MATH 130: STATISTICS

Jupei Hsaio
Rio Hondo

Rio Hondo College

Introduction to Statistics

Fendi He, Jupei Hsiao, Erin Irwin

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.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

TABLE OF CONTENTS

Licensing

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

CHAPTER OVERVIEW

1: Introduction to Statistics

1.1: Statistics Vocabulary

The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives.

Collaborative Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5; 5.5; 6; 6; 6; 6.5; 6.5; 6.5; 6.5; 7; 7; 8; 8; 9

The dot plot for this data would be as follows:

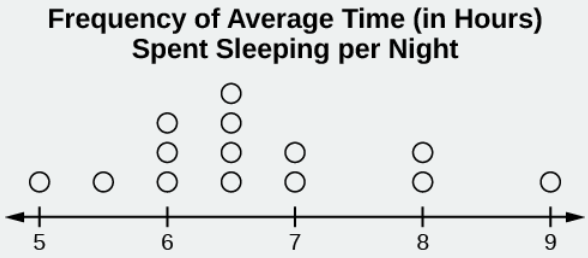


Figure 1.1.1

- Does your dot plot look the same as or different from the example? Why?
- If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?
- Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or

not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

> The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example 1.1.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Answer

- The **population** is all first year students attending ABC College this term.
- The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

- The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.
- The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.
- The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.
- The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Exercise 1.1.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Answer

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.
- The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

Example 1.1.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1._____ Population 2._____ Statistic 3._____ Parameter 4._____ Sample 5._____ Variable 6._____ Data

a. all students who attended the college last year
b. the cumulative GPA of one student who graduated from the college last year
c. 3.65, 2.80, 1.50, 3.90
d. a group of students who graduated from the college last year, randomly selected
e. the average cumulative GPA of students who graduated from the college last year
f. all students who graduated from the college last year
g. the average cumulative GPA of students in the study who graduated from the college last year

Answer

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

Example 1.1.3

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of "drive" (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple

random sample of 75 cars.

Answer

- The **population** is all cars containing dummies in the front seat.
- The **sample** is the 75 cars, selected by a simple random sample.
- The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.
- The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.
- The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.
- The **data** are either: yes, had head injury, or no, did not.

Example 1.1.4

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Answer

- The **population** is all medical doctors listed in the professional directory.
- The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.
- The **sample** is the 500 doctors selected at random from the professional directory.
- The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.
- The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.
- The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

1. The Data and Story Library, https://dasl.datadescription.com/ (accessed May 1, 2013).

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:

3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B:

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

Exercise 1.1.2

population

Answer

AIDS patients.

Exercise 1.1.3

sample

Exercise 1.1.4

parameter

Answer

The average length of time (in months) AIDS patients live after treatment.

Exercise 1.1.5

statistic

Exercise 1.1.6

variable

Answer

$X =$ the length of time (in months) AIDS patients live after treatment

Glossary

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Average

also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative**(an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable

a characteristic of interest for each person or object in a population

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.

1.2: Sampling Techniques

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of *counting* or *measuring* attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either *discrete* or *continuous*.

All data that are the result of counting are called *quantitative discrete data*. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are *quantitative continuous data* assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the **quantitative discrete data**.

Exercise 1.2.1

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Answer

quantitative discrete data

Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are **quantitative continuous data** because weights are measured.

Exercise 1.2.2

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Answer

quantitative continuous data

Exercise 1.2.3

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and two pounds (32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

Sample of qualitative data

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are **qualitative data**.

Exercise 1.2.4

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Answer

qualitative data

Collaborative Exercise 1.2.1

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. where you go on vacation
d. the distance it is from your home to the nearest grocery store
e. the number of classes you take per school year.
f. the tuition for your classes
g. the type of calculator you use
h. movie ratings
i. political party preferences
j. weights of sumo wrestlers
k. amount of money (in dollars) won playing poker
l. number of correct answers on a quiz
m. peoples' attitudes toward the government
n. IQ scores (This may cause some discussion.)

Answer

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Exercise 1.2.5

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Answer

quantitative discrete

Exercise 1.2.6

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart Figure 1.2.1. What type of data does this graph show?

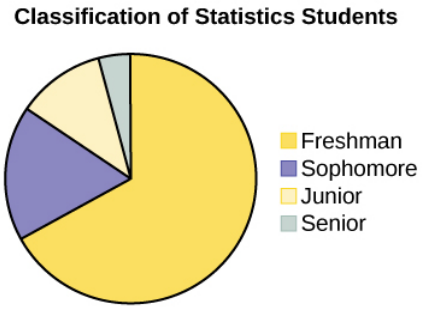


Figure 1.2.1

Answer

This pie chart shows the students in each year, which is **qualitative data**.

Exercise 1.2.7

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.

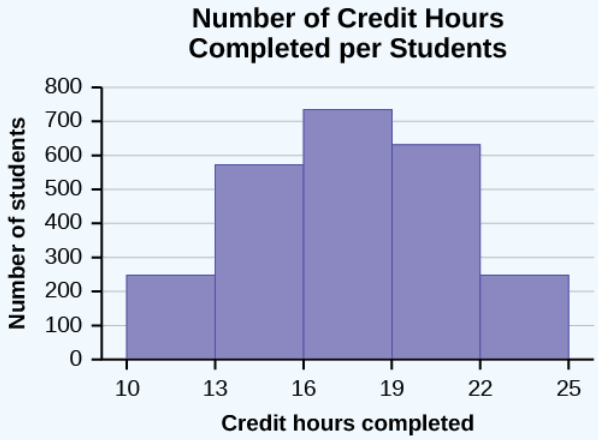


Figure 1.2.2

What type of data does this graph show?

Answer

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Table 1.2.1: Fall Term 2007 (Census day)

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

- In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.
- In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.
- A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figures 1.2.3 and 1.2.4 and determine which graph (pie or bar) you think displays the comparisons better.

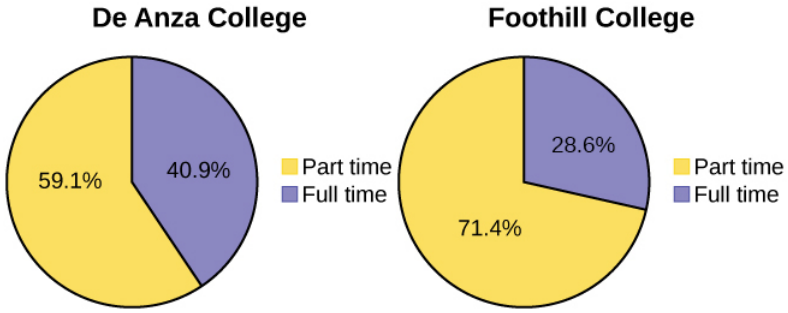


Figure 1.2.3: Pie Charts

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for.

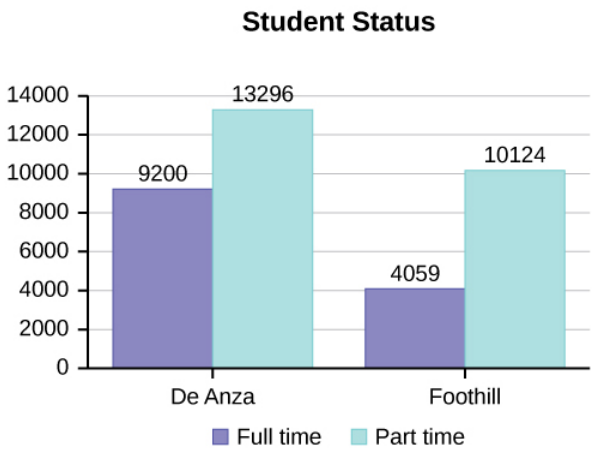


Figure 1.2.4: Bar chart

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Table 1.2.2: De Anza College Spring 2010

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

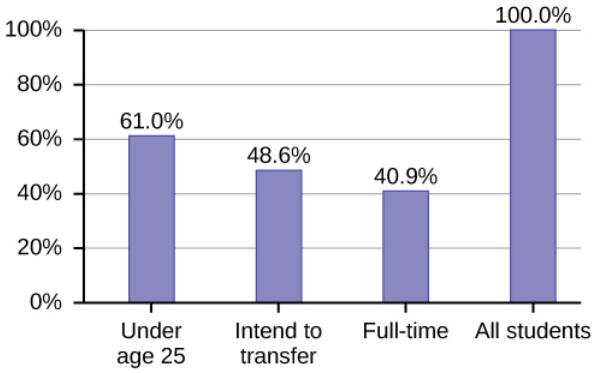


Figure 1.2.2: Bar chart of data in Table 1.2.2.

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Table 1.2.2: Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%

	Frequency	Percent
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

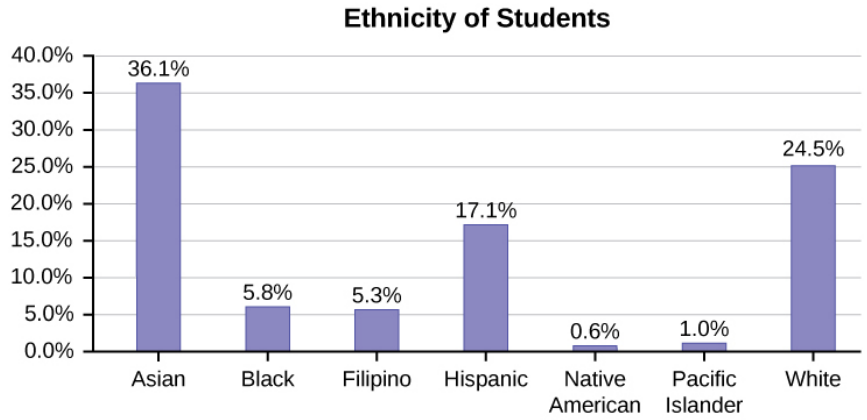


Figure 1.2.3: Enrollment of De Anza College (Spring 2010)

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 1.2.4 can be difficult to understand visually. The graph in Figure 1.2.5 is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

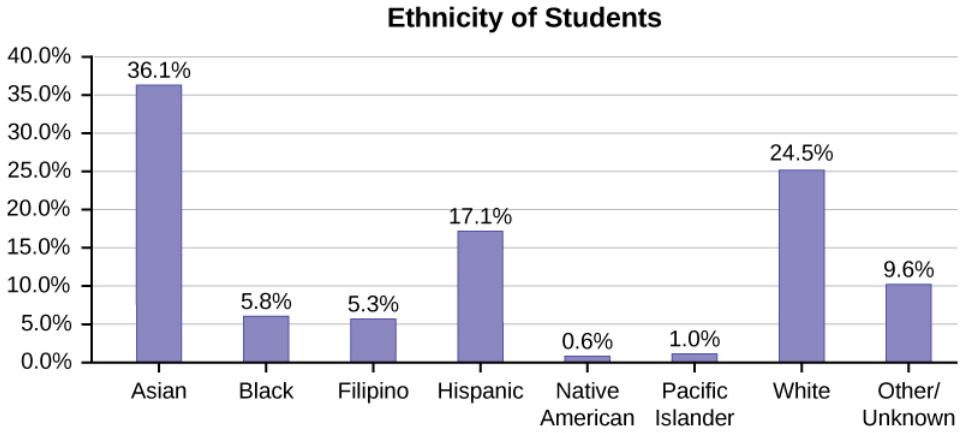


Figure 1.2.4: Bar Graph with Other/Unknown Category

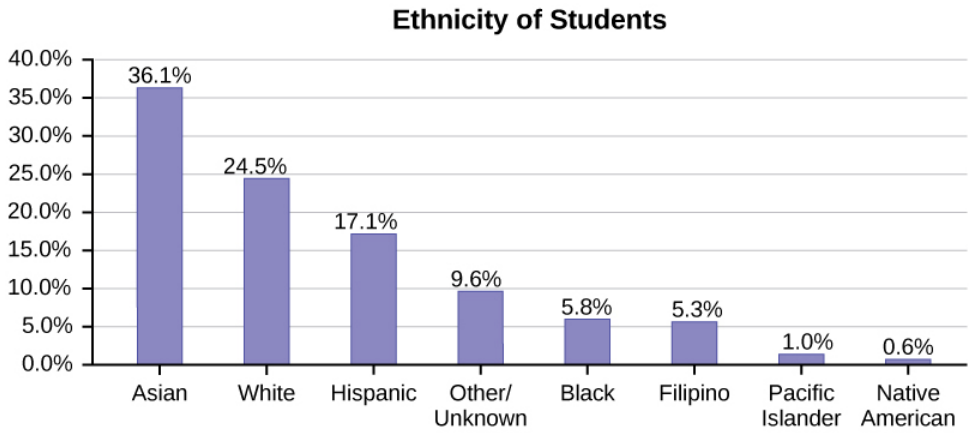


Figure 1.2.5: Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure 1.2.6 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 1.2.6.

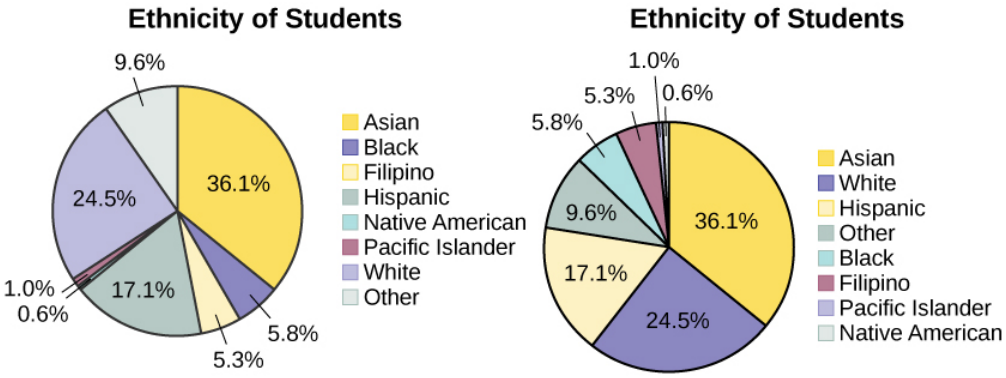


Figure 1.2.6.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 1.2.2:

Table 1.2.3: Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell

ID	Name	ID	Name	ID	Name
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.

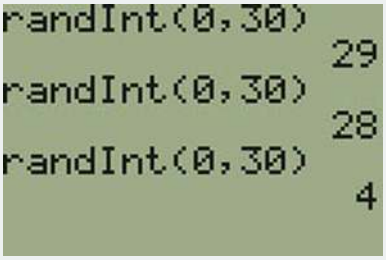


Figure 1.2.7

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and

do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Exercise 1.2.8

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Answer

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Example 1.2.9: Calculator

You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.
   - Number each row one through ten.
   - On your calculator, press Math and arrow over to PRB.

- For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
- Repeat for columns two through six.
- These 18 quiz scores are a stratified sample.

b. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.

- Press MATH and arrow over to PRB.
- Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
- The two numbers are for two of the columns.
- The quiz scores (20 of them) in these 2 columns are the cluster sample.

c. Create a simple random sample of 15 quiz scores.

- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER 15 times and record the numbers.
- Record the quiz scores that correspond to these numbers.
- These 15 quiz scores are the systematic sample.

d. Create a systematic sample of 12 quiz scores.

- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example 1.2.10

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Answer

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

Exercise 1.2.11

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Answer

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 1.2.12: Sampling

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task. Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Answer

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Answer

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. Is the sample biased?

Answer

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Exercise 1.2.12

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Answer

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Collaborative Exercise 1.2.8

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This *variability in samples* cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Collaborative Exercise 1.2.8

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in the following table ("frequency" is the number of times a particular face of the die occurs):

First Experiment (20 rolls)		Second Experiment (20 rolls)	
Face on Die	**Frequency**	**Face on Die**	**Frequency**
1			
2			
3			
4			
5			
6			

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from www.bookofodds.com/Relationsh...-the-President
5. Dominic Lusinchi, "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

6. “The Literary Digest Poll,” Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
7. “Gallup Presidential Election Trial-Heat Trends, 1936–2008,” Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
8. The Data and Story Library, lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).
9. LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
10. Data from San Jose Mercury News

Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Footnotes

1. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. Frequently Asked Questions, Pew Research Center for the People & the Press, www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).

Glossary

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

CHAPTER OVERVIEW

2: Descriptive Statistics

2.4: Applications of Standard Deviation

Learning Objectives

- To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the *Empirical Rule* and *Chebyshev's Theorem*.
- To use the Empirical Rule and Chebyshev's Theorem to draw conclusions about a data set.

You probably have a good intuitive grasp of what the average of a data set says about that data set. In this section we begin to learn what the standard deviation has to tell us about the nature of the data set.

The Empirical Rule

We start by examining a specific set of data. Table 2.4.1 shows the heights in inches of 100 randomly selected adult men. A relative frequency histogram for the data is shown in Figure 2.4.1. The mean and standard deviation of the data are, rounded to two decimal places, $\bar{x} = 69.92$ and $\sigma = 1.70$.

Table 2.4.1: Heights of Men

68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1
68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between $69.92 - 1.70 = 68.22$ and $69.92 + 1.70 = 71.62$ inches, there are 69 of them. If we count the number of observations that are within two standard deviations of the mean, that is, that are between $69.92 - 2(1.70) = 66.52$ and $69.92 + 2(1.70) = 73.32$ inches, there are 95 of them. All of the measurements are within three standard deviations of the mean, that is, between $69.92 - 3(1.70) = 64.822$ and $69.92 + 3(1.70) = 75.02$ inches. These tallies are not coincidences, but are in agreement with the following result that has been found to be widely applicable.

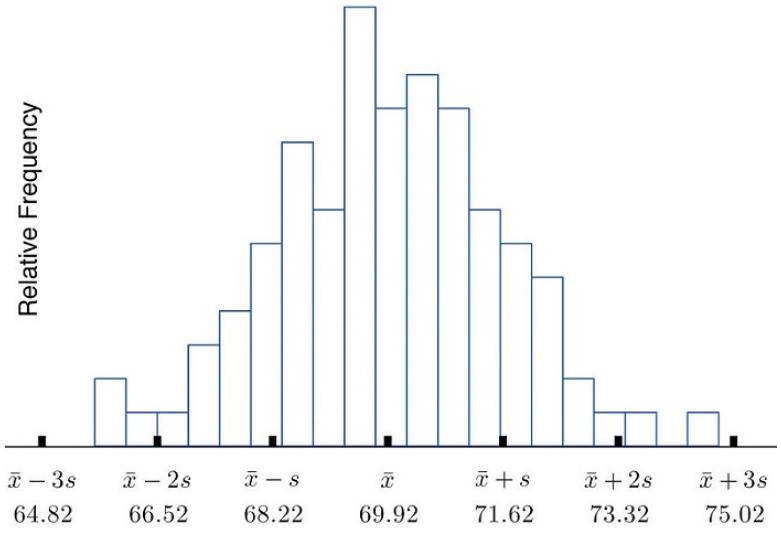


Figure 2.4.1: Heights of Adult Men

The Empirical Rule

Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations; if a data set has an approximately bell-shaped relative frequency histogram, then (Figure 2.4.2)

- approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations; and
- approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

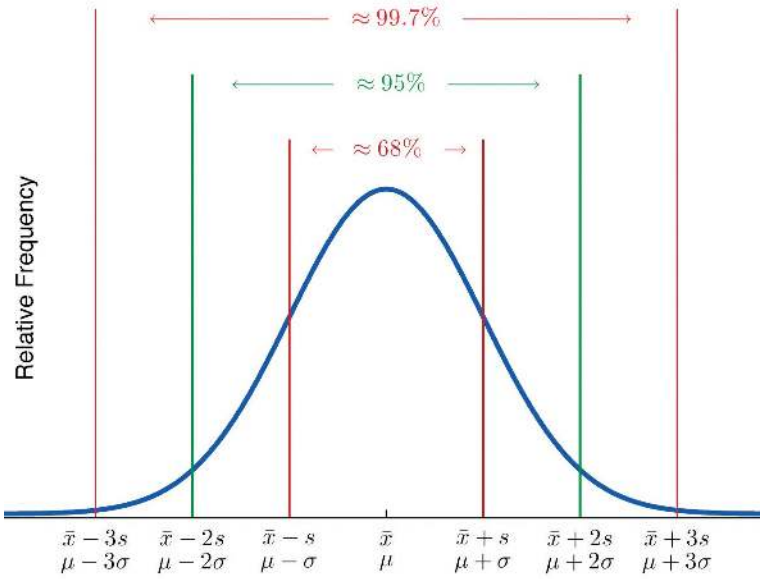


Figure 2.4.2: The Empirical Rule

Two key points in regard to the Empirical Rule are that the data distribution must be approximately bell-shaped and that the percentages are only approximately true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule. We see this with the example of the heights of the men: the Empirical Rule suggested 68 observations between 68.22 and 71.62 inches, but we counted 69.

Example 2.4.1

Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.

1. About what proportion of all such men are between 68.2 and 71 inches tall?
2. What interval centered on the mean should contain about 95% of all such men?

Solution

A sketch of the distribution of heights is given in Figure 2.4.3.

1. Since the interval from 68.2 to 71.0 has endpoints $\bar{x}-s$ and $\bar{x}+s$, by the Empirical Rule about 68% of all 18-year-old males should have heights in this range.
2. By the Empirical Rule the shortest such interval has endpoints $\bar{x}-2s$ and $\bar{x}+2s$. Since

$$\bar{x}-2s=69.6-2(1.4)=66.8$$

and

$$\bar{x}+2s=69.6+2(1.4)=72.4$$

the interval in question is the interval from 66.8 inches to 72.4 inches.

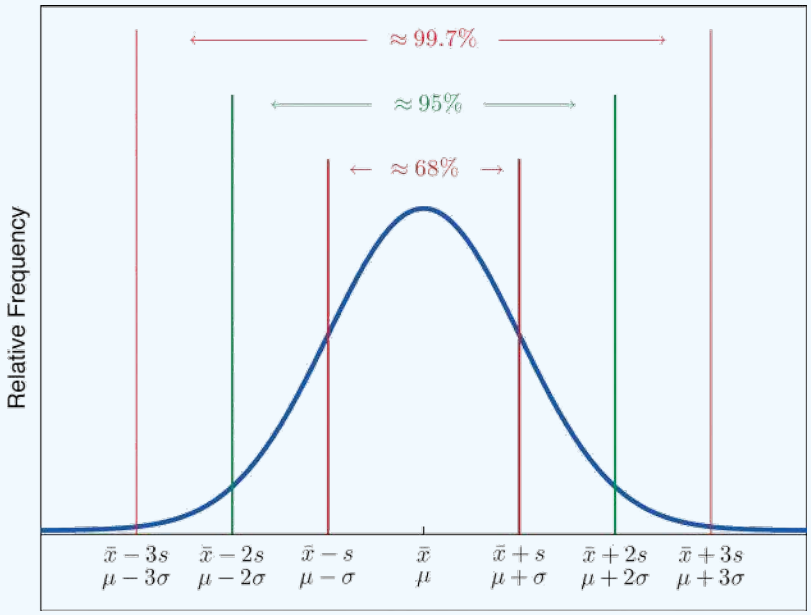


Figure 2.4.3: Distribution of Heights

Example 2.4.2

Scores on IQ tests have a bell-shaped distribution with mean $\mu=100$ and standard deviation $\sigma=10$. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

Solution

A sketch of the IQ distribution is given in Figure 2.4.3. The Empirical Rule states that

1. approximately 68% of the IQ scores in the population lie between 90 and 110,
2. approximately 95% of the IQ scores in the population lie between 80 and 120, and
3. approximately 99.7% of the IQ scores in the population lie between 70 and 130.

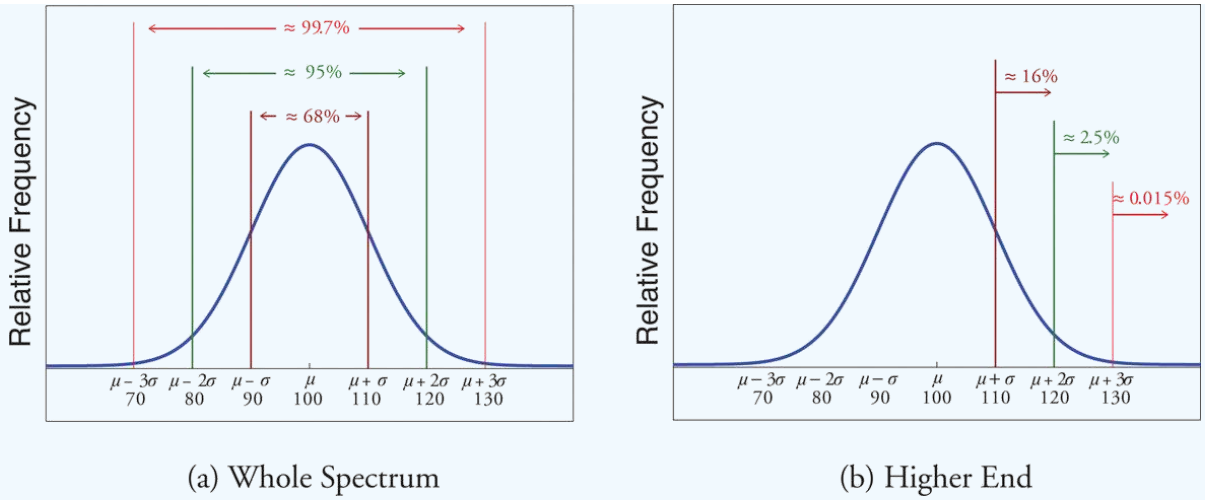


Figure 2.4.3: Distribution of IQ Scores.

1. Since 68% of the IQ scores lie *within* the interval from 90 to 110, it must be the case that 32% lie *outside* that interval. By symmetry approximately half of that 32%, or 16% of all IQ scores, will lie above 110. If 16% lie above 110, then 84% lie below. We conclude that the IQ score 110 is the 84^{th} percentile.
2. The same analysis applies to the score 120. Since approximately 95% of all IQ scores lie within the interval form 80 to 120, only 5% lie outside it, and half of them, or 2.5% of all scores, are above 120. The IQ score 120 is thus higher than 97.5% of all IQ scores, and is quite a high score.
3. By a similar argument, only $15/100$ of 1% of all adults, or about one or two in every thousand, would have an IQ score above 130. This fact makes the score 130 extremely high.

Chebyshev's Theorem

The Empirical Rule does not apply to all data sets, only to those that are bell-shaped, and even then is stated in terms of approximations. A result that applies to every data set is known as Chebyshev's Theorem.

Chebyshev's Theorem

For any numerical data set,

- at least $3/4$ of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations;
- at least $8/9$ of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations;
- at least $1 - 1/k^2$ of the data lie within k standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm ks$ for samples and with endpoints $\mu \pm k\sigma$ for populations, where k is any positive whole number that is greater than 1.

Figure $2.4.4$ gives a visual illustration of Chebyshev's Theorem.

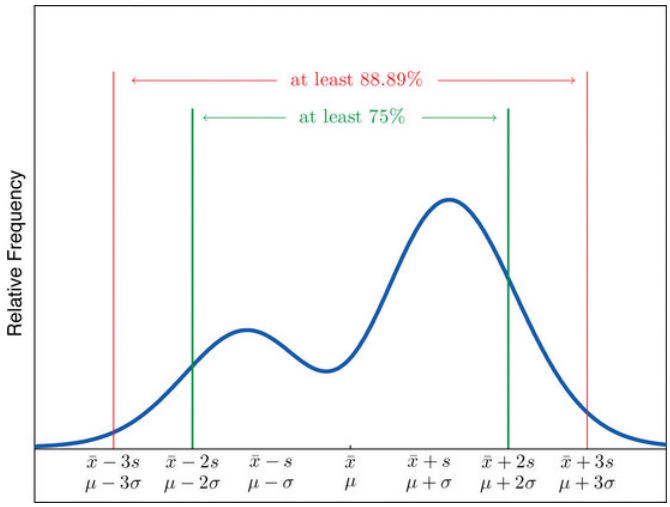


Figure 2.4.4: Chebyshev's Theorem

It is important to pay careful attention to the words **"at least"** at the beginning of each of the three parts of Chebyshev's Theorem. The theorem gives the *minimum* proportion of the data which must lie within a given number of standard deviations of the mean; the true proportions found within the indicated regions could be greater than what the theorem guarantees.

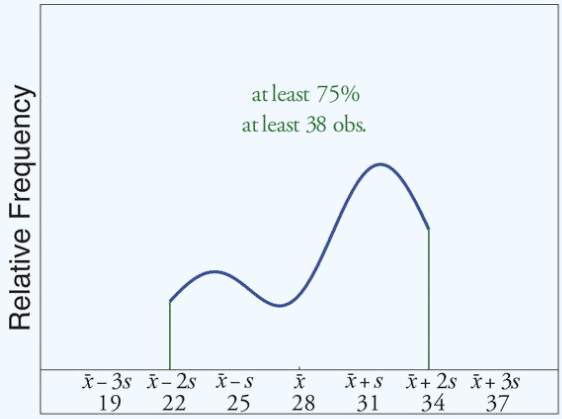
Example 2.4.3

A sample of size $n = 50$ has mean $\bar{x} = 28$ and standard deviation $s = 3$. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval $(22, 34)$? What can be said about the number of observations that lie outside that interval?

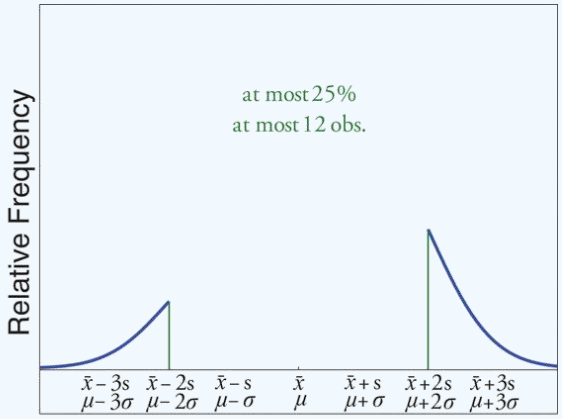
Solution

The interval $(22, 34)$ is the one that is formed by adding and subtracting two standard deviations from the mean. By Chebyshev's Theorem, at least $3/4$ of the data are within this interval. Since $3/4$ of 50 is 37.5, this means that at least 37.5 observations are in the interval. But one cannot take a fractional observation, so we conclude that at least 38 observations must lie inside the interval $(22, 34)$.

If at least $3/4$ of the observations are in the interval, then at most $1/4$ of them are outside it. Since $1/4$ of 50 is 12.5, at most 12.5 observations are outside the interval. Since again a fraction of an observation is impossible, x $(22, 34)$.



