, 67.3, 62.5, 91, 93.8, 61.9, 91.9, 84.5, 66.4`. Into `L2`, put the expected frequencies `82, 82, 82, 82, 82, 82, 82, 82, 82, 82`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`. `Enter L3`. Rounded to two decimal places, you should see `26.38`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press `7`). Press `ENTER`. Enter `26.38,1E99,9)`. Rounded to four places, you should see `.0018`, which is the *p*-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

References

1. Data from the U.S. Census Bureau
2. Data from the College Board. Available online at http://www.collegeboard.com.
3. Data from the U.S. Census Bureau, Current Population Reports.
4. Ma, Y., E.R. Bertone, E.J. Stanek III, G.W. Reed, J.R. Hebert, N.L. Cohen, P.A. Merriam, I.S. Ockene, "Association between Eating Patterns and Obesity in a Free-living US Adult Population." *American Journal of Epidemiology* volume 158, no. 1, pages 85-92.
5. Ogden, Cynthia L., Margaret D. Carroll, Brian K. Kit, Katherine M. Flegal, "Prevalence of Obesity in the United States, 2009–2010." NCHS Data Brief no. 82, January 2012. Available online at http://www.cdc.gov/nchs/data/databriefs/db82.pdf (accessed May 24, 2013).
6. Stevens, Barbara J., "Multi-family and Commercial Solid Waste and Recycling Survey." Arlington Count, VA. Available online at www.arlingtonva.us/department.../file84429.pdf (accessed May 24,2013).

Review

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is almost always right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_k \frac{(O-E)^2}{E}$ goodness-of-fit test statistic where:

O: observed values

E: expected value

k: number of different data cells or categories

$df = k-1$ degrees of freedom

Determine the appropriate test to be used in the next three exercises.

? Exercise 11.2.2.5

An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

? Exercise 11.2.2.6

An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.

Answer

a goodness-of-fit test

? Exercise 11.2.2.7

A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises: A teacher predicts that the distribution of grades on the final exam will be and they are recorded in the table below.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

The actual distribution for a class of 20 is in the table below.

Grade	Frequency
A	7
B	7
C	5
D	1

? Exercise 11.2.2.8

$df =$ ______

Answer

3

? Exercise 11.2.2.9

State the null and alternative hypotheses.

? Exercise 11.2.2.10

$\chi^2 \text{test statistic} =$ ______

Answer

2.04

Exercise 11.2.2.11

$p\text{-value} = \_\_\_\_\_\_$

Exercise 11.2.2.12

At the 5% significance level, what can you conclude?

Answer

We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

Use the following information to answer the next nine exercises: The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in the table below.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

The percentage of each ethnic group in Santa Clara County is as in the table below.

Ethnicity	Percentage of total county population	Number expected (round to two decimal places)
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Exercise 11.2.2.13

If the ethnicities of AIDS victims followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.

Perform a goodness-of-fit test to determine whether the occurrence of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 11.2.2.14

