BFE 1201 STATISTICAL METHODS FOR FINANCE (KUTER)

LibreTexts™

BFE 1201 Statistical Methods for Finance
(Kuter)

This text was compiled on 03/18/2025

TABLE OF CONTENTS

4: Random Variables

5: Point Estimates

8: Linear Regression and Correlation

9: Apppendices

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

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

CHAPTER OVERVIEW

1: Sampling and Data

1.1: Introduction to Sampling and Data

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Figure 1.1.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

1.2: Definitions of Statistics, Probability, and Key Terms

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

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 330 ml can contains 330 ml 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, in this case the mean. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. 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**, or random variable, usually notated by capital letters such as X and Y, is a characteristic or measurement that can be determined for each member of 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. Numerical variables are also referred to as **quantitative** variables, and categorical variables may be referred to as **qualitative** variables. 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 in the US, 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.

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."

Example 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 surveyed 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Answer

Solution 1.1

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 population mean.

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

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.

? Example 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

Solution 1.2

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

? Example 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 Traffic Safety Council of Zimbabwe collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Table 1.1

Speed at which cars crashed	Location of "drive" (i.e. dummies)
55 km/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 55 km 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

Solution 1.3

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 1.4

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

Solution 1.4

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.

1.3: Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample. Lowercase 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. **Qualitative data** are also often called categorical data. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative(categorical) data. Qualitative(categorical) 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(categorical) 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.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

? Example 1.3.1: 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 1.3.1

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?

? Example 1.3.2: 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.

? Exercise 1.3.2

The data are the areas of lawns in square meters. You sample five houses. The areas of the lawns are 13 sq. meters, 15 sq. meters, 17.5 sq. meters, 17 sq. meters, and 19.5 sq. meters. What type of data is this?

? Example 1.3.3

You go to the supermarket and purchase three cans of soup (560 ml of tomato bisque, 420 ml of lentil, and 560 ml of Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetables (broccoli, cauliflower, spinach, and carrots), and two desserts 450 grams of pistachio ice cream and 900 grams of chocolate chip cookies).

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

Answer

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 (560 ml, 420 ml, 560 ml) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative(categorical) data because they are categorical.

Try to identify additional data sets in this example.

? Example 1.3.4

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(categorical) data.

? Exercise 1.3.4

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?

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as letter grades A, B, C, D, or F.

? Example 1.3.5

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. the distance from your home to the nearest grocery store
d. the number of classes you take per school year
e. the type of calculator you use
f. weights of sumo wrestlers
g. number of correct answers on a quiz
h. IQ scores (This may cause some discussion.)

Answer

Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative, or categorical.

? Exercise 1.3.5

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.

? Example 1.3.6

A statistics professor collects information about the classification of her students as freshmen (Part 1), sophomores (Part 2), juniors (Part 3), or seniors (Part 4). The data she collects are summarized in the pie chart in Figure 1.2 (pie charts will be reviewed in a later chapter). What type of data does this graph show?

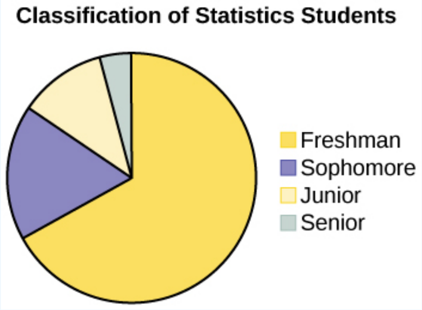


Figure 1.2

Answer

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

Exercise 1.3.6

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 in Figure 1.3 (histograms will be reviewed in a later chapter). 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.

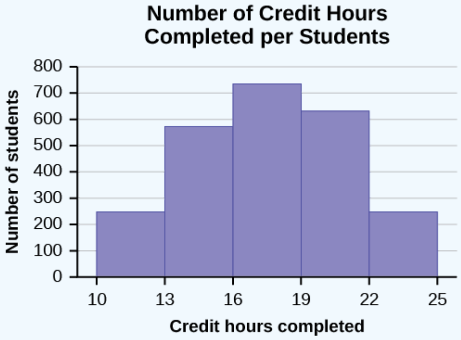


Figure 1.3

What type of data does this graph show?

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 term. 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 1.3.1: 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(categorical) 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 Figure 1.5 and determine which graph (pie or bar) you think displays the comparisons better.

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.

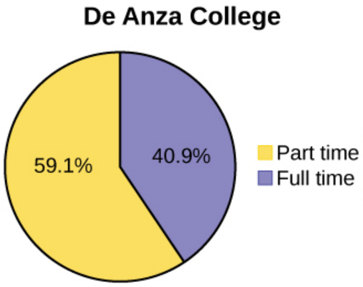


Figure 1.4a

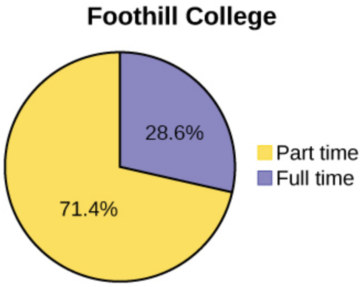


Figure 1.4B

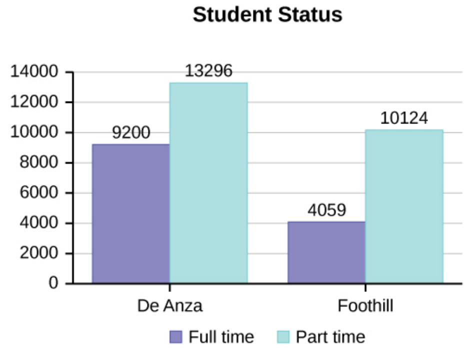


Figure 1.5

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 1.3.2: 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%

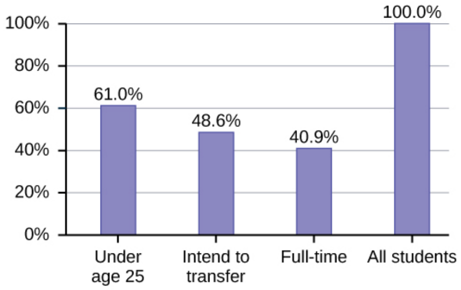


Figure 1.3.2

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

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

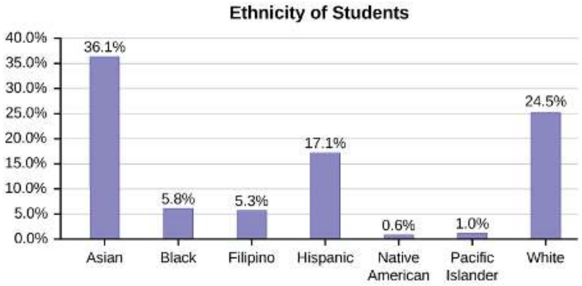


Figure 1.3.3

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 1.9 is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

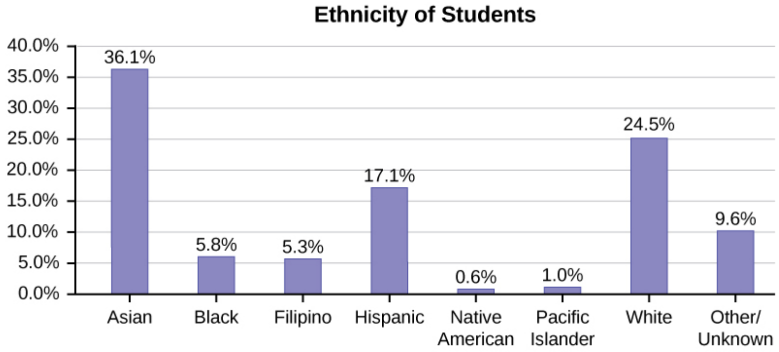


Figure 1.3.4: **Bar Graph with Other/Unknown Category**

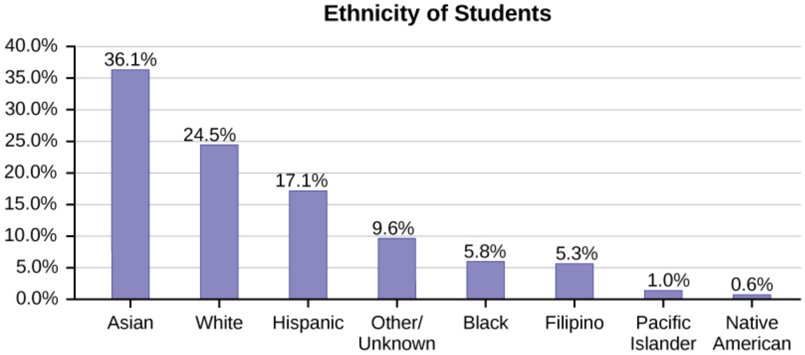


Figure 1.3.4: **Pareto Chart With Bars Sorted by Size**

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 1.10.

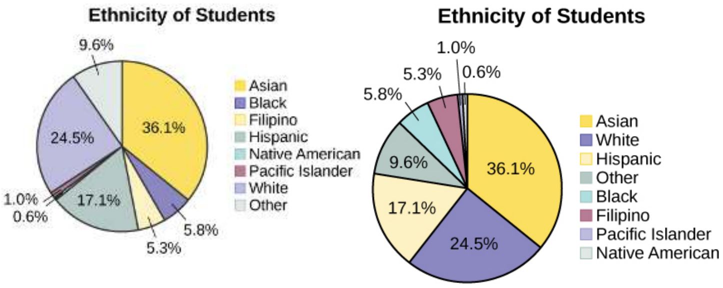


Figure 1.3.5: Pie Charts with no missing data

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 as 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.

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 k^{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.

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.

Example 1.3.7

A study is done to determine the average tuition that Saint Mary's College 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 Saint Mary's 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 1.3.8

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

a. A football 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 football club.
b. A pollster interviews all human resource personnel in five different 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

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 1.3.8

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 Physical Education 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

It is unlikely that any student is in both samples.

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 part-time 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

Solution 1.13

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

Solution 1.13

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

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 1.3.8

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?

Variation in Data

Variation is present in any set of data. For example, 330 ml cans of beverage may contain more or less than 330 ml of liquid. In one study, eight 330 ml cans were measured and produced the following amount (in ml) of beverage:

328.1; 330.3; 335.4; 330.0; 315.9; 328.7; 329.1; 363.5

Measurements of the amount of beverage in a 330 ml can may vary because different people make the measurements or because the exact amount, 330 ml of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 330 ml 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, usually given the symbol n 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. Later we will find that even much smaller sample sizes will give very good results. 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.

1.4: Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **independent variable** or **explanatory variable**. The affected variable is called the **dependent variable** or **response variable**: stimulus, response. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.* (McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.)

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

? Example 1.4.19

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.
d. Is it possible to use blinding in this study?

Answer

Solution 1.19

The explanatory variable is scent, and the response variable is the time it takes to complete the maze. There are two treatments: a floral-scented mask and an unscented mask. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

1.5: Sampling Experiment (Worksheet)

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this exercise, you will be asked to pick several random samples of companies from the list below. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the companies in the sample you obtained.

The following table contains companies stratified by industry type into rows and grouped vertically by Stock Price at close on 20 September 2023 (clusters)

Companies Used in Sample

Stock Price	Under $50	$50 to under $100	$100 to under $200	Over $200
apparel, fitness, footwear, retail	american eagle outfitters footlocker nordstrom	puma nike adidas crocs	target	the home depot costco
automotive, aviation, logistics, manufacturing	porsche honda southwest airlines delta airlines/td>	bmw group toyota		tesla fedex
consumer goods, food & beverage, food, healthcare	unilever	colgate palmolive the coca cola company starbucks	proctor & gamble johnson & johnson	"mcdonald's chipotle hershey company
cryptocurrency, e-commerce, financial services, finance	block	coinbase shopify	amazon jpmorgan chase & co american express	visa mastercard
entertainment, gaming, hospitality, luxury goods, music	ubisoft nintendo roblox	the walt disney company	airbnb hilton marriott lvmh spotify	netflix
social media, technology	pinterest zoominfo philips uber	cisco zoom video communications logitech	amd google apple	salesforce / slack microsoft nvidia adobe

A Simple Random Sample

Pick a **simple random sample** of 15 companies.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________	11. __________
2. __________	7. __________	12. __________

3. __________	8. __________	13. __________
4. __________	9. __________	14. __________
5. __________	10. __________	15. __________

A Systematic Sample

Pick a **systematic sample** of 15 companies.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________	11. __________
2. __________	7. __________	12. __________
3. __________	8. __________	13. __________
4. __________	9. __________	14. __________
5. __________	10. __________	15. __________

A Stratified Sample

Pick a **stratified sample**, by stock price, of 10 companies. Round any decimals to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________
2. __________	7. __________
3. __________	8. __________
4. __________	9. __________
5. __________	10. __________

A Stratified Sample

Pick a **stratified sample**, by industry grouping, of 20 companies. Round any decimals to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

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

A Cluster Sample

Pick a **cluster sample** of companies from two industry groupings. The number of companies will vary.

1. Describe your procedure.
2. Complete the table with your sample.

1. ________	6. ________	11. ________	16. ________	21. ________
2. ________	7. ________	12. ________	17. ________	22. ________

3. ________	8. ________	13. ________	18. ________	23. ________
4. ________	9. ________	14. ________	19. ________	24. ________
5. ________	10. ________	15. ________	20. ________	25. ________

1.6: Chapter Key Terms

Key Term	Definition
Average	also called mean or arithmetic mean; a number that describes the central tendency of the data
Categorical Variable	variables that take on values that are names or labels, also known as qualitative variables
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.
Control Group	a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups
Convenience Sampling	a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.
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)
Discrete Random Variable	a random variable (RV) whose outcomes are counted
Experimental Unit	any individual or object to be measured
Explanatory Variable	the **independent variable** in an experiment; the value controlled by researchers
Frequency	the number of times a value of the data occurs
Informed Consent	Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.
Institutional Review Board	a committee tasked with oversight of research programs that involve human subjects
Lurking Variable	a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable
Mathematical Models	a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

Key Term	Definition
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.
Numerical Variable	variables that take on values that are indicated by numbers, also known as quantitative variable
Observational Study	a study in which the independent variable is not manipulated by the researcher
Parameter	a number that is used to represent a population characteristic and that generally cannot be determined easily
Placebo	an inactive treatment that has no real effect on the explanatory variable
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
Qualitative Data	See Data.
Quantitative Data	See Data.
Random Assignment	the act of organizing experimental units into treatment groups using random methods
Random Sampling	a method of selecting a sample that gives every member of the population an equal chance of being selected.
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 to the total number of outcomes
Representative Sample	a subset of the population that has the same characteristics as the population
Response Variable	the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment
Sample	a subset of the population studied
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.

Key Term	Definition
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.
Statistic	a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.
Statistical Models	a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations.
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.
Survey	a study in which data is collected as reported by individuals.
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.
Treatments	different values or components of the explanatory variable applied in an experiment
Variable	a characteristic of interest for each person or object in a population

---

1.7: Chapter References

Definitions of Statistics, Probability, and Key Terms

- The Data and Story Library, lib.stat.cmu.edu/DASL/Stories...stDummies.html (accessed May 1, 2013).

Data, Sampling, and Variation in Data and Sampling

- Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
- Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
- Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
- Data from www.bookofodds.com/Relationsh...-the-President
- 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).
- "The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
- "Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
- The Data and Story Library, lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).
- LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
- Data from San Jose Mercury News

Experimental Design and Ethics

- "Vitamin E and Health," Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritio...rce/vitamin-e/ (accessed May 1, 2013).
- Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).
- Ankita Mehta. "Daily Dose of Aspiring Helps Reduce Heart Attacks: Study," International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-as...s-study-300443 (accessed May 1, 2013).
- The Data and Story Library, lib.stat.cmu.edu/DASL/Stories...dLearning.html (accessed May 1, 2013).
- M.L. Jacskon et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors," Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).
- "Earthquake Information by Year," U.S. Geological Survey. http://earthquake.usgs.gov/earthquak...archives/year/ (accessed May 1, 2013).
- "Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).
- Data from www.businessweek.com (accessed May 1, 2013).
- Data from www.forbes.com (accessed May 1, 2013).
- "America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).
- U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.
- "April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april...onsumer-report (accessed May 1, 2013).
- Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).
- Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

1.H: Sampling and Data (Homework)

1.2 Definitions of Statistics, Probability, and Key Terms

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

1.

A fitness center is interested in the mean amount of time a client exercises in the center each week.

2.

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

3.

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

4.

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

5.

A politician is interested in the proportion of voters in his district who think he is doing a good job.

6.

A marriage counselor is interested in the proportion of clients she counsels who stay married.

7.

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

8.

A marketing company is interested in the proportion of people who will buy a particular product.

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

9.

What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

10.

Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

11.

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

1.3 Data, Sampling, and Variation in Data and Sampling

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

12.

number of tickets sold to a concert

13.

percent of body fat

14.

favorite baseball team

15.

time in line to buy groceries

16.

number of students enrolled at Evergreen Valley College

17.

most-watched television show

18.

brand of toothpaste

19.

distance to the closest movie theatre

20.

age of executives in Fortune 500 companies

21.

number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

22.

“Number of times per week” is what type of data?

a. qualitative (categorical)
b. quantitative discrete
c. quantitative continuous

23.

“Duration (amount of time)” is what type of data?

a. qualitative (categorical)
b. quantitative discrete
c. quantitative continuous

24.

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.
b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

25.

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

26.

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

27.

List some practical difficulties involved in getting accurate results from a telephone survey.

28.

List some practical difficulties involved in getting accurate results from a mailed survey.

29.

With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

30.

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

31.

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

32.

Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.

e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

33.

A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why? Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

34.

The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative(categorical), quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

35.

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or nonsampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. These researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

36.

Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's*Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in Example $1.H.4$ could explain this connection?

37.

YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

“Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?” (lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).)

As of April 25, 11 people responded to this question. Each participant answered “NO!”

Which of the potential problems with samples discussed in this module could explain this connection?

38.

A scholarly article about response rates begins with the following quote:

“Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research.” (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).)

The Pew Research Center for People and the Press admits:

“The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more.” (Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).)

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

1.S: Sampling and Data (Solutions)

2.

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

4.

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

6.

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

8.

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

10.

a

12.

quantitative discrete, 150

14.

qualitative, Oakland A's

16.

quantitative discrete, 11,234 students

18.

qualitative, Crest

20.

quantitative continuous, 47.3 years

22.

b

24.

a. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

26.

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

28.

Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

30.

b

32.

convenience cluster stratified systematic simple random

34.

a. qualitative(categorical)
b. quantitative discrete
c. quantitative discrete
d. qualitative(categorical)

36.

Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

38.

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

CHAPTER OVERVIEW

2: Descriptive Statistics

2.1: Introduction to Descriptive Statistics

Figure $2.1.1$ When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

2.2: Display Data

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table $2.2.1$ lists the different data values in ascending order and their frequencies.

Data value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

Table $2.2.1$ Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to Table $2.2.1$, there are three students who work 2 hours, five students who work 3 hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

Data value	Frequency	Relative frequency
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Table $2.2.2$ Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of Table $2.2.2$ is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table $2.2.3$.

Data value	Frequency	Relative frequency	Cumulative relative frequency
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55

Data value	Frequency	Relative frequency	Cumulative relative frequency
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Table 2.2.3 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

NOTE

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

Table 2.2.4 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Heights (inches)	Frequency	Relative frequency	Cumulative relative frequency
59.95–61.94	5	$\frac{5}{10} = 0.05$	0.05
61.95–63.94	3	$\frac{3}{100} = 0.03$	0.05 + 0.03 = 0.08
63.95–65.94	15	$\frac{15}{100} = 0.15$	0.08 + 0.15 = 0.23
65.95–67.94	40	$\frac{40}{100} = 0.40$	0.23 + 0.40 = 0.63
67.95–69.94	17	$\frac{17}{100} = 0.17$	0.63 + 0.17 = 0.80
69.95–71.94	12	$\frac{12}{100} = 0.12$	0.80 + 0.12 = 0.92
71.95–73.94	7	$\frac{7}{100} = 0.07$	0.92 + 0.07 = 0.99
73.95–75.94	1	$\frac{1}{100} = 0.01$	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Table 2.2.4 Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.94 inches
- 61.95 to 63.94 inches
- 63.95 to 65.94 inches
- 65.95 to 67.94 inches
- 67.95 to 69.94 inches
- 69.95 to 71.94 inches
- 71.95 to 73.94 inches
- 73.95 to 75.94 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.94 inches, **three** players whose heights fall within the interval 61.95–63.94 inches, **15** players whose heights fall within the interval 63.95–65.94 inches, **40** players whose heights fall within the interval 65.95–67.94 inches, **17** players whose heights fall within the interval 67.95–69.94 inches, **12** players whose heights fall within the interval 69.95–71.94, **seven** players whose heights fall within the interval 71.95–73.94, and **one** player whose heights fall within the interval 73.95–75.94. All heights fall between the endpoints of an interval and not at the endpoints.

Exercise 2.2.1

From Table 2.2.4, find the percentage of heights that are less than 65.95 inches.

Example 2.2.1

From Table 2.2.5, find the percentage of heights that fall between 61.95 and 65.95 inches.

Answer

Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$or 18%.

Example 2.2.2

Use the heights of the 100 male semiprofessional soccer players in Table 2.2.4. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
2. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
3. The percentage of heights that are more than 65.95 inches is: ____.
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Answer

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Exercise 2.2.2

Table 2.2.5 shows the amount, in inches, of annual rainfall in a sample of towns.

Table 2.2.5

Rainfall (inches)	Frequency	Relative frequency	Cumulative relative frequency
2.95–4.96	6	$\frac{6}{50}$ = 0.12	0.12
4.97–6.98	7	$\frac{7}{50}$ = 0.14	0.12 + 0.14 = 0.26
6.99–9.00	15	$\frac{15}{50}$ = 0.30	0.26 + 0.30 = 0.56
9.01–11.02	8	$\frac{8}{50}$ = 0.16	0.56 + 0.16 = 0.72
11.03–13.04	9	$\frac{9}{50}$ = 0.18	0.72 + 0.18 = 0.90
13.05–15.07	5	$\frac{5}{50}$ = 0.10	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

From Table 2.2.5, find the percentage of rainfall that is less than 9.01 inches.

Exercise 2.2.3

From Table 2.2.5, find the percentage of rainfall that is between 6.99 and 13.05 inches.

Exercise 2.2.4

Table 2.2.5 represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Example 2.2.3

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table 2.2.6 was produced:

Table 2.2.6 Frequency of Commuting Distances

Data	Frequency	Relative frequency	Cumulative relative frequency
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute five or seven miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Answer

1. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
3. $\frac{5}{19}$
4. \(\frac{7}{19}, \frac{12}{19}, \frac{7}{19}\)

Example 2.2.4

Table 2.2.7 contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total number of deaths
2000	231

Year	Total number of deaths
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Table $2.2.7$

Answer the following questions.

1. What is the frequency of deaths measured from 2006 through 2009?
2. What percentage of deaths occurred after 2009?
3. What is the relative frequency of deaths that occurred in 2003 or earlier?
4. What is the percentage of deaths that occurred in 2004?
5. What kind of data are the numbers of deaths?
6. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Answer

1. 97,118 (11.8%)
2. 41.6%
3. 67,092/823,356 or 0.081 or 8.1 %
4. 27.8%
5. Quantitative discrete
6. Quantitative continuous

? Exercise $2.2.5$

Table $2.2.8$ contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total number of crashes	Year	Total number of crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435

Year	Total number of crashes	Year	Total number of crashes
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Table 2.2.8

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem 2 and leaf 3. The number 432 has stem 43 and leaf 2. Likewise, the number 5,432 has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.2.5

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

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

Table 2.2.9 Stem-and-Leaf Graph

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% (8/31) were in the 90s or 100, a fairly high number of As.

Exercise 2.2.6

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example 2.2.6

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

NOTE

The leaves are to the right of the decimal.

Answer

The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

Table 2.2.10

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Exercise 2.2.7

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Example 2.2.7

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Table 2.2.11 and Table 2.2.12 show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Answer

Table 2.2.13

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 8 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Table 2.2.11 Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51	Trump	70

Table 2.2.12 Presidential Age at Death

President	Age	President	Age	President	Age

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example 2.2.8, the **x-axis**(horizontal axis) consists of **data values** and the **y-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example 2.2.8

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table 2.2.14 and in Figure 2.2.1.

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Table 2.2.14

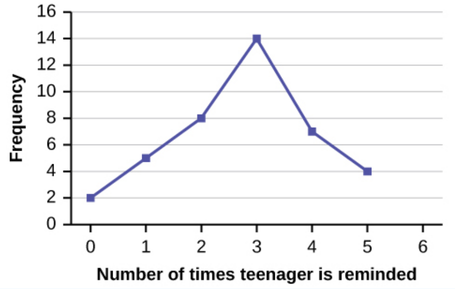


Figure 2.2.1

? Exercise 2.2.8

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table 2.2.15. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Table 2.2.15

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in Example 2.2.9 has age groups represented on the **x-axis** and proportions on the **y-axis**.

✓ Example 2.2.9

By the end of 2011, Facebook had over 146 million users in the United States. Table 2.2.16 shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Table 2.2.16

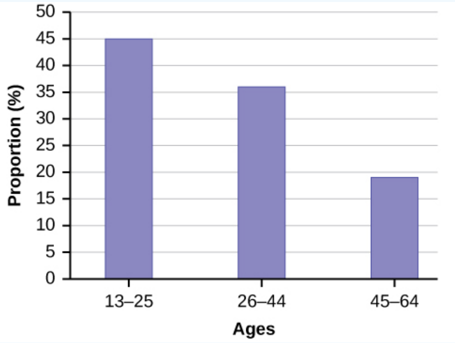


Figure 2.2.2

✓ Exercise 2.2.9

The population in Park City is made up of children, working-age adults, and retirees. Table $2.2.17$ shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Table $2.2.17$

? Example 2.2.10

The columns in Table $2.2.18$ contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/ethnicity	AP examinee population	Overall student population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Table $2.2.18$

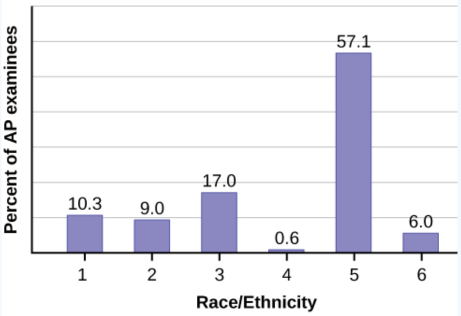


Figure $2.2.3$

? Exercise 2.2.10

Park city is broken down into six voting districts. Table $2.2.19$ shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

Table 2.2.19

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Example 2.2.11

Below is a two-way table showing the types of pets owned by men and women:

Table 2.2.20

	Dogs	**Cats**	**Fish**	**Total**
Men	4	2	2	8
Women	4	6	2	12
Total	8	8	4	20

Given these data, calculate the conditional distributions for the subpopulation of men who own each pet type.

Answer

- Men who own dogs = 4/8 = 0.5
- Men who own cats = 2/8 = 0.25
- Men who own fish = 2/8 = 0.25

Note: The sum of all of the conditional distributions must equal one. In this case, 0.5 + 0.25 + 0.25 = 1; therefore, the solution "checks".

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 term. 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 2.2.21: 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(categorical) 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 2.2.4 and 2.2.5 and determine which graph (pie or bar) you think displays the comparisons better.

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.

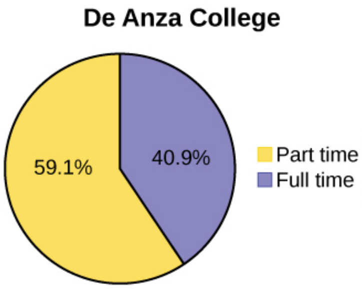


Figure 2.2.4A

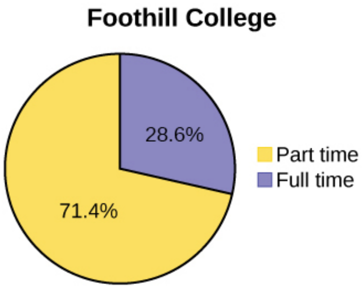


Figure 2.2.4B

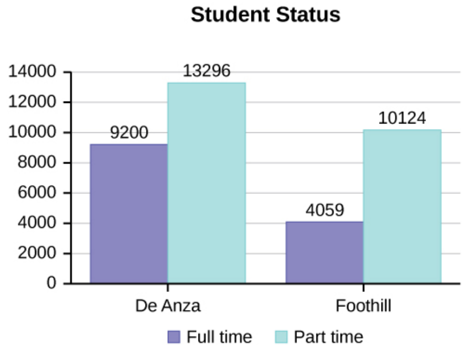


Figure 2.2.5

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 2.2.22: 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%

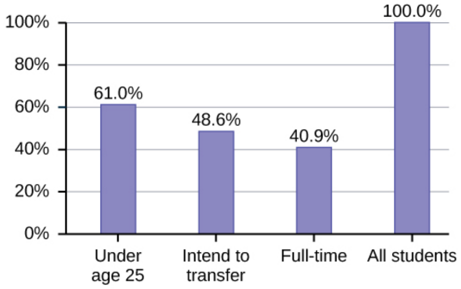


Figure 2.2.6

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

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
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 2.2.7

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 $2.2.9$ is a **Pareto chart**. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

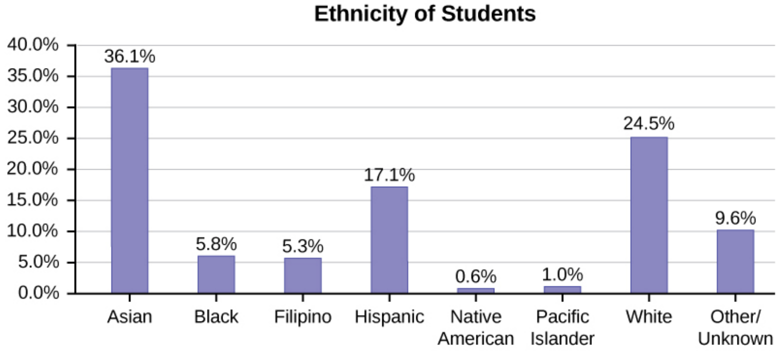


Figure 2.2.8: **Bar Graph with Other/Unknown Category**

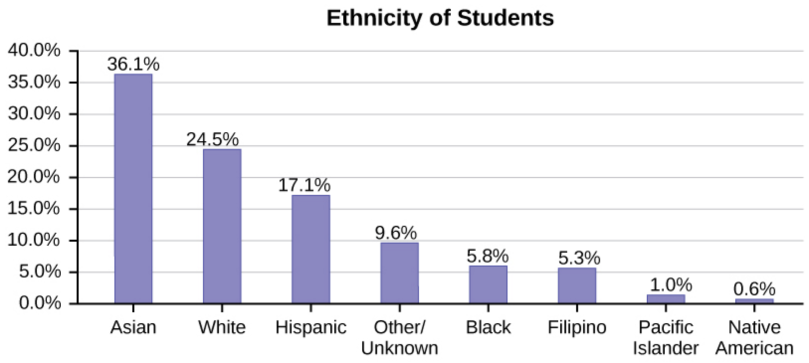


Figure $2.2.9$: **Pareto Chart With Bars Sorted by Size**

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 $2.2.10$.

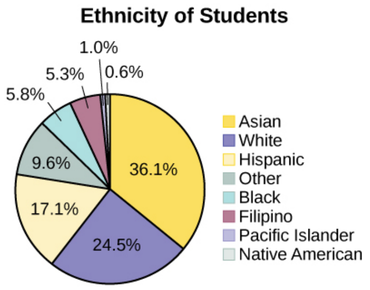


Figure $2.2.10$: Pie Charts with no missing data

Histograms, Frequency Polygons, and Time Series Graphs

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 a value occurs.) If:

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

then:

\[\RF=\frac{f}{n}\nonumber]

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 $RF = \frac{f}{n} = \frac{3}{40} = 0.075$. In other words, 7.5% of the students received 90–100%, and 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 2, 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 2.2.12

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; 66.5; 67; 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.

$$\frac{74.05 - 59.95}{8} = 1.76\backslash\text{non}$$

NOTE

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 59.95–61.94. The heights that are 63.5 are in the interval 61.95–63.94. The heights that are 64 through 64.5 are in the interval 63.95–65.94. The heights 66 through 67.5 are in the interval 65.95–67.94. The heights 68 through 69.5 are in the interval 67.95–69.94. The heights 70 through 71 are in the interval 69.95–71.94. The heights 72 through 73.5 are in the interval 71.95–73.94. The height 74 is in the interval 73.95–75.94.

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

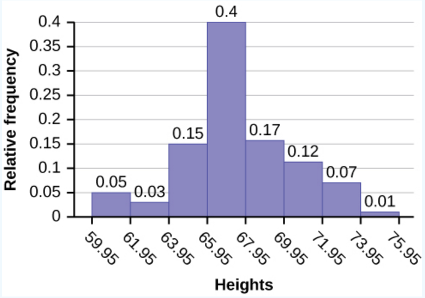


Figure 2.2.11

? Example 2.2.13

Create a histogram for the following data: 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; 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; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 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 _______ .

Solution

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.

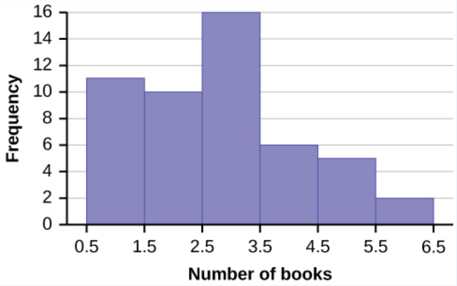


Figure 2.2.12

Example 2.2.14

Using this data set, construct a histogram.

Table 2.2.24

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

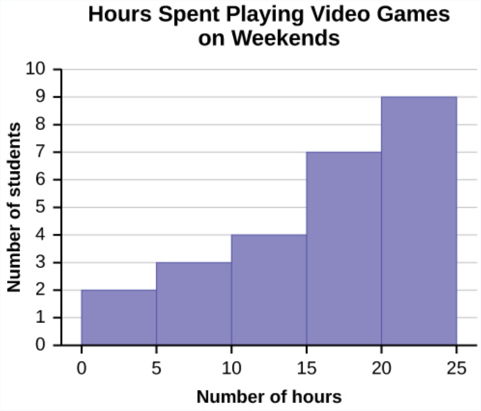


Figure 2.2.13

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.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example 2.2.15

A frequency polygon was constructed from the frequency table below.

Table 2.2.25: Frequency distribution for calculus final test scores

Lower bound	Upper bound	Frequency	Cumulative frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85

Lower bound	Upper bound	Frequency	Cumulative frequency
89.5	99.5	15	100

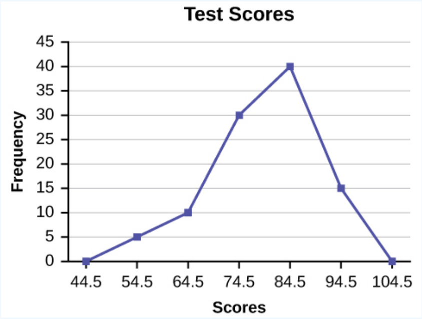


Figure 2.2.7

The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

? Exercise 2.2.11

Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in Table 2.2.26.

Age at inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Table 2.2.26

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

? Example 2.2.16

We will construct an overlay frequency polygon comparing the scores from Example 2.2.15 with the students' final numeric grade.

Table 2.2.27: Frequency distribution for calculus final test scores

Lower bound	Upper bound	Frequency	Cumulative frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45

Lower bound	Upper bound	Frequency	Cumulative frequency
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.2.28: Frequency distribution for calculus final grades

Lower bound	Upper bound	Frequency	Cumulative frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

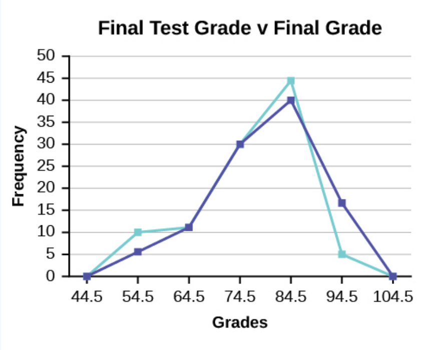


Figure 2.2.8

Constructing a Time Series Graph

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with these data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we do not have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

? Example 2.2.17

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Table 2.2.29

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Table 2.2.30

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Answer

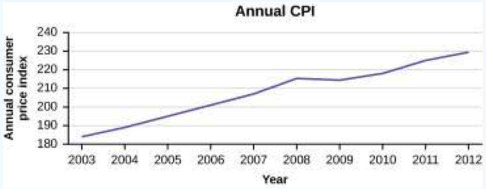


Figure 2.2.9

? Exercise 2.2.18

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO$_2$emissions for the United States.

Table 2.2.20: CO$_2$ emissions

Year	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394

Year	Ukraine	United Kingdom	United States
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

How NOT to Lie with Statistics

It is important to remember that the very reason we develop a variety of methods to present data is to develop insights into the subject of what the observations represent. We want to get a "sense" of the data. Are the observations all very much alike or are they spread across a wide range of values, are they bunched at one end of the spectrum or are they distributed evenly and so on. We are trying to get a visual picture of the numerical data. Shortly we will develop formal mathematical measures of the data, but our visual graphical presentation can say much. It can, unfortunately, also say much that is distracting, confusing and simply wrong in terms of the impression the visual leaves. Many years ago Darrell Huff wrote the book *How to Lie with Statistics*. It has been through 25 plus printings and sold more than one and one-half million copies. His perspective was a harsh one and used many actual examples that were designed to mislead. He wanted to make people aware of such deception, but perhaps more importantly to educate so that others do not make the same errors inadvertently.

Again, the goal is to enlighten with visuals that tell the story of the data. Pie charts have a number of common problems when used to convey the message of the data. Too many pieces of the pie overwhelm the reader. More than perhaps five or six categories ought to give an idea of the relative importance of each piece. This is after all the goal of a pie chart, what subset matters most relative to the others. If there are more components than this then perhaps an alternative approach would be better or perhaps some can be consolidated into an "other" category. Pie charts cannot show changes over time, although we see this attempted all too often. In federal, state, and city finance documents pie charts are often presented to show the components of revenue available to the governing body for appropriation: income tax, sales tax motor vehicle taxes and so on. In and of itself this is interesting information and can be nicely done with a pie chart. The error occurs when two years are set side-by-side. Because the total revenues change year to year, but the size of the pie is fixed, no real information is provided and the relative size of each piece of the pie cannot be meaningfully compared.

Histograms can be very helpful in understanding the data. Properly presented, they can be a quick visual way to present probabilities of different categories by the simple visual of comparing relative areas in each category. Here the error, purposeful or not, is to vary the width of the categories. This of course makes comparison to the other categories impossible. It does embellish the importance of the category with the expanded width because it has a greater area, inappropriately, and thus visually "says" that that category has a higher probability of occurrence.

Time series graphs perhaps are the most abused. A plot of some variable across time should never be presented on axes that change part way across the page either in the vertical or horizontal dimension. Perhaps the time frame is changed from years to months. Perhaps this is to save space or because monthly data was not available for early years. In either case this confounds the presentation and destroys any value of the graph. If this is not done to purposefully confuse the reader, then it certainly is either lazy or sloppy work.

Changing the units of measurement of the axis can smooth out a drop or accentuate one. If you want to show large changes, then measure the variable in small units, penny rather than thousands of dollars. And of course to continue the fraud, be sure that the axis does not begin at zero, zero. If it begins at zero, zero, then it becomes apparent that the axis has been manipulated.

Perhaps you have a client that is concerned with the volatility of the portfolio you manage. An easy way to present the data is to use long time periods on the time series graph. Use months or better, quarters rather than daily or weekly data. If that doesn't get the volatility down then spread the time axis relative to the rate of return or portfolio valuation axis. If you want to show "quick" dramatic growth, then shrink the time axis. Any positive growth will show visually "high" growth rates. Do note that if the growth is negative then this trick will show the portfolio is collapsing at a dramatic rate.

Again, the goal of descriptive statistics is to convey meaningful visuals that tell the story of the data. Purposeful manipulation is fraud and unethical at the worst, but even at its best, making these type of errors will lead to confusion on the part of the analysis.

2.3: Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25^{th} percentile, and the third quartile, Q_3, is the same as the 75^{th} 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 90^{th} 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 75^{th} 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.

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

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

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 7. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.
1; 1; 2; 2; 4; 6; 6.8

The number 2, which is part of the data, is the **first quartile**. One-fourth of the entire set of values are the same as or less than 2 and three-fourths of the values are more than 2.

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 9.

The **third quartile**, Q_3, is 9. Three-fourths (75%) of the ordered data set are less than 9. One-fourth (25%) of the ordered data set are greater than 9. 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).

$IQR=Q_3-Q_1$

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

potential outlier

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 2.3.1

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**.

Example 2.3.2

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

Table 2.3.1

	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.

Example 2.3.3

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

Table 2.3.2

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

Amount of sleep per school night (hours)	Frequency	Relative frequency	Cumulative relative frequency
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 75^{th} 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 $bf75^{th}$ 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 2.3.2

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

Table 2.3.3

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

? Example 2.3.4

Using Table $2.3.22$

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

Answer

Solution 2.17

Using the data from the frequency table, we have:

a. The 80^{th} percentile is between the last eight and the first nine in the table (between the 40^{th} and 41^{st} values). Therefore, we need to take the mean of the 40^{th} an 41^{st} values. The 80^{th} percentile $=\frac{8+9}{2}=8.5$

b. The 90^{th} percentile will be the 45^{th} data value (location is $0.90(50)=45$) and the 45th data value is nine.

c. Q_1 is also the 25th percentile. The 25^{th} percentile location calculation: $P_{25}=0.25(50)=12.5\approx 13$ the 13^{th} data value. Thus, the 25^{th} percentile is six.

A Formula for Finding the kth Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

$k =$ the k^{th} 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 points, or observations

- 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 2.3.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
$18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77$

1. Find the 70^{th} percentile.
2. Find the 83^{rd} percentile.

Answer

1.

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

$i = \frac{k}{100}(n+1) = \left(\frac{70}{100}\right)(29+1) = 21$. 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.

2.

- $k = 83^{rd}$ 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 24^{th} position is 71 and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

Exercise 2.3.3

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.

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 2.3.6

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
$18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77$

1. Find the percentile for 58.
2. Find the percentile for 25.

Answer

1. 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 64^{th} percentile.

2. 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 12^{th} 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.

NOTE

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.

Example 2.3.7

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.)

Example 2.3.8

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th 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 2.3.4

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.

Example 2.3.9

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.

? Example 2.3.10

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.

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.

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

- **2.2: Measures of the Location of the Data** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

2.4: Measures of the Center of the Data

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. Technically this is the arithmetic mean. We will discuss the geometric mean later. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts meaning an equal number of observations on each side. The weight of 25 people are below this weight and 25 people are heavier than this weight. 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. Formally, the arithmetic mean is called the first moment of the distribution by mathematicians. 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”): $\overline{x}$.

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

$$\overline{x}=\frac{1+1+1+2+2+3+4+4+4+4+4}{11}=2.7$$

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

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

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.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

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 $\frac{n+1}{2}=\frac{97+1}{2}=49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2}=\frac{100+1}{2}=50.5$. 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 2.4.1

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:

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

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

$\frac{n+1}{2}=\frac{40+1}{2}=20.5$

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

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

$M=\frac{24+24}{2}=24$

Example 2.4.2

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?

Answer

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

$M=30,000$

(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.

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 2.4.3

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.

? Example 2.4.4

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.

NOTE

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.

Calculating the Arithmetic 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* = $\frac{\text{data sum}}{\text{number of data values}}$ 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 $\frac{\text{lower boundary}+\text{upper boundary}}{2}$. We can now modify the mean definition to be $\textbf{Mean of Frequency Table} = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

? Example 2.4.5

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

Table 2.24

Grade interval	Number of students
50–56.4	1
56.5–62.4	0
62.5–68.4	4
68.5–74.4	4
74.5–80.4	2
80.5–86.4	3
86.5–92.4	4
92.5–98.4	1

Answer

Find the midpoints for all intervals

Table 2.25

Grade interval	Midpoint
50–56.4	53.25
56.5–62.4	59.5
62.5–68.4	65.5
68.5–74.4	71.5
74.5–80.4	77.5
80.5–86.4	83.5
86.5–92.4	89.5
92.5–98.4	95.5

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

Exercise 2.4.1

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:

Table 2.26

Hours teenagers spend on video games	Number of teenagers
0–3.4	3
3.5–7.4	7
7.5–11.4	12
11.5–15.4	7
15.5–19.4	9

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

2.5: Sigma Notation and Calculating the Arithmetic Mean

This unit is here to remind you of material that you once studied and said at the time “I am sure that I will never need this!”

Here are the formulas for a population mean and the sample mean.

Formula for Population Mean

$$\mu=\frac{1}{N}\sum_{i=1}^{N}x_i$$

Formula for Sample Mean

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

The Greek letter μ is the symbol for the population mean and $\bar{x}$ is the symbol for the sample mean. Both formulas have a mathematical symbol that tells us how to make the calculations. It is called Sigma notation because the symbol is the Greek capital letter sigma: Σ. Like all mathematical symbols it tells us what to do: just as the plus sign tells us to add and the $\times$ sign tells us to multiply. These are called mathematical operators. The Σ symbol tells us to add a specific list of numbers.

Let’s say we have a sample of animals from the local animal shelter and we are interested in their average age. If we list each value, or observation, in a column, you can give each one an index number. The first number will be number 1 and the second number 2 and so on.

Table 2.5.1

Animal	**Age**
1	9
2	1
3	8.5
4	10.5
5	10
6	8.5
7	12
8	8
9	1
10	9.5

Each observation represents a particular animal in the sample. Purr is animal number one and is a 9 year old cat, Toto is animal number 2 and is a 1 year old puppy and so on.

To calculate the mean we are told by the formula to add up all these numbers, ages in this case, and then divide the sum by 10, the total number of animals in the sample.

Animal number one, the cat Purr, is designated as x_1, animal number 2, Toto, is designated as x_2 and so on through Dundee who is animal number 10 and is designated as x_{10}.

The i in the formula tells us which of the observations to add together. In this case it is x_1 through x_{10} which is all of them. We know which ones to add by the indexing notation, the $i=1$ and the n or capital N for the population. For this example the indexing notation would be $i=1$ and because it is a sample we use a small n on the top of the Σ which would be 10.

$$\bar{x}=\frac{1}{10}\sum_{i=1}^{10}x_i=\frac{1}{10}(9+1+8.5+10.5+10+8.5+12+8+1+9.5)$$

The sum of the ages is found to be 78 and dividing by 10 gives us the sample mean age as 7.8 years.

The standard deviation requires the same mathematical operator and so it would be helpful to recall this knowledge from your past.

2.6: Geometric Mean

The mean (arithmetic), median and mode are all measures of the “center” of the data, the “average” or "typical" value. They are all in their own way trying to measure the “common” point within the data, that which is “normal”. In the case of the arithmetic mean this is solved by finding the value from which all points are equal linear distances. We can imagine that all the data values are combined through addition and then distributed back to each data point in equal amounts. The sum of all the values is what is redistributed in equal amounts such that the total sum remains the same.

The geometric mean redistributes not the sum of the values but the product of multiplying all the individual values and then redistributing them in equal portions such that the total product remains the same. This can be seen from the formula for the geometric mean, $\tilde{x}$ *(Pronounced x-tilde)x:*

$$\tilde{x}=\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}=\sqrt[n]{x_1 \cdot x_2 \cdots x_n}=\left(x_1 \cdot x_2 \cdots x_n\right)^{\frac{1}{n}}$$

where $\prod$ (capital Greek pi) is another mathematical operator, that tells us to multiply all the x_i numbers in the same way capital Greek sigma tells us to add all the x_i numbers. Remember that a fractional exponent is calling for the nth root of the number, thus an exponent of 1/3 is the cube root of the number.

The geometric mean answers the question, "if all the quantities had the same value, what would that value have to be in order to achieve the same product?” The geometric mean gets its name from the fact that when redistributed in this way the sides form a geometric shape for which all sides have the same length. To see this, take the example of the numbers 10, 51.2 and 8. The geometric mean is the product of multiplying these three numbers together (4,096) and taking the cube root because there are three numbers among which this product is to be distributed. Thus the geometric mean of these three numbers is 16. This describes a cube 16x16x16 and has a volume of 4,096 units.

The geometric mean is relevant in Economics and Finance for dealing with growth: growth of markets, in investment, population and other variables of growth in which there is an interest. Imagine that our box of 4,096 units (perhaps dollars) is the value of an investment after three years and that the investment returns in percents were the three numbers in our example. The geometric mean will provide us with the answer to the question, what is the average rate of return: 16 percent. The arithmetic mean of these three numbers is 23.6 percent. The reason for this difference, 16 versus 23.6, is that the arithmetic mean is additive and thus does not account for the interest on the interest, compound interest, embedded in the investment growth process. The same issue arises when asking for the average rate of growth of a population or sales or market penetration, etc., knowing the annual rates of growth. The formula for the geometric mean rate of return, or any other growth rate, is:

$$r_s=\left(x_1 \cdot x_2 \cdots x_n\right)^{\frac{1}{n}}-1$$

Manipulating the formula for the geometric mean can also provide a calculation of the average rate of growth between two periods knowing only the initial value a_0 and the ending value a_n and the number of periods, n. The following formula provides this information:

$$\left(\frac{a_n}{a_0}\right)^{\frac{1}{n}}=\tilde{x}$$

Finally, we note that the formula for the geometric mean requires that all numbers be positive, greater than zero. The reason of course is that the root of a negative number is undefined for use outside of mathematical theory. There are ways to avoid this problem however. In the case of rates of return and other simple growth problems we can convert the negative values to meaningful positive equivalent values. Imagine that the annual returns for the past three years are +12%, -8%, and +2%. Using the decimal multiplier equivalents of 1.12, 0.92, and 1.02, allows us to compute a geometric mean of 1.0167. Subtracting 1 from this value gives the geometric mean of +1.67% as a net rate of population growth (or financial return). From this example we can see that the geometric mean provides us with this formula for calculating the geometric (mean) rate of return for a series of annual rates of return:

$$r_s=\tilde{x}-1$$

where r_s is average rate of return and $\tilde{x}$ is the geometric mean of the returns during some number of time periods. Note that the length of each time period must be the same.

As a general rule one should convert the percent values to its decimal equivalent multiplier. It is important to recognize that when dealing with percents, the geometric mean of percent values does not equal the geometric mean of the decimal multiplier equivalents and it is the decimal multiplier equivalent geometric mean that is relevant.

2.7: Skewness and the Mean, Median, and Mode

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.

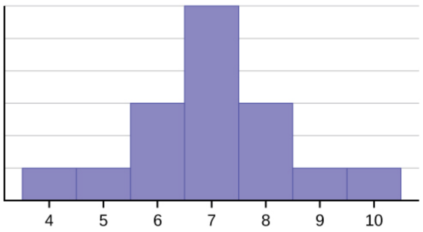


Figure $2.7.1$

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. We can formally measure the skewness of a distribution just as we can mathematically measure the center weight of the data or its general "speadness". The mathematical formula for skewness is:

$$a_3 = \sum \frac{(x_i - \overline{x})^3}{ns^3}.$$

The greater the deviation from zero indicates a greater degree of skewness. If the skewness is negative then the distribution is skewed left as in Figure $2.7.2$.