(a) Within $\bar{x} \pm 2s$



(b) Outside $\bar{x} \pm 2$

Example 2.4.4

The number of vehicles passing through a busy intersection between $8:00\ a.m.$ and $10:00\ a.m.$ was observed and recorded on every weekday morning of the last year. The data set contains $n = 251$ numbers. The sample mean is $\bar{x} = 725$ and the sample standard deviation is $s = 25$. Identify which of the following statements *must* be true.

1. On approximately 95% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
2. On at least 75% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
3. On at least 189 weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
4. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was either less than 675 or greater than 775.
5. On at most 12.5% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.
6. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.

Solution

1. Since it is not stated that the relative frequency histogram of the data is bell-shaped, the Empirical Rule does not apply. Statement (1) is based on the Empirical Rule and therefore it might not be correct.
2. Statement (2) is a direct application of part (1) of Chebyshev's Theorem because $\bar{x} - 2s$, $\bar{x} + 2s = (675, 775)$. It must be correct.
3. Statement (3) says the same thing as statement (2) because 75% of 251 is 188.25, so the minimum whole number of observations in this interval is 189. Thus statement (3) is definitely correct.
4. Statement (4) says the same thing as statement (2) but in different words, and therefore is definitely correct.
5. Statement (4), which is definitely correct, states that at most 25% of the time either fewer than 675 or more than 775 vehicles passed through the intersection. Statement (5) says that half of that 25% corresponds to days of light traffic. This would be correct if the relative frequency histogram of the data were known to be symmetric. But this is not stated; perhaps all of the observations outside the interval $(675, 775)$ are less than 75. Thus statement (5) might not be correct.
6. Statement (4) is definitely correct and statement (4) implies statement (6): even if every measurement that is outside the interval $(675, 775)$ is less than 675 (which is conceivable, since symmetry is not known to hold), even so at most 25% of all observations are less than 675. Thus statement (6) must definitely be correct.

Key Takeaway

- The Empirical Rule is an approximation that applies only to data sets with a bell-shaped relative frequency histogram. It estimates the proportion of the measurements that lie within one, two, and three standard deviations of the mean.
- Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within one, two, or more standard deviations of the mean.

2.1.1: Five Number Summary and Box Plots Part 1

Box plots (also called *box-and-whisker plots* or *box-whisker plots*) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately *the middle 50 percent of the data fall inside the box*. The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

> You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1; 1; 2; 2; 4; 6; 6;.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

> See the calculator instructions on the TI web site or in the appendix.

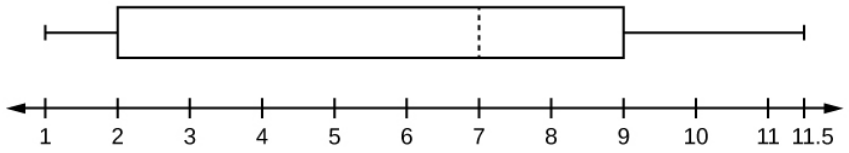


Figure 2.1.1.1

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

> It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example 2.1.1.1

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- $Q1$: First quartile = 64.5
- $Q2$: Second quartile or median= 66
- $Q3$: Third quartile = 70

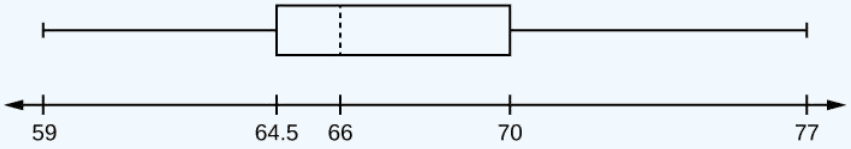


Figure 2.1.1.2

a. Each quarter has approximately 25% of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. $\text{Range} = \text{maximum value} - \text{the minimum value} = 77 - 59 = 18$
d. Interquartile Range: $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$.
e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Calculator

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4: Plotsoff. Press ENTER.

Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

Press TRACE, and use the arrow keys to examine the box plot.

Exercise 2.1.1.1

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

Answer

Figure 2.1.1.3

$IQR = 158$

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:

Figure 2.1.1.4

In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example 2.1.1.2

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

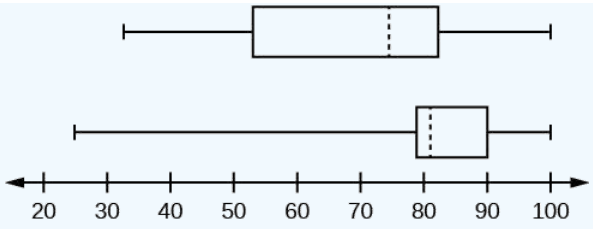
a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Answer

a.
- Min = 32
- Q_1 = 56
- M = 74.5
- Q_3 = 82.5
- Max = 99

b.
- Min = 25.5
- Q_1 = 78
- M = 81
- Q_3 = 89
- Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:



d.

e. The first data set has the wider spread for the middle 50% of the data. The *IQR*for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

Exercise 2.1.1.2

The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Answer

Figure 2.1.1.6

IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

Example 2.1.1.3

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.

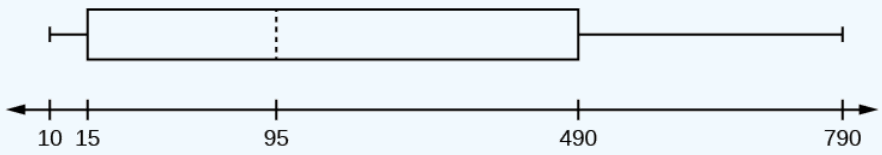


Figure 2.1.1.7

Exercise 2.1.1.3

Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

Answer

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q_1: 15
- Med: 50
- Q_3: 110
- Max: 330

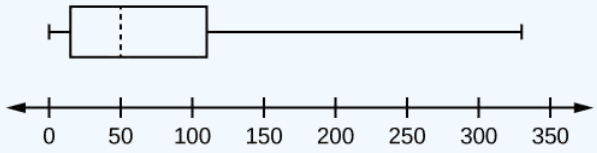


Figure 2.1.1.8

To find the IQR and create a box plot on the TI-83/84:

1. Go into the STAT menu, and then Choose 1:Edit

Figure 2.1.1.5 *: STAT Menu on TI-83/84*

2. Type your data values into L1. If L1 has data in it, arrow up to the name L1, click CLEAR and then press ENTER. The column will now be cleared and you can type the data in.
3. Go into the STAT menu, move over to CALC and choose 1-Var Stats. Press ENTER, then type L1 (2nd 1) and then ENTER. This will give you the summary statistics. If you press the down arrow, you will see the five-number summary.
4. To draw the box plot press 2nd STAT PLOT.

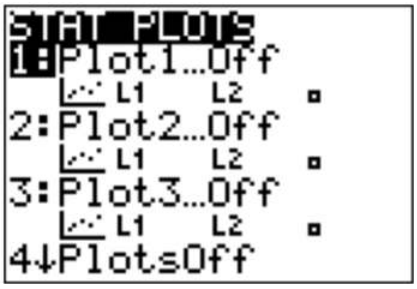


Figure 2.1.1.6 *: STAT PLOT Menu on TI-83/84*

5. Use Plot1. Press ENTER

Figure 2.1.1.7 : Plot1 Menu on TI-83/84 Setup for Box Plot

6. Put the cursor on On and press Enter to turn the plot on. Use the down arrow and the right arrow to highlight the boxplot in the middle of the second row of types then press ENTER. Set Data List to L1 (it might already say that) and leave Freq as 1.
7. Now tell the calculator the set up for the units on the x-axis so you can see the whole plot. The calculator will do it automatically if you press ZOOM, which is in the middle of the top row.

Figure 2.1.1.8 : ZOOM Menu on TI-83/84

Then use the down arrow to get to 9:ZoomStat and press ENTER. The box plot will be drawn.

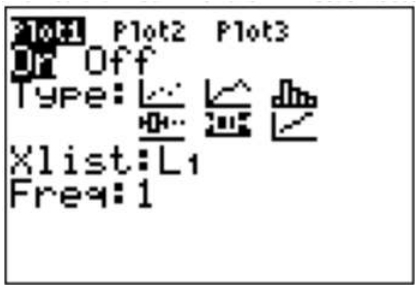


Figure 2.1.1.9 : ZOOM Menu on TI-83/84 with ZoomStat

References

1. Data from *West Magazine*.

Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

? Exercise 2.5.4

Construct a box plot below. Use a ruler to measure and scale accurately.

? Exercise 2.5.5

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Answer

More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Bringing It Together

Exercise 2.5.6

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under age 35?
c. Which box plot most resembles the information above?

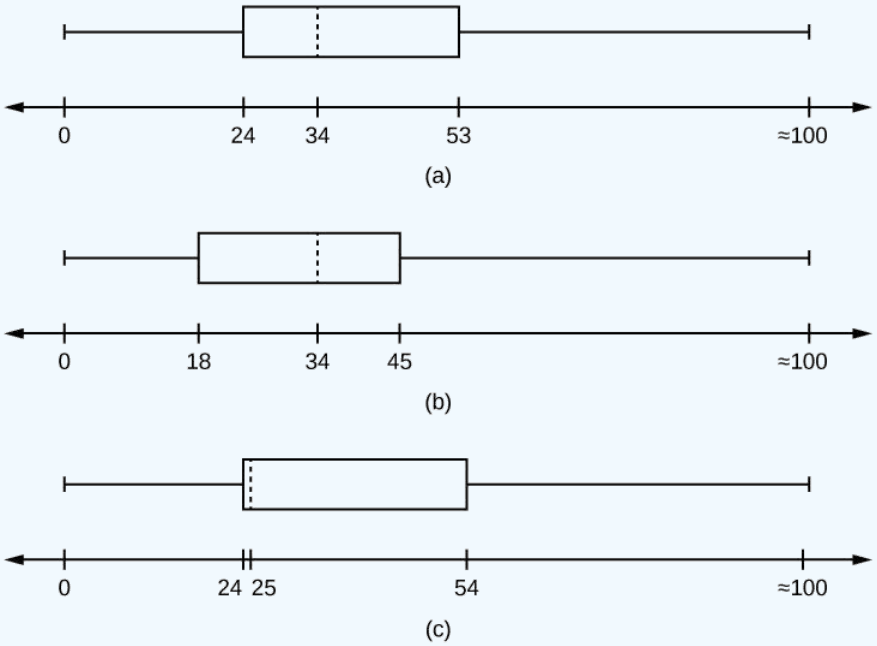


Figure 2.1.1.9 .

Answer

a. For graph, check student's solution.
b. 49.7% of the community is under the age of 35.
c. Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the of the lower half of the ordered data set

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

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.

2.1.2: Five Number Summary and Box Plots Part 2

The common measures of location are **quartiles** and **percentiles.** Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, *M*, is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8+7.2}{2}=7 \tag{2.1.2.1}$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, *Q*3, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1 \tag{2.4.1}$$

The *IQR* can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(*IQR*) below the first quartile or more than (1.5)(*IQR*) above the third quartile**. Potential outliers always require further investigation.

Definition: Outliers

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example 2.4.1

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Answer

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Exercise 2.1.2.1

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

$33,000; $64,500; $28,000; $54,000; $72,000; $68,500; $69,000; $42,000; $54,000; $120,000; $40,500

Answer

Order the data from smallest to largest.

$28,000; $33,000; $40,500; $42,000; $54,000; $54,000; $64,500; $68,500; $69,000; $72,000; $120,000

Median = $54,000

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

$$IQR = \$69,000 - \$40,500 = \$28,500$$

$$(1.5)(IQR) = (1.5)(\$28,500) = \$42,750$$

$$Q_1 - (1.5)(IQR) = \$40,500 - \$42,750 = -\$2,250$$

$$Q_3 + (1.5)(IQR) = \$69,000 + \$42,750 = \$111,750$$

No salary is less than –$2,250. However, $120,000 is more than $11,750, so $120,000 is a potential outlier.

Example 2.4.2

For the two data sets in the test scores example, find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.

Answer

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

a. The *IQR* for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

The *IQR* for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56 - 26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR(1.5) = 78 - 11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Exercise 2.1.2.2

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class *B*

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Answer

Class *A*

Order the data from smallest to largest.

65; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99

$Median = \frac{80+81}{2} = 80.5$

$Q_1 = \frac{69+76}{2} = 72.5$

$Q_3 = \frac{90+91}{2} = 90.5$

$IQR = 90.5 - 72.5 = 18$

Class B

Order the data from smallest to largest.

68; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92; 95; 95; 97; 99; 100

$Median = \frac{80+80}{2} = 80$

$Q_1 = \frac{72+73}{2} = 72.5$

$Q_3 = \frac{92+95}{2} = 93.5$

$IQR = 93.5 - 72.5 = 21$

The data for Class B has a larger IQR, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class B are more spread out and not clustered about the median.

Example 2.1.2.3

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Exercise 2.1.2.3

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Answer

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Example 2.4.4

Using the table above in Example 2.1.2.3

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

Exercise 2.1.2.4

Refer to the table above in Exercise 2.1.2.3 Find the third quartile. What is another name for the third quartile?

Answer

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

COLLABORATIVE STATISTICS

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

a. How many students were surveyed?
b. What kind of sampling did you do?
c. Construct two different histograms. For each, starting value = _____ ending value = ____.
d. Find the median, first quartile, and third quartile.
e. Construct a table of the data to find the following:
   i. the 10th percentile
   ii. the 70th percentile
   iii. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the kth percentile. Here is one of them.

- $k =$ the kth percentile. It may or may not be part of the data.
- $i =$ the index (ranking or position of a data value)
- $n =$ the total number of data

Order the data from smallest to largest.

Calculate $i = \frac{k}{100}(n+1)$

If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.

If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 2.4.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution

a.
- $k = 70$
- i = the index
- $n = 29$

$i = \frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

b.
- k = 83rd percentile
- i= the index
- $n = 29$

$i = \frac{k}{100}(n+1) = (\frac{83}{100})(29+1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Exercise 2.1.2.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Calculate the 20th percentile and the 55th percentile.

Answer

$k = 20$. Index $= i = \frac{k}{100}(n+1) = \frac{20}{100}(29+1) = 6$. The age in the sixth position is 27. The 20th percentile is 27 years.

$k = 55$. Index $= i = \frac{k}{100}(n+1) = \frac{55}{100}(29+1) = 16.5$. Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55. The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

Note 2.4.2

You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- $x =$ the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- $y =$ the number of data values equal to the data value for which you want to find the percentile.
- $n =$ the total number of data.
- Calculate $\frac{x+0.5y}{n}(100)$. Then round to the nearest integer.

Example 2.4.6

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

$x = 18$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{18+0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

$x = 3$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{3+0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12thpercentile.

Exercise 2.1.2.6

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the percentiles for 47 and 31.

Answer

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$x = 15$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{15+0.5(1)}{30}(100) = 51.67$. 47 is the 52nd percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$x = 8$ and $y = 2$. $\frac{x+0.5y}{n}(100) = \frac{8+0.5(2)}{30}(100) = 30$. 31 is the 30th percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Answer

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

? Exercise 2.1.2.7

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Answer

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

On a 20 question math test, the 70^{th} percentile for number of correct answers was 16. Interpret the 70^{th} percentile in the context of this situation.

Answer

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

? Exercise 2.1.2.8

On a 60 point written assignment, the 80^{th} percentile for the number of points earned was 49. Interpret the 80^{th} percentile in the context of this situation.

Answer

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

✓ Example 2.4.9

At a community college, it was found that the 30^{th} percentile of credit units that students are enrolled for is seven units. Interpret the 30^{th} percentile in the context of this situation.

Answer

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

? Exercise 2.1.2.9

During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Answer

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

✓ Example 2.4.10

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \quad (2.1.2.2)$$

.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q_1 = 20
- Q_3 = 60
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

1. Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at usatoday30.usatoday.com/news/...sus/55029100/1 (accessed April 3, 2013).

2. Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).
3. "1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).
4. Data from *San Jose Mercury News*.
5. Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$$i = \frac{k}{100}(n+1)$$

where i = the ranking or position of a data value,

- k = the kth percentile,
- n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x+0.5y}{n}\right)(100)$

where $x =$ the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

$y =$ the number of data values equal to the data value for which you want to find the percentile,

$n =$ total number of data

Glossary

Interquartile Range
or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier
an observation that does not fit the rest of the data

Percentile
a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles
the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

2.2.1: Histograms Part 1

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample.(Remember, frequency is defined as the number of times an answer occurs.) If:

- f is frequency
- n is total number of data values (or the sum of the individual frequencies), and
- RF is relative frequency,

then:

$$RF = \frac{f}{n} \tag{2.2.1.1}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and RF = fn = 340 = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many bars or intervals, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is $6.05 (6.1 – 0.05 = 6.05)$. We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is $1.495 (1.5 – 0.005 = 1.495)$. If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is $0.9995 (1.0 – 0.0005 = 0.9995)$. If all the data happen to be integers and the smallest value is two, then a convenient starting point is $1.5 (2 – 0.5 = 1.5)$. Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example 2.2.1.1

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72; 72.5; 72.5; 73; 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76 \quad (2.2.1.2)$$

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.

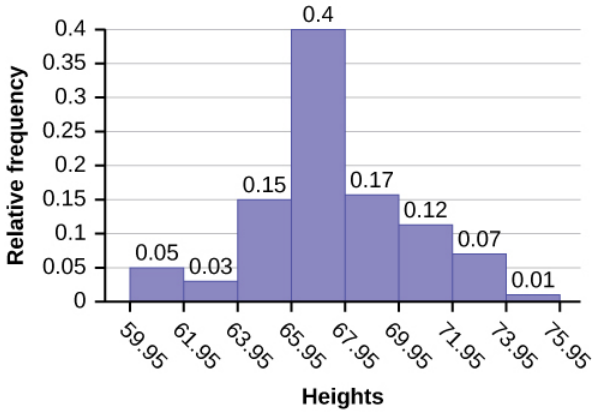


Figure 2.2.1.1 : Histogram of something

Exercise 2.2.1.1

The following data are the shoe sizes of 50 male students. The sizes are discrete data since shoe size is measured in whole and half units only. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Answer

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

$\frac{14.05-8.95}{6} = 0.85$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Example 2.2.1.2

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

Answer

Calculate the number of bars as follows:

$\frac{6.5-0.5}{\text{number of bars}} = 1$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.

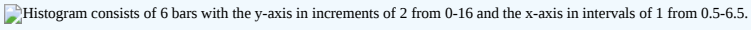


Figure 2.2.1.2 .

Note

Go to [link]. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example.

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.

- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rdpicture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

Exercise 2.2.1.2

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 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

20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

Answer

1.5

1.5 to 2.5

2.5 to 3.5

Example 2.2.1.3

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends

9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Answer

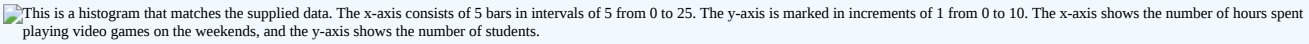

This is a histogram that matches the supplied data. The x-axis consists of 5 bars in intervals of 5 from 0 to 25. The y-axis is marked in increments of 1 from 0 to 10. The x-axis shows the number of hours spent playing video games on the weekends, and the y-axis shows the number of students.

Figure 2.2.1.3 .

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Exercise 2.2.1.3

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22; 35; 15; 26; 40; 28; 18; 20; 25; 34; 39; 42; 24; 22; 19; 27; 22; 34; 40; 20; 38 and 28

Use 10–19 as the first interval.

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

To create a histogram on the TI-83/84:

1. Go into the STAT menu, and then Chose 1: Edit

Figure 2.2.1.1 *: STAT Menu on TI-83/84*

2. Type your data values into L1.
3. Now click STAT PLOT (2^{nd} $Y=$).

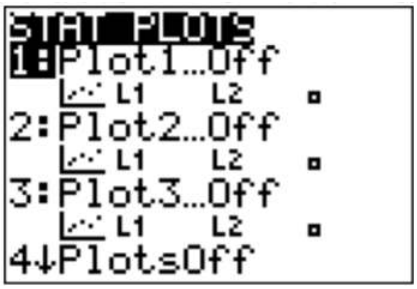


Figure 2.2.1.2 *: STAT PLOT Menu on TI-83/84*

4. Use 1:Plot1. Press ENTER.

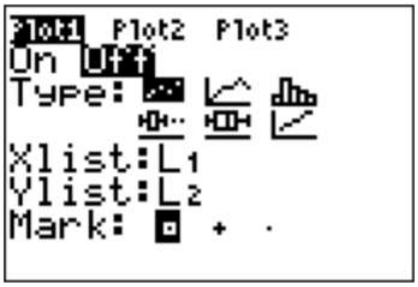


Figure 2.2.1.3 *: Plot1 Menu on TI-83/84*

5. You will see a new window. The first thing you want to do is turn the plot on. At this point you should be on On, just press ENTER. It will make On dark.
6. Now arrow down to Type: and arrow right to the graph that looks like a histogram (3rd one from the left in the top row).
7. Now arrow down to Xlist. Make sure this says L1. If it doesn't, then put L1 there (2nd number 1). Freq: should be a 1.

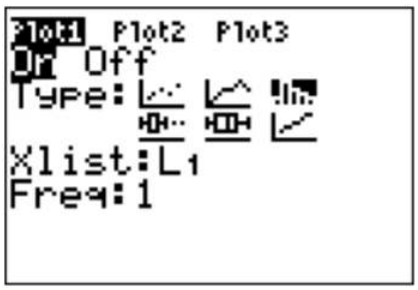


Figure 2.2.1.4 *: Plot1 Menu on TI-83/84 Setup for Histogram*

8. Now you need to set up the correct window to graph on. Click on WINDOW. You need to set up the settings for the x variable. Xmin should be your smallest data value. Xmax should just be a value sufficiently above your highest data value, but not too high. Xscl is your class width that you calculated. Ymin should be 0 and Ymax should be above what you think the highest frequency is going to be. You can always change this if you need to. Yscl is just how often you would like to see a tick mark on the y-axis.
9. Now press GRAPH. You will see a histogram.

Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.Glossary

References

1. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker
2. "Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at www.scholastic.com/teachers/a...-us-presidents (accessed April 3, 2013).
3. "Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).
4. "Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).
5. " Births Time Series Data." General Register Office For Scotland, 2013. Available online at www.gro-scotland.gov.uk/stati...me-series.html (accessed April 3, 2013).
6. "Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).
7. Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.
8. "Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Frequency

the number of times a value of the data occurs

Histogram

a graphical representation in $x - y$ form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

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.

2.2.2: Histograms Part 2

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

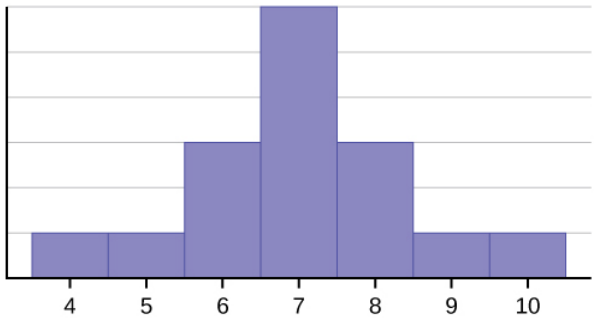


Figure 2.2.2.1

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

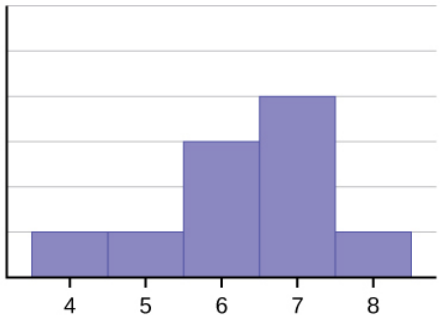


Figure 2.2.2.2

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**.

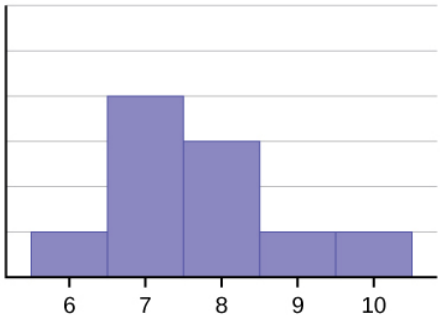


Figure 2.2.2.3

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

> *Generally, if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.*

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example 2.2.2.1

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

- Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2
- Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1
- Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

Solution

a.

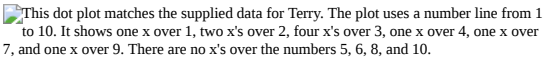


Figure 2.2.2.4: This dot plot matches the supplied data for Terry. The plot uses a number line from 1 to 10. It shows one x over 1, two x's over 2, four x's over 3, one x over 4, one x over 7, and one x over 9. There are no x's over the numbers 5, 6, 8, and 10.

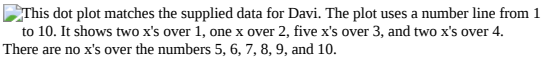


Figure 2.2.2.5: Copy and Paste Caption here. (Copyright; author via source)

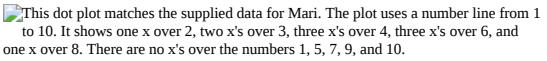


Figure 2.2.2.6: Copy and Paste Caption here. (Copyright; author via source)

- Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
- Terry's median is three, Davis' median is three. Maris' median is four.
- It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Exercise 2.2.2.1

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

Figure 2.2.2.7: This dot plot matches the supplied data. The plot uses a number line from 0 to 14. It shows two x's over 0, four x's over 1, three x's over 2, one x over 3, two x's over the number 4, 5, 6, and 9, and 1 x each over 10 and 14. There are no x's over the numbers 7, 8, 11, 12, and 13.

b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 7 8

The Ages Former U.S Presidents Died	
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

c.

Figure 2.2.2.8: This is a histogram titled Hours Spent Playing Video Games on Weekends. The x-axis shows the number of hours spent playing video games with bars showing values at intervals of 5. The y-axis shows the number of students. The first bar for 0 - 4.99 hours has a height of 2. The second bar from 5 - 9.99 has a height of 3. The third bar from 10 - 14.99 has a height of 4. The fourth bar from 15 - 19.99 has a height of 7. The fifth bar from 20 - 24.99 has a height of 9.

Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **left (or negative) skewed** distribution has a shape like Figure 2.2.2.2 A **right (or positive) skewed** distribution has a shape like Figure 2.2.2.3 A **symmetrical** distribution looks like Figure 2.2.2.1.

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

Exercise 2.7.2

1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

Answer

The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

Exercise 2.7.3

16; 17; 19; 22; 22; 22; 22; 22; 23

Exercise 2.7.4

87; 87; 87; 87; 87; 88; 89; 89; 90; 91

Answer

The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercise 2.7.5

When the data are skewed left, what is the typical relationship between the mean and median?

Exercise 2.7.6

When the data are symmetrical, what is the typical relationship between the mean and median?

Answer

When the data are symmetrical, the mean and median are close or the same.

Exercise 2.7.7

What word describes a distribution that has two modes?

Exercise 2.7.8

Describe the shape of this distribution.

Figure 2.2.2.9: This is a historgram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right.

Answer

The distribution is skewed right because it looks pulled out to the right.

Exercise 2.7.9

Describe the relationship between the mode and the median of this distribution.

Figure 2.2.2.10: This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heighs from left to right are: 8, 4, 2, 2, 1.

Exercise 2.7.10

Describe the relationship between the mean and the median of this distribution.

Figure 2.2.2.11: This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

Answer

The mean is 4.1 and is slightly greater than the median, which is four.

Exercise 2.7.11

Describe the shape of this distribution.

Figure 2.2.2.12

Exercise 2.7.12

Describe the relationship between the mode and the median of this distribution.

Figure 2.2.2.13

Answer

The mode and the median are the same. In this case, they are both five.

Exercise 2.7.13

Are the mean and the median the exact same in this distribution? Why or why not?

Figure 2.2.2.14

Exercise 2.7.14

Describe the shape of this distribution.

Figure 2.2.2.15

Answer

The distribution is skewed left because it looks pulled out to the left.

Exercise 2.7.15

Describe the relationship between the mode and the median of this distribution.

Figure 2.2.2.16: Copy and Paste Caption here. (Copyright; author via source)

Exercise 2.7.16

Describe the relationship between the mean and the median of this distribution.

Figure 2.2.2.17

Answer

The mean and the median are both six.

Exercise 2.7.17

The mean and median for the data are the same.

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

Is the data perfectly symmetrical? Why or why not?

Exercise 2.7.18

Which is the greatest, the mean, the mode, or the median of the data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

Answer

The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.

Exercise 2.7.19

Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

Exercise 2.7.20

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Answer

The mean tends to reflect skewing the most because it is affected the most by outliers.

Exercise 2.7.21

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

2.3.1: Measures of Center and Spread Part 1

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the median. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced “x bar”): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

$$1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4$$

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \quad (2.3.1.1)$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \quad (2.3.1.2)$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression

$$\frac{n+1}{2} \quad (2.3.1.3)$$

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \quad (2.3.1.4)$$

The median is the 49th value in the ordered data. If the total number of data values is 100, then

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \quad (2.3.1.5)$$

The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.3.1.1

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

Calculate the mean and the median.

Answer

The calculation for the mean is:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+\ldots+35+37+40+(44)(2)+47]}{40} = 23.6 \quad (2.3.1.6)$$

To find the median, M, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \quad (2.3.1.7)$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

$$M = \frac{24+24}{2} = 24 \quad (2.3.1.8)$$

Calculator

To find the mean and the median:

Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.

Enter data into the list editor. Press STAT 1:EDIT.

Put the data values into list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

$\bar{x}$ = 23.6, M = 24

Exercise 2.3.1.1

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11; 12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21; 21; 22; 22; 23; 24; 24; 24; 24

Answer

Mean: $3+4+5+7+7+7+7+8+8+9+9+10+10+10+10+10+11+12+12+13+14+14+15+15+17+17+18+19+19+19+21+21+22+22+23+24+24+24 = 544$

$$\frac{544}{39} = 13.95 \quad (2.3.1.9)$$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example 2.3.1.2

Suppose that in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the "center": the mean or the median?

Solution

$$\bar{x} = \frac{5,000,000+49(30,000)}{50} = 129,400 \quad (2.3.1.10)$$

$M = 30,000$

(There are 49 people who earn \$30,000 and one person who earns \$5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Exercise 2.3.1.2

In a sample of 60 households, one house is worth \$2,500,000. Half of the rest are worth \$280,000, and all the others are worth \$315,000. Which is the better measure of the “center”: the mean or the median?