H_0: _______

Answer

H_0: the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 11.2.2.15

H_a: _______

Exercise 11.2.2.16

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

right-tailed

Exercise 11.2.2.17

degrees of freedom = _______

Exercise 11.2.2.18

χ^2test statistic = _______

Answer

88,621

Exercise 11.2.2.19

p-value $=$ _______

Exercise 11.2.2.20

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

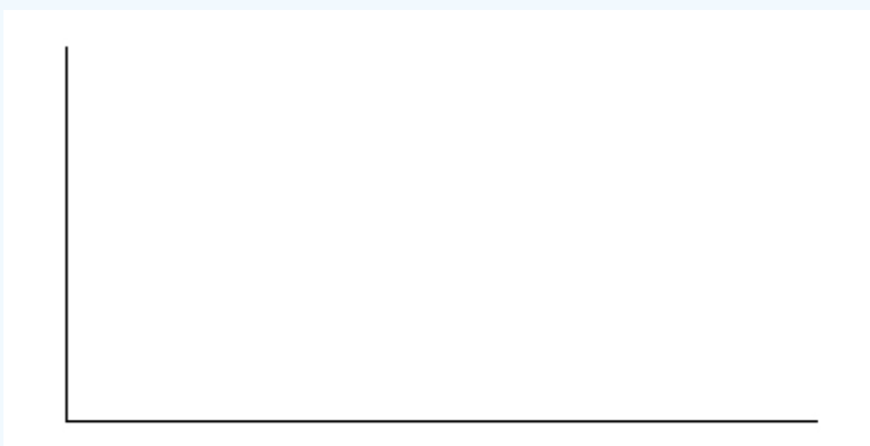


Figure 11.2.2.5.

Let $\alpha = 0.05$

Decision: ________________

Reason for the Decision: ________________

Conclusion (write out in complete sentences): ________________

Answer

Graph: Check student's solution.

Decision: Reject the null hypothesis.

Reason for the Decision: $p\text{-value} < \alpha$

Conclusion (write out in complete sentences): The make-up of AIDS cases does not fit the ethnicities of the general population of Santa Clara County.

Exercise 11.2.2.21

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

11.2.3: Test of Independence

Tests of independence involve using a contingency table of observed (data) values.

The test statistic for a *test of independence* is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{11.2.3.1}$$

where:

- $O =$ observed values
- $E =$ expected values
- $i =$ the number of rows in the table
- $j =$ the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

> The expected value for each cell needs to be at least five in order for you to use this test.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Probability Topics. As a review, consider the following example.

Example 11.2.3.1

Suppose $A =$ a speeding violation in the last year and $B =$ a cell phone user while driving. If A and B are independent then $P(A \text{ AND } B) = P(A)P(B)$. A AND B is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year, and who used cell phone while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not.

Let $y =$ expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right)$$

Solve for y:

$$y = \frac{(70)(305)}{755} = 28.3$$

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternative hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example, then the null hypothesis is:

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

$$df = (\text{number of columns} - 1)(\text{number of rows} - 1)$$

The following formula calculates the **expected number** (E):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

Exercise 11.2.3.1

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Answer

About 16 students are expected to be music students and on the honor roll.

Example 11.2.3.2

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. In Table 11.2.3.1is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Table 11.2.3.1: Number of Hours Worked Per Week by Volunteer Type (Observed). The table contains **observed (O)** values (data).

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Is the number of hours volunteered **independent** of the type of volunteer?

Answer

The **observed table** and the question at the end of the problem, "Is the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

- H_0: The number of hours volunteered is **independent** of the type of volunteer.
- H_a: The number of hours volunteered is **dependent** on the type of volunteer.

The expected results are in Table 11.2.3.2

Table 11.2.3.2: Number of Hours Worked Per Week by Volunteer Type (Expected). The table contains **expected**(E) values (data).

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

For example, the calculation for the expected frequency for the top left cell is

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)(298)}{839} = 90.57$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$$df = (3 \text{ columns} - 1)(3 \text{ rows} - 1) = (2)(2) = 4$$

Graph:

Nonsymmetrical chi-square curve with values of 0 and 12.99 on the x-axis representing the test statistic of number of hours worked by volunteers of different types. A vertical upward line extends from 12.99 to the curve and the area to the right of this is equal to the p-value.

Figure 11.2.3.1.

Probability statement: $p\text{-value} = P(\chi^2 > 12.99) = 0.0113$

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. $p\text{-value} = 0.0113$. $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the example in Table, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Table. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. If necessary, use the arrow keys to move the cursor after `Observed:` and press `2nd MATRX`. Press `1:[A]` to select matrix A. It is not necessary to enter expected values. The matrix listed after `Expected:` can be blank. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the *p*-value = 0.0113. Do the procedure a second time, but arrow down to `Draw` instead of `calculate`.