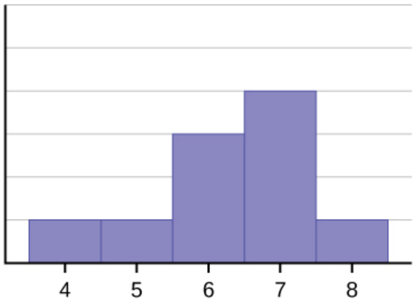


Figure $2.7.2$

The mean is 6.3, the median is 6.5, and the mode is 7. **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 given in Figure $2.7.3$. It is **skewed to the right**.

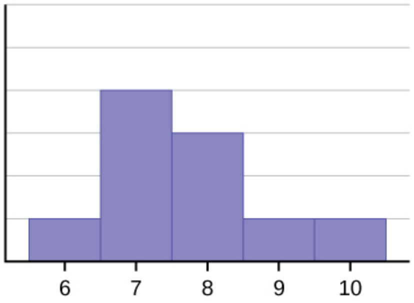


Figure 2.7.3

The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, 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.

As with the mean, median and mode, and as we will see shortly, the variance, there are mathematical formulas that give us precise measures of these characteristics of the distribution of the data. Again looking at the formula for skewness we see that this is a relationship between the mean of the data and the individual observations cubed.

$$a_3 = \sum \frac{(x_i - \overline{x})^3}{ns^3}$$

where s is the sample standard deviation of the data, x_i, and $\overline{x}$ is the arithmetic mean and n is the sample size.

Formally the arithmetic mean is known as the first moment of the distribution. The second moment we will see is the variance, and skewness is the third moment. The variance measures the squared differences of the data from the mean, and skewness measures the cubed differences of the data from the mean. While a variance can never be a negative number, the measure of skewness can and this is how we determine if the data are skewed right of left. The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. The skewness characterizes the degree of asymmetry of a distribution around its mean. While the mean and standard deviation are *dimensional* quantities (this is why we will take the square root of the variance) that is, have the same units as the measured quantities x_i, the skewness is conventionally defined in such a way as to make it *nondimensional*. It is a pure number that characterizes only the shape of the distribution. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive x and a negative value signifies a distribution whose tail extends out towards more negative x. A zero measure of skewness will indicate a symmetrical distribution.

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

- **2.6: Skewness and the Mean, Median, and Mode** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

2.8: Measures of the Spread of the Data

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.

Calculating the Standard Deviation

If x_i is a data value, then the difference "x_i minus the 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 $x_i-\mu$. For sample data, in symbols a deviation is $x_i-\overline{x}$.

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 $x_i-\overline{x}$ values for a sample, or the $x_i-\mu$ 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. Formally, the variance is the second moment of the distribution or the first moment around the mean. Remember that the mean is the first moment of the distribution.

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 $\mathbf{n-1}$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$ or $s=\sqrt{\frac{\Sigma f_i(x_i-\overline{x})^2}{n-1}}$ or $s=\sqrt{\frac{\left(\sum_{i=1}^{n}x_i^2\right)-n\overline{x}^2}{n-1}}$
- For the sample standard deviation, the denominator is $\mathbf{n-1}$, that is the sample size minus 1.

Formulas for the Population Standard Deviation

- $\boldsymbol{\sigma}=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$ or $\sigma=\sqrt{\frac{\Sigma f_i(x_i\mu)^2}{N}}$ or $\sigma=\sqrt{\frac{\sum_{i=1}^{N}x_i^2}{N}-\mu^2}$
- For the population standard deviation, the denominator is N, the number of items in the population.

In these formulas, f_i represents the frequency with which a value appears. For example, if a value appears once, $f_i = 1$. If a value appears three times in the data set or population, $f_i = 3$.

Two important observations concerning the variance and standard deviation: the deviations are measured from the mean and the deviations are squared. In principle, the deviations could be measured from any point, however, our interest is measurement from the center weight of the data, what is the "normal" or most usual value of the observation. Later we will be trying to measure the "unusualness" of an observation or a sample mean and thus we need a measure from the mean. The second observation is that the deviations are squared. This does two things, first it makes the deviations all positive and second it changes the units of measurement from that of the mean and the original observations. If the data are weights then the mean is measured in pounds, but the variance is measured in pounds-squared. One reason to use the standard deviation is to return to the original units of measurement by taking the square root of the variance. Further, when the deviations are squared it explodes their value. For example, a deviation of 10 from the mean when squared is 100, but a deviation of 100 from the mean is 10,000. What this does is place great weight on outliers when calculating the variance.

Types of Variability in Samples

When trying to study a population, a sample is often used, either for convenience or because it is not possible to access the entire population. Variability is the term used to describe the differences that may occur in these outcomes. Common types of variability include the following:

- Observational or measurement variability
- Natural variability
- Induced variability
- Sample variability

Here are some examples to describe each type of variability.

Example 1: Measurement variability

Measurement variability occurs when there are differences in the instruments used to measure or in the people using those instruments. If we are gathering data on how long it takes for a ball to drop from a height by having students measure the time of the drop with a stopwatch, we may experience measurement variability if the two stopwatches used were made by different manufacturers: For example, one stopwatch measures to the nearest second, whereas the other one measures to the nearest tenth of a second. We also may experience measurement variability because two different people are gathering the data. Their reaction times in pressing the button on the stopwatch may differ; thus, the outcomes will vary accordingly. The differences in outcomes may be affected by measurement variability.

Example 2: Natural variability

Natural variability arises from the differences that naturally occur because members of a population differ from each other. For example, if we have two identical corn plants and we expose both plants to the same amount of water and sunlight, they may still grow at different rates simply because they are two different corn plants. The difference in outcomes may be explained by natural variability.

Example 3: Induced variability

Induced variability is the counterpart to natural variability; this occurs because we have artificially induced an element of variation (that, by definition, was not present naturally): For example, we assign people to two different groups to study memory, and we induce a variable in one group by limiting the amount of sleep they get. The difference in outcomes may be affected by induced variability.

Example 4: Sample variability

Sample variability occurs when multiple random samples are taken from the same population. For example, if I conduct four surveys of 50 people randomly selected from a given population, the differences in outcomes may be affected by sample variability.

Example 2.8.1

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;

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

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.

Table 2.8.1

Data	Freq.	Deviations	Deviations^2	(Freq.)(Deviations^2)
x_i	f_i	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$f_i(x_{-}\overline{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)2 = 0.950625$	$3 \times 0.950625 = 2.851875$
				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:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, $s = 0.72$.

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 9. **If you add the deviations, the sum is always zero**. (For Example 2.8.1, 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. By squaring the deviations we are placing an extreme penalty on observations that are far from the mean; these observations get greater weight in the calculations of the variance. We will see later on that the variance (standard deviation) plays the critical role in determining our conclusions in inferential statistics. We can begin now by using the standard deviation as a measure of "unusualness." "How did you do on the test?" "Terrific! Two standard deviations above the mean." This, we will see, is an unusually good exam grade.

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.** This estimate requires us to use an estimate of the population mean rather than the actual population mean. Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

The standard deviation, s or σ, is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the 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.

? Example 2.8.2

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:
   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. IQR

Answer

a. See Table 2.8.2

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. $IQR = 90 - 61 = 29$

Table 2.8.2

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

Data	Frequency	Relative frequency	Cumulative relative frequency
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
100	1	0.032	0.998 (Why isn't this value 1? Answer: Rounding)

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: $\text{Mean of Frequency Table } = \frac{\sum f_i m_i}{\sum f_i}$ where $f_i =$ interval frequencies and m_i = 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 2.8.3

Find the standard deviation for the data in Table 2.8.3.

Table 2.8.3

Class	Frequency, f_i	Midpoint, m_i	f_i\cdot m_	$f_i(m_i - \bar{x})^2$
0–2	1	1	$1 \cdot 1 = 1$	$1(1 - 6.88)^2 = 34.57$
3–5	6	4	$6 \cdot 4 = 24$	$6(4 - 6.88)^2 = 49.77$
6-8	10	7	$10 \cdot 7 = 70$	$10(7 - 6.88)^2 = 0.14$
9-11	7	10	$7 \cdot 10 = 70$	$7(10 - 6.88)^2 = 68.14$
12-14	0	13	$0 \cdot 13 = 0$	$0(13 - 6.88)^2 = 0$
	n = 24		$\boldsymbol{\bar{x} = 16524 = 6.88}$	$\boldsymbol{s^2 = 152.6224 - 1 = 6.64}$

For this data set, we have the mean, $\bar{x} = 6.88$ and the standard deviation, $s_x = 2.58$. This means that a randomly selected data value would be expected to be 2.58 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost three standard deviations from the mean. While the formula for calculating the standard deviation is not complicated,

$$s_x = \sqrt{\frac{\sum f_i (m_i - \bar{x})^2}{n-1}}$$

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.

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

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

Table 2.8.4

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example 2.8.4

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?

Table 2.8.5

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 (#of STDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$z = \#\text{ of STDE Vs} = \frac{\text{value - mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$

For John, $z = \#\text{ ofSTDEV } s = \frac{2.85 \cdot 3.0}{0.7} = -0.21$

For Ali, $z = \#\text{ ofSTDEV } s = \frac{77 - 80}{10} = -0.3$

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.

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 Normal Distribution, which we will examine in great detail later:

- 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.

Coefficient of Variation

Another useful way to compare distributions besides simple comparisons of means or standard deviations is to adjust for differences in the scale of the data being measured. Quite simply, a large variation in data with a large mean is different than the same variation in data with a small mean. To adjust for the scale of the underlying data the Coefficient of Variation (CV) has been developed. Mathematically:

$$CV = \frac{s}{\overline{x}} * 100 \text{ conditioned upon } \overline{x} \neq 0, \text{ where } s \text{ is the standard deviation of the data and } \overline{x} \text{ is the mean}$$

We can see that this measures the variability of the underlying data as a percentage of the mean value; the center weight of the data set. This measure is useful in comparing risk where an adjustment is warranted because of differences in scale of two data sets. In effect, the scale is changed to common scale, percentage differences, and allows direct comparison of the two or more magnitudes of variation of different data sets.

2.9: Descriptive Statistics (Worksheet)

Work in groups on these problems. You should try to answer the questions without referring to your notes. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will organize and construct visual displays of qualitative and quantitative data.
- The student will calculate univariate descriptive statistics.
- The student will analyze graphs and numeric summaries to interpret what the data implies.

The Data

The table contains data from a 2014 African Financial Survey on a random sample of 18 African countries. The four columns provide the following information:

1. **Country**
2. **Region** - which region of the continent the country is in: North, East, South, West, Central
3. **Inflation Rate** - the inflation rate of the country in 2014, reported as a %
4. **No. of Bank Branches** - the number of bank branches in 2014 located in the country

African Financial Survey Data

Country	Region	Inflation Rate	No. Bank Branches
Algeria	North Africa	4.24	1438
Cameroon	Central Africa	2.45	185
Cote d'Ivoire	West Africa	2.86	600
Equatorial Guinea	Central Africa	5.83	15
Ethiopia	East Africa	7.89	2208
Gabon	Central Africa	2.67	58
Ghana	West Africa	9.89	967
Lesotho	Southern Africa	4.94	46
Madagascar	Southern Africa	7.2	233
Mali	West Africa	2.07	516
Mauritius	Southern Africa	3.61	223
Namibia	Southern Africa	5.78	111
Rwanda	East Africa	4.44	161
Seychelles	East Africa	4.17	30
South Africa	Southern Africa	5.68	3697
Swaziland	Southern Africa	5.95	49
Togo	West Africa	2.27	200
Zimbabwe	Southern Africa	3.97	742

Organize the Data

For each of the three variables - Region, Inflation Rate, No. Bank Branches - complete each of the following steps.

1. Identify the variable type: qualitative, quantitative - discrete, quantitative - continuous.

2. Construct a frequency distribution table.
3. Visualize the frequency distribution with either a bar graph, pie chart, or histogram. Be sure to select the most appropriate visual display for the variable type.
4. Describe the shape of the distribution.
5. Are there any potential outliers? List the value(s) that could be outliers. Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the appropriate measure of center for the variable **Region**. Can a measure of variation be computed for this variable
2. For the variable **Inflation Rate**, determine each of the following.
    - measures of center: mean, median, mode
    - measures of variation: range, variance, sd
    - 5-number summary
    - outliers
    - Comment on which measure of center and which measure of variation is most appropriate.
3. Repeat the steps in #2. for the variable **No. Bank Branches**.

Interpret the Data

Write a few sentences to indicate what information is included in the sample. Consider commenting on which region is most represented in the data, what the typical inflation rate is, and what the typical number of bank branches is based on this sample. Add any other insights you noticed in analyzing the data.

2.10: Chapter Key Terms

Key Term	Definition
Cumulative Relative Frequency	The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.
Frequency	the number of times a value of the data occurs
Frequency Table	a data representation in which grouped data is displayed along with the corresponding frequencies
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.
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.
Mean (arithmetic)	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 $\boldsymbol{\mu} = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$
Mean (geometric)	a measure of central tendency that provides a measure of average geometric growth over multiple time periods.
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
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.
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
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.

Key Term	Definition
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 – $\overline{x}$ where x is a value of the data and $\overline{x}$ 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.

2.H: Descriptive Statistics (Homework)

2.2 Display Data

39.

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of courses	Frequency	Relative frequency	Cumulative relative frequency
1	30	0.6	
2	15		
3			

Table1.13 Part-time Student Course Loads

a. Fill in the blanks in Table $2.H.13$.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

40.

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in **Table $2.H.14$**.

# flossing per week	Frequency	Relative frequency	Cumulative relative frequency
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Table1.14 Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in Table $2.H.14$.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

41.

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2;5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 4; 5; 10 .

Table $2.H.15$ was produced.

Table $2.H.15$ Frequency of Immigrant Survey Responses

Data	Frequency	Relative frequency	Cumulative relative frequency
0	2	219219	0.1053
2	3	319319	0.2632
4	1	119119	0.3158
5	3	319319	0.4737
7	2	219219	0.5789

Data	Frequency	Relative frequency	Cumulative relative frequency
10	2	219219	0.6842
12	2	219219	0.7895
15	1	119119	0.8421
20	1	119119	1.0000

a. Fix the errors in Table $2.H.15$. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

42.

How much time does it take to travel to work? **Table $2.H.16$** shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

Table $2.H.16$

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

43.

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. **Table $2.H.17$** shows the ages of the chief executive officers for the first 60 ranked firms.

Table $2.H.17$

Age	Frequency	Relative frequency	Cumulative relative frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?

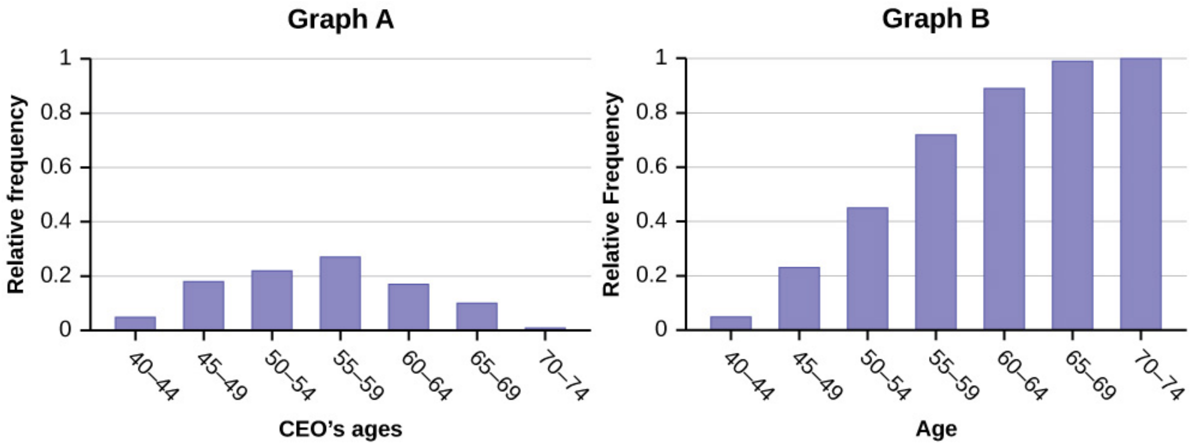


Figure $2.H.11$

Use the following information to answer the next two exercises: Table $2.H.18$ contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of direct hits	Relative frequency	Cumulative frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Table1.18 Frequency of Hurricane Direct Hits

44.

What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

45.

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

84.

Table $2.H.63$ contains the 2010 obesity rates in U.S. states and Washington, DC.

Table $2.H.63$

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."
c. Construct a bar graph for all the states beginning with the letter "M."

85.

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Table 2.H.64 Publisher A

# of books	Freq.	Rel. freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Table 2.H.65 Publisher B

# of books	Freq.	Rel. freq.

# of books	Freq.	Rel. freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.H.66 Publisher C

# of books	Freq.	Rel. freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

a. Find the relative frequencies for each survey. Write them in the charts.
b. Use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

86.

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Table 2.H.67 Singles

Amount($)	Frequency	Rel. frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Table 2.H.68 Couples

Amount($)	Frequency	Rel. frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the x-axis by \$50 widths. Use relative frequency on the y-axis.
c. Construct a histogram for the couples group. Scale the x-axis by \$50 widths. Use relative frequency on the y-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by \$50, scale it by \$100. Use relative frequency on the y-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

87.

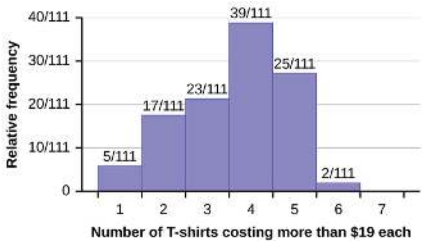
Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

Table 2.H.69

# of movies	Frequency	Relative frequency	Cumulative relative frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than \$19 each.



88.

The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21
b. 59
c. 41
d. Cannot be determined

89.

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

90.

Following are the 2010 obesity rates by U.S. states and Washington, DC.

Table 2.H.70

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.

2.3 Measures of the Location of the Data

91.

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

92.

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in Table 2.71. Also, include left endpoint, but not the right endpoint.

Table $2.H.71$

Salary ($)	Relative frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.
    i. Should all bars have the same width, based on the data? Why or why not?
    ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40th and 80th percentiles
e. Construct a bar graph of the data

2.4 Measures of the Center of the Data

93.

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Table $2.H.72$

Percent of population obese	Number of countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2

Percent of population obese	Number of countries
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

94.

Table $2.H.73$ gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Table 2:73

Percent of underweight children	Number of countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

2.5 Sigma Notation and Calculating the Arithmetic Mean

95.

A sample of 10 prices is chosen from a population of 100 similar items. The values obtained from the sample, and the values for the population, are given in Table $2.H.74$ and Table $2.H.75$ respectively.

a. Is the mean of the sample within $1 of the population mean?
b. What is the difference in the sample and population means?

Table $2.H.74$

Prices of the sample
$21
$23
$21
$24
$22
$22
$25
$21
$20
$24

Table $2.H.75$

Prices of the population	Frequency
$20	20
$21	35
$22	15
$23	10
$24	18
$25	2

96.

A standardized test is given to ten people at the beginning of the school year with the results given in Table $2.H.76$ below. At the end of the year the same people were again tested.

a. What is the average improvement?
b. Does it matter if the means are subtracted, or if the individual values are subtracted?

Table $2.H.76$

Student	Beginning score	Ending score
1	1100	1120
2	980	1030
3	1200	1208
4	998	1000
5	893	948
6	1015	1030
7	1217	1224
8	1232	1245
9	967	988
10	988	997

97.

A small class of 7 students has a mean grade of 82 on a test. If six of the grades are 80, 82,86, 90, 90, and 95, what is the other grade?

98.

A class of 20 students has a mean grade of 80 on a test. Nineteen of the students has a mean grade between 79 and 82, inclusive.

a. What is the lowest possible grade of the other student?
b. What is the highest possible grade of the other student?

99.

If the mean of 20 prices is $10.39, and 5 of the items with a mean of $10.99 are sampled, what is the mean of the other 15 prices?

2.6 Geometric Mean

100.

An investment grows from $10,000 to $22,000 in five years. What is the average rate of return?

101.

An initial investment of $20,000 grows at a rate of 9% for five years. What is its final value?

102.

A culture contains 1,300 bacteria. The bacteria grow to 2,000 in 10 hours. What is the rate at which the bacteria grow per hour to the nearest tenth of a percent?

103.

An investment of $3,000 grows at a rate of 5% for one year, then at a rate of 8% for three years. What is the average rate of return to the nearest hundredth of a percent?

104.

An investment of $10,000 goes down to $9,500 in four years. What is the average return per year to the nearest hundredth of a percent?

2.7 Skewness and the Mean, Median, and Mode

105.

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

2.8 Measures of the Spread of the Data

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- $\text{median } = 1,014$ FTES
- $\sigma = 474$ FTES
- $\text{first quartile } = 528.5$ FTES
- $\text{third quartile } = 1,447.5$ FTES
- $n = 29$ years

106.

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

107.

75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

108.

The population standard deviation = _____

109.

What percent of the FTES were from 528.5 to 1447.5? How do you know?

110.

What is the IQR? What does the IQR represent?

111.

How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Table $2.H.77$

Year	**2005–06**	**2006–07**	**2007–08**	**2008–09**	**2009–10**	**2010–11**
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

112.

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the IQR. Round to one decimal place.

113.

Compare the IQR for the FTES for 1976–77 through 2004–2005 with the IQR for the FTES for 2005-2006 through 2010–2011. Why do you suppose the IQRs are so different?

114.

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Table $2.H.78$

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

115.

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

116.

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

117.

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in Table $2.H.79$.

Table $2.H.79$

Percent of population obese	Number of countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0

Percent of population obese	Number of countries
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How "unusual" is the United States' obesity rate compared to the average rate? Explain.

118.

Table $2.H.80$ gives the percent of children under five considered to be underweight.

Table $2.H.80$

Percent of underweight children	Number of countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

1.1: Homework

119.

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.

Table 1.1.81

	Javier	Ercilia
$\overline{x}$	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?

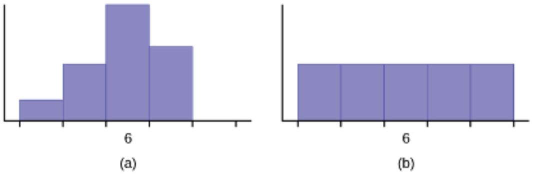


Figure 2.24

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.

Table 1.1.82

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
20	3		
			Total = 20

120.

What is the IQR?

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

121.

What is the mode?

a. 19
b. 19.5
c. 14 and 20

d. 22.65

122.

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

123.

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

Table 1.1.83

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

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

124.

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:

Table 1.1.84

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

a. Find the sample mean $\overline{x}$
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. What percent of the students owned at least five pairs?
i. Find the 40th percentile.
j. Find the 90th percentile.

k. Construct a line graph of the data
l. Construct a stemplot of the data

125.

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. The middle 50% of the weights are from _______ to _______.
f. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
g. 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?
h. Assume the population was the San Francisco 49ers. Find:
    i. the population mean, μ.
    ii. the population standard deviation, $sigma$.
    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?
i. 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?

126.

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.

127.

Refer to **Figure $\mathbf{1.1.25}$** determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

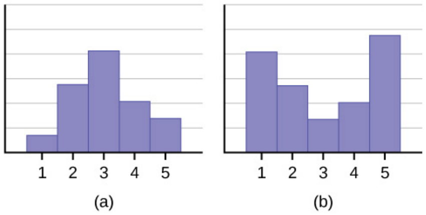


Figure **2.25**

a. The medians for both graphs are the same.
b. We cannot determine if any of the means for both 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 both graphs is different.

128.

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 65th percentile.
d. Find the 10th percentile.
e. The middle 50% of the conferences last from _______ days to _______ days.
f. Calculate the sample mean of days of engineering conferences.
g. Calculate the sample standard deviation of days of engineering conferences.
h. Find the mode.
i. 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.
j. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

129.

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?

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.

Table 1.1.85

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

130.

The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

131.

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

132.

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 1.1.86.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table **2.86**

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?

2.S: Descriptive Statistics (Solutions)

40.

Table 1.19

a.

# flossing per week	Frequency	Relative frequency	Cumulative relative frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

42.

The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

44.

b

1.

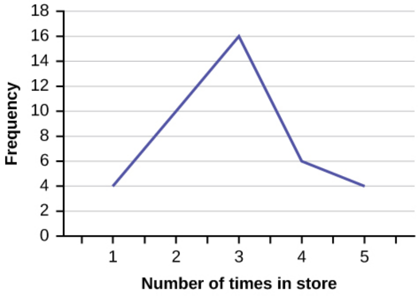


Figure $2.S.26$

3.

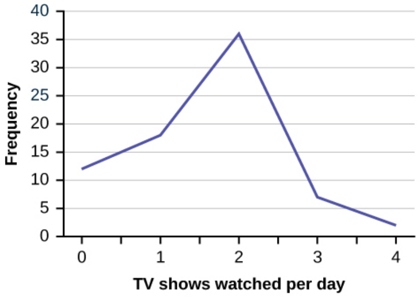


Figure $2.S.27$

5.

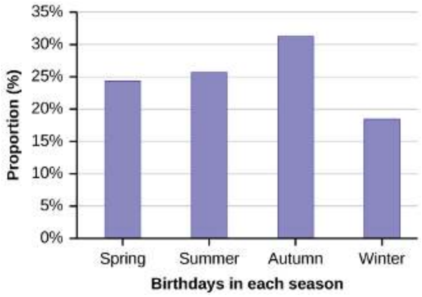


Figure 2.S.28

7.

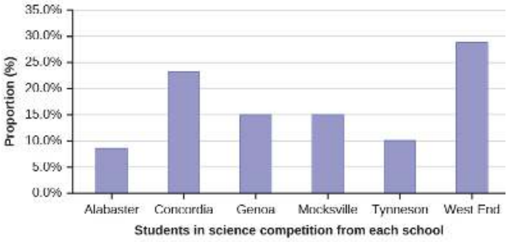


Figure 2.S.29

9.

65

11.

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

13.

Answers will vary. One possible histogram is shown:

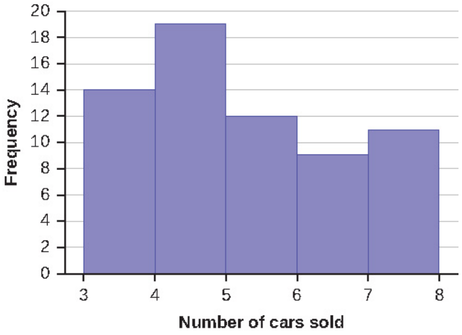


Figure 2.S.30

15.

Find the midpoint for each class. These will be graphed on the x-axis. The frequency values will be graphed on the y-axis values.

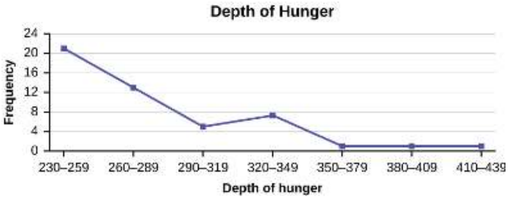


Figure 2.S.31

17.

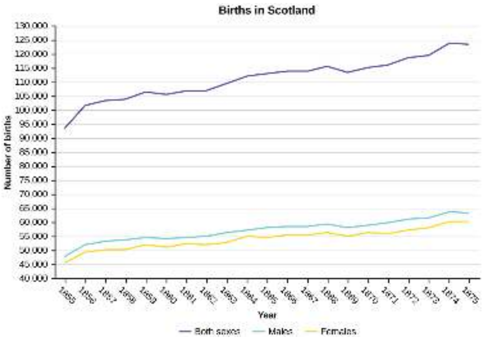


Figure $2.S.32$

19.

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

21.

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{143+0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

23.

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

25.

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

27.

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of \$1700 or less; only 10% had damage repair costs of \$1700 or more.

29.

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost \$240,000 or less. 66% of houses cost \$240,000 or more.

31.

4

33.

$6 – 4 = 2$

35.

6

37.

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

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

39.

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

41.

4

44.

39.48 in.

45.

$21,574

46.

15.98 ounces

47.

81.56

48.

4 hours

49.

2.01 inches

50.

18.25

51.

10

52.

14.15

53.

14

54.

14.78

55.

44%

56.

100%

57.

6%

58.

33%

59.

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.

61.

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.

63.

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

65.

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

67.

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

69.

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

71.

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

73.

The mean and the median are both six.

75.

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

77.

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

79.

$s = 34.5$

81.

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.

83.

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$
b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{38045.3}{101} - 60.94^2} = 7.62$
c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

84.

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.
    - Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)
    - Press MATH
    - Arrow over to PRB
    - Press 5:randInt(
    - Enter 51,1,8)

   Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

   Corresponding percents are $\{30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1\}$

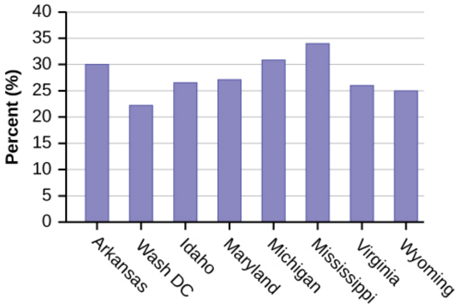


Figure 2.S.33

b.

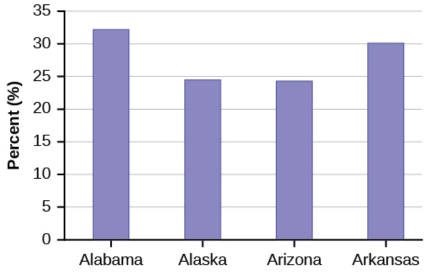


Figure 2.S.34

c.

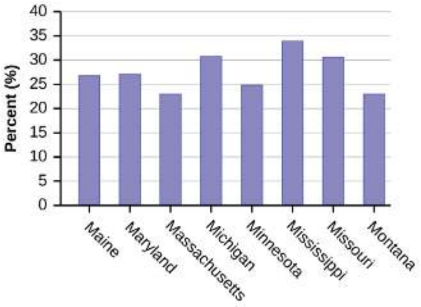


Figure 2.S.35

86.

Amount($)	Frequency	Relative frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Table2.87 Singles

Amount($)	Frequency	Relative frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07

Amount($)	Frequency	Relative frequency
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Table2.88 Couples

a. See Table $2.S.87$ and Table $2.S.88$.
b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

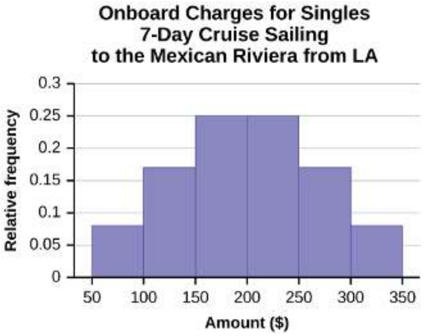


Figure $2.S.36$

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).

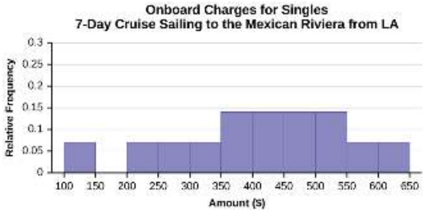


Figure $2.S.37$

d. Compare the two graphs:
    i. Answers may vary. Possible answers include:
        - Both graphs have a single peak.
        - Both graphs use class intervals with width equal to $50.
    ii. Answers may vary. Possible answers include:
        - The couples graph has a class interval with no values.
        - It takes almost twice as many class intervals to display the data for couples.
    iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.
e. Check student's solution.
f. Compare the graph for the Singles with the new graph for the Couples:
    i.
        - Both graphs have a single peak.
        - Both graphs display 6 class intervals.
        - Both graphs show the same general pattern.

    ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.
h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

88.

c

90.

Answers will vary.

92.

a. $1-(0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01)=0.06$
b. $0.19+0.26+0.18=0.63$
c. Check student’s solution.
d. 40th percentile will fall between 30,000 and 40,000

   80th percentile will fall between 50,000 and 75,000

e. Check student’s solution.

94.

The mean percentage, $\bar{x}=\frac{1328.65}{50}=26.75$

95.

a. Yes
b. The sample is 0.5 higher.

96.

a. 20
b. No

97.

51

98.

a. 42
b. 99

99.

$10.19

100.

17%

101.

$30,772.48

102.

4.4%

103.

7.24%

104.

-1.27%

106.

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

108.

474 FTES

110.

919

112.

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- $IQR = 245$

113.

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

115.

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

117.

- $\bar{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that $23.32 + 12.95 = 36.27$is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not hav e an unusually high percentage of obese people.

120.

a

122.

b

123.

a. 1.48
b. 1.12

125.

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. 205.5, 272.5
f. sample
g. population
h.
  i. 236.34
  ii. 37.50
  iii. 161.34
  iv. 0.84 std. dev. below the mean
i. Young

127.

a. True
b. True
c. True
d. False

129.

Table $2.S.89$

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

131.

a

CHAPTER OVERVIEW

3: Probability Topics

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

3.1: Introduction to Probability

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

Figure 3.1 Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

3.2: Probability Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** or **probability** experiment. Flipping one fair coin twice is an example of a probability 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: (1) to list the possible outcomes, (2) to create a tree diagram, or (3) to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S=\{H,T\}$ where $H=$ heads and $T=$ 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 $P(A)$.

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). $P(A)=0$ means the event A can never happen. $P(A)=1$ means the event A always happens. $P(A)=0.5$ 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 $\{HH,TH,HT,TT\}$ where $T=$ tails and $H=$ heads. The sample space has four outcomes. If we consider the event A = getting one head, there are two outcomes that meet this condition $\{HT,TH\}$, so $P(A)=\frac{2}{4}=0.5$.

Suppose you roll one fair six-sided die, with the numbers $\{1,2,3,4,5,6\}$on its faces. Let event $E=$ rolling a number that is at least five. There are two outcomes $\{5,6\}$ that satisfy this condition, so $P(E)=\frac{2}{6}$. 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.

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.

"$\cup$" Event: The Union

An outcome is in the event $A\cup B$ 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 $B=\{4,5,6,7,8\}$, then $A\cup B=\{1,2,3,4,5,6,7,8\}$. Notice that 4 and 5 are NOT listed twice.

"$\cap$" Event: The Intersection

An outcome is in the event $A \cap B$ 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 $A \cap B = \{4,5\}$.

The **complement** of event A is denoted A^C (read "A complement"). A^C consists of all outcomes that are **NOT** in A. Notice that $P(A)+P(A^C)=1$. For example, let $S=\{1,2,3,4,5,6\}$and let $A=\{1,2,3,4\}$. Then, $A^C=\{5,6\}$. Note that $P(A)=\frac{4}{6}$, so $P(A^C)=\frac{2}{6}$, and $P(A)+P(A^C)=\frac{4}{6}+\frac{2}{6}=1$

The **conditional probability** of A given B is written $P(A|B)$. The conditional probability $P(A|B)$ 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 $P(A|B)$ is $P(A|B)=\frac{P(A\cap B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $S=\{1,2,3,4,5,6\}$. Let $A=$ face is 2 or 3 and $B=$ face is even $(2,4,6)$. To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B=\{2,4,6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B)=\frac{\frac{(\text{ the number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{ the number of outcomes that are even in } S)}{6}}=\frac{\frac{1}{6}}{\frac{3}{6}}=\frac{1}{3}$$

Odds

The odds of an event presents the probability as a ratio of success to failure. This is common in various gambling formats. Mathematically, the odds of an event can be defined as:

$$\frac{P(A)}{1-P(A)}$$

where $P(A)$ is the probability of success and of course $1-P(A)$ is the probability of failure. Odds are always quoted as "numerator to denominator," e.g. 2 to 1. Here the probability of winning is twice that of losing; thus, the probability of winning is 0.66. A probability of winning of 0.60 would generate odds in favor of winning of 3 to 2. While the calculation of odds can be useful in gambling venues in determining payoff amounts, it is not helpful for understanding probability or statistical theory.

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

3.3: Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a **sample space**, then: $P(A \cap B) = P(B)P(A|B)$. We can think of the intersection symbol as substituting for the word "and".

This rule may also be written as: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

This equation is read as the probability of A given B equals the probability of A and B divided by the probability of B.

If A and B are **independent**, then $P(A|B) = P(A)$. Then $P(A \cap B) = P(A|B)P(B)$ becomes $P(A \cap B) = P(A)(B)$ because the $P(A|B) = P(A)$ if A and B are independent.

One easy way to remember the multiplication rule is that the word "and" means that the event has to satisfy two conditions. For example the name drawn from the class roster is to be both a female and a Part 2 student. It is harder to satisfy two conditions than only one and of course when we multiply fractions the result is always smaller. This reflects the increasing difficulty of satisfying two conditions.

The Addition Rule

If A and B are defined on a sample space, then: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can think of the union symbol substituting for the word "or". The reason we subtract the intersection of A and B is to keep from double counting elements that are in both A and B.

If A and B are **mutually exclusive**, then $P(A \cap B) = 0$. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes $P(A \cup B) = P(A) + P(B)$.

? Exercise 3.3.1

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

1. Find $P(B^C)$.
2. Find $P(D \cap B)$.
3. Find $P(B|D)$.
4. Find $P(D \cap B^C)$.
5. Find $P(D|B^C)$.

3.4: Contingency Tables and Probability Trees

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example 3.4.1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Table 3.4.1

	Speeding violation in the last year	No speeding violation in the last year	Total
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

a. Find P(Driver is a cell phone user).

Answer

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$

b. Find P(Driver had no violation in the last year).

Answer

b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

c. Find P(Driver had no violation in the last year $\cap$ was a cell phone user).

Answer

c. $\frac{280}{755}$

d. Find P(Driver is a cell phone user $\cup$ driver had no violation in the last year).

Answer

d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$

e. Find P(Driver is a cell phone user $\mid$ driver had a violation in the last year).

Answer

e. $\frac{25}{70}$ (The sample space is reduced to the number of drivers who had a violation.)

f. Find P(Driver had no violation last year | driver was not a cell phone user)

Answer

f. $\frac{405}{450}$ (The sample space is reduced to the number of drivers who were not cell phone users.)

Exercise 3.4.1

Table $3.4.2$ 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

Table 3.4.2

1. What is P(athlete stretches before exercising)?
2. What is P(athlete stretches before exercising||no injury in the last year)?

Example 3.4.2

Table $3.4.3$ shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The coastline	Near lakes and streams	On mountain peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Table 3.4.3 Hiking Area Preference

a. Complete the table.

Answer

a.

Table 3.4.4 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. Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \cap C)$.
2. Find P(F)P(C)

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Answer

b.

1. $P(F \cap C) = \frac{18}{100}$ = 0.18
2. P(F)P(C) = $\left(\frac{45}{100}\right)\left(\frac{34}{100}\right)$ = (0.45)(0.34) = 0.153

$P(F \cap C)$ ≠ P(F)P(C), so the events F and C are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: P(___|___) = ___.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Answer

c.

1. The word 'given' tells you that this is a conditional.
2. $P(M|L) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d. 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(F).
2. Find P(P).
3. Find $P(F \cap P)$.
4. Find $P(F \cup P)$.

Answer

d.

1. P(F) = $\frac{45}{100}$
2. P(P) = $\frac{25}{100}$
3. $P(F \cap P)$ = $\frac{11}{100}$
4. $P(F \cup P)$ = $\frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Exercise 3.4.2

Table $3.4.5$ shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Table 3.4.5

Gender	Lake path	Hilly path	Wooded path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Example 3.4.3

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Table 3.4.6 Door Choice

Caught or not	Door one	Door two	Door three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One} \cap \text{Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One} \cap \text{Not Caught})$

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Answer

a.

Table 3.4.7 Door Choice

Caught or not	Door one	Door two	Door three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	**\(\frac{41}{60}\)**
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. What is the probability that Alissa does not catch Muddy?

Answer

b. $\frac{41}{60}$

c. What is the probability that Muddy chooses $\text{Door One} \cup \text{Door Two}$ given that Muddy is caught by Alissa?

Answer

c. $\frac{1/15+1/12}{19/60} = \frac{9}{19}$

Example 3.4.4

Table $3.4.8$ contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Table 3.4.8 United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	

Year	Robbery	Burglary	Rape	Vehicle	Total
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Note that total data = 4,520.7

1. Find $P(2009 \cap \text{ Robbery})$.
2. Find $P(2010 \cap \text{ Burglary})$.
3. Find $P(2010 \cup \text{ Burglary})$.
4. Find $P(2011 | \text{ Rape})$.
5. Find $P(\text{Vehicle } | 2008)$.

Answer

1. 0.0294
2. 0.1551
3. 0.7165
4. 0.2365
5. 0.2575

Exercise $3.4.3$

Table $3.4.9$ relates the weights and heights of a group of individuals participating in an observational study.

Table 3.4.9

Weight/height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

1. Find the total for each row and column
2. Find the probability that a randomly chosen individual from this group is Tall.
3. Find the probability that a randomly chosen individual from this group is Obese and Tall.
4. Find the probability that a randomly chosen individual from this group is Tall given that the individual is Obese.
5. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
6. Find the probability a randomly chosen individual from this group is Tall and Underweight.
7. Are the events Obese and Tall independent?

Tree Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 3.4.5

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.

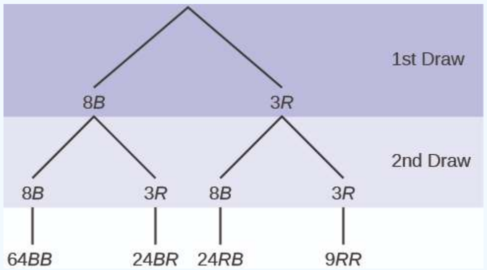


Figure 3.4.1 Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the nine RR outcomes can be written as:

R1R1; R1R2; R1R3; R2R1; R2R2; R2R3; R3R1; R3R2; R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the **sample space**.

a. List the 24 BR outcomes: B1R1, B1R2, B1R3, ...

Answer

a. B1R1; B1R2; B1R3; B2R1; B2R2; B2R3; B3R1; B3R2; B3R3; B4R1; B4R2; B4R3; B5R1; B5R2; B5R3; B6R1; B6R2; B6R3; B7R1; B7R2; B7R3; B8R1; B8R2; B8R3

b. Using the tree diagram, calculate $P(RR)$.

Answer

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. Using the tree diagram, calculate $P(RB \cup BR)$.

Answer

c. $P(RB \cup BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. Using the tree diagram, calculate $P(R \text{ on 1st draw } \cap B \text{ on 2nd draw})$.

Answer

d. $P(R \text{ on 1st draw } \cap B \text{ on 2nd draw}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. Using the tree diagram, calculate $P(R \text{ on 2nd draw} | B \text{ on 1st draw})$.

Answer

e. $P(R \text{ on 2nd draw} | B \text{ on 1st draw}) = P(R \text{ on 2nd} | B \text{ on 1st}) = \backslash(\frac{24}{88} = \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88}=\frac{3}{11}$.

f. Using the tree diagram, calculate $P(BB)$.

Answer

f. $P(BB)=\frac{64}{121}$

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw} | R \text{ on the first draw})$.

Answer

g. $P(B \text{ on 2nd draw} | R \text{ on 1st draw})=\frac{8}{11}$

There are 9 + 24 outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Exercise 3.4.3

In a standard deck, there are 52 cards, 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate the probability that both are face cards.

Figure 3.4.2

Example 3.4.5

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right)=\frac{6}{110}$.

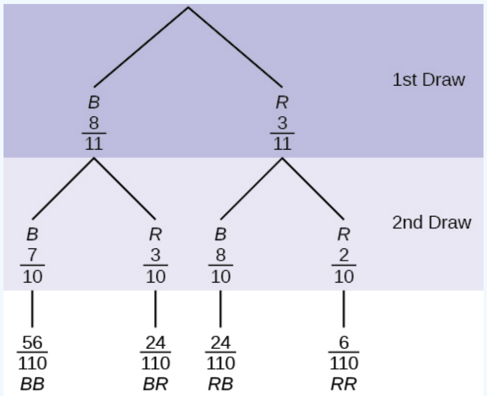


Figure 3.4.3 Total = $\frac{56+24+24+6}{110}=\frac{110}{110}=1$

NOTE

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without replacement**, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

a. P(RR) = ________

Answer

a. P(RR) = $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right)=\frac{6}{110}$

b. Fill in the blanks:

$P(RB \cup BR)=\left(\frac{3}{11}\right)\left(\frac{8}{10}\right)$ + (___)(___) = $\frac{48}{110}$

Answer

b. $P(RB \cup BR)=\left(\frac{3}{11}\right)\left(\frac{8}{10}\right)+\left(\frac{8}{11}\right)\left(\frac{3}{10}\right)=\frac{48}{110}$

c. P(R on 2nd|B on 1st) =

Answer

c. P(R on 2nd|B on 1st) = $\frac{3}{10}$

d. Fill in the blanks.

$P(R \text{ on 1st } \cap B \text{ on 2nd})$ = (___)(___) = $\frac{24}{100}$

Answer

d. $P(R \text{ on 1st } \cap B \text{ on 2nd})=\left(\frac{3}{11}\right)\left(\frac{8}{10}\right)=\frac{24}{110}$

e. Find P(BB).

Answer

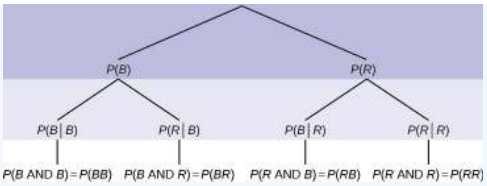
e. P(BB) = $\left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

f. Find P(B on 2nd|R on 1st).

Answer

f. Using the tree diagram, P(B on 2nd|R on 1st) = P(R|B) = $\frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.



- P(R|R) here means P(R on 2nd|R on 1st)
- P(B|R) here means P(B on 2nd|R on 1st)
- P(R|B) here means P(R on 2nd|B on 1st)
- P(B|B) here means P(B on 2nd|B on 1st)

Exercise 3.4.4

In a standard deck, there are 52 cards. Twelve cards are face cards (*F*) and 40 cards are not face cards (*N*). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

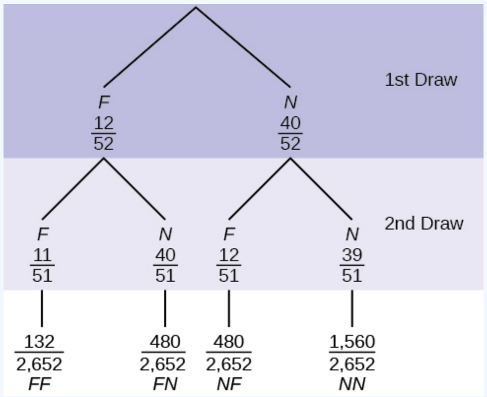
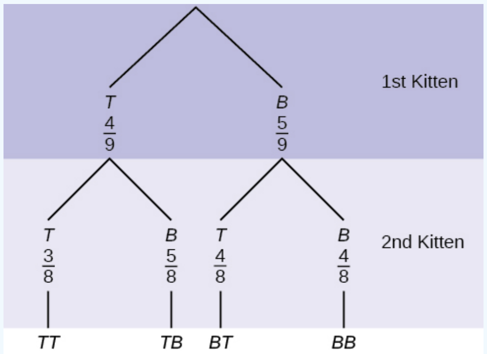


Figure 3.4.4

1. Find $P(FN \cup NF)$.
2. Find P(N|F).
3. Find P(at most one face card).
   Hint: "At most one face card" means zero or one face card.
4. Find P(at least on face card).
   Hint: "At least one face card" means one or two face cards.

Example 3.4.6

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



1. What is the probability that both kittens are tabby?

   a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
2. What is the probability that one kitten of each coloring is selected?

   a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$
3. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
4. What is the probability of choosing two kittens of the same color?

Answer

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

? Exercise 3.4.5

Suppose there are four red balls and three yellow balls in a box. Two balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

3.5: Venn Diagrams

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events. Venn diagrams also help us to convert common English words into mathematical terms that help add precision.

Venn diagrams are named for their inventor, John Venn, a mathematics professor at Cambridge and an Anglican minister. His main work was conducted during the late 1870's and gave rise to a whole branch of mathematics and a new way to approach issues of logic. We will develop the probability rules just covered using this powerful way to demonstrate the probability postulates including the Addition Rule, Multiplication Rule, Complement Rule, Independence, and Conditional Probability.

Example $3.5.1$

Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event $A=\{1,2,3,4,5,6\}$ and event $B=\{6,7,8,9\}$. Then A intersect $B=A\cap B=\{6\}$ and A union $B=A\cup B=\{1,2,3,4,5,6,7,8,9\}$. The Venn diagram is as follows:

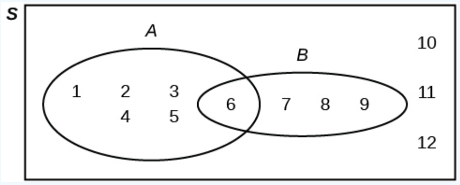


Figure $3.5.1$

Figure $3.5.1$ shows the most basic relationship among these numbers. First, the numbers are in groups called sets; set A and set B. Some number are in both sets; we say in set A ∩ in set B. The English word "and" means inclusive, meaning having the characteristics of both A and B, or in this case, being a part of both A and B. This condition is called the INTERSECTION of the two sets. All members that are part of both sets constitute the intersection of the two sets. The intersection is written as $A\cap B$ where ∩ is the mathematical symbol for intersection. The statement $A\cap B$ is read as "A intersect B." You can remember this by thinking of the intersection of two streets.

There are also those numbers that form a group that, for membership, the number must be in either one or the other group. The number does not have to be in BOTH groups, but instead only in either one of the two. These numbers are called the UNION of the two sets and in this case they are the numbers 1-5 (from A exclusively), 7-9 (from set B exclusively) and also 6, which is in both sets A and B. The symbol for the UNION is ∪, thus $A\cup B=$ numbers 1-9, but excludes number 10, 11, and 12. The values 10, 11, and 12 are part of the universe, but are not in either of the two sets.

Translating the English word "AND" into the mathematical logic symbol ∩, intersection, and the word "OR" into the mathematical symbol ∪, union, provides a very precise way to discuss the issues of probability and logic. The general terminology for the three areas of the Venn diagram in Figure $3.5.1$ is shown in Figure $3.5.2$, below.

Exercise $3.5.1$

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event C = {green, blue, purple} and event P = {red, yellow, blue}. Then $C\cap P=\{blue\}$ and $C\cup P=\{$ green, blue, purple, red, yellow $\}$. Draw a Venn diagram representing this situation.

Example $3.5.2$

Flip two fair coins. Let A = tails on the first coin. Let B = tails on the second coin. Then $A=TT,TH$ and $B=TT,HT$. Therefore, $A\cap B=\{TT\}$. $A\cup B=\{TH,TT,HT\}$.