Answer

The median is the better measure of the “center” than the mean because 59 of the values are \$280,000 and one is \$2,500,000. The \$2,500,000 is an outlier. Either \$280,000 or \$315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 2.3.1.3

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Answer

The most frequent score is 72, which occurs five times. Mode = 72.

Exercise 2.3.1.3

The number of books checked out from the library from 25 students are as follows:

0; 0; 0; 1; 2; 3; 3; 4; 4; 5; 5; 7; 7; 7; 7; 8; 8; 8; 9; 10; 10; 11; 11; 12; 12

Find the mode.

Answer

The most frequent number of books is 7, which occurs four times. Mode = 7.

Example 2.3.1.4

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

> The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Exercise 2.3.1.4

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the “center”?

Answer

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don’t represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$

# of movies	Relative Frequency
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$mean = \frac{\text{data sum}}{\text{number of data values}}. \tag{2.3.1.11}$$

We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is

$$\frac{\text{lower boundary}+\text{upper boundary}}{2}. \tag{2.3.1.12}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{2.3.1.13}$$

where f is the frequency of the interval and m is the midpoint of the interval.

Example 2.3.1.5

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution

- Find the midpoints for all intervals

Grade Interval	Midpoint

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint.
  $53.25(1)+59.5(0)+65.5(4)+71.5(4)+77.5(2)+83.5(3)+89.5(4)+95.5(1)=1460.25$
- $\mu=\dfrac{\sum fm}{\sum f}=\dfrac{1460.25}{19}=76.86$

Exercise 2.3.1.5

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Answer

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.

$$Mean=(1.75)(3)+(5.5)(7)+(9.5)(12)+(13.5)(7)+(17.5)(9)=409.75 \quad (2.3.1.14)$$

References

1. Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).
2. “Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

$$\mu = \frac{\sum fm}{\sum f} \qquad (2.3.1.15)$$

where f = interval frequencies and m = interval midpoints.

Exercise 2.6.6

Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

Exercise 2.6.7

Calculate the mean.

Answer

Mean: $16+17+19+20+20+21+23+24+25+25+25+26+26+27+27+27+28+29+30+32+33+33$; $+34+35+37+39+40=738$

$\frac{738}{27}=27.33$

Exercise 2.6.8

Identify the median.

Exercise 2.6.9

Identify the mode.

Answer

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

Exercise 2.6.10

sample mean = $\bar{x}$ = _____

Exercise 2.6.11

median = _____

Answer

4

Bringing It Together

Exercise 2.6.12

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\bar{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

Figure 2.3.1.1: This shows two histograms. The first histogram shows a fairly symmetrical distribution with a mode of 6. The second histogram shows a uniform distribution.

1. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

Figure 2.3.1.2: This shows two horizontal boxplots. The first boxplot is graphed over a number line from 0 to 21. The first whisker extends from 0 to 1. The box begins at the first quartile, 1, and ends at the third quartile, 14. A vertical, dashed line marks the median at 6. The second whisker extends from the third quartile to the largest value, 21. The second boxplot is graphed over a number line from 0 to 12. The first whisker extends from 0 to 4. The box begins at the first quartile, 4, and ends at the third quartile, 9. A vertical, dashed line marks the median at 6. The second whisker extends from the third quartile to the largest value, 12.

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
			Total = 20

Number of years	Frequency	Number of years	Frequency
20	3		
			Total = 20

Exercise 2.6.13

What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Answer

a

Exercise 2.6.14

What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise 2.6.15

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Answer

b

Glossary

Frequency Table
a data representation in which grouped data is displayed along with the corresponding frequencies

Mean
a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted 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 $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median
a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint
the mean of an interval in a frequency table

Mode
the value that appears most frequently in a set of data

2.3.2: Measures of Center and Spread Part 2

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.

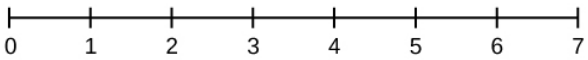


Figure 2.3.2.1

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer

- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation value = mean + (#ofSTDEVs)(standard deviation) can be expressed for a sample and for a population.

- sample:

$$x = \bar{x} + (\#\text{ofSTDEV})(s) \tag{2.3.2.1}$$

- Population:

$$x = \mu + (\#\text{ofSTDEV})(s) \tag{2.3.2.2}$$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol $\bar{x}$ is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The variance is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by ***n* – 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{2.3.2.3}$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \tag{2.3.2.4}$$

For the sample standard deviation, the denominator is $n - 1$, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{2.3.2.5}$$

or

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \tag{2.3.2.6}$$

For the population standard deviation, the denominator is N, the number of items in the population.

In Equations 2.3.2.4 and 2.3.2.6, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Section 2.6. How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the **sampling variability of a statistic** by its standard error.

The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in Chapter 7. The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example 2.3.2.1

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of n = 20 fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	Freq.	Deviations	*Deviations*^{2}	(Freq.)(*Deviations*^{2})
x	f	$(x - \bar{x})$	$(x - \bar{x})2$	$(f)(x - \bar{x})2$
9	1	9 – 10.525 = –1.525	(–1.525)2 = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20-1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891$$

and this is rounded to two decimal places, $s = 0.72$.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: $x = \bar{x}$ + (#ofSTDEVs)(s)
- For a population: $x = \mu$ + (#ofSTDEVs)σ
- For this example, use $x = \bar{x}$ + (#ofSTDEVs)(s) because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
c. Find the value that is two standard deviations below the mean. Find ($\bar{x}$ – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

a.
- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- $\bar{x}$ = 10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

b. $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$
c. $(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$
d.
- $(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

? Exercise 2.8.1

On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Answer

μ = 30.68

$s = 6.09$

$(\bar{x} + 2s = 30.68 + (2)(6.09) = 42.86$.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value

nine. **If you add the deviations, the sum is always zero**. (For Example 2.3.2.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by ($n - 1$) gives a better estimate of the population variance.

> Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example 2.3.2.2

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. *IQR*
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Answer

a. See Table
b.
   i. The sample mean = 73.5
   ii. The sample standard deviation = 17.9
   iii. The median = 73
   iv. The first quartile = 61
   v. The third quartile = 90
   vi. *IQR* = 90 – 61 = 29
c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5 =$ the ending value; No data values fall on an interval boundary.

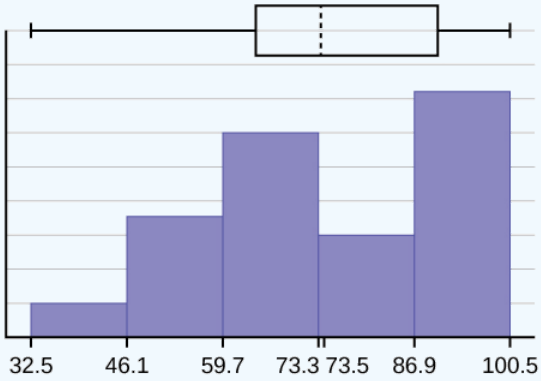


Figure 2.3.2.2.

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater $(73 - 33 = 40)$ than the spread in the upper 50% $(100 - 73 = 27)$. The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
100	1	0.032	0.998 (Why isn't this value 1?)

? Exercise 2.3.2.2

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Answer

$\mu = 9.3$ and $s = 2.2$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \qquad (2.3.2.7)$$

where f interval frequencies and $m =$ interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

✓ Example 2.3.2.3

Find the standard deviation for the data in Table 2.3.2.3

Table 2.3.2.3

Class	Frequency, *f*	Midpoint, *m*	m^2	$\bar{x}$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $\bar{x}$ = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-\bar{x})^2}{n-1}}$ where s_x = sample standard deviation, $\bar{x}$ = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Find the standard deviation for the data from the previous example

Class	0-2	3-5	6-8	9–11	12–14	15–17
Frequency, f	1	6	10	7	0	2

First, press the **STAT** key and select **1:Edit**

Figure 2.3.2.3

Input the midpoint values into **L1** and the frequencies into **L2**

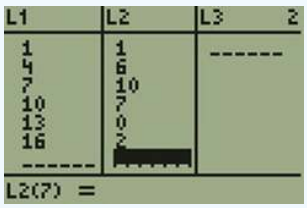


Figure 2.3.2.4

Select **STAT**, **CALC**, and **1: 1-Var Stats**

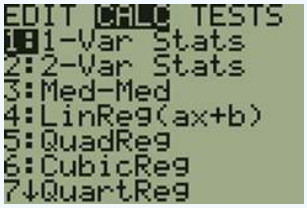


Figure 2.3.2.5

Select **2nd** then **1** then , **2nd** then **2 Enter**

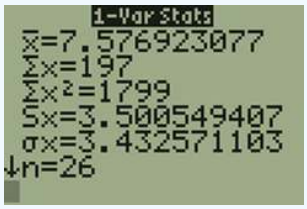


Figure 2.3.2.6

You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\text{ofSTDEVs} = \dfrac{\text{value-mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = \dfrac{x - \bar{x}}{s}$

Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example 2.3.2.4

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Answer

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \#\text{ofSTDEVs} = \left(\frac{\text{value-mean}}{\text{standard deviation}}\right) = \left(\frac{x + \mu}{\sigma}\right)$$

For John,

$$z = \#\text{ofSTDEVs} = \left(\frac{2.85 - 3.0}{0.7}\right) = -0.21$$

For Ali,

$$z = \#\text{ofSTDEVs} = \left(\frac{77 - 80}{10}\right) = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Exercise 2.3.2.4

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Answer

For Angie:

$$z = \left(\frac{26.2 - 27.2}{0.8}\right) = -1.25$$

For Beth:

$$z = \left(\frac{27.3 - 30.1}{1.4}\right) = -2$$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

1. Data from Microsoft Bookshelf.
2. King, Bill."Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$. .

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \tag{2.3.2.8}$$

where $s_x = \text{sample standard deviation}$ and $\bar{x} = \text{sample mean}$

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

? Exercise 2.8.4

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Answer

s = 34.5

Exercise 2.8.5

Find the value that is one standard deviation below the mean.

Exercise 2.8.6

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Answer

For Fredo:

$$z = \frac{0.158 - 0.166}{0.012} = -0.67$$

For Karl:

$$z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's *z*-score of –0.67 is higher than Karl's *z*-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise 2.8.7

Use Table to find the value that is three standard deviations:

- above the mean
- below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise 2.8.5

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15

Daily Low Temperature	Frequency
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Answer

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Bringing It Together

Exercise 2.8.7

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

a. Find the sample mean $\bar{x}$.
b. Find the approximate sample standard deviation, s.

Answer

a. 1.48
b. 1.12

Exercise 2.8.8

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let $X =$ the number of pairs of sneakers owned. The results are as follows:

X	Frequency

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40^{th} percentile.
k. Find the 90^{th} percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

Exercise 2.8.9

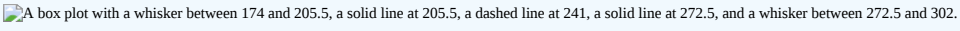
Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _____ to _____.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:
   i. the population mean, μ.
   ii. the population standard deviation, σ.
   iii. the weight that is two standard deviations below the mean.
   iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Answer

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5
e.  A box plot with a whisker between 174 and 205.5, a solid line at 205.5, a dashed line at 241, a solid line at 272.5, and a whisker between 272.5 and 302.
f. 205.5, 272.5
g. sample
h. population
i. i. 236.34
ii. 37.50
iii. 161.34
iv. 0.84 std. dev. below the mean
j. Young

Exercise 2.8.10

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; –1; 2; 0; 5; –3; 1; –1; 6; 5; –2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

Exercise 2.8.11

Refer to Figure determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

<figure >

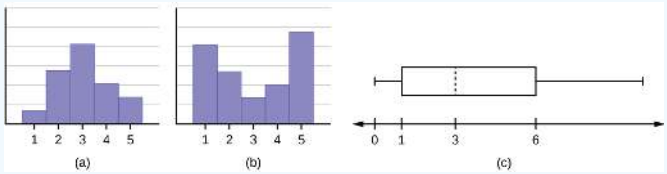


Figure 2.3.2.6.

</figure>

a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

Answer

a. True
b. True
c. True
d. False

Exercise 2.8.12

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65^{th} percentile.
d. Find the 10^{th} percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _____ days to _____ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise 2.8.13

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Answer

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.
c. mode
d. 8628.74
e. 6943.88

f. –0.09

Use the following information to answer the next two exercises. $X =$ the number of days per week that 100 clients use a particular exercise facility.

x	**Frequency**
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise 2.8.14

The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise 2.8.15

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Answer

a

Exercise 2.8.16

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the Table.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	

# of books	Freq.	Rel. Freq.
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR*to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

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.

Contributors and Attributions

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

CHAPTER OVERVIEW

3: Probability

3.1: Basics of Probability

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, $S = \{\text{H, T}\}$ where $\text{H} =$ heads and $\text{T} =$ 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 $P(\text{A})$.

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).

- $P(\text{A}) = 0$ means the event A can never happen.
- $P(\text{A}) = 1$ means the event A always happens.
- $P(\text{A}) = 0.5$ 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 $\{\text{HH, TH, HT,TT}\}$ where $\text{T} =$ tails and $\text{H} =$ heads. The sample space has four outcomes. $\text{A} =$ getting one head. There are two outcomes that meet this condition $\{\text{HT, TH}\}$, so $P(\text{A}) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event $\text{E} =$ rolling a number that is at least five. There are two outcomes {5, 6}. $P(\text{E}) = \frac{2}{6}$. 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 $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. 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 A AND B = 4, 5 .

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.

Example 3.1.1

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ _____
b. Let event $A =$ the even numbers. $A =$ _____
c. The probability of event A is $P(A) =$ _____.
d. Let event $B =$ numbers greater than 13. $B =$ _____
e. The probability of event B is $P(B) =$ _____
f. A AND B = _____
g. The probability of event A AND B is $P(\text{A AND B}) =$ _____.
h. A OR B = _____
i. The probability of event A OR B is $P(\text{A OR B}) =$ _____.

Answer

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$
c. $P(A) = \frac{9}{19}$
d. $B = \{14, 15, 16, 17, 18, 19\}$
e. $P(B) = \frac{6}{19}$
f. A AND B $= \{14, 16, 18\}$,
g. $P(\text{A AND B}) = \frac{3}{19}$
h. A OR $B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
i. $P(\text{A OR B}) = \frac{12}{19}$

Example 3.1.2

A fair, six-sided die is rolled. For parts (b) through (h), 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. Sample space S =.
b. Event T = the outcome is two.
c. Event A = the outcome is an even number.
d. Event B = the outcome is less than four.
e. A AND B

f. A OR B
g. Event N = the outcome is a prime number.
h. Event I = the outcome is seven.

Solution

a. $\mathrm{S} = \{1, 2, 3, 4, 5, 6\}$
b. $\mathrm{T} = \{2\}$, $P(\mathrm{T}) = \frac{1}{6}$
c. $A = \{2, 4, 6\}$, $P(\mathrm{A}) = \frac{1}{2}$
d. $\mathrm{B} = \{1, 2, 3\}$, $P(\mathrm{B}) = \frac{1}{2}$
e. A AND B $= 2$, $P(\mathrm{A\ AND\ B}) = \frac{1}{6}$
f. A OR B $= \{1, 2, 3, 4, 6\}$, $P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
g. $\mathrm{N} = \{2, 3, 5\}$, $P(\mathrm{N}) = \frac{1}{2}$
h. A six-sided die does not have seven dots. $P(7) = 0$.

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

$0 \leq P(\mathrm{A}) \leq 1$

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 $\mathrm{P(A\ OR\ B)} = P(\mathrm{A}) + P(\mathrm{B})$.
- $P(\mathrm{S}) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Conditional Probability of *A* GIVEN *B*

$P(\text{A}|\text{B})$ 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 A OR B if the outcome is in A or is in B or is in both A and B.

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.

Exercise 3.1.1

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. $P(\text{A}) =$ _____, $P(\text{B}) =$ _____
d. A AND B = _____, A OR B = _____
e. $P(\text{A AND B}) =$ _____, $P(\text{A OR B}) =$ _____

Answer

a. $\text{S} = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4),(3,1),(3,2),(3,3),(3,4)\}$
b. $\text{A} = \{(1,1),(1,3),(2,2),(2,4),(3,1),(3,3)\}$
$\text{B} = \{(2,1),(2,2),(2,3),(2,4),(3,1),(3,2),(3,3),(3,4)\}$
c. $P(\text{A}) = \frac{1}{2}$, $P(\text{B}) = \frac{2}{3}$
d. $\text{A AND B} = \{(2,2),(2,4),(3,1),(3,3)\}$
$\text{A OR B} = \{(1,1),(1,3),(2,1),(2,2),(2,3),(2,4),(3,1),(3,2),(3,3),(3,4)\}$
e. $P(\text{A AND B}) = \frac{1}{3}$, $P(\text{A OR B}) = \frac{5}{6}$

Exercise 3.1.2

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 is male or has short hair.
b. The probability that a student is a female and has long hair.
c. The probability that a student is male, given that the student has long hair.
d. The probability that a student has long hair, given that the student is male.
e. Of all the female students, the probability that a student has short hair.
f. Of all students with long hair, the probability that a student is female.

g. The probability that a student is female or has long hair.
h. The probability that a randomly selected student is a male student with short hair.
i. The probability that a student is female.

Answer

a. $P(\text{M OR S})$
b. $P(\text{F AND L})$
c. $P(\text{M}|\text{L})$
d. $P(\text{L}|\text{M})$
e. $P(\text{S}|\text{F})$
f. $P(\text{F}|\text{L})$
g. $P(\text{F OR L})$
h. $P(\text{M AND S})$
i. $P(\text{F})$

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 3.1.3

Find $P(\text{H})$.

? Exercise 3.1.4

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{15}{42} = \frac{5}{14} = 0.36$

? Exercise 3.1.5

Find $P(\text{F})$.

? Exercise 3.1.6

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 3.1.7

Find $P(\text{B})$.

Exercise 3.1.8

Find $P(\text{G})$.

Answer

$P(\text{G}) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise 3.1.9

Find $P(\text{P})$.

Exercise 3.1.10

Find $P(\text{R})$.

Answer

$P(\text{R}) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise 3.1.11

Find $P(\text{Y})$.

Exercise 3.1.12

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 3.1.13

Find $P(\text{A})$.

Exercise 3.1.14

Find $P(\text{E})$.

Answer

$P(\text{E}) = \frac{47}{194} = 0.24$

Exercise 3.1.15

Find $P(\text{F})$.

Exercise 3.1.16

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{23}{194} = 0.12$

Exercise 3.1.17

Find $P(\text{O})$.

Exercise 3.1.18

Find $P(\text{S})$.

Answer

$P(\text{S}) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise 3.1.18

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise 3.1.20

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 3.1.21

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise 3.1.22

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.

Figure 3.1.1 .

Let B = the event of landing on blue.

Let R = the event of landing on red.

Let G = the event of landing on green.

Let Y = the event of landing on yellow.

Exercise 3.1.23

If you land on Y, you get the biggest prize. Find $P(\text{Y})$.

Exercise 3.1.24

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 I = the event that a player in an infielder.

Let O = the event that a player is an outfielder.

Let H = the event that a player is a great hitter.

Let N = the event that a player is not a great hitter.

Exercise 3.1.25

Write the symbols for the probability that a player is not an outfielder.

Exercise 3.1.26

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Answer

$P(\text{O OR H})$

Exercise 3.1.27

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise 3.1.28

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 3.1.29

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise 3.1.30

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 3.1.31

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise 3.1.32

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 3.1.33

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise 3.1.34

Write the symbols for the probability that a player is an infielder.

Answer

$P(\text{I})$

Exercise 3.1.35

What is the word for the set of all possible outcomes?

Exercise 3.1.36

What is conditional probability?

Answer

The likelihood that an event will occur given that another event has already occurred.

Exercise 3.1.37

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 3.1.38

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.

What does $P(\text{E OR M})$ mean in words?

Answer

the probability of landing on an even number or a multiple of three

---

3.2: The Addition Rules of Probability

Your favorite professional basketball team either won or lost their last game. Winning and losing are mutually exclusive events.

Mutually Exclusive Events

A and B are mutually exclusive events (or disjoint 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 $\text{X} = \{1, 2, 3, 4, 5\}$, $\text{Y} = \{4, 5, 6, 7, 8\}$ and $\text{Z} = \{7, 9\}$. Events X and Y both have 4 and 5. Thus, $\text{X AND Y} = \{4, 5\}$.

Because there are two shared outcomes from the sample space S, the probability of X AND Y is

$$P(\text{X AND Y}) = \frac{2}{10}$$

.

Since $P(\text{X AND Y})$ is not equal to zero, X and Y **are not** mutually exclusive.

However, events X and Z have no outcomes (numbers) in common. So, $P(\text{X AND Z}) = \frac{0}{10} = 0$. Therefore, X and Y **are** mutually exclusive.

The Addition Rule of Probability

The probability of two mutually exclusive events A OR B (two events that share no outcomes) is

$$P(A \text{ OR } B) = P(A) + P(B)$$

The probability of two **non**-mutually exclusive events A OR B (two events that share outcomes) is

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

Using the example from above, where the sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and events $\text{X} = \{1, 2, 3, 4, 5\}$, $\text{Y} = \{4, 5, 6, 7, 8\}$ and $\text{Z} = \{7, 9\}$.

Since events X and Z are mutually exclusive then the probability of X OR Z

$$P(X \text{ OR } Z) = P(X) + P(Z) = \frac{5}{10} + \frac{2}{10} = \frac{7}{10}. \tag{3.2.1}$$

Since events X and Y are **not** mutually exclusive then the probability of X OR Y

$$P(X \text{ OR } Y) = P(X) + P(Y) - P(X \text{ AND } Y) = \frac{5}{10} + \frac{5}{10} - \frac{2}{10} = \frac{8}{10}. \tag{3.2.2}$$

Below we will see our first **contingency table**, which is a table with categories in both the horizontal and vertical direction. As you see below, the contingency table describes cell phone users versus speeding violations.

Example 3.2.1

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

	Speeding violation in the last year	No speeding violation in the last year	Total
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 $P(\text{Person is a cell phone user OR person had no violation in the last year})$.

Answer

a. $\dfrac{\text{number of cell phone users}}{\text{total number in study}} = \dfrac{305}{755}$
b. $\dfrac{\text{number that had no violation}}{\text{total number in study}} = \dfrac{685}{755}$
c. $\dfrac{280}{755}$
d. $\left(\dfrac{305}{755} + \dfrac{685}{755}\right) - \dfrac{280}{755} = \dfrac{710}{755}$

Exercise 3.2.2

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 $P(\text{athlete stretches before exercising and no injury in the last year})$?
c. What is $P(\text{athlete stretches before exercising or no injury in the last year})$?

Answer

a. $P(\text{athlete stretches}) = \dfrac{350}{800} = 0.4375$
b. $P(\text{athlete stretches AND no injury in the last year}) = \dfrac{295}{800} = 0.3688$
c. $P(\text{athlete stretches OR no injury in the last year})$
$= P(\text{athlete stretches }) + P(\text{no injurty in the last year}) - P(\text{athlete stretches AND no injury in the last year})$
$= \dfrac{350}{800} + \dfrac{514}{800} - \dfrac{295}{800}$
$= \dfrac{569}{800} = 0.7113$

Example 3.2.3

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. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.
   1. Find $P(\text{F})$.
   2. Find $P(\text{M})$.
   3. Find $P(\text{F AND M})$.
   4. Find $P(\text{F OR M})$.

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.

1. $P(\text{F}) = \frac{45}{100}$
2. $P(\text{M}) = \frac{25}{100}$
3. $P(\text{F AND M}) = \frac{11}{100}$
4. $P(\text{F OR M}) = P(\text{F}) + P(\text{M}) - P(\text{F AND M}) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Complement of an event

The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that

$$P(\text{A}) + P(\text{A}') = 1.$$

In other words,

$$P(\text{A}') = 1 - P(\text{A})$$

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

$$P(\text{A}) + P(\text{A}') = \frac{4}{6} + \frac{2}{6} = 1.$$

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 3.2.1

Find P(musician is a female).

Exercise 3.2.2

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 3.2.3

Find $P(\text{musician is a female OR is self taught})$.

Answer

\(P(\text{musician is a female OR is self taught}) \\

= P(\text{musician is a female}) + P(\text{self taught}) - P(\text{musician is a female OR is self taught})\\

=\frac{72}{130} + \frac{31}{130} - \frac{12}{130}\\

=\frac{91}{130}\)

Exercise 3.2.4

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.

References

1. “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).

Glossary

mutually exclusive (or disjoint) events
events that cannot happen at the same time

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.

complement of an event

The complement of event A consists of all outcomes that are NOT in A.

3.3: Multiplication Rule for Independent Events

On days that your favorite basketball team plays, does it affect your ability to find parking on campus? If not, then we call these two events independent.

Independent Events

Two events are independent if the following are true:

- For the probably of event A given event B, $P(\text{A}|\text{B}) = P(\text{A})$
- For the probably of event B given event A, $P(\text{B}|\text{A}) = P(\text{B})$
- $P(\text{A AND B}) = P(\text{A})P(\text{B})$

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.

Note, *independent* and *mutually exclusive* do not mean the same thing. For example, campus parking availability is not affected by your favorite basketball team game days (i.e., independent events). However, that does not mean they cannot happen at the same time (i.e., not mutually exclusive).

Example 3.3.1

Let event 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

$$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{P(\text{B})} = \frac{0.08}{0.2} = 0.4 = P(\text{A}) \tag{3.3.1}$$

The events are independent because $P(\text{A}|\text{B}) = P(\text{A})$.

Example 3.3.2

Let event G = taking a math class. Let event H = taking a science class. Then, 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. Show that $P(\text{G}|\text{H}) = P(\text{G})$.
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.

Multiplication Rule for Independent Events

Given events A and B are independent, then

$$P(\text{A AND B}) = P(\text{A}) \cdot P(\text{B})$$

Example 3.3.3

In a bag of colored blocks, 5 are red, 4 are green, and 3 are blue. Draw three blocks. After each draw, you replace the block into the bag. For parts (b)-(c), find the probabilities of the events.

a. Is event F described below independent?
b. Let F = the event drawing a red, green, and blue.
c. Let G = the event drawing two red and one blue.
d. Let H = the event of drawing all red.

Answer a

Yes. After each draw, you replace the block into the bag. So, after drawing a red block, you replace it into the bag, which does not affect the probability of drawing any subsequent blocks. This is also true for events G and H.

Answer b

$$P(\text{G}) = P(\text{red AND green AND blue}) = P(\text{red}) \cdot P(\text{green}) \cdot P(\text{blue}) = \frac{5}{12} \cdot \frac{4}{12} \cdot \frac{3}{12} = 0.0347$$

Answer c

$$P(\text{G}) = P(\text{red AND red AND blue}) = P(\text{red}) \cdot P(\text{red}) \cdot P(\text{blue}) = \frac{5}{12} \cdot \frac{5}{12} \cdot \frac{3}{12} = 0.0434$$

Answer d

$$P(\text{H}) = P(\text{red AND red AND red}) = P(\text{red}) \cdot P(\text{red}) \cdot P(\text{red}) = \frac{5}{12} \cdot \frac{5}{12} \cdot \frac{5}{12} = \left(\frac{5}{12}\right)^3 = 0.0723$$

Suppose you flip a coin three times. The sample space $S = \{HHH, HHT, HTT, HTH, THT, THH, TTH, TTT\}$. We can see that there are seven outcomes with *at least one* head. (The only outcome that does not have at least one head is "TTT".)

What is the probability of getting at least one head when a coin is flipped three times?

\(P(\text{at least one head when a coin is flipped three times}) = \dfrac{7}{8}=0.875\)

Luckily for us, this example required us to flip a coin three times which made finding the probability above easier by counting the outcomes. What if we were asked to flip the coin 5 times? or 10 times? or 20 times? This would be troublesome as listing the sample space would take a while.

Turns out, there is a formula for such situations.

"At least one" Probability

$$P(\text{at least one}) = 1 - P(\text{none})$$

Example 3.3.4

You are asked to flip a coin three times. What is the probability that you flip at least one head?

Answer

Using the formula above, we get

$$P(\text{at least one head}) = 1 - P(\text{none are heads})$$

On the right side of the formula, we need to find the probability that none of the three flips are heads. This is the same thing as saying find the probability that all three flips were tails. So,

$$
\begin{aligned}
&P(\text{none are heads}) \\
&= P(\text{tail AND tail AND tail}) \\
&= P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \\
&= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
&= 0.125
\end{aligned}
$$

Now,

$$P(\text{at least one head}) = 1 - P(\text{none are heads}) = 1 - 0.125 = 0.875$$

Which, you'll notice is the same answer as above when this problem was done by counting.

Example 3.3.5

You are asked to flip a coin ten times. What is the probability that you flip at least one head?

Answer

Using the formula above, we get

$$P(\text{at least one head}) = 1 - P(\text{none are heads})$$

On the right side of the formula, we need to find the probability that none of the ten flips are heads. This is the same thing as saying find the probability that all ten flips were tails. So,

$$
\begin{aligned}
&P(\text{none are heads}) \\
&= P(\text{tail AND tail AND tail AND tail AND tail AND tail AND tail AND tail AND tail AND tail}) \\
&= P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \cdot P(\text{tail}) \\
&= \left(\frac{1}{2}\right)^{10} = 0.001
\end{aligned}
$$

Now,

$$P(\text{at least one head}) = 1 - P(\text{none are heads}) = 1 - 0.001 = 0.999$$

Example 3.3.6

At the local pet shop, the manager has noticed a that a fish food manufacture has been sending expired fish food in the big bulk orders to the store. On average, they send 5 out of 100 bags that are expired. Before putting the product on the store floor, she will randomly select three bags of fish food. If at least one of the randomly selected three bags is expired, she will return the entire order to the manufacturer. What is the probability that the entire order will be returned?

Answer

Using the formula above, we get

$$P(\text{at least one of the randomly selected bags is bad}) = 1 - P(\text{none of the randomly selected bags are bad})$$

On the right side of the formula, we need to find the probability that none of the randomly selected bags are bad. This is the same thing as saying find the probability that all three bags are good. So,

$$P(\text{none are heads}) = P(\text{good AND good AND good}) = P(\text{good}) \cdot P(\text{good}) \cdot P(\text{good}) = \frac{95}{100} \cdot \frac{95}{100} \cdot \frac{95}{100} = 0.8574$$

Now,

$$P(\text{at least one of the randomly selected bags is bad}) = 1 - P(\text{none of the randomly selected bags are bad}) \quad = 1\text{-}0.8574\text{=}0.1426\backslash)$$

The probability that the manager will send back the entire order to the manufacture is 0.1426

Glossary

Independent Events

If one event does not affect the other, then the two events are independent.