Exercise 11.2.3.2

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table 11.2.3.3shows the results:

Table 11.2.3.3

Industry Sector	2000	2010	2020	Total
Nonagriculture wage and salary	13,243	13,044	15,018	41,305
Goods-producing, excluding agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, forestry, fishing, and hunting	240	214	201	655
Nonagriculture self-employed and unpaid family worker	931	894	972	2,797
Secondary wage and salary jobs in agriculture and private household industries	14	11	11	36

Industry Sector	2000	2010	2020	Total
Secondary jobs as a self-employed or unpaid family worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Answer

- H_0: The number of jobs is independent of the year.
- H_a: The number of jobs is dependent on the year.

$df = 12$

alt

Figure 11.2.3.2.

Press the `MATRX` key and arrow over to `EDIT` . Press `1:[A]` . Press `3 ENTER 3 ENTER` . Enter the table values by row. Press `ENTER` after each. Press `2nd QUIT` . Press `STAT` and arrow over to `TESTS` . Arrow down to c:χ^2-TEST. Press `ENTER` . You should see `Observed:[A] and Expected:[B]` . Arrow down to `Calculate` . Press `ENTER` . The test statistic is 227.73 and the p-value $= 5.90E - 42 = 0$. Do the procedure a second time but arrow down to `Draw` instead of `calculate` .

Example 11.2.3.3

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. Table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

a. How many high anxiety level students are expected to have a high need to succeed in school?
b. If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?
c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} =$ ________
d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about ________.

Solution

a. The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09 \tag{11.2.3.2}$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

d. 8

? Exercise 11.2.3.3

Refer back to the information in Note. How many service providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

Answer

12,727, 14,965

References

1. DiCamilo, Mark, Mervin Field, "Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs." The Field Poll, released Feb. 14, 2013. Available online at field.com/fieldpollonline/sub...rs/Rls2436.pdf (accessed May 24, 2013).
2. Harris Interactive, "Favorite Flavor of Ice Cream." Available online at http://www.statisticbrain.com/favori...r-of-ice-cream (accessed May 24, 2013)
3. "Youngest Online Entrepreneurs List." Available online at http://www.statisticbrain.com/younge...repreneur-list (accessed May 24, 2013).

Review

To assess whether two factors are independent or not, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values. The test is right-tailed. Each observation or cell category must have an expected value of at least 5.

Formula Review

Test of Independence

- The number of degrees of freedom is equal to $(\text{number of columns - 1})(\text{number of rows - 1})$.
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where $O =$ observed values, $E =$ expected values, $i =$ the number of rows in the table, and $j =$ the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

Determine the appropriate test to be used in the next three exercises.

? Exercise 11.2.3.4

A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.

Answer

a test of independence

? Exercise 11.2.3.5

The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

Exercise 11.2.3.6

A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times as well as the brand of shoes they were wearing.

Answer

a test of independence

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. Table $11.2.3.4$shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance they must travel.

Table 11.2.3.4

Traveling Distance	Third class	Second class	First class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

Exercise 11.2.3.7

State the hypotheses.

- H_0: _______
- H_a: _______

Exercise 11.2.3.8

$df =$ _______

Answer

8

Exercise 11.2.3.9

How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

Exercise 11.2.3.10

How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

Answer

6.6

Exercise 11.2.3.11

What is the test statistic?

Exercise 11.2.3.12

What is the p-value?

Answer

0.0435

Exercise 11.2.3.13

What can you conclude at the 5% level of significance?

Use the following information to answer the next eight exercises: An article in the New England Journal of Medicine, discussed a study on smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans and 7,650 whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

Exercise 11.2.3.14

Complete the table.

Table 11.2.3.5: Smoking Levels by Ethnicity (Observed)

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Answer

Table 11.2.3.5B

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1-10	9,886	2,745	12,831	8,378	7,650	41,490
11-20	6,514	3,062	4,932	10,680	9,877	35,065
21-30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

Exercise 11.2.3.15

State the hypotheses.

- H_0: _______

- H_a: _______

? Exercise 11.2.3.16

Enter expected values in Table. Round to two decimal places.