The sample space when you flip two fair coins is $X=HH,HT,TH,TT$. The outcome HH is in NEITHER A NOR B. The Venn diagram is as follows:

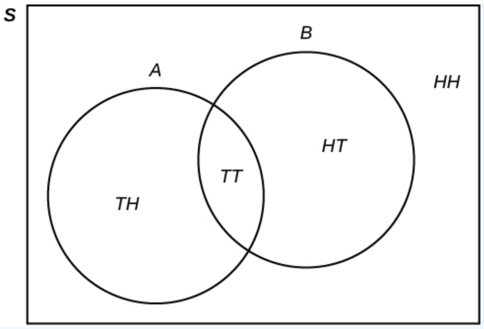


Figure 3.5.2

Exercise 3.5.2

Roll a fair, six-sided die. Let A = a prime number of dots is rolled. Let B = an odd number of dots is rolled. Then $A = 2, 3, 5$ and $B = 1, 3, 5$. Therefore, $A \cap B = \{3, 5\}$. $A \cup B = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $S = 1, 2, 3, 4, 5, 6$. Draw a Venn diagram representing this situation.

Example 3.5.3

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5−10% of African Americans have the Rh- factor, and 51% have type O blood.

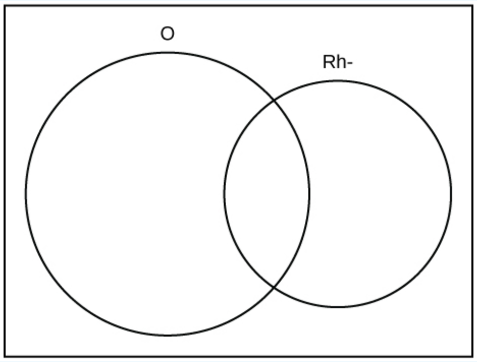


Figure 3.5.3

The “O” circle represents the African Americans with type O blood. The “Rh-“ oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let O = African American with Type O blood and R = African American with Rh- factor.

1. $P(O)$ = ____________
2. $P(R)$ = ____________
3. $P(O \cap R) =$ ____________
4. $P(O \cup R) =$ _____________
5. In the Venn Diagram, describe the overlapping area using a complete sentence.
6. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Answer

1. 0.51; 2. 0.075; 3. 0.04; 4. 0.545; 5. The area represents the African Americans that have type O blood and the Rh- factor.
6. The area represents the African Americans that have neither type O blood nor the Rh- factor.

Example $3.5.4$

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let W = works a second job and S = spouse also works.

Answer

Forty percent of the students at a local college belong to a club and **50%** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let C = student belongs to a club and PT = student works part time.

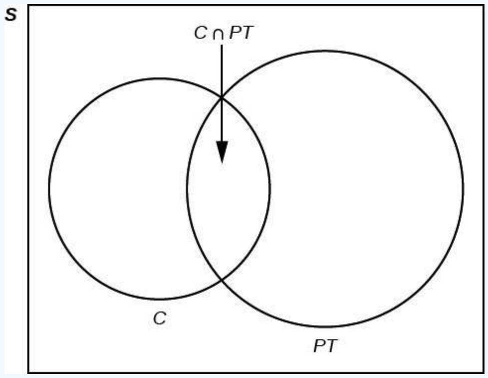


Figure $3.5.4$

If a student is selected at random, find

- the probability that the student belongs to a club. $P(C)=0.40$
- the probability that the student works part time. $P(PT)=0.50$
- the probability that the student belongs to a club AND works part time. $P(C\cap PT)=0.05$
- the probability that the student belongs to a club given that the student works part time. $P(C|PT)=\frac{P(C\cap PT)}{P(PT)}=\frac{0.05}{0.50}=0.1$
- the probability that the student belongs to a club OR works part time. $P(C\cup PT)=P(C)+P(PT)-P(C\cap PT)=0.40+0.50-0.05=0.85$

In order to solve Example $3.5.4$ we had to draw upon the concept of conditional probability from the previous section. There we used tree diagrams to track the changes in the probabilities, because the sample space changed as we drew without replacement. In short, conditional probability is the chance that something will happen given that some other event has already happened. Put another way, the probability that something will happen conditioned upon the situation that something else is also true. In Example $3.5.4$ the probability $P(C|PT)$ is the conditional probability that the randomly drawn student is a member of the club, conditioned upon the fact that the student also is working part time. This allows us to see the relationship between Venn diagrams and the probability postulates.

Exercise $3.5.3$

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

1. Draw a Venn diagram representing the situation.
2. Find the probability that the customer buys either a novel or a non-fiction book.
3. In the Venn diagram, describe the overlapping area using a complete sentence.
4. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Example 3.5.5

A set of 20 German Shepherd dogs is observed. 12 are male, 8 are female, 10 have some brown coloring, and 5 have some white sections of fur. Answer the following using Venn Diagrams.

Draw a Venn diagram simply showing the sets of male and female dogs.

Answer

The Venn diagram below demonstrates the situation of mutually exclusive events where the outcomes are independent events. If a dog cannot be both male and female, then there is no intersection. Being male precludes being female and being female precludes being male: in this case, the characteristic gender is therefore mutually exclusive. A Venn diagram shows this as two sets with no intersection. The intersection is said to be the null set using the mathematical symbol $\emptyset$.

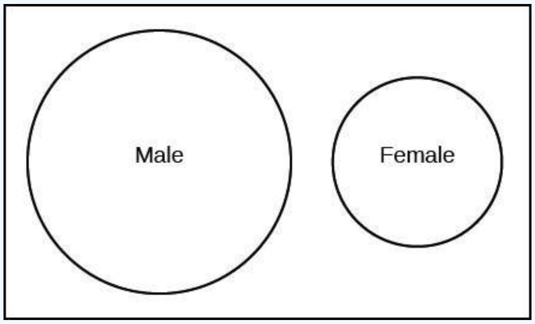


Figure 3.5.5

Draw a second Venn diagram illustrating that 10 of the male dogs have brown coloring.

Answer

The Venn diagram below shows the overlap between male and brown where the number 10 is placed in it. This represents $\text{Male} \cap \text{Brown}$: both male and brown. This is the intersection of these two characteristics. To get the union of Male and Brown, then it is simply the two circled areas minus the overlap. In proper terms, $\text{Male} \cup \text{Brown} = \text{Male} + \text{Brown} - \text{Male} \cap \text{Brown}$ will give us the number of dogs in the union of these two sets. If we did not subtract the intersection, we would have double counted some of the dogs.

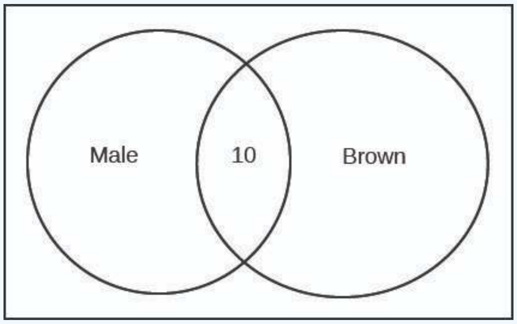


Figure 3.5.6

Now draw a situation depicting a scenario in which the non-shaded region represents "No white fur and female," or $\text{White Fur}^C \cap \text{Female}$. The ^{C} above "fur" indicates "not white fur." The ^{C} above a set means not in that set.

Answer

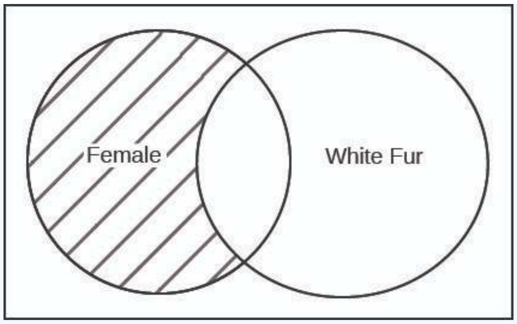


Figure ❓ 3.5.7

The Addition Rule of Probability

We met the addition rule earlier but without the help of Venn diagrams. Venn diagrams help visualize the counting process that is inherent in the calculation of probability. To restate the Addition Rule of Probability:

$$P(A \cup B) = P(\mathrm{A}) + P(B) - P(A \cap B)$$

Remember that probability is simply the proportion of the objects we are interested in relative to the total number of objects. This is why we can see the usefulness of the Venn diagrams. Example $3.5.5$ shows how we can use Venn diagrams to count the number of dogs in the union of brown and male by reminding us to subtract the intersection of brown and male. We can see the effect of this directly on probabilities in the addition rule.

❓ Example $3.5.6$

Let's sample 50 students who are in a statistics class. 20 are freshmen and 30 are sophomores. 15 students get a "B" in the course, and 5 students both get a "B" and are freshmen.

Find the probability of selecting a student who either earns a "B" OR is a freshmen. We are translating the word OR to the mathematical symbol for the addition rule, which is the union of the two sets.

Answer

We know that there are 50 students in our sample, so we know the denominator of our fraction to give us probability. We need only to find the number of students that meet the characteristics we are interested in, i.e. any freshman and any student who earned a grade of "B." With the Addition Rule of probability, we can skip directly to probabilities.

Let "A" = the number of freshmen, and let "B" = the grade of "B." Below we can see the process for using Venn diagrams to solve this.

The $P(A) = \frac{20}{50} = 0.40, P(B) = \frac{15}{50} = 0.30,$ and $P(A \cap B) = \frac{5}{50} = 0.10$

Therefore, $P(A \cap B) = 0.40 + 0.30 - 0.10 = 0.60$

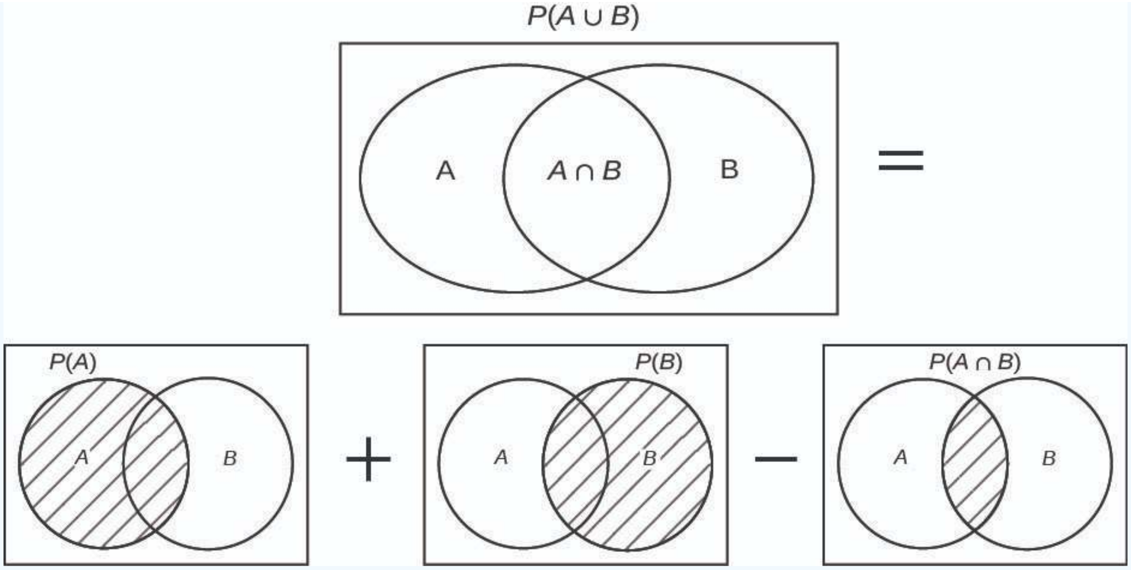


Figure ? 3.5.7

If two events are mutually exclusive, then, like the example where we diagram the male and female dogs, the addition rule is simplified to just $P(A \cup B)=P(A)+P(B)-0$. This is true because, as we saw earlier, the union of mutually exclusive events is the null set, ∅. The diagrams below demonstrate this.

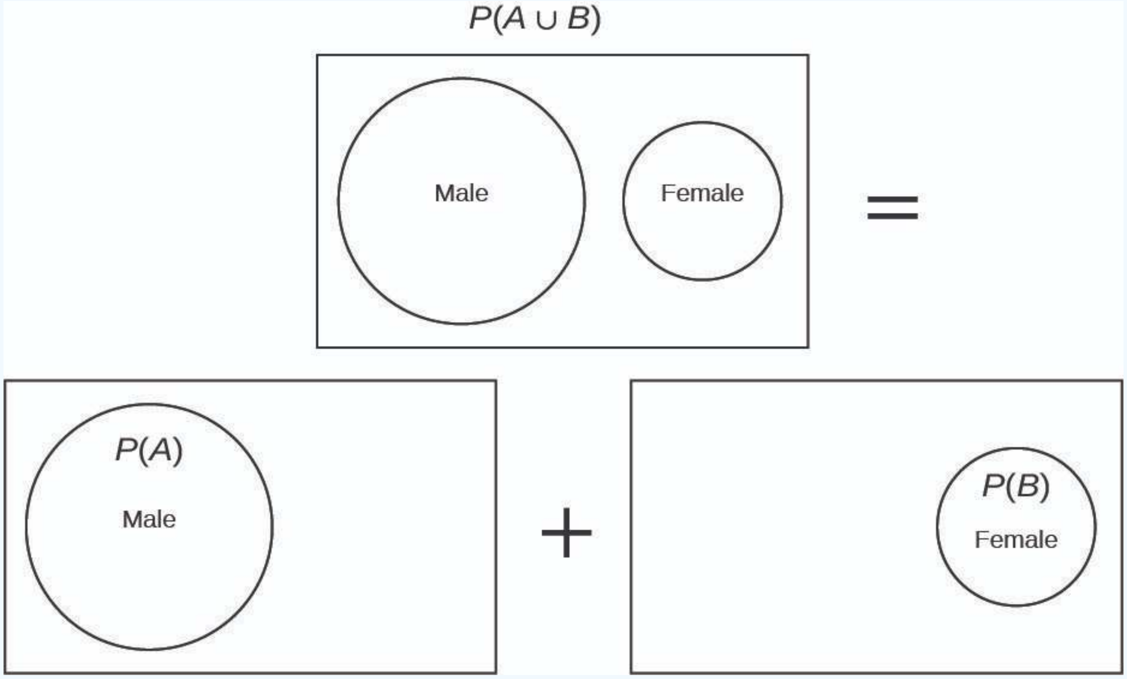


Figure ? 3.5.8

The Multiplication Rule of Probability

Restating the Multiplication Rule of Probability using the notation of Venn diagrams, we have:

$$P(A \cap B)=P(A|B) \cdot P(B)$$

The multiplication rule can be modified with a bit of algebra into the following conditional rule. Then Venn diagrams can then be used to demonstrate the process.

The conditional rule: $P(A|B)=\frac{P(A \cap B)}{P(B)}$

Using the same facts from Example $3.5.6$ above, find the probability that someone will earn a "B" if they are a "freshman."

$$P(A|B) = \frac{0.10}{0.30} = \frac{1}{3}$$

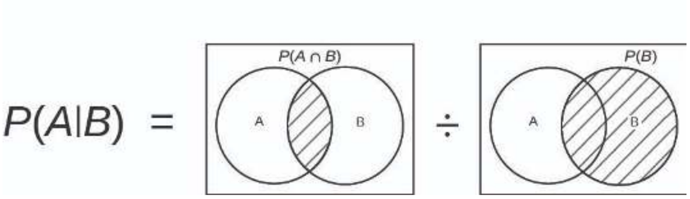


Figure 3.5.9

The multiplication rule must also be altered if the two events are independent. Independent events are defined as a situation where the conditional probability is simply the probability of the event of interest. Formally, independence of events is defined as $P(A|B) = P(A)$ or $P(B|A) = P(B)$. When flipping coins, the outcome of the second flip is independent of the outcome of the first flip; coins do not have memory. The Multiplication Rule of Probability for independent events thus becomes:

$$P(A \cap B) = P(A) \cdot P(B)$$

One easy way to remember this is to consider what we mean by the word "and." We see that the Multiplication Rule has translated the word "and" to the Venn notation for intersection. Therefore, the outcome must meet the two conditions of freshmen and grade of "B" in the above example. It is harder, less probable, to meet two conditions than just one or some other one. We can attempt to see the logic of the Multiplication Rule of probability due to the fact that fractions multiplied times each other become smaller.

The development of the Rules of Probability with the use of Venn diagrams can be shown to help as we wish to calculate probabilities from data arranged in a contingency table.

? Example 3.5.7

Table ? $3.5.1$ is from a sample of 200 people who were asked how much education they completed. The columns represent the highest education they completed, and the rows separate the individuals by male and female.

Table ? $3.5.1$

	Less than high school grad	**High school grad**	**Some college**	**College grad**	**Total**
Male	5	15	40	60	120
Female	8	12	30	30	80
Total	13	27	70	90	200

Now, we can use this table to answer probability questions. The following examples are designed to help understand the format above while connecting the knowledge to both Venn diagrams and the probability rules.

What is the probability that a selected person both finished college and is female?

Answer

This is a simple task of finding the value where the two characteristics intersect on the table, and then applying the postulate of probability, which states that the probability of an event is the proportion of outcomes that match the event in which we are interested as a proportion of all total possible outcomes.

$$P(\text{College Grad} \cap \text{Female}) = \frac{30}{200} = 0.15$$

What is the probability of selecting either a female or someone who finished college?

Answer

This task involves the use of the addition rule to solve for this probability.

$$P(\text{ College Grad } \cup \text{ Female }) = P(F) + P(CG) - P(F \cap CG)$$

$$P(\text{ College Grad } \cup \text{ Female }) = \frac{80}{200} + \frac{90}{200} - \frac{30}{200} = \frac{140}{200} = 0.70$$

What is the probability of selecting a high school graduate if we only select from the group of males?

Answer

Here we must use the conditional probability rule (the modified multiplication rule) to solve for this probability.

$$P(\text{HS Grad} \mid \text{Male }) \quad = \frac{P(\text{HS Grad} \cap \text{Male})}{\text{P(Male)}} = \frac{\left(\frac{15}{200}\right)}{\left(\frac{120}{200}\right)} = \frac{15}{120} = 0.125$$

Can we conclude that the level of education attained by these 200 people is independent of the gender of the person?

Answer

There are two ways to approach this test. The first method seeks to test if the intersection of two events equals the product of the events separately remembering that if two events are independent than $P(A)^* P(B) = P(A \cap B)$. For simplicity's sake, we can use calculated values from above.

Does $P(\text{ College Grad } \cap \text{ Female }) = P(CG) \cdot P(F)$?

$\frac{30}{200} \neq \frac{90}{200} \cdot \frac{80}{200}$ because 0.15 ≠ 0.18.

Therefore, gender and education here are **not** independent.

The second method is to test if the conditional probability of A given B is equal to the probability of A. Again for simplicity, we can use an already calculated value from above.

Does $P(HS\text{ Grad} \mid \text{ Male }) = P(HS\text{ Grad})$?

$\frac{15}{120} \neq \frac{27}{200}$ because 0.125 ≠ 0.135.

Therefore, again gender and education here are **not** independent.

3.6: Conditional Probability and Bayes' Rule

In many situations, additional information about the result of a probability experiment is known (or at least assumed to be known) and given that information the probability of some other event is desired. For this scenario, we compute what is referred to as *conditional probability*.

Definition 3.6.1

For events A and B, with $P(B) > 0$, the ***conditional probability*** of A given B, denoted $P(A \mid B)$, is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

In computing a conditional probability we *assume* that we know the outcome of the experiment is in event B and then, given that additional information, we calculate the probability that the outcome is also in event A. This is useful in practice given that partial information about the outcome of an experiment is often known, as the next example demonstrates.

Example 3.6.1

Continuing in the context of Example 1.2.1, where we considered tossing a fair coin twice, define D to be the event that at least one tails is recorded:

$$D = \{ht, th, tt\}$$

Let's calculate the conditional probability of A given D, i.e., the probability that at least one heads is recorded (event A) assuming that at least one tails is recorded (event D). Recalling that outcomes in this sample space are equally likely, we apply the definition of conditional probability (Definition 2.1.1) and find

$$P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{P(\{ht, th\})}{P(\{ht, th, tt\})} = \frac{(2/4)}{(3/4)} = \frac{2}{3} \approx 0.67.$$

Note that in Example 1.2.1 we found the ***un***-conditional probability of A to be $P(A) = 0.75$. So, knowing that at least one tails was recorded, i.e., assuming event D occurred, the conditional probability of A given D decreased. This is because, if event D occurs, then the outcome hh in A cannot occur, thereby decreasing the chances that event A occurs.

Exercise 3.6.1

Suppose we randomly draw a card from a standard deck of 52 playing cards.

a. If we know that the card is a King, what is the probability that the card is a club?
b. If we instead know that the card is black, what is the probability that the card is a club?

Answer

In order to compute the necessary probabilities, first note that the sample space is given by the set of cards in a standard deck of playing cards. So the number of outcomes in the sample space is 52. Next, note that the outcomes are equally likely, since we are *randomly* drawing the card from the deck.

For part (a), we are looking for the conditional probability that the randomly selected card is club, given that it is a King. If we let C denote the event that the card is a club and K the event that it is a King, then we are looking to compute

$$P(C \mid K) = \frac{P(C \cap K)}{P(K)}. \tag{3.6.1}$$

To compute these probabilities, we count the number of outcomes in the following events:

$$\#\text{ of outcomes in } C = \#\text{ of clubs in standard deck } = 13$$

$$\#\text{ of outcomes in } K = \#\text{ of Kings in standard deck } = 4$$

$$\#\text{ of outcomes in } C\cap K = \#\text{ of King of clubs in standard deck } = 1$$

The probabilities in Equation 3.6.1are then given by dividing the counts of outcomes in each event by the total number of outcomes in the sample space (by the boxed Equation 1.3.2 in Section 1.3):

$$P(C \mid K) = \frac{P(C\cap K)}{P(K)} = \frac{(1/52)}{(4/52)} = \frac{1}{4} = 0.25.$$

For part (b), we are looking for the conditional probability that the randomly selected card is club, given that it is instead black. If we let B denote the event that the card is black, then we are looking to compute

$$P(C \mid B) = \frac{P(C\cap B)}{P(B)}. \tag{3.6.2}$$

To compute these probabilities, we count the number of outcomes in the following events:

$$\#\text{ of outcomes in } B = \#\text{ of black cards in standard deck } = 26$$

$$\#\text{ of outcomes in } C\cap B = \#\text{ of black clubs in standard deck } = 13$$

The probabilities in Equation 3.6.2are then given by dividing the counts of outcomes in each event by the total number of outcomes in the sample space:

$$P(C \mid B) = \frac{P(C\cap B)}{P(B)} = \frac{(13/52)}{(26/52)} = \frac{13}{26} = 0.5.$$

Remark: Exercise $3.6.1$ demonstrates the following fact. For sample spaces with equally likely outcomes, conditional probabilities are calculated using

$$\boxed{P(A \mid B) = \frac{\text{number of outcomes in } A\cap B}{\text{number of outcomes in } B}.} \tag{3.6.3}$$

In other words, if we know that the outcome of the probability experiment is in the event B, then we restrict our focus to the outcomes in that event that are also in A. We can think of this as event B taking the place of the sample space, since we know the outcome must lie in that event.

Properties of Conditional Probability

As with unconditional probability, we also have some useful properties for conditional probabilities. The first property below, referred to as the *Multiplication Law*, is simply a rearrangement of the probabilities used to define conditional probability. The Multiplication Law provides a way for computing the probability of an intersection of events when the conditional probabilities are known.

Multiplication Law

$$P(A\cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

The next two properties are useful when a *partition* of the sample space exists, where a partition is a way of dividing up the outcomes in the sample space into non-overlapping sets. A partition is formally defined in the *Law of Total Probability* below. In many cases, when a partition exists, it is easy to compute the conditional probability of an event in the sample space given an event in the partition. The Law of Total Probability then provides a way of using those conditional probabilities of an event, given the partition to compute the unconditional probability of the event. Following the Law of Total Probability, we state *Bayes' Rule*, which is really just an application of the Multiplication Law. Bayes' Rule is used to calculate what are informally referred to as "reverse conditional probabilities", which are the conditional probabilities of an event in a partition of the sample space, given any other event.

Law of Total Probability

Suppose events $B_1, B_2, \ldots, B_n$, satisfy the following:

1. $\Omega = B_1 \cup B_2 \cup \cdots \cup B_n$
2. $B_i \cap B_j = \varnothing$, for every $i \neq j$
3. $P(B_i) > 0$, for $i = 1, \ldots, n$

We say that the events $B_1, B_2, \ldots, B_n$, **partition** the sample space Ω. Then for any event A, we can write

$$P(A) = P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_n)P(B_n).$$

Bayes' Rule

Let $B_1, B_2, \ldots, B_n$, partition the sample space Ω and let A be an event with $P(A) > 0$. Then, for $j = 1, \ldots, n$, we have

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{P(A)}.$$

A common application of the Law of Total Probability and Bayes' Rule is in the context of medical diagnostic testing.

? Exercise 3.6.1

Consider a test that can diagnose kidney cancer. The test correctly detects when a patient has cancer 90% of the time. Also, if a person does not have cancer, the test correctly indicates so 99.9% of the time. Finally, suppose it is known that 1 in every 10,000 individuals has kidney cancer. We find the probability that a patient has kidney cancer, given that the test indicates she does.

First, note that we are finding a conditional probability. If we let A denote the event that the patient tests positive for cancer, and we let B_1 denote the event that the patient actually has cancer, then we want

$$P(B_1 \mid A).$$

If we let $B_2 = B_1^c$, then we have a partition of all patients (which is the sample space) given by B_1 and B_2.

In the first paragraph of this example, we are given the following probabilities:

$$\text{test correctly detects cancer 90\% of time:} \quad P(A \mid B_1) = 0.9$$

$$\text{test correctly detects no cancer 99.9\% of time:} \quad P(A^c \mid B_2) = 0.999 \Rightarrow P(A \mid B_2) = 1 - P(A^c \mid B_2) = 0.001$$

$$\text{1 in every 10,000 individuals has cancer:} \quad P(B_1) = 0.0001 \Rightarrow P(B_2) = 1 - P(B_1) = 0.9999$$

Since we have a partition of the sample space, we apply the Law of Total Probability to find $P(A)$:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) = (0.9)(0.0001) + (0.001)(0.9999) = 0.0010899$$

Next, we apply Bayes' Rule to find the desired conditional probability:

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A)} = \frac{(0.9)(0.0001)}{0.0010899} \approx 0.08$$

This implies that only about 8% of patients that test positive under this particular test actually have kidney cancer, which is not very good.

3.7: Independent Events

In this section we consider a property of events that relates to conditional probability, namely *independence*. First, we define what it means for a pair of events to be independent, and then we consider collections of more than two events.

Independence for Pairs of Events

The following definition provides an intuitive definition of the concept of independence for two events, and then we look at an example that provides a computational way for determining when events are independent.

Definition 3.7.1

Events A and B are ***independent*** if knowing that one occurs does not affect the probability that the other occurs, i.e., \[P(A|B) = P(A) \quad\text{and}\quad P(B|A) = P(B). \label{indep}]

Using the definition of conditional probability (Definition 2.2.1), we can derive an alternate way to the Equations ??? for determining when two events are independent, as the following example demonstrates.

Example 3.7.1

Suppose that events A and B are independent. We rewrite Equations ??? using the definition of conditional probability:

\[P(A|B) = P(A) \quad \Rightarrow\quad \frac{P(A\cap B)}{P(B)} = P(A) \notag]

and

\[P(B|A) = P(B) \quad \Rightarrow\quad \frac{P(A\cap B)}{P(A)} = P(B) \notag]

In each of the expressions on the right-hand side above we isolate $P(A\cap B)$:

\[\frac{P(A\cap B)}{P(B)} = P(A) \quad \Rightarrow\quad P(A\cap B) = P(A)P(B) \notag]

and

\[\frac{P(A\cap B)}{P(A)} = P(B) \quad \Rightarrow\quad P(A\cap B) = P(A)P(B) \notag]

Both expressions result in $P(A\cap B) = P(A)P(B)$. Thus, we have shown that if events A and B are independent, then the probability of their intersection is equal to the product of their individual probabilities. We state this fact in the next definition.

Definition 3.7.2

Events A and B are ***independent*** if $P(A\cap B) = P(A)P(B)$.

Generally speaking, Definition 2.3.2 tends to be an easier condition than Definition 2.3.1 to verify when checking whether two events are independent.

Example 3.7.1

Consider the context of Exercise 2.2.1, where we randomly draw a card from a standard deck of 52 and C denotes the event of drawing a club, K the event of drawing a King, and B the event of drawing a black card.

Are C and K independent events? Recall that $P(C\cap K) = 1/52$, and note that $P(C) = 13/52$ and $P(K) = 4/52$. Thus, we have

\[P(C\cap K) = \frac{1}{52} = P(C)P(K) = \frac{13}{52}\times\frac{4}{52},\notag]

indicating that C and K **are independent**.

Are C and B independent events? Recall that $P(C\cap B) = 13/52$, and note that $P(B) = 26/52$. Thus, we have

\[P(C\cap B) = \frac{13}{52} \neq P(C)P(B) = \frac{13}{52}\times\frac{26}{52},\notag]

indicating that C and B **are *not* independent**.

Let's think about the results of this example intuitively. To say that C and K are independent means that knowing that one of the events occurs does not affect the probability of the other event occurring. In other words, knowing that the card drawn is a King does not influence the probability of the card being a club. The proportion of clubs in the entire deck of 52 is the same as the proportion of clubs in just the collection of Kings: $1/4$. On the other hand, C and B are not independent (AKA *dependent*) because knowing that the card drawn is club indicates that the card *must be black*, i.e., the probability that the card is black is 1. Alternately, knowing that the card drawn is black increases the probability that the card is a club, since the proportion of clubs in the entire deck is $1/4$, but the proportion of clubs in the collection of black cards is $1/2$.

Independence for 3 or More Events

For collections of 3 or more events, there are two different types of independence.

Definition: (\PageIndex{3}\)

Let $A_1, A_2, \ldots, A_k$, where $k \geq 3$, be a collection of events.

1. The events are ***pairwise independent*** if every pair of events in the collection is independent.
2. The events are ***mutually independent*** if every sub-collection of events, say $A_{i_1}, A_{i_2}, \ldots, A_{i_n}$, satisfy the following:

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}) = P(A_{i_1}) \times P(A_{i_2}) \times \ldots \times P(A_{i_n})$$

Mutually independent is a stronger type of independence, since it *implies* pairwise independent. But pairwise independence does NOT imply mutual independence, as the following example will demonstrate.

Example 3.7.1

Consider again the context of Example 1.1.1, i.e., tossing a fair coin twice, and define the following events:

$$A = \text{first toss is heads}$$

$$B = \text{second toss is heads}$$

$$C = \text{exactly one head is recorded}$$

We show that this collection of events - A, B, C - is pairwise independent, but NOT mutually independent. First, we note that the individual probabilities of each event are 0.5:

$$P(A) = P(\{hh, ht\}) = 0.5$$

$$P(B) = P(\{hh, th\}) = 0.5$$

$$P(C) = P(\{ht, th\}) = 0.5$$

Next, we look at the probabilities of all pairwise intersections to establish pairwise independence:

$$P(A \cap B) = P(hh) = 0.25 = P(A)P(B)$$

$$P(A \cap C) = P(ht) = 0.25 = P(A)P(C)$$

$$P(B \cap C) = P(th) = 0.25 = P(B)P(C)$$

> However, note that the three events do not have any outcomes in common, i.e., $A \cap B \cap C = \varnothing$. Thus, we have $P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$, and so the events are not mutually independent.

3.8: Discrete Distribution Experiment (Worksheet)

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if an everyday experiment fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

The Experiment and Random Variable

The experimental procedure is to select a playing card from a full set of denominations, i.e., Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, and King. The procedure is repeated three times and the number of Kings picked is recorded.

1. Shuffle the set of cards.
2. Pick one card from the set at random.
3. Record whether it was a King or not a King.
4. Put the card back and reshuffle the set.
5. Do this a total of three times.
6. Record the number of Kings picked.

We consider the **random variable *X* = number of Kings picked.**

Empirical Distribution of *X*

1. Record the number of Kings picked by the class in a table like the one below and calculate the relative frequencies.
2. Using the table as the empirical probability distribution of *X*, calculate the **mean** and **standard deviation** for *X*.
3. Construct a **histogram** for *X* using the empirical distribution.
4. Calculate the following **probabilities** using the empirical distribution.
   a. the probability of picking exactly 3 Kings: $P(X=3)$
   b. the probability of picking between 0 and 3 Kings, not inclusive: $P(0<X<3)$
   c. the probability of picking at least 2 Kings: $P(X \geq 2)$

Empirical Distribution of *X*

x	Frequency	Relative Frequency $\approx p(x)$
0		
1		
2		
3		

Theoretical Distribution of *X*

The distribution of *X* can be modeled theoretically with the **binomial distribution**. The binomial distribution applies when

- we have n identical trials - each with two possible outcomes (called success or failure),
- the probability of success, p, is the same for each trial, and
- the trials are independent.

The pdf of a binomial distribution is given by the following formula:

$$p(x)=\binom{n}{x} p^{x}(1-p)^{n-x},$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}. \tag{3.8.1}$$

Remember $n! = n(n-1)\cdots 2\cdot 1$. For example, $3! = 3\cdot 2\cdot 1 = 6$. Note that $0! = 1$.

1. **Explain** why the random variable X = the number of Kings picked can be modeled using a binomial distribution. Identify the values of n, the number of trials, and p, the probability of "success" for X.
2. Present the theoretical PDF in a table like the one below.
3. Using the theoretical distribution of X, calculate the **mean** and **standard deviation** for X.
4. Construct a **histogram** for X using the theoretical distribution.
5. Calculate the following **probabilities** using the theoretical distribution.
    a. the probability of picking exactly 3 Kings: $P(X = 3)$
    b. the probability of picking between 0 and 3 Kings, not inclusive: $P(0 < X < 3)$
    c. the probability of picking at least 2 Kings: $P(X \geq 2)$

Theoretical Distribution of X

x	$p(x)$
0	
1	
2	
3	

Reflection Questions

For questions 1 and 2 below, think about the shapes of the two graphs, the probabilities you calculated, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions.
3. Using your answers from questions 1 and 2, does it appear that the data fit the theoretical distribution? In complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Would you expect the Empirical Distribution or the Theoretical Distribution to change, and how would it change? Why? Why wouldn't the other distribution change?

3.9: Chapter Reference

3.2 Terminology

“Countries List by Continent.” Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

3.3 Independent and Mutually Exclusive Events

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/te...workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

3.4 Two Basic Rules of Probability

DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at www.field.com/fieldpollonline...rs/Rls2443.pdf (accessed May 2, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011..._suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at http://en.Wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at www.census.gov/hhes/socdemo/l...acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

3.5 Contingency Tables and Probability Trees

“Blood Types.” American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

“Human Blood Types.” Unite Blood Services, 2011. Available online at https://www.vitalant.org/Donate/Blood-Donation/Donate-Blood-Overview.aspx (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_st...ive-blood.html (accessed May 2, 2013).

“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).

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at https://ropercenter.cornell.edu/?s=S...h+for+Datasets (accessed February 6, 2019).

3.10: Chapter Key Terms

Key Term	Definition
Conditional Probability	the likelihood that an event will occur given that another event has already occurred
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.
Dependent Events	If two events are NOT independent, then we say that they are dependent.
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
Independent Events	The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true: 1. $P(A\|B)=P(A)$ 2. $P(B\|A)=P(B)$ 3. $P(A\cap B)=P(A)P(B)$
Mutually Exclusive	Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A\cap B)=0$.
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: • $0\leq P(A)\leq 1$ • If A and B are any two mutually exclusive events, then $P(A\cup B)=P(A)+P(B)$. • $P(S)=1$
Sample Space	the set of all possible outcomes of an experiment
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 Complement Event	The complement of event A consists of all outcomes that are NOT in A.
The Conditional Probability of $A\|B$	P(A||B) is the probability that event A will occur given that the event B has already occurred.
The Intersection: the ∩ Event	An outcome is in the event |(A \cap B\) if the outcome is in both $A\cap B$ at the same time.
The Union: the ∪ Event	An outcome is in the event $A\cup B$ if the outcome is in A or is in B or is in both A and B.
Tree Diagram	the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)
Venn Diagram	the visual representation of a sample space and events in the form of circles or ovals showing their intersections

3.H: Probability (Homework)

3.2 Terminology

72. The graph in **Figure $3.H.17$** displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

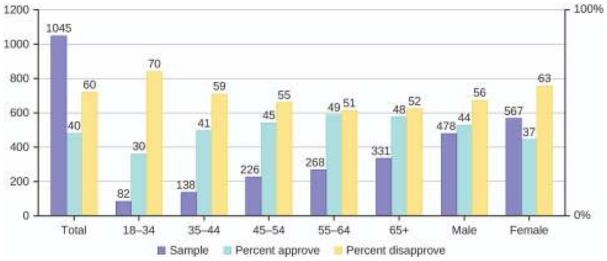


Figure $3.H.17$

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find P(Approve|Male).
i. Out of the age groups, what percent are more than 44 years old?
j. Find P(Approve|Age < 35).

73. Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

3.3 Independent and Mutually Exclusive Events

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.

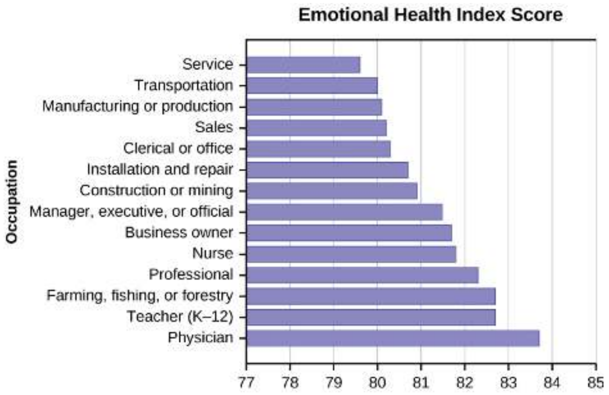


Figure $3.H.18$

74. Find the probability that an Emotional Health Index Score is 82.7.

75. Find the probability that an Emotional Health Index Score is 81.0.

76. Find the probability that an Emotional Health Index Score is more than 81?

77. Find the probability that an Emotional Health Index Score is between 80.5 and 82?

78. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

79. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

80. What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

81. What occupation has the highest emotional index score?

82. What occupation has the lowest emotional index score?

83. What is the range of the data?

84. Compute the average EHIS.

85. If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

3.4 Two Basic Rules of Probability

86. On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- C = California registered voters who support same-sex marriage.
- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18 to 39 years old.

a. Find $P(C)$.
b. Find $P(B)$.
c. Find $P(C|A)$.
d. Find $P(B|C)$.
e. In words, what is $C|A$?
f. In words, what is $B|C$?
g. Find $P(C \cap B)$.
h. In words, what is $C \cap B$?
i. Find $P(C \cup B)$.
j. Are C and B mutually exclusive events? Show why or why not.

87. After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.

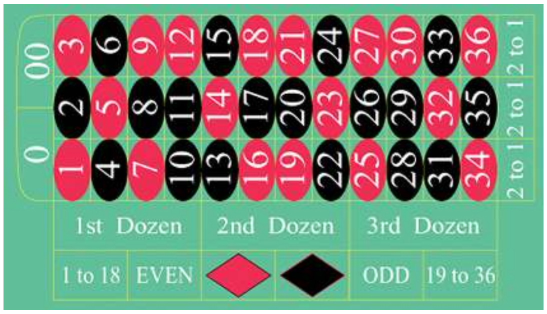


Figure $3.H.19$ (credit: film8ker/wikibooks)

88.

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find P(red).
c. You bet on -1st 12- (1st Dozen). Find P(-1st 12-).
d. You bet on an even number. Find P(even number).
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

89. Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

90. Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

91. Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

  a. List the sample space.
  b. $P(G) =$ _____
  c. $P(G|E) =$ _____
  d. $P(G \cap E) =$ _____
  e. $P(G \cup E) =$ _____
  f. Are G and E mutually exclusive? Justify your answer numerically.

92. Roll two fair dice separately. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.

d. In words, explain what “$P(A|B)$” represents. Find $P(A|B)$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

93. A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find P(A).
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

94. An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find P(A).
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

95. An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find P(A).
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

96. Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Let $P(C|D) = 0.6$.

a. Find $P(C \cap D)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(C \cup D)$.
e. Find $P(D|C)$.

97. Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(Z)$ if $P(Y \cup Z) = 0.71$ and $P(Y) = 0.42$.

98. G and H are mutually exclusive events. $P(G) = 0.5 P(H) = 0.3$

a. Explain why the following statement MUST be false: $P(H|G) = 0.4$.
b. Find $P(H \cup G)$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

99. Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E′ = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Table $3.H.14$

Probability Statements	Answers

Probability Statements	Answers
a. $P(E') =$	i. 0.8043
b. $P(E) =$	ii. 0.623
c. $P(S \cap E') =$	iii. 0.1957
d. $P(S\|E') =$	iv. 0.1219

100. 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

101. Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.
c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

102. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Table $3.H.15$

Name	Single	Double	Triple	Home run	Total hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because P(hit by Hank Aaron|hit is a double) = P(hit by Hank Aaron)
b. No, because P(hit by Hank Aaron|hit is a double) ≠ P(hit is a double)
c. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron)
d. Yes, because P(hit is by Hank Aaron|hit is a double) = P(hit is a double)

103. United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.

b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

104. At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

a. Find the probability that a course has a final exam or a research project.
b. Find the probability that a course has NEITHER of these two requirements.

105. In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

106. A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class andE = event that a student is a part time student

a. Find $P(D \cap E)$.
b. Find $P(E|D)$.
c. Find $P(D \cup E)$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

3.6 Venn Diagrams

Use the information in the Table $3.H.16$ to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Table $3.H.16$

Up for reelection:	Democratic party	Republican party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

107. What is the probability that a randomly selected senator has an “Other” affiliation?

108. What is the probability that a randomly selected senator is up for reelection in November 2016?

109. What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

110. What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

111. Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

112. Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

113. The events “Republican” and “Up for reelection in 2016” are ________

1. mutually exclusive.
2. independent.
3. both mutually exclusive and independent.
4. neither mutually exclusive nor independent.

114. The events “Other” and “Up for reelection in November 2016” are ________

a. mutually exclusive.
b. independent.

c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

115. Table $3.H.17$ gives the number of participants in the recent National Health Interview Survey who had been treated for cancer in the previous 12 months. The results are sorted by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Table $3.H.17$

Race and sex	15–24	25–40	41–65	Over 65	TOTALS
White, male	1,165	2,036	3,703		8,395
White, female	1,076	2,242	4,060		9,129
Black, male	142	194	384		824
Black, female	131	290	486		1,061
All others					
TOTALS	2,792	5,279	9,354		21,081

Do not include "all others" for parts f and g.

a. Fill in the column for cancer treatment for individuals over age 65.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a white male.
d. Find the probability that a randomly selected individual was a black female.
e. Find the probability that a randomly selected individual was black
f. Find the probability that a randomly selected individual was male.
g. Out of the individuals over age 65, find the probability that a randomly selected individual was a black or white male.

Use the following information to answer the next two exercises. The table of data obtained from *www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

Table $3.H.18$

Name	Single	Double	Triple	Home run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

116. Find P(hit was made by Babe Ruth).

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

117. Find P(hit was made by Ty Cobb|The hit was a Home Run).

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

118. Table $3.H.19$ identifies a group of children by one of four hair colors, and by type of hair.

Table $3.H.19$

Hair type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If B is the event of a child having brown hair, find the probability of the complement of B.
g. In words, what does the complement of B represent?

119. In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the*San Jose Mercury News*. The factual data were compiled into the following table.

Table $3.H.20$

Shirt #	≤ 210	211–250	251–290	> 290
1–33	21	5	0	290" class="lt-stats-5547">290" class="">290" class=" ">0
34–66	6	18	7	290" class="lt-stats-5547">290" class="">290" class=" ">4
66–99	6	12	22	290" class="lt-stats-5547">290" class="">290" class=" ">5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (R), four yellow (Y) and five blue (B) beads. For the coin, P(H) = $\frac{2}{3}$ and P(T) = $\frac{1}{3}$ where H is heads and T is tails.

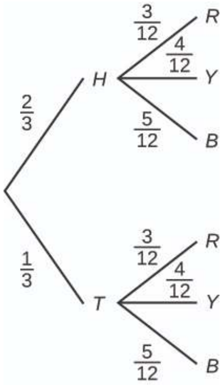


Figure $3.H.20$

120. Find P(tossing a Head on the coin AND a Red bead)

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

121. Find P(Blue bead).

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

122. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find P(S).
e. Let T be the event that the cookies selected were different flavors. Find P(T) by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find P(U).

1.1: Chapter Practice

1.2: Chapter More Practice

3.S: Probability (Solutions- Practice + Homework)

1.

1. $P(L') = P(S)$
2. $P(M \cup S)$
3. $P(F \cap L)$
4. $P(M|L)$
5. $P(L|M)$
6. $P(S|F)$
7. $P(F|L)$
8. $P(F \cup L)$
9. $P(M \cap S)$
10. $P(F)$

3.

$P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

5.

$P(C) = \frac{5}{42} = 0.12$

7.

$P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

9.

$P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

11.

$P(O) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

13.

$P(E) = \frac{47}{194} = 0.24$

15.

$P(N) = \frac{23}{194} = 0.12$

17.

$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

19.

$\frac{13}{52} = \frac{1}{4} = 0.25$

21.

$\frac{3}{6} = \frac{1}{2} = 0.5$

23.

$P(R) = \frac{4}{8} = 0.5$

25.

$P(O \cup H)$

27.

$P(H|I)$

29.

$P(N|O)$

31.

$P(I \cup N)$

33.

$P(I)$

35.

The likelihood that an event will occur given that another event has already occurred.

37.

1

39.

the probability of landing on an even number or a multiple of three

41.

$P(J) = 0.3$

43.

$P(Q \cap R) = P(Q)P(R)$

$0.1 = (0.4)P(R)$

$P(R) = 0.25$

45.

0.376

47.

C|L means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

49.

L \cap C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

51.

0.6492

53.

No, because P(L \cap C) does not equal 0.

55.

$P(\text{ musician is a male } \cap \text{ had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12.$

57.

The events are not mutually exclusive. It is possible to be a female musician who learned music in school.

58.

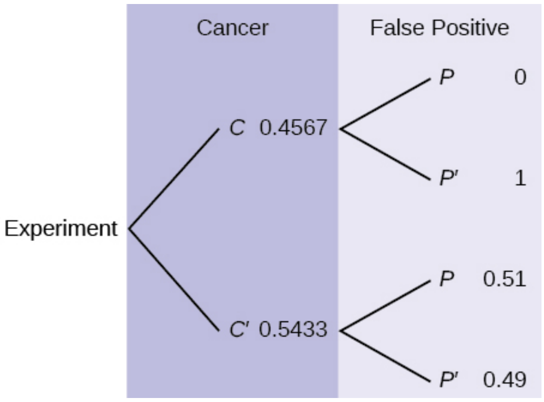


Figure $3.S.21$

60.

$\frac{35,065}{100,450}$

62.

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

64.

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

66.

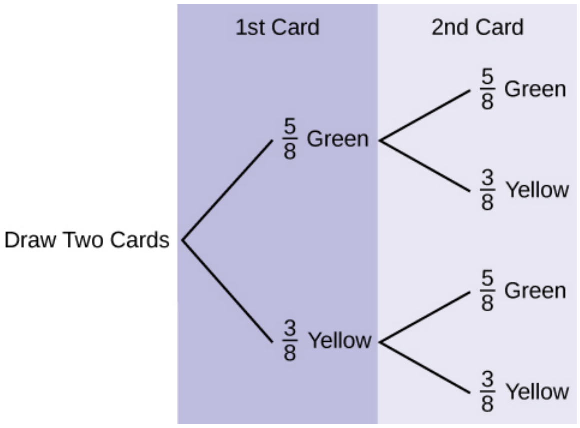
1. 

   Figure $3.S.22$
2. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
3. $P(\text{ at least one green }) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
4. $P(G|G) = \frac{5}{8}$
5. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

68.

1.

	<20>	20–64	>64	Totals

	<20>	20–64	>64	Totals
Female	" class=" ">" class=" ">0.0244	0.3954	64" class=" ">64" class=" ">64">0.0661	0.486
Male	" class=" ">" class=" ">0.0259	0.4186	64" class=" ">64" class=" ">64">0.0695	0.514
Totals	" class=" ">" class=" ">0.0503	0.8140	64" class=" ">64" class=" ">64">0.1356	1

Table3.22

2. $P(F) = 0.486$
3. $P(>64|F) = 0.1361$
4. $P(>64 \text{ and } F) = P(F)P(>64|F) = (0.486)(0.1361) = 0.0661$
5. $P(>64|F)$ is the percentage of female drivers who are 65 or older and P(>64 \cap F) is the percentage of drivers who are female and 65 or older.
6. $P(>64) = P(>64 \cap F) + P(>64 \cap M) = 0.1356$
7. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \cap F) = 0.0661$.

70.

1.

	Car, truck or van	Walk	Public transportation	Other	Totals
Alone	0.7318				
Not alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

Table3.23

2. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$
3. Make the same assumptions as in (b) we have: $(0.7708)(1,000) = 771$
4. $(0.1332)(1,000) = 133$

73.

1. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
2. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

75.

0

77.

0.3571

79.

0.2142

81.

Physician (83.7)

83.

$83.7 - 79.6 = 4.1$

85.

$P(\text{Occupation} < 81.3) = 0.5$

87.

1. The Forum Research surveyed 1,046 Torontonians.
2. 58%
3. 42% of 1,046 = 439 (rounding to the nearest integer)
4. 0.57
5. 0.60.

89.

1. $P(\text{ Betting on two line that touch each other on the table}) = \frac{6}{38}$.
2. $P(\text{ Betting on three numbers in a line }) = \frac{3}{38}$
3. $P(\text{ Betting on one number }) = \frac{1}{38}$
4. $P(\text{ Betting on four number that touch each other to form a square}) = \frac{4}{38}$.
5. $P(\text{ Betting on two number that touch each other on the table }) = \frac{2}{38}$
6. $P(\text{ Betting on } 0-00-1-2-3) = \frac{5}{38}$
7. $P(\text{ Betting on } 0-1-2;\ \text{or } 0-00-2;\ \text{or } 00-2-3) = \frac{3}{38}$

91.

1. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
2. $\frac{5}{8}$
3. $\frac{2}{3}$
4. $\frac{2}{8}$
5. $\frac{6}{8}$
6. No, because $P(G \cap E)$ does not equal 0.

93.

NOTE

The coin toss is independent of the card picked first.

1. $\{(G,H)(G,T)(B,H)(B,T)(R,H)(R,T)\}$
2. $P(A) = P(\text{ blue })P(\text{ head }) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
3. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \cap B) = 0$
4. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(A \cap C) = P(A) = \frac{3}{20}$

95.

1. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
2. $\frac{4}{8}$
3. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \cap B) = 0$.

97.

1. If Y and Z are independent, then $P(Y \cap Z) = P(Y)P(Z)$, so $P(Y \cup Z) = P(Y) + P(Z) - P(Y)P(Z)$.
2. 0.5

99.

iii i iv ii

101.

1. $P(R) = 0.44$
2. $P(R|E) = 0.56$

3. $P(R|O) = 0.31$
4. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
5. No, this study definitely does not support that notion; in fact, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

103.