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.

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: General Multiplication Probability

It is a hot summery day. You are looking forward to an unopened pint of (cow milk based) ice cream that you have waiting for you in the freezer. When you reach for the carton and take off the lid, you see that there are only two spoonfuls left. Who is the ice cream thief in your household?

You live with your family, which consists of your grandma, both two parents, and three siblings; a total of six suspects. At this point, you have a probability of $\frac{1}{6}$ of guessing correctly who is the thief.

However, you know your grandma's gums are sensitive to cold food. Also, one parent and two siblings are lactose intolerant. That leaves only two possible suspects: the other parent and one sibling. At this point, you have a probability of $\frac{1}{2}$ of guessing correctly who is the thief. This is an example of **conditional probability**, where instead of looking at the entire sample space, we look at a smaller group to find the correct probability.

Conditional Probability

The conditional probability of A given B is written $P(\mathrm{A}|\mathrm{B})$. $P(\mathrm{A}|\mathrm{B})$ 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 $P(\mathrm{A}|\mathrm{B})$ is

$$P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}}$$

where $P(\mathrm{B})$ is greater than zero.

Although the formula can be intimidating, there are times when we can find conditional probabilities without it.

Example

Suppose we toss one fair, six-sided die. The sample space $\mathrm{S} = \{1, 2, 3, 4, 5, 6\}$. Let event A be rolling a 2 or 3, $\mathrm{A} = \{2, 3\}$. Let event B be rolling an even number, $\mathrm{B} = \{2, 4, 6\}$.

What is the probability of getting event A, given we know event B happened? In other words, what is $P(\mathrm{A}|\mathrm{B})$? Since there three outcomes in event B. Both events A and B only share one outcome, "2". Then $P(\mathrm{A}|\mathrm{B}) = \frac{1}{3}$.

Example

Let's do the example above, but with using the formula.

Remember that S has six outcomes.

$$P(\mathrm{A}|\mathrm{B})$$
$$= \frac{\mathrm{P(A\ AND\ B)}}{P(\mathrm{B})}$$
$$= \frac{\frac{\text{the number of outcomes that are 2 or 3 AND even in S}}{6}}{\frac{\text{the number of outcomes that are even in S}}{6}}$$
$$= \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Notice, we get the same answer as in Example 1A.

Notation: A|B is equivalent to A GIVEN B

Example 3.4.2

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 A = the outcome is an even number.
b. Event B = the outcome is less than four.
c. A GIVEN B
d. B GIVEN A

Answer

a. $A = \{2, 4, 6\}$, $P(\text{A}) = \frac{1}{2}$
b. $\text{B} = \{1, 2, 3\}$, $P(\text{B}) = \frac{1}{2}$
c. $\text{A|B} = \{2\}$, $P(\text{A|B}) = \frac{1}{3}$
d. $\text{B|A} = \{2\}$, $P(\text{B|A}) = \frac{1}{3}$

With or Without Replacement

Sampling may be done with replacement or without replacement (Figure 3.4.1):

- **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*.

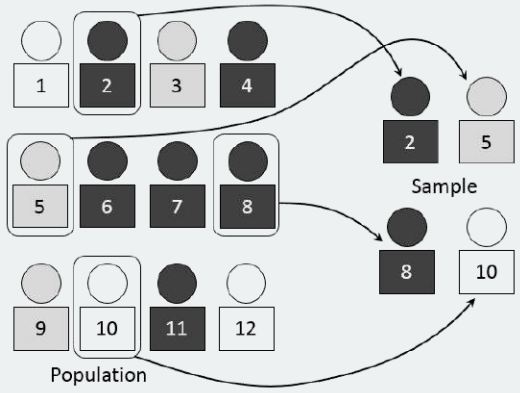


Figure 3.4.1 : A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". (CC BY-SA 4.0; Dan Kernler).

If it is not known whether A and B are independent or dependent, assume they are dependent until you can show otherwise.

Example 3.4.3

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. Each time you drew a card and replaced it before drawing another card is **sampling with replacement**.

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. Each time you drew a card and did not replace it into the deck before drawing another card is **sampling without replacement.**

Example 3.4.4

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 3.4.5

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

General Multiplication Rule

In Section 3.3, we learned the Multiplication Rule when events A and B are independent:

$$P(\text{A AND B}) = P(\text{A}) \cdot P(\text{B})$$

However, if events A and B are dependent, then

$$P(\text{A AND B}) = P(\text{A}) \cdot P(\text{B}|\text{A})$$

Example 3.4.6

A gum ball machine has 5 red balls, 6 green balls, and 5 yellow balls. Suppose you Find the following probabilities.

a. $P(\text{draw a red ball and then a green})$ **with** replacement
b. $P(\text{draw a red ball and then a green})$ **without** replacement
c. $P(\text{draw three red balls})$ **with** replacement

d. $P(\text{draw three red balls})$ **without** replacement
e. $P(\text{draw a red ball, two green balls, and one yellow ball})$ **with** replacement
f. $P(\text{draw a red ball, two green balls, and one yellow ball})$ **without** replacement

Answer a

$P(\text{draw a red ball and then a green replacement}) = P(\text{red}) \cdot P(\text{green}) = \frac{5}{16} \cdot \frac{6}{16} = 0.1172$

Since we are replacing this the ball after each draw, you recover the total number of gumballs each time, which was 16.

Answer b

$P(\text{draw a red ball and then a green}) = P(\text{red}) \cdot P(\text{green}) = \frac{5}{16} \cdot \frac{6}{15} = 0.128$

Since we are **not** replacing this the ball after each draw, the total number of gumballs each time decreases by one.

Answer c

$P(\text{draw three red balls}) = P(\text{red}) \cdot P(\text{red}) \cdot P(\text{red}) = \frac{5}{16} \cdot \frac{5}{16} \cdot \ frac516 = 0.3052$

Answer d

$P(\text{draw three red balls}) = P(\text{red}) \cdot P(\text{red}) \cdot P(\text{red}) = \frac{5}{16} \cdot \frac{4}{15} \cdot \frac{3}{14} = 0.0179$

Answer e

\(P(\text{draw a red ball, two green balls, and one yellow ball}) = P(\text{red}) \cdot P(\text{green}) \cdot P(\text{green}) \cdot P(\text{yellow})= \frac{5}{16} \cdot \frac{6}{16} \cdot \frac{6}{16} \cdot \frac{5}{16} = 0.0137\)

Answer f

\(P(\text{draw a red ball, two green balls, and one yellow ball}) = P(\text{red}) \cdot P(\text{green}) \cdot P(\text{green}) \cdot P(\text{yellow})= \frac{5}{16} \cdot \frac{6}{15} \cdot \frac{5}{14}\cdot \frac{5}{13} = 0.0172\)

Glossary

Independent Events

If one event does not affect the other, then the two events are independent.

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.

CHAPTER OVERVIEW

4: Discrete Probability Distributions

4.2: The Binomial Distribution

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn *with replacement* from a population of size N.

Three characteristics of a binomial experiment

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable $X =$ the number of successes obtained in the n independent trials. The mean, μ, and variance, σ^2, for the binomial probability distribution are

$$\mu = np \tag{4.2.1}$$

and

$$\sigma^2 = npq. \tag{4.2.2}$$

The standard deviation, σ, is then

$$\sigma = \sqrt{npq}. \tag{4.2.3}$$

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example 4.2.1

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable $X =$ the number of students who withdraw from the randomly selected elementary physics class.

Exercise 4.2.1

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a "success" be in this case?

Answer

a school that offers fruit in their lunch every day

Example 4.2.2

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Exercise 4.2.2

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

Answer

$P(x = 12)$

Example 4.2.3

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let $X =$ the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically.

Solution

$P(x > 10)$

Exercise 4.2.4

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

Answer

$P(x > 3)$

Example 4.2.5

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

a. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.
b. If we are interested in the number of students who do their homework on time, then how do we define X?
c. What values does x take on?
d. What is a "failure," in words?
e. If $p + q = 1$, then what is q?
f. The words "at least" translate as what kind of inequality for the probability question $P(x$ ____ $40)$.

Solution

a. failure
b. X = the number of statistics students who do their homework on time
c. 0, 1, 2, …, 50
d. Failure is defined as a student who does not complete his or her homework on time. The probability of a success is $p = 0.70$. The number of trials is $n = 50$.
e. $q = 0.30$
f. greater than or equal to ($\geq$). The probability question is $P(x \geq 40)$.

Exercise 4.2.5

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

Answer

This is a binomial problem because there is only a success or a failure, and there are a definite number of trials. The probability of a success stays the same for each trial.

Notation for the Binomial: $B =$ Binomial Probability Distribution Function

$$X \sim B(n, p) \tag{4.2.4}$$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p; $n =$ number of trials, $p =$ probability of a success on each trial.

Example 4.2.6

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20, p = 0.41$, and $q = 1 - 0.41 = 0.59$. $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (calculator or computer)

Go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate $(x =$ value$)$: binompdf$(n, p,$ number **)** if "number" is left out, the result is the binomial probability table.

To calculate $P(x \leq$ value$)$: binomcdf$(n, p,$ number$)$ if "number" is left out, the result is the cumulative binomial probability table.

For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,0.41,12). The result is $P(x \leq 12) = 0.9738$.

If you want to find $P(x = 12)$, use the pdf (binompdf). If you want to find $P(x > 12)$, use $1 - \text{binomcdf}(20, 0.41, 12)$.

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

The graph of $X \sim B(20, 0.41)$ is as follows:

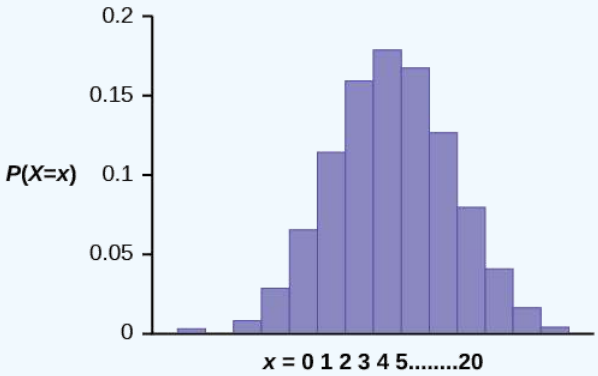


Figure 4.2.1 : The graph of $X \sim B(20, 0.41)$.

The *y*-axis contains the probability of x, where $X =$ the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$.

$$\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20. \quad (4.2.5)$$

Exercise 4.4.5

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

Answer

$P(x \leq 14) = 0.9695$

Example 4.2.7

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let $X =$ the number of pages that feature signature artists.

a. What values does x take on?
b. What is the probability distribution? Find the following probabilities:
   i. the probability that two pages feature signature artists
   ii. the probability that at most six pages feature signature artists
   iii. the probability that more than three pages feature signature artists.
c. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Answer

a. $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
b. $X \sim B(100, 8560)(100, 8560)$
   i. $P(x = 2) = \text{binompdf}\left(100, \frac{8}{560}, 2\right) = 0.2466$
   ii. $P(x \leq 6) = \text{binomcdf}\left(100, \frac{8}{560}, 6\right) = 0.9994$
   iii. $P(x > 3) = 1 - P(x \leq 3) = 1 - \text{binomcdf}\left(100, \frac{8}{560}, 3\right) = 1 - 0.9443 = 0.0557$
c. i. Mean $= np = (100)\left(\frac{8}{560}\right) = \frac{800}{560} \approx 1.4286$
   ii. Standard Deviation $= \sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

Exercise 4.2.7

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

a. What is the probability distribution for X?
b. Use your calculator to find the following probabilities:
   i. the probability that 25 adults in the sample prefer saving over spending
   ii. the probability that at most 20 adults prefer saving
   iii. the probability that more than 30 adults prefer saving
c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

Answer

a. $X \sim B(50, 0.6)$
b. Using the TI-83, 83+, 84 calculator with instructions as provided in Example:

i. $P(x = 25) = \text{binompdf}(50, 0.6, 25) = 0.0405$
ii. $P(x \leq 20) = \text{binomcdf}(50, 0.6, 20) = 0.0034$
iii. $(x > 30) = 1 - \text{binomcdf}(50, 0.6, 30) = 1 - 0.5535 = 0.4465$

c. i. Mean $= np = 50(0.6) = 30$
ii. Standard Deviation $= \sqrt{npq} = \sqrt{50(0.6)(0.4)} \approx 3.4641$

Example 4.2.8

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that at most eight people develop pancreatic cancer
d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Answer

a. $X \sim B(200, 0.0128)$

b. i. Mean $= np = 200(0.0128) = 2.56$
ii. Standard Deviation $= \sqrt{npq} = \sqrt{(200)(0.0128)(0.9872)} \approx 1.5897$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in Example:
$P(x \leq 8) = \text{binomcdf}(200, 0.0128, 8) = 0.9988$

d. $P(x = 5) = \text{binompdf}(200, 0.0128, 5) = 0.0707$
$P(x = 6) = \text{binompdf}(200, 0.0128, 6) = 0.0298$
So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

Exercise 4.2.8

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let $X =$ the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
d. Find the probability that DeAndre scored with more than 50 of these shots.

Answer

a. $X \sim B(80, 0.613)$

b. i. Mean $= np = 80(0.613) = 49.04$
ii. Standard Deviation $= \sqrt{npq} = \sqrt{80(0.613)(0.387)} \approx 4.3564$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in Example:
$P(x = 60) = \text{binompdf}(80, 0.613, 60) = 0.0036$

d. $P(x > 50) = 1 - P(x \leq 50) = 1 - \text{binomcdf}(80, 0.613, 50) = 1 - 0.6282 = 0.3718$

Example 4.2.9

The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the

second draw is $\frac{5}{15}$, when the first draw selects a student. The probability is $\frac{6}{15}$, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

? Exercise 4.2.9

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

Answer

This is not binomial because the names are not replaced, which means the probability changes for each time a name is drawn. This violates the condition of independence.

References

1. "Access to electricity (% of population)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...first&sort=asc (accessed May 15, 2015).
2. "Distance Education." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Distance_education (accessed May 15, 2013).
3. "NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).
4. Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," GALLUP® Economy, 2013. Available online at http://www.gallup.com/poll/162368/am...-spending.aspx (accessed May 15, 2013).
5. Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/N...eshman2011.pdf (accessed May 15, 2013).
6. "The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/publicat...k/geos/af.html (accessed May 15, 2013).
7. "What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at www.cancer.org/cancer/pancrea...key-statistics (accessed May 15, 2013).

Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

There are a fixed number of trials, n.

There are only two possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.

The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable $X =$ the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

Formula Review

- $X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p.
- $X =$ the number of successes in n independent trials
- $n =$ the number of independent trials
- X takes on the values $x = 0, 1, 2, 3, \ldots, n$
- $p =$ the probability of a success for any trial
- $q =$ the probability of a failure for any trial

- $p+q=1$
- $q=1-p$

The mean of X is $\mu=np$. The standard deviation of X is $\sigma=\sqrt{npq}$.

Use the following information to answer the next eight exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

Exercise 4.4.9

In words, define the random variable X.

Answer

$X=$ the number that reply "yes"

Exercise 4.4.10

$X \sim$ _____(_____,_____)

Exercise 4.4.11

What values does the random variable X take on?

Answer

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

Exercise 4.4.12

Construct the probability distribution function (PDF).

x	$P(x)$

Exercise 4.4.13

On average (μ), how many would you expect to answer yes?

Answer

5.7

Exercise 4.4.14

What is the standard deviation (σ)?

Exercise 4.4.15

What is the probability that at most five of the freshmen reply "yes"?

Answer

0.4151

Exercise 4.4.16

What is the probability that at least two of the freshmen reply “yes”?

Glossary

Binomial Experiment

a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

Bernoulli Trials

an experiment with the following characteristics:

1. There are only two possible outcomes called “success” and “failure” for each trial.
2. The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Binomial Probability 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 one) 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: $X\ B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

4.1.1: Discrete Probability Distributions Part 1

A discrete probability distribution function has two characteristics:

a. Each probability is between zero and one, inclusive.
b. The sum of the probabilities is one.

Example 4.1.1.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let $X =$ the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$

$P(x) =$ probability that X takes on a value x.

x	$P(x)$
0	$P(x=0) = \frac{2}{50}$
1	$P(x=1) = \frac{11}{50}$
2	$P(x=2) = \frac{23}{50}$
3	$P(x=3) = \frac{9}{50}$
4	$P(x=4) = \frac{4}{50}$
5	$P(x=5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because:

a. Each $P(x)$ is between zero and one, inclusive.
b. The sum of the probabilities is one, that is,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \qquad (4.1.1.1)$$

Exercise 4.1.1.1

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. Let $X =$ the number of times a patient rings the nurse during a 12-hour shift. For this exercise, $x = 0, 1, 2, 3, 4, 5$ $P(x) =$ the probability that X takes on value x. Why is this a discrete probability distribution function (two reasons)?

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	$P(x=4) = \frac{6}{50}$

X	$P(x)$
5	$P(x=5)=\frac{2}{50}$

Answer

Each $P(x)$ is between 0 and 1, inclusive, and the sum of the probabilities is 1, that is:

$$\frac{4}{50}+\frac{8}{50}+\frac{16}{50}+\frac{14}{50}+\frac{6}{50}+\frac{2}{50}=1 \quad (4.1.1.2)$$

Example 4.1.1.2

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

a. Let X = the number of days Nancy _____.
b. X takes on what values?
c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in Example. The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Solutions

a. Let X = the number of days Nancy attends class per week.

b. 0, 1, 2, and 3

c

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

Exercise 4.1.1.2

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is *X* and what values does it take on?

Answer

X is the number of days Jeremiah attends basketball practice per week. *X* takes on the values 0, 1, and 2.

Review

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

Use the following information to answer the next five exercises: A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution.

Let $X =$ the number of years a new hire will stay with the company.

Let $P(x) =$ the probability that a new hire will stay with the company *x* years.

Exercise 4.2.3

Complete Table using the data provided.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

Answer

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

Exercise 4.2.4

$P(x = 4) =$ _____

Exercise 4.2.5

$P(x \geq 5) =$ _____

Answer

0.10 + 0.05 = 0.15

Exercise 4.2.6

On average, how long would you expect a new hire to stay with the company?

Exercise 4.2.7

What does the column "$P(x)$" sum to?

Answer

1

Use the following information to answer the next six exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

Exercise 4.2.8

Define the random variable X.

Exercise 4.2.9

What is the probability the baker will sell more than one batch? $P(x > 1) =$ _____

Answer

0.35 + 0.40 + 0.10 = 0.85

Exercise 4.2.10

What is the probability the baker will sell exactly one batch? $P(x = 1) =$ _____

Exercise 4.2.11

On average, how many batches should the baker make?

Answer

1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45

Use the following information to answer the next four exercises: Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

Exercise 4.2.12

Define the random variable X.

Exercise 4.2.13

Construct a probability distribution table for the data.

Answer

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

Exercise 4.2.14

We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Use the following information to answer the next five exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

Exercise 4.2.15

Define the random variable X.

Answer

Let $X =$ the number of events Javier volunteers for each month.

Exercise 4.2.16

What values does x take on?

Exercise 4.2.17

Construct a PDF table.

Answer

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

Exercise 4.2.18

Find the probability that Javier volunteers for less than three events each month. $P(x < 3) =$ _____

Exercise 4.2.19

Find the probability that Javier volunteers for at least one event each month. $P(x > 0) =$ _____

Answer

1 – 0.05 = 0.95

Glossary

Probability Distribution Function (PDF)

a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

4.1.2: Discrete Probability Distributions Part 2

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 4.1.2.1

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 4.1.2.2

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	$P(x = 0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$
1	$P(x = 1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$

x	$P(x)$	$x * P(x)$	$(x-\mu)^2 \cdot P(x)$
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 4.1.2.2

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

$$(0)\frac{4}{50}+(1)\frac{8}{50}+(2)\frac{16}{50}+(3)\frac{14}{50}+(4)\frac{6}{50}+(5)\frac{2}{50}=0+\frac{8}{50}+\frac{32}{50}+\frac{42}{50}+\frac{24}{50}+\frac{10}{50}=\frac{116}{50}=2.32 \quad (4.1.2.1)$$

Example 4.1.2.2

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

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)=(1)(10^{-5})$$
$$=0.00001.$$

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 4.1.2.3

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 0.23cents.

Example 4.1.2.4

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 4.1.2.4

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	$\frac{2}{5}$	$\frac{0}{5}$
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 $(x) - \mu^2 P(x)$ and take the square root.

x	$P(x)$	$x * P(x)$	$(x-\mu)^2 P(x)$
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 4.1.2.5

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:

$$\sigma = \sqrt{\frac{18}{36}} \approx 0.7071. \quad (4.1.2.2)$$

Example 4.1.2.6

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 4.1.2.6

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).

CHAPTER OVERVIEW

5: Normal Probability Distribution

5.1: 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:

$$z = \frac{x - \mu}{\sigma} \tag{5.1.1}$$

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 5.1.1 produces 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 5.1.1

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 5.1.1):

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

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 5.1.1

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 5.1.2

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 5.1.2

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.

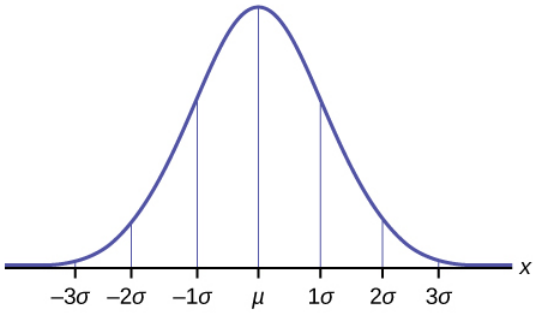


Figure 5.1.1

Example 5.1.3

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 5.1.3

Use the information in Example 5.1.3 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 $z = \dfrac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$z = \dfrac{176 - 170}{6.28}$, 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 $z = \dfrac{x - \mu}{\sigma}$ 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 5.1.4

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 5.1.1) 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 5.1.4

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 5.1.1) for $x_1 = 325$ is $z_1 = -1.15$.

The *z*-score (Equation 5.1.1) 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 5.1.5

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 5.1.5

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 5.1.6

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 5.1.6

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).

5.2: Area Under Any Normal Curve

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 $P(X < x)$.

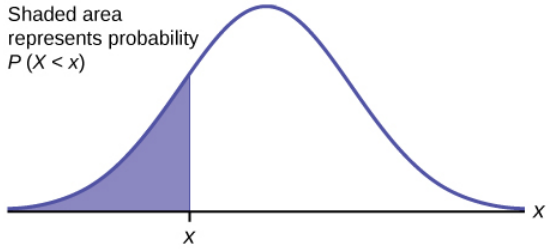


Figure 5.2.1.

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 5.2.1

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$

Exercise 5.2.1

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 5.2.2

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 $P(x > 65)$.

$$P(x > 65) = 0.3446$$

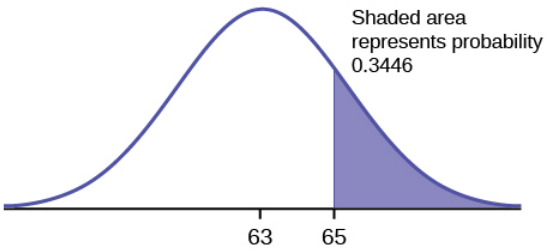


Figure 5.2.2.

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 $P(x < 85) = 1$.

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.

$k = 69.4$

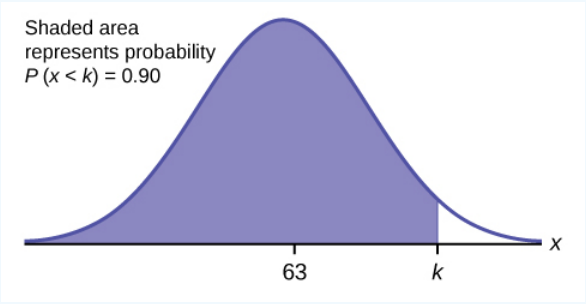


Figure 5.2.3.

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, 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 5.2.2

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

normalcdf$(10^{99}, 65, 68, 3) = 0.1587$

Example 5.2.3

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 $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

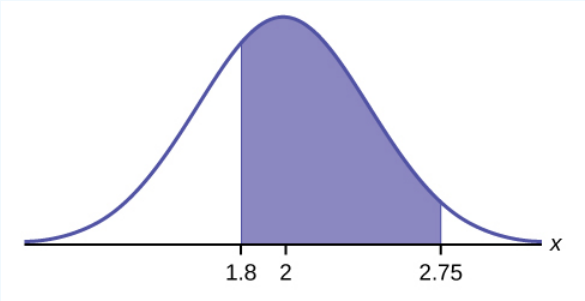


Figure 5.2.4.

$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

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 25th percentile,** k, where $P(x < k) = 0.25$.

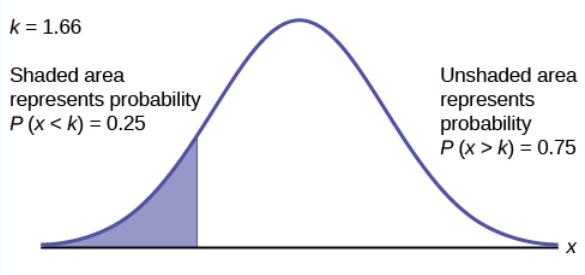


Figure 5.2.5.

$$\text{invNorm}(0.25, 2, 0.5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Exercise 5.2.3

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 5.2.4

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 80th 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 80th 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 5.2.4

a. Find the 30th 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 30th 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.)

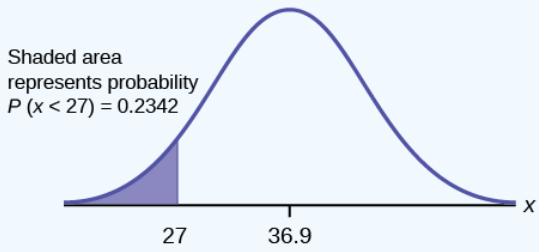


Figure 5.2.6.

$$\text{normalcdf}(0, 27, 36.9, 13.9) = 0.2342$$

Example 5.2.5

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 5.2.5

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 5.2.6

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. $\text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$

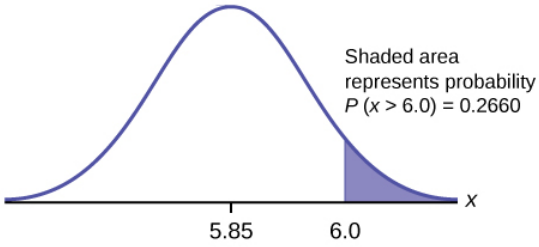


Figure 5.2.7.

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 40th percentile, and $k2$, the 60th 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 5.2.6

Using the information from Example, answer the following:

a. The middle 45% of mandarin oranges from this farm are between _____ and _____.
b. Find the 16th 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 30th percentile and $k2$, the 70th 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 5.2.7

How would you represent the area to the left of one in a probability statement?

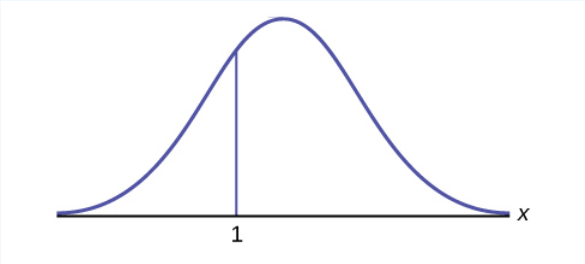


Figure 5.2.8.

Answer

$P(x < 1)$

? Exercise 5.2.8

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 5.2.9

How would you represent the area to the left of three in a probability statement?

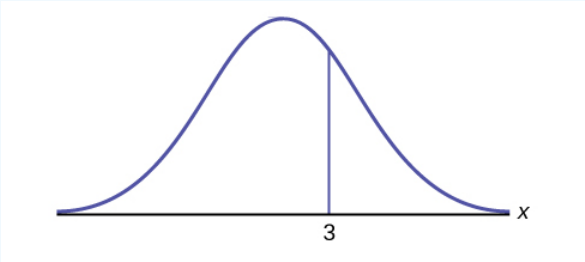


Figure 5.2.10.

? Exercise 5.2.10

What is the area to the right of three?

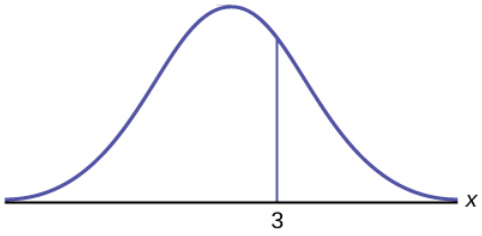


Figure 5.2.11.

Answer

$1 - P(x < 3)$ or $P(x > 3)$

Exercise 5.2.11

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 5.2.12

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 5.2.13

Find the probability that $x > 56$.

Exercise 5.2.14

Find the probability that $x < 30$.

Answer

0.0013

Exercise 5.2.15

Find the 80^{th} percentile.

Exercise 5.2.16

Find the 60^{th} percentile.

Answer

56.03

Exercise 5.2.17

$X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise 5.2.18

$X \sim N(-3, 4)$

Find the probability that x is between one and four.

Answer

0.1186

Exercise 5.2.19

$X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

Exercise 5.2.20

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.

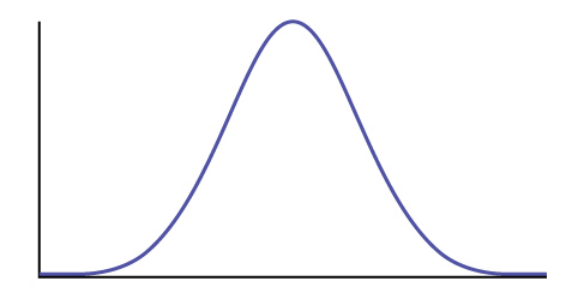


***Figure* 5.2.12.**

$P(0 < x <$ _____$)=$ _____ (Use zero for the minimum value of x.)

Answer

a. Check student's solution.
b. 3, 0.1979

Exercise 5.2.21

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.