Calculate the following values:

Answer

Table 11.2.3.6

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
1-10	7777.57	3310.11	8248.02	10771.29	11383.01
11-20	6573.16	2797.52	6970.76	9103.29	9620.27
21-30	2863.02	1218.49	3036.20	3965.05	4190.23
31+	1616.25	687.87	1714.01	2238.37	2365.49

? Exercise 11.2.3.17

$df =$ _______

? Exercise 11.2.3.18

χ^2test statistic $=$ ______

Answer

10,301.8

? Exercise 11.2.3.19

p-value $=$ ______

? Exercise 11.2.3.20

Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

Answer

right

? Exercise 11.2.3.21

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Blank graph with vertical and horizontal axes.

Figure 11.2.3.3.

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α.

Exercise 11.2.3.22

$\alpha = 0.05$

a. Decision: ____________________
b. Reason for the decision: ____________________
c. Conclusion (write out in a complete sentence): ____________________

Answer

a. Reject the null hypothesis.
b. p-value $< \alpha$
c. There is sufficient evidence to conclude that smoking level is dependent on ethnic group.

Exercise 11.2.3.23

$\alpha = 0.05$

a. Decision: ____________________
b. Reason for the decision: ____________________
c. Conclusion (write out in a complete sentence): ____________________

Glossary

Contingency Table
: a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

11.2.4: Test for Homogeneity

The goodness–of–fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the test for homogeneity, can be used to draw a conclusion about whether two populations have the same distribution. To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.

> The expected value for each cell needs to be at least five in order for you to use this test.

Hypotheses

- H_0: The distributions of the two populations are the same.
- H_a: The distributions of the two populations are not the same.

Test Statistic

- Use a χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of Freedom (df)

- $df = \text{number of columns} - 1$

Requirements

- All values in the table must be greater than or equal to five.

Common Uses

Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

Example 11.2.4.1

Do male and female college students have the same distribution of living arrangements? Use a level of significance of 0.05. Suppose that 250 randomly selected male college students and 300 randomly selected female college students were asked about their living arrangements: dormitory, apartment, with parents, other. The results are shown in Table $11.2.4.1$ Do male and female college students have the same distribution of living arrangements?

Table 11.2.4.1: Distribution of Living Arragements for College Males and College Females

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

Answer

- H_0: The distribution of living arrangements for male college students is the same as the distribution of living arrangements for female college students.
- H_a: The distribution of living arrangements for male college students is not the same as the distribution of living arrangements for female college students.

Degrees of Freedom (df):

$df = \text{number of columns} - 1 = 4 - 1 = 3$

Distribution for the test: χ^2_3

Calculate the test statistic: $\chi^2 = 10.1287$ (calculator or computer)

Probability statement: $p\text{-value} = P(\chi^2 > 10.1287) = 0.0175$

Press the

```
MATRX
```

key and arrow over to

```
EDIT
```

. Press

```
1:[A]
```

. Press

```
2 ENTER 4 ENTER
```

. Enter the table values by row. Press

```
ENTER
```

after each. Press

```
2nd QUIT
```

. Press

```
STAT
```

and arrow over to

```
TESTS
```

. Arrow down to

```
C:χ2-TEST
```

. Press

```
ENTER
```

. You should see

```
Observed:[A] and Expected:[B]
```

. Arrow down to

```
Calculate
```

. Press

```
ENTER
```

. The test statistic is 10.1287 and the $p\text{-value} = 0.0175$. Do the procedure a second time but arrow down to

```
Draw
```

instead of

```
calculate
```

.

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. $p\text{-value} = 0.0175$. $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0. This means that the distributions are not the same.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students are not the same.

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

Exercise 11.2.4.1

Do families and singles have the same distribution of cars? Use a level of significance of 0.05. Suppose that 100 randomly selected families and 200 randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, van/SUV. The results are shown in Table $11.2.4.2$ Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

Table 11.2.4.1

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Answer

With a p-value of almost zero, we reject the null hypothesis. The data show that the distribution of cars is not the same for families and singles.

Example 11.5.2

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. Table shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Answer

H_0: The distribution of voter preferences was the same before and after the earthquake.

H_a: The distribution of voter preferences was not the same before and after the earthquake.