1. $P(\text{ type O} \cup \text{Rh}-) = P(\text{ type O}) + P(\text{Rh}-) - P(\text{ type O} \cap \text{Rh}-)$

   $0.52 = 0.43 + 0.15 - P(\text{ type } O \cap \text{Rh}-)$; solve to find $P(\text{ type O} \cap \text{Rh}-) = 0.06$

   6% of people have type O, Rh- blood
2. $P(\text{ NOT (type O } \cap \text{Rh}-)) = 1 - P(\text{ type O} \cap \text{Rh}-) = 1 - 0.06 = 0.94$

   94% of people do not have type O, Rh- blood

105.

1. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
2. $P(C \cup N) = P(C) + P(N) - P(C \cap N) = 0.36 + 0.12 - 0.08 = 0.40$
3. $P(\text{ NEITHER chocolate NOR nuts}) = 1 - P(C \cup N) = 1 - 0.40 = 0.60$

107.

0

109.

$\frac{10}{67}$

111.

$\frac{10}{34}$

113.

d

115.

1.

Race and sex	1–14	15–24	25–64	Over 64	TOTALS
White, male	210	3,360	13,610	4,870	22,050
White, female	80	580	3,380	890	4,930
Black, male	10	460	1,060	140	1,670
Black, female	0	40	270	20	330
All others				100	
TOTALS	310	4,650	18,780	6,020	29,760

Table3.24

2.

Race and sex	1–14	15–24	25–64	Over 64	TOTALS
White, male	210	3,360	13,610	4,870	22,050
White, female	80	580	3,380	890	4,930
Black, male	10	460	1,060	140	1,670
Black, female	0	40	270	20	330

Race and sex	1–14	15–24	25–64	Over 64	TOTALS
All others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

Table3.25

3. $\frac{22{,}050}{29{,}760}$
4. $\frac{330}{29{,}760}$
5. $\frac{2{,}000}{29{,}760}$
6. $\frac{23{,}720}{29{,}760}$
7. $\frac{5{,}010}{6{,}020}$

117.

b

119.

1. $\frac{26}{106}$
2. $\frac{33}{106}$
3. $\frac{21}{106}$
4. $\left(\frac{26}{106}\right)+\left(\frac{33}{106}\right)-\left(\frac{21}{106}\right)=\left(\frac{38}{106}\right)$
5. $\frac{21}{33}$

121.

a

CHAPTER OVERVIEW

4: Random Variables

4.1: Introduction to Discrete Random Variables

Figure 4.1.1 You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the historical average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count, that is, the random variable can only take on whole number values. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment, often called a trial.

Random Variable Notation

The upper case letter X denotes a random variable. Lower case letters like x or y denote the value of a random variable. If **X is a random variable, then X is written in words, and x is given as a number.**

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*; *THH*; *HTH*; *HHT*; *HTT*; *THT*; *TTH*; *HHH*. Then, x = 0, 1, 2, 3. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values as whole numbers that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Probability Density Functions (PDF) for a Random Variable

A **probability density function** or **probability distribution function** has two characteristics:

1. A probability density function is a mathematical formula that calculates probabilities for specific types of events, what we have been calling experiments. There is a sort of magic to a probability density function (PDF) partially because the same formula often describes very different types of events. For example, the binomial PDF will calculate probabilities for flipping coins, yes/no questions on an exam, opinions of voters in an up or down opinion poll, indeed any binary event. Other probability density functions will provide probabilities for the time until a part will fail, when a customer will arrive at the turnpike booth, the number of telephone calls arriving at a central switchboard, the growth rate of a bacterium, and on and on. There are whole families of probability density functions that are used in a wide variety of applications, including medicine, business and finance, physics and engineering, among others.

   For our needs here we will concentrate on only a few probability density functions as we develop the tools of inferential statistics.

   ### Counting Formulas and the Combinational Formula

   As an equation this is:

$$P(A)=\frac{\text{number of ways to get A}}{\text{Total number of possible outcomes}} \tag{4.1.1}$$

When we looked at the sample space for flipping 3 coins we could easily write the full sample space and thus could easily count the number of events that met our desired result, e.g. $x = 1$, where X is the random variable defined as the number of heads.

As we have larger numbers of items in the sample space, such as a full deck of 52 cards, the ability to write out the sample space becomes impossible.

We see that probabilities are nothing more than counting the events in each group we are interested in and dividing by the number of elements in the universe, or sample space. This is easy enough if we are counting sophomores in a Stat class, but in more complicated cases listing all the possible outcomes may take a life time. There are, for example, 36 possible outcomes from throwing just two six-sided dice where the random variable is the sum of the number of spots on the up-facing sides. If there were four dice then the total number of possible outcomes would become 1,296. There are more than 2.5 MILLION possible 5 card poker hands in a standard deck of 52 cards. Obviously keeping track of all these possibilities and counting them to get at a single probability would be tedious at best.

An alternative to listing the complete sample space and counting the number of elements we are interested in, is to skip the step of listing the sample space, and simply figuring out the number of elements in it and doing the appropriate division. If we are after a probability we really do not need to see each and every element in the sample space, we only need to know how many elements are there. Counting formulas were invented to do just this. They tell us the number of unordered subsets of a certain size that can be created from a set of unique elements. By unordered it is meant that, for example, when dealing cards, it does not matter if you got {ace, ace, ace, ace, king} or {king, ace, ace, ace, ace} or {ace, king, ace, ace, ace} and so on. Each of these subsets are the same because they each have 4 aces and one king.

Combinational Formula

$$\binom{n}{x} = {}_n C_x = \frac{n!}{x!(n-x)!}$$

This is the formula that tells the number of unique unordered subsets of size x that can be created from n unique elements. The formula is read “n combinatorial x”. Sometimes it is read as “n choose x." The exclamation point "!" is called a factorial and tells us to take all the numbers from 1 through the number before the ! and multiply them together thus 4! is 1·2·3·4=24. By definition 0! = 1. The formula is called the Combinatorial Formula. It is also called the Binomial Coefficient, for reasons that will be clear shortly. While this mathematical concept was understood long before 1653, Blaise Pascal is given major credit for his proof that he published in that year. Further, he developed a generalized method of calculating the values for combinatorials known to us as the Pascal Triangle. Pascal was one of the geniuses of an era of extraordinary intellectual advancement which included the work of Galileo, Rene Descartes, Isaac Newton, William Shakespeare and the refinement of the scientific method, the very rationale for the topic of this text.

Let’s find the hard way the total number of combinations of the four aces in a deck of cards if we were going to take them two at a time. The sample space would be:

S={Spade,Heart),(Spade, Diamond),(Spade,Club), (Diamond,Club),(Heart,Diamond),(Heart,Club)}

There are 6 combinations; formally, six unique unordered subsets of size 2 that can be created from 4 unique elements. To use the combinatorial formula we would solve the formula as follows:

$$\binom{4}{2} = \frac{4!}{(4-2)!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6$$

If we wanted to know the number of unique 5 card poker hands that could be created from a 52 card deck we simply compute:

$$\binom{52}{5}$$

where 52 is the total number of unique elements from which we are drawing and 5 is the size group we are putting them into.

With the combinatorial formula we can count the number of elements in a sample space without having to write each one of them down, truly a lifetime's work for just the number of 5 card hands from a deck of 52 cards. We can now apply this tool to a very important probability density function, the hypergeometric distribution.

Remember, a probability density function computes probabilities for us. We simply put the appropriate numbers in the formula and we get the probability of specific events. However, for these formulas to work they must be applied only to cases for which they were designed.

4.2: Binomial Distribution

A more valuable probability density function with many applications is the binomial distribution. This distribution will compute probabilities for any binomial process. A binomial process, often called a Bernoulli process after the first person to fully develop its properties, is any case where there are only two possible outcomes in any one trial, called successes and failures. It gets its name from the binary number system where all numbers are reduced to either 1's or 0's, which is the basis for computer technology and CD music recordings.

Binomial Formula

$$b(x)=\binom{n}{x} p^x q^{n-x}$$

where $b(x)$ is the probability of x successes in n trials when the probability of a success in ANY ONE TRIAL is p. And of course $q=(1-p)$ and is the probability of a failure in any one trial.

We can see now why the combinatorial formula is also called the binomial coefficient because it reappears here again in the binomial probability function. For the binomial formula to work, the probability of a success in any one trial must be the same from trial to trial, or in other words, the outcomes of each trial must be independent. Flipping a coin is a binomial process because the probability of getting a head in one flip does not depend upon what has happened in PREVIOUS flips. (At this time it should be noted that using p for the parameter of the binomial distribution is a violation of the rule that population parameters are designated with Greek letters. In many textbooks θ (pronounced theta) is used instead of p and this is how it should be.

Just like a set of data, a probability density function has a mean and a standard deviation that describes the data set. For the binomial distribution these are given by the formulas:

$$\mu=np$$

$$\sigma=\sqrt{npq}$$

Notice that p is the only parameter in these equations. The binomial distribution is thus seen as coming from the one-parameter family of probability distributions. In short, we know all there is to know about the binomial once we know p, the probability of a success in any one trial.

In probability theory, under certain circumstances, one probability distribution can be used to approximate another. We say that one is the limiting distribution of the other. If a small number is to be drawn from a large population, even if there is no replacement, we can still use the binomial even thought this is not a binomial process. If there is no replacement it violates the independence rule of the binomial. Nevertheless, we can use the binomial to approximate a probability that is really a hypergeometric distribution if we are drawing fewer than 10 percent of the population, i.e. n is less than 10 percent of N in the formula for the hypergeometric function. The rationale for this argument is that when drawing a small percentage of the population we do not alter the probability of a success from draw to draw in any meaningful way. Imagine drawing from not one deck of 52 cards but from 6 decks of cards. The probability of say drawing an ace does not change the conditional probability of what happens on a second draw in the same way it would if there were only 4 aces rather than the 24 aces now to draw from. This ability to use one probability distribution to estimate others will become very valuable to us later.

There are 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. The random variable, X, number of successes, is discrete.
3. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on any one trial, and q denotes the probability of a failure on any one trial. $p+q=1$.
4. The n trials are independent and are repeated using identical conditions. Think of this as drawing WITH replacement. 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 a 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 $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ, is then \sigma = $\sqrt{npq}$.

Any experiment that has characteristics three and four 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

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 4.2.1

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. Find the $P(X = 12)$ using the binomial PDF

? Example 4.2.2

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.

Answer

$P(x > 10)$

? Example 4.2.3

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.

Answer

a. failure

b. If we are interested in the number of students who do their homework on time, then how do we define X?

Answer

b. X = the number of statistics students who do their homework on time

c. What values does x take on?

Answer

c. 0, 1, 2, …, 50

d. What is a "failure," in words?

Answer

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. If $p + q = 1$, then what is q?

Answer

e. $q = 0.30$

f. The words "at least" translate as what kind of inequality for the probability question $P(x$ ____ 40).

Answer

f. greater than or equal to $(\geq)$
The probability question is $P(x \geq 40)$.

? Exercise 4.2.2

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

? Exercise 4.2.3

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. 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.

4.3: Poisson Distribution

Another useful probability distribution is the Poisson distribution, or waiting time distribution. This distribution is used to determine how many checkout clerks are needed to keep the waiting time in line to specified levels, how may telephone lines are needed to keep the system from overloading, and many other practical applications. A modification of the Poisson, the Pascal, invented nearly four centuries ago, is used today by telecommunications companies worldwide for load factors, satellite hookup levels and Internet capacity problems. The distribution gets its name from Simeon Poisson who presented it in 1837 as an extension of the binomial distribution which we will see can be estimated with the Poisson.

There are two main characteristics of a Poisson experiment.

1. The **Poisson probability distribution** gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate.
2. The events are independent of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages and it is assumed that there is no relationship between when misspellings occur.
3. The random variable X = the number of occurrences in the interval of interest.

? Example 4.3.1

A bank expects to receive six bad checks per day, on average. What is the probability of the bank getting fewer than five bad checks on any given day? Of interest is the number of checks the bank receives in one day, so the time interval of interest is one day. Let X = the number of bad checks the bank receives in one day. If the bank expects to receive six bad checks per day then the average is six checks per day. Write a mathematical statement for the probability question.

Answer

$P(X<5)$

? Example 4.3.2

You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast.

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

a. What is the interval of interest?

Answer

a. one broadcast measured in minutes

b. What is the average number of times the news reporter says "uh" during one broadcast?

Answer

b. 2

c. Let X = ____________. What values does X take on?

Answer

c. Let X = the number of times the news reporter says "uh" during one broadcast.
$x=0,1,2,3$, ...

d. The probability question is P (______).

Answer

d. $P(X>2)$

Poisson Probability Distribution Function

$X \sim \text{Poisson}(\mu)$

Read this as "X is a random variable with a Poisson distribution." The parameter is μ; μ = the mean for the interval of interest. The mean is the number of occurrences that occur on average during the interval period.

The formula for computing probabilities that are from a Poisson process is:

$$p(x)=\frac{\mu^{x} e^{-\mu}}{x!}$$

where $p(x)=P(X=x)$ is the probability of x occurrences, μ is the expected number of occurrences based upon historical data, e is the natural logarithm approximately equal to 2.718.

In order to use the Poisson distribution, certain assumptions must hold. These are: the probability of an occurrence is unchanged within the interval, there cannot be simultaneous occurrences within the interval, and finally, that the probability of a occurrence among intervals is independent, the same assumption of the binomial distribution.

In a way, the Poisson distribution can be thought of as a clever way to convert a continuous random variable, usually time, into a discrete random variable by breaking up time into discrete independent intervals. This way of thinking about the Poisson helps us understand why it can be used to estimate the probability for the discrete random variable from the binomial distribution. The Poisson is asking for the probability of a number of successes during a period of time while the binomial is asking for the probability of a certain number of successes for a given number of trials.

? Example 4.3.3

Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call **in the next 15 minutes?**

Let X = the number of calls Leah receives in 15 minutes. (The **interval of interest** is 15 minutes or $\frac{1}{4}$ hour.)

$x=0,1,2,3,\ldots$

If Leah receives, on the average, six telephone calls in two hours, and there are eight 15 minute intervals in two hours, then Leah receives

$\left(\frac{1}{8}\right)$(6)= 0.75 calls in 15 minutes, on average. So, $\mu=0.75$ for this problem.

$X \sim \text{Poisson}(0.75)$

Find $P(X>1)$:

$$P(X>1)=1-P(X=0)=1-\frac{0.75^{0} e^{-0.75}}{0!}=0.1734$$

Probability that Leah receives more than one telephone call in the next 15 minutes is about 0.1734.

The graph of $X \sim \text{Poisson}(0.75)$ is:

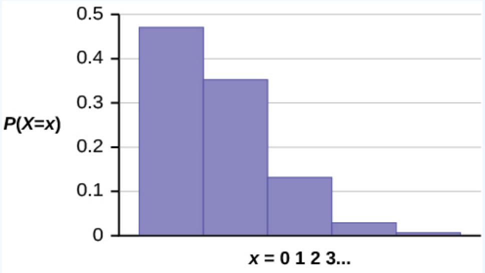


Figure 4.3.1

The y-axis contains the probability of X = the number of calls in 15 minutes.

Example 4.3.4

According to a survey a university professor gets, on average, 7 emails per day. Let X = the number of emails a professor receives per day. The discrete random variable X takes on the values x = 0, 1, 2 …. The random variable X has a Poisson distribution: $X \sim \text{Poisson}(7)$. The mean is 7 emails.

a. What is the probability that an email user receives exactly 2 emails per day?
b. What is the probability that an email user receives at most 2 emails per day?
c. What is the standard deviation?

Answer

a. $P(X=2)=p(2)=\frac{\mu^x e^{-\mu}}{x!}=\frac{7^2 e^{-7}}{2!}=0.022$

b. $P(X \leq 2)=p(0)+p(1)+p(2)=\frac{7^0 e^{-7}}{0!}+\frac{7^1 e^{-7}}{1!}+\frac{7^2 e^{-7}}{2!}=0.029$

c. Standard Deviation = $\sigma=\sqrt{\mu}=\sqrt{7} \approx 2.65$

Example 4.3.5

Text message users receive or send an average of 41.5 text messages per day.

a. How many text messages does a text message user receive or send per hour?
b. What is the probability that a text message user receives or sends two messages per hour?
c. What is the probability that a text message user receives or sends more than two messages per hour?

Answer

a.Let X = the number of texts that a user sends or receives in one hour. The average number of texts received per hour is $\frac{41.5}{24} \approx 1.7292$.

b.$P(X=2)=p(2)=\frac{\mu^x e^{-\mu}}{x!}=\frac{1.729^2 e^{-1.729}}{2!}=0.265$

c.$P(>x2)=1-P(X \leq 2)=1-(p(0)+p(1)+p(2))=1-\left[\frac{7^0 e^{-7}}{0!}+\frac{7^1 e^7}{1!}+\frac{7^2 e^{-7}}{2!}\right]=0.250$

Estimating the Binomial Distribution with the Poisson Distribution

The Poisson distribution can provide an approximation for the binomial. We say that the binomial distribution approaches the Poisson. The binomial distribution approaches the Poisson distribution is as *n* gets larger and *p* is small such that *np* becomes a constant value. There are several rules of thumb for when one can say they will use a Poisson to estimate a binomial. One suggests that *np*, the mean of the binomial, should be less than 25. Another author suggests that it should be less than 7. And another, noting that the mean and variance of the Poisson are both the same, suggests that *np* and *npq*, the mean and variance of the binomial, should be greater than 5. There is no one broadly accepted rule of thumb for when one can use the Poisson to estimate the binomial.

As we move through these probability distributions we are getting to more sophisticated distributions that, in a sense, contain the less sophisticated distributions within them. This proposition has been proven by mathematicians. This gets us to the highest level of sophistication in the next probability distribution which can be used as an approximation to all of those that we have discussed so far. This is the normal distribution.

Example 4.3.6

On May 13, 2013, starting at 4:30 PM, the probability of low seismic activity for the next 48 hours in Alaska was reported as about 1.02%. Use this information for the next 200 days to find the probability that there will be low seismic activity in ten of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Answer

Let X = the number of days with low seismic activity.

Using the binomial distribution:

$$P(X=10)=p(10)=\frac{200!}{10!(200-10)!}\times .0102^{10}\times .9898^{190}=0.000039$$

Using the Poisson distribution:

Calculate $\mu=np=200(0.0102)\approx 2.04$

$$P(X=10)=p(10)=\frac{\mu^{x}e^{-\mu}}{x!}=\frac{2.04^{10}e^{-2.04}}{10!}=0.000045$$

We expect the approximation to be good because n is large (greater than 20) and p is small (less than 0.05). The results are close—both probabilities reported are almost 0.

Example 4.3.7

A survey of 500 seniors in the Price Business School yields the following information. 75% go straight to work after graduation. 15% go on to work on their MBA. 9% stay to get a minor in another program. 1% go on to get a Master's in Finance.

What is the probability that more than 2 seniors go to graduate school for their Master's in finance?

Answer

This is clearly a binomial probability distribution problem. The choices are binary when we define the results as "Graduate School in Finance" versus "all other options." The random variable is discrete, and the events are, we could assume, independent. Solving as a binomial problem, we have:

Binomial Solution

$$n\cdot p=500\cdot 0.01=5=\mu$$

$$p(0)=\frac{500!}{0!(500-0)!}0.01^{0}(1-0.01)^{500^{-0}}=0.00657$$

$$p(1)=\frac{500!}{1!(500-1)!}0.01^{1}(1-0.01)^{500}=0.03318$$

$$p(2)=\frac{500!}{2!(500-2)!}0.01^{2}(1-0.01)^{500^{2}}=0.08363$$

Adding all 3 together = 0.12339

$$1-0.12339=0.87661$$

Poisson approximation

$$n\cdot p=500\cdot 0.01=5=\mu$$

$$n\cdot p\cdot(1-p)=500\cdot 0.01\cdot(0.99)\approx 5=\sigma^{2}=\mu$$

$$p(x)=\frac{e^{-np}(np)^{x}}{x!}=\left\{p(0)=\frac{e^{-5}\cdot 5^{0}}{0!}\right\}+\left\{p(1)=\frac{e^{-5}\cdot 5^{1}}{1!}\right\}+\left\{p(2)=\frac{e^{-5}\cdot 5^{2}}{2!}\right\}$$

$$0.0067+0.0337+0.0842=0.1247$$

$$1-0.1247=0.8753$$

An approximation that is off by 1 one thousandth is certainly an acceptable approximation.

4.4: Expected Value of Discrete Random Variables

4.4.1: Variance of Discrete Random Variables

4.5: Introduction to Continuous Random Variables

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, rates of return from an investment, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables, as do all areas of risk analysis.

Figure 4.5.1 The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

> **Note**
>
> The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

4.5.1: Properties of Continuous Probability Density Functions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve. We have already met this concept when we developed relative frequencies with histograms in Chapter 2. The relative area for a range of values was the probability of drawing at random an observation in that group. Again with the Poisson distribution, the graph in Example 4.3.3 used boxes to represent the probability of specific values of the random variable. In this case, we were being a bit casual because the random variables of a Poisson distribution are discrete, whole numbers, and a box has width. Notice that the horizontal axis, the random variable X, purposefully did not mark the points along the axis. The probability of a specific value of a continuous random variable will be zero because the area under a point is zero. Probability is area.

The curve is called the **probability density function** (abbreviated as **pdf**). We use the symbol $f(x))$ to represent the curve. $f(x))$ is the function that corresponds to the graph; we use the density function $f(x))$ to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as **cdf**). The cumulative distribution function is used to evaluate probability as area. Mathematically, the cumulative probability density function is the integral of the pdf, and the probability between two values of a continuous random variable will be the integral of the pdf between these two values: the area under the curve between these values. Remember that the area under the pdf for all possible values of the random variable is one, certainty. Probability thus can be seen as the relative percent of certainty between the two values of interest.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for *intervals* of x values rather than for individual x values.
- $P(c<X<d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c<X<d)$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(X=c)=0$ The probability that X takes on any single individual value is zero. The area below the curve, above the x-axis, and between $x=c$ and $x=c$ has no width, and therefore no area ($\text{area }=0$). Since the probability is equal to the area, the probability is also zero.
- $P(c<X<d)$ is the same as $P(c\leq x\leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, integral calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

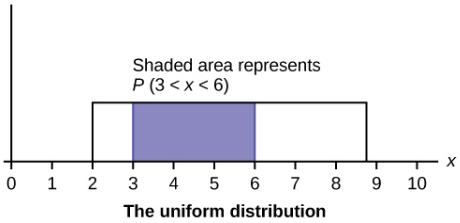


Figure 4.5.1.1 The graph shows a Uniform Distribution with the area between $x=3$ and $x=6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

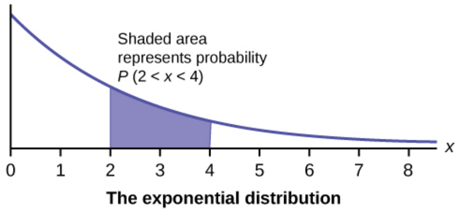


Figure 4.5.1.2 The graph shows an Exponential Distribution with the area between $x=2$ and $x=4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

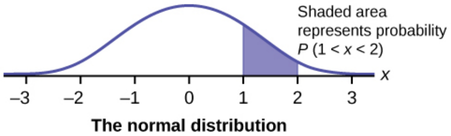


Figure 4.5.1.3 The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

For continuous probability distributions, PROBABILITY = AREA.

? Example 4.5.1.1

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.

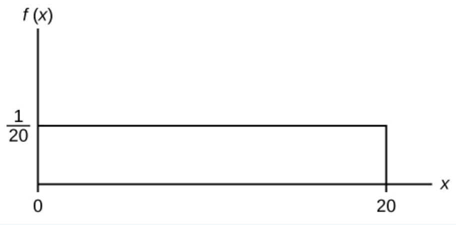


Figure 4.5.1.4

$f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$.

The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x-axis is the area of a rectangle with base $= 20$ and height $= \frac{1}{20}$.

$$\text{AREA} = 20\left(\frac{1}{20}\right) = 1$$

Suppose we want to find the area between \(bf{f(x)) = \frac{1}{20}}\) **and the *x*-axis where $\mathbf{0 < x < 2}$.**

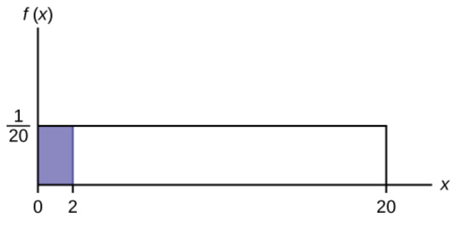


Figure 4.5.1.5

$$\text{AREA} = (2 - 0)\left(\frac{1}{20}\right) = 0.1$$

$$(2 - 0) = 2 = \text{base of rectangle}$$

REMINDER

area of a rectangle = (base)(height).

The area corresponds to a probability. The probability that x is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

Suppose we want to find the area between $\mathbf{f(x) = \frac{1}{20}}$ and the *x*-axis where $\mathbf{4 < x < 15}$.

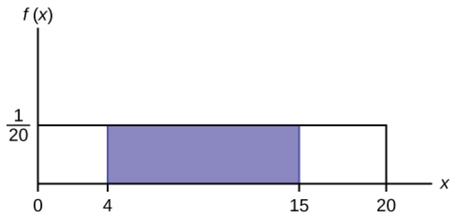


Figure 4.5.1.6

$\text{AREA} = (15-4)\left(\frac{1}{20}\right) = 0.55$

$(15-4) = 11 = \text{the base of a rectangle}$

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find $P(x = 15)$. On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or zero width). Therefore, $P(x = 15) = \text{(base)(height)} = (0)\left(\frac{1}{20}\right) = 0$

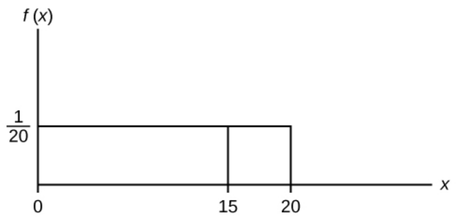


Figure 4.5.1.7

$P(X \leq x)$, which can also be written as $P(X < x)$ for continuous distributions, is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can also use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X < x)$.

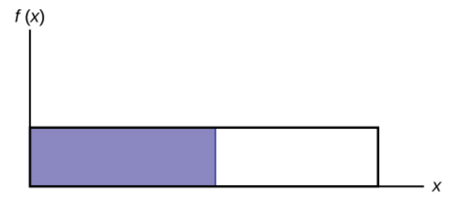


Figure 4.5.1.8

Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}, 0 \leq x \leq 20$.

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.

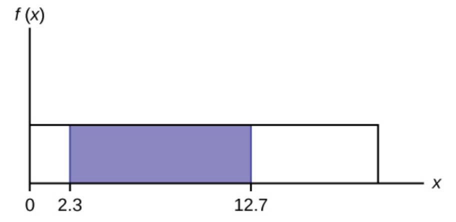


Figure 4.5.1.9

$P(2.3 < x < 12.7) = (\text{ base })(\text{ height }) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$

Exercise 4.5.1.1

Consider the function $f(x)=\frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$) and find $P(2.5<x<7.5)$.

4.6: The Uniform Distribution

The uniform distribution is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive of endpoints.

The mathematical statement of the uniform distribution is

$f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$

where $a =$ the lowest value of x and $b =$ the highest value of x.

Formulas for the theoretical mean and standard deviation are

$\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Exercise 4.6.1

The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a uniform distribution where all values between and including zero and 14 are equally likely. State the values of a and b. Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

1	12	4	10	4	14	11
7	11	4	13	2	4	6
3	10	0	12	6	9	10
5	13	4	10	14	12	11
6	10	11	0	11	13	2

Table 4.6.1

Example 4.6.1

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

a. What is the probability that a person waits fewer than 12.5 minutes?

Answer

a. Let X = the number of minutes a person must wait for a bus. $a = 0$ and $b = 15$. $X \sim U(0, 15)$. Write the probability density function. $f(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \leq x \leq 15$.

Find $P(x < 12.5)$. Draw a graph.

$$P(x < k) = \text{(base)}\ \text{(height)} = (12.5 - 0)\left(\frac{1}{15}\right) = 0.8333$$

The probability a person waits less than 12.5 minutes is 0.8333.

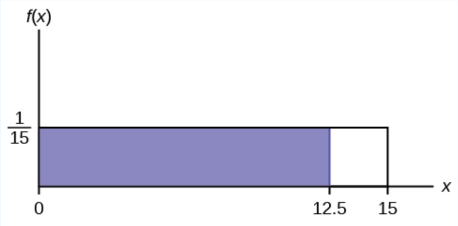


Figure 4.6.1

b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

Answer

b. $\mu = \frac{a+b}{2} = \frac{15+0}{2} = 7.5$. On the average, a person must wait 7.5 minutes.

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-\theta)^2}{12}} = 4.3$. The Standard deviation is 4.3 minutes.

c. Ninety percent of the time, the time a person must wait falls below what value?

Note

This asks for the 90th percentile.

Answer

c. Find the 90th percentile. Draw a graph. Let $k =$ the 90th percentile.

$P(x < k) > \backslash(0.90 = (k)\left(\frac{1}{15}\right)$

$k = (0.90)(15) = 13.5$

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.

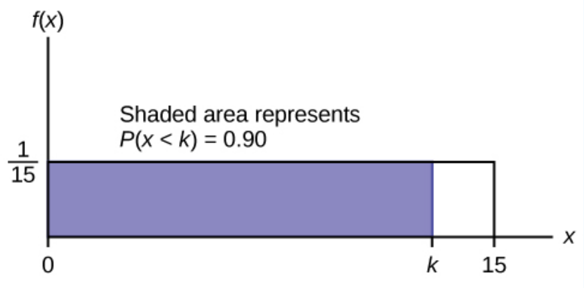


Figure 4.6.2

Exercise 4.6.2

The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

1. Find a and b and describe what they represent.
2. Write the distribution.
3. Find the mean and the standard deviation.
4. What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?

4.7: The Exponential Distribution

The **exponential distribution** is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length of time, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an exponential distribution.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, marketing studies have shown that the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

Exponential distributions are commonly used in calculations of product reliability, or the length of time a product lasts.

The random variable for the exponential distribution is continuous and often measures a passage of time, although it can be used in other applications. Typical questions may be, “what is the probability that some event will occur within the next x hours or days, or what is the probability that some event will occur between x_1 hours and x_2 hours, or what is the probability that the event will take more than x_1 hours to perform?” In short, the random variable X equals (a) the time between events or (b) the passage of time to complete an action, e.g. wait on a customer. The probability density function is given by:

$$f(x)=\frac{1}{\mu}e^{-\frac{x}{\mu}}$$

where μ is the historical average waiting time.

An alternative form of the exponential distribution formula recognizes what is often called the decay factor. The decay factor simply measures how rapidly the probability of an event declines as the random variable X increases. When the notation using the decay parameter λ is used, the probability density function is presented as:

$$f(x)=\lambda e^{-\lambda x}$$

where $\lambda=\frac{1}{\mu}$

In order to calculate probabilities for specific probability density functions, the cumulative density function is used. The cumulative density function (cdf) is simply the integral of the pdf and is:

$$F(x)=\int_0^x\left[\frac{1}{\mu}e^{-\frac{t}{\mu}}\right]dt=1-e^{-\frac{x}{\mu}}$$

? Example 4.7.1

Let X = amount of time (in minutes) a postal clerk spends with a customer. The time is known from historical data to have an average amount of time equal to four minutes.

It is given that $\mu=4$ minutes, that is, the average time the clerk spends with a customer is 4 minutes. Remember that we are still doing probability and thus we have to be told the population parameters such as the mean. To do any calculations, we need to know the mean of the distribution: the historical time to provide a service, for example. Knowing the historical mean allows the calculation of the decay parameter, λ.

$\lambda=\frac{1}{\mu}$. Therefore, $\lambda=\frac{1}{4}=0.25$.

When the notation used the decay parameter, λ, the probability density function is presented as $f(x)=\lambda e^{-\lambda x}$, which is simply the original formula with λ substituted for $\frac{1}{\mu}$, or $f(x)=\frac{1}{\mu}e^{-\frac{1}{\mu}x}$.

To calculate probabilities for an exponential probability density function, we need to use the cumulative density function. As shown below, the curve for the cumulative density function is:

$f(x)=0.25e^{-0.25x}$ where x is at least zero and $\lambda=0.25$.

For example, $f(5)=0.25e^{(-0.25)(5)}=0.072$. In other words, the function has a value of .072 when $x=5$.

The graph is as follows:

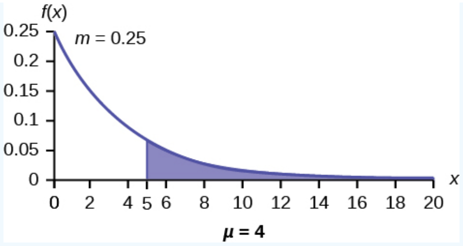


Figure 4.7.1

Notice the graph is a declining curve. When $x=0$,

$f(x)=0.25e^{(-0.25)(0)}=(0.25)(1)=0.25=m$. The maximum value on the y-axis is always m, one divided by the mean.

Exercise 4.7.1

The amount of time spouses shop for anniversary cards can be modeled by an exponential distribution with the average amount of time equal to eight minutes. Write the distribution, state the probability density function, and graph the distribution.

Example 4.7.2

a. Using the information in Example 4.7.1, find the probability that a clerk spends four to five minutes with a randomly selected customer.

Answer

a. Find $P(4<X<5)$.
The cumulative distribution function (CDF) gives the area to the left.
$P(X<x)=1–e^{-\lambda x}$
$P(X<5)=1–e^{(-0.25)(5)}=0.7135$ and $P(x<4)=1–e^{(-0.25)(4)}=0.6321$
$P(4<X<5)=0.7135–0.6321=0.0814$

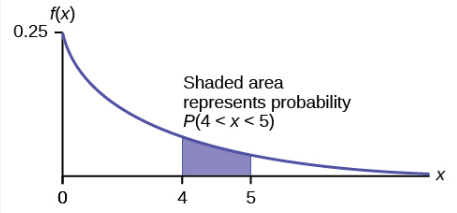


Figure 4.7.2

Exercise 4.7.2

The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days. Find the probability that a traveler will purchase a ticket fewer than ten days in advance. How many days do half of all travelers wait?

Example 4.7.3

On the average, a certain computer part lasts ten years. The length of time the computer part lasts is exponentially distributed.

a. What is the probability that a computer part lasts more than 7 years?

Answer

a. Let $x=$ the amount of time (in years) a computer part lasts.
$\mu=10$ so $\lambda=\frac{1}{\mu}=\frac{1}{10}=0.1$
Find $P(X>7)$. Draw the graph.

$P(X > 7) = 1 - P(X < 7)$.

Since $P(X < x) = 1 - e^{-\lambda x}$ then $P(X > x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$

$P(X > 7) = e^{(-0.1)(7)} = 0.4966$. The probability that a computer part lasts more than seven years is 0.4966

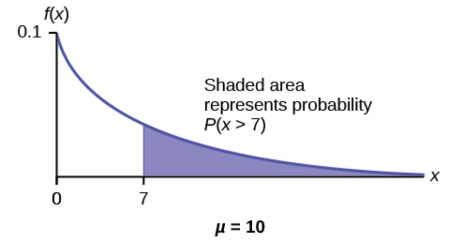


Figure 4.7.3

b. On the average, how long would five computer parts last if they are used one after another?

Answer

b. On the average, one computer part lasts ten years. Therefore, five computer parts, if they are used one right after the other would last, on the average, (5)(10) = 50 years.

d. What is the probability that a computer part lasts between 9 and 11 years?

Answer

d. Find $P(9 < X < 11)$. Draw the graph.

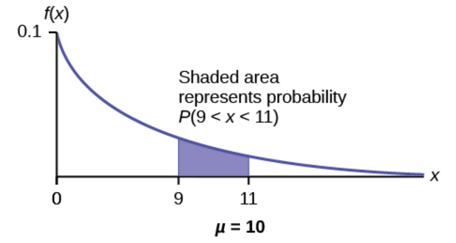


Figure 4.7.4

$P(9 < X < 11) = P(X < 11) - P(X < 9) = (1 - e^{(-0.1)(11)}) - (1 - e^{(-0.1)(9)}) = 0.6671 - 0.5934 = 0.0737$. The probability that a computer part lasts between 9 and 11 years is 0.0737.

? Exercise 4.7.3

On average, a pair of running shoes can last 18 months if used every day. The length of time running shoes last is exponentially distributed. What is the probability that a pair of running shoes last more than 15 months? On average, how long would six pairs of running shoes last if they are used one after the other? Eighty percent of running shoes last at most how long if used every day?

? Example 4.7.4

Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter $\frac{1}{12}$. If another person arrives at a public telephone just before you, find the probability that you will have to wait more than five minutes. Let X = the length of a phone call, in minutes.

What is $lambda$, μ, and σ? The probability that you must wait more than 5 minutes is _______ .

Answer

$\lambda = \frac{1}{12}$

$\mu = 12$

$\sigma = 12$

$P(X > 5) = e^{-5/12} = 0.6592$

Example 4.7.5

The time spent waiting between events is often modeled using the exponential distribution. For example, suppose that an average of 30 customers per hour arrive at a store and the time between arrivals is exponentially distributed.

1. On average, how many minutes elapse between two successive arrivals?
2. When the store first opens, how long on average does it take for three customers to arrive?
3. After a customer arrives, find the probability that it takes less than one minute for the next customer to arrive.
4. After a customer arrives, find the probability that it takes more than five minutes for the next customer to arrive.
5. Is an exponential distribution reasonable for this situation?

Answer

a.Since we expect 30 customers to arrive per hour (60 minutes), we expect on average one customer to arrive every two minutes on average.

b.Since one customer arrives every two minutes on average, it will take six minutes on average for three customers to arrive.

c.Let $X =$ the time between arrivals, in minutes. By part a, $\mu = 2$, so $\lambda = \frac{1}{2} = 0.5$.
The cumulative distribution function is $P(X < x) = 1 - e^{(-0.5)(x)}$
Therefore $P(X < 1) = 1 - e^{(-0.5)(1)} = 0.3935$.

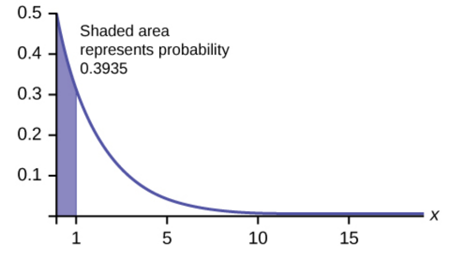


Figure 4.7.5

d.$P(X > 5) = 1 - P(X < 5) = 1 - (1 - e^{(-0.5)(5)}) = e^{-2.5} \approx 0.0821$.

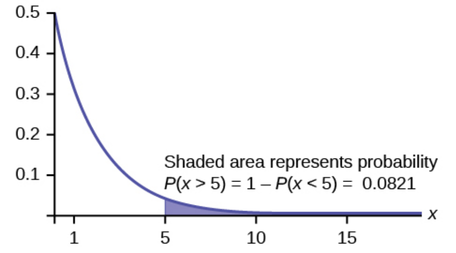


Figure 4.7.6

This model assumes that a single customer arrives at a time, which may not be reasonable since people might shop in groups, leading to several customers arriving at the same time. It also assumes that the flow of customers does not change throughout the day, which is not valid if some times of the day are busier than others.

Memoryless Property of the Exponential Distribution

Recall that the amount of time between customers for the postal clerk discussed earlier is exponentially distributed with a mean of two minutes. Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case–the additional time spent waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the **memoryless property**. The exponential and geometric probability density functions are the only probability functions that have the memoryless property. Specifically, the **memoryless property** says that

$$P(X > r + t | X > r) = P(X > t) \quad \text{for all } r \geq 0 \text{ and } t \geq 0$$

For example, if five minutes have elapsed since the last customer arrived, then the probability that more than one minute will elapse before the next customer arrives is computed by using r = 5 and t = 1 in the foregoing equation.

$$P(X > 5 + 1 | X > 5) = P(X > 1) = e^{(-0.5)(1)} = 0.6065 \ .$$

This is the same probability as that of waiting more than one minute for a customer to arrive after the previous arrival.

The exponential distribution is often used to model the longevity of an electrical or mechanical device. In Example $4.7.3$, the lifetime of a certain computer part has the exponential distribution with a mean of ten years. The **memoryless property** says that knowledge of what has occurred in the past has no effect on future probabilities. In this case it means that an old part is not any more likely to break down at any particular time than a brand new part. In other words, the part stays as good as new until it suddenly breaks. For example, if the part has already lasted ten years, then the probability that it lasts another seven years is $P(X > 17 | X > 10) = P(X > 7) = 0.4966$, where the vertical line is read as "given".

Example $4.7.6$

Refer back to the postal clerk again where the time a postal clerk spends with his or her customer has an exponential distribution with a mean of four minutes. Suppose a customer has spent four minutes with a postal clerk. What is the probability that he or she will spend at least an additional three minutes with the postal clerk?

The decay parameter of X is $\lambda = \frac{1}{4} = 0.25$.

The cumulative distribution function is $P(X < x) = 1 - e^{-0.25x}$.

We want to find $P(X > 7 | X > 4)$. The **memoryless property** says that $P(X > 7 | X > 4) = P(X > 3)$, so we just need to find the probability that a customer spends more than three minutes with a postal clerk.

This is $P(X > 3) = 1 - P(X < 3) = 1 - (1 - e^{-0.25 \cdot 3}) = e^{-0.75} \approx 0.4724$.

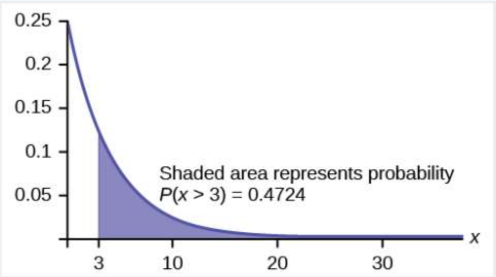


Figure $4.7.7$

Relationship between the Poisson and the Exponential Distribution

There is an interesting relationship between the exponential distribution and the Poisson distribution. Suppose that the time that elapses between two successive events follows the exponential distribution with a mean of μ units of time. Also assume that these times are independent, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit time follows a Poisson distribution with mean μ. Recall that if X has the Poisson distribution with mean μ, then $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$.

The formula for the exponential distribution: $P(X = x) = me^{-mx} = \frac{1}{\mu} e^{-\frac{1}{\mu} x}$ Where $m =$ the rate parameter, or $\mu =$ average time between occurrences.

We see that the exponential is the cousin of the Poisson distribution and they are linked through this formula. There are important differences that make each distribution relevant for different types of probability problems.

First, the Poisson has a discrete random variable, X, where time; a continuous variable is artificially broken into discrete pieces. We saw that the number of occurrences of an event in a given time interval, X, follows the Poisson distribution.

For example, the **number** of times the telephone rings per hour. By contrast, the time **between** occurrences follows the exponential distribution. For example. The telephone just rang, how long will it be until it rings again? We are measuring length of time of the interval, a continuous random variable, exponential, not events during an interval, Poisson.

The Exponential Distribution v. the Poisson Distribution

A visual way to show both the similarities and differences between these two distributions is with a time line.

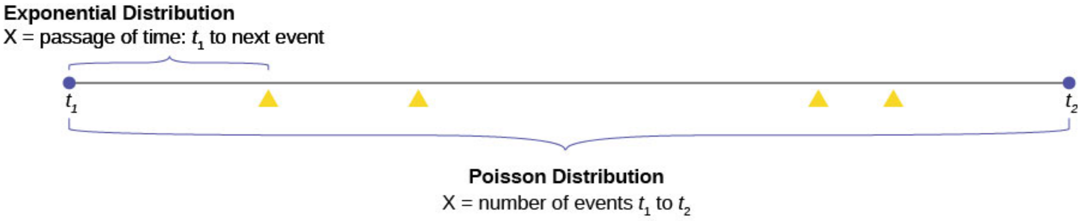

Exponential Distribution
X = passage of time: t_1 to next event
t_1
t_2
Poisson Distribution
X = number of events t_1 to t_2

Figure 4.7.8

The random variable for the Poisson distribution is discrete and thus counts events during a given time period, t_1 to t_2 on Figure 4.7.8, and calculates the probability of that number occurring. The number of events, four in the graph, is measured in counting numbers; therefore, the random variable of the Poisson is a discrete random variable.

The exponential probability distribution calculates probabilities of the passage of time, a continuous random variable. In Figure 4.7.8 this is shown as the bracket from t_1 to the next occurrence of the event marked with a triangle.

Classic Poisson distribution questions are "how many people will arrive at my checkout window in the next hour?".

Classic exponential distribution questions are "how long it will be until the next person arrives," or a variant, "how long will the person remain here once they have arrived?".

Again, the formula for the exponential distribution is:

$$f(x)=\lambda e^{-\lambda x} \text{ or } f(x)=\frac{1}{\mu} e^{-\frac{1}{\mu} x}$$

We see immediately the similarity between the exponential formula and the Poisson formula.

$$P(x)=\frac{\mu^{x} e^{-\mu}}{x!}$$

Both probability density functions are based upon the relationship between time and exponential growth or decay. The “e” in the formula is a constant with the approximate value of 2.71828 and is the base of the natural logarithmic exponential growth formula. When people say that something has grown exponentially this is what they are talking about.

An example of the exponential and the Poisson will make clear the differences been the two. It will also show the interesting applications they have.

Poisson Distribution

Suppose that historically 10 customers arrive at the checkout lines each hour. Remember that this is still probability so we have to be told these historical values. We see this is a Poisson probability problem.

We can put this information into the Poisson probability density function and get a general formula that will calculate the probability of **any** specific number of customers arriving in the next hour.

The formula is for any value of the random variable we chose, and so the x is put into the formula. This is the formula:

$$f(x)=\frac{10^{x} e^{-10}}{x!}$$

As an example, the probability of 15 people arriving at the checkout counter in the next hour would be

$$P(x=15)=\frac{10^{15} e^{-10}}{15!}=0.0611$$

Here we have inserted x = 15 and calculated the probability that in the next hour 15 people will arrive is .061.

Exponential Distribution

If we keep the same historical facts that 10 customers arrive each hour, but we now are interested in the service time a person spends at the counter, then we would use the exponential distribution. The exponential probability function for any value of x, the random variable, for this particular checkout counter historical data is:

$$f(x) = \frac{1}{.1} e^{-x/1} = 10e^{-10x}$$

To calculate μ, the historical average service time, we simply divide the number of people that arrive per hour, 10 , into the time period, one hour, and have $\mu = 0.1$. Historically, people spend 0.1 of an hour at the checkout counter, or 6 minutes. This explains the .1 in the formula.

There is a natural confusion with μ in both the Poisson and exponential formulas. They have different meanings, although they have the same symbol. The mean of the exponential is one divided by the mean of the Poisson. If you are given the historical number of arrivals you have the mean of the Poisson. If you are given an historical length of time between events you have the mean of an exponential.

Continuing with our example at the checkout clerk; if we wanted to know the probability that a person would spend 9 minutes or less checking out, then we use this formula. First, we convert to the same time units which are parts of one hour. Nine minutes is 0.15 of one hour. Next we note that we are asking for a range of values. This is always the case for a continuous random variable. We write the probability question as:

$$p(x \leq 9) = 1 - 10e^{-10x}$$

We can now put the numbers into the formula and we have our result.

$$p(x = .15) = 1 - 10e^{-10(.15)} = 0.7769$$

The probability that a customer will spend 9 minutes or less checking out is 0.7769.

We see that we have a high probability of getting out in less than nine minutes and a tiny probability of having 15 customers arriving in the next hour.

4.8: Introduction to Normal Distribution

The normal probability density function, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

Figure 4.8.1 If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

The normal distribution is extremely important, but it cannot be applied to everything in the real world. Remember here that we are still talking about the distribution of population data. This is a discussion of probability and thus it is the population data that may be normally distributed, and if it is, then this is how we can find probabilities of specific events just as we did for population data that may be binomially distributed or Poisson distributed. This caution is here because in the next chapter we will see that the normal distribution describes something very different from raw data and forms the foundation of inferential statistics.

The normal distribution has two parameters (two numerical descriptive measures): the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing the following formula of the normal probability density function:

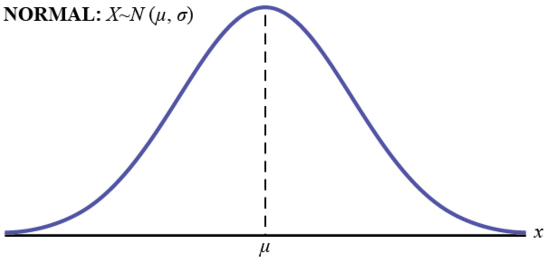


Figure 4.8.2

The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

$$f(x)=\frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \mathrm{e}^{-\frac{1}{2} \cdot\left(\frac{x-\mu}{\sigma}\right)^{2}}$$

The curve is symmetric about a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Note that this is unlike several probability density functions we have already studied, such as the Poisson, where the mean is equal to μ and the standard deviation simply the square root of the mean, or the binomial, where p is used to determine both the mean and standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the normal curve; the curve becomes fatter and wider or skinnier and taller depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

4.8.1: The Standard Normal Distribution

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.**

The mean for the standard normal distribution is 0, and the standard deviation is 1. What this does is dramatically simplify the mathematical calculation of probabilities. Take a moment and substitute 0 and 1 in the appropriate places in the above formula and you can see that the equation collapses into one that can be much more easily solved using integral calculus. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0,1)$. The value x in the given equation comes from a known normal distribution with known mean μ and known standard deviation σ. The z-score tells how many standard deviations a particular x is away from the mean.

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score for a particular value x is:

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

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.

Example 4.8.1.1

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:

$$z = \frac{x-\mu}{\sigma} = \frac{17-5}{6} = 2$$

This means that $x = 17$ is **two standard deviations** (2σ) above or to the right of the mean $\mu = 5$.

Now suppose $x = 1$. Then: $z = \frac{x-\mu}{\sigma} = \frac{1-5}{6} = -0.67$ (rounded to two decimal places)

This means that $\mathbf{x = 1}$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$.

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** states the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between -3σ and $+3\sigma$ 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 $+1\sigma$ and -1σ are $+1$ and -1, respectively.
- The z-scores for $+2\sigma$ and -2σ are $+2$ and -2, respectively.
- The z-scores for $+3\sigma$ and -3σ are $+3$ and -3 respectively.

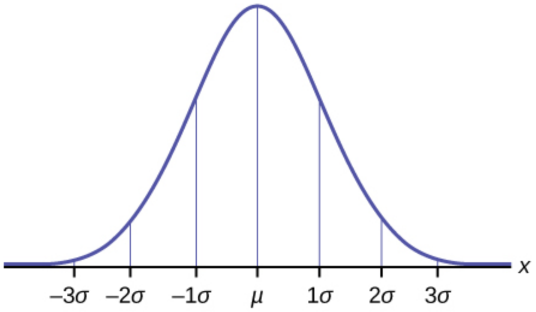


Figure 4.8.1.1

? Example 4.8.1.2

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 $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ 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 $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. 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 $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations from the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

4.8.2: Using the Normal Distribution

The shaded area in the following graph indicates the area to the right of x_1. This area is represented by the probability $P(X > x_1)$. Some normal tables provide the probability between the mean, 0 for the standard normal distribution, and a specific value such as x_1. This is the unshaded part of the graph from the mean to x_1.