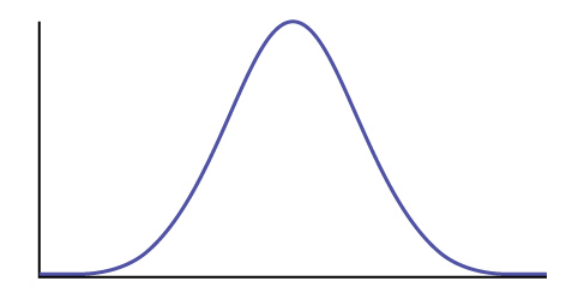


***Figure* 5.2.13.**

$P(\_\_\_\_\_ < x < \_\_\_\_\_) = \_\_\_\_\_$

Exercise 5.2.22

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%.

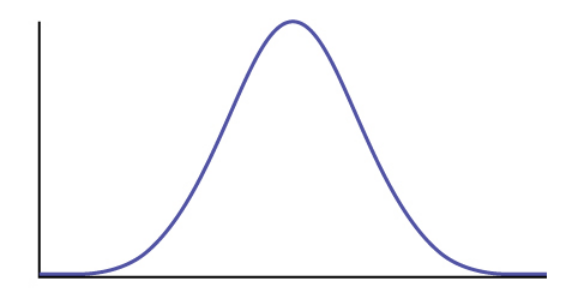


***Figure* 5.2.14.**

$P(x < k) = \_\_\_\_\_$ Therefore, $k = \_\_\_\_\_$

Answer

a. Check student's solution.
b. 0.70, 4.78 years

CHAPTER OVERVIEW

6: Sampling Distribution

6.1: The Sampling Distribution of Means

Learning Objectives

- To learn what the sampling distribution of $\overline{X}$ is when the sample size is large.
- To learn what the sampling distribution of $\overline{X}$ is when the population is normal.

In Example 6.1.1, we constructed the probability distribution of the sample mean for samples of size two drawn from the population of four rowers. The probability distribution is:

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Figure 6.1.1 shows a side-by-side comparison of a histogram for the original population and a histogram for this distribution. Whereas the distribution of the population is uniform, the sampling distribution of the mean has a shape approaching the shape of the familiar bell curve. This phenomenon of the sampling distribution of the mean taking on a bell shape even though the population distribution is not bell-shaped happens in general. Here is a somewhat more realistic example.

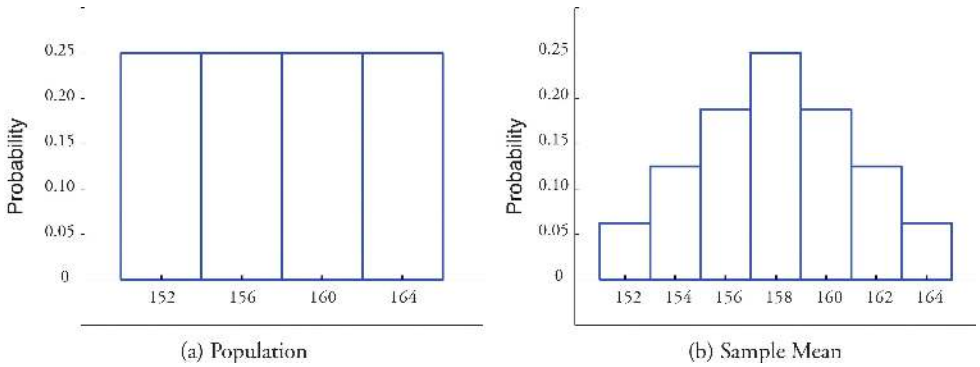


Figure 6.1.1: Distribution of a Population and a Sample Mean

Suppose we take samples of size 1, 5, 10, or 20 from a population that consists entirely of the numbers 0 and 1, half the population 0, half 1, so that the population mean is 0.5. The sampling distributions are:

$n = 1$:

$\bar{x}$	0	1
$P(\bar{x})$	0.5	0.5

$n = 5$:

$\bar{x}$	0	0.2	0.4	0.6	0.8	1
$P(\bar{x})$	0.03	0.16	0.31	0.31	0.16	0.03

$n = 10$:

$\bar{x}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P(\bar{x})$	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00

$n = 20$:

$\bar{x}$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$P(\bar{x})$	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.07	0.12	0.16	0.18

and

$\bar{x}$	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1
$P(\bar{x})$	0.16	0.12	0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00

Histograms illustrating these distributions are shown in Figure 6.1.2.

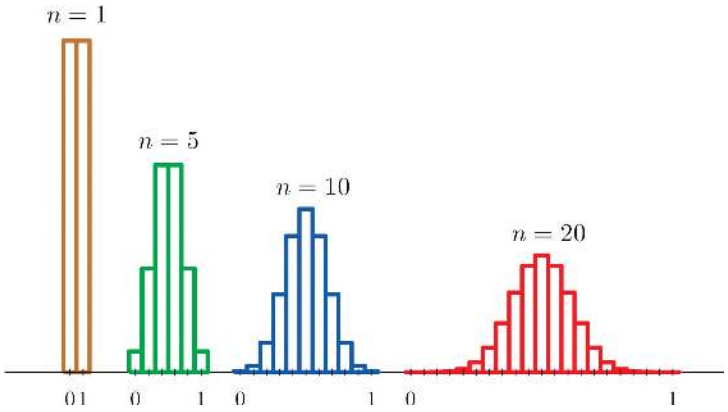


Figure 6.1.2: Distributions of the Sample Mean

As n increases the sampling distribution of $\overline{X}$ evolves in an interesting way: the probabilities on the lower and the upper ends shrink and the probabilities in the middle become larger in relation to them. If we were to continue to increase n then the shape of the sampling distribution would become smoother and more bell-shaped.

What we are seeing in these examples does not depend on the particular population distributions involved. In general, one may start with any distribution and the sampling distribution of the sample mean will increasingly resemble the bell-shaped normal curve as the sample size increases. This is the content of the Central Limit Theorem.

The Central Limit Theorem

For samples of size 30 or more, the sample mean is approximately normally distributed, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$, where n is the sample size. The larger the sample size, the better the approximation. The **Central Limit Theorem** is illustrated for several common population distributions in Figure 6.1.3.

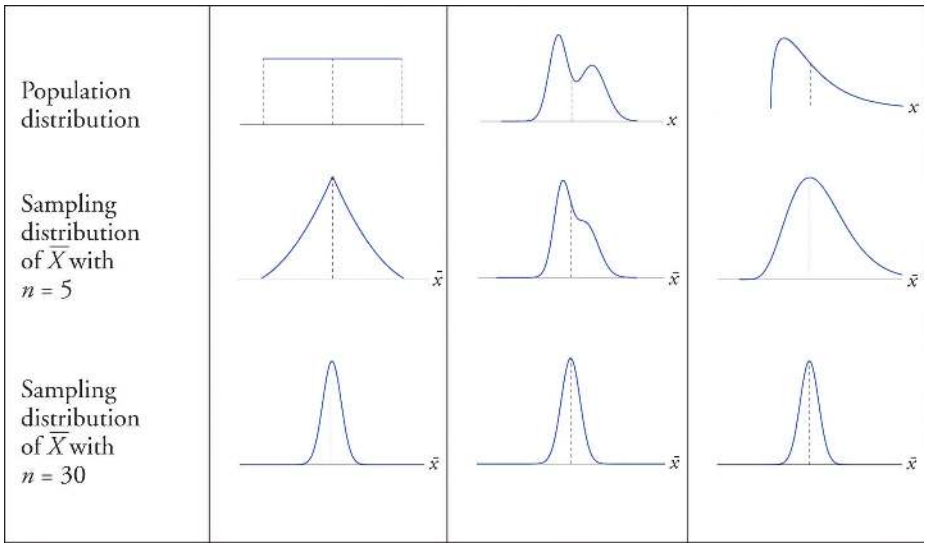


Figure 6.1.3: Distribution of Populations and Sample Means

The dashed vertical lines in the figures locate the population mean. Regardless of the distribution of the population, as the sample size is increased the shape of the sampling distribution of the sample mean becomes increasingly bell-shaped, centered on the population mean. Typically by the time the sample size is 30 the distribution of the sample mean is practically the same as a normal distribution.

The importance of the Central Limit Theorem is that it allows us to make probability statements about the sample mean, specifically in relation to its value in comparison to the population mean, as we will see in the examples. But to use the result properly we must first realize that there are two separate random variables (and therefore two probability distributions) at play:

1. X, the measurement of a single element selected at random from the population; the distribution of X is the distribution of the population, with mean the population mean μ and standard deviation the population standard deviation σ;
2. $\overline{X}$, the mean of the measurements in a sample of size n; the distribution of $\overline{X}$ is its sampling distribution, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

Example 6.1.1

Let $\overline{X}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40.

1. Find the mean and standard deviation of $\overline{X}$.
2. Find the probability that $\overline{X}$ assumes a value between 110 and 114.
3. Find the probability that $\overline{X}$ assumes a value greater than 113.

Solution

1. By the formulas in the previous section

$$\mu_{\overline{X}} = \mu = 112$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.65685$$

2. Since the sample size is at least 30, the Central Limit Theorem applies: $\overline{X}$ is approximately normally distributed. We compute probabilities using Figure 5.3.1 in the usual way, just being careful to use $\sigma_{\overline{X}}$ and not σ when we standardize:

$$\begin{aligned} P(110 < \overline{X} < 114) &= P\left(\frac{110 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{114 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{110 - 112}{5.65685} < Z < \frac{114 - 112}{5.65685}\right) \\ &= P(-0.35 < Z < 0.35) \\ &= 0.6368 - 0.3632 \\ &= 0.2736 \end{aligned}$$

3. Similarly

$$\begin{aligned} P(\overline{X} > 113) &= P\left(Z > \frac{113 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z > \frac{113 - 112}{5.65685}\right) \\ &= P(Z > 0.18) \\ &= 1 - P(Z < 0.18) \\ &= 1 - 0.5714 \\ &= 0.4286 \end{aligned}$$

Note that if in the above example we had been asked to compute the probability that the value of a single randomly selected element of the population exceeds 113, that is, to compute the number $P(X > 113)$, we would not have been able to do so, since we do not know the distribution of X, but only that its mean is 112 and its standard deviation is 40. By contrast we could compute

$P(\overline{X} > 113)$ even without complete knowledge of the distribution of X because the Central Limit Theorem guarantees that $\overline{X}$ is approximately normal.

Example 6.1.2

The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

Solution

The sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 2.61$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.5}{10} = 0.05$, so

$$\begin{aligned} P(2.51 < \overline{X} < 2.71) &= P\left(\frac{2.51 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{2.71 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{2.51 - 2.61}{0.05} < Z < \frac{2.71 - 2.61}{0.05}\right) \\ &= P(-2 < Z < 2) \\ &= P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 \\ &= 0.9544 \end{aligned}$$

Normally Distributed Populations

The Central Limit Theorem says that no matter what the distribution of the population is, as long as the sample is "large," meaning of size 30 or more, the sample mean is approximately normally distributed. If the population is normal to begin with then the sample mean also has a normal distribution, regardless of the sample size.

For samples of any size drawn from a normally distributed population, the sample mean is normally distributed, with mean $\mu_X = \mu$ and standard deviation $\sigma_X = \sigma/\sqrt{n}$, where n is the sample size.

The effect of increasing the sample size is shown in Figure $6.1.4$.

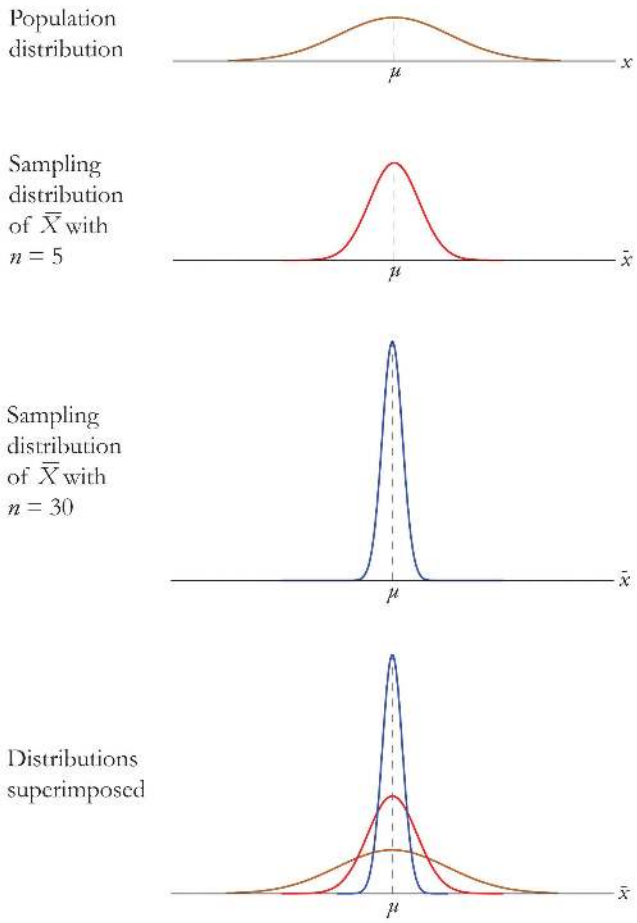


Figure 6.1.4: Distribution of Sample Means for a Normal Population

Example 6.1.3

A prototype automotive tire has a design life of $38,500$ miles with a standard deviation of $2,500$ miles. Five such tires are manufactured and tested. On the assumption that the actual population mean is $38,500$ miles and the actual population standard deviation is $2,500$ miles, find the probability that the sample mean will be less than $36,000$ miles. Assume that the distribution of lifetimes of such tires is normal.

Solution

For simplicity we use units of thousands of miles. Then the sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 38.5$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2.5}{\sqrt{5}} = 1.11803$. Since the population is normally distributed, so is $\overline{X}$, hence

$$\begin{aligned} P(\overline{X} < 36) &= P\left(Z < \frac{36 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z < \frac{36 - 38.5}{1.11803}\right) \\ &= P(Z < -2.24) \\ &= 0.0125 \end{aligned}$$

That is, if the tires perform as designed, there is only about a 1.25% chance that the average of a sample of this size would be so low.

Example 6.1.4

An automobile battery manufacturer claims that its midgrade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal.

a. On the assumption that the manufacturer's claims are true, find the probability that a randomly selected battery of this type will last less than 48 months.
b. On the same assumption, find the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Solution

a. Since the population is known to have a normal distribution

$$\begin{aligned} P(X<48) &= P\left(Z<\frac{48-\mu}{\sigma}\right) \\ &= P\left(Z<\frac{48-50}{6}\right) \\ &= P(Z<-0.33) \\ &= 0.3707 \end{aligned}$$

b. The sample mean has mean $\mu_{\overline{X}}=\mu=50$ and standard deviation $\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}=\frac{6}{\sqrt{36}}=1$. Thus

$$\begin{aligned} P(\overline{X}<48) &= P\left(Z<\frac{48-\mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z<\frac{48-50}{1}\right) \\ &= P(Z<-2) \\ &= 0.0228 \end{aligned}$$

Key Takeaway

- When the sample size is at least 30 the sample mean is normally distributed.
- When the population is normal the sample mean is normally distributed regardless of the sample size.

6.2: The Sampling Distribution for Proportions

Learning Objectives

- To recognize that the sample proportion $\hat{p}$ is a random variable.
- To understand the meaning of the formulas for the mean and standard deviation of the sample proportion.
- To learn what the sampling distribution of $\hat{p}$ is when the sample size is large.

Often sampling is done in order to estimate the proportion of a population that has a specific characteristic, such as the proportion of all items coming off an assembly line that are defective or the proportion of all people entering a retail store who make a purchase before leaving. The population proportion is denoted p and the sample proportion is denoted $\hat{p}$. Thus if in reality 43% of people entering a store make a purchase before leaving,

$$p = 0.43$$

if in a sample of 200 people entering the store, 78 make a purchase,

$$\hat{p} = \frac{78}{200} = 0.39.$$

The sample proportion is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. Viewed as a random variable it will be written $\hat{P}$. It has a mean $\mu_{\hat{P}}$ and a standard deviation $\sigma_{\hat{P}}$. Here are formulas for their values.

mean and standard deviation of the sample proportion

Suppose random samples of size n are drawn from a population in which the proportion with a characteristic of interest is p. The mean $\mu_{\hat{P}}$ and standard deviation $\sigma_{\hat{P}}$ of the sample proportion $\hat{P}$ satisfy

$$\mu_{\hat{P}} = p \tag{6.2.1}$$

and

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}} \tag{6.2.2}$$

where $q = 1 - p$.

The Central Limit Theorem has an analogue for the population proportion $\hat{p}$. To see how, imagine that every element of the population that has the characteristic of interest is labeled with a 1, and that every element that does not is labeled with a 0. This gives a numerical population consisting entirely of zeros and ones. Clearly the proportion of the population with the special characteristic is the proportion of the numerical population that are ones; in symbols,

$$p = \frac{\text{number of 1s}}{N} \tag{6.2.3}$$

But of course the sum of all the zeros and ones is simply the number of ones, so the mean μ of the numerical population is

$$\mu = \frac{\sum x}{N} = \frac{\text{number of 1s}}{N} \tag{6.2.4}$$

Thus the population proportion p is the same as the mean μ of the corresponding population of zeros and ones. In the same way the sample proportion $\hat{p}$ is the same as the sample mean $\bar{x}$. Thus the Central Limit Theorem applies to $\hat{p}$. However, the condition that the sample be large is a little more complicated than just being of size at least 30.

The Sampling Distribution of the Sample Proportion

For large samples, the sample proportion is approximately normally distributed, with mean $\mu_{\hat{P}} = p$ and standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$.

A sample is large if the interval $\left[p-3\sigma_{\hat{p}},\ p+3\sigma_{\hat{p}}\right]$ lies wholly within the interval $[0,1]$.

In actual practice p is not known, hence neither is $\sigma_{\hat{P}}$. In that case in order to check that the sample is sufficiently large we substitute the known quantity $\hat{p}$ for p. This means checking that the interval

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{6.2.5}$$

lies wholly within the interval $[0,1]$. This is illustrated in the examples.

Figure 6.2.1 shows that when $p=0.1$, a sample of size 15 is too small but a sample of size 100 is acceptable.

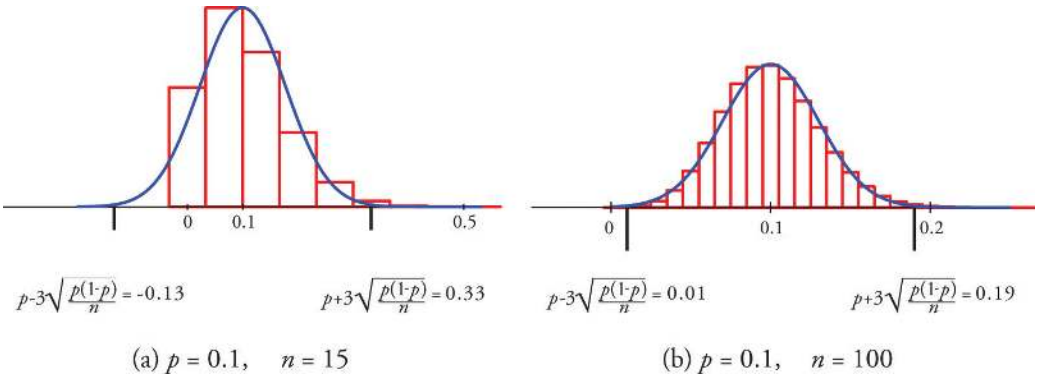


Figure 6.2.1 : *Distribution of Sample Proportions*

Figure 6.2.2 shows that when $p=0.5$ a sample of size 15 is acceptable.

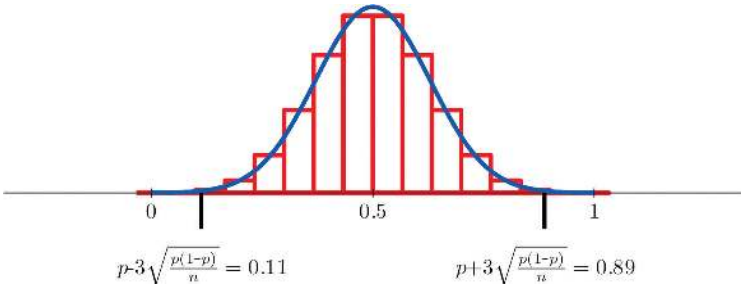


Figure 6.2.2 : Distribution of Sample Proportions for $p=0.5$ *and* $n=15$

Example 6.2.1

Suppose that in a population of voters in a certain region 38% are in favor of particular bond issue. Nine hundred randomly selected voters are asked if they favor the bond issue.

1. Verify that the sample proportion $\hat{p}$ computed from samples of size 900 meets the condition that its sampling distribution be approximately normal.
2. Find the probability that the sample proportion computed from a sample of size 900 will be within 5 percentage points of the true population proportion.

Solution:

1. The information given is that $p=0.38$, hence $q=1-p=0.62$. First we use the formulas to compute the mean and standard deviation of $\hat{p}$:

$$\mu_{\hat{p}}=p=0.38 \text{ and } \sigma_{\hat{P}}=\sqrt{\frac{pq}{n}}=\sqrt{\frac{(0.38)(0.62)}{900}}=0.01618$$

Then $3\sigma_{\hat{P}}=3(0.01618)=0.04854\approx 0.05$so

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]=[0.38-0.05,0.38+0.05]=[0.33,0.43]$$

which lies wholly within the interval $[0, 1]$, so it is safe to assume that $\hat{p}$ is approximately normally distributed.

2. To be within 5 percentage points of the true population proportion 0.38 means to be between $0.38 - 0.05 = 0.33$ and $0.38 + 0.05 = 0.43$. Thus

$$\begin{aligned} P(0.33 < \hat{P} < 0.43) &= P\left(\frac{0.33 - \mu_{\hat{P}}}{\sigma_{\hat{P}}} < Z < \frac{0.43 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\ &= P\left(\frac{0.33 - 0.38}{0.01618} < Z < \frac{0.43 - 0.38}{0.01618}\right) \\ &= P(-3.09 < Z < 3.09) \\ &= P(3.09) - P(-3.09) \\ &= 0.9990 - 0.0010 \\ &= 0.9980 \end{aligned}$$

Example 6.2.2

An online retailer claims that 90% of all orders are shipped within 12 hours of being received. A consumer group placed 121 orders of different sizes and at different times of day; 102 orders were shipped within 12 hours.

1. Compute the sample proportion of items shipped within 12 hours.
2. Confirm that the sample is large enough to assume that the sample proportion is normally distributed. Use $p = 0.90$, corresponding to the assumption that the retailer's claim is valid.
3. Assuming the retailer's claim is true, find the probability that a sample of size 121 would produce a sample proportion so low as was observed in this sample.
4. Based on the answer to part (c), draw a conclusion about the retailer's claim.

Solution:

1. The sample proportion is the number x of orders that are shipped within 12 hours divided by the number n of orders in the sample:

$$\hat{p} = \frac{x}{n} = \frac{102}{121} = 0.84$$

2. Since $p = 0.90$, $q = 1 - p = 0.10$, and $n = 121$,

$$\sigma_{\hat{P}} = \sqrt{\frac{(0.90)(0.10)}{121}} = 0.0\overline{27}$$

hence

$$\left[p - 3\sigma_{\hat{P}},\, p + 3\sigma_{\hat{P}}\right] = [0.90 - 0.08, 0.90 + 0.08] = [0.82, 0.98]$$

Because

$$[0.82, 0.98] \subset [0, 1]$$

it is appropriate to use the normal distribution to compute probabilities related to the sample proportion $\hat{P}$.

3. Using the value of $\hat{P}$ from part (a) and the computation in part (b),

$$\begin{aligned} P(\hat{P} \leq 0.84) &= P\left(Z \leq \frac{0.84 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\ &= P\left(Z \leq \frac{0.84 - 0.90}{0.0\overline{27}}\right) \\ &= P(Z \leq -2.20) \\ &= 0.0139 \end{aligned}$$

4. The computation shows that a random sample of size 121 has only about a 1.4% chance of producing a sample proportion as the one that was observed, $\hat{p} = 0.84$, when taken from a population in which the actual proportion is 0.90. This is so unlikely that it is reasonable to conclude that the actual value of p is less than the 90% claimed.

Key Takeaway

- The sample proportion is a random variable $\hat{P}$.
- There are formulas for the mean $\mu_{\hat{P}}$, and standard deviation $\sigma_{\hat{P}}$ of the sample proportion.
- When the sample size is large the sample proportion is normally distributed.

CHAPTER OVERVIEW

7: Confidence Intervals

7.1: Confidence Intervals Concepts

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 7.1.1 . 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

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \tag{7.1.1}$$

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

and

$$\bar{x} + 0.2 = 2 + 0.2 = 2.2 \tag{7.1.2}$$

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

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.

7.2: Confidence Interval for a Proportion

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03,0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The build a confidence interval for population proportion p, we use:

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7.2.1}$$

where

- $\hat{p} = \dfrac{x}{n}$, the estimated proportion of successes $\hat{p}$ is a point estimate for p, the true proportion.)
- $x =$ the number of successes
- $n =$ the size of the sample
- $z_{\frac{\alpha}{2}} \cdot \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ is called the **margin of error**

In the margin of error formula, the sample proportions $\hat{p}$ and 1-{\hat p} are estimates of the unknown population proportions p and 1-p. The estimated proportions $\hat{p}$ and $1-\hat{p}$ are used because p and $1-\hat{p}$ are not known. The sample proportions $\hat{p}$ and $1-\hat{p}$ are calculated from the data: $\hat{p}$ is the estimated proportion of successes, and $1-\hat{p}$ is the estimated proportion of failures.

Example 7.2.1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

- $n = 500$
- $x =$ the number of successes $= 421$

$$\hat{p} = \frac{x}{n} = \frac{421}{500} = 0.842$$

- $\hat{p} = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$1-\hat{p} = 1-0.842 = 0.158$$

Since the confidence level $CL = 0.95$, then

$$\alpha = 1-CL = 1-0.95 = 0.05$$

$$\text{So, } \frac{\alpha}{2} = 0.025.$$

Then

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$\text{margin of error} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 1.96 \cdot \sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$\hat{p} - \text{margin of error} = 0.842 - 0.032 = 0.81$$

$$\hat{p} + \text{margin of error} = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(\hat{p} - \text{margin of error}, \hat{p} + \text{margin of error}) = (0.810, 0.874)$.

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 421.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (0.81003, 0.87397).

Example 7.2.2

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

- $x = 300$ and
- $n = 500$

$$\hat{p} = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$1 - \hat{p} = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then

$$\alpha = 1 - CL = 1 - 0.90 = 0.10$$

$$\text{So, } \frac{\alpha}{2} = 0.05.$$

$$z\frac{\alpha}{2} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$\text{margin of error} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 1.645 \cdot \sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$\hat{p} - \text{ margin of error } = 0.60 - 0.036 = 0.564$$

$$\hat{p} + \text{margin of error} = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(\hat{p} - \text{margin of error}, \hat{p} + \text{margin of error}) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 300.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press `ENTER` .

The confidence interval is (0.564, 0.636).

Example 7.2.3

To estimate the proportion of students at a large college who are female, a random sample of 120 students is selected. There are 69 female students in the sample. Construct a 90% confidence interval for the proportion of all students at the college who are female.

Solution A

The proportion of students in the sample who are female is

$$\hat{p} = 69/120 = 0.575$$

Confidence level 90% means that $\alpha = 1 - 0.90 = 0.10$ so $\alpha/2 = 0.05$. From the last line of Figure 7.1.6 we obtain $z_{0.05} = 1.645$.

Thus

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.575 \pm 1.645\sqrt{\frac{(0.575)(0.425)}{120}} = 0.575 \pm 0.074$$

One may be 90% confident that the true proportion of all students at the college who are female is contained in the interval $(0.575 - 0.074, 0.575 + 0.074) = (0.501, 0.649)$.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to x and enter 69.

Arrow down to n and enter 120.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.501,0.649).

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: Confidence Interval for a Mean

You have determined to turn your dusty backyard into a flower paradise. Before you start any work, you want to know if this idea is worth it. You visit the local park and have noticed that certain flowers do not last for very long after blooming. Seeking scientific proof to make your decision, you have collected data on the number of days you notice the flowers are in full bloom. What should you do with this data?

You can quickly calculate the mean number of days of full bloom. But you have decided that this single number does not determine the range of *possible* days that a full bloom occurs. To do so, you will need to build a confidence interval for mean.

> ***The following is the confidence interval for a population mean:***
>
> $$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \qquad (7.3.1)$$
>
> *where the* ***lower bound*** $= \bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ *and the* ***upper bound*** $= \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ *, and a margin of error* $= \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$.
>
> *Requirement:* X *is normally distributed or* $n \geq 30$

Notice, the new notation of $t_{\alpha/2}$. This critical value refers to the Student's t-distribution which is described in more detail below.

Student's t-distribution with $n-1$ degrees of freedom. Student's t-distribution is very much like the standard normal distribution in that it is centered at 0 and has the same qualitative bell shape, but it has heavier tails than the standard normal distribution does, as indicated by Figure 7.3.1 , in which the curve (in brown) that meets the dashed vertical line at the lowest point is the t-distribution with two degrees of freedom, the next curve (in blue) is the t-distribution with five degrees of freedom, and the thin curve (in red) is the standard normal distribution. As also indicated by the figure, as the sample size n increases, Student's t-distribution ever more closely resembles the standard normal distribution. Although there is a different t-distribution for every value of n, once the sample size is 30 or more it is typically acceptable to use the standard normal distribution instead, as we will always do in this text.

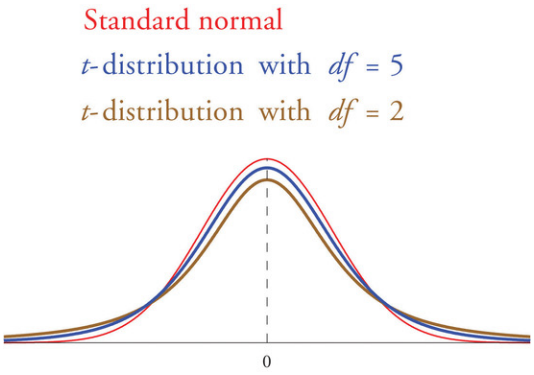


Figure 7.3.1 : Student's t-*DistributionJ*

Student's t-distribution table relies on the degree of freedom ($df = n-1$) and the area in the right tail. In the table, the first column gives to the degree of freedom. The first row indicates the subscript value for the Student's t critical values. This subscript represents the area in the right tail.