Degrees of Freedom (df):

$df = \text{number of columns} - 1 = 3 - 1 = 2$

Distribution for the test: χ^2_2

Calculate the test statistic: $\chi^2 = 3.2603$ (calculator or computer)

Probability statement: $p\text{-value} = P(\chi^2 > 3.2603) = 0.1959$

Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `2 ENTER 3 ENTER`. Enter the table values by row. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 3.2603 and the *p*-value = 0.1959. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Compare α and the p-value: $\alpha = 0.05$ and the $p\text{-value} = 0.1959$. $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At a 5% level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

? Exercise 11.2.4.2

Ivy League schools receive many applications, but only some can be accepted. At the schools listed in Table, two types of applications are accepted: regular and early decision.

Application Type Accepted	Brown	Columbia	Cornell	Dartmouth	Penn	Yale
Regular	2,115	1,792	5,306	1,734	2,685	1,245
Early Decision	577	627	1,228	444	1,195	761

We want to know if the number of regular applications accepted follows the same distribution as the number of early applications accepted. State the null and alternative hypotheses, the degrees of freedom and the test statistic, sketch the graph of the *p*-value, and draw a conclusion about the test of homogeneity.

Answer

H_0: The distribution of regular applications accepted is the same as the distribution of early applications accepted.

H_a: The distribution of regular applications accepted is not the same as the distribution of early applications accepted.

$df = 5$

$\chi^2 \text{test statistic} = 430.06$

This is a nonsymmetric chi-square curve with df = 5. The values 0, 5, and 430.06 are labeled on the horizontal axis. The value 5 coincides with the peak of the curve. A vertical upward line extends from 430.06 to the curve, and the region to the right of this line is shaded. The shaded area is equal to the p-value.

Figure 11.2.4.1.

Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 430.06 and the $p\text{-value} = 9.80E-91$. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

References

1. Data from the Insurance Institute for Highway Safety, 2013. Available online at www.iihs.org/iihs/ratings (accessed May 24, 2013).

2. "Energy use (kg of oil equivalent per capita)." The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...G.OE/countries (accessed May 24, 2013).
3. "Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES)," U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubsearch/pubsinf...?pubid=2009030 (accessed May 24, 2013).
4. "Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES)," U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubs2009/2009030_sup.pdf (accessed May 24, 2013).

Review

To assess whether two data sets are derived from the same distribution—which need not be known, you can apply the test for homogeneity that uses the chi-square distribution. The null hypothesis for this test states that the populations of the two data sets come from the same distribution. The test compares the observed values against the expected values if the two populations followed the same distribution. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_{i \cdot j} \frac{(O-E)^2}{E}$ Homogeneity test statistic where: $O =$ observed values

$E =$ expected values

$i =$ number of rows in data contingency table

$j =$ number of columns in data contingency table

$df = (i-1)(j-1)$ Degrees of freedom

Exercise 11.2.4.3

A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

Answer

test for homogeneity

Exercise 11.2.4.4

What are the null and alternative hypotheses for Exercise?

Exercise 11.2.4.5

A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?

Answer

test for homogeneity

Exercise 11.2.4.6

A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

Exercise 11.2.4.7

What condition must be met to use the test for homogeneity?

Answer

All values in the table must be greater than or equal to five.

Use the following information to answer the next five exercises: Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in Table.

	20–30	30–40	40–50	50–60
Private Practice	16	40	38	6
Hospital	8	44	59	39

Exercise 11.2.4.8

State the null and alternative hypotheses.

Exercise 11.2.4.9

$df =$ _______

Answer

3

Exercise 11.2.4.10

What is the test statistic?

Exercise 11.2.4.11

What is the p-value?

Answer

0.00005

Exercise 11.2.4.12

What can you conclude at the 5% significance level?