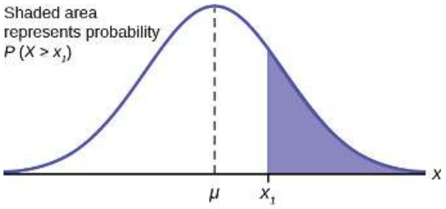


Figure 4.8.2.1

Because the normal distribution is symmetrical , if x_1 were the same distance to the left of the mean the area, probability, in the left tail, would be the same as the shaded area in the right tail. Also, bear in mind that because of the symmetry of this distribution, 0.5 of the probability is to the right of the mean and 0.5 is to the left of the mean.

Calculations of Probabilities

To find the probability for probability density functions with a continuous random variable we need to calculate the area under the function across the values of X we are interested in. For the normal distribution this seems a difficult task given the complexity of the formula. There is, however, a simply way to get what we want. Here again is the formula for the normal distribution:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \mathrm{e}^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

Looking at the formula for the normal distribution it is not clear just how we are going to solve for the probability doing it the same way we did it with the previous probability functions. There we put the data into the formula and did the math.

To solve this puzzle we start knowing that the area under a probability density function is the probability.

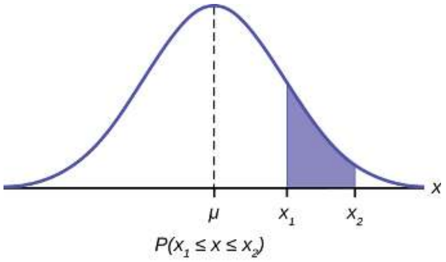


Figure 4.8.2.2

This shows that the area between x_1 and x_2 is the probability as stated in the formula: $P(x_1 \leq X \leq x_2)$

The mathematical tool needed to find the area under a curve is integral calculus. The integral of the normal probability density function between the two points x_1 and x_2 is the area under the curve between these two points and is the probability between these two points.

Doing these integrals is no fun and can be very time consuming. But now, remembering that there are an infinite number of normal distributions out there, we can consider the one with a mean of 0 and a standard deviation of 1. This particular normal distribution is given the name Standard Normal Distribution. Putting these values into the formula it reduces to a very simple equation. We can now quite easily calculate all probabilities for any value of x, for this particular normal distribution, that has a mean of 0 and a standard deviation of 1. These have been produced and are available here in the appendix to the text or everywhere on the web. They are presented in various ways. The table used in this text is the most common presentation and is set up with probabilities for half of the distribution beginning with 0, the mean, and moving outward. The shaded area in the graph at the top of the table in Statistical Tables represents the probability from zero to the specific z value noted on the horizontal axis.

The only problem is that even with this table, it would be a ridiculous coincidence that our data had a mean of 0 and a standard deviation of 1. The solution is to convert the distribution we have with its mean and standard deviation to this new Standard Normal Distribution. The Standard Normal has a random variable called Z.

Using the standard normal table, typically called the normal table, to find the probability of one standard deviation, go to the z column, reading down to 1.0 and then read at column 0. That number, 0.3413 is the probability from 0 to 1 standard deviation. At the top of the table is the shaded area in the distribution which is the probability for one standard deviation. The table has solved our integral calculus problem. But only if our data has a mean of 0 and a standard deviation of 1.

However, the essential point here is, the probability for one standard deviation on one normal distribution is the same on every normal distribution. If the population data set has a mean of 10 and a standard deviation of 5 then the probability from 10 to 15, one standard deviation, is the same as from 0 to 1, one standard deviation on the standard normal distribution. To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution and look up the answer in the tables. As review, here again is the **standardizing formula**:

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

where z is the value on the standard normal distribution, x is the value from a normal distribution one wishes to convert to the standard normal, μ and σ are, respectively, the mean and standard deviation of that population. Note that the equation uses μ and σ which denotes population parameters. This is still dealing with probability so we always are dealing with the population, with **known** parameter values and a **known** distribution. It is also important to note that because the normal distribution is symmetrical it does not matter if the z-score is positive or negative when calculating a probability. One standard deviation to the left (negative z-score) covers the same area as one standard deviation to the right (positive z-score). This fact is why the Standard Normal tables do not provide areas for the left side of the distribution. Because of this symmetry, the z-score formula is sometimes written as:

$$z=\frac{|x-\mu|}{\sigma}$$

Where the vertical lines in the equation means the absolute value of the number.

What the standardizing formula is really doing is computing the number of standard deviations x is from the mean of its own distribution. The standardizing formula and the concept of counting standard deviations from the mean is the secret of all that we will do in this statistics class. The reason this is true is that **all** of statistics boils down to variation, and the counting of standard deviations is a measure of variation.

This formula, in many disguises, will reappear over and over throughout this course.

? Example 4.8.2.1

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

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.

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 $P(X>65)$.

$P(X>65)=0.3446$

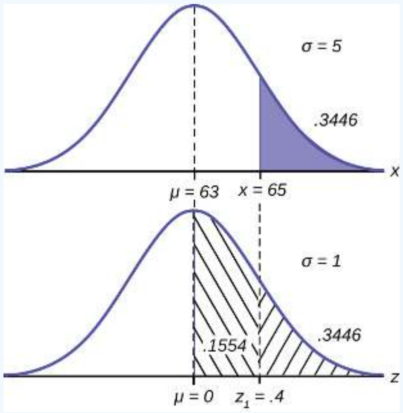


Figure 4.8.2.3

$$z_1=\frac{x_1-\mu}{\sigma}=\frac{65-63}{5}=0.4$$

$P\left(X \geq x_1\right)=P\left(Z \geq z_1\right)=0.3446$

The probability that any student selected at random scores more than 65 is 0.3446. Here is how we found this answer.

Answer b

The normal table provides probabilities from zero to the value z_1. For this problem the question can be written as: $P(X \geq 65)=P(Z \geq z_1)$, which is the area in the tail. To find this area the formula would be $0.5 - P(X \leq 65)$. One half of the probability is above the mean value because this is a symmetrical distribution. The graph shows how to find the area in the tail by subtracting that portion from the mean, zero, to the z_1 value. The final answer is: $P(X \geq 63)=P(Z \geq 0.4)=0.3446$

$z_1=\frac{65-63}{5}=0.4$

Area to the left of z_1 to the mean of zero is 0.1554

$P(X>65)=P(Z>0.4)=0.5–0.1554=0.3446$

$z=\frac{x-\mu}{\sigma}=\frac{85-63}{5}=4.4$ which is larger than the maximum value on the Standard Normal Table. Therefore, the probability that one student scores less than 85 is approximately one or 100%.

A score of 85 is 4.4 standard deviations from the mean of 63 which is beyond the range of the standard normal table. Therefore, the probability that one student scores less than 85 is approximately one (or 100%).

? Exercise $4.8.2.1$

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.

? Example $4.8.2.2A$

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.

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 $X=2.75$. $P(1.8<X<2.75)=0.5886$

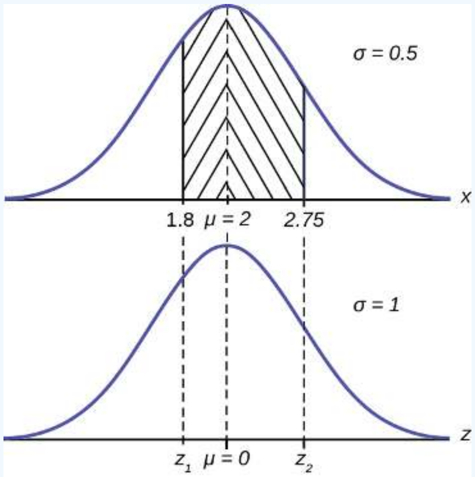


Figure 4.8.2.4

$P(1.8\leq X\leq 2.75)=P(Z_1\leq Z\leq Z_2)$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

? Example $4.8.2.2B$

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer

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 $P(X<k)=0.25$.

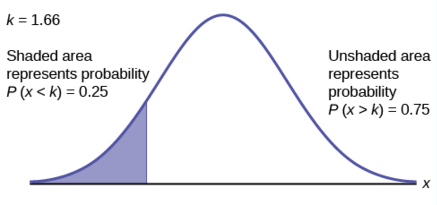


Figure 4.8.2.5

$f(z)=0.5-0.25=0.25,$ therefore $z\approx -0.675($ or just 0.67 using the table$)$ $z=\frac{x-\mu}{\sigma}=\frac{x-2}{0.5}=-0.675,$ therefore $x=-0.675*0.5+2=1.66$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

? Exercise 4.8.2.2

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 golfer scored between 66 and 70.

? Example 4.8.2.3

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.

Answer

Answer

a. 0.8186

b. 0.8413

? Example 4.8.2.4

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.

Answer

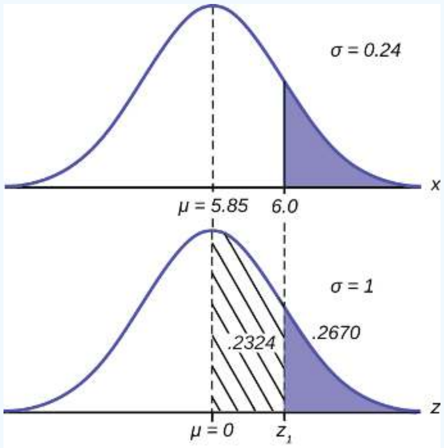


Figure 4.8.2.6

$$z_1=\frac{6-5.85}{.24}=.625$$

$P(X\geq 6)=P(Z\geq 0.625)=0.2670$

b. The middle 20% of mandarin oranges from this farm have diameters between ______ and ______.

$f(z)=\frac{0.20}{2}=0.10,$ therefore $z\approx\pm 0.25$

$z=\frac{x-\mu}{\sigma}=\frac{x-5.85}{0.24}=\pm 0.25\rightarrow\pm 0.25\cdot 0.24+5.85=(5.79,5.91)$

4.9: Normal Distribution - Pinkie Length (Worksheet)

Name: ______________________________

Section: _____________________________

Student ID#:__________________________

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if data from the experiment follow a continuous distribution.

Collect the Data

Measure the length of your pinky finger (in centimeters).

1. Randomly survey 30 adults for their pinky finger lengths. Round the lengths to the nearest 0.5 cm.

_______	_______	_______	_______	_______
_______	_______	_______	_______	_______
_______	_______	_______	_______	_______
_______	_______	_______	_______	_______
_______	_______	_______	_______	_______
_______	_______	_______	_______	_______

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

Blank graph with frequency on the vertical axis and length of finger on the horizontal axis.

Figure 6.5.1.

3. Calculate the following.
   a. $\bar{x}$ = _______
   b. s = _______
4. Draw a smooth curve through the top of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram, what was the approximate theoretical distribution of the data you collected?

- $X \sim$ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Using the data you collected complete the following statements. (Hint: order the data)

REMEMBER

($IQR = Q_3 - Q_1$)

- IQR = _______
- The 15th percentile is _______.
- The 85th percentile is _______.

- Median is _______.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. Use a normal approximation based on the sample mean and standard deviation.

- IQR = _______
- The 15th percentile is _______.
- The 85th percentile is _______.
- Median is _______.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Discussion Questions

Do the data you collected give a close approximation to the theoretical distribution? In complete sentences and comparing the results in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

4.10: Chapter Key Items

Key Terms	Definition
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 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.
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 mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.
Geometric Distribution	a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula: $P(X = x) = p(1 - p)^{x-1}$ where one wants to know probability for the number of trials until the first success: the xth trail is the first success. An alternative formulation of the geometric distribution asks the question: what is the probability of x failures until the first success? In this formulation the trial that resulted in the first success is not counted. The formula for this presentation of the geometric is: $P(X = x) = p(1 - p)^x$ The expected value in this form of the geometric distribution is $\mu = \frac{1-p}{p}$ The easiest way to keep these two forms of the geometric distribution straight is to remember that p is the probability of success and $(1 - p)$ is the probability of failure. In the formula the exponents simply count the number of successes and number of failures of the desired outcome of the experiment. Of course the sum of these two numbers must add to the number of trials in the experiment.

Key Terms	Definition
Geometric Experiment	a statistical experiment with the following properties: 1. There are one or more Bernoulli trials with all failures except the last one, which is a success. 2. In theory, the number of trials could go on forever. There must be at least one trial. 3. The probability, p, of a success and the probability, q, of a failure do not change from trial to trial.
Hypergeometric Experiment	a statistical experiment with the following properties: 1. You take samples from two groups. 2. You are concerned with a group of interest, called the first group. 3. You sample without replacement from the combined groups. 4. Each pick is not independent, since sampling is without replacement.
Hypergeometric Probability	a discrete random variable (RV) that is characterized by: 1. A fixed number of trials. 2. The probability of success is not the same from trial to trial. We sample from two groups of items when we are interested in only one group. X is defined as the number of successes out of the total number of items chosen.
Poisson Probability Distribution	a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable: The probability that the event occurs in a given interval is the same for all intervals. The events occur with a known mean and independently of the time since the last event. The distribution is defined by the mean μ of the event in the interval. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(x) = \frac{\mu^x e^{-\mu}}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is “large” and p is “small” (a general rule is that np should be greater than or equal to 25 and p should be less than or equal to 0.01).
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.

Key Terms	Definition
Random Variable (RV)	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, and 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 the random variable X takes only after performing the experiment.

4.10.1: Chapter Key Terms

Key Terms	Definition
Conditional Probability	the likelihood that an event will occur given that another event has already occurred.
decay parameter	The decay parameter describes the rate at which probabilities decay to zero for increasing values of x. It is the value m in the probability density function $f(x)=me^{(-mx)}$ of an exponential random variable. It is also equal to $m=\frac{1}{\mu}$, where μ is the mean of the random variable.
Exponential Distribution	a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. The mean is $\mu=\frac{1}{m}$ and the standard deviation is $\sigma=\frac{1}{m}$. The probability density function is $f(x)=me^{-mx}$ or $f(x)=\frac{1}{\mu}e^{-\frac{1}{\mu}x}, x\geq 0$ and the cumulative distribution function is $P(X\leq x)=1-e^{-mx}$ or $P(X\leq x)=1-e^{-\frac{1}{\mu}x}$.
memoryless property	For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds $x+t$, given that it has exceeded x, is the same as the probability that X would exceed t if we had no knowledge about it. In symbols we say that $P(X>x+t \mid X>x)=P(X>t)$.
Poisson distribution	If there is a known average of \mu events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution. The probability of x events occurring in one unit time is equal to $P(X=x)=\frac{\mu^{x}e^{-\mu}}{x!}$.
Uniform Distribution	a continuous random variable (RV) that has equally likely outcomes over the domain, $a<x<b$; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. The mean is $\mu=\frac{a+b}{2}$ and the standard deviation is $\sigma=\sqrt{\frac{(b-a)^{2}}{12}}$. The probability density function is \(f(x)=\frac{1}{b-a} \text { for } a

4.10.2: Chapter Key Items

Key Terms	Definition
Normal Distribution	a continuous random variable (RV) with pdf $f(x) =$ $$\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}}$$ where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV, Z, is called the **standard normal distribution**.
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 $z = \frac{x-\mu}{\sigma}$ or written as $z = \frac{\lvert x-\mu\rvert}{\sigma}$; 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. A z-score is the number of standard deviations a particular x is away from its mean value.

- **6.5: Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

4.11: Chapter Review

4.1 Introduction

The characteristics of a probability distribution or density function (PDF) 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.

4.2 Hypergeometric Distribution

The combinatorial formula can provide the number of unique subsets of size x that can be created from n unique objects to help us calculate probabilities. The combinatorial formula is $\left(\begin{array}{l}n\\x\end{array}\right) ={}_n C_x = \frac{n!}{x!(n-x)!}$

A **hypergeometric experiment** is a statistical experiment with the following properties:

1. You take samples from two groups.
2. You are concerned with a group of interest, called the first group.
3. You sample without replacement from the combined groups.
4. Each pick is not independent, since sampling is without replacement.

The outcomes of a hypergeometric experiment fit a hypergeometric probability distribution. The random variable $X =$ the number of items from the group of interest. $h(x)=\frac{\left(\begin{array}{l}A\\x\end{array}\right)\left(\begin{array}{l}N-A\\n-x\end{array}\right)}{\left(\begin{array}{l}N\\n\end{array}\right)}$.

4.3 Binomial Distribution

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

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.

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 $\sigma = \sqrt{npq}$.

The formula for the Binomial probability density function is

$$P(x)=\frac{n!}{x!(n-x)!}\cdot p^{x} q^{(n-x)}$$

4.4 Geometric Distribution

There are three characteristics of a geometric experiment:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p, of a success and the probability, q, of a failure are the same for each trial.

In a geometric experiment, define the discrete random variable X as the number of independent trials until the first success. We say that X has a geometric distribution and write $X \sim G(p)$ where p is the probability of success in a single trial.

The mean of the geometric distribution $X \sim G(p)$ is $\mu = 1/p$ where $x =$ number of trials until first success for the formula $P(X=x)=(1-p)^{x-1}p$ where the number of trials is up and including the first success.

An alternative formulation of the geometric distribution asks the question: what is the probability of x failures until the first success? In this formulation the trial that resulted in the first success is not counted. The formula for this presentation of the geometric is:

$$P(X=x)=p(1-p)^x$$

The expected value in this form of the geometric distribution is

$$\mu=\frac{1-p}{p}$$

The easiest way to keep these two forms of the geometric distribution straight is to remember that p is the probability of success and $(1-p)$ is the probability of failure. In the formula the exponents simply count the number of successes and number of failures of the desired outcome of the experiment. Of course the sum of these two numbers must add to the number of trials in the experiment.

4.5 Poisson Distribution

A **Poisson probability distribution** of a discrete random variable gives the probability of a number of events occurring in a fixed interval of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to approximate the binomial, if the probability of success is "small" (less than or equal to 0.01) and the number of trials is "large" (greater than or equal to 25). Other rules of thumb are also suggested by different authors, but all recognize that the Poisson distribution is the limiting distribution of the binomial as n increases and p approaches zero.

The formula for computing probabilities that are from a Poisson process is:

$$P(x)=\frac{\mu^x e^{-\mu}}{x!}$$

where $P(X)$ is the probability of successes, μ (pronounced mu) is the expected number of successes, e is the natural logarithm approximately equal to 2.718, and X is the number of successes per unit, usually per unit of time.

4.12: Chapter References

Binomial Distribution

- “Access to electricity (% of population),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...first&sort=asc (accessed May 15, 2015).
- “Distance Education.” Wikipedia. Available online at http://en.Wikipedia.org/wiki/Distance_education (accessed May 15, 2013).
- “NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).
- 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).
- 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).
- “The World FactBook,” Central Intelligence Agency. Available online at https://www.cia.gov/library/publicat...k/geos/af.html (accessed May 15, 2013).
- “What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at http://www.cancer.org/cancer/pancrea...key-statistics (accessed May 15, 2013).

Geometric Distribution

- “Millennials: A Portrait of Generation Next,” PewResearchCenter. Available online at http://www.pewsocialtrends.org/files...-to-change.pdf (accessed May 15, 2013).
- “Millennials: Confident. Connected. Open to Change.” Executive Summary by PewResearch Social & Demographic Trends, 2013. Available online at http://www.pewsocialtrends.org/2010/...pen-to-change/ (accessed May 15, 2013).
- “Prevalence of HIV, total (% of populations ages 15-49),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...last&sort=desc (accessed May 15, 2013).
- 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).
- “Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan,” The European Union and ICON-Institute. Available online at ec.europa.eu/europeaid/where/...summary_en.pdf (accessed May 15, 2013).
- “The World FactBook,” Central Intelligence Agency. Available online at https://www.cia.gov/library/publicat...k/geos/af.html (accessed May 15, 2013).
- “UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] and writing skills,” UNICEF Television. Video available online at http://www.unicefusa.org/assets/vide...y-centers.html (accessed May 15, 2013).

Poisson Distribution

- “ATL Fact Sheet,” Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at www.atl.com/about-atl/atl-factsheet/ (accessed February 6, 2019).
- Center for Disease Control and Prevention. “Teen Drivers: Fact Sheet,” Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafet...factsheet.html (accessed May 15, 2013).
- “Children and Childrearing,” Ministry of Health, Labour, and Welfare. Available online at http://www.mhlw.go.jp/english/policy...ing/index.html (accessed May 15, 2013).
- “Eating Disorder Statistics,” South Carolina Department of Mental Health, 2006. Available online at http://www.state.sc.us/dmh/anorexia/statistics.htm (accessed May 15, 2013).
- “Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day,” theguardian, 2013. Available online at http://www.theguardian.com/world/gal...471900&index=2 (accessed May 15, 2013).
- “How Americans Use Text Messaging,” Pew Internet, 2013. Available online at http://pewinternet.org/Reports/2011/...in-Report.aspx (accessed May 15, 2013).

- Lenhart, Amanda. “Teens, Smartphones & Testing: Texting volum is up while the frequency of voice calling is down. About one in four teens say they own smartphones,” Pew Internet, 2012. Available online at www.pewinternet.org/~/media/F...nd_Texting.pdf (accessed May 15, 2013).
- “One born every minute: the maternity unit where mothers are THREE to a bed,” MailOnline. Available online at http://www.dailymail.co.uk/news/arti...thers-bed.html (accessed May 15, 2013).
- Vanderkam, Laura. “Stop Checking Your Email, Now.” CNNMoney, 2013. Available online at management.fortune.cnn.com/20...our-email-now/ (accessed May 15, 2013).
- “World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. www.world-earthquakes.com/ind...thq_prediction (accessed May 15, 2013).

4.12.1: Chapter References

5.3 The Uniform Distribution

McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

5.4 The Exponential Distribution

Data from the United States Census Bureau.

Data from World Earthquakes, 2013. Available online at http://www.world-earthquakes.com/ (accessed June 11, 2013).

“No-hitter.” Baseball-Reference.com, 2013. Available online at http://www.baseball-reference.com/bullpen/No-hitter (accessed June 11, 2013).

Zhou, Rick. “Exponential Distribution lecture slides.” Available online at www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf (accessed June 11, 2013).

4.12.2: Chapter References

The Standard Normal Distribution

- “Blood Pressure of Males and Females.” StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewre...reportid=11960 (accessed May 14, 2013).
- “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).
- “2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at http://media.collegeboard.com/digita...Group-2012.pdf (accessed May 14, 2013).
- “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 http://nces.ed.gov/programs/digest/d...s/dt09_147.asp (accessed May 14, 2013).
- Data from the *San Jose Mercury News*.
- Data from *The World Almanac and Book of Facts*.
- “List of stadiums by capacity.” Wikipedia. Available online at https://en.Wikipedia.org/wiki/List_o...ms_by_capacity (accessed May 14, 2013).
- Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Using the Normal Distribution

- “Naegele’s rule.” Wikipedia. Available online at http://en.Wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).
- “403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radi...sode/403/nummi (accessed May 14, 2013).
- “Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at www.winatthelottery.com/publi...partment40.cfm (accessed May 14, 2013).
- “Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).
- “Facebook Statistics.” Statistics Brain. Available online at http://www.statisticbrain.com/facebo...tics/(accessed May 14, 2013).

4.H: Random Variables (Homework)

4.2 Hypergeometric Distribution

47. A group of Martial Arts students is planning on participating in an upcoming demonstration. Six are students of Tae Kwon Do; seven are students of Shotokan Karate. Suppose that eight students are randomly picked to be in the first demonstration. We are interested in the number of Shotokan Karate students in that first demonstration

Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

53. Define the random variable and list its possible values.

54. State the distribution of X.

55. Find the probability that at least four of the 25 patients actually have the flu.

56. On average, for every 25 patients calling in, how many do you expect to have the flu?

57. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given Table $4.H.5$. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

Table $4.H.1$

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

Use the following information to answer the next two exercises: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.59.

The expected number of wins for that upcoming month is:

Let X = the number of games won in that upcoming month.

60.What is the probability that the San Jose Sharks win six games in that upcoming month?

Use the following information to answer the next two exercises: The average number of times per week that Mrs. Plum's cats wake her up at night because they want to play is ten. We are interested in the number of times her cats wake her up each week.**93**.

In words, the random variable $X =$ _________________

94. Find the probability that her cats will wake her up no more than five times next week.

1. the number of times Mrs. Plum's cats wake her up each week.
2. the number of times Mrs. Plum's cats wake her up each hour.
3. the number of times Mrs. Plum's cats wake her up each night.
4. the number of times Mrs. Plum's cats wake her up.
5. 0.5000

6. 0.9329
7. 0.0378
8. 0.0671

1.1: Discrete RVs Practice

1.2: Uniform and Exponential Homework

5.2 Properties of Continuous Probability Density Functions

For each probability and percentile problem, draw the picture.

70. Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer “yes” or “no.” You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.

1. What part of the experiment will yield discrete data?
2. What part of the experiment will yield continuous data?

71. When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

5.3 The Uniform Distribution

For each probability and percentile problem, draw the picture.

72. Births are approximately uniformly distributed between the 52 weeks of the year. They can be said to follow a uniform distribution from one to 53 (spread of 52 weeks).

a. Graph the probability distribution.
b. $f(x) =$ _________
c. $\mu =$ _________
d. $\sigma =$ _________
e. Find the probability that a person is born at the exact moment week 19 starts. That is, find $P(x = 19) =$ _________
f. $P(2 < x < 31) =$ _________
g. Find the probability that a person is born after week 40.
h. $P(12 < x | x < 28) =$ _________

73. A random number generator picks a number from one to nine in a uniform manner.

a. Graph the probability distribution.
b. $f(x) =$ _________
c. $\mu =$ _________
d. $\sigma =$ _________
e. $P(3.5 < x < 7.25) =$ _________
f. $P(x > 5.67)$
g. $P(x > 5 | x > 3) =$ _________

74. According to a study by Dr. John McDougall of his live-in weight loss program at St. Helena Hospital, the people who follow his program lose between six and 15 pounds a month until they approach trim body weight. Let’s suppose that the weight loss is uniformly distributed. We are interested in the weight loss of a randomly selected individual following the program for one month.

a. Define the random variable. $X =$ _________
b. Graph the probability distribution.
c. $f(x) =$ _________
d. $\mu =$ _________
e. $\sigma =$ _________
f. Find the probability that the individual lost more than ten pounds in a month.
g. Suppose it is known that the individual lost more than ten pounds in a month. Find the probability that he lost less than 12 pounds in the month.
h. $P(7 < x < 13 | x > 9) =$ __________. State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

75. A subway train on the Red Line arrives every eight minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution.

a. Define the random variable. $X =$ _______

b. Graph the probability distribution.
c. $f(x) =$ _______
d. $\mu =$ _______
e. $\sigma =$ _______
f. Find the probability that the commuter waits less than one minute.
g. Find the probability that the commuter waits between three and four minutes.

76. The age of a first grader on September 1 at Garden Elementary School is uniformly distributed from 5.8 to 6.8 years. We randomly select one first grader from the class.

a. Define the random variable. $X =$ _________
b. Graph the probability distribution.
c. $f(x) =$ _________
d. $\mu =$_________
e. $\sigma =$ _________
f. Find the probability that she is over 6.5 years old.
g. Find the probability that she is between four and six years old.

Use the following information to answer the next three exercises. The Sky Train from the terminal to the rental–car and long–term parking center is supposed to arrive every eight minutes. The waiting times for the train are known to follow a uniform distribution.

77. What is the average waiting time (in minutes)?

a. zero
b. two
c. three
d. four

78. The probability of waiting more than seven minutes given a person has waited more than four minutes is?

a. 0.125
b. 0.25
c. 0.5
d. 0.75

79. The time (in minutes) until the next bus departs a major bus depot follows a distribution with *f*(*x*) = 120120 where *x* goes from 25 to 45 minutes.

a. Define the random variable. $X =$ ________
b. Graph the probability distribution.
c. The distribution is ______________ (name of distribution). It is _____________ (discrete or continuous).
d. $\mu =$ ________
e. $\sigma =$ ________
f. Find the probability that the time is at most 30 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
g. Find the probability that the time is between 30 and 40 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
h. $P(25 < x < 55) =$ _________. State this in a probability statement, similarly to parts g and h, draw the picture, and find the probability.

80. Suppose that the value of a stock varies each day from \$16 to \$25 with a uniform distribution.

a. Find the probability that the value of the stock is more than \$19.
b. Find the probability that the value of the stock is between \$19 and \$22.
c. Given that the stock is greater than \$18, find the probability that the stock is more than \$21.

81. A fireworks show is designed so that the time between fireworks is between one and five seconds, and follows a uniform distribution.

a. Find the average time between fireworks.

b. Find probability that the time between fireworks is greater than four seconds.

82. The number of miles driven by a truck driver falls between 300 and 700, and follows a uniform distribution.

a. Find the probability that the truck driver goes more than 650 miles in a day.
b. Find the probability that the truck drivers goes between 400 and 650 miles in a day.

5.4 The Exponential Distribution

83. Suppose that the length of long distance phone calls, measured in minutes, is known to have an exponential distribution with the average length of a call equal to eight minutes.

a. Define the random variable. $X =$ ________________.
b. Is X continuous or discrete?
c. $\mu =$ ________
d. $\sigma =$ ________
e. Draw a graph of the probability distribution. Label the axes.
f. Find the probability that a phone call lasts less than nine minutes.
g. Find the probability that a phone call lasts more than nine minutes.
h. Find the probability that a phone call lasts between seven and nine minutes.
i. If 25 phone calls are made one after another, on average, what would you expect the total to be? Why?

84. Suppose that the useful life of a particular car battery, measured in months, decays with parameter 0.025. We are interested in the life of the battery.

a. Define the random variable. $X =$ ________________________________.
b. Is X continuous or discrete?
c. On average, how long would you expect one car battery to last?
d. On average, how long would you expect nine car batteries to last, if they are used one after another?
e. Find the probability that a car battery lasts more than 36 months.
f. Seventy percent of the batteries last at least how long?

85. The percent of persons (ages five and older) in each state who speak a language at home other than English is approximately exponentially distributed with a mean of 9.848. Suppose we randomly pick a state.

a. Define the random variable. $X =$ ________________________________.
b. Is X continuous or discrete?
c. $\mu =$ ________
d. $\sigma =$ ________
e. Draw a graph of the probability distribution. Label the axes.
f. Find the probability that the percent is less than 12.
g. Find the probability that the percent is between eight and 14.
h. The percent of all individuals living in the United States who speak a language at home other than English is 13.8.
    - Why is this number different from 9.848%?
    - What would make this number higher than 9.848%?

86. The time (in years) **after** reaching age 60 that it takes an individual to retire is approximately exponentially distributed with a mean of about five years. Suppose we randomly pick one retired individual. We are interested in the time after age 60 to retirement.

a. Define the random variable. $X =$ ________________________________.
b. Is X continuous or discrete?
c. $\mu =$ ________
d. $\sigma =$ ________
e. Draw a graph of the probability distribution. Label the axes.
f. Find the probability that the person retired after age 70.
g. Do more people retire before age 65 or after age 65?
h. In a room of 1,000 people over age 80, how many do you expect will NOT have retired yet?

87. The cost of all maintenance for a car during its first year is approximately exponentially distributed with a mean of $150.

a. Define the random variable. $X =$ ________________________________.
b. $\mu =$ ________
c. $\sigma =$ ________
d. Draw a graph of the probability distribution. Label the axes.
e. Find the probability that a car required over $300 for maintenance during its first year.

Use the following information to answer the next three exercises. The average lifetime of a certain new cell phone is three years. The manufacturer will replace any cell phone failing within two years of the date of purchase. The lifetime of these cell phones is known to follow an exponential distribution.

88. The decay rate is:

a. 0.3333
b. 0.5000
c. 2
d. 3

89. What is the probability that a phone will fail within two years of the date of purchase?

a. 0.8647
b. 0.4866
c. 0.2212
d. 0.9997

90. What is the median lifetime of these phones (in years)?

a. 0.1941
b. 1.3863
c. 2.0794
d. 5.5452

91. At a 911 call center, calls come in at an average rate of one call every two minutes. Assume that the time that elapses from one call to the next has the exponential distribution.

a. On average, how much time occurs between five consecutive calls?
b. Find the probability that after a call is received, it takes more than three minutes for the next call to occur.
c. Ninety-percent of all calls occur within how many minutes of the previous call?
d. Suppose that two minutes have elapsed since the last call. Find the probability that the next call will occur within the next minute.
e. Find the probability that less than 20 calls occur within an hour.

92**.** In major league baseball, a no-hitter is a game in which a pitcher, or pitchers, doesn't give up any hits throughout the game. No-hitters occur at a rate of about three per season. Assume that the duration of time between no-hitters is exponential.

a. What is the probability that an entire season elapses with a single no-hitter?
b. If an entire season elapses without any no-hitters, what is the probability that there are no no-hitters in the following season?
c. What is the probability that there are more than 3 no-hitters in a single season?

93. During the years 1998–2012, a total of 29 earthquakes of magnitude greater than 6.5 have occurred in Papua New Guinea. Assume that the time spent waiting between earthquakes is exponential.

a. What is the probability that the next earthquake occurs within the next three months?
b. Given that six months has passed without an earthquake in Papua New Guinea, what is the probability that the next three months will be free of earthquakes?
c. What is the probability of zero earthquakes occurring in 2014?
d. What is the probability that at least two earthquakes will occur in 2014?

94. According to the American Red Cross, about one out of nine people in the U.S. have Type B blood. Suppose the blood types of people arriving at a blood drive are independent. In this case, the number of Type B blood types that arrive roughly follows the Poisson distribution.

a. If 100 people arrive, how many on average would be expected to have Type B blood?

b. What is the probability that over 10 people out of these 100 have type B blood?
c. What is the probability that more than 20 people arrive before a person with type B blood is found?

95. A web site experiences traffic during normal working hours at a rate of 12 visits per hour. Assume that the duration between visits has the exponential distribution.

a. Find the probability that the duration between two successive visits to the web site is more than ten minutes.
b. The top 25% of durations between visits are at least how long?
c. Suppose that 20 minutes have passed since the last visit to the web site. What is the probability that the next visit will occur within the next 5 minutes?
d. Find the probability that less than 7 visits occur within a one-hour period.

96. At an urgent care facility, patients arrive at an average rate of one patient every seven minutes. Assume that the duration between arrivals is exponentially distributed.

a. Find the probability that the time between two successive visits to the urgent care facility is less than 2 minutes.
b. Find the probability that the time between two successive visits to the urgent care facility is more than 15 minutes.
c. If 10 minutes have passed since the last arrival, what is the probability that the next person will arrive within the next five minutes?
d. Find the probability that more than eight patients arrive during a half-hour period.

1.3: Uniform and Exponential Practice

5.2 Properties of Continuous Probability Density Functions

1. Which type of distribution does the graph illustrate?

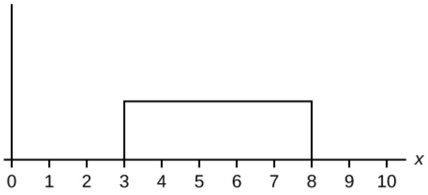


Figure 1.3.1

2. Which type of distribution does the graph illustrate?

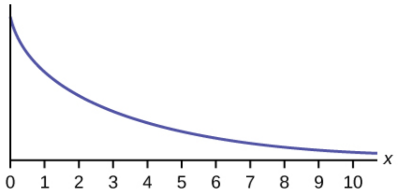


Figure 1.3.2

3. Which type of distribution does the graph illustrate?

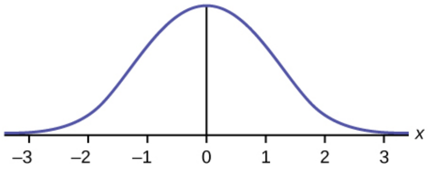


Figure 1.3.3

4. What does the shaded area represent? $P(\_\_\_ < x < \_\_\_)$

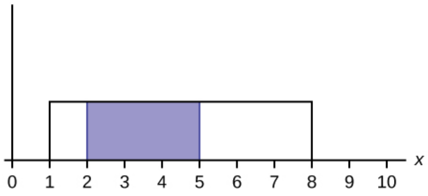


Figure 1.3.4

5. What does the shaded area represent? $P(\_\_\_ < x < \_\_\_)$

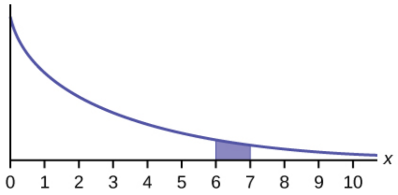


Figure 1.3.5

6. For a continuous probablity distribution, $0 \leq x \leq 15$. What is $P(x > 15)$?

7. What is the area under $f(x)$ if the function is a continuous probability density function?

8. For a continuous probability distribution, $0 \leq x \leq 10$. What is $P(x = 7)$?

9. A **continuous** probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?

10. $f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \leq x \leq 5$. What is $P(x < 0)$?

11. $f(x)$, a continuous probability function, is equal to $\frac{1}{12}$, and the function is restricted to $0 \leq x \leq 12$. What is $P(0 < x < 12)$?

12. Find the probability that x falls in the shaded area.

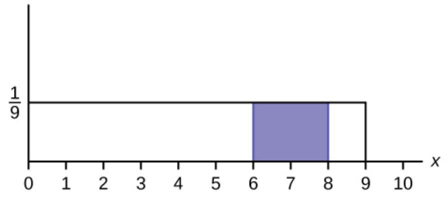


Figure 1.3.6

13. Find the probability that x falls in the shaded area.

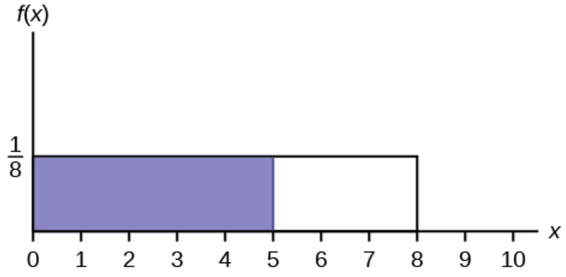


Figure 1.3.7

14. Find the probability that x falls in the shaded area.

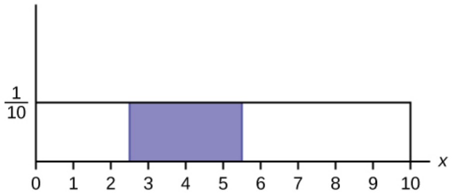


Figure 1.3.8

15. $f(x)$, a continuous probability function, is equal to $\frac{1}{3}$ and the function is restricted to $1 \leq x \leq 4$. Describe $P(x > \frac{3}{2})$.

5.3 The Uniform Distribution

Use the following information to answer the next ten questions. The data that follow are the square footage (in 1,000 feet squared) of 28 homes.

Table 1.3.2

1.5	2.4	3.6	2.6	1.6	2.4	2.0
3.5	2.5	1.8	2.4	2.5	3.5	4.0
2.6	1.6	2.2	1.8	3.8	2.5	1.5
2.8	1.8	4.5	1.9	1.9	3.1	1.6

The sample mean = 2.50 and the sample standard deviation = 0.8302.

The distribution can be written as $X \sim U(1.5, 4.5)$.

16. What type of distribution is this?

17. In this distribution, outcomes are equally likely. What does this mean?

18. What is the height of $f(x)$ for the continuous probability distribution?

19. What are the constraints for the values of x?

20. Graph $P(2 < x < 3)$.

21. What is $P(2 < x < 3)$?

22. What is $P(x < 3.5|x < 4)$?

23. What is $P(x = 1.5)$?

24. Find the probability that a randomly selected home has more than 3,000 square feet given that you already know the house has more than 2,000 square feet.

Use the following information to answer the next eight exercises. A distribution is given as $X \sim U(0, 12)$.

25. What is a? What does it represent?

26. What is b? What does it represent?

27. What is the probability density function?

28. What is the theoretical mean?

29. What is the theoretical standard deviation?

30. Draw the graph of the distribution for $P(x > 9)$.

31. Find $P(x > 9)$.

Use the following information to answer the next eleven exercises. The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

32. What is being measured here?

33. In words, define the random variable X.

34. Are the data discrete or continuous?

35. The interval of values for x is ______.

36. The distribution for X is ______.

37. Write the probability density function.

38. Graph the probability distribution.

1. Sketch the graph of the probability distribution.

Figure 1.3.9

2. Identify the following values:
    - Lowest value for $\overline{x}$: _______
    - Highest value for $\overline{x}$: _______
    - Height of the rectangle: _______
    - Label for x-axis (words): _______
    - Label for y-axis (words): _______

39. Find the average age of the cars in the lot.

40. Find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, and shade the area of interest.

Blank graph with vertical and horizontal axes.

Figure 1.3.10

b. Find the probability. $P(x<4)$ = _______

41. Considering only the cars less than 7.5 years old, find the probability that a randomly chosen car in the lot was less than four years old.

1. Sketch the graph, shade the area of interest.

This is a blank graph template. The vertical and horizontal axes are unlabeled.

Figure 1.3.11

2. Find the probability. $P(x<4 | x<7.5)=$ _______

42. What has changed in the previous two problems that made the solutions different?

43. Find the third quartile of ages of cars in the lot. This means you will have to find the value such that $\frac{3}{4}$, or 75%, of the cars are at most (less than or equal to) that age.

1. Sketch the graph, and shade the area of interest.

Figure 1.3.12

2. Find the value k such that $P(x<k)=0.75$.
3. The third quartile is _______

5.4 The Exponential Distribution

Use the following information to answer the next ten exercises. A customer service representative must spend different amounts of time with each customer to resolve various concerns. The amount of time spent with each customer can be modeled by the following distribution: $X \sim Exp(0.2)$

44. What type of distribution is this?

45.Are outcomes equally likely in this distribution? Why or why not?

46. What is m? What does it represent?

47. What is the mean?

48. What is the standard deviation?

49. State the probability density function.

50. Graph the distribution.

51. Find $P(2 < x < 10)$.

52. Find $P(x > 6)$.

53. Find the 70th percentile.

Use the following information to answer the next seven exercises. A distribution is given as $X \sim Exp(0.75)$.

54. What is *m*?

55. What is the probability density function?

56. What is the cumulative distribution function?

57. Draw the distribution.

58. Find $P(x < 4)$.

59. Find the 30th percentile.

60. Find the median.

61. Which is larger, the mean or the median?

Use the following information to answer the next 16 exercises. Carbon-14 is a radioactive element with a half-life of about 5,730 years. Carbon-14 is said to decay exponentially. The decay rate is 0.000121. We start with one gram of carbon-14. We are interested in the time (years) it takes to decay carbon-14.

62. What is being measured here?

63. Are the data discrete or continuous?

64. In words, define the random variable X.

65. What is the decay rate (m)?

66. The distribution for X is ______.

67. Find the amount (percent of one gram) of carbon-14 lasting less than 5,730 years. This means, find $P(x < 5,730)$.

1. Sketch the graph, and shade the area of interest.

Figure 1.3.13

2. Find the probability. $P(x < 5,730) =$ __________

68. Find the percentage of carbon-14 lasting longer than 10,000 years.

1. Sketch the graph, and shade the area of interest.

Figure 1.3.14

2. Find the probability. $P(x > 10,000) =$ ________

69. Thirty percent (30%) of carbon-14 will decay within how many years?

1. Sketch the graph, and shade the area of interest.

Figure 1.3.15

Find the value k such that $P(x < k) = 0.30$.

1.4: Normal Homework

6.1 The Standard Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

65. What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

66. What is the *z*-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

67. The length of time to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution..
III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only
d. I, II, and III

68. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.

69. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.

a. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

70. Kyle’s doctor told him that the z-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If $X =$ a systolic blood pressure score then $X \sim$ N (125, 14).

a. Which answer(s) is/are correct?

- Kyle’s systolic blood pressure is 175.
- Kyle’s systolic blood pressure is 1.75 times the average blood pressure of men his age.
- Kyle’s systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
- Kyles’s systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.

b. Calculate Kyle’s blood pressure.

71. Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. $X \sim$N(10.2, 0.8). Calculate the z-scores that correspond to the following weights and interpret them.

a. 11 kg
b. 7.9 kg
c. 12.2 kg

72. In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.

a. Calculate the z-score for an SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

6.3 Estimating the Binomial with the Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

73. What is the probability of spending more than two days in recovery?

a. 0.0580
b. 0.8447
c. 0.0553
d. 0.9420

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

74. Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?

a. Yes
b. No
c. Unable to determine

75. Find the probability that it takes at least eight minutes to find a parking space.

a. 0.0001
b. 0.9270
c. 0.1862
d. 0.0668

76. Seventy percent of the time, it takes more than how many minutes to find a parking space?

a. 1.24
b. 2.41
c. 3.95
d. 6.05

77. According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let $X =$ height of the individual.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
c. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.

d. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

78. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X= IQ of an individual.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.

79. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let $X =$ percent of fat calories.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
c. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

80. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If $X =$ distance in feet for a fly ball, then $X \sim$ _____(_____,_____)
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.

81. In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percent of the children spend over ten hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

82. In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let $X =$ number of votes for President Clinton for an election district.

a. State the approximate distribution of X.
b. Is 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
e. Find the third quartile for votes for President Clinton.

83. Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

84. Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3% of her laps are under _____.
e. The middle 80% of her laps are from _______ seconds to _______ seconds.

85. Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let $X =$ time in line. Table 1.4.1 displays the ordered real data (in minutes):

Table 1.4.1

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate \sigma. The distribution of X can then be approximated by $X \sim$ _____(_____,_____)
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren't the answers to part 6 and part 7 exactly the same?
i. Why are the answers to part 6 and part 7 as close as they are?
j. If only ten customers has been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

86. Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

a. Ricardo's actual GPA is lower than Anita's actual GPA.
b. Ricardo is not passing because his z-score is zero.
c. Anita is in the 70^{th} percentile of students at her college.

87. An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the z-scores first, and then use those to calculate the probability.

88. A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one

standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

89. We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

a. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

90. A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

a. somewhere between 34 and 54 prizes.
b. somewhere between 54 and 64 prizes.
c. more than 64 prizes.

91. Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site.

On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

92. A hospital has 49 births in a year. It is considered equally likely that a birth be a boy as it is the birth be a girl.

a. What is the mean?
b. What is the standard deviation?
c. Can this binomial distribution be approximated with a normal distribution?
d. If so, use the normal distribution to find the probability that at least 23 of the 49 births were boys.

93. Historically, a final exam in a course is passed with a probability of 0.9. The exam is given to a group of 70 students.

1. What is the mean of the binomial distribution?
2. What is the standard deviation?
3. Can this binomial distribution be approximate with a normal distribution?
4. If so, use the normal distribution to find the probability that at least 60 of the students pass the exam?

94. A tree in an orchard has 200 oranges. Of the oranges, 40 are not ripe. Use the normal distribution to approximate the binomial distribution, and determine the probability a box containing 35 oranges has at most two oranges that are not ripe.

95. In a large city one in ten fire hydrants are in need of repair. If a crew examines 100 fire hydrants in a week, what is the probability they will find nine of fewer fire hydrants that need repair? Use the normal distribution to approximate the binomial distribution.

96. On an assembly line it is determined 85% of the assembled products have no defects. If one day 50 items are assembled, what is the probability at least 4 and no more than 8 are defective. Use the normal distribution to approximate the binomial distribution.

1.5: Normal Practice

6.1 The Standard Normal Distribution

1. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ ____________.

2. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

3. $X \sim N(1, 2)$

$\sigma =$ _______

4. A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. $X =$ ______________.

5. $X \sim N(-4, 1)$

What is the median?

6. $X \sim N(3, 5)$

$\sigma =$ _______

7. $X \sim N(-2, 1)$

$\mu =$ _______

8. What does a z-score measure?

9. What does standardizing a normal distribution do to the mean?

10. Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

11. What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

12. What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

13. What is the z-score of $x =-2$, if it is 2.78 standard deviations to the right of the mean?

14. What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

15. Suppose $X \sim N(2, 6)$. What value of x has a z-score of three?

16. Suppose $X \sim N(8, 1)$. What value of x has a z-score of –2.25?

17. Suppose $X \sim N(9, 5)$. What value of x has a z-score of –0.5?

18. Suppose $X \sim N(2, 3)$. What value of x has a z-score of –0.67?

19. Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

20. Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

21. Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

22. Suppose $X \sim N(-1, 2)$. What is the z-score of $x = 2$?

23. Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

24. Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

25. Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

26. In a normal distribution, $x = 5$ and $z =-1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

27. In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

28. In a normal distribution, $x =-2$ and $z = 6$. This tells you that $x =-2$ is ____ standard deviations to the ____ (right or left) of the mean.

29. In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.

30. In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

31. About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

32. About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

33. About what percent of x values lie between the second and third standard deviations (both sides)?

34. Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

35. Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution(i.e., –3).

36. Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

37. About what percent of x values lie between the mean and three standard deviations?

38. About what percent of x values lie between the mean and one standard deviation?

39. About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

40. About what percent of x values lie betwween the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

41. Define the random variable X in words. $X =$ _______________.

42. $X \sim$ _____(_____,_____)

6.4 Estimating the Binomial with the Normal Distribution

43. How would you represent the area to the left of one in a probability statement?

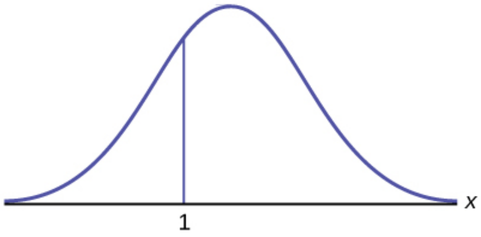


Figure 1.5.1

44. What is the area to the right of one?

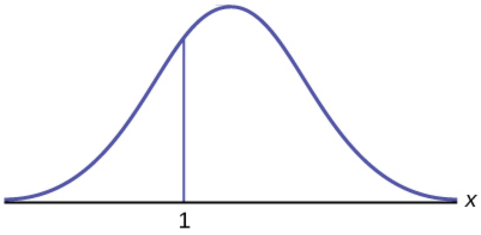


Figure 1.5.2

45. Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

46. How would you represent the area to the left of three in a probability statement?

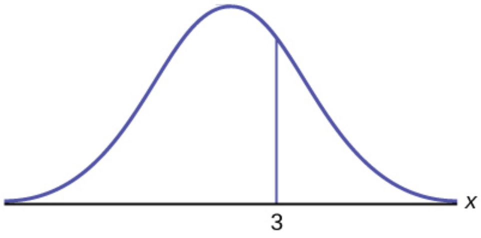


Figure 1.5.3

47. What is the area to the right of three?

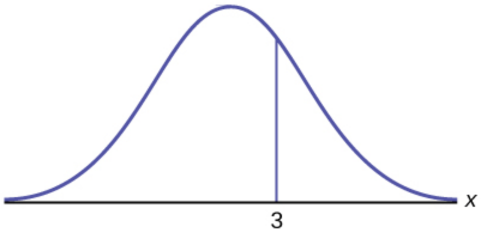


Figure 1.5.4

48. If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

49. If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

50. Find the probability that $x > 56$.

51.

Find the probability that $x < 30$.

52. $X \sim N(6, 2)$

Find the probability that x is between three and nine.

53. $X \sim N(-3, 4)$

Find the probability that x is between one and four.

54. $X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

55. *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.

1. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

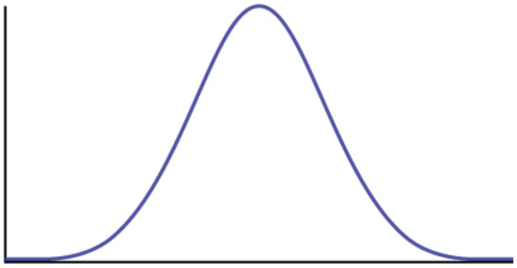


Figure 1.5.5

2. $P(0 < x <$ ________________) = ______________ (Use zero for the minimum value of x.)