Example 7.3.1

Use Figure 7.3.1 below to find the number $t_{\alpha/2}$ needed in construction of a confidence interval:

1. when the level of confidence is 90% with $n = 15$

2. when the level of confidence is 99% with $n = 23$

Area in the Right Tail of a t-Distribution

df	0.1	0.05	0.025	0.01	0.005	0.0025	0.0005
1	3.078	6.314	12.706	31.821	63.657	127.321	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.646
31	1.309	1.696	2.040	2.453	2.744	3.022	3.633
32	1.309	1.694	2.037	2.449	2.738	3.015	3.622
33	1.308	1.692	2.035	2.445	2.733	3.008	3.611
34	1.307	1.691	2.032	2.441	2.728	3.002	3.601
35	1.306	1.690	2.030	2.438	2.724	2.996	3.591
36	1.306	1.688	2.028	2.434	2.719	2.990	3.582
37	1.305	1.687	2.026	2.431	2.715	2.985	3.574
38	1.304	1.686	2.024	2.429	2.712	2.980	3.566
39	1.304	1.685	2.023	2.426	2.708	2.976	3.558
40	1.303	1.684	2.021	2.423	2.704	2.971	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.496
60	1.296	1.671	2.000	2.390	2.660	2.915	3.460
70	1.294	1.667	1.994	2.381	2.648	2.899	3.435
80	1.292	1.664	1.990	2.374	2.639	2.887	3.416
90	1.291	1.662	1.987	2.368	2.632	2.878	3.402
100	1.290	1.660	1.984	2.364	2.626	2.871	3.390
z	1.282	1.645	1.960	2.326	2.576	2.807	3.291

Figure 7.3.1 : Critical Values of t

Solution:

1. Figure 7.3.1 gives the critical values which is the cut-off value for which area in the right tail. The area in the right tail is defined by $\alpha/2$. We will start with finding α, which is $1 -$ confidence level $= 1 - 0.90 = 0.10$. Now to get $\alpha/2$, we divide 0.10 by 2. So, $\alpha/2 = 0.05$, which is the column we In other words, we need to find $t_{0.05}$. The subscript 0.05 is the column which we will look under. And, we will look at the degree of freedom$= n - 1$ row. In this example our degree of freedom is 14 (because $15 - 1$). Thus, $t_{0.05} = 1.761$.
2. In Figure 7.3.1 $t_{0.005}$ is the number in the 22nd row and in the column headed 0.005, namely 2.819.

Example 7.3.2

A sample of size 15 drawn from a normally distributed population has sample mean 35 and sample standard deviation 14. Construct a 95% confidence interval for the population mean, and interpret its meaning.

Solution:

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is Equation ???.

Confidence level 95% means that

$$\alpha = 1 - 0.95 = 0.05 \tag{7.3.2}$$

so $\alpha/2 = 0.025$. Since the sample size is $n = 15$, there are $n - 1 = 14$ degrees of freedom. By Figure 7.1.6 $t_{0.025} = 2.145$. Thus

$$\overline{x} = \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.3.3}$$

$$= 35 \pm 2.145\left(\frac{14}{\sqrt{15}}\right) \tag{7.3.4}$$

$$= 35 \pm 7.8 \tag{7.3.5}$$

One may be 95% confident that the true value of μ is contained in the interval

$$(35 - 7.8, 35 + 7.8) = (27.2, 42.8). \tag{7.3.6}$$

Example 7.3.3

A random sample of 12 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university. Assume that the numerical population of GPAs from which the sample is taken has a normal distribution.

Solution:

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is Equation ???

Confidence level 90% means that

$$\alpha = 1 - 0.90 = 0.10 \tag{7.3.7}$$

so $\alpha/2 = 0.05$. Since the sample size is $n = 12$, there are $n - 1 = 11$ degrees of freedom. By Figure 7.1.6 $t_{0.05} = 1.796$. Thus

$$\overline{x} = \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.3.8}$$

$$= 2.71 \pm 1.796\left(\frac{0.51}{\sqrt{12}}\right) \tag{7.3.9}$$

$$= 2.71 \pm 0.26 \tag{7.3.10}$$

AutoAttribution

- **7.1: Large Sample Estimation of a Population Mean** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

7.4: Confidence Interval for Standard Deviation

When you open a bag of dog kibble, there is an expectation of how much is in the bag. Kibble filling machines do not necessarily fill each bag to the exact weight on the label, there may be a bit more or even a little bit less. In instances like this, we would be interested in checking if a filling machine is consistent in its filling capabilities. Examples like this measure consistency (i.e., how spread out are the data values); thus, in this section, we will discuss how to construct confidence intervals for a population standard deviation.

Recall, to find the confidence interval for a population proportion, you used the normal distribution and to find the confidence interval for population mean, you used the Student's t-distribution. In this section, we will need to introduce a new distribution called the χ^2 distribution, which we will use a table.

First, the χ^2 distribution curve is not symmetrical and depends on the degree of freedom ($df = n - 1$). The smaller the degree of free, the curve looks more skewed right. The bigger the degree of freedom, the curve looks more symmetric.

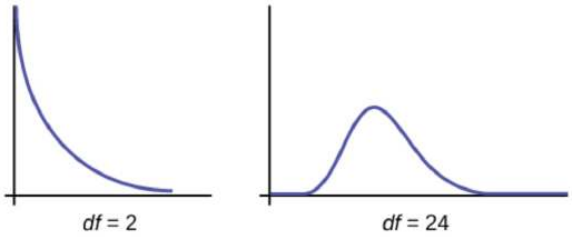


Figure 7.4.1

Second, χ^2-distribution table relies on the degree of freedom ($df = n - 1$) and the area in the right tail. In the table, the first column gives to the degree of freedom. The first row indicates the subscript value for the χ^2 critical values. This subscript represents the area in the right tail.

Figure 7.4.2

Example 7.4.1

Given $n = 12$, find $\chi^2_{0.01}$.

Answer

Notice that α (the subscript) is 0.01. This represents the area in the right tail. Draw a χ^2 curve, shade the right tail as 0.01. Since $n = 12$, that means the degree of freedom $df = 11$.

Figure 7.4.3

Under the column labeled as 0.01 and across the row of degree of freedom $df = 12 - 1 = 11$

Area in the Right Tail of a X^2-Distribution

DF	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757

Figure 7.4.4

Thus, we can label that the tail begins at $\chi^2_{0.01} = 24.725$.

Figure 7.4.5

The following is the confidence interval for a population standard deviation:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} < \sigma^2 < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} \tag{7.4.1}$$

where the ***lower bound*** $\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$ *and the* ***upper bound*** $= \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}$

Requirement: X *is normally distributed.*

Notice that the formula does not look like the previous confidence interval format of "point estimate $\pm$ margin of error." Also, notice that the critical values in the denominators look similar but subscripts are not the same. Let's talk about critical values.

Example 7.4.2

Example 7.4.2

Find the critical values given a 95% *level of confidence and the sample size* $n = 5$.

Answer

Given $n = 5$, *then the degree of freedom* $df = 5 - 1 = 4$. *Thus, we will look at the 4th row of the table.*

Notice, no α *is explicitly given here. What do we do? To find* α, *subtract the level of confidence from* 1. *So,* $\alpha = 1 - 0.95 = 0.05$. *Which means,* $\alpha/2 = 0.025$ *and* $1 - \alpha/2 = 0.975$.

To find $\chi^2_{\alpha/2}$, we will look for $\chi^2_{0.025}$. To find $1-\chi^2_{\alpha/2}$, we will look for $\chi^2_{0.975}$. We look for the subscript value in the top row of the table.

The critical values are $\chi^2_{0.025}=11.143$ and $\chi^2_{0.975}=0.484$

Figure 7.4.6

Figure 7.4.7

In the next example, we will find a confidence interval for population standard deviation.

Example 7.4.3

You buy in bulk 12 bags of dog kibble and weigh each bag. The following data is the weight in pounds.

(a) Find the confidence interval for the standard deviation at a 90% level of confidence.

(b) Give an interpretation of your confidence interval.

9.63	9.60	8.98	9.86
9.52	9.53	9.78	9.56
9.12	9.25	9.52	9.34

Answers:

(a) First find the critical values. Since $\alpha=0.10$, then $\alpha/2=0.05$ and $1-\alpha/2=0.95$. Looking at the columns labeled 0.05 and 0.95 with the df row of 11, we see that $\chi^2_{0.05}=19.675$ and $\chi^2_{0.95}=4.575$

Figure 7.4.8

Now, find the sample standard deviation. The sample standard deviation $s = 0.2585$.

Lastly, putting everything together:

$$\text{lower bound} = \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} = \sqrt{\frac{(12-1)0.2585^2}{19.675}} = 0.1933$$

$$\text{upper bound} = \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} = \sqrt{\frac{(12-1)0.2585^2}{4.575}} = 0.4008$$

The confidence interval is $0.1933 < \sigma < 0.4008$ pounds. (Notice in this example, we are not moving the decimal to the left twice because these values do not represent percentages, but pounds.)

(b) We are 90% confident that the true standard deviation of potato chip bag weight is between 0.1933 and 0.4008 ounces.

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

CHAPTER OVERVIEW

8: Hypothesis Testing with One Sample

8.1.1: Introduction to Hypothesis Testing Part 1

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 8.1.1.1: 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 8.1.1.1

- 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 8.1.1.1

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 8.1.1.2

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 8.1.1.2

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 : μ_66
- H_a : μ_66

Answer

- $H_0 : \mu = 66$
- $H_a : \mu \neq 66$

Example 8.1.1.3

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 8.1.1.3

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 ($=, \neq, \geq, <, \leq, >$) for the null and alternative hypotheses.

a. H_0 : μ_45
b. H_a : μ_45

Answer

a. $H_0 : \mu \geq 45$
b. $H_a : \mu < 45$

Example 8.1.1.4

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 8.1.1.4

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).

8.1.2: Introduction to Hypothesis Testing Part 2

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 8.1.2.1: 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 8.1.2.1

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 8.1.2.2

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 8.1.2.2

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 8.1.2.3

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 8.1.2.3

"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 800μg $H_0 : \mu_0 \leq 800\mu$g.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800μg 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 800μg). 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 8.1.2.4

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 8.1.2.4

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.

8.2: Hypothesis Testing of Single Proportion

Learning Objectives

- To learn how to apply the five-step critical value test procedure for test of hypotheses concerning a population proportion.
- To learn how to apply the five-step p-value test procedure for test of hypotheses concerning a population proportion.

Both the critical value approach and the p-value approach can be applied to test hypotheses about a population proportion p. The null hypothesis will have the form $H_0 : p = p_0$ for some specific number p_0 between 0 and 1. The alternative hypothesis will be one of the three inequalities

- $p < p_0$,
- $p > p_0$, or
- $p \neq p_0$

for the same number p_0 that appears in the null hypothesis.

The information in Section 6.3 gives the following formula for the test statistic and its distribution. In the formula p_0 is the numerical value of p that appears in the two hypotheses, $q_0 = 1 - p_0, \hat{p}$ is the sample proportion, and n is the sample size. Remember that the condition that the sample be large is not that n be at least 30 but that the interval

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lie wholly within the interval $[0, 1]$.

Standardized Test Statistic for Large Sample Hypothesis Tests Concerning a Single Population Proportion

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_o}{n}}} \tag{8.2.1}$$

The test statistic has the standard normal distribution.

The distribution of the standardized test statistic and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.2.1.

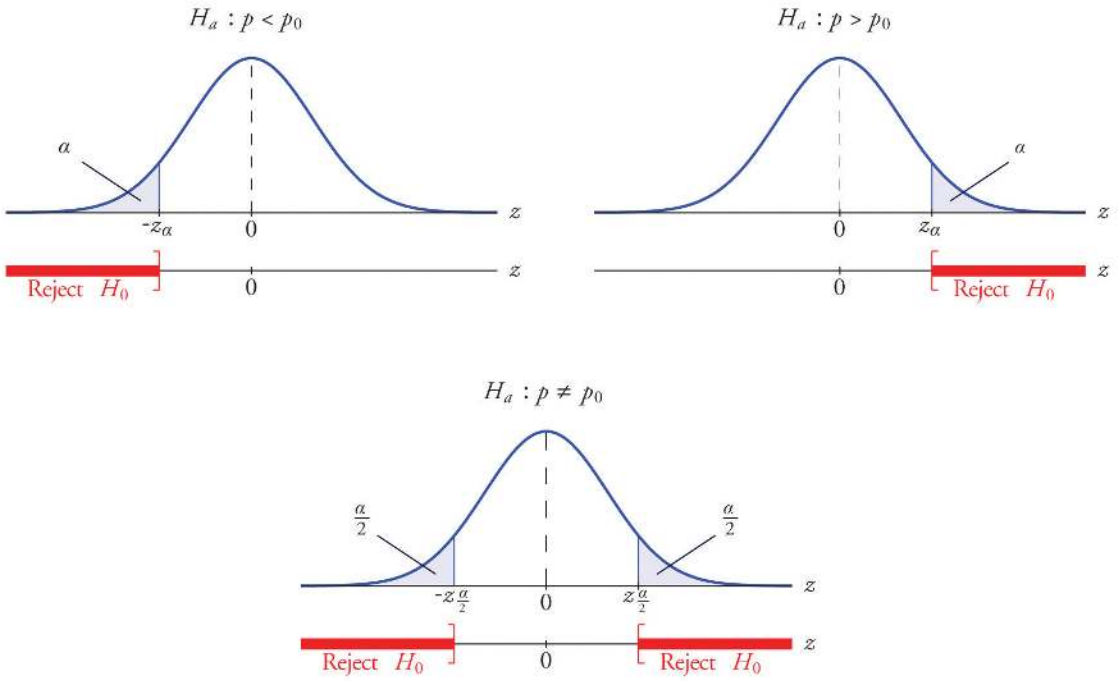


Figure 8.2.1: Distribution of the Standardized Test Statistic and the Rejection Region

Example 8.2.1

A soft drink maker claims that a majority of adults prefer its leading beverage over that of its main competitor's. To test this claim 500 randomly selected people were given the two beverages in random order to taste. Among them, 270 preferred the soft drink maker's brand, 211 preferred the competitor's brand, and 19 could not make up their minds. Determine whether there is sufficient evidence, at the 5% level of significance, to support the soft drink maker's claim against the default that the population is evenly split in its preference.

Solution

We will use the critical value approach to perform the test. The same test will be performed using the p-value approach in Example 8.2.3.

We must check that the sample is sufficiently large to validly perform the test. Since $\hat{p} = 270/500 = 0.54$,

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{500}} \approx 0.02$$

hence

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.2.2}$$

$$= [0.54 - (3)(0.02), 0.54 + (3)(0.02)] \tag{8.2.3}$$

$$= [0.48, 0.60] \subset [0, 1] \tag{8.2.4}$$

so the sample is sufficiently large.

- **Step 1.** The relevant test is

$$H_0 : p = 0.50$$

$$vs.$$

$$H_a : p > 0.50 \,@\, \alpha = 0.05$$

where p denotes the proportion of all adults who prefer the company's beverage over that of its competitor's beverage.

- **Step 2.** The test statistic (Equation 8.2.1) is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

and has the standard normal distribution.

- **Step 3.** The value of the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{8.2.5}$$

$$= \frac{0.54 - 0.50}{\sqrt{\frac{(0.50)(0.50)}{500}}} \tag{8.2.6}$$

$$= 1.789 \tag{8.2.7}$$

- **Step 4.** Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $z_\alpha = z_{0.05}$. Reading from the last line in Figure 7.1.6 its value is 1.645. The rejection region is $[1.645, \infty)$
- **Step 5.** As shown in Figure 8.2.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that a majority of adults prefer the company's beverage to that of their competitor's.

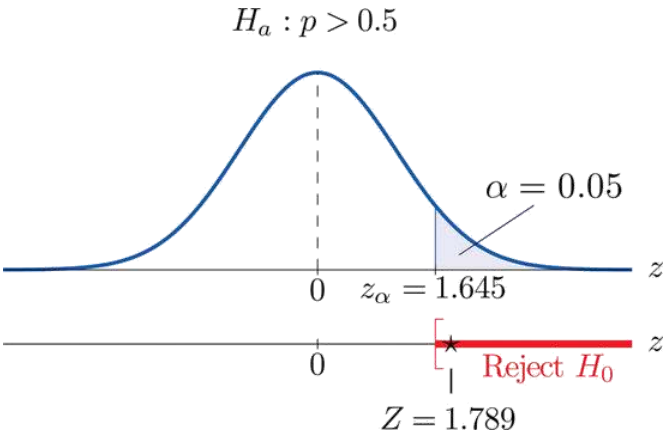


Figure 8.2.2: Rejection Region and Test Statistic for Example 8.2.1

Example 8.2.2

Globally the long-term proportion of newborns who are male is 51.46%. A researcher believes that the proportion of boys at birth changes under severe economic conditions. To test this belief randomly selected birth records of $5,000$ babies born during a period of economic recession were examined. It was found in the sample that 52.55% of the newborns were boys. Determine whether there is sufficient evidence, at the 10% level of significance, to support the researcher's belief.

Solution

We will use the critical value approach to perform the test. The same test will be performed using the p-value approach in Example 8.2.1.

The sample is sufficiently large to validly perform the test since

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.5255)(0.4745)}{5000}} \approx 0.01$$

hence

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.2.8}$$

$$= [0.5255 - 0.03, 0.5255 + 0.03] \tag{8.2.9}$$

$$= [0.4955, 0.5555] \subset [0, 1] \tag{8.2.10}$$

- **Step 1**. Let p be the true proportion of boys among all newborns during the recession period. The burden of proof is to show that severe economic conditions change it from the historic long-term value of 0.5146rather than to show that it stays the same, so the hypothesis test is

$$H_0 : p = 0.5146$$

$$vs.$$

$$H_a : p \neq 0.5146 \,@\, \alpha = 0.10$$

- **Step 2.** The test statistic (Equation 8.2.1) is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

and has the standard normal distribution.

- **Step 3.** The value of the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{8.2.11}$$

$$= \frac{0.5255 - 0.5146}{\sqrt{\frac{(0.5146)(0.4854)}{5000}}} \tag{8.2.12}$$

$$= 1.542 \tag{8.2.13}$$

- **Step 4.** Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are a pair of critical values, $\pm z_{\alpha/2} = \pm z_{0.05} = \pm 1.645$. The rejection region is $(-\infty, -1.645] \cup [1.645, \infty)$.
- **Step 5.** As shown in Figure $8.2.3$ the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

  The data do not provide sufficient evidence, at the 10% level of significance, to conclude that the proportion of newborns who are male differs from the historic proportion in times of economic recession.

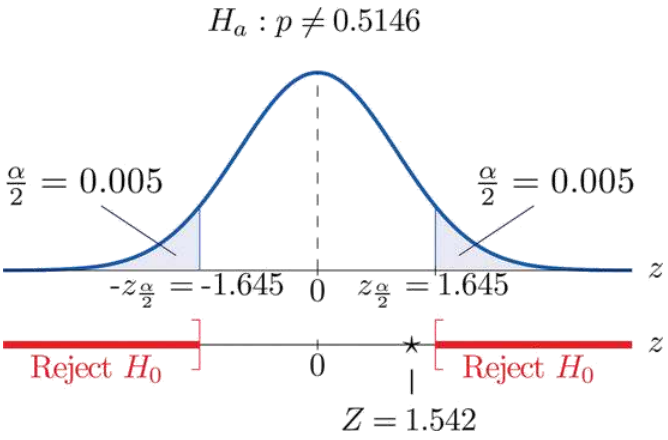


Figure 8.2.3: Rejection Region and Test Statistic for Example 8.2.2.

Example 8.2.3

Perform the test of Example 8.2.1 using the p-value approach.

Solution

We already know that the sample size is sufficiently large to validly perform the test.

- **Steps 1–3** of the five-step procedure described in Section 8.3 have already been done in Example 8.2.1 so we will not repeat them here, but only say that we know that the test is right-tailed and that value of the test statistic is $Z = 1.789$.
- **Step 4.** Since the test is right-tailed the p-value is the area under the standard normal curve cut off by the observed test statistic, $Z = 1.789$, as illustrated in Figure 8.2.4. By Figure 7.1.5 that area and therefore the p-value is $1 - 0.9633 = 0.0367$.
- **Step 5.** Since the p-value is less than $\alpha = 0.05$ the decision is to reject H_0.

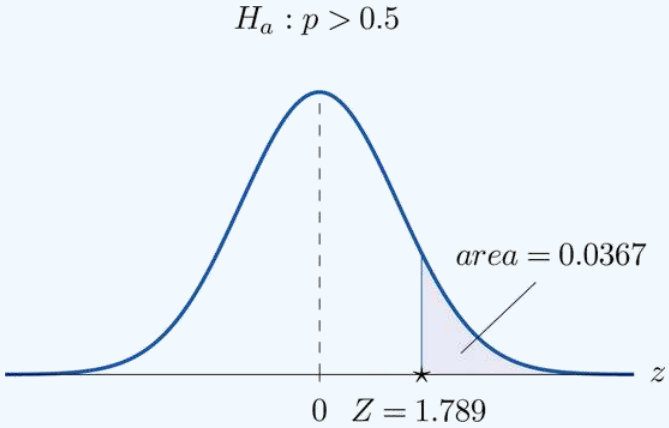


Figure 8.2.4: P-Value for Example 8.2.3

Example 8.2.4

Perform the test of Example 8.2.2 using the p-value approach.

Solution

We already know that the sample size is sufficiently large to validly perform the test.

- **Steps 1–3** of the five-step procedure described in Section 8.3 have already been done in Example 8.2.2. They tell us that the test is two-tailed and that value of the test statistic is $Z = 1.542$.
- **Step 4.** Since the test is two-tailed the p-value is the double of the area under the standard normal curve cut off by the observed test statistic, $Z = 1.542$. By Figure 7.1.5 that area is $1 - 0.9382 = 0.0618$, as illustrated in Figure 8.2.5, hence the p-value is $2 \times 0.0618 = 0.1236$.
- **Step 5.** Since the p-value is greater than $\alpha = 0.10$ the decision is not to reject H_0.

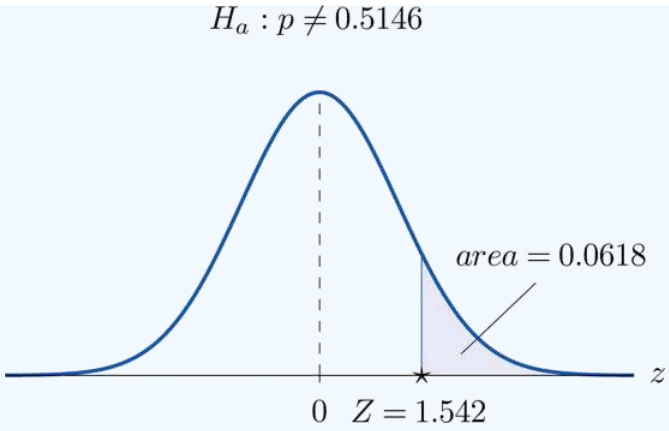


Figure 8.2.5: P-Value for Example 8.2.4

Key Takeaway

- There is one formula for the test statistic in testing hypotheses about a population proportion. The test statistic follows the standard normal distribution.
- Either five-step procedure, critical value or p-value approach, can be used.

8.3: Hypothesis Testing of Single Mean

Learning Objectives

- To learn how to apply the five-step test procedure for test of hypotheses concerning a population mean when the sample size is small.

In the previous section hypotheses testing for population means was described in the case of large samples. The statistical validity of the tests was insured by the Central Limit Theorem, with essentially no assumptions on the distribution of the population. When sample sizes are small, as is often the case in practice, the Central Limit Theorem does not apply. One must then impose stricter assumptions on the population to give statistical validity to the test procedure. One common assumption is that the population from which the sample is taken has a normal probability distribution to begin with. Under such circumstances, if the population standard deviation is known, then the test statistic

$$\frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}}$$

still has the standard normal distribution, as in the previous two sections. If σ is unknown and is approximated by the sample standard deviation s, then the resulting test statistic

$$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$$

follows Student's t-distribution with $n-1$ degrees of freedom.

Standardized Test Statistics for Small Sample Hypothesis Tests Concerning a Single Population Mean

If σ is known:

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

If σ is unknown:

$$T=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

- The first test statistic (σ known) has the standard normal distribution.
- The second test statistic (σ unknown) has Student's t-distribution with $n-1$ degrees of freedom.
- The population must be normally distributed.

The distribution of the second standardized test statistic (the one containing s) and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.3.1. This is just like Figure 8.2.1 except that now the critical values are from the t-distribution. Figure 8.2.1 still applies to the first standardized test statistic (the one containing (σ) since it follows the standard normal distribution.

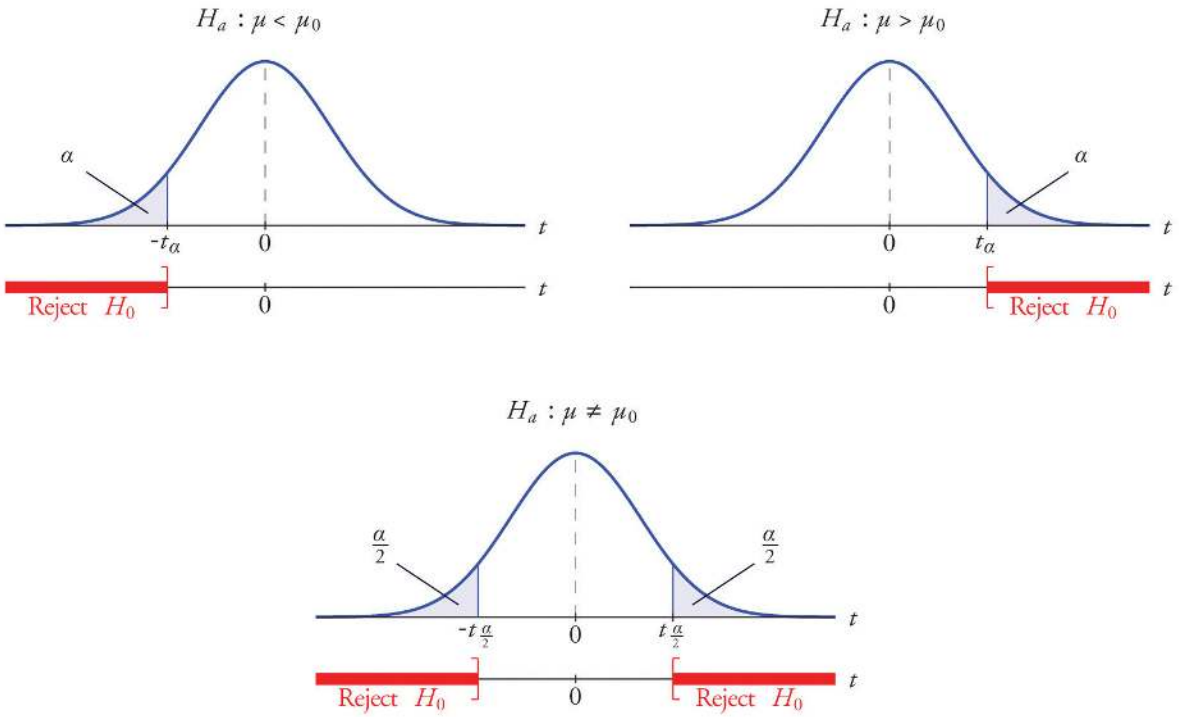


Figure 8.3.1: Distribution of the Standardized Test Statistic and the Rejection Region

The p-value of a test of hypotheses for which the test statistic has Student's t-distribution can be computed using statistical software, but it is impractical to do so using tables, since that would require 30 tables analogous to Figure 7.1.5, one for each degree of freedom from 1 to 30. Figure 7.1.6 can be used to approximate the p-value of such a test, and this is typically adequate for making a decision using the p-value approach to hypothesis testing, although not always. For this reason the tests in the two examples in this section will be made following the critical value approach to hypothesis testing summarized at the end of Section 8.1, but after each one we will show how the p-value approach could have been used.

Example 8.3.1

The price of a popular tennis racket at a national chain store is $\$179$. Portia bought five of the same racket at an online auction site for the following prices:

$$155 \quad 179 \quad 175 \quad 175 \quad 161$$

Assuming that the auction prices of rackets are normally distributed, determine whether there is sufficient evidence in the sample, at the 5% level of significance, to conclude that the average price of the racket is less than $\$179$ if purchased at an online auction.

Solution

- **Step 1**. The assertion for which evidence must be provided is that the average online price μ is less than the average price in retail stores, so the hypothesis test is

$$H_0 : \mu = 179$$
$$\text{vs}$$
$$H_a : \mu < 179 \ @ \ \alpha = 0.05$$

- **Step 2**. The sample is small and the population standard deviation is unknown. Thus the test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the Student t-distribution with $n - 1 = 5 - 1 = 4$ degrees of freedom.

- **Step 3**. From the data we compute $\bar{x} = 169$ and $s = 10.39$. Inserting these values into the formula for the test statistic gives

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{169 - 179}{10.39/\sqrt{5}} = -2.152$$

- **Step 4**. Since the symbol in H_a is "$<$" this is a left-tailed test, so there is a single critical value, $-t_\alpha = -t_{0.05}[df = 4]$. Reading from the row labeled $df = 4$ in Figure 7.1.6 its value is -2.132. The rejection region is $(-\infty, -2.132]$.
- **Step 5**. As shown in Figure $8.3.2$ the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the average price of such rackets purchased at online auctions is less than $\$179$.

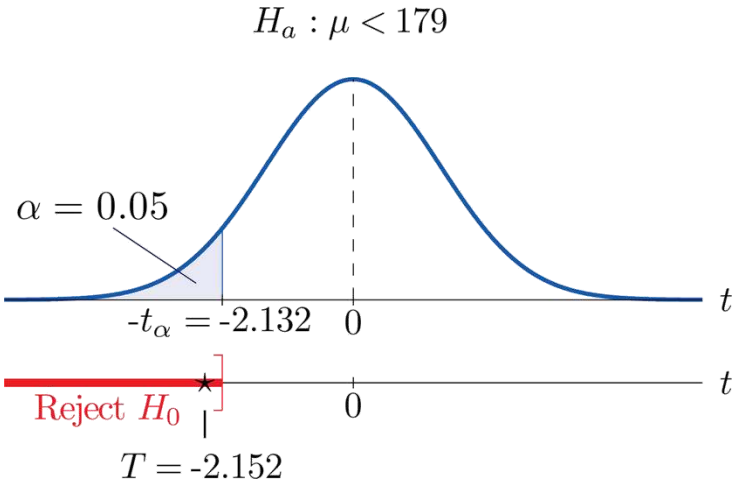


Figure 8.3.2: *Rejection Region and Test Statistic for "Example* 8.3.1*"*

To perform the test in Example $8.3.1$ using the p-value approach, look in the row in Figure 7.1.6 with the heading $df = 4$ and search for the two t-values that bracket the unsigned value 2.152 of the test statistic. They are 2.132 and 2.776, in the columns with headings $t_{0.050}$ and $t_{0.025}$. They cut off right tails of area 0.050 and 0.025, so because 2.152 is between them it must cut off a tail of area between 0.050 and 0.025. By symmetry -2.152 cuts off a left tail of area between 0.050 and 0.025, hence the p-value corresponding to $t = -2.152$ is between 0.025 and 0.05. Although its precise value is unknown, it must be less than $\alpha = 0.05$, so the decision is to reject H_0.