11.3: PowerPoints

https://professormo.com/holistic/Powerpoint/ch11.pptx

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: STAT 300: My Introductory Statistics Textbook (Mirzaagha)

Webpages: 201

All licenses found:

- CC BY 4.0: 26.4% (53 pages)
- CC BY-SA 3.0: 22.9% (46 pages)
- CC BY-SA 4.0: 20.4% (41 pages)
- Undeclared: 15.9% (32 pages)
- Public Domain: 14.4% (29 pages)

By Page

- STAT 300: My Introductory Statistics Textbook (Mirzaagha) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Basic Ideas - *Undeclared*
    - 1.1: Videos - *CC BY-SA 4.0*
    - 1.2: Introduction - *CC BY-SA 4.0*
      - 1.2.1: A Classroom Story and an Inspiration - *CC BY-SA 4.0*
      - 1.2.2: The Blind Man and the Elephant - *CC BY-SA 4.0*
      - 1.2.3: What can go Wrong in Research - Two Stories - *CC BY-SA 4.0*
    - 1.3: Displaying and Analyzing Data with Graphs - *CC BY-SA 4.0*
      - 1.3.1: Introduction and Examples - *CC BY-SA 4.0*
      - 1.3.2: Types of Data - *CC BY-SA 4.0*
      - 1.3.3: Levels of Data - *CC BY-SA 4.0*
      - 1.3.4: Graphs of Categorical Data - *CC BY-SA 4.0*
      - 1.3.5: Graphs of Numeric Data - *CC BY-SA 4.0*
        - 1.3.5.1: Stem and Leaf Plots - *CC BY-SA 4.0*
        - 1.3.5.2: Dot Plots - *CC BY-SA 4.0*
        - 1.3.5.3: Grouping Numeric Data - *CC BY-SA 4.0*
        - 1.3.5.4: Histograms - *CC BY-SA 4.0*
        - 1.3.5.5: Cumulative Frequency and Relative Frequency - *CC BY-SA 4.0*
        - 1.3.5.6: Using Ogives to find Percentiles - *CC BY-SA 4.0*
    - 1.4: Introduction to Statistics - *Public Domain*
      - 1.4.1: What are Statistics? - *Public Domain*
      - 1.4.2: Importance of Statistics - *Public Domain*
      - 1.4.3: Descriptive Statistics - *Public Domain*
      - 1.4.4: Inferential Statistics - *Public Domain*
      - 1.4.5: Sampling Demonstration - *Public Domain*
      - 1.4.6: Variables - *Public Domain*
      - 1.4.7: Percentiles - *Public Domain*
      - 1.4.8: Levels of Measurement - *Public Domain*
      - 1.4.9: Measurements - *Public Domain*
      - 1.4.10: Distributions - *Public Domain*
      - 1.4.11: Summation Notation - *Public Domain*
      - 1.4.12: Linear Transformations - *Public Domain*
      - 1.4.13: Logarithms - *Public Domain*
      - 1.4.14: Statistical Literacy - *Public Domain*
      - 1.4.E: Introduction to Statistics (Exercises) - *Public Domain*
    - 1.5: PowerPoints - *Undeclared*
  - 2: Descriptive Statistics - *Undeclared*
    - 2.1: Videos - *CC BY-SA 4.0*
    - 2.2: Graphing Distributions - *Public Domain*
      - 2.2.1: Graphing Qualitative Variables - *Public Domain*
      - 2.2.2: Quantitative Variables - *Public Domain*
      - 2.2.3: Stem and Leaf Displays - *Public Domain*
      - 2.2.4: Histograms - *Public Domain*
      - 2.2.5: Frequency Polygons - *Public Domain*
      - 2.2.6: Box Plots - *Public Domain*
      - 2.2.7: Box Plot Demo - *Public Domain*
      - 2.2.8: Bar Charts - *Public Domain*
      - 2.2.9: Line Graphs - *Public Domain*
      - 2.2.10: Dot Plots - *Public Domain*
      - 2.2.11: Statistical Literacy - *Public Domain*
      - 2.2.E: Graphing Distributions (Exercises) - *Public Domain*
    - 2.3: PowerPoints - *Undeclared*
  - 3: Regression Analysis - *Undeclared*
    - 3.1: Videos - *CC BY-SA 4.0*