56. Find the probability that a CD player will last between 2.8 and six years.

1. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

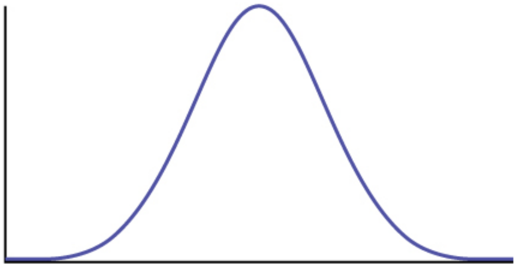


Figure 1.5.18

2. $P($_____________ $< x <$ _____________) = _____________

57. An experiment with a probability of success given as 0.40 is repeated 100 times. Use the normal distribution to approximate the binomial distribution, and find the probability the experiment will have at least 45 successes.

58. An experiment with a probability of success given as 0.30 is repeated 90 times. Use the normal distribution to approximate the binomial distribution, and find the probability the experiment will have at least 22 successes.

59. An experiment with a probability of success given as 0.40 is repeated 100 times. Use the normal distribution to approximate the binomial distribution, and find the probability the experiment will have from 35 to 45 successes.

60. An experiment with a probability of success given as 0.30 is repeated 90 times. Use the normal distribution to approximate the binomial distribution, and find the probability the experiment will have from 26 to 30 successes.

61. An experiment with a probability of success given as 0.40 is repeated 100 times. Use the normal distribution to approximate the binomial distribution, and find the probability the experiment will have at most 34 successes.

62. An experiment with a probability of success given as 0.30 is repeated 90 times. Use the normal distribution to approximate the binomial distribution, and find the probability the experiment will have at most 34 successes.

63. A multiple choice test has a probability any question will be guesses correctly of 0.25. There are 100 questions, and a student guesses at all of them. Use the normal distribution to approximate the binomial distribution, and determine the probability at least 30, but no more than 32, questions will be guessed correctly.

64. A multiple choice test has a probability any question will be guesses correctly of 0.25. There are 100 questions, and a student guesses at all of them. Use the normal distribution to approximate the binomial distribution, and determine the probability at least 24, but no more than 28, questions will be guessed correctly.

4.S: Random Variables (Solutions- Practice + Homework)

1.

Table $4.S.1$

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

3. 0.10 + 0.05 = 0.15

5. 1

7. 0.35 + 0.40 + 0.10 = 0.85

9. 1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45

11.

Table $4.S.2$

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

13. Let $X =$ the number of events Javier volunteers for each month.

15.

Table $4.S.3$

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

17. 1 – 0.05 = 0.95

18. $X =$ the number of business majors in the sample.

19. 2, 3, 4, 5, 6, 7, 8, 9

20. $X =$ the number that reply “yes”

22. 0, 1, 2, 3, 4, 5, 6, 7, 8

24. 5.7

26. 0.4151

28. $X =$ the number of freshmen selected from the study until one replied "yes" that same-sex couples should have the right to legal marital status.

30. 1,2,…

32. 1.4

35. 0, 1, 2, 3, 4, …

37. 0.0485

39. 0.0214

41. $X =$ the number of U.S. teens who die from motor vehicle injuries per day.

43. 0, 1, 2, 3, 4, ...

45. No

48.

a. X = the number of pages that advertise footwear
b. 0, 1, 2, 3, ..., 20
c. 3.03
d. 1.5197

50.

a. X = the number of Patriots picked
b. 0, 1, 2, 3, 4
c. Without replacement

53. $X =$ the number of patients calling in claiming to have the flu, who actually have the flu. X = 0, 1, 2, ...25

55. 0.0165

57.

a. X = the number of DVDs a Video to Go customer rents
b. 0.12
c. 0.11
d. 0.77

59. 4.43

61. c

63.

- X = number of questions answered correctly
- $X \sim B(32, 13)(32, 13)$
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event "more than 24" is the complement of "less than or equal to 24."
- $P(x > 24) = 0$
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

65.

a. X = the number of college and universities that offer online offerings.
b. 0, 1, 2, …, 13

c. $X \sim B(13, 0.96)$
d. 12.48
e. 0.0135
f. $P(x = 12) = 0.3186$ $P(x = 13) = 0.5882$ More likely to get 13.

67.

- X = the number of fencers who do **not** use the foil as their main weapon
- 0, 1, 2, 3,... 25
- $X \sim B(25,0.40)$
- 10
- 0.0442
- The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

69.

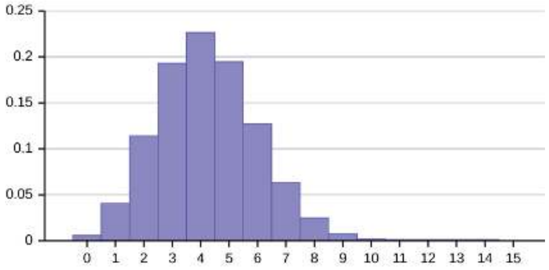
a. X = the number of audits in a 20-year period
b. 0, 1, 2, …, 20
c. $X \sim B(20, 0.02)$
d. 0.4
e. 0.6676
f. 0.0071

71.

1. X = the number of matches
2. 0, 1, 2, 3
3. In dollars: −1, 1, 2, 3
4. 1212
5. The answer is −0.0787. You lose about eight cents, on average, per game.
6. The house has the advantage.

73.

a. $X \sim B(15, 0.281)$



b. i. Mean = $\mu = np$ = 15(0.281) = 4.215
ii. Standard Deviation = σ = npq−−−√npq = 15(0.281)(0.719)−−−−−−−−−−−−√15(0.281)(0.719) = 1.7409

c. $P(x > 5)=1 – 0.7754 = 0.2246$
$P(x = 3) = 0.1927$
$P(x = 4) = 0.2259$
It is more likely that four people are literate that three people are.

75.

a. X = the number of adults in America who are surveyed until one says he or she will watch the Super Bowl.
b. $X \sim G(0.40)$
c. 2.5
d. 0.0187
e. 0.2304

77.

a. X = the number of pages that advertise footwear
b. X takes on the values 0, 1, 2, ..., 20
c. X ~ B(20, 2919229192)
d. 3.02
e. No
f. 0.9997
g. X = the number of pages we must survey until we find one that advertises footwear. X ~ G(2919229192)
h. 0.3881
i. 6.6207 pages

79. 0, 1, 2, and 3

81.

a. X ~ G(0.25)
b. i. Mean = μ = 1p1p = 10.2510.25 = 4
 ii. Standard Deviation = σ = 1−pp2−−−√1−$p$$p$2 = 1−0.250.252−−−−−√1−0.250.252 ≈ 3.4641
c. $P(x = 10) = 0.0188$
d. $P(x = 20) = 0.0011$
e. $P(x \leq 5) = 0.7627$

82.

a. X ~ P(5.5); μ = 5.5; σ = 5.5−−−√σ = 5.5 ≈ 2.3452
b. $P(x \leq 6) \approx 0.6860$
c. There is a 15.7% probability that the law staff will receive more calls than they can handle.
d. $P(x > 8) = 1 - P(x \leq 8) \approx 1 - 0.8944 = 0.1056$

84.

Let X = the number of defective bulbs in a string.

Using the Poisson distribution:

- $\mu = np = 100(0.03) = 3$
- X ~ P(3)
- $P(x \leq 4) \approx 0.8153$

Using the binomial distribution:

- X ~ B(100, 0.03)
- $P(x \leq 4) = 0.8179$

The Poisson approximation is very good—the difference between the probabilities is only 0.0026.

86.

a. X = the number of children for a Spanish woman
b. 0, 1, 2, 3,...
c. 0.2299
d. 0.5679
e. 0.4321

88.

a. X = the number of fortune cookies that have an extra fortune
b. 0, 1, 2, 3,... 144
c. 4.32
d. 0.0124 or 0.0133
e. 0.6300 or 0.6264
f. As n gets larger, the probabilities get closer together.

90.

a. X = the number of people audited in one year
b. 0, 1, 2, ..., 100
c. 2
d. 0.1353
e. 0.3233

92.

a. X = the number of shell pieces in one cake
b. 0, 1, 2, 3,...
c. 1.5
d. 0.2231
e. 0.0001
f. Yes

94. d

1.1: Uniform + Exponential Solution (Practice + Homework)

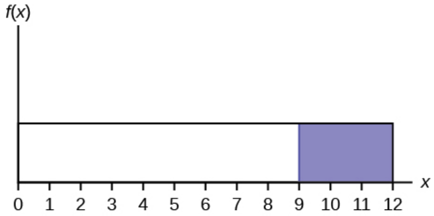


Figure 1.1.38

Figure 1.1.41

1.2: Normal Solution (Practice + Homework)

1. ounces of water in a bottle

3. 2

5. –4

7. –2

9. The mean becomes zero.

11. $z = 2$

13. $z = 2.78$

15. $x = 20$

17. $x = 6.5$

19. $x = 1$

21. $x = 1.97$

23. $z = -1.67$

25. $z \approx -0.33$

27. 0.67, right

29. 3.14, left

31. about 68%

33. about 4%

35. between –5 and –1

37. about 50%

39. about 27%

41. The lifetime of a Sunshine CD player measured in years.

43. $P(x < 1)$

45. Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

47. $1 - P(x < 3)$ or $P(x > 3)$

49. $1 - 0.543 = 0.457$

51. 0.0013

53. 0.1186

55.

57.0.154 0.874

59. 0.693

60. 0.346

61. 0.110

62. 0.946

63. 0.071

64. 0.347

66. c

68.

a. Use the *z*-score formula. *z* = –0.5141. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.
b. Use the *z*-score formula. *z* = 1.5424. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
c. Height = 79 + 3.5(3.89) = 92.615 inches, which is taller than 7 feet, 8 inches. There are very few NBA players this tall so the answer is no, not likely.

70.

a. iv
b. Kyle's blood pressure is equal to 125 + (1.75)(14) = 149.5.

72.

Let *X* = an SAT math score and *Y* = an ACT math score.

a. *X* = 720 720 – 52015720 – 52015 = 1.74 The exam score of 720 is 1.74 standard deviations above the mean of 520.
b. *z* = 1.5
The math SAT score is 520 + 1.5(115) ≈ 692.5. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
c. X – μσ$X - \mu\sigma$ = 700 – 514117700 – 514117 ≈ 1.59, the *z*-score for the SAT. Y – μσ$Y - \mu\sigma$ = 30 – 215.330 – 215.3 ≈ 1.70, the *z*-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher *z*-score).

75. d

79.

a. *X* ~ *N*(66, 2.5)
b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082

81.

a. *X* = number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. *X* ~ *N*(3, 1.5)
c. The probability that the child spends less than one hour a day unsupervised is 0.0918.
d. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
e. 2.21 hours

83.

a. *X* = the distribution of the number of days a particular type of criminal trial will take
b. *X* ~ *N*(21, 7)
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

85.

a. mean = 5.51, *s* = 2.15
b. Check student's solution.
c. Check student's solution.
d. Check student's solution.
e. *X* ~ *N*(5.51, 2.15)
f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit normal curve as well.

88.

- $n=100; p=0.1; q=0.9$
- $\mu=np=(100)(0.10)=10$
- $\sigma=\sqrt{npq}=\sqrt{(100)(0.1)(0.9)}=3$

i. $z=\pm 1: x_1=\mu+z\sigma=10+1(3)=13$ and $x2=\mu-z\sigma=10-1(3)=7.68\%$ of the defective cars will fall between seven and 13 .
ii. $z=\pm 2: x_1=\mu+z\sigma=10+2(3)=16$ and $x2=\mu-z\sigma=10-2(3)=4.95\%$ of the defective cars will fall between four and 16
iii. $z=\pm 3: x_1=\mu+z\sigma=10+3(3)=19$ and $x2=\mu-z\sigma=10-3(3)=1.99.7\%$ of the defective cars will fall between one and 19.

90.

- $n=190; p=\frac{1}{5}=0.2; q=0.8$
- $\mu=np=(190)(0.2)=38$
- $\sigma=\sqrt{npq}=\sqrt{(190)(0.2)(0.8)}=5.5136$

a. For this problem: $P(34<x<54)=0.7641$
b. For this problem: $P(54<x<64)=0.0018$
c. For this problem: $P(x>64)=0.0000012$(approximately 0)

92.

a. 24.5
b. 3.5
c. Yes
d. 0.67

93.

a. 63
b. 2.5
c. Yes
d. 0.88

94. 0.02

95. 0.37

96. 0.50

CHAPTER OVERVIEW

5: Point Estimates

5.1: Introduction to the Central Limit Theorem

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **Central Limit Theorem**.

The **Central Limit Theorem** is one of the most powerful and useful ideas in all of statistics. The Central Limit Theorem is a theorem which means that it is NOT a theory or just somebody's idea of the way things work. As a theorem it ranks with the Pythagorean Theorem, or the theorem that tells us that the sum of the angles of a triangle must add to 180. These are facts of the ways of the world rigorously demonstrated with mathematical precision and logic. As we will see this powerful theorem will determine just what we can, and cannot say, in inferential statistics. The Central Limit Theorem is concerned with drawing finite samples of size n from a population with a known mean, μ, and a known standard deviation, σ. The conclusion is that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram (distribution) of those means, then the resulting distribution will tend to have an approximate normal distribution.

The astounding result is that it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means tend to follow the normal distribution.

Figure 5.1.1 If you want to figure out the distribution of the change people carry in their pockets, using the Central Limit Theorem and assuming your sample is large enough, you will find that the distribution is the normal probability density function. (credit: John Lodder)

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means. **Sampling is done randomly and with replacement in the theoretical model.**

5.2: The Central Limit Theorem for Sample Means

The **sampling distribution** is a theoretical distribution. Specifically, the sampling distribution is the probability distribution of the mean of a random sample. Since the sample is obtained randomly, the value of the sample mean is a random variable. The sampling distribution is created by taking many samples of size n from a population. Each sample mean is then treated like a single observation of this new distribution, the sampling distribution. The genius of thinking this way is that it recognizes that when we sample we are creating an observation and that observation must come from some particular distribution. The Central Limit Theorem answers the question: from what distribution did a sample mean come? If this is discovered, then we can treat a sample mean just like any other random variable and calculate probabilities about what values it might take on. We have effectively moved from the world of statistics where we know only what we have from the sample, to the world of probability where we know the distribution from which the sample mean came and the parameters of that distribution.

The reasons that one samples a population are obvious. The time and expense of checking every invoice to determine its validity or every shipment to see if it contains all the items may well exceed the cost of errors in billing or shipping. For some products, sampling would require destroying them, called destructive sampling. One such example is measuring the ability of a metal to withstand saltwater corrosion for parts on ocean going vessels.

Sampling thus raises an important question; just which sample was drawn. Even if the sample were randomly drawn, there are theoretically an almost infinite number of samples. With just 100 items, there are more than 75 million unique samples of size 5 that can be drawn. If 6 are in the sample, the number of possible samples increases to just more than one billion. Of the 75 million possible samples, then, which one did you get? If there is variation in the items to be sampled, there will be variation in the samples. One could draw an "unlucky" sample and make very wrong conclusions concerning the population. This recognition that any sample we draw is really only one from a distribution of samples provides us with what is probably the single most important theorem is statistics: **the Central Limit Theorem**. Without the Central Limit Theorem it would be impossible to proceed to inferential statistics from simple probability theory. In its most basic form, the Central Limit Theorem states that **regardless** of the underlying probability density function of the population data, the theoretical distribution of the means of samples from the population will be normally distributed. In essence, this says that the mean of a sample should be treated like an observation drawn from a normal distribution. The Central Limit Theorem only holds if the sample size is "large enough" which has been shown to be only 30 or more.

Figure $5.2.1$ graphically displays this very important proposition.

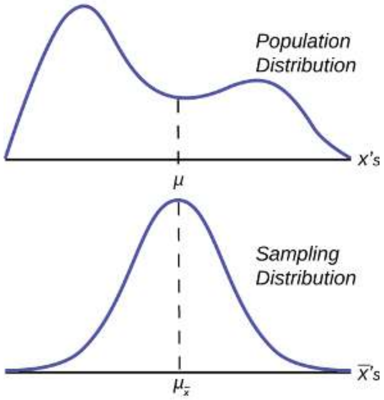


Figure 5.2.1

Notice that the horizontal axis in the top panel is labeled x. These are the individual observations of the population. This is the **unknown** distribution of the population values. The graph is purposefully drawn all squiggly to show that it does not matter just how odd ball it really is. Remember, we will never know what this distribution looks like, or its mean or standard deviation for that matter.

The horizontal axis in the bottom panel is labeled $\overline{x}$'s. This is the theoretical distribution called the sampling distribution of the sample mean. Each observation on this distribution is a sample mean. All these sample means were calculated from individual samples with the same sample size. The theoretical sampling distribution contains all of the sample mean values from all the possible samples that could have been taken from the population. Of course, no one would ever actually take all of these samples, but if they did this is how they would look. And the Central Limit Theorem says that they will be normally distributed.

The Central Limit Theorem goes even further and tells us the mean and standard deviation of this theoretical distribution, as detailed in Table $5.2.1$.

Table 5.2.1

Parameter	Population distribution	Sample	Sampling distribution of $\overline{X}$
Mean	μ	$\overline{X}$	$\mu_{\overline{X}} = \mathrm{E}\left[\overline{X}\right] = \mu$
Standard deviation	σ	s	$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

The practical significance of The Central Limit Theorem is that now we can compute probabilities for drawing a sample mean, $\overline{X}$, in just the same way as we did for drawing specific observations, X's, when we knew the population mean and standard deviation and that the population data were normally distributed. The standardizing formula has to be amended to recognize that the mean and standard deviation of the sampling distribution, sometimes, called the standard error of the mean, are different from those of the population distribution, but otherwise nothing has changed. The new standardizing formula is

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Notice that $\mu_{\overline{X}}$ in the first formula has been changed to simply μ in the second version. The reason is that mathematically it can be shown that the expected value of $\overline{X}$, the mean of $\overline{X}$, is equal to μ. This was stated in Table $5.2.1$ above. This formula will be used in the next chapter to provide interval estimates of the **unknown** population parameter μ.

5.3: Using the Central Limit Theorem

Examples of the Central Limit Theorem

Law of Large Numbers

The **Law of Large Numbers** says that if you take samples of larger and larger size from any population, then the mean of the sampling distribution, $\mu_{\overline{X}}$, tends to get closer and closer to the true population mean, μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation of the sampling distribution gets. (Remember that the standard deviation for the sampling distribution of $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{X}$ must be closer to the population mean μ as n increases. We can say that μ is the value that the sample means approach as n gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

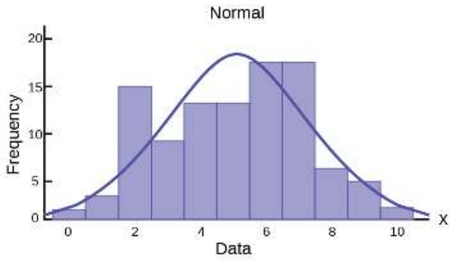
This concept is so important and plays such a critical role in what follows it deserves to be developed further. Indeed, there are two critical issues that flow from the Central Limit Theorem and the application of the Law of Large Numbers to it. These are

1. The probability density function of the sampling distribution of the sample mean is normally distributed **regardless** of the underlying distribution of the population observations and
2. standard deviation of the sampling distribution decreases as the size of the samples that were used to calculate the means for the sampling distribution increases.

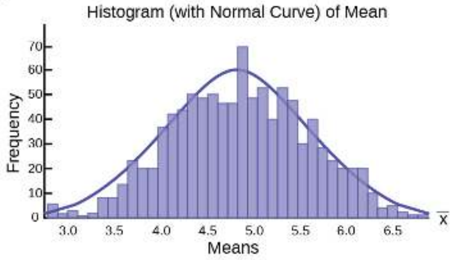
Taking these in order. It would seem counterintuitive that the population may have **any** distribution and the distribution of means coming from it would be normally distributed. With the use of computers, experiments can be simulated that show the process by which the sampling distribution changes as the sample size is increased. These simulations show visually the results of the mathematical proof of the Central Limit Theorem.

We consider three examples of very different population distributions and the evolution of the sampling distribution to a normal distribution as the sample size increases. The top panel in these cases represents the histogram for the original data. The three panels show the histograms for 1,000 randomly drawn samples for different sample sizes: $n = 10$, $n = 25$ and $n = 50$. As the sample size increases, and the number of samples taken remains constant, the distribution of the 1,000 sample means becomes closer to the smooth line that represents the normal distribution.

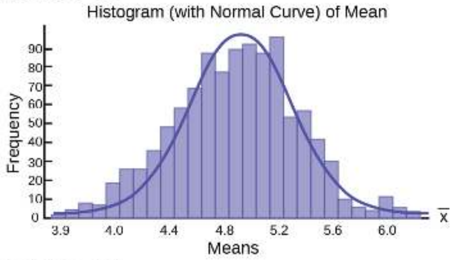
Figure $5.3.1$ is for a normal distribution of individual observations and we would expect the sampling distribution to converge on the normal quickly. The results show this and show that even at a very small sample size the distribution is close to the normal distribution.



Sample Size *n* = 10



Sample Size *n* = 25



Sample Size *n* = 50

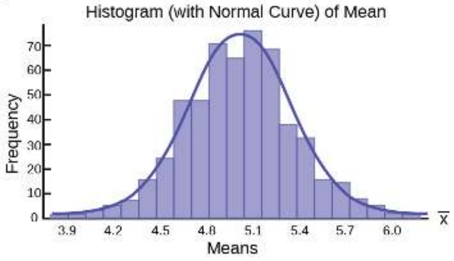


Figure $5.3.1$

Figure $5.3.2$ is a uniform distribution which, a bit amazingly, quickly approached the normal distribution even with only a sample of 10.

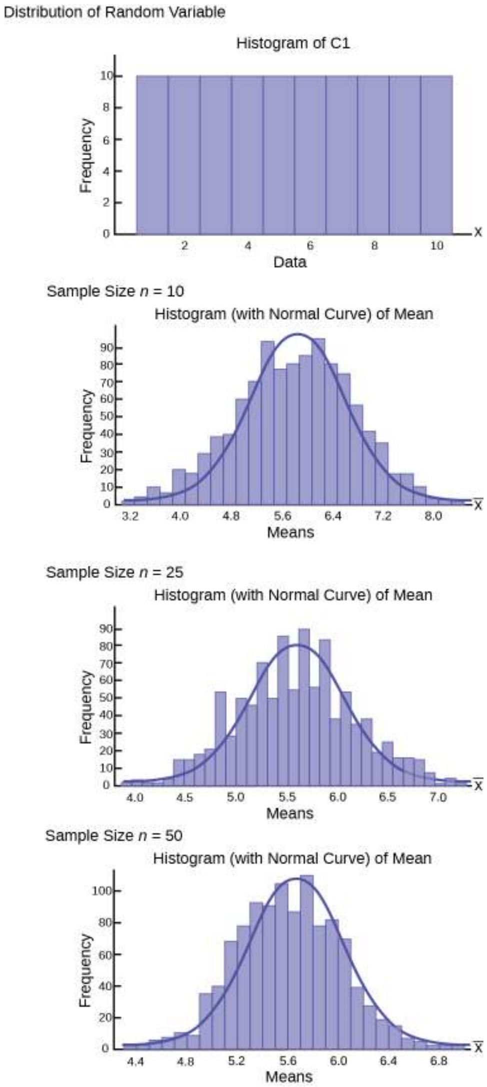
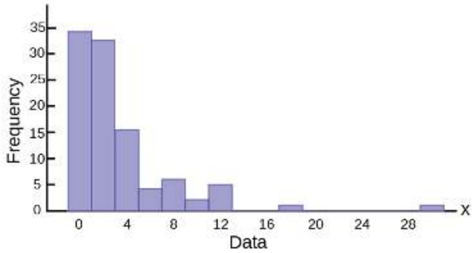
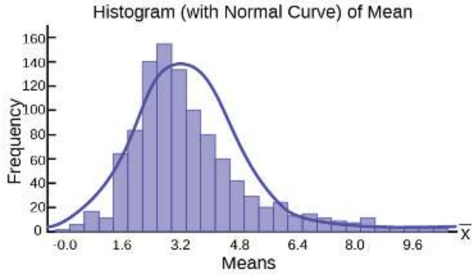


Figure $5.3.2$

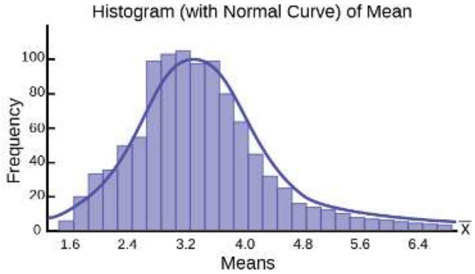
Figure $5.3.3$ is a skewed distribution. This last one could be an exponential, or binomial with a small probability of success creating the skew in the distribution. For skewed distributions our intuition would say that this will take larger sample sizes to move to a normal distribution and indeed that is what we observe from the simulation. Nevertheless, at a sample size of 50, not considered a very large sample, the distribution of sample means has very decidedly gained the shape of the normal distribution.



Distribution of Sample means with n = 10



Distribution of Sample means with n = 25



Distribution of Sample means with n = 50

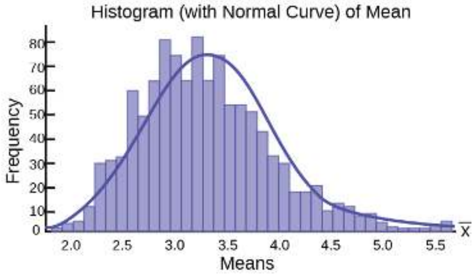


Figure $5.3.3$

The Central Limit Theorem provides more than the proof that the sampling distribution of the sample mean is normally distributed. It also provides us with the mean and standard deviation of this distribution. As discussed above, the mean of the sample mean (its expected value, in other words) is equal to the mean of the population of the original data, which is what we are interested in estimating from the sample we took. We have already inserted this conclusion of the Central Limit Theorem into the formula we use for standardizing from the sampling distribution to the standard normal distribution. And finally, the Central Limit Theorem has also provided the standard deviation of the sampling distribution, $\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}$, and this is critical to have in order to calculate probabilities of values of the new random variable, $\overline{X}$.

Figure $5.3.4$ shows a sampling distribution. The mean has been marked on the horizontal axis of the $\overline{x}$'s and the standard deviation has been written to the right above the distribution. Notice that the standard deviation of the sampling distribution is the original standard deviation of the population, divided by the square root of the sample size. We have already seen that as the sample size increases the sampling distribution becomes closer and closer to the normal distribution. As this happens, the standard deviation of the sampling distribution changes in another way; the standard deviation decreases as n increases. At very very large n, the standard deviation of the sampling distribution becomes very small and at infinity it collapses on top of the population mean. This is what it means that the expected value of $\overline{X}$ is the population mean, μ.

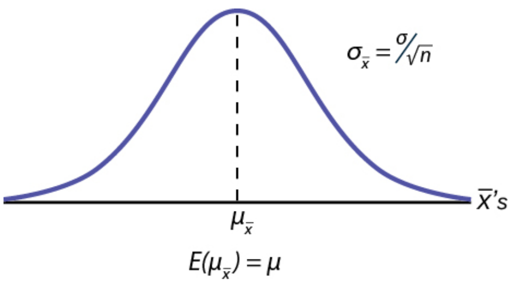


Figure 5.3.4

At non-extreme values of n, this relationship between the standard deviation of the sampling distribution and the sample size plays a very important part in our ability to estimate the parameters we are interested in.

Figure 5.3.5 shows three sampling distributions. The only change that was made is the sample size that was used to get the sample means for each distribution. As the sample size increases, n goes from 10 to 30 to 50, the standard deviations of the respective sampling distributions decrease because the sample size is in the denominator of the standard deviations of the sampling distributions.

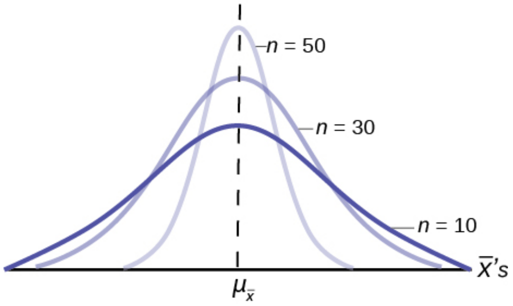


Figure 5.3.5

The implications for this are very important. Figure 5.3.6 shows the effect of the sample size on the confidence we will have in our estimates. These are two sampling distributions from the same population. One sampling distribution was created with samples of size 10 and the other with samples of size 50. All other things constant, the sampling distribution with sample size 50 has a smaller standard deviation that causes the graph to be higher and narrower. The important effect of this is that for the same probability of one standard deviation from the mean, this distribution covers much less of a range of possible values than the other distribution. One standard deviation is marked on the $\overline{X}$ axis for each distribution. This is shown by the two arrows that are plus or minus one standard deviation for each distribution. If the probability that the true mean is one standard deviation away from the mean, then for the sampling distribution with the smaller sample size, the possible range of values is much greater. A simple question is, would you rather have a sample mean from the narrow, tight distribution, or the flat, wide distribution as the estimate of the population mean? Your answer tells us why people intuitively will always choose data from a large sample rather than a small sample. The sample mean they are getting is coming from a more compact distribution. This concept will be the foundation for what will be called level of confidence in the next unit.

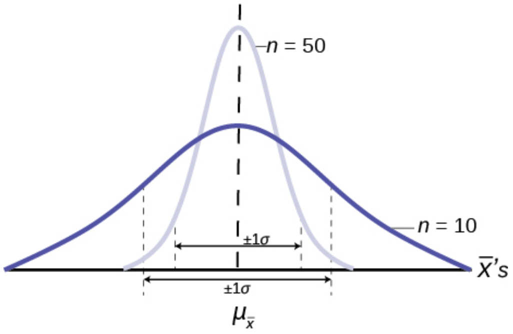


Figure 5.3.6

5.4: The Central Limit Theorem for Proportions

The Central Limit Theorem tells us that the point estimate for the sample mean, $\overline{X}$, comes from a normal distribution of $\overline{X}$'s. This theoretical distribution is called the sampling distribution of $\overline{X}$'s. We now investigate the sampling distribution for another important parameter we wish to estimate; p from the binomial probability density function.

If the random variable is discrete, such as for categorical data, then the parameter we wish to estimate is the population proportion. This is, of course, the probability of drawing a success in any one random draw. Unlike the case just discussed for a continuous random variable where we did not know the population distribution of X's, here we actually know the underlying probability density function for these data; it is the binomial. The random variable is $X =$ the number of successes in the sample and the parameter we wish to know is p, the probability of drawing a success which is of course the proportion of successes in the population. The question at issue is: from what distribution was the sample proportion, $\hat{p} = \frac{X}{n}$ drawn? The sample size is n and X is the number of successes found in that sample. This is a parallel question that was just answered by the Central Limit Theorem: from what distribution was the sample mean, $\overline{X}$, drawn? We saw that once we knew that the distribution was the Normal distribution then we were able to create confidence intervals for the population parameter, μ. We will also use this same information to test hypotheses about the population mean later. We wish now to be able to develop confidence intervals for the population parameter "p" from the binomial probability density function.

In order to find the distribution from which sample proportions come we need to develop the sampling distribution of sample proportions just as we did for sample means. So again imagine that we randomly sample say 50 people and ask them if they support the new school bond issue. From this we find a sample proportion, $\hat{p}$, and graph it on the axis of p's. We do this again and again until we have the theoretical distribution of $\hat{p}$. Some sample proportions will show high favorability toward the bond issue and others will show low favorability because random sampling will reflect the variation of views within the population. What we have done can be seen in Figure $5.4.1$. The top panel is the population distributions of probabilities for each possible value of the random variable X. While we do not know what the specific distribution looks like because we do not know p, the population parameter, we do know that it must look something like this. In reality, we do not know either the mean or the standard deviation of this population distribution, the same difficulty we faced when analyzing the X's previously.

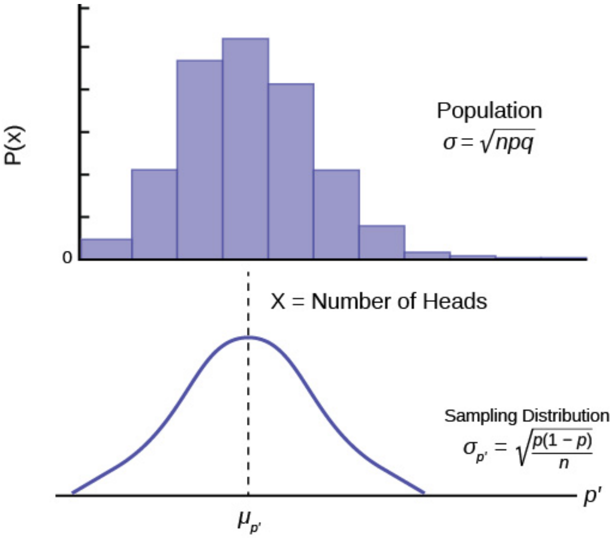


Figure $5.4.1$

Figure $5.4.1$ places the mean on the distribution of population probabilities as $\mu = np$ but of course we do not actually know the population mean because we do not know the population probability of success, p. Below the distribution of the population values is the sampling distribution of $\hat{p}$. Again the Central Limit Theorem tells us that this distribution is normally distributed just like the case of the sampling distribution for $\overline{X}$. This sampling distribution also has a mean, the mean of the p's, and a standard deviation, $\sigma_{\hat{p}}$.

Importantly, in the case of the analysis of the distribution of sample means, the Central Limit Theorem told us the expected value of the mean of the sample means in the sampling distribution, and the standard deviation of the sampling distribution. Again the Central Limit Theorem provides this information for the sampling distribution for proportions. The answers are:

1. The mean of the sampling distribution of the sample proportion, $\mu_{\hat{p}}$, is the population proportion, p.
2. The standard deviation of the sampling distribution of the sample proportion, $\sigma_{\hat{p}}$, is the population standard deviation divided by the square root of the sample size, n.

Both these conclusions are the same as we found for the sampling distribution for sample means. However in this case, because the mean and standard deviation of the binomial distribution both rely upon pp, the formula for the standard deviation of the sampling distribution requires algebraic manipulation to be useful. We will take that up in the next chapter. The proof of these important conclusions from the Central Limit Theorem is provided below.

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \left(\frac{1}{n}\right) E[X] = \left(\frac{1}{n}\right) np = p$$

(The expected value of X, $E[X]$, is simply the mean of the binomial distribution which we know to be np.)

$$\sigma_{\hat{p}}^2 = \operatorname{Var}(\hat{p}) = \operatorname{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}(\operatorname{Var}(X)) = \frac{1}{n^2}(np(1-p)) = \frac{p(1-p)}{n}$$

The standard deviation of the sampling distribution for proportions is thus:

$$\sigma_{\hat{p}}, = \sqrt{\frac{p(1-p)}{n}}$$

Table 5.4.1

Parameter	Population distribution	Sample	Sampling distribution of p's
Mean	$\mu = np$	$\hat{p} = \frac{X}{n}$ \)	$\mu_{\hat{p}} = E[\hat{p}] = p$
Standard Deviation	$\sigma = \sqrt{np(1-p)}$		$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Table $5.4.1$ summarizes these results and shows the relationship between the population, sample and sampling distribution.

Reviewing the formula for the standard deviation of the sampling distribution for proportions we see that as n increases the standard deviation decreases. This is the same observation we made for the standard deviation for the sampling distribution for means. Again, as the sample size increases, the point estimate for either μ or p is found to come from a distribution with a narrower and narrower distribution.

- **7.3: The Central Limit Theorem for Proportions** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

5.5: Finite Population Correction Factor

We saw that the sample size has an important effect on the variance and thus the standard deviation of the sampling distribution. Also of interest is the proportion of the total population that has been sampled. We have assumed that the population is extremely large and that we have sampled a small part of the population. As the population becomes smaller and we sample a larger number of observations the sample observations are not independent of each other. To correct for the impact of this, the **Finite Population Correction Factor** can be used to adjust the variance of the sampling distribution. It is appropriate when more than 5% of the population is being sampled and the population has a known population size. There are cases when the population is known, and therefore the correction factor must be applied. The issue arises for both the sampling distribution of the means and the sampling distribution of proportions. The Finite Population Correction Factor for the variance of the means shown in the standardizing formula is:

$$Z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}}$$

and for the variance of proportions is:

$$\sigma_{\hat{p}}=\sqrt{\frac{p(1-p)}{n}}\times\sqrt{\frac{N-n}{N-1}}$$

The following examples show how to apply the factor. Sampling variances get adjusted using the above formula.

Example 5.5.1

It is learned that the population of White German Shepherds in the USA is 4,000 dogs and the mean weight for German Shepherds is 75.45 pounds. It is also learned that the population standard deviation is 10.37 pounds. If the sample size is 100 dogs, then find the probability that a sample will have a mean that differs from the true population mean by less than 2 pounds.

Answer

We are given the following information:

$$N=4000,\quad n=100,\quad \sigma=10.37,\quad \mu=75.45,\quad \text{want } |\overline{X}-\mu|<\pm 2$$

Note that "differs by less" references the area on both sides of the mean within 2 pounds right or left. We apply the finite population correction factor as follows:

$$P(|\overline{X}-\mu|<2)=P\left(\frac{|\overline{X}-\mu|}{\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}}<\frac{2}{\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}}\right)\approx P\left(\frac{-2}{\frac{10.37}{\sqrt{100}}\cdot\sqrt{\frac{4000-100}{4000-1}}}<Z<\frac{2}{\frac{10.37}{\sqrt{100}}\cdot\sqrt{\frac{4000-100}{4000-1}}}\right)=P(-1.95<Z<1.95)$$

From the Z Table, we find that $P(0<Z<1.95)=0.4744$, which gives the following:

$$P(-1.95<Z<1.95)=0.4744\cdot 2=0.9488$$

We have found that there is approximately a 94.88% chance that the mean weight of a sample of 100 dogs will differ from the true mean by less than 2 pounds.

Example 5.5.2

When a customer places an order with Rudy's On-Line Office Supplies, a computerized accounting information system (AIS) automatically checks to see if the customer has exceeded his or her credit limit. Past records indicate that the probability of customers exceeding their credit limit is .06.

Suppose that on a given day, 3,000 orders are placed in total. If we randomly select 360 orders, what is the probability that between 10 and 20 customers will exceed their credit limit?

Answer

We are given the following information:

$$N=3000,\quad n=360,\quad p=0.06$$

We want to find $P(10<X<20)=P\left(\frac{10}{360}<\hat{p}<\frac{20}{360}\right)=P(0.0278<\hat{p}<0.0556)$. We first find the standard deviation of $\hat{p}$ applying the finite population correction factor (because $n/N=360/3000=0.12>0.05$

$$\sigma_{\hat{p}}=\sqrt{\frac{p(1-p)}{n}}\times\sqrt{\frac{N-n}{N-1}}=\sqrt{\frac{0.06(1-0.06)}{360}}\times\sqrt{\frac{3000-360}{3000-1}}=0.0117$$

Note that we know the mean of $\hat{p}$ since we know the population proportion, $p=0.06$. Thus, the sampling distribution of $\hat{p}$ can be approximated by $N(0.06, 0.0117$.)So we compute:

$$P(0.0278<\hat{p}<0.0556)=P\left(\frac{0.0278-0.06}{0.0117}<\frac{\hat{p}-0.06}{0.0117}<\frac{0.0556-0.06}{0.0117}\right)\approx P(-2.75<Z<-0.38)=0.4970+0.1480=0.6450.$$

There is approximately a 64.5% chance that between 10 and 20 customers in a sample of size 360 will exceed their credit limit.

5.6: Central Limit Theorem (Worksheet)

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will compute normal probabilities using the Z Table.
- The student will demonstrate an understanding of the Central Limit Theorem.

I. Calculating Standard Normal Probabilities

Recall that we denote a random variable with the standard normal distribution as Z, meaning that $Z \sim N(0,1)$. Use the Z Table to calculate each of the following probabilities of Z.

1. $P(0 < Z < 1.07)$
2. $P(-1.07 < Z < 0)$
3. $P(Z < 1.07)$
4. $P(Z > 1.07)$
5. $P(1.07 < Z < 2.89)$
6. $P(-0.43 < Z < 1.07)$
7. $P(-2.89 < Z < -0.43)$
8. Find the value of z, such that $P(Z > z) = 0.0250$.

II. Calculating Non-Standard Normal Probabilities

Suppose that $X \sim N(\mu = 5, \sigma = 2)$. Use the Z Table to calculate each of the following probabilities of Z. You will need to convert each probability into an equivalent probability of Z.

1. $P(5 < X < 6.3)$
2. $P(1.7 < X < 6.3)$
3. Find the value of x, such that $P(X < x) = 0.0020$.

III. Applying the Central Limit Theorem

Let the random variable $Y =$ the monthly income of a randomly selected employed adult in Zimbabwe. We can model the distribution of Y as exponential with parameter $\mu =$ \$477.5 USD.

1. Determine the mean and standard deviation of Y.
2. Suppose a random sample of size $n = 100$ of employed adults in Zimbabwe is obtained. Use the Central Limit Theorem to approximate the probability that the mean of the sample (that is, the sample mean $\overline{Y}$) will be within \$25 USD of the population mean that you found in the previous question.

5.7: Chapter Key Terms

Key Term	Definition
Average	a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.
Central Limit Theorem	Given a random variable with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\overline{X}$. If the size (n) of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size (n) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.
Finite Population Correction Factor	adjusts the variance of the sampling distribution if the population is known and more than 5% of the population is being sampled.
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 (denoted by $\overline{x}$) is $\overline{x} = \overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.
Normal Distribution	a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation.; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the random variable, Z, is called the **standard normal distribution**.
Sampling Distribution	Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.
Standard Error of the Mean	the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.
Standard Error of the Proportion	the standard deviation of the sampling distribution of proportions

---

- **7.6: Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

5.8: Chapter References

7.1 The Central Limit Theorem for Sample Means

Baran, Daya. “20 Percent of Americans Have Never Used Email.”WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-...ver-used-email (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at blog.flurry.com (accessed May 17, 2013).

Data from the United States Department of Agriculture.

5.9: Chapter Formula Review

7.1 The Central Limit Theorem for Sample Means

The Central Limit Theorem for Sample Means:

$$\overline{X} \sim N\left(\mu_{\overline{x}}, \frac{\sigma}{\sqrt{n}}\right)$$

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{X}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

The Mean $\overline{X}$: $\mu_{\overline{x}}$

Central Limit Theorem for Sample Means z-score $z = \frac{\overline{x} - \mu_{\overline{x}}}{\left(\frac{\sigma}{\sqrt{n}}\right)}$

Standard Error of the Mean (Standard Deviation $(\overline{X})$) : $\frac{\sigma}{\sqrt{n}}$

Finite Population Correction Factor for the sampling distribution of means: $Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}}$

Finite Population Correction Factor for the sampling distribution of proportions: $\sigma_{\mathrm{p}'} = \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$

5.10: Chapter Review

7.1 The Central Limit Theorem for Sample Means

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

7.2 Using the Central Limit Theorem

The Central Limit Theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer the sample mean $\overline{x}$ gets to μ.

7.3 The Central Limit Theorem for Proportions

The Central Limit Theorem can also be used to illustrate that the sampling distribution of sample proportions is normally distributed with the expected value of p and a standard deviation of $\sigma_{\mathrm{p}'}=\sqrt{\frac{p(1-p)}{n}}$

5.H: Point Estimates (Homework + Practice)

7.2 The Central Limit Theorem for Sample Means

49 Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of $0.88. Suppose that we randomly pick 25 daytime statistics students.

a. In words, X = ____________
b. X $\sim$ _____(_____,_____)
c. In words, $\overline{X}$ = ____________
d. $\overline{X} \sim$ ______ (______, ______)
e. Find the probability that an individual had between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
f. Find the probability that the average of the 25 students was between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
g. Explain why there is a difference in part e and part f.

Answer

a. X = amount of change students carry
b. X $\sim E(0.88, 0.88)$
c. $\overline{X}$ = average amount of change carried by a sample of 25 students.
d. $\overline{X} \sim N(0.88, 0.176)$
e. 0.0819
f. 0.1882
g. The distributions are different. Part 1 is exponential and part 2 is normal.

50. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If $\overline{X}$ = average distance in feet for 49 fly balls, then $\overline{X} \sim$ _______(_______,_______)
b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for $\overline{X}$. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of the average of 49 fly balls.

51. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. In words, X $=$ _____________
b. In words, $\overline{X}$ = _____________
c. $\overline{X} \sim$ _____(_____,_____)
d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
e. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

52. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let $\overline{X}$ the average of the 49 races.

a. $\overline{X} \sim$ _____(_____,_____)
b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
c. Find the 80^{th} percentile for the average of these 49 marathons.
d. Find the median of the average running times.

53. The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

a. In words, X = _________
b. X ∼ _____________
c. In words, $\overline{X}$ = _____________
d. $\overline{X}$ ∼ _____(_____,_____)
e. Find the first quartile for the average song length.
f. The IQR (interquartile range) for the average song length is from _______–_______.

54. In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

a. In words, X = _____________
b. In words, $\overline{X}$ = _____________
c. $\overline{X}$ ∼ _____(_____,_____)
d. The IQR for $\overline{X}$ is from _______ acres to _______ acres.

55. Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of $\overline{X}$ is approximately equal to the mean of X.
b. When the sample size is large, $\overline{X}$ is approximately normally distributed.
c. When the sample size is large, the standard deviation of $\overline{X}$ is approximately the same as the standard deviation of X.

56. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen. Let $\overline{X}$ = average percent of fat calories.

a. $\overline{X}$ ∼ ______(______, ______)
b. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percent of fat calories.

57. The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.

a. In words, X = _____________
b. In words, $\overline{X}$ = _____________
c. $\overline{X}$ ∼ _____(_____,_____)
d. How is it possible for the standard deviation to be greater than the average?
e. Why is it more likely that the average of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to $2,200?

58. Which of the following is NOT TRUE about the distribution for averages?

a. The mean, median, and mode are equal.
b. The area under the curve is one.
c. The curve never touches the x-axis.
d. The curve is skewed to the right.

59. The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations. The distribution to use for the average cost of gasoline for the 16 gas stations is:

a. $\overline{X} \sim N(4.59, 0.10)$

b. $\overline{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$

c. $\overline{X} \sim N\left(4.59, \frac{16}{0.10}\right)$

d. $\overline{X} \sim N\left(4.59, \frac{\sqrt{16}}{0.10}\right)$

7.3 Using the Central Limit Theorem

60. A large population of 5,000 students take a practice test to prepare for a standardized test. The population mean is 140 questions correct, and the standard deviation is 80. What size samples should a researcher take to get a distribution of means of the samples with a standard deviation of 10?

61. A large population has skewed data with a mean of 70 and a standard deviation of 6. Samples of size 100 are taken, and the distribution of the means of these samples is analyzed.

a. Will the distribution of the means be closer to a normal distribution than the distribution of the population?
b. Will the mean of the means of the samples remain close to 70?
c. Will the distribution of the means have a smaller standard deviation?
d. What is that standard deviation?

62. A researcher is looking at data from a large population with a standard deviation that is much too large. In order to concentrate the information, the researcher decides to repeatedly sample the data and use the distribution of the means of the samples? The first effort used sample sized of 100. But the standard deviation was about double the value the researcher wanted. What is the smallest size samples the researcher can use to remedy the problem?

63. A researcher looks at a large set of data, and concludes the population has a standard deviation of 40. Using sample sizes of 64, the researcher is able to focus the mean of the means of the sample to a narrower distribution where the standard deviation is 5. Then, the researcher realizes there was an error in the original calculations, and the initial standard deviation is really 20. Since the standard deviation of the means of the samples was obtained using the original standard deviation, this value is also impacted by the discovery of the error. What is the correct value of the standard deviation of the means of the samples?

64. A population has a standard deviation of 50. It is sampled with samples of size 100. What is the variance of the means of the samples?

7.4 The Central Limit Theorem for Proportions

65. A farmer picks pumpkins from a large field. The farmer makes samples of 260 pumpkins and inspects them. If one in fifty pumpkins are not fit to market and will be saved for seeds, what is the standard deviation of the mean of the sampling distribution of sample proportions?

66. A store surveys customers to see if they are satisfied with the service they received. Samples of 25 surveys are taken. One in five people are unsatisfied. What is the variance of the mean of the sampling distribution of sample proportions for the number of unsatisfied customers? What is the variance for satisfied customers?

67. A company gives an anonymous survey to its employees to see what percent of its employees are happy. The company is too large to check each response, so samples of 50 are taken, and the tendency is that three-fourths of the employees are happy. For the mean of the sampling distribution of sample proportions, answer the following questions, if the sample size is doubled.

a. How does this affect the mean?
b. How does this affect the standard deviation?
c. How does this affect the variance?

68. A pollster asks a single question with only yes and no as answer possibilities. The poll is conducted nationwide, so samples of 100 responses are taken. There are four yes answers for each no answer overall. For the mean of the sampling distribution of sample proportions, find the following for yes answers.

a. The expected value.
b. The standard deviation.
c. The variance.

69. The mean of the sampling distribution of sample proportions has a value of p of 0.3, and sample size of 40.

a. Is there a difference in the expected value if p and q reverse roles?
b. Is there a difference in the calculation of the standard deviation with the same reversal?

Finite Population Correction Factor

70. A company has 1,000 employees. The average number of workdays between absence for illness is 80 with a standard deviation of 11 days. Samples of 80 employees are examined. What is the probability a sample has a mean of workdays with no absence for illness of at least 78 days and at most 84 days?

71. Trucks pass an automatic scale that monitors 2,000 trucks. This population of trucks has an average weight of 20 tons with a standard deviation of 2 tons. If a sample of 50 trucks is taken, what is the probability the sample will have an average weight within one-half ton of the population mean?

72. A town keeps weather records. From these records it has been determined that it rains on an average of 12% of the days each year. If 30 days are selected at random from one year, what is the probability that at most 3 days had rain?

73. A maker of greeting cards has an ink problem that causes the ink to smear on 7% of the cards. The daily production run is 500 cards. What is the probability that if a sample of 35 cards is checked, there will be ink smeared on at most 5 cards?

74. A school has 500 students. Usually, there are an average of 20 students who are absent. If a sample of 30 students is taken on a certain day, what is the probability that at least 2 students in the sample will be absent?

1.1: Chapter Practice

7.3 Using the Central Limit Theorem

Use the following information to answer the next ten exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is taken.

1.

a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deivation?
b. What is the distribution for the mean weight of 100 25-pound lifting weights?
c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

2. Draw the graph from **Exercise $\mathbf{1.1.1}$**

3. Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

4. Draw the graph from **Exercise $\mathbf{1.1.3}$**

5. Find the 90th percentile for the mean weight for the 100 weights.

6. Draw the graph from **Exercise $\mathbf{1.1.5}$**

7.

a. What is the distribution for the sum of the weights of 100 25-pound lifting weights?
b. Find $P(\Sigma x < 2,450)$.

8. Draw the graph from **Exercise $\mathbf{1.1.7}$**

9. Find the 90th percentile for the total weight of the 100 weights.

10. Draw the graph from **Exercise $\mathbf{1.1.9}$**

Use the following information to answer the next five exercises: The length of time a particular smartphone's battery lasts follows an exponential distribution with a mean of ten months. A sample of 64 of these smartphones is taken.