Example 8.3.2

A small component in an electronic device has two small holes where another tiny part is fitted. In the manufacturing process the average distance between the two holes must be tightly controlled at 0.02 mm, else many units would be defective and wasted. Many times throughout the day quality control engineers take a small sample of the components from the production line, measure the distance between the two holes, and make adjustments if needed. Suppose at one time four units are taken and the distances are measured as

$$0.021 \quad 0.019 \quad 0.023 \quad 0.020$$

Determine, at the 1% level of significance, if there is sufficient evidence in the sample to conclude that an adjustment is needed. Assume the distances of interest are normally distributed.

Solution

- **Step 1**. The assumption is that the process is under control unless there is strong evidence to the contrary. Since a deviation of the average distance to either side is undesirable, the relevant test is

$$H_0 : \mu = 0.02$$
$$\text{vs}$$
$$H_a : \mu \neq 0.02 \ @ \ \alpha = 0.01$$

  where μ denotes the mean distance between the holes.
- **Step 2**. The sample is small and the population standard deviation is unknown. Thus the test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the Student t-distribution with $n - 1 = 4 - 1 = 3$ degrees of freedom.

- **Step 3**. From the data we compute $\bar{x} = 0.02075$ and $s = 0.00171$. Inserting these values into the formula for the test statistic gives

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.02075 - 0.02}{0.00171\sqrt{4}} = 0.877$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm t_{\alpha/2} = -t_{0.005}[df = 3]$. Reading from the row in Figure 7.1.6 labeled $df = 3$ their values are ± 5.841. The rejection region is $(-\infty, -5.841] \cup [5.841, \infty)$.
- **Step 5**. As shown in Figure 8.3.3 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean distance between the holes in the component differs from 0.02 mm.

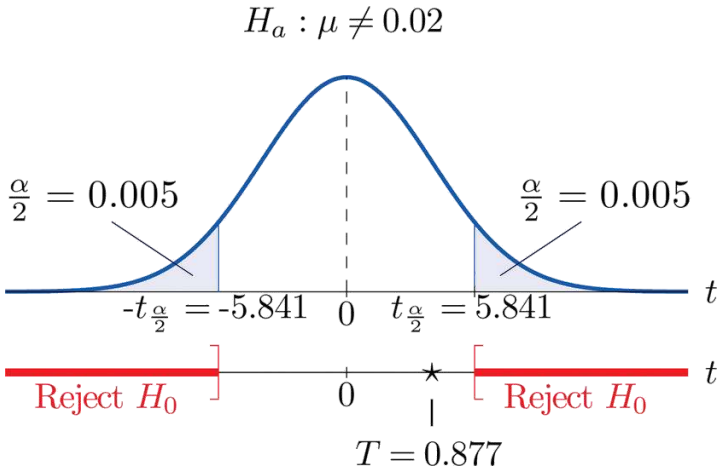


Figure 8.3.3: Rejection Region and Test Statistic for "Example 8.3.2"

To perform the test in "Example 8.3.2" using the p-value approach, look in the row in Figure 7.1.6 with the heading $df = 3$ and search for the two t-values that bracket the value 0.877 of the test statistic. Actually 0.877 is smaller than the smallest number in the row, which is 0.978, in the column with heading $t_{0.200}$. The value 0.978 cuts off a right tail of area 0.200, so because 0.877 is to its left it must cut off a tail of area greater than 0.200. Thus the p-value, which is the double of the area cut off (since the test is two-tailed), is greater than 0.400. Although its precise value is unknown, it must be greater than $\alpha = 0.01$, so the decision is not to reject H_0.

Key Takeaway

- There are two formulas for the test statistic in testing hypotheses about a population mean with small samples. One test statistic follows the standard normal distribution, the other Student's t-distribution.
- The population standard deviation is used if it is known, otherwise the sample standard deviation is used.
- Either five-step procedure, critical value or p-value approach, is used with either test statistic.

8.4: Hypothesis Test on a Single Standard Deviation

A test of a single standard deviation assumes that the underlying distribution is **normal**. The null and alternative hypotheses are stated in terms of the population standard deviation (or population variance). The test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{8.4.1}$$

where:

- n is the the total number of data
- s^2 is the sample variance
- σ^2 is the population variance

The requirements to be able to perform a hypothesis test on a population standard deviation are:

- the sample must be obtained from a simple random sample or from a randomized experiment
- the population has a normal distribution

You may think of s as the random variable in this test. The number of degrees of freedom is $df = n - 1$. **A test of a single standard deviation may be right-tailed, left-tailed, or two-tailed.** The next example will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

Example 8.4.1

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the standard deviation may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Answer

- $H_0 : \sigma = 5$
- $H_a : \sigma > 5$

Exercise 8.4.2

A SCUBA instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. Suppose the instructor finds a random sample of 25 SCUBA students and found that the sample standard deviation is 2.8 feet.

With a significance level of 5%, test the claim that the diving depths is less than 3 feet.

Answer

$H_0 : \sigma = 3$

$H_a : \sigma < 3$

The word **"less"** tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where $n =$ the number of customers sampled $df = n - 1 = 25 - 1 = 24$

Calculate the test statistic (Equation 8.4.1):

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(2.8)^2}{3^2} = 20.91$$

where $n = 25$, $s = 2.8$, and $\sigma = 3$.

Graph:

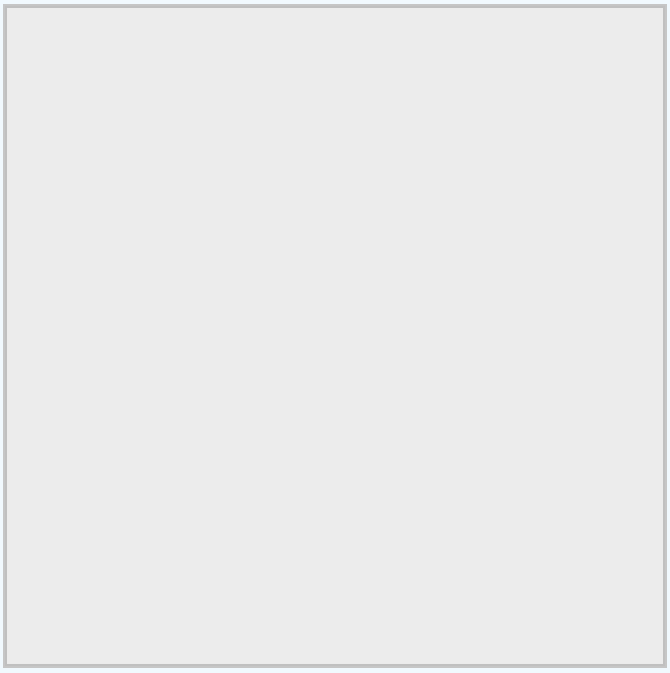


Figure 8.4.1 .

Probability statement: p-value $= P(\chi^2 < 20.91) = 0.356$

In `2nd DISTR` , use `7:χ2cdf` . The syntax is `(lower, upper, df)` for the parameter list. For Example, `χ2cdf(-1E99,20.91,24)` . The p-value $= 0.356$.

Compare α and the p-value:

Given $\alpha = 0.05, p\text{-value} = 0.356\}$, then $\alpha < p$-value

Make a decision: Since $\alpha < p$-value, fail to reject H_0. This means that you are not reject $\sigma = 3$. In other words, you do think the standard deviation of diving depth less than 3 feet.

Conclusion: At a 5% level of significance, from the data, there is not sufficient evidence to conclude that a SCUBA students' collective depth is less than 3 feet.

References

1. "AppleInsider Price Guides." Apple Insider, 2013. Available online at http://appleinsider.com/mac_price_guide (accessed May 14, 2013).
2. Data from the World Bank, June 5, 2012.

Review

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

Formula Review

$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ Test of a single variance statistic where:

n : sample size

s : sample standard deviation

σ : population standard deviation

$df = n - 1$ Degrees of freedom

Test of a Single Standard Deviation

- Use the test to determine standard deviation.
- The degrees of freedom is the number of samples $- 1$.
- The test statistic is $\frac{(n-1)\cdot s^2}{\sigma^2}$, where $n =$ the total number of data , $s^2 =$ sample variance, and $\sigma^2 =$ population variance.
- The test may be left-, right-, or two-tailed.

8.5: Hypothesis Test on a Single Variance

A test of a single variance assumes that the underlying distribution is **normal**. The null and alternative hypotheses are stated in terms of the population variance (or population standard deviation). The test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{8.5.1}$$

where:

- n is the the total number of data
- s^2 is the sample variance
- σ^2 is the population variance

You may think of s as the random variable in this test. The number of degrees of freedom is $df = n - 1$. **A test of a single variance may be right-tailed, left-tailed, or two-tailed.** The next example will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

Example 8.5.1

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Answer

Even though we are given the population standard deviation, we can set up the test using the population variance as follows.

- $H_0 : \sigma^2 = 5^2$
- $H_a : \sigma^2 > 5^2$

Exercise 8.5.1

A SCUBA instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. If the instructor were to conduct a test, what would the null and alternative hypotheses be?

Answer

- $H_0 : \sigma^2 = 3^2$
- $H_a : \sigma^2 > 3^2$

Example 8.5.2

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5%, test the claim that **a single line causes lower variation among waiting times (shorter waiting times) for customers**.

Answer

Since the claim is that a single line causes less variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable. Let $s =$ standard deviation for the waiting times.

- $H_0 : \sigma^2 = 7.2^2$
- $H_a : \sigma^2 < 7.2^2$

The word "**less**" tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where:

- $n =$ the number of customers sampled
- $df = n - 1 = 25 - 1 = 24$

Calculate the test statistic (Equation 8.5.1):

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(3.5)^2}{7.2^2} = 5.67$$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph:

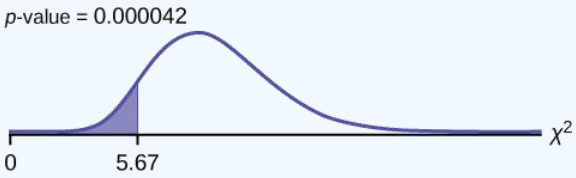


Figure 8.5.1 .

Probability statement: $p\text{-value} = P(\chi^2 < 5.67) = 0.000042$

Compare α and the p-value:

$$\alpha = 0.05 (p\text{-value} = 0.000042 \alpha > p\text{-value}$$

Make a decision: Since $\alpha > p$-value, reject H_0. This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes; you think the variation in waiting times is less.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times **or** with a single line, the customer waiting times vary less than 7.2 minutes.

In `2nd DISTR`, use `7:χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For Example, `χ2cdf(-1E99,5.67,24)`. The p-value $= 0.000042$

Exercise 8.5.2

The FCC conducts broadband speed tests to measure how much data per second passes between a consumer's computer and the internet. As of August of 2012, the standard deviation of Internet speeds across Internet Service Providers (ISPs) was 12.2 percent. Suppose a sample of 15 ISPs is taken, and the standard deviation is 13.2. An analyst claims that the standard deviation of speeds is more than what was reported. State the null and alternative hypotheses, compute the degrees of freedom, the test statistic, sketch the graph of the p-value, and draw a conclusion. Test at the 1% significance level.

Answer

- $H_0 : \sigma^2 = 12.2^2$
- $H_a : \sigma^2 > 12.2^2$

In `2nd DISTR`, use7: `χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. `χ2cdf(16.39,10^99,14)`. The p-value $= 0.2902$

$df = 14$

$$\text{chi}^2 \text{test statistic} = 16.39$$

Figure 8.5.2 .

The p-value is 0.2902 so we decline to reject the null hypothesis. There is not enough evidence to suggest that the variance is greater than 12.2^2.

References

1. "AppleInsider Price Guides." Apple Insider, 2013. Available online at http://appleinsider.com/mac_price_guide (accessed May 14, 2013).
2. Data from the World Bank, June 5, 2012.

Review

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

Formula Review

$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ Test of a single variance statistic where:

n : sample size

s : sample standard deviation

σ : population standard deviation

$df = n - 1$ Degrees of freedom

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom is the number of samples $- 1$.
- The test statistic is $\frac{(n-1)\cdot s^2}{\sigma^2}$, where $n =$ the total number of data , $s^2 =$ sample variance, and $\sigma^2 =$ population variance.
- The test may be left-, right-, or two-tailed.

Use the following information to answer the next three exercises: An archer's standard deviation for his hits is six (data is measured in distance from the center of the target). An observer claims the standard deviation is less.

Exercise 8.5.3

What type of test should be used?

Answer

a test of a single variance

Exercise 8.5.4

State the null and alternative hypotheses.

Exercise 8.5.5

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

a left-tailed test

Use the following information to answer the next three exercises: The standard deviation of heights for students in a school is 0.81. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the standard deviation of heights for the school is greater than 0.81.

Exercise 8.5.6

What type of test should be used?

Exercise 8.5.5

State the null and alternative hypotheses.

Answer

$H_0 : \sigma^2 = 0.81^2$;

$H_a : \sigma^2 > 0.81^2$

Exercise 8.5.6

$df =$ _____

Use the following information to answer the next four exercises: The average waiting time in a doctor's office varies. The standard deviation of waiting times in a doctor's office is 3.4 minutes. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.

Exercise 8.5.7

What type of test should be used?

Answer

a test of a single variance

Exercise 8.5.8

What is the test statistic?

Exercise 8.5.9

What is the p-value?

Answer

0.0542

Exercise 8.5.10

What can you conclude at the 5% significance level?

CHAPTER OVERVIEW

9: More Hypothesis Tests

9.1: 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:

$$\sum_{k} \frac{(O-E)^2}{E} \tag{9.1.1}$$

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 9.1.1.1

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 9.1.1.2

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 9.1.1

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 $(15, 12, 9, 9, 15)$
- $(O - E)$
- $(O - E)^2$
- $\frac{(O-E)^2}{E}$

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 $p\text{-value} = 0.5578$.)

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.)

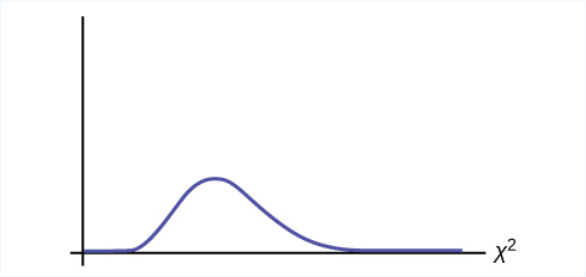


Figure 9.1.1.

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 9.1.2

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-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 9.1.3.1

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:

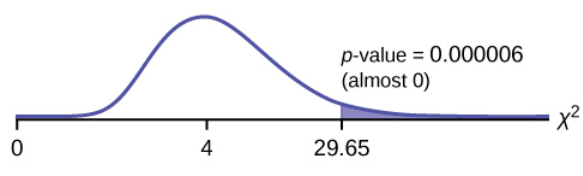


Figure 9.1.2.

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 $\alpha > p$-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 9.1.3

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-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:

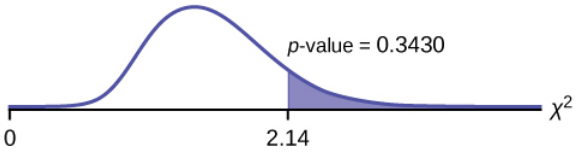


Figure 9.1.3.

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 $\alpha < p$-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 9.1.4

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 $= 26.38$

Figure 9.1.4.

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 9.1.5

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 9.1.6

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 9.1.7

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 9.1.8

$df =$ _____

Answer

3

Exercise 9.1.9

State the null and alternative hypotheses.

Exercise 9.1.10

$\chi^2 \text{test statistic} =$ _____

Answer

2.04

Exercise 9.1.11

p-value = _____

Exercise 9.1.12

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 9.1.13

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 9.1.14

H_0: _____

Answer

H_0: the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 9.1.15

H_a: _____

Exercise 9.1.16

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

right-tailed

Exercise 9.1.17

degrees of freedom = _____

Exercise 9.1.18

χ^2test statistic = _____

Answer

88,621

Exercise 9.1.19

p-value = _____

Exercise 9.1.20

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 9.1.5.

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-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 9.1.21

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?

9.2: 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:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{9.2.1}$$

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 9.2.1

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 9.2.1

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 9.2.2

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 9.2.1 is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Table 9.2.1: 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 9.2.2.

Table 9.2.2: 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:

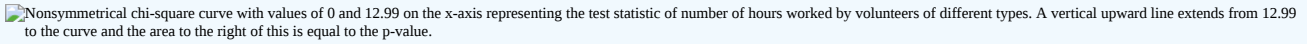


Figure 9.2.1.

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 $\alpha > p$-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 9.2.2

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 9.2.3 shows the results:

Table 9.2.3

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.

$df = 12$

Figure 9.2.2.

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 9.2.3

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.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09 \quad (9.2.2)$$

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 9.2.3

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 9.2.4

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 9.2.5

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 9.2.6

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 9.2.4 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 9.2.4

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 9.2.7

State the hypotheses.

- H_0: _____
- H_a: _____

Exercise 9.2.8

$df =$ _____

Answer

8

Exercise 9.2.9

How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

Exercise 9.2.10

How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

Answer

6.6

Exercise 9.2.11

What is the test statistic?

Exercise 9.2.12

What is the p-value?

Answer

0.0435

Exercise 9.2.13

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 9.2.14

Complete the table.

Table 9.2.5: 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 9.2.5B

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 9.2.15

State the hypotheses.

- H_0: _____

- H_a: _____

Exercise 9.2.16

Enter expected values in Table. Round to two decimal places.

Calculate the following values:

Answer

Table 9.2.6

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 9.2.17

$df =$ _____

Exercise 9.2.18

χ^2test statistic $=$ _____

Answer

10,301.8

Exercise 9.2.19

p-value $=$ _____

Exercise 9.2.20

Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

Answer

right

Exercise 9.2.21

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Figure 9.2.3.

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α.

Exercise 9.2.22

$\alpha = 0.05$

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 9.2.23

$\alpha = 0.05$

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.

9.3: ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses variances to help determine if the means are equal or not. To perform a one-way ANOVA test, there are several basic assumptions to be fulfilled:

Five basic assumptions of one-way ANOVA to be fulfilled

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).
4. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups:

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
- H_a : At least two of the group means$\mu_2 = \mu_3 = \ldots = \mu_k$are not equal

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), $H_0 : \mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

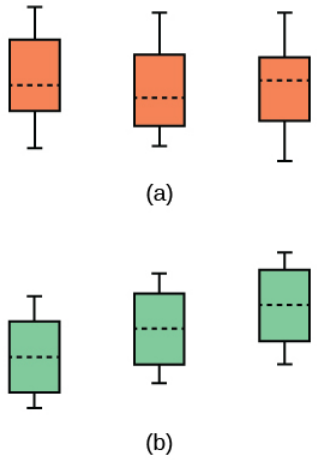


Figure 9.3.1: (a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

Review

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent, and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the F distribution with two different degrees of freedom.

Assumptions:

a. Each population from which a sample is taken is assumed to be normal.

b. All samples are randomly selected and independent.
c. The populations are assumed to have equal standard deviations (or variances).

Glossary

Analysis of Variance

also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

One-WayANOVA

a method of testing whether or not the means of three or more populations are equal; the method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

Variance

mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Contributors and Attributions

CHAPTER OVERVIEW

10: Hypothesis Testing with Two Samples

10.1: 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:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \tag{10.1.1}$$

The test statistic (t-score) is calculated as follows:

$$\frac{(\bar{x} - \bar{x}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{10.1.2}$$

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

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{10.1.3}$$

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 10.1.1 : 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 10.1.1 . Each populations has a normal distribution.

Table 10.1.1

	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: $p\text{-value} = 0.0054$

Graph:

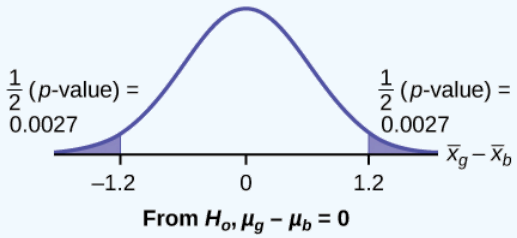


Figure 10.1.1 : Normal distribution curve representing the difference in the average amount of time girls and boys play sports all day

$$s_g = 0.866 \tag{10.1.4}$$

$$s_b = 1 \tag{10.1.5}$$

So,

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).

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 10.1.2

	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.

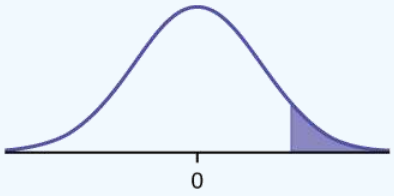
Example 10.1.2

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 $H_a : \mu_A > \mu_B$



Note: $\bar{x}_A - \bar{x}_B = 4 - 3.5 = 0.5$

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 10.1.2

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-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 10.1.3

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 10.1.3 and Table 10.1.4 .

Table 10.1.3 : 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 10.1.4 : 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-value $= 0.0011$

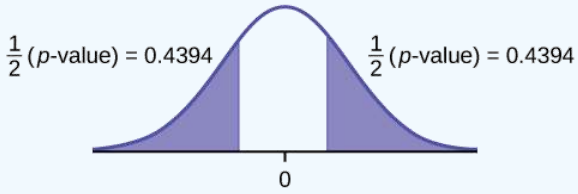


Figure 10.1.3 .

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 10.1.5 : 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 10.1.4

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 10.1.5

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-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.

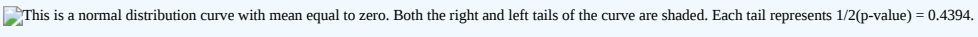


Figure 10.1.4 .

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:

$$SE = \sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}} \tag{10.1.6}$$

Test statistic (*t*-score):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{10.1.7}$$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{10.1.8}$$

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:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \tag{10.1.9}$$

where

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{10.1.10}$$

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.

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.

10.2: 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:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{10.2.1}$$

The distribution for the differences is:

$$P\_A - P_b' N \left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{10.2.2}$$

The test statistic (*z*-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{10.2.3}$$

Example 10.2.1

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:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 8001 - p_c = 0.92 \tag{10.2.4}$$

$$P'_A - P'_B - N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200}+\frac{1}{200}\right)}\right] \tag{10.2.5}$$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution: p-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:

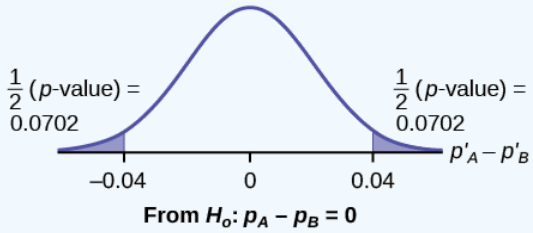


Figure 10.4.1.

$P'_A - P'_B = 0.1 - 0.06 = 0.04$.

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 10.2.1

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 10.2.2 : 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
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:

$$p_C = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077 \tag{10.2.6}$$

$$1 - p_C = 0.923 \tag{10.2.7}$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right) \tag{10.2.8}$$

$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:

p-value $= 0.1045$

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:

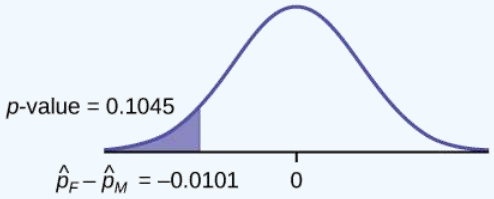


Figure 10.4.2.

Decision: Since $\alpha < p$-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

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:

$$p_C = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927 \tag{10.2.9}$$

$$1 - p_C = 0.9073 \tag{10.2.10}$$

Therefore,

$$p'_W - p'_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right) \tag{10.2.11}$$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:

p-value $= 0.0077$

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05

Graph:

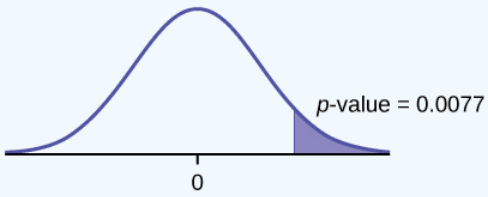


Figure 10.4.3.

Decision: Since $\alpha > p$-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

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 $= 0.00086$

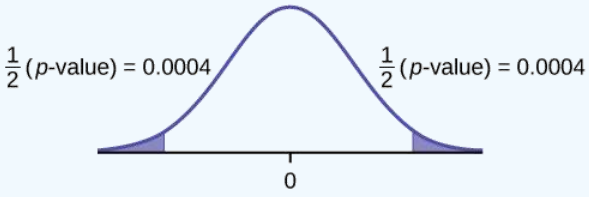


Figure 10.4.4.

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:

$$p_c = \frac{x_F + x_M}{n_F + n_M} \tag{10.2.12}$$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}\right] \tag{10.2.13}$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

Test Statistic (*z*-score):

$$z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \tag{10.2.14}$$

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.

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.

10.3.1: Matched or Paired Samples Part 1

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:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{10.3.1.1}$$

The distribution for the differences is:

$$p_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{10.3.1.2}$$

The test statistic (*z*-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{10.3.1.3}$$

Example 10.3.1.1

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:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 8001 - p_c = 0.92 \tag{10.3.1.4}$$

$$p'_A - p'_B \sim N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200}+\frac{1}{200}\right)}\right] \quad (10.3.1.5)$$

$p'_A - p'_B$ 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:

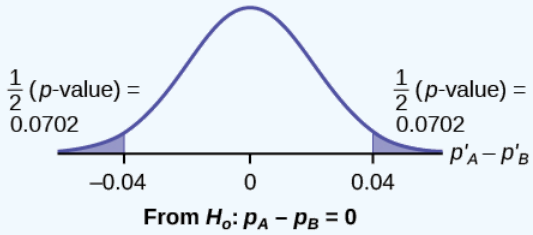


Figure 10.4.1.

$p'_A - p'_B = 0.1 - 0.06 = 0.04$.

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 $p\text{-value} = 0.1404$. $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, 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 10.3.1.1

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 10.3.1.2: 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:

$$p_C = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077 \tag{10.3.1.6}$$

$$1 - p_C = 0.923 \tag{10.3.1.7}$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right) \tag{10.3.1.8}$$

$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:

p-value $= 0.1045$

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:

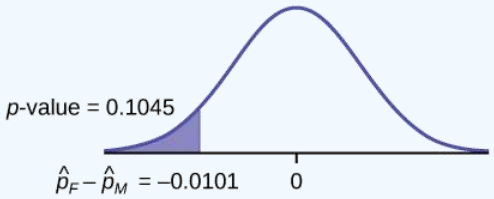


Figure 10.4.2.

Decision: Since $\alpha < p$-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 10.3.1.3

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:

$$p_C = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927 \tag{10.3.1.9}$$

$$1 - p_C = 0.9073 \tag{10.3.1.10}$$

Therefore,

$$p'_W - p'_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right) \tag{10.3.1.11}$$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:

p-value $= 0.0077$

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05

Graph:

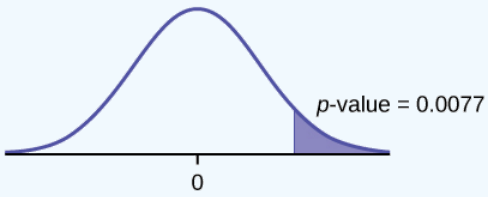


Figure 10.4.3.

Decision: Since $\alpha > p$-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 10.3.1.3

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 $= 0.00086$

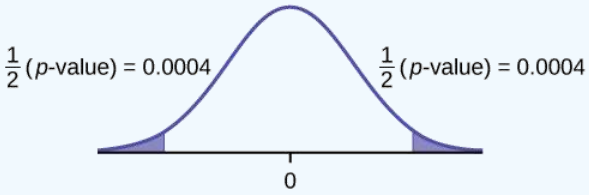


Figure 10.4.4.

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:

$$p_c = \frac{x_F + x_M}{n_F + n_M} \qquad (10.3.1.12)$$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}\right] \quad (10.3.1.13)$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

Test Statistic (*z*-score):

$$z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \quad (10.3.1.14)$$

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.

10.3.2: Matched or Paired Samples Part 2

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:

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \qquad (10.3.2.1)$$

Example 10.3.2.1

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

$$H_0 : \mu_d \geq 0 \tag{10.3.2.2}$$

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.

$$H_a : \mu_d < 0 \tag{10.3.2.3}$$

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:

$$p\text{-value} = 0.0095 \tag{10.3.2.4}$$

Graph:

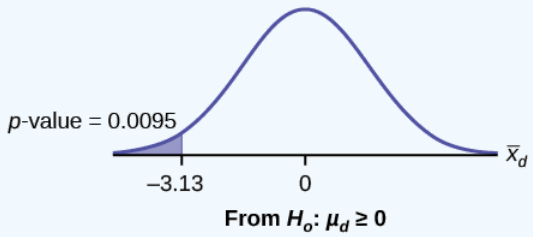


Figure 10.5.1.

$\bar{X}_d$ 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 10.3.2.1

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 10.3.2.2

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.

$$\bar{x}_d = 21.3 \tag{10.3.2.5}$$

and

$$s_d = 46.7 \tag{10.3.2.6}$$

> 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:

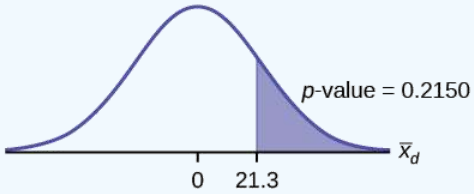


Figure 10.5.2.

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 $\alpha < p$-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 10.3.2.2

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 10.3.2.3

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:

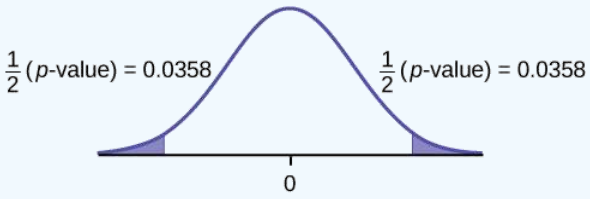


Figure 10.5.3.