11.

a. What is the standard deviation?
b. What is the parameter m?

12. What is the distribution for the length of time one battery lasts?

13. What is the distribution for the mean length of time 64 batteries last?

14. What is the distribution for the total length of time 64 batteries last?

15. Find the probability that the sample mean is between seven and 11.

16. Find the 80th percentile for the total length of time 64 batteries last.

17. Find the IQR for the mean amount of time 64 batteries last.

18. Find the middle 80% for the total amount of time 64 batteries last.

Use the following information to answer the next eight exercises: A uniform distribution has a minimum of six and a maximum of ten. A sample of 50 is taken.

19. Find $P(\Sigma x > 420)$.

20. Find the 90th percentile for the sums.

21. Find the 15th percentile for the sums.

22. Find the first quartile for the sums.

23. Find the third quartile for the sums.

24. Find the 80th percentile for the sums.

25. A population has a mean of 25 and a standard deviation of 2. If it is sampled repeatedly with samples of size 49, what is the mean and standard deviation of the sample means?

26. A population has a mean of 48 and a standard deviation of 5. If it is sampled repeatedly with samples of size 36, what is the mean and standard deviation of the sample means?

27. A population has a mean of 90 and a standard deviation of 6. If it is sampled repeatedly with samples of size 64, what is the mean and standard deviation of the sample means?

28. A population has a mean of 120 and a standard deviation of 2.4. If it is sampled repeatedly with samples of size 40, what is the mean and standard deviation of the sample means?

29. A population has a mean of 17 and a standard deviation of 1.2. If it is sampled repeatedly with samples of size 50, what is the mean and standard deviation of the sample means?

30. A population has a mean of 17 and a standard deviation of 0.2. If it is sampled repeatedly with samples of size 16, what is the expected value and standard deviation of the sample means?

31. A population has a mean of 38 and a standard deviation of 3. If it is sampled repeatedly with samples of size 48, what is the expected value and standard deviation of the sample means?

32. A population has a mean of 14 and a standard deviation of 5. If it is sampled repeatedly with samples of size 60, what is the expected value and standard deviation of the sample means?

7.4 The Central Limit Theorem for Proportions

33. A question is asked of a class of 200 freshmen, and 23% of the students know the correct answer. If a sample of 50 students is taken repeatedly, what is the expected value of the mean of the sampling distribution of sample proportions?

34. A question is asked of a class of 200 freshmen, and 23% of the students know the correct answer. If a sample of 50 students is taken repeatedly, what is the standard deviation of the mean of the sampling distribution of sample proportions?

35. A game is played repeatedly. A player wins one-fifth of the time. If samples of 40 times the game is played are taken repeatedly, what is the expected value of the mean of the sampling distribution of sample proportions?

36. A game is played repeatedly. A player wins one-fifth of the time. If samples of 40 times the game is played are taken repeatedly, what is the standard deviation of the mean of the sampling distribution of sample proportions?

37. A virus attacks one in three of the people exposed to it. An entire large city is exposed. If samples of 70 people are taken, what is the expected value of the mean of the sampling distribution of sample proportions?

38. A virus attacks one in three of the people exposed to it. An entire large city is exposed. If samples of 70 people are taken, what is the standard deviation of the mean of the sampling distribution of sample proportions?

39. A company inspects products coming through its production process, and rejects detected products. One-tenth of the items are rejected. If samples of 50 items are taken, what is the expected value of the mean of the sampling distribution of sample proportions?

40. A company inspects products coming through its production process, and rejects detected products. One-tenth of the items are rejected. If samples of 50 items are taken, what is the standard deviation of the mean of the sampling distribution of sample proportions?

7.5 Finite Population Correction Factor

41. A fishing boat has 1,000 fish on board, with an average weight of 120 pounds and a standard deviation of 6.0 pounds. If sample sizes of 50 fish are checked, what is the probability the fish in a sample will have mean weight within 2.8 pounds the true mean of the population?

42. An experimental garden has 500 sunflowers plants. The plants are being treated so they grow to unusual heights. The average height is 9.3 feet with a standard deviation of 0.5 foot. If sample sizes of 60 plants are taken, what is the probability the plants in a given sample will have an average height within 0.1 foot of the true mean of the population?

43. A company has 800 employees. The average number of workdays between absence for illness is 123 with a standard deviation of 14 days. Samples of 50 employees are examined. What is the probability a sample has a mean of workdays with no absence for illness of at least 124 days?

44. Cars pass an automatic speed check device that monitors 2,000 cars on a given day. This population of cars has an average speed of 67 miles per hour with a standard deviation of 2 miles per hour. If samples of 30 cars are taken, what is the probability a given sample will have an average speed within 0.50 mile per hour of the population mean?

45. A town keeps weather records. From these records it has been determined that it rains on an average of 37% of the days each year. If 30 days are selected at random from one year, what is the probability that at least 5 and at most 11 days had rain?

46. A maker of yardsticks has an ink problem that causes the markings to smear on 4% of the yardsticks. The daily production run is 2,000 yardsticks. What is the probability if a sample of 100 yardsticks is checked, there will be ink smeared on at most 4 yardsticks?

47. A school has 300 students. Usually, there are an average of 21 students who are absent. If a sample of 30 students is taken on a certain day, what is the probability that at most 2 students in the sample will be absent?

48. A college gives a placement test to 5,000 incoming students each year. On the average 1,213 place in one or more developmental courses. If a sample of 50 is taken from the 5,000, what is the probability at most 12 of those sampled will have to take at least one developmental course?

5.S: Point Estimates (Solutions- Practice + Homework)

1.

a. $U(24, 26)$, 25, 0.5774
b. $N(25, 0.0577)$
c. 0.0416

3. 0.0003

5. 25.07

7.

a. $N(2{,}500, 5.7735)$
b. 0

9. 2,507.40

13. $N\left(10, \frac{10}{8}\right))$

15. 0.7799

17. 1.69

19. 0.0072

21. 391.54

23. 405.51

25. Mean = 25, standard deviation = 2/7

26. Mean = 48, standard deviation = 5/6

27. Mean = 90, standard deviation = 3/4

28. Mean = 120, standard deviation = 0.38

29. Mean = 17, standard deviation = 0.17

30. Expected value = 17, standard deviation = 0.05

31. Expected value = 38, standard deviation = 0.43

32. Expected value = 14, standard deviation = 0.65

33. 0.23

34. 0.060

35. 1/5

36. 0.063

37. 1/3

38. 0.056

39. 1/10

40. 0.042

41. 0.999

42. 0.901

43. 0.301

44. 0.832

45. 0.483

46. 0.500

47. 0.502

48. 0.519

49.

a. X = amount of change students carry
b. $X \sim E(0.88, 0.88)$
c. X–X– = average amount of change carried by a sample of 25 students.
d. X–X– ~ $N(0.88, 0.176)$
e. 0.0819
f. 0.1882
g. The distributions are different. Part a is exponential and part b is normal.

51.

a. length of time for an individual to complete IRS form 1040, in hours.
b. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
c. N(10.53, 13)(10.53, 13)
d. Yes. I would be surprised, because the probability is almost 0.
e. No. I would not be totally surprised because the probability is 0.2312

53.

a. the length of a song, in minutes, in the collection
b. $U(2, 3.5)$
c. the average length, in minutes, of the songs from a sample of five albums from the collection
d. $N(2.75, 0.066)$
e. 2.74 minutes
f. 0.03 minutes

55.

a. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
c. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.

57.

a. X = the yearly income of someone in a third world country
b. the average salary from samples of 1,000 residents of a third world country
c. X–X– ∼ N(2000, 80001000√)(2000, 80001000)
d. Very wide differences in data values can have averages smaller than standard deviations.
e. The distribution of the sample mean will have higher probabilities closer to the population mean.
P(2000 < X–X– < 2100) = 0.1537
P(2100 < X–X– < 2200) = 0.1317

59.

b

60. 64

61.

a. Yes
b. Yes
c. Yes
d. 0.6

62. 400

63. 2.5

64. 25

65. 0.0087

66. 0.0064, 0.0064

67.

a. It has no effect.
b. It is divided by 2–√2.
c. It is divided by 2.

68.

69.

70. 0.955

71. 0.927

72. 0.648

73. 0.101

74. 0.273

CHAPTER OVERVIEW

6: Interval Estimates

6.1: Introduction to Confidence Intervals

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 the parameter p in the binomial probability density function.

Figure 6.1.1 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)

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. What statistics provides us beyond a simple average, or point estimate, is an estimate to which we can attach a measure of accuracy, what we will call a confidence level. We make inferences with a known level of confidence.

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 (until a specifi interval is calculated). 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, $\overline{x}$, and the sample standard deviation, s. You would use $\overline{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\overline{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, s, is the **point estimate** for the population standard deviation, σ.

$\overline{x}$ and s are each 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 the 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 of the sampling distribution of the sample means is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **Empirical Rule**, which applies to the normal distribution, says that in approximately 95% of the samples, the sample mean, $\overline{x}$, 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 $\overline{x}$ is likely to be within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\overline{x}$ with 95% probability. 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 $\overline{x}-0.2$ and $\overline{x}+0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\overline{x}=2$. Then with 95% **confidence** the unknown population mean μ is between

$$\overline{x}-0.2=2-0.2=1.8 \text{ and } \overline{x}+0.2=2+0.2=2.2$$

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).** Please note that we talked in terms of 95% confidence using the Empirical Rule. The Empirical Rule for two standard deviations is only approximately 95% of the probability under the normal distribution. To be precise, two standard deviations under a normal distribution is actually 95.44% of the probability. To calculate the exact 95% confidence level we would use 1.96 standard deviations.

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ, or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (100% minus 95% = 5%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ.

For the confidence interval for a mean the formula would be:

$$\mu=\overline{X} \pm z_{\alpha/2} \sigma / \sqrt{n}$$

Or written another way as:

\[\overline{X}-z_{\alpha/2} \frac{\sigma}{{\sqrt{n}} \leq \mu \leq \overline{X}+z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\nonumber\]

Where $\overline{X}$ is the sample mean. The **critical value**, $z_{\alpha/2}$, is determined by the level of confidence desired by the analyst, and $\sigma / \sqrt{n}$ is the standard deviation of the sampling distribution for means given to us by the Central Limit Theorem.

6.2: A Confidence Interval for a Population Mean- Standard Deviation Known or Large Sample Size

A confidence interval for a population mean with a known population standard deviation is based on the conclusion of the Central Limit Theorem that the sampling distribution of the sample mean follows an approximately normal distribution.

Calculating the Confidence Interval

Consider the standardizing formula for the sampling distribution developed in the discussion of the Central Limit Theorem:

$$Z=\frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}}=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

Notice that μ is substituted for $\mu_{\overline{x}}$ because we know that the expected value (mean) of $\overline{x}$ is μ from the Central Limit Theorem and $\sigma_{\overline{x}}$ is replaced with $\sigma/\sqrt{n}$, also from the Central Limit Theorem.

In this formula, given a specific sample, we will know $\overline{X}$, $\sigma_{\overline{x}}$ and n, the sample size. (In actuality we do not know the population standard deviation, but we do have a point estimate for it, s, from the sample we took. More on this later.) What we do not know is μ or Z. We can solve for either one of these in terms of the other. Solving for μ in terms of Z gives:

$$\mu=\overline{X}\pm Z\frac{\sigma}{\sqrt{n}}$$

Remembering that the Central Limit Theorem tells us that the distribution of $\overline{X}$, the sampling distribution for the sample mean, is normal, and that the normal distribution is symmetrical, we can rearrange terms as follows:

$$\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

This is the formula for a confidence interval for the mean of a population.

Notice that $z_{\alpha/2}$ has been substituted for Z in this equation. This is where a choice must be made by the statistician. The analyst must decide the **level of confidence** they wish to impose on the confidence interval. The value, α is the probability that the interval will not contain the true population mean. The confidence level is defined as $(1-\alpha)$. The **critical value**, $z_{\alpha/2}$ is the number of standard deviations $\overline{X}$ lies from the mean with a certain probability. If we chose $z_{\alpha/2}=1.96$, we are asking for the 95% confidence interval because we are setting the probability that the true mean lies within the range at 0.95. If we set $z_{\alpha/2}$ at 1.64, we are asking for the 90% confidence interval because we have set the probability at 0.90. These numbers can be verified by consulting the Standard Normal Table. Divide either 0.95 or 0.90 in half and find that probability inside the body of the table. Then read on the top and left margins the number of standard deviations it takes to get this level of probability.

In reality, we can set whatever level of confidence we desire simply by changing the $z_{\alpha/2}$ value in the formula. It is the analyst's choice. Common convention in Economics and most social sciences sets confidence intervals at either 90, 95, or 99 percent levels. Levels less than 90% are considered of little value. The level of confidence of a particular interval estimate is called by $(1-\alpha)$.

Let us again consider the example in the introduction concerning the number of downloads from iTunes. That case was for a 95% confidence interval, but other levels of confidence could have just as easily been chosen depending on the need of the analyst. However, the level of confidence MUST be pre-set and not subject to revision as a result of the calculations. For this example, let's say we know that the actual population mean number of iTunes downloads is 2.1. The true population mean falls within the range of the 95% confidence interval. There is absolutely nothing to guarantee that this will happen. **Further, if the true mean falls outside of the interval we will never know it. We must always remember that we will never ever know the true mean.** Statistics simply allows us, with a given level of probability (confidence), to say that the true mean is within the range calculated. This is referred to as the "level of ignorance admitted".

Changing the Confidence Level or Sample Size

Here again is the formula for a confidence interval for an unknown population mean assuming we know the population standard deviation:

$$\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

It is clear that the confidence interval is driven by two things, the chosen level of confidence, $z_{\alpha/2}$, and the standard deviation of the sampling distribution. The standard deviation of the sampling distribution is further affected by two things, the standard deviation of the population and the sample size we chose for our data. Here we wish to examine the effects of each of the choices we have made on the calculated confidence interval, the confidence level and the sample size.

For a moment we should ask just what we desire in a confidence interval. Our goal was to estimate the population mean from a sample. We have forsaken the hope that we will ever find the true population mean, and population standard deviation for that matter, for any case except where we have an extremely small population and the cost of gathering the data of interest is very small. In all other cases we must rely on samples. With the Central Limit Theorem we have the tools to provide a meaningful confidence interval with a given level of confidence, meaning a known probability of being wrong. By meaningful confidence interval we mean one that is useful. Imagine that you are asked for a confidence interval for the ages of your classmates. You have taken a sample and find a mean of 19.8 years. You wish to be very confident so you report an interval between 9.8 years and 29.8 years. This interval would certainly contain the true population mean and have a very high confidence level. However, it hardly qualifies as meaningful. **The very best confidence interval is narrow while having high confidence.** There is a natural tension between these two goals. The higher the level of confidence the wider the confidence interval as the case of the students' ages above. We can see this tension in the equation for the confidence interval.

$$\mu=\overline{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

The confidence interval will increase in width as $z_{\alpha/2}$ increases, $z_{\alpha/2}$ increases as the level of confidence increases. There is a tradeoff between the level of confidence and the width of the interval.

Now let's look at the formula again and we see that the sample size also plays an important role in the width of the confidence interval. The sample size, n, shows up in the denominator of the standard deviation of the sampling distribution. As the sample size increases, the standard deviation of the sampling distribution decreases and thus the width of the confidence interval, while holding constant the level of confidence. Again we see the importance of having large samples for our analysis although we then face a second constraint, the cost of gathering data.

Calculating the Confidence Interval: An Alternative Approach

Another way to approach confidence intervals is through the use of something called the **Margin of Error**. The margin of error gets its name from the recognition that it provides the boundary of the interval derived from the standard error of the sampling distribution. In the equations above it is seen that the interval is simply the estimated mean, sample mean, plus or minus something. That something is the margin of error and is driven by the level of confidence we desire to maintain in our estimate, $z_{\alpha/2}$, times the standard deviation of the sampling distribution. The margin of error for a mean is given the name, **Error Bound Mean**, or EBM.

To construct a confidence interval for a single unknown population mean μ, **where the population standard deviation is known**, we need $\overline{X}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated **EBM**). The sample mean \(\overline X) is the **point estimate** of the unknown population mean μ.

The confidence interval estimate will have the form:

(point estimate - error bound, point estimate + error bound) or, in symbols, $(\overline{x}-EBM, \overline{x}+EBM)$

The mathematical formula for this confidence interval is:

$$\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.2.1}$$

The margin of error (EBM) depends on the **confidence level** (abbreviated **CL**). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called "alpha", (α). This quantity α is related to the confidence level, CL. The value α is the probability that the interval does not contain the unknown population parameter.
Mathematically, $1-\alpha=CL$.

A confidence interval for a population mean with a **known** standard deviation is based on the fact that the sampling distribution of the sample mean follows an approximately normal distribution. Suppose that our sample has a mean of $\overline{x}=10$, and we have constructed the 90% confidence interval $(5,15)$ where $EBM=5$.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha=10$ in both tails, or 5% in each tail, of the normal distribution.

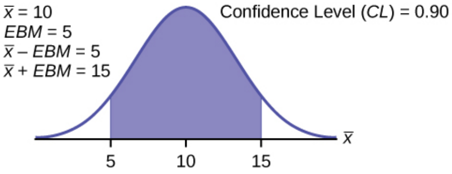


Figure 6.2.1

To capture the central 90%, we must go out 1.645 standard deviations on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail. It is denoted as $z_{0.05}=1.645$.

It is important that the standard deviation used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to the sampling distribution for means which we studied with the Central Limit Theorem and is, $\frac{\sigma}{\sqrt{n}}$.

Calculating the Confidence Interval Using EMB

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\overline{x}$ from the sample data. Remember, in this section we know the population standard deviation σ.
- Find the z-score from the standard normal table that corresponds to the confidence level desired, $z_{\alpha/2}$.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem.

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z\sim N(0,1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. Given that $CL=1-\alpha$, we know that α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right equal to $\frac{\alpha}{2}$ is denoted by $z_{\alpha/2}$.

For example, when $CL=0.95$, $\alpha=0.05$ and $\frac{\alpha}{2}=0.025$; we write $z_{\alpha/2}$ = z_{0.025}\).

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1-0.025=0.975$ and so the area from 0 and left of $z_{0.025}$ is $0.975-0.05=0.475$.

$z_{\alpha/2}=z_{0.025}=1.96$, using a standard normal probability table.

Calculating the Error Bound (EBM)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

- $EBM=z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(\overline{x} - EBM, \overline{x} + EBM)$ or the formula: $\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

The graph gives a picture of the entire situation.

$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$.

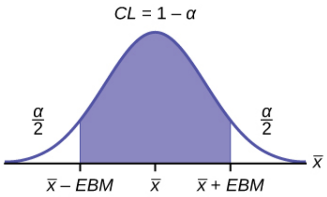


Figure 6.2.2

Example 6.2.1

Suppose we are interested in the mean for scores on an exam. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68 ($\overline{x} = 68$). In this example we have the unusual knowledge that the population standard deviation is $\sigma = 3$ points. Do not count on knowing the population parameters outside of textbook examples. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Specifically, find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

To find the confidence interval, you need the sample mean, $\overline{x}$, and the EBM.

- $\overline{x} = 68$
- $EBM = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; $n = 36$; The confidence level is 90% $(CL = 0.90)$

$CL = 0.90$ so $\alpha = 1 – CL = 1 – 0.90 = 0.10$

$\frac{\alpha}{2} = 0.05, z_{\alpha/2} = z_{0.05}$

The area to the right of $z_{0.05}$ is $0.05$and the area to the left of $Z_{0.05}$ is $1 – 0.05 = 0.95$

$z_{\alpha/2} = z_{0.05} = 1.645$

This can be found using a computer, or using a probability table for the standard normal distribution. Because the common levels of confidence in the social sciences are 90%, 95% and 99% it will not be long until you become familiar with the numbers , 1.645, 1.96, and 2.56.

$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$

$\overline{x} - EBM = 68 - 0.8225 = 67.1775$

$\overline{x} + EBM = 68 + 0.8225 = 68.8225$

The 90% confidence interval is **(67.1775, 68.8225).**

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Example 6.2.2

Suppose we change the original problem in Example 6.2.1 by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

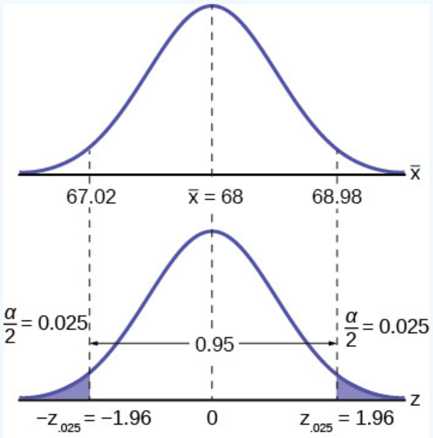


Figure 6.2.3

$$\mu = \overline{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3$; $n = 36$; The confidence level is 95% ($CL = 0.95$).

$CL = 0.95$ so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$z_{\alpha/2} = z_{0.025} = 1.96$

$$\mu = 68 \pm 1.96\left(\frac{3}{\sqrt{36}}\right)$$

$$67.02 \leq \mu \leq 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider. This demonstrates a very important principle of confidence intervals. There is a trade off between the level of confidence and the width of the interval. Our desire is to have a narrow confidence interval, huge wide intervals provide little information that is useful. But we would also like to have a high level of confidence in our interval. This demonstrates that we cannot have both.

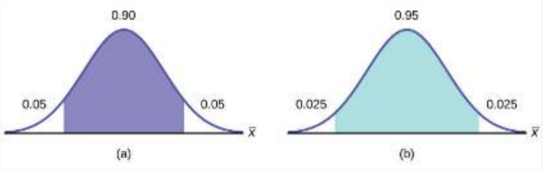


Figure 6.2.4

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level makes the confidence interval wider.
- Decreasing the confidence level makes the confidence interval narrower.

And again here is the formula for a confidence interval for an unknown mean assuming we have the population standard deviation:

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The standard deviation of the sampling distribution was provided by the Central Limit Theorem as $\frac{\sigma}{\sqrt{n}}$. While we infrequently get to choose the sample size it plays an important role in the confidence interval. Because the sample size is in the denominator of the equation, as n increases it causes the standard deviation of the sampling distribution to decrease and thus the width of the confidence interval to decrease. We have met this before as we reviewed the effects of sample size on the Central Limit Theorem. There we saw that as n increases the sampling distribution narrows until in the limit it collapses on the true population mean.

? Example 6.2.3

Suppose we change the original problem in Example 6.2.1 to see what happens to the confidence interval if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the confidence interval if we increase the sample size and use $n=100$ instead of $n=36$? What happens if we decrease the sample size to $n=25$ instead of $n=36$?

Answer

Solution A

$$\mu=\overline{x} \pm z_{\alpha / 2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\mu=68 \pm 1.645\left(\frac{3}{\sqrt{100}}\right)$$

$$67.5065 \leq \mu \leq 68.4935$$

If we **increase** the sample size n to 100, we **decrease** the width of the confidence interval relative to the original sample size of 36 observations.

Answer

Solution B

$$\mu=\overline{x} \pm z_{\alpha / 2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\mu=68 \pm 1.645\left(\frac{3}{\sqrt{25}}\right)$$

$$67.013 \leq \mu \leq 68.987$$

If we **decrease** the sample size n to 25, we **increase** the width of the confidence interval by comparison to the original sample size of 36 observations.

Summary: Effect of Changing the Sample Size

- Increasing the sample size makes the confidence interval narrower.
- Decreasing the sample size makes the confidence interval wider.

We have already seen this effect when we reviewed the effects of changing the size of the sample, *n*, on the Central Limit Theorem. Before we saw that as the sample size increased the standard deviation of the sampling distribution decreases. This was why we chose the sample mean from a large sample as compared to a small sample, all other things held constant.

Thus far we assumed that we knew the population standard deviation. This will virtually never be the case. We will have the sample standard deviation, *s*, however. This is a point estimate for the population standard deviation and can be substituted into the formula for confidence intervals for a mean under certain circumstances. We just saw the effect the sample size has on the width of confidence interval and the impact on the sampling distribution for our discussion of the Central Limit Theorem. We can invoke this to substitute the point estimate for the standard deviation if the sample size is large "enough". Simulation studies indicate that 30 observations or more will be sufficient to eliminate any meaningful bias in the estimated confidence interval.

Example 6.2.4

Spring break can be a very expensive holiday. A sample of 80 students is surveyed, and the average amount spent by students on travel and beverages is \$593.84. The sample standard deviation is approximately \$369.34.

Construct a 92% confidence interval for the population mean amount of money spent by spring breakers.

Answer

We begin with the confidence interval for a mean. We use the formula for a mean because the random variable is dollars spent and this is a continuous random variable. The point estimate for the population standard deviation, s, has been substituted for the true population standard deviation because with 80 observations there is no concern for bias in the estimate of the confidence interval.

$$\mu=\overline{x} \pm\left[z_{\alpha / 2} \frac{s}{\sqrt{n}}\right]$$

Substituting the values into the formula, we have:

$$\mu=593.84 \pm\left[1.75 \frac{369.34}{\sqrt{80}}\right]$$

Note that $z_{\alpha/2}$ is found on the standard normal table by looking up 0.46 in the body of the table and finding the number of standard deviations on the side and top of the table; 1.75. The solution for the interval is thus:

$$\mu=593.84 \pm 72.2636=(521.57,666.10)$$

$$\$521.58 \leq \mu \leq \$666.10$$

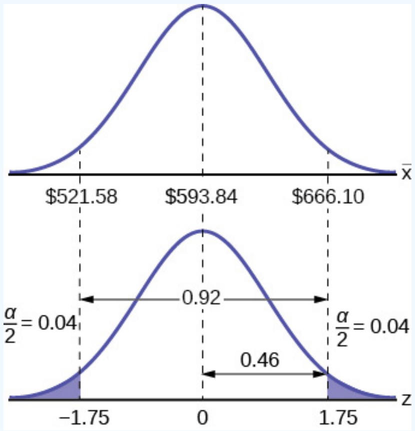


Figure 6.2.5

Formula Review

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by $\overline{X}-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ This formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha=1-CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\alpha/2}$ = the z-score with the property that the area to the right of the z-score is $\alpha/2$ this is the z-score used in the calculation of "EBM" where $\alpha=1-CL$.

- **8.1: A Confidence Interval for a Population Standard Deviation Known or Large Sample Size** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

6.3: A Confidence Interval for a Population Mean- Standard Deviation Unknown, Small Sample Case

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. This is what we did in an example in the previous section. The point estimate for the standard deviation, s, was substituted in the formula for the confidence interval for the population standard deviation. In this case there were 80 observations, well above the suggested 30 observations to eliminate any bias from a small sample. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's *t*-distribution**. The name comes from the fact that Gosset wrote under the pen name "A Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30 observations.

If you draw a simple random sample of size n from a population with mean μ and unknown population standard deviation σ and calculate the ***t*-score**

$$t=\frac{\overline{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} \tag{6.3.1}$$

then the *t*-scores follow a **Student's *t*-distribution with $\mathbf{n-1}$ degrees of freedom**. The *t*-score has the same interpretation as the *z*-score. It measures how far in standard deviation units $\overline{x}$ is from its mean μ. For each sample size n, there is a different Student's *t*-distribution.

The **degrees of freedom**, $\mathbf{n-1}$, come from the calculation of the sample standard deviation $\mathbf{s}$. Remember when we first calculated a sample standard deviation we divided the sum of the squared deviations by $n-1$, but we used n deviations ($x_i-\overline{x}$ values) to calculate $\mathbf{s}$. Because the sum of the deviations is zero, we can find the last deviation once we know the other $\mathbf{n-1}$ deviations. The other $\mathbf{n-1}$ deviations can change or vary freely. **We call the number $\mathbf{n-1}$ the degrees of freedom ($\boldsymbol{df}$)** in recognition that one is lost in the calculations. The effect of losing a degree of freedom is that the *t*-value increases and the confidence interval increases in width.

Properties of the Student's t-Distribution

- The graph for the Student's *t*-distribution is similar to the standard normal curve and at infinite degrees of freedom it is the normal distribution.
- The mean for the Student's *t*-distribution is 0 and the distribution is symmetric about 0, again like the standard normal distribution.
- The Student's *t*-distribution has more probability in its tails than the standard normal distribution because the spread of the *t*-distribution is greater than the spread of the standard normal. So the graph of the Student's *t*-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's *t*-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's *t*-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean \ mu and unknown population standard deviation σ. This assumption comes from the Central Limit Theorem because the individual observations in this case are the $\overline{x}$s of the sampling distribution. The size of the underlying population is generally not relevant unless it is very small. If it is normal then the assumption is met and doesn't need discussion.

A probability table for the Student's *t*-distribution is used to calculate *t*-values at various commonly-used levels of confidence. The table gives *t*-scores that correspond to the confidence level (column) and degrees of freedom (row). When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that at the bottom the table will show the *t*-value for infinite degrees of freedom. Mathematically, as the degrees of freedom increase, the t distribution approaches the standard normal distribution. You can find familiar Z-values by looking in the relevant α column and reading values in the last row.

The Student's t distribution has one of the most desirable properties of the normal: it is symmetrical. What the Student's t distribution does is spread out the horizontal axis so it takes a larger number of standard deviations to capture the same amount of probability. In reality there are an infinite number of Student's t distributions, one for each adjustment to the sample size. As the sample size increases, the Student's t distribution become more and more like the normal distribution. When the sample size reaches 30 the normal distribution is usually substituted for the Student's t because they are so much alike. This relationship between the Student's t distribution and the normal distribution is shown in Figure $6.3.1$.

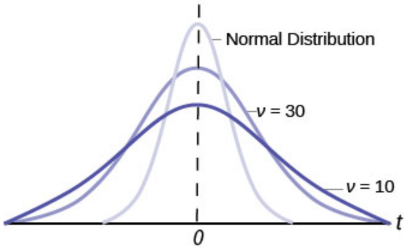


Figure 6.3.1

This is another example of one distribution limiting another one, in this case the normal distribution is the limiting distribution of the Student's t when the degrees of freedom in the Student's t approaches infinity. This conclusion comes directly from the derivation of the Student's t distribution by Mr. Gosset. He recognized the problem as having few observations and no estimate of the population standard deviation. He was substituting the sample standard deviation and getting volatile results. He therefore created the Student's t distribution as a ratio of the normal distribution and Chi squared distribution. The Chi squared distribution is itself a ratio of two variances, in this case the sample variance and the unknown population variance. The Student's t distribution thus is tied to the normal distribution, but has degrees of freedom that come from those of the Chi squared distribution. The algebraic solution demonstrates this result.

Development of Student's t-distribution:

1. $T=\frac{Z}{\sqrt{\frac{X^2}{v}}}$

   Where Z is the standard normal distribution and X^2 is the chi-squared distribution with v degrees of freedom.

2. $T=\frac{\frac{(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{\frac{S^2}{(n-1)}}{\frac{\sigma^2}{(n-1)}}}}$

   by substitution, and thus Student's t with $v=n-1$ degrees of freedom is:

3. $T=\frac{\overline{X}-\mu}{\frac{X}{\sqrt{n}}}$

Restating the formula for a confidence interval for the mean for cases when the sample size is smaller than 30 and we do not know the population standard deviation, σ:

$$\overline{x}-t_{\nu,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\leq\mu\leq\overline{x}+t_{\nu,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Here the point estimate of the population standard deviation, s, has been substituted for the population standard deviation, σ, and $t_{\nu,\alpha/2}$ has been substituted for $z_{\alpha/2}$. The Greek letter ν (pronounced "nu") is placed in the general formula in recognition that there are many Student t_ν distributions, one for each sample size. The symbol ν is the symbol for the degrees of freedom of the distribution and depends on the size of the sample. Often df is used to abbreviate degrees of freedom. **For this type of problem,** the degrees of freedom is $\nu=n-1$, where n is the sample size. To look up a probability in the Student's t table we have to know the degrees of freedom in the problem.

Example 6.3.1

The average earnings per share (EPS) for 10 industrial stocks randomly selected from those listed on the Dow-Jones Industrial Average was found to be $\overline{x}=1.85$ with a standard deviation of $s=0.395$. Calculate a 99% confidence interval for the average EPS of all the industrials listed on the $DJIA$.

$$\overline{x}-t_{\nu,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\leq\mu\leq\overline{x}+t_{\nu,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Answer

To help visualize the process of calculating a confidence interval we draw the appropriate distribution for the problem. In this case this is the Student's *t* because we do not know the population standard deviation and the sample is small, less than 30.

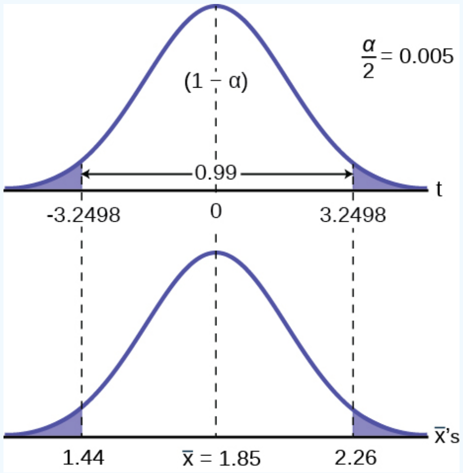


Figure 6.3.2

To find the appropriate *t*-value requires two pieces of information, the level of confidence desired and the degrees of freedom. The question asked for a 99% confidence level. On the graph this is shown where $(1-\alpha)$, the level of confidence , is in the unshaded area. The tails, thus, have .005 probability each, $\alpha/2$. The degrees of freedom for this type of problem is $n-1=9$. From the Student's *t* table, at the row marked 9 and column marked .005, is the number of standard deviations to capture 99% of the probability, 3.2498. These are then placed on the graph remembering that the Student's *t* is symmetrical and so the *t*-value is both plus or minus on each side of the mean.

Inserting these values into the formula gives the result. These values can be placed on the graph to see the relationship between the distribution of the sample mean, $\overline{X}$, and the Student's *t* distribution.

$$\mu=\overline{X}\pm t_{\alpha/2,\mathrm{df}=n-1}\frac{s}{\sqrt{n}}=1.85\pm 3.2498\frac{0.395}{\sqrt{10}}=1.85\pm 0.406$$

$$1.445\leq\mu\leq 2.257$$

We state the formal conclusion as :

With 99% confidence level, the average EPS of all the industries listed at $DJIA$ is from \$1.44 to \$2.26.

Exercise 6.3.2

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

6.4: A Confidence Interval for A Population 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.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell solar panels are interested in the proportion of households in Zimbabwe that have solar panels. 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 Zimbabwe that own solar panels.

The procedure to find the confidence interval for a population proportion is similar to that for the population mean, but the formulas are a bit different although conceptually identical. While the formulas are different, they are based upon the same mathematical foundation given to us by the Central Limit Theorem. Because of this we will see the same basic format using the same three pieces of information: the sample value of the parameter in question, the standard deviation of the relevant sampling distribution, and the number of standard deviations we need to have the confidence in our estimate that we desire.

How do you know you are dealing with a proportion problem? First, the underlying **distribution has a binary random variable and therefore is a binomial distribution**. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim \text{binomial}(n, p)$ where n is the number of trials and p is the probability of a success. To form a sample proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable $\hat{p}$ (read "p hat") is the sample proportion,

$$\hat{p} = \frac{X}{n}$$

(Sometimes the random variable is denoted as p', read "p prime".)

- $\hat{p}$ = the **estimated proportion** of successes or sample proportion of successes. ($\hat{p}$ is a **point estimate** for p, the true population proportion, and thus $q = 1 - p$ is the probability of a failure in any one trial.)
- X = the **number** of successes in the sample
- n = the size of the sample

The formula for the confidence interval for a population proportion follows the same format as that for an estimate of a population mean. Remembering the sampling distribution for the proportion from Chapter 5, the standard deviation was found to be:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

The confidence interval for a population proportion, therefore, becomes:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$z_{\alpha/2}$ is set according to our desired degree of confidence and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the standard deviation of the sampling distribution.

The **sample proportions $\hat{p}$ and $\hat{q}$ are estimates of the unknown population proportions p and q**. The estimated proportions $\hat{p}$ and $\hat{q}$ are used because p and q are not known.

Remember that as p moves further from 0.5 the binomial distribution becomes less symmetrical. Because we are estimating the binomial with the symmetrical normal distribution the further away from symmetrical the binomial becomes the less confidence we have in the estimate.

This conclusion can be demonstrated through the following analysis. Proportions are based upon the binomial probability distribution. The possible outcomes are binary, either “success” or “failure”. This gives rise to a proportion, meaning the percentage of the outcomes that are “successes”. It was shown that the binomial distribution could be fully understood if we knew only the probability of a success in any one trial, called p. The mean and the standard deviation of the binomial were found to be:

$$\mu = np$$

$$\sigma=\sqrt{npq}$$

It was also shown that the binomial could be estimated by the normal distribution if BOTH np AND nq were greater than 5. From the discussion above, it was found that the standardizing formula for the binomial distribution is:

$$Z=\frac{\hat{p}-p}{\sqrt{\left(\frac{pq}{n}\right)}}$$

which is nothing more than a restatement of the general standardizing formula with appropriate substitutions for μ and σ from the binomial. We can use the standard normal distribution, the reason Z is in the equation, because the normal distribution is the limiting distribution of the binomial. This is another example of the Central Limit Theorem. We have already seen that the sampling distribution of the sample mean is normally distributed. Recall the extended discussion in Chapter 5 concerning the sampling distribution of proportions and the conclusions of the Central Limit Theorem.

We can now manipulate this formula in just the same way we did for finding the confidence intervals for a mean, but to find the confidence interval for the binomial population parameter, p.

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\leq p\leq\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Where $\hat{p}=\frac{x}{n}$, the point estimate of p taken from the sample. Notice that $\hat{p}$ has replaced p in the formula. This is because we do not know p, indeed, this is just what we are trying to estimate.

Unfortunately, there is no correction factor for cases where the sample size is small so $n\hat{p}$ and $n\hat{q}$ must always be greater than 5 to develop an interval estimate for p.

? Example 6.4.1

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 sampled, 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

Let X = the number of people in the sample who have cell phones. X is binomial: the random variable is binary, people either have a cell phone or they do not.

To calculate the confidence interval, we must find $\hat{p}$, $\hat{q}$.

$n=500$

$x=$ the number of successes in the sample $=421$

$\hat{p}=\frac{x}{n}=\frac{421}{500}=0.842$

$\hat{p}=0.842$ is the sample proportion; this is the point estimate of the population proportion.

$\hat{q}=1-\hat{p}=1-0.842=0.158$

Since the requested confidence level is $CL=0.95$, then $\alpha=1-CL=1-0.95=0.05$ and $\frac{\alpha}{2}=0.025$.

Then $z_{\alpha/2}=z_{0.025}=1.96$

This can be found using the Standard Normal probability table.

The confidence interval for the true binomial population proportion is

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\leq p\leq\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Substituting in the values from above we find the confidence interval is: $0.810\leq p\leq 0.874$

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.

? Exercise 6.4.1

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

? Example 6.4.2

The Dundee Dog Training School has a larger than average proportion of clients who compete in competitive professional events. A confidence interval for the population proportion of dogs that compete in professional events from 150 different training schools is constructed. The lower limit is determined to be 0.08 and the upper limit is determined to be 0.16. Determine the level of confidence used to construct the interval of the population proportion of dogs that compete in professional events.

Answer

We begin with the formula for a confidence interval for a proportion because the random variable is binary; either the client competes in professional competitive dog events or they do not.

\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}\right]\nonumber

Next we find the sample proportion:

$$\hat{p}=\frac{0.08+0.16}{2}=0.12$$

The $\pm$ that makes up the confidence interval is thus 0.04; $0.12+0.04=0.16$ and $0.12-0.04=0.08$; the boundaries of the confidence interval. Finally, we solve for $z_{\alpha/2}$.

$z_{\alpha/2}\sqrt{\frac{0.12(1-0.12)}{150}}=0.04,$ $\textbf{therefore } \mathbf{z_{\alpha/2}=1.51}$

And then look up the probability for 1.51 standard deviations on the standard normal table.

$P(0<Z<1.51)=0.4345 \Rightarrow 0.4345\cdot 2=0.8690$ **or** 86.90%

? Example 6.4.3

A financial officer for a company wants to estimate the percent of accounts receivable that are more than 30 days overdue. He surveys 500 accounts and finds that 300 are more than 30 days overdue. Compute a 90% confidence interval for the true percent of accounts receivable that are more than 30 days overdue, and interpret the confidence interval.

Answer

$x=300$ and $n=500$

$\hat{p}=\frac{x}{n}=\frac{300}{500}=0.600$

$\hat{q}=1-\hat{p}=1-0.600=0.400$

Since confidence level = 0.90, then $\alpha=1-$ confidence level $=(1-0.90)=0.10$, and so $\left(\frac{\alpha}{2}\right)=0.05$

$z_{\alpha/2}=z_{0.05}=1.645$

This z-value can be found using a standard normal probability table. We use this formula for a confidence interval for a proportion:

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\leq p\leq\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Substituting in the values from above we find the confidence interval for the true binomial population proportion is $0.564\leq p\leq 0.636$

Interpretation

We estimate with 90% confidence that the true percent of all accounts receivable overdue 30 days is between 56.4% and 63.6%. Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL accounts are overdue 30 days.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of accounts receivable that are overdue 30 days.

? Exercise 6.4.2

a. A student polls her school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.
b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

6.5: Calculating the Sample Size n- Continuous and Binary Random Variables

Continuous Random Variables

Usually we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly in both time and product. Simple phone surveys will cost approximately $30.00 USD each, for example, and some sampling requires the destruction of the product.

If we go back to our standardizing formula for the sampling distribution for means, we can see that it is possible to solve it for n. If we do this we have $(\overline{X}-\mu)$ in the denominator.

$$n=\frac{z_{\alpha/2}^{2}\sigma^{2}}{(\overline{X}-\mu)^{2}}=\frac{z_{\alpha/2}^{2}\sigma^{2}}{\varepsilon^{2}}$$

Because we have not taken a sample yet we do not know any of the variables in the formula except that we can set $z_{\alpha/2}$ to the level of confidence we desire just as we did when determining confidence intervals. If we set a predetermined acceptable **error**, or **tolerance**, for the difference between $\overline{X}$ and μ, called ε in the formula, we are much further in solving for the sample size n. We still do not know the population standard deviation, σ. In practice, a pre-survey is usually done which allows for fine tuning the questionnaire and will give a sample standard deviation that can be used. In other cases, previous information from other surveys may be used for σ in the formula. While crude, this method of determining the sample size may help in reducing cost significantly. It will be the actual data gathered that determines the inferences about the population, so caution in the sample size is appropriate calling for high levels of confidence and small sampling errors.

Binary Random Variables

What was done in cases when looking for the mean of a distribution can also be done when sampling to determine the population parameter p for proportions. Manipulation of the standardizing formula for proportions gives:

$$n=\frac{z_{\alpha/2}^{2}pq}{\varepsilon^{2}}$$

where $\varepsilon=(\hat{p}-p)$, and is the acceptable sampling error, or tolerance, for this application. This will be measured in percentage points.

In this case the very object of our search is in the formula, p, and of course q because $q=1-p$. This result occurs because the binomial distribution is a one parameter distribution. If we know p then we know the mean and the standard deviation. Therefore, p shows up in the standard deviation of the sampling distribution which is where we got this formula. If, in an abundance of caution, we substitute 0.5 for p we will draw the largest required sample size that will provide the level of confidence specified by $z_{\alpha/2}$ and the tolerance we have selected. This is true because of all combinations of two fractions that add to 1, the largest multiple is when each is 0.5. Without any other information concerning the population parameter p, this is the common practice. This may result in oversampling, but certainly not under sampling, thus, this is a cautious approach.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. Table $6.5.1$ shows the appropriate sample size at different levels of confidence and different levels of the acceptable error, or tolerance.

Table 6.5.1

Required sample size (90%)	Required sample size (95%)	Tolerance level (ε)
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

This table is designed to show the maximum sample size required at different levels of confidence given an assumed $p = 0.5$ and $q = 0.5$ as discussed above.

The acceptable error, called tolerance in the table, is measured in plus or minus values from the actual proportion. For example, an acceptable error of 5% means that if the sample proportion was found to be 26 percent, the conclusion would be that the actual population proportion is between 21 and 31 percent with a 90 percent level of confidence if a sample of 271 had been taken. Likewise, if the acceptable error was set at 2%, then the population proportion would be between 24 and 28 percent with a 90 percent level of confidence, but would require that the sample size be increased from 271 to 1,691. If we wished a higher level of confidence, we would require a larger sample size. Moving from a 90 percent level of confidence to a 95 percent level at a plus or minus 5% tolerance requires changing the sample size from 271 to 384. A very common sample size often seen reported in political surveys is 384. With the survey results it is frequently stated that the results are good to a plus or minus 5% level of "accuracy".

❓ Example 6.5.1

Suppose a mobile phone company wants to determine the current percentage of customers aged 50 and over who use text messaging on their cell phones. How many customers aged 50 and over should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50 and over who use text messaging on their cell phones.

Answer

From the problem, we know that the acceptable error, ε, is **0.03** (3%=0.03) and $z_{\alpha/2} = z_{0.05} = 1.645$ because the confidence level is 90%. The acceptable error, ε, is the difference between the actual population proportion p, and the sample proportion we expect to get from the sample.

However, in order to find n, we need to know the estimated (sample) proportion $\hat{p}$. Remember that $\hat{q} = 1 - \hat{p}$. But, we do not know $\hat{p}$ yet. Since we multiply $\hat{p}$ and $\hat{q}$ together, we make them both equal to 0.5 because $\hat{p}\hat{q} = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within 3 percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{\varepsilon^2} \text{ gives } n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50 and over in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers aged 50 and over who use text messaging on their cell phones.

❓ Exercise 6.5.1

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within 5 percentage points of the true population proportion of customers who click on ads on their smartphones?

6.6: Confidence Intervals - Inflation Rates (Worksheet)

Work in groups on these problems. You should try to answer the questions without referring to your notes. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will be able to find *t* critical values using the Student's *t* Table.
- The student will be able to construct confidence intervals for a population mean and population proportion.
- The student will understand the factors that influence the width of a confidence interval.

The Data

Recall the data used in the Descriptive Statistics Activity, which was data from a 2014 African Financial Survey on a random sample of $n=18$ African countries. In this activity, we will focus on the **Inflation Rate** variable, which gives the inflation rate of the country in 2014, reported as a %. The data are provided in the following table.

African Financial Survey Data

Country	Inflation Rate	Country	Inflation Rate
Algeria	4.24	Mali	2.07
Cameroon	2.45	Mauritius	3.61
Cote d'Ivoire	2.86	Namibia	5.78
Equatorial Guinea	5.83	Rwanda	4.44
Ethiopia	7.89	Seychelles	4.17
Gabon	2.67	South Africa	5.68
Ghana	9.89	Swaziland	5.95
Lesotho	4.94	Togo	2.27
Madagascar	7.2	Zimbabwe	3.97

I. Finding *t* Critical Values

Before we calculate any confidence intervals, let's practice using the t table by calculating the following.

1. Find $t_{\text{df, }0.10}$ assuming a sample size of $n=3$.
2. Find $t_{\text{df, }0.025}$ assuming a sample size of $n=9$.
3. Find $t_{\text{df, }0.01}$ assuming a sample size of $n=12$.

II. Confidence Intervals for a Population Mean

We will calculate confidence intervals for the population mean inflation rate for African countries using the sample data provided.

Recall that the mean and standard deviation of the sample of inflation rates are equal to the following:

$$\bar{x}=4.77$$

$$s=2.11$$

1. First, using the following z-interval formula for a confidence interval of μ, construct a 95% confidence interval for the mean inflation rate of African countries. Note that σ is not known, so use the sample standard deviation instead.

$$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

2. Now, using the following t-interval formula for a confidence interval of μ, construct a 95% confidence interval for the mean inflation rate of African countries.

$$\bar{x} \pm t_{df,\alpha/2} \times \frac{s}{\sqrt{n}}$$

3. Compare the intervals you found in questions 1 and 2. Which one is wider? Briefly explain why.
4. Repeat questions 1 and 2, but this time use a confidence level of 90%.
5. Compare the 90% confidence intervals you computed in question 4 to the 95% confidence intervals you found in questions 1 and 2. Which intervals are wider? Briefly explain why.
6. Given that the sample size is less than 30, the population standard deviation is not known, and the histogram of the sample data is as below in Figure $6.6.1$, do you think a z- or t-interval is more appropriate?

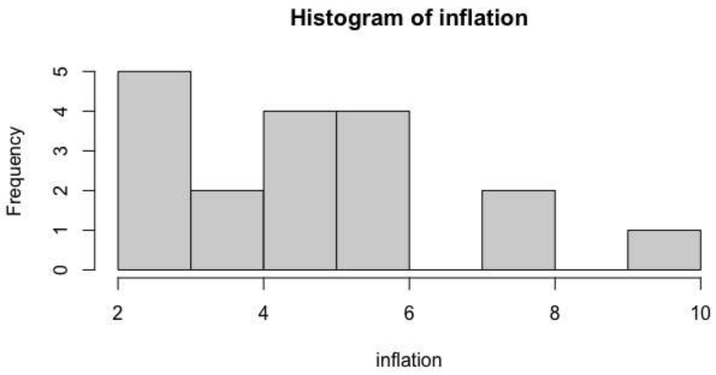


Figure 6.6.1

III. Confidence Interval for a Population Proportion

Many countries target an inflation rate of 2%; however, there is debate amongst economists whether the target needs to be that low for a healthy economy. Some economists are now arguing for countries to target a rate of 3%. Using this new target of 3% and the sample data, you will construct a confidence interval for the population proportion of African countries that have an inflation rate below 3%.

1. First, using the sample data, find a point estimate for the population proportion of African countries that have an inflation rate below 3%.
2. Now, use the following formula to construct a 95% confidence interval for the population proportion of African countries that have an inflation rate below 3%.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

3. What is the margin of error in the confidence interval you found in the previous question? Give 2 ways to **decrease** the margin of error.

IV. Interpret the Interval Estimates

1. Pick the confidence interval from the Section II questions that you think is the most appropriate estimate for the population mean inflation rate of African countries and write a sentence providing an interpretation of the interval.
2. Write a sentence providing an interpretation of the confidence interval you found for the population proportion of African countries with inflation rate below 3% in question 2 of Section III.

6.7: Chapter Key Terms

Key Term	Definition
Binomial Distribution	a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(\mathbf{n}, \mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.
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.
Confidence Level (CL)	the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.
Degrees of Freedom (df)	the number of objects in a sample that are free to vary
Error Bound for a Population Mean (EBM)	the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.
Error Bound for a Population Proportion (EBP)	the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.
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.
Normal Distribution	a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.
Parameter	a numerical characteristic of a population
Point Estimate	a single number computed from a sample and used to estimate a population parameter
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 \sigma for population standard deviation

Key Term	Definition
Student's t-Distribution	investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of this random variable (RV) are: • It is continuous and assumes any real values. • The pdf is symmetrical about its mean of zero. • It approaches the standard normal distribution as n get larger. • There is a "family" of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which depends upon the application for which the t is being used.

6.8: Chapter References

A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

- “American Fact Finder.” U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...html?refresh=t (accessed July 2, 2013).
- “Disclosure Data Catalog: Candidate Summary Report 2012.” U.S. Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2, 2013).
- “Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall.” Foothill De Anza Community College District. Available online at research.fhda.edu/factbook/FH...phicTrends.htm (accessed September 30,2013).
- Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. “2000 CDC Growth Charts for the United States: Methods and Development.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/growthcharts/2000...thchart-us.pdf (accessed July 2, 2013).
- La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).
- “Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates.” American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...prodType=table (accessed July 2, 2013).
- “Metadata Description of Candidate Summary File.” U.S. Federal Election Commission. Available online at www.fec.gov/finance/disclosur...esummary.shtml (accessed July 2, 2013).
- “National Health and Nutrition Examination Survey.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

- “America’s Best Small Companies.” Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).
- Data from *Microsoft Bookshelf*.
- Data from http://www.businessweek.com/.
- Data from http://www.forbes.com/.
- “Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2,2013).
- “Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at www.ewg.org/sites/humantoxome...tero%2Fnewborn (accessed July 2, 2013).
- “Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

A Confidence Interval for A Population Proportion

- Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons.” Public Policy Polling. Available online at www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).
- Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. “Teens, Social Media, and Privacy.” PewInternet, 2013. Available online at www.pewinternet.org/Reports/2...d-Privacy.aspx (accessed July 2, 2013).
- Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at www.pewinternet.org/~/media//...al%20Media.pdf (accessed July 2, 2013).
- Saad, Lydia. “Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity.” Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/th...ement-age.aspx (accessed July 2, 2013).
- The Field Poll. Available online at field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).
- Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/2...analytics-poll (accessed July 2, 2013).

- “52% Say Big-Time College Athletics Corrupt Education Process.” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/publ...cation_process (accessed July 2, 2013).

6.9: Chapter Formula Review

A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

s = the standard deviation of sample values.

$t=\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student's t-distribution

$df=n-1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student's t-distribution with df degrees of freedom

The general form for a confidence interval for a single mean, population standard deviation unknown, and sample size less than 30 Student's t is given by: $\overline{x}-t_{\mathrm{v}, \alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x}+t_{\mathrm{v}, \alpha}\left(\frac{s}{\sqrt{n}}\right)$

A Confidence Interval for A Population Proportion

$p^{\prime}=\frac{x}{n}$ where x represents the number of successes in a sample and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q^{\prime}=1-p^{\prime} \tag{6.9.1}$$

The variable $p^{\prime}$ has a binomial distribution that can be approximated with the normal distribution shown here. The confidence interval for the true population proportion is given by the formula:

$$\mathrm{p}^{\prime}-Z_\alpha \sqrt{\frac{\mathrm{p}^{\prime} \mathrm{q}^{\prime}}{n}} \leq p \leq \mathrm{p}^{\prime}+Z_\alpha \sqrt{\frac{\mathrm{p}^{\prime} \mathrm{q}^{\prime}}{n}} \tag{6.9.2}$$

$n=\frac{Z_{\frac{\alpha}{2}}^2 p^{\prime} q^{\prime}}{e^2}$ provides the number of observations needed to sample to estimate the population proportion, p, with confidence $1-\alpha$ and margin of error e. Where e = the acceptable difference between the actual population proportion and the sample proportion.

Calculating the Sample Size n: Continuous and Binary Random Variables

$n=\frac{Z^2 \sigma^2}{(\overline{x}-\mu)^2}$ = the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence for a continuous random variable

$n=\frac{Z_\alpha^2 \mathrm{pq}}{e^2}$ = the formula used to determine the sample size if the random variable is binary

---

- **8.8: Formula Review** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

6.10: Chapter Review

8.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of \sigma. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t=\frac{\bar{x}-\mu}{s / \sqrt{n}} \tag{6.10.1}$$

The t-score follows the Student's t-distribution with $n-1$ degrees of freedom. The confidence interval under this distribution is calculated with $\bar{x} \pm\left(t_{\frac{\alpha}{2}}\right) \frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

8.4 A Confidence Interval for A Population Proportion

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n, where x represents the number of successes and n represents the sample size. Let $q'=1-p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\mathrm{p}'-Z_\alpha \sqrt{\frac{\mathrm{p}' \mathrm{q}'}{n}} \leq p \leq \mathrm{p}'+Z_\alpha \sqrt{\frac{\mathrm{p}' \mathrm{q}'}{n}} \tag{6.10.2}$$

8.5 Calculating the Sample Size n: Continuous and Binary Random Variables

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the relevant confidence interval formula for n to discover the size of the sample that is needed to achieve this goal:

$$n=\frac{Z_\alpha^2 \sigma^2}{(\bar{x}-\mu)^2} \tag{6.10.3}$$

If the random variable is binary then the formula for the appropriate sample size to maintain a particular level of confidence with a specific tolerance level is given by

$$n=\frac{Z_\alpha^2 \mathrm{pq}}{e^2} \tag{6.10.4}$$

6.H: Interval Estimates (Homework + Practice)

8.2 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

102. In six packages of “The Flintstones® Real Fruit Snacks” there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice
c. Calculate p'.
d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

103. A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a. i. x−x- = ___________
   ii. sxs_x = ___________
   iii. n = ___________
   iv. n – 1 = ___________
b. Define the random variables XX and X−X- in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
   i. State the confidence interval.
   ii. Sketch the graph.
e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

104. Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. i. x−x- = ___________
   ii. sxs_x = ___________
   iii. n = ___________
   iv. n – 1 = ___________
b. Define the random variables XX and X−X- in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean time wasted.
   i. State the confidence interval.
   ii. Sketch the graph.
e. Explain in a complete sentence what the confidence interval means.

105. A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

a. i. x¯$\bar{x}$ = ___________
   ii. sxs_x = ___________
   iii. n = ___________
   iv. n – 1 = ___________
b. Define the random variable XX in words.

c. Define the random variable X−X- in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 95% confidence interval for the population mean length of time.
    i. State the confidence interval.
    ii. Sketch the graph.
f. What does it mean to be “95% confident” in this problem?

106. Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. i. x−x- = ___________
    ii. sxs_x = ___________
    iii. n = ___________
    iv. n – 1 = ___________
b. Define the random variable XX in words.
c. Define the random variableX−X- in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99% confidence interval for the population mean length of time using training wheels.
    i. State the confidence interval.
    ii. Sketch the graph.
f. Why would the error bound change if the confidence level were lowered to 90%?

107. The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates’ campaigns.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs.

Table 6.H.1

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80
$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

$\boxed{\text{x\_\bar}}$=$251,854.23

s= $521,130.41

Use this sample data to construct a 95% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

108. *Forbes* magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The Table 6.$H$.2 shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46

55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the Student's t-distribution.

109. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

a. i. $\bar{x}\bar{x}$ = ___________
  ii. $s_x s_x$ = ___________
  iii. n = ___________
  iv. n-1 = ___________

b. Define the random variables XX and $\bar{X}\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.

  i. State the confidence interval.
  ii. Sketch the graph.

110. In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable $\bar{X}\bar{X}$ in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.

  i. State the confidence interval.
  ii. Sketch the graph.

d. Explain what a "95% confidence interval" means for this study.

111. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.

  i. State the confidence interval.
  ii. Sketch the graph.

b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

112. A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

a. i. $\bar{x}\bar{x}$ = ___________
  ii. $s_x s_x$ = ___________
  iii. n = ___________

iv. n-1 = __________

b. Define the random variables XX and X−X- in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean worth of coupons.
   i. State the confidence interval.
   ii. Sketch the graph.
e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

113. Find the 95% Confidence Interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

8.4 A Confidence Interval for A Population Proportion

114. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

115. Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

116. According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

117. An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

118. Refer to the information in Table $6.H.5$ shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest \$100. The standard deviation for this data to the nearest hundred is σ = \$909,200.

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
   i. State the confidence interval.
   ii. Sketch the graph.
d. Explain what a “97% confidence interval” means for this study.

120. A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
   i. State the confidence interval.

  - ii. Sketch the graph.
- d. Suppose we want to lower the sampling error. What is one way to accomplish that?
- e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

121. Refer to Exercise 8.120. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

- a. Define the random variables X and P' in words.
- b. Which distribution should you use for this problem? Explain your choice.
- c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
  - i. State the confidence interval.
  - ii. Sketch the graph.
- d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

122. A point estimate for the true population proportion is:

- a. 0.90
- b. 1.27
- c. 0.79
- d. 400

123. A 90% confidence interval for the population proportion is _______.

- a. (0.761, 0.820)
- b. (0.125, 0.188)
- c. (0.755, 0.826)
- d. (0.130, 0.183)

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

124. Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

- a. (0.2975, 0.3796)
- b. (0.6270, 0.6959)
- c. (0.3041, 0.3730)
- d. (0.6204, 0.7025)

125. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is ______.

- a. 0.6614
- b. 0.3386
- c. 173
- d. 338

126. On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ±3% margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify *CL* and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

127. A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from part a and n = 1,000 to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

128. Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey’s outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

129. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

8.5 Calculating the Sample Size n: Continuous and Binary Random Variables

130. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a. i. x−x- =________
  ii. σ =________
  iii. n =________
b. In words, define the random variables X and X−X-.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean height of male Swedes.
  i. State the confidence interval.
  ii. Sketch the graph.
e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

131. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

a. In words, define the random variables X and X−X-.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population mean length of engineering conferences.
  i. State the confidence interval.
  ii. Sketch the graph.

132. Suppose that an accounting firm does a study to determine the time needed to complete one person’s tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. i. x−X- =_________
   ii. σ =_________
   iii. n =_________

b. In words, define the random variables X and X−X-.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90% confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.

e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?

g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

133. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. x−X- =_________
   ii. σ =_________
   iii. s_x =_________

b. In words, define the random variable X.

c. In words, define the random variable X−X-.

d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 90% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.

f. Construct a 98% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.

h. In complete sentences, give an interpretation of what the interval in part f means.

134. A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

a. i. x−X- =_________
   ii. σ =_________
   iii. n =_________

b. Define the random variables X and X−X- in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90% confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.

e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

135. What is meant by the term “90% confident” when constructing a confidence interval for a mean?

a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.

b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.

c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

136. The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. Table 8.5 shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

Table $6.H.3$

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

137. The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

138. The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

139. If the confidence interval is change to a higher probability, would this cause a lower, or a higher, minimum sample size?

140. If the tolerance is reduced by half, how would this affect the minimum sample size?

141. If the value of p is reduced, would this necessarily reduce the sample size needed?

142. Is it acceptable to use a higher sample size than the one calculated by $\frac{z^2pq}{e^2}$?

143. A company has been running an assembly line with 97.42%% of the products made being acceptable. Then, a critical piece broke down. After the repairs the decision was made to see if the number of defective products made was still close enough to the long standing production quality. Samples of 500 pieces were selected at random, and the defective rate was found to be 0.025%.

1.1: Chapter Practice

8.3 A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

1. Identify the following:

Identify the following:

a. x−x- =_______
b. sxsx =_______
c. n =_______
d. $n - 1$ =_______

2. Define the random variables X and X−X- in words.
3. Which distribution should you use for this problem?
4. Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.
5. Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

6. Identify the following:

a. x−x- =_______
b. sxsx =_______
c. n =_______
d. $n - 1$ =_______

7. Define the random variable X in words.
8. Define the random variable X−X- in words.
9. Which distribution should you use for this problem?
10. Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.
11. Why would the error bound change if the confidence level were lowered to 95%?

12. Use the following information to answer the next 13 exercises: The data in Table 1.1.2 are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

12. Calculate the following:

a. x−x- =______
b. sxsx =______
c. n =______

13. Define the random variable X−X- in words.
14. What is x−x- estimating?
15. Is σxσx known?
16. As a result of your answer to Exercise 8.15, state the exact distribution to use when calculating the confidence interval.

Construct a 95% confidence interval for the true mean number of colors on national flags

17. How much area is in both tails (combined)?

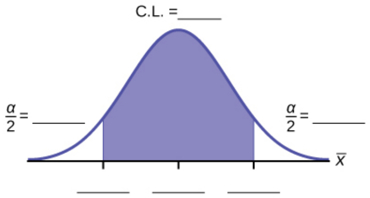
18. How much area is in each tail?

19. Calculate the following:

a. lower limit
b. upper limit
c. error bound

20. The 95% confidence interval is_____.
21. Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

mean.



22. In one complete sentence, explain what the interval means.
23. Using the same x−x-, sx$s x$, and level of confidence, suppose that *n* were 69 instead of 39. Would the error bound become larger or smaller? How do you know?
24. Using the same x−x-, sx$s x$, and *n* = 39, how would the error bound change if the confidence level were reduced to 90%? Why?

8.4 A Confidence Interval for A Population Proportion

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

25. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?
26. If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

27. Identify the following:

a. x = _______
b. n = _______
c. p' = _______

28. Define the random variables X and P' in words.
29. Which distribution should you use for this problem?
30. Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.
31. List two difficulties the company might have in obtaining random results, if this survey were done by email.

Identify the following:

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

32. We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and p' in words.

33. Which distribution should you use for this problem?

34. Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

35. Suppose we want to lower the sampling error. What is one way to accomplish that?

36. The sampling error given in the survey is ±2%. Explain what the ±2% means.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

37. Define the random variable X in words.

38. Define the random variable p' in words.

39. Which distribution should you use for this problem?

40. Construct a 90% confidence interval, and state the confidence interval and the error bound.

41. What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

42. What is being counted?

43. In words, define the random variable X.

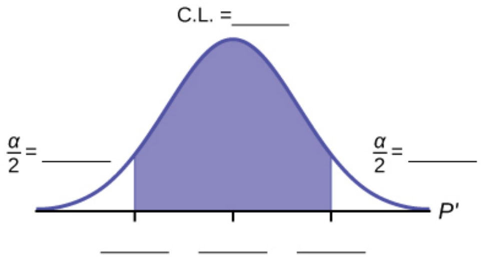
44. Calculate the following:

a. x = _______
b. n = _______
c. p' = _______

45. State the estimated distribution of X. X~________
46. Define a new random variable P'. What is p' estimating?
47. In words, define the random variable P'.
48. State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.
49. How much area is in both tails (combined)?
50. How much area is in each tail?
51. Calculate the following:

a. lower limit
b. upper limit
c. error bound

52. The 92% confidence interval is _______.
53. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.



54. In one complete sentence, explain what the interval means.
55. Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?
56. Using the same p' and n = 80, how would the error bound change if the confidence level were increased to 98%? Why?
57. If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

8.5 Calculating the Sample Size n: Continuous and Binary Random Variables

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

58. Identify the following:

59. In words, define the random variables *X* and X−*X*-.
60. Which distribution should you use for this problem?
61. Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.
62. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

63.

Identify the following:

a. x−x- = ______
b. σ = ______
c. n = ______

64. In words, define the random variables *X* and X−*X*-.
65. Which distribution should you use for this problem?
66. Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.
67. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
68. If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?
69. Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

70. Identify the following:

a. x−x- = _______
b. σ = _______
c. n = _______

71. In words, define the random variable X.
72. In words, define the random variable X−X-.
73. Which distribution should you use for this problem?
74. Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.
75. Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.
76. In complete sentences, explain why the confidence interval in Exercise 8.74 is larger than in Exercise 8.75.
77. In complete sentences, give an interpretation of what the interval in Exercise 8.75 means.
78. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?
79. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student

.80. $\bar{x}$ = _____

81. n = _____

82. _________ = 15

83. In words, define the random variable $\bar{X}$.

84. What is $\bar{x}$ estimating?

85. Is σ_x known?

86. As a result of your answer to **Exercise 1.1.83**, state the exact distribution to use when calculating the confidence interval.

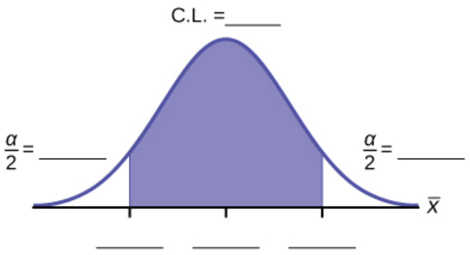
87. How much area is in both tails (combined)? α =_________

88. How much area is in each tail? $\frac{\alpha}{2}$ =_________

89. Identify the following specifications:

90. The 95% confidence interval is:____________________.

91. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.



92. In one complete sentence, explain what the interval means.

93. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

94. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

95. Find the value of the sample size needed to if the confidence interval is 90% that the sample proportion and the population proportion are within 4% of each other. The sample proportion is 0.60. Note: Round all fractions up for n.

96. Find the value of the sample size needed to if the confidence interval is 95% that the sample proportion and the population proportion are within 2% of each other. The sample proportion is 0.650. Note: Round all fractions up for n.

97. Find the value of the sample size needed to if the confidence interval is 96% that the sample proportion and the population proportion are within 5% of each other. The sample proportion is 0.70. Note: Round all fractions up for n.

98. Find the value of the sample size needed to if the confidence interval is 90% that the sample proportion and the population proportion are within 1% of each other. The sample proportion is 0.50. Note: Round all fractions up for n.

99. Find the value of the sample size needed to if the confidence interval is 94% that the sample proportion and the population proportion are within 2% of each other. The sample proportion is 0.65. Note: Round all fractions up for n.

100. Find the value of the sample size needed to if the confidence interval is 95% that the sample proportion and the population proportion are within 4% of each other. The sample proportion is 0.45. Note: Round all fractions up for n.

101. Find the value of the sample size needed to if the confidence interval is 90% that the sample proportion and the population proportion are within 2% of each other. The sample proportion is 0.3. Note: Round all fractions up for n.

6.S: Interval Estimates (Solutions- Practice + Homework)

2.

X is the number of hours a patient waits in the emergency room before being called back to be examined. X−X- is the mean wait time of 70 patients in the emergency room.

4.

CI: (1.3808, 1.6192)

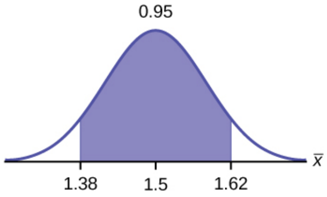


Figure **8.13**

EBM = 0.12

6.

- a. x−x- = 151
- b. sxsx = 32
- c. *n* = 108
- d. *n* – 1 = 107

8.

X−X- is the mean number of hours spent watching television per month from a sample of 108 Americans.

10.

CI: (142.92, 159.08)

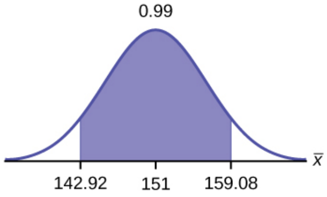


Figure **8.14**

EBM = 8.08

12.

- a. 3.26
- b. 1.02
- c. 39

14.

μ

16.

t_{38}

18.

0.025

20.

(2.93, 3.59)

22.

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

23.

The error bound would become EBM = 0.245. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

26.

The sample size needed would increase. As the confidence level increases, $\alpha\alpha$ decreases and $z(a2)z(a2)$ increases. To maintain the same error bound, the size of the sample needs to increase.

28.

X is the number of “successes” where the woman makes the majority of the purchasing decisions for the household. *P*′ is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

30.

CI: (0.5321, 0.6679)

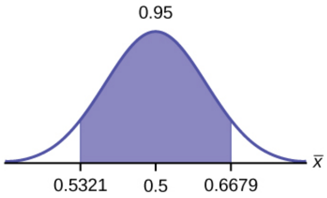


Figure **8.15**

EBM: 0.0679

32.

X is the number of “successes” where an executive prefers a truck. *P*′ is the percentage of executives sampled who prefer a truck.

34.

CI: (0.19432, 0.33068)

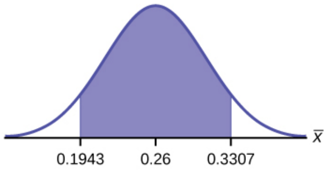


Figure **8.16**

36.

The sampling error means that the true mean can be 2% above or below the sample mean.

38.

P′ is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

40.

CI: (0.62735, 0.67265)

EBM: 0.02265

42.

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

44.

a. $x = 64$
b. $n = 80$
c. $p' = 0.8$

46.

p

48.

P′~N(0.8,(0.8)(0.2)80−−−−−−√)$P' \sim N(0.8,(0.8)(0.2)80)$. (0.72171, 0.87829).

50.

0.04

52.

(0.72; 0.88)

54.

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

56.

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

58.

a. 244
b. 15
c. 50

60.

N(244,1550√)$N(244,1550)$

62.

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

64.

X is the time in minutes it takes to complete the U.S. Census short form. X−X- is the mean time it took a sample of 200 people to complete the U.S. Census short form.

66.

CI: (7.9441, 8.4559)

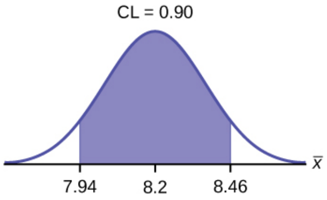


Figure **8.17**

68.

The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

70.

a. x−*x*- = 2.2
b. σ = 0.2
c. n = 20

72.

X−*X*- is the mean weight of a sample of 20 heads of lettuce.

74.

EBM = 0.07
CI: (2.1264, 2.2736)

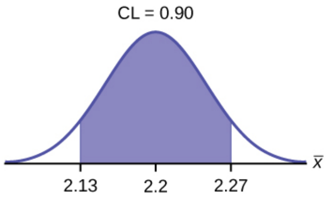


Figure **8.18**

76.

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

78.

The confidence level would increase.

80.

30.4

82.

σ

84.

μ

86.

normal

88.

0.025

90.

(24.52,36.28)

92.

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

94.

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

96.

2,185

98.

6,765

100.

595

103.

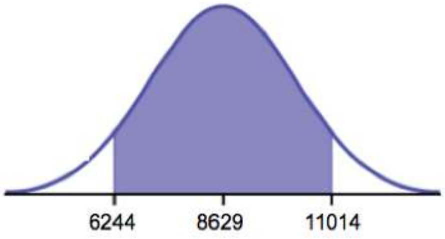
- a. i. 8629
  - ii. 6944
  - iii. 35
  - iv. 34
- b. t_{34}
- c. i. CI: (6244, 11,014)
  - ii. 

    Figure **8.19**
- d. It will become smaller

105.

- a. i. $\bar{x}$ = 2.51
  - ii. s_x = 0.318
  - iii. n = 9
  - iv. n - 1 = 8
- b. the effective length of time for a tranquilizer
- c. the mean effective length of time of tranquilizers from a sample of nine patients
- d. We need to use a Student's-t distribution, because we do not know the population standard deviation.
- e. i. CI: (2.27, 2.76)
  - ii. Check student's solution.
- f. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

107.

$\bar{x}$=$251,854.23

s= $521,130.41

Note that we are not given the population standard deviation, only the standard deviation of the sample.

There are 30 measures in the sample, so n = 30, and df = 30 - 1 = 29

CL = 0.96, so α = 1 - CL = 1 - 0.96 = 0.04

$\frac{\alpha}{2}=0.02 \; t_{\frac{\alpha}{2}}=t_{0.02}$ = 2.150

$EBM=t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)=2.150\left(\frac{521,130.41}{\sqrt{30}}\right)$ ~ $204,561.66

$\bar{x}$ - EBM = $251,854.23 - $204,561.66 = $47,292.57

$\bar{x}$ + EBM = $251,854.23+ $204,561.66 = $456,415.89

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between $47,292.57 and $456,415.89.

109.

- a.
  - i. $\bar{x} = 11.6$
  - ii. $s_x = 4.1$
  - iii. $n = 225$
  - iv. $n - 1 = 224$
- b. X is the number of unoccupied seats on a single flight. $\bar{X}$ is the mean number of unoccupied seats from a sample of 225 flights.
- c. We will use a Student’s-t distribution, because we do not know the population standard deviation.
- d.
  - i. CI: (11.12 , 12.08)
  - ii. Check student's solution.

111.

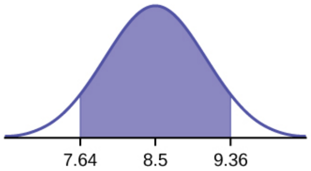
- a.
  - i. CI: (7.64 , 9.36)
  - ii. 

    Figure **8.20**
- b. The sample should have been increased.
- c. Answers will vary.
- d. Answers will vary.
- e. Answers will vary.

113.

b

114.

- a. 1,068
- b. The sample size would need to be increased since the critical value increases as the confidence level increases.

116.

- a. X = the number of people who feel that the president is doing an acceptable job;

  P' = the proportion of people in a sample who feel that the president is doing an acceptable job.
- b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
- c.
  - i. CI: (0.59, 0.63)
  - ii. Check student’s solution

118.

- a.
  - i. (0.72, 0.82)
  - ii. (0.65, 0.76)
  - iii. (0.60, 0.72)
- b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
- c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a White person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
- d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a White person into their families and the proportion of Asian adults who say that their families would welcome a Black person into their families.

120.

a. X = the number of adult Americans who feel that crime is the main problem; P' = the proportion of adult Americans who feel that crime is the main problem

b. Since we are estimating a proportion, given P' = 0.2 and n = 1000, the distribution we should use is $N\left(0.2,\sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c. i. CI: (0.18, 0.22)
   ii. Check student’s solution.

d. One way to lower the sampling error is to increase the sample size.

e. The stated “± 3%” represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

122.

c

125.

a

127.

a. $p' = \frac{(0.55 + 0.49)}{2} = 0.52$; $EBP = 0.55 - 0.52 = 0.03$

b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.

c. $CL = 0.75$, so $\alpha = 1 – 0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125$ $z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is $1 – 0.125 = 0.875$.)
$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1,000}} \approx 0.018$
$(p' - EBP, p' + EBP) = (0.52 – 0.018, 0.52 + 0.018) = (0.502, 0.538)$

d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

130.

a. i. 71
   ii. 2.8
   iii. 48

b. X is the height of a male Swede, and $\bar{x}$ is the mean height from a sample of 48 male Swedes.

c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.85)
   ii.

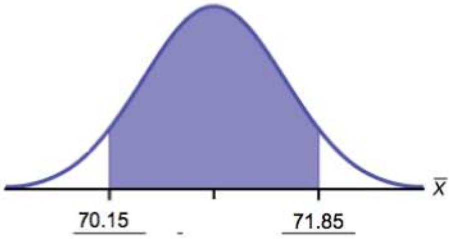


Figure **8.21**

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

132.

a. i. $\bar{x}$ = 23.6
   ii. σ = 7
   iii. n = 100

b. X is the time needed to complete an individual tax form. $\bar{X}$ is the mean time to complete tax forms from a sample of 100 customers.

c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.

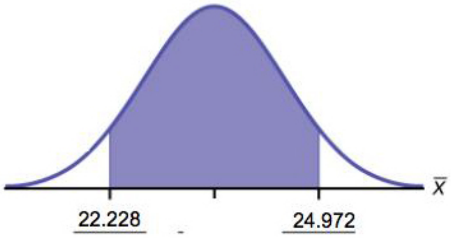
d.
  i. (22.228, 24.972)
  ii.


Figure **8.22**

e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.

f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.

g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

134.

a.
  i. 7.9
  ii. 2.5
  iii. 20

b. X is the number of letters a single camper will send home. $\bar{X}$ is the mean number of letters sent home from a sample of 20 campers.

c. $N\ 7.9\left(\frac{2.5}{\sqrt{20}}\right)$

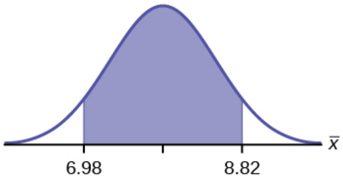
d.
  i. CI: (6.98, 8.82)
  ii.


Figure **8.23**

e. The error bound and confidence interval will decrease.

136.

a. $\bar{x}$ = \$568,873

b. $CL = 0.95$ $\alpha = 1 - 0.95 = 0.05$ $z_{\frac{\alpha}{2}} = 1.96$
$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\frac{909200}{\sqrt{40}}$ = \$281,764

c. $\bar{x} - EBM = 568,873 - 281,764 = 287,109$
$\bar{x} + EBM = 568,873 + 281,764 = 850,637$

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between \$287,109 and \$850,637.

139.

Higher

140.

It would increase to four times the prior value.

141.

No, It could have no affect if it were to change to 1 – p, for example. If it gets closer to 0.5 the minimum sample size would increase.

142.

Yes

143.

a. No
b. No

CHAPTER OVERVIEW

7: Hypothesis Testing

7.1: Introduction to Hypothesis Testing

Now we are down to the bread and butter work of the statistician: developing and testing hypotheses. It is important to put this material in a broader context so that the method by which a hypothesis is formed is understood completely. Using textbook examples often clouds the real source of statistical hypotheses.

Statistical testing is part of a much larger process known as the scientific method. This method was developed more than two centuries ago as the accepted way that new knowledge could be created. Until then, and unfortunately even today, among some, "knowledge" could be created simply by some authority saying something was so, *ipso dicta*. Superstition and conspiracy theories were (are?) accepted uncritically.

Figure 7.1.1 You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

The scientific method, briefly, states that only by following a careful and specific process can some assertion be included in the accepted body of knowledge. This process begins with a set of assumptions upon which a theory, sometimes called a model, is built. This theory, if it has any validity, will lead to predictions; what we call hypotheses.

As an example, in Microeconomics the theory of consumer choice begins with certain assumptions concerning human behavior. From these assumptions a theory of how consumers make choices using indifference curves and the budget line. This theory gave rise to a very important prediction, namely, that there was an inverse relationship between price and quantity demanded. This relationship was known as the demand curve. The negative slope of the demand curve is really just a prediction, or a hypothesis, that can be tested with statistical tools.

Unless hundreds and hundreds of statistical tests of this hypothesis had not confirmed this relationship, the so-called Law of Demand would have been discarded years ago. This is the role of statistics, to test the hypotheses of various theories to determine if they should be admitted into the accepted body of knowledge; how we understand our world. Once admitted, however, they may be later discarded if new theories come along that make better predictions.

Not long ago two scientists claimed that they could get more energy out of a process than was put in. This caused a tremendous stir for obvious reasons. They were on the cover of *Time* and were offered extravagant sums to bring their research work to private industry and any number of universities. It was not long until their work was subjected to the rigorous tests of the scientific method and found to be a failure. No other lab could replicate their findings. Consequently they have sunk into obscurity and their theory discarded. It may surface again when someone can pass the tests of the hypotheses required by the scientific method, but until then it is just a curiosity. Many pure frauds have been attempted over time, but most have been found out by applying the process of the scientific method.

This discussion is meant to show just where in this process statistics falls. Statistics and statisticians are not necessarily in the business of developing theories, but in the business of testing others' theories. Hypotheses come from these theories based upon an explicit set of assumptions and sound logic. The hypothesis comes first, before any data are gathered. Data do not create hypotheses; they are used to test them. If we bear this in mind as we study this section the process of forming and testing hypotheses will make more sense.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about the value of a specific parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "**hypothesis testing**." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

7.2: Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

- H_0: **The null hypothesis:** It is a statement of no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0. 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 cannot accept H_0. The alternative hypothesis is the contender and must win with significant evidence to overthrow the status quo. This concept is sometimes referred to the tyranny of the status quo because as we will see later, to overthrow the null hypothesis takes usually 90 or greater confidence that this is the proper decision.

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 "cannot accept 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. These conclusions are all based upon a level of probability, a significance level, that is set my the analyst.

Table 9.1 presents the various hypotheses in the relevant pairs. For example, if the null hypothesis is equal to some value, the alternative has to be not equal to that value.

Table 9.1

H_0	H_a
equal (=)	not equal ($\neq$)
greater than or equal to ($\geq$)	less than (<)
less than or equal to ($\leq$)	more than (>)

Note

As a mathematical convention H_0 always has a symbol with an equal in it. Ha never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test.

Example 9.1

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. $p > 30$

Example 9.2

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:
$H_0 : \mu = 2.0$
$H_a : \mu \neq 2.0$

Example 9.3

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:
$H_0 : \mu \geq 5$
$H_a : \mu < 5$

7.3: Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

Table 9.2

Statistical Decision	**$\mathrm{H_0}$ is actually...**	
	True	**False**
Cannot accept H_0	Type I error	Correct outcome
Cannot reject H_0	Correct outcome	Type II error

The four possible outcomes in the table are:

1. The decision is **cannot reject $\mathbf{H_0}$** when **$\mathbf{H_0}$ is true (correct decision).**
2. The decision is **cannot accept $\mathbf{H_0}$** when **$\mathbf{H_0}$ is true** (incorrect decision known as a **Type I error**). This case is described as "rejecting a good null". As we will see later, it is this type of error that we will guard against by setting the probability of making such an error. The goal is to NOT take an action that is an error.
3. The decision is **cannot reject $\mathbf{H_0}$** when, in fact, **$\mathbf{H_0}$ is false** (incorrect decision known as a **Type II error**). This is called "accepting a false null". In this situation you have allowed the status quo to remain in force when it should be overturned. As we will see, the null hypothesis has the advantage in competition with the alternative.
4. The decision is **cannot accept $\mathbf{H_0}$** when **$\mathbf{H_0}$ is false** (**correct decision**).

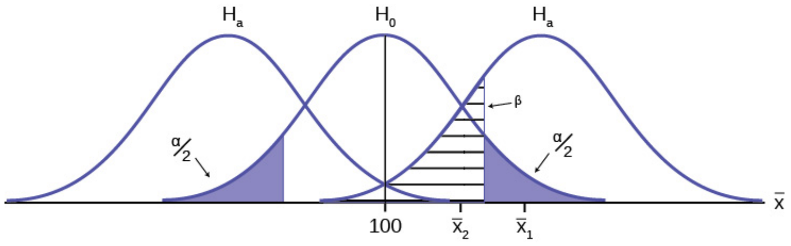
Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

- α = probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true: rejecting a good null.
- β = probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false. $(1-\beta)$ is called the **Power of the Test**.

α and β should be as small as possible because they are probabilities of errors.

Statistics allows us to set the probability that we are making a Type I error. The probability of making a Type I error is α. Recall that the confidence intervals in the last unit were set by choosing a value called Z_α (or t_α) and the alpha value determined the confidence level of the estimate because it was the probability of the interval failing to capture the true mean (or proportion parameter p). This alpha and that one are the same.

The easiest way to see the relationship between the alpha error and the level of confidence is with the following figure.



H_0: μ = 100
H_a: μ ≠ 100

Figure 9.2

In the center of Figure 9.2 is a normally distributed sampling distribution marked H_0. This is a sampling distribution of $\overline{X}$ and by the Central Limit Theorem it is normally distributed. The distribution in the center is marked H_0 and represents the distribution for the null hypotheses H_0: $\mu = 100$. This is the value that is being tested. The formal statements of the null and alternative hypotheses are listed below the figure.

The distributions on either side of the H_0 distribution represent distributions that would be true if H_0 is false, under the alternative hypothesis listed as Ha. We do not know which is true, and will never know. There are, in fact, an infinite number of distributions from which the data could have been drawn if Ha is true, but only two of them are on Figure 9.2 representing all of the others.

To test a hypothesis we take a sample from the population and determine if it could have come from the hypothesized distribution with an acceptable level of significance. This level of significance is the alpha error and is marked on Figure 9.2 as the shaded areas in each tail of the H_0 distribution. (Each area is actually \alpha/2 because the distribution is symmetrical and the alternative hypothesis allows for the possibility for the value to be either greater than or less than the hypothesized value--called a two-tailed test).

If the sample mean marked as $\overline{X}_1$ is in the tail of the distribution of H_0, we conclude that the probability that it could have come from the H_0 distribution is less than alpha. We consequently state, "the null hypothesis cannot be accepted with (\alpha) level of significance". The truth **may** be that this $\overline{X}_1$ did come from the H_0 distribution, but from out in the tail. If this is so then we have falsely rejected a true null hypothesis and have made a Type I error. What statistics has done is provide an estimate about what we know, and what we control, and that is the probability of us being wrong, α.

We can also see in Figure 9.2 that the sample mean could be really from an Ha distribution, but within the boundary set by the alpha level. Such a case is marked as $\overline{X}_2$. There is a probability that $\overline{X}_2$ actually came from Ha but shows up in the range of H_0 between the two tails. This probability is the beta error, the probability of accepting a false null.

Our problem is that we can only set the alpha error because there are an infinite number of alternative distributions from which the mean could have come that are not equal to H_0. As a result, the statistician places the burden of proof on the alternative hypothesis. That is, we will not reject a null hypothesis unless there is a greater than 90, or 95, or even 99 percent probability that the null is false: the burden of proof lies with the alternative hypothesis. This is why we called this the tyranny of the status quo earlier.

By way of example, the American judicial system begins with the concept that a defendant is "presumed innocent". This is the status quo and is the null hypothesis. The judge will tell the jury that they can not find the defendant guilty unless the evidence indicates guilt beyond a "reasonable doubt" which is usually defined in criminal cases as 95% certainty of guilt. If the jury cannot accept the null, innocent, then action will be taken, jail time. The burden of proof always lies with the alternative hypothesis. (In civil cases, the jury needs only to be more than 50% certain of wrongdoing to find culpability, called "a preponderance of the evidence").

The example above was for a test of a mean, but the same logic applies to tests of hypotheses for all statistical parameters one may wish to test.

The following are examples of Type I and Type II errors.

? Example 9.4

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.)

This is a situation described as "accepting a false null".

? Example 9.5

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. This is the status quo and requires no action if it is true. If the null hypothesis cannot be accepted then action is required and the hospital will begin appropriate procedures.

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(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(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

? Exercise 9.5

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?

? Example 9.6

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. The status quo is that the claim is false. The burden of proof always falls to the person making the claim, in this case the Genetics Lab.

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, \alpha.

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, \beta.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

? Exercise 9.6

"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.

? Example 9.7

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.

7.4: One-Sample Test

Earlier, we discussed sampling distributions. Particular distributions are associated with hypothesis testing.We will perform hypotheses tests of a population mean using a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the sample size is small, where small is considered to be less than 30 observations.) We perform tests of a population proportion using a normal distribution when we can assume that the distribution is normally distributed. We consider this to be true if the sample proportion, p', times the sample size is greater than 5 and $1-p'$ times the sample size is also greater then 5. This is the same rule of thumb we used when developing the formula for the confidence interval for a population proportion.

Hypothesis Test for the Mean

Going back to the standardizing formula we can derive the **test statistic** for testing hypotheses concerning means.

$$Z_c=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

The standardizing formula can not be solved as it is because we do not have μ, the population mean. However, if we substitute in the hypothesized value of the mean, μ_0 in the formula as above, we can compute a Z value. This is the test statistic for a test of hypothesis for a mean and is presented in Figure 9.3. We interpret this Z value as the associated probability that a sample with a sample mean of $\overline{X}$ could have come from a distribution with a population mean of H_0 and we call this Z value Z_c for “calculated”. Figure 9.3 and Figure 9.4 show this process.

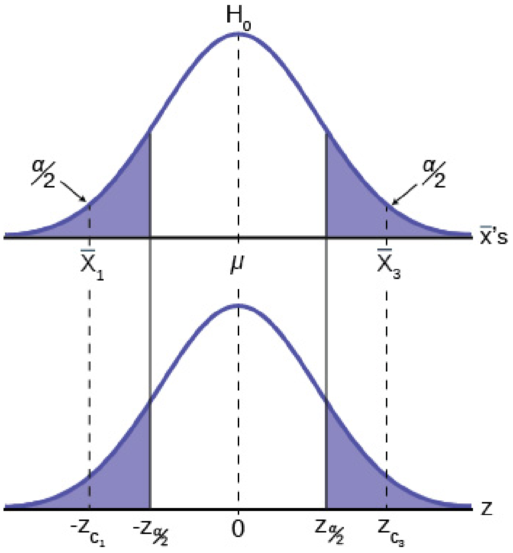


Figure 9.3

In Figure 9.3 two of the three possible outcomes are presented. $\overline{X}_1$ and $\overline{X}_3$ are in the tails of the hypothesized distribution of H_0. Notice that the horizontal axis in the top panel is labeled $\overline{X}$'s. This is the same theoretical distribution of $\overline{X}$'s, the sampling distribution, that the Central Limit Theorem tells us is normally distributed. This is why we can draw it with this shape. The horizontal axis of the bottom panel is labeled Z and is the standard normal distribution. $Z_{\frac{\alpha}{2}}$ and $-Z_{\frac{\alpha}{2}}$, called the **critical values**, are marked on the bottom panel as the Z values associated with the probability the analyst has set as the level of significance in the test, (α). The probabilities in the tails of both panels are, therefore, the same.

Notice that for each $\overline{X}$ there is an associated Z_c, called the calculated Z, that comes from solving the equation above. This calculated Z is nothing more than the number of standard deviations that the **hypothesized** mean is from the sample mean. If the sample mean falls "too many" standard deviations from the hypothesized mean we conclude that the **sample** mean could not have come from the distribution with the hypothesized mean, given our pre-set required level of significance. It **could** have come from H_0, but it is deemed just too unlikely. In Figure 9.3 both $\overline{X}_1$ and $\overline{X}_3$ are in the tails of the distribution. They are deemed "too far" from the hypothesized value of the mean given the chosen level of alpha. If in fact this sample mean it did come from H_0, but from in the tail, we have made a Type I error: we have rejected a good null. Our only real comfort is that we know the probability of making such an error, \alpha, and we can control the size of α.

Figure 9.4 shows the third possibility for the location of the sample mean, $\overline{x}$. Here the sample mean is within the two critical values. That is, within the probability of $(1-\alpha)$ and we cannot reject the null hypothesis.

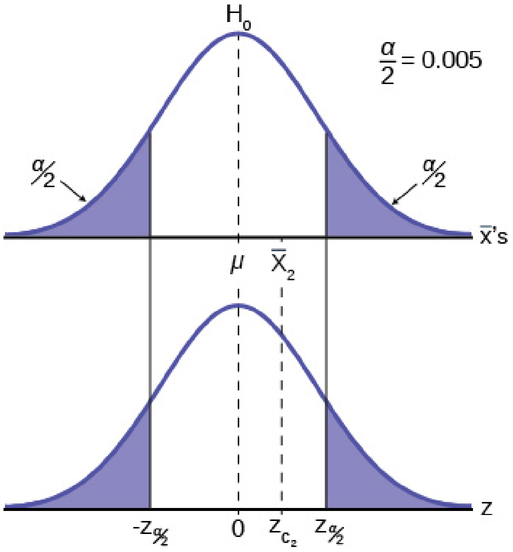


Figure 9.4

This gives us the decision rule for testing a hypothesis for a two-tailed test:

Table 9.3

Decision rule: two-tail test
If $\lvert \mathrm{Z}_c \rvert < \mathrm{Z}_{\frac{\alpha}{2}}$: then do not REJECT H_0
If $\lvert \mathrm{Z}_c \rvert > \mathrm{Z}_{\frac{\alpha}{2}}$: then REJECT H_0

This rule will always be the same no matter what hypothesis we are testing or what formulas we are using to make the test. The only change will be to change the Z_c to the appropriate symbol for the test statistic for the parameter being tested. Stating the decision rule another way: if the sample mean is unlikely to have come from the distribution with the hypothesized mean we cannot accept the null hypothesis. Here we define "unlikely" as having a probability less than alpha of occurring.

P-value Approach

An alternative decision rule can be developed by calculating the probability that a sample mean could be found that would give a test statistic larger than the test statistic found from the current sample data assuming that the null hypothesis is true. Here the notion of "likely" and "unlikely" is defined by the probability of drawing a sample with a mean from a population with the hypothesized mean that is either larger or smaller than that found in the sample data. Simply stated, the p-value approach compares the desired significance level, α, to the p-value which is the probability of drawing a sample mean further from the hypothesized value than the actual sample mean. A large p-value calculated from the data indicates that we should not reject the **null hypothesis**. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it. The relationship between the decision rule of comparing the calculated test statistics, Z_c, and the Critical Value, Z_α, and using the p-value can be seen in Figure 9.5.

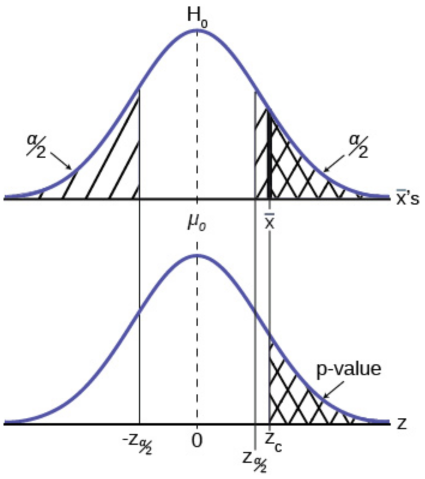


Figure 9.5

The calculated value of the test statistic is Z_c in this example and is marked on the bottom graph of the standard normal distribution because it is a Z value. In this case the calculated value is in the tail and thus we cannot accept the null hypothesis, the associated $\overline{X}$ is just too unusually large to believe that it came from the distribution with a mean of μ_0 with a significance level of \alpha.

If we use the p-value decision rule we need one more step. We need to find in the standard normal table the probability associated with the calculated test statistic, Z_c. We then compare that to the \alpha associated with our selected level of confidence. In Figure 9.5 we see that the p-value is less than \alpha and therefore we cannot accept the null. We know that the p-value is less than \alpha because the area under the p-value is smaller than $\alpha/2$. It is important to note that two researchers drawing randomly from the same population may find two different p-values from their samples. This occurs because the p-value is calculated as the probability in the tail beyond the sample mean assuming that the null hypothesis is correct. Because the sample means will in all likelihood be different this will create two different p-values. Nevertheless, the conclusions as to the null hypothesis should be different with only the level of probability of α.

Here is a systematic way to make a decision of whether you cannot accept or cannot reject a null **hypothesis** if using the **p-value** and a **preset or preconceived** \(\bf{\alpha}\) (the "**significance level**"). A preset α is the probability of a **Type I** error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem. In any case, the value of α is the decision of the analyst. When you make a decision to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, cannot accept H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, Ha, may be correct.
- If $\alpha \leq p$-value, cannot reject H_0. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, Ha, may be correct. In this case the status quo stands.
- When you "cannot reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_0. Remember that the null is the status quo and it takes high probability to overthrow the status quo. This bias in favor of the null hypothesis is what gives rise to the statement "tyranny of the status quo" when discussing hypothesis testing and the scientific method.

Both decision rules will result in the same decision and it is a matter of preference which one is used.

One and Two-tailed Tests

The discussion of Figure 9.3-Figure 9.5 was based on the null and alternative hypothesis presented in Figure 9.3. This was called a two-tailed test because the alternative hypothesis allowed that the mean could have come from a population which was either larger or smaller than the hypothesized mean in the null hypothesis. This could be seen by the statement of the alternative hypothesis as $\mu \neq 100$, in this example.

It may be that the analyst has no concern about the value being "too" high or "too" low from the hypothesized value. If this is the case, it becomes a one-tailed test and all of the alpha probability is placed in just one tail and not split into $\alpha/2$ as in the above case of a two-tailed test. Any test of a claim will be a one-tailed test. For example, a car manufacturer claims that their Model 17B provides gas mileage of greater than 25 miles per gallon. The null and alternative hypothesis would be:

- $H_0 : \mu \leq 25$

- $H_a : \mu > 25$

The claim would be in the alternative hypothesis. The burden of proof in hypothesis testing is carried in the alternative. This is because failing to reject the null, the status quo, must be accomplished with 90 or 95 percent significance that it cannot be maintained. Said another way, we want to have only a 5 or 10 percent probability of making a Type I error, rejecting a good null; overthrowing the status quo.

This is a one-tailed test and all of the alpha probability is placed in just one tail and not split into $\alpha/2$ as in the above case of a two-tailed test.

Figure 9.6 shows the two possible cases and the form of the null and alternative hypothesis that give rise to them.

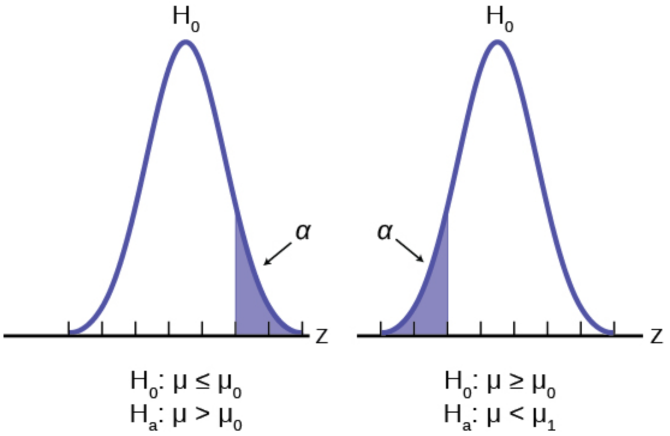


Figure 9.6

where μ_0 is the hypothesized value of the population mean.

Table 9.4 Test Statistics for Test of Means, Varying Sample Size, Population Standard Deviation Known or Unknown

Sample size	Test statistic
< 30 (σ unknown)	$t_c = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$
< 30 (σ known)	$Z_c = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$
> 30 (σ unknown)	$Z_c = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$
> 30 (σ known)	$Z_c = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$

Effects of Sample Size on Test Statistic

In developing the confidence intervals for the mean from a sample, we found that most often we would not have the population standard deviation, σ. If the sample size were less than 30, we could simply substitute the point estimate for σ, the sample standard deviation, s, and use the student's t-distribution to correct for this lack of information.

When testing hypotheses we are faced with this same problem and the solution is exactly the same. Namely: If the population standard deviation is unknown, and the sample size is less than 30, substitute s, the point estimate for the population standard deviation, σ, in the formula for the test statistic and use the student's t distribution. All the formulas and figures above are unchanged except for this substitution and changing the Z distribution to the student's t distribution on the graph. Remember that the student's t distribution can only be computed knowing the proper degrees of freedom for the problem. In this case, the degrees of freedom is computed as before with confidence intervals: $df = (n-1)$. The calculated t-value is compared to the t-value associated with the pre-set level of confidence required in the test, $t_{\alpha,df}$ found in the student's t tables. If we do not know σ, but the sample size is 30 or more, we simply substitute s for σ and use the normal distribution.

Table 9.4 summarizes these rules.

A Systematic Approach for Testing A Hypothesis

A systematic approach to hypothesis testing follows the following steps and in this order. This template will work for all hypotheses that you will ever test.

- Set up the null and alternative hypothesis. This is typically the hardest part of the process. Here the question being asked is reviewed. What parameter is being tested, a mean, a proportion, differences in means, etc. Is this a one-tailed test or two-tailed test? Remember, if someone is making a claim it will always be a one-tailed test.
- Decide the level of significance required for this particular case and determine the critical value. These can be found in the appropriate statistical table. The levels of confidence typical for businesses are 80, 90, 95, 98, and 99. However, the level of significance is a policy decision and should be based upon the risk of making a Type I error, rejecting a good null. Consider the consequences of making a Type I error.

  Next, on the basis of the hypotheses and sample size, select the appropriate test statistic and find the relevant critical value: Z_α, t_α, etc. Drawing the relevant probability distribution and marking the critical value is always big help. Be sure to match the graph with the hypothesis, especially if it is a one-tailed test.

- Take a sample(s) and calculate the relevant parameters: sample mean, standard deviation, or proportion. Using the formula for the test statistic from above in step 2, now calculate the test statistic for this particular case using the parameters you have just calculated.
- Compare the calculated test statistic and the critical value. Marking