Calculate the *p*-value: The p-value is 0.0716 (using the data directly).

(test statistic = 2.18. p-value $= 0.0719$ using $(\bar{x}_d = 3.71, s_d = 4.5$.

Decision: Assume $\alpha = 0.05$. Since $\alpha < p$-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 10.3.2.3

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):

$$t = \frac{\bar{x}_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \qquad (10.3.2.7)$$

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.

10.4: Test of Two Variances

Another of the uses of the F distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

to perform a F test of two variances, it is important that the following are true:

- The populations from which the two samples are drawn are *normally* distributed.
- The two populations are *independent* of each other.

Unlike most other tests in this book, the F test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher p-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the F ratio:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} \tag{10.4.1}$$

F has the distribution

$$F \sim F(n_1 - 1, n_2 - 1) \tag{10.4.2}$$

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F Ratio becomes

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}. \tag{10.4.3}$$

The F ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on H_a and on which sample variance is larger.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to one. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than one. If s_1^2 and s_2^2 are far apart, then

$$F = \frac{(s_1)^2}{(s_2)^2} \tag{10.4.4}$$

is a large number.

Therefore, if F is close to one, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than one, then the evidence is against the null hypothesis. A test of two variances may be left, right, or two-tailed.

> *A test of two variances may be left, right, or two-tailed.*

Example 10.4.1

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

Answer

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

- $n_1 = n_2 = 30$.
- $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_a : \sigma_1^2 < \sigma_2^2$

Calculate the test statistic: By the null hypothesis $\sigma_1^2 = \sigma_2^2$), the F statistic is:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(s_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818 \tag{10.4.5}$$

Distribution for the test: $F_{29,29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

Graph: This test is left tailed.

Draw the graph labeling and shading appropriately.

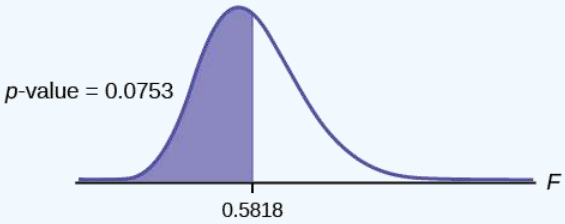


Figure 10.4.1

Probability statement: p-value $= P(F < 0.5818) = 0.0753$

Compare α and the p-value: $\alpha = 0.10 \alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

Conclusion: With a 10% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Press `STAT` and arrow over to `TESTS`. Arrow down to `D:2-SampFTest`. Press `ENTER`. Arrow to `Stats` and press `ENTER`. For `Sx1`, `n1`, `Sx2`, and `n2`, enter `(52.3)-----√(52.3)`, `30`, `(89.9)-----√(89.9)`, and `30`. Press `ENTER` after each. Arrow to `σ1:` and `<σ2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. F = 0.5818 and p-value = 0.0753. Do the procedure again and try `Draw` instead of `Calculate`.

Exercise 10.4.1

The New York Choral Society divides male singers up into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the variance of the heights of singers in each of these two groups (Tenor1 and Bass2) are different?

Tenor1	Bass2	Tenor 1	Bass 2	Tenor 1	Bass 2
69	72	67	72	68	67

Tenor1	Bass2	Tenor 1	Bass 2	Tenor 1	Bass 2
72	75	70	74	67	70
71	67	65	70	64	70
66	75	72	66		69
76	74	70	68		72
74	72	68	75		71
71	72	64	68		74
66	74	73	70		75
68	72	66	72		

Answer

The histograms are not as normal as one might like. Plot them to verify. However, we proceed with the test in any case.

Subscripts: $\mathrm{T1}$ = tenor 1 and $\mathrm{B2}$ = bass 2

The standard deviations of the samples are $s_{\mathrm{T1}} = 3.3302$ and $s_{\mathrm{B2}} = 2.7208$.

The hypotheses are

$H_0 : \sigma^2_{\mathrm{T1}} = \sigma^2_{\mathrm{B2}}$ and $H_0 : \sigma^2_{\mathrm{T1}} \neq \sigma^2_{\mathrm{B2}}$ (two tailed test)

The F statistic is $1.4894$with 20 and 25 degrees of freedom.

The p-value is 0.3430 If we assume alpha is 0.05, then we cannot reject the null hypothesis.

We have no good evidence from the data that the heights of Tenor1 and Bass2 singers have different variances (despite there being a significant difference in mean heights of about 2.5 inches.)

References

1. “MLB Vs. Division Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/y...ion/order/true.

Review

The F test for the equality of two variances rests heavily on the assumption of normal distributions. The test is unreliable if this assumption is not met. If both distributions are normal, then the ratio of the two sample variances is distributed as an F statistic, with numerator and denominator degrees of freedom that are one less than the samples sizes of the corresponding two groups. A **test of two variances** hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the F distribution with two different degrees of freedom.

Assumptions:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

Formula Review

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

If $\sigma_1 = \sigma_2$, then $F = \frac{s_1^2}{s_2^2}$

CHAPTER OVERVIEW

11: Correlation

11.1.1: Correlation Concepts Part 1

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 11.1.1.1

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 11.1.1.1: 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

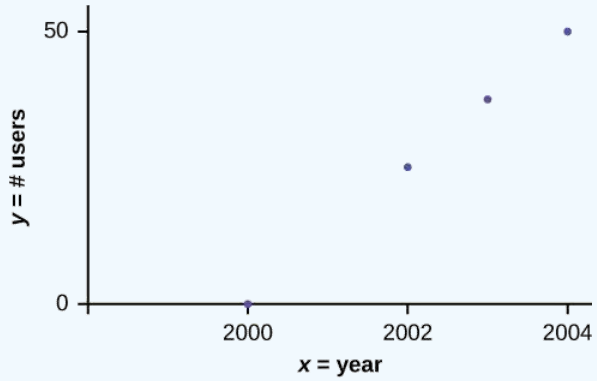


Figure 11.1.1.1: Scatter plot showing the number of m-commerce users (in millions) by year.

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 11.1.1.1

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

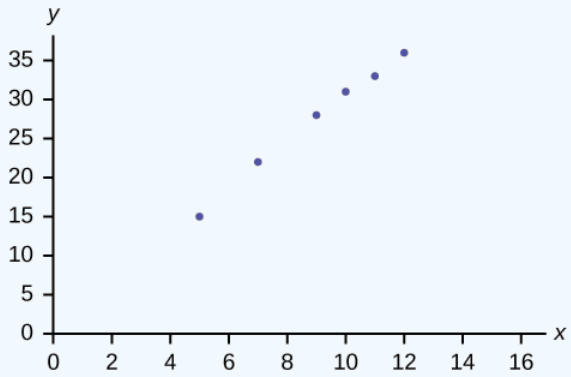


Figure 11.1.1.2

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.

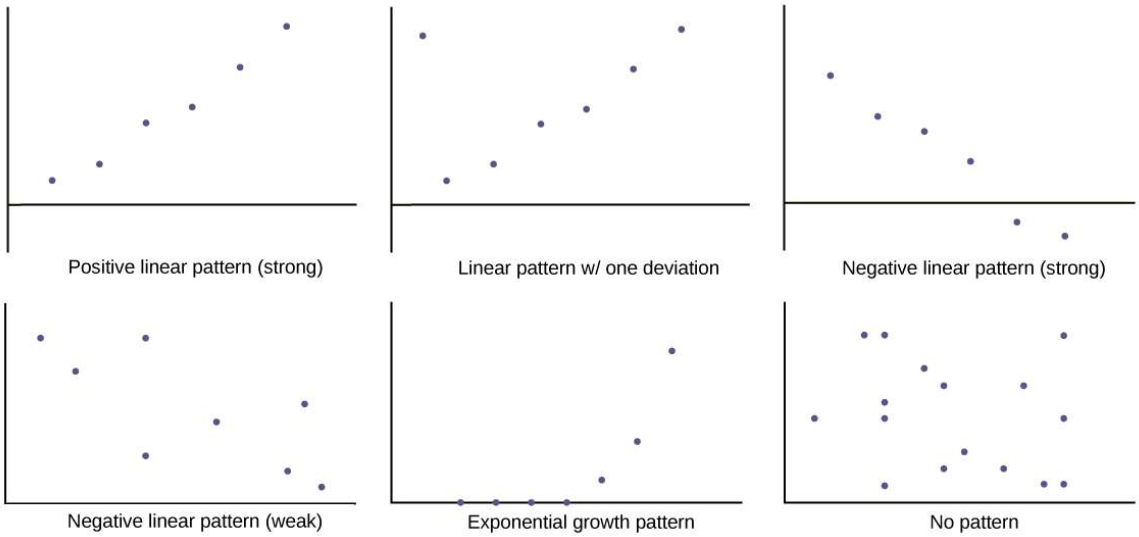


Figure 11.1.1.3:

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.

11.1.2: Correlation Concepts Part 2

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to "**fit**" a straight line. This is called a Line of Best Fit **or** Least-Squares Line.

COLLABORATIVE EXERCISE

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 11.1.2.1

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

1a: Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

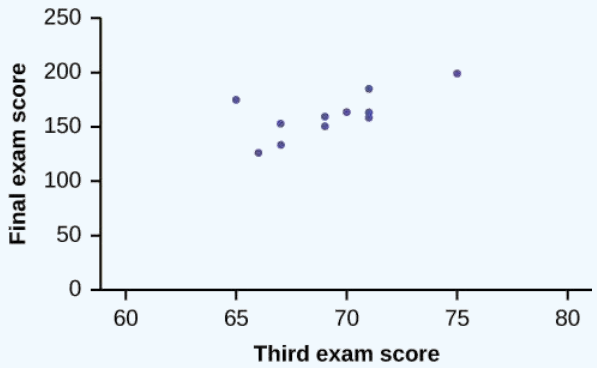


Figure 11.1.2.1: Scatter plot showing the scores on the final exam based on scores from the third exam.

Exercise 11.1.2.1

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in Table show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

X (depth in feet)	Y (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Answer

$\hat{y} = 127.24 - 1.11x$

At 110 feet, a diver could dive for only five minutes.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a least-squares regression line to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read "y **hat**" and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

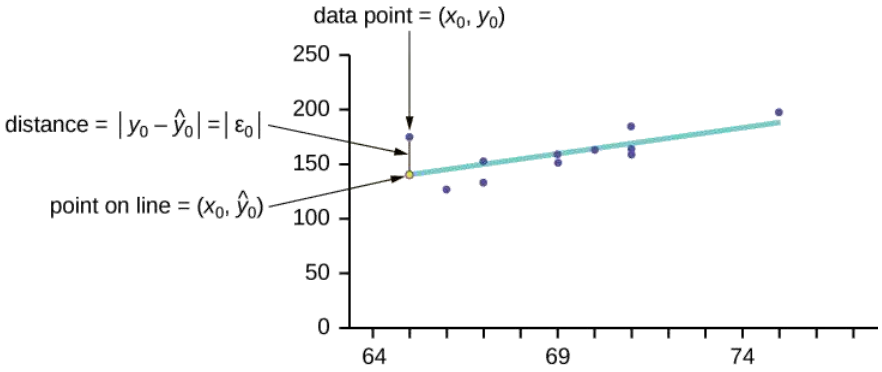


Figure 11.1.2.2

The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the **"error" or** residual. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram in Figure, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

$\varepsilon =$ the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, \ldots, 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + \ldots + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2 \tag{11.1.2.1}$$

Equation11.1.2.1is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx \tag{11.1.2.2}$$

where

- $a = \bar{y} - b\bar{x}$ and
- $b = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$.

The sample means of the x values and the x values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r\left(\dfrac{s_y}{s_x}\right)$ where $s_y =$ the standard deviation of the y values and $s_x =$ the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

Least Square Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line** .

> **Note**
>
> Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third-exam/final-exam example is as follows:

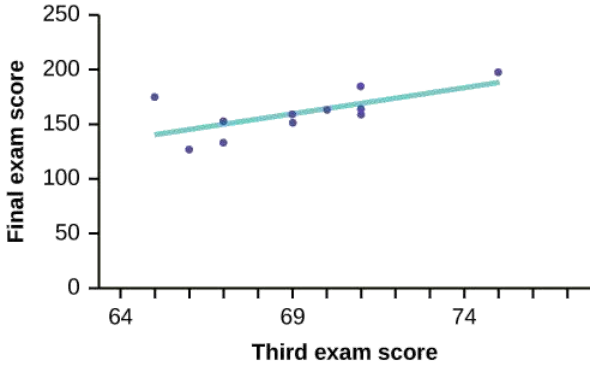


Figure 11.1.2.3

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \tag{11.1.2.3}$$

REMINDER

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for *x*-values outside that domain.** You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

Understanding Slope

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Using the Linear Regression T Test: LinRegTTest

a. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x, y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
b. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
c. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
d. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER
e. Leave the line for "RegEq:" blank
f. Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen

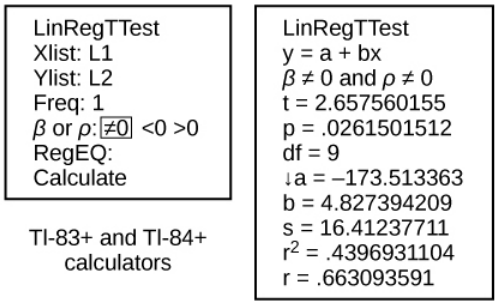


Figure 11.1.2.4

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$, and $b = 4.8273$; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are $r_2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1

3. On the input screen for PLOT 1, highlight **On**, and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
8. To graph the best-fit line, press the "$Y =$" key and type the equation $-173.5 + 4.83X$ into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

Note

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

a. Make sure you have done the scatter plot. Check it on your screen.
b. Go to LinRegTTest and enter the lists.
c. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
d. Press $Y = (\text{you will see the regression equation})$.
e. Press GRAPH. The line will be drawn."

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y. The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \qquad (11.1.2.4)$$

where $n =$ the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

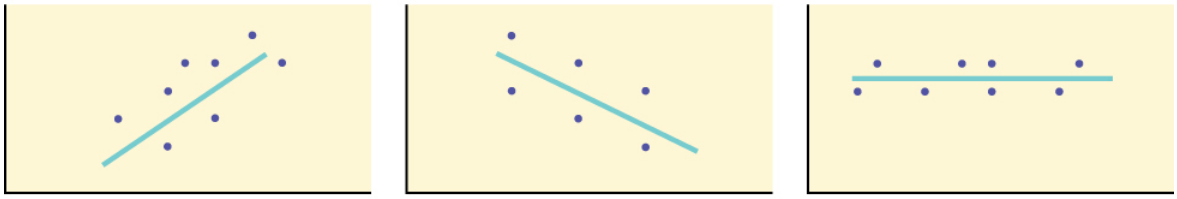
- The value of r is always between –1 and +1: $-1 \le r \le 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to –1 or to +1 indicate a stronger linear relationship between x and y.
- If $r = 0$ there is absolutely no linear relationship between x and y **(no linear correlation)**.
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course,in the real world, this will not generally happen.

What the SIGN of r tells us:

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best-fit line.

Note

Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."**



(a) Positive correlation (b) Negative correlation (c) Zero correlation

Figure 11.1.2.5: (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called *the coefficient of determination* and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation $(1 - 0.44 = 0.56)$ in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Summary

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Glossary

Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{11.1.2.5}$$

where n is the number of data points. The coefficient cannot be more than 1 or less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

11.2: Correlation Hypothesis Test

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together. We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- $\rho =$ population correlation coefficient (unknown)
- $r =$ sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

NOTE

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

- **Null Hypothesis:** $H_0 : \rho = 0$
- **Alternate Hypothesis:** $H_a : \rho \neq 0$

WHAT THE HYPOTHESES MEAN IN WORDS:

- **Null Hypothesis H_0:** The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis H_a:** The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:There are two methods of making the decision. The two methods are equivalent and give the same result.

- **Method 1: Using the p-value**
- **Method 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

NOTE

Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

Using the TI83, 83+, 84, 84+ CALCULATOR

To calculate the p-value using LinRegTTEST:

On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq 0$"

The output screen shows the p-value on the line that reads "$p =$".

(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the p-value. The following describes the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t-distribution with $n - 2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.
- The p-value is the combined area in both tails.

An alternative way to calculate the p-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: p-value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.
- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 11.2.1

Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and $+0.632$. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

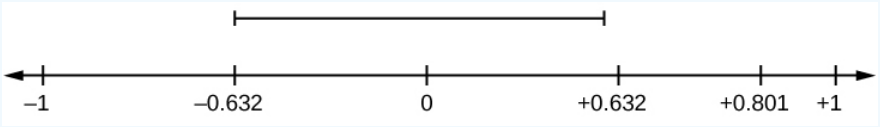


Figure 11.2.1. r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

Exercise 11.2.1

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points and the critical value is 0.576. Can the line be used for prediction? Why or why not?

Answer

If the scatter plot looks linear then, yes, the line can be used for prediction, because $r >$ the positive critical value.

Example 11.2.2

Suppose you computed $r = -0.624$ with 14 data points. $df = 14 - 2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line can be used for prediction

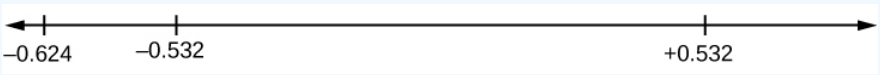


Figure 11.2.2. $r = -0.624 - 0.532$. Therefore, r is significant.

Exercise 11.2.2

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

Answer

No, the line cannot be used for prediction, because $r <$ the positive critical value.

Example 11.2.3

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.

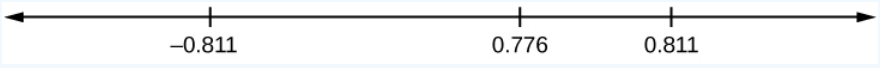


Figure 11.2.3. $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

Exercise 11.2.3

For a given line of best fit, you compute that $r = -0.7204$ using $n = 8$ data points, and the critical value is $= 0.707$. Can the line be used for prediction? Why or why not?

Answer

Yes, the line can be used for prediction, because $r <$ the negative critical value.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: critical value method

Consider the third exam/final exam example. The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$
- $\alpha = 0.05$

- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$.
- The critical values are -0.602 and $+0.602$
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion:There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Example 11.2.4

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

a. $r = -0.567$ and the sample size, n, is 19. The $df = n - 2 = 17$. The critical value is -0.456. $-0.567 < -0.456$ so r is significant.
b. $r = 0.708$ and the sample size, n, is 9. The $df = n - 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
c. $r = 0.134$ and the sample size, n, is 14. The $df = 14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not significant.
d. $r = 0$ and the sample size, n, is five. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

Exercise 11.2.4

For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

Answer

No, the line cannot be used for prediction no matter what the sample size is.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatter plot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.

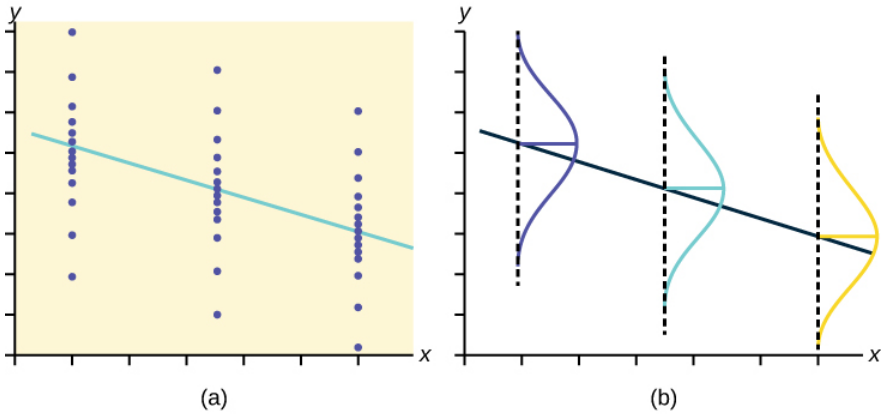


Figure 11.2.4. The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

Summary

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.
- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ, use the standard deviation of the residuals, s. $s = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis $H_0 : \rho =$ *hypothesized value*, use a linear regression t-test. The most common null hypothesis is $H_0 : \rho = 0$ which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

$$\hat{y} = a + bx \tag{11.2.1}$$

where

$$a = y\text{-intercept} \tag{11.2.2}$$

$$b = \text{slope} \tag{11.2.3}$$

Standard deviation of the residuals:

$$s = \sqrt{\frac{SSE}{n-2}} \tag{11.2.4}$$

where

$$SSE = \text{sum of squared errors} \tag{11.2.5}$$

$$n = \text{the number of data points} \tag{11.2.6}$$

11.3: Normal Probability Plots

The distributions you have seen up to this point have been assumed to be normally distributed, but how do you determine if it is normally distributed? One way is to take a sample and look at the sample to determine if it appears normal. If the sample looks normal, then most likely the population is also. Here are some guidelines that are use to help make that determination.

Normal quantile plot (or normal probability plot): This plot is provided through statistical software on a computer or graphing calculator. If the points lie close to a line, the data comes from a distribution that is approximately normal. If the points do not lie close to a line or they show a pattern that is not a line, the data are likely to come from a distribution that is not normally distributed.

To create a normal quantile plot on the TI-83/84

1. Go into the STAT menu, and then Chose 1:Edit

Figure 11.3.10 *: STAT Menu on TI-83/84*

2. Type your data values into L1. If L1 has data in it, arrow up to the name L1, click CLEAR and then press ENTER. The column will now be cleared and you can type the data in.
3. Now click STAT PLOT ($2^{nd}\ Y=$). You have three stat plots to choose from.

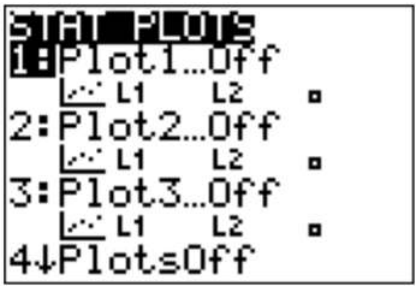


Figure 11.3.11 *: STAT PLOT Menu on TI-83/84*

4. Use 1:Plot1. Press ENTER.
5. Put the cursor on the word On and press ENTER. This turns on the plot. Arrow down to Type: and use the right arrow to move over to the last graph (it looks like an increasing linear graph). Set Data List to L1 (it might already say that) and set Data Axis to Y. The Mark is up to you.

Figure 11.3.12 *: Plot1 Menu on TI-83/84 Setup for Normal Quantile Plot*

6. Now you need to set up the correct window on which to graph. Click on WINDOW. You need to set up the settings for the *x* variable. Xmin should be -4. Xmax should be 4. Xscl should be 1. Ymin and Ymax are based on your data, the Ymin should be below your lowest data value and Ymax should be above your highest data value. Yscl is just how often you would like to see a tick mark on the *y*-axis.

7. Now press GRAPH. You will see the normal quantile plot.

Example 11.3.1 is it normal?

In Kiama, NSW, Australia, there is a blowhole. The data in table #6.4.1 are times in seconds between eruptions ("Kiama blowhole eruptions," 2013). Do the data come from a population that is normally distributed?

Table 11.3.1 : Time (in Seconds) Between Kiama Blowhole Eruptions

83	51	87	60	28	95	8	27
15	10	18	16	29	54	91	8
17	55	10	35	47	77	36	17
21	36	18	40	10	7	34	27
28	56	8	25	68	146	89	18
73	69	9	37	10	82	29	8
60	61	61	18	169	25	8	26
11	83	11	42	17	14	9	12

a. State the random variable
b. Draw the normal scatterplot.
c. Do the data come from a population that is normally distributed?

Solution

a. x = time in seconds between eruptions of Kiama Blowhole

b. The normal scatterplot is in *Figure 11.3.15* .

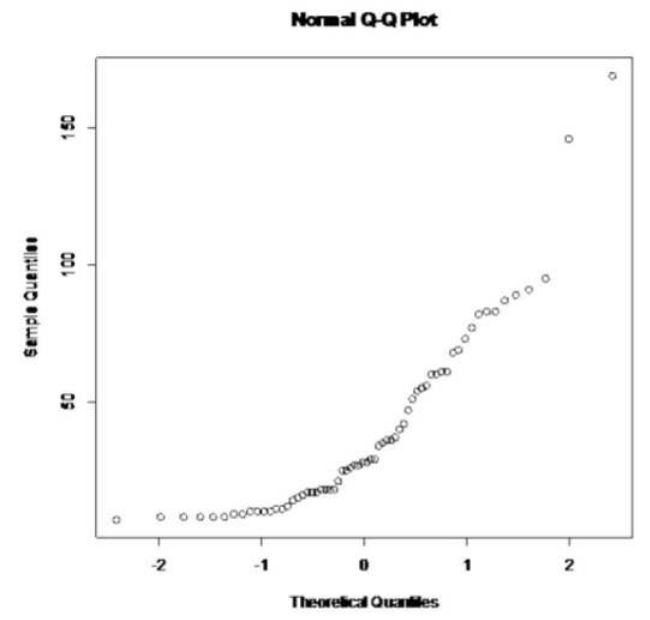


Figure 11.3.15 : Normal Probability Plot

This graph looks more like an exponential growth than linear.

c. Considering the histogram is skewed right, there are two extreme outliers, and the normal probability plot does not look linear, then the conclusion is that this sample is not from a population that is normally distributed.

Example 11.3.2 is it normal?

One way to measure intelligence is with an IQ score. Example 11.3.2 contains 50 IQ scores. Determine if the sample comes from a population that is normally distributed.

Table 11.3.2 : IQ Scores

78	92	96	100	67	105	109	75	127	111
93	114	82	100	125	67	94	74	81	98
102	108	81	96	103	91	90	96	86	92
84	92	90	103	115	93	85	116	87	106
85	88	106	104	102	98	116	107	102	89

a. State the random variable.
b. Draw the normal scatterplot.
c. Do the data come from a population that is normally distributed?

Solution

a. x = IQ score

b. The normal scatterplot is in *Figure 11.3.18* .

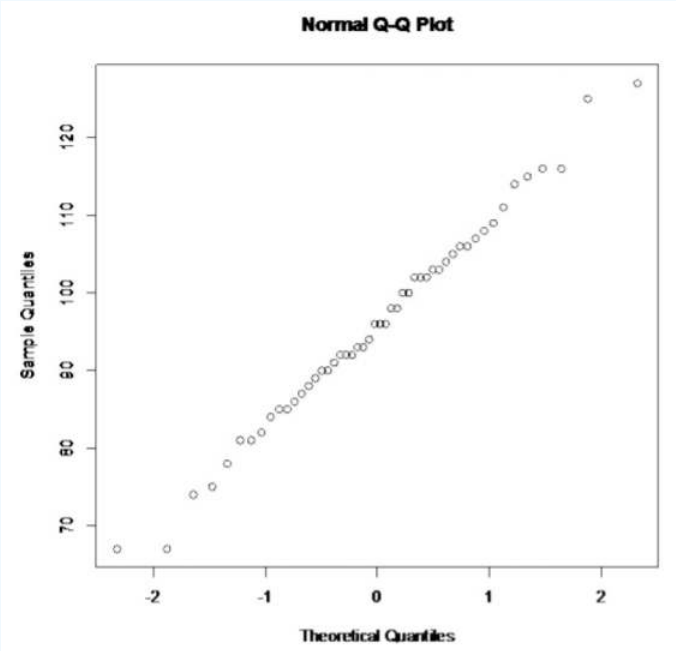


Figure 11.3.18 : Normal Quantile Plot

This graph looks fairly linear.

c. Considering the histogram is somewhat symmetric, there are no outliers, and the normal probability plot looks linear, then the conclusion is that this sample is from a population that is normally distributed.

Hypothesis Test on Normality:

A Normal Probability Plot is a scatterplot that show the relationship between a data value (x-value) and its predicted z-score (y-value). If the normal probability plot shows a linear relationship and a hypothesis test for ρ shows that there is a linear relationship, we can assume the population is approximately normal. (Recall from Section 11.3, if two variables do show a linear relationship, then $\rho \neq 0$.)

Exercise 11.3.1

1. Cholesterol data was collected on patients four days after having a heart attack. The data is in Example 11.3.3 . Determine if the data is from a population that is normally distributed.

Table 11.3.3 : Cholesterol Data Collected Four Days After a Heart Attack

218	234	214	116	200	276	146
182	238	288	190	236	244	258
240	294	220	200	220	186	352
202	218	248	278	248	270	242

2. The size of fish is very important to commercial fishing. A study conducted in 2012 collected the lengths of Atlantic cod caught in nets in Karlskrona (Ovegard, Berndt & Lunneryd, 2012). Data based on information from the study is in Example 11.3.4 . Determine if the data is from a population that is normally distributed.

Table 11.3.4 : Atlantic Cod Lengths

48	50	50	55	53	50	49	52
61	48	45	47	53	46	50	48
42	44	50	60	54	48	50	49
53	48	52	56	46	46	47	48
48	49	52	47	51	48	45	47

3. The WHO MONICA Project collected blood pressure data for people in China (Kuulasmaa, Hense & Tolonen, 1998). Data based on information from the study is in Example 11.3.5 . Determine if the data is from a population that is normally distributed.

Table 11.3.5 : Blood Pressure Values for People in China

114	141	154	137	131	132	133	156	119
138	86	122	112	114	177	128	137	140
171	129	127	104	97	135	107	136	118
92	182	150	142	97	140	106	76	115
119	125	162	80	138	124	132	143	119

4. Annual rainfalls for Sydney, Australia are given in Example 11.3.6 . ("Annual maximums of," 2013). Can you assume rainfall is normally distributed?

Table 11.3.6 : Annual Rainfall in Sydney, Australia

146.8	383	90.9	178.1	267.5	95.5	156.5	180
90.9	139.7	200.2	171.7	187.2	184.9	70.1	58
84.1	55.6	133.1	271.8	135.9	71.9	99.4	110.6
47.5	97.8	122.7	58.4	154.4	173.7	118.8	88
84.6	171.5	254.3	185.9	137.2	138.9	96.2	85
45.2	74.7	264.9	113.8	133.4	68.1	156.4	

Answer

1. Normally distributed
3. Normally distributed

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Math 130: Statistics

Webpages: 63

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 52.4% (33 pages)
- CC BY-NC 4.0: 41.3% (26 pages)
- Undeclared: 4.8% (3 pages)
- CC BY-SA 4.0: 1.6% (1 page)

By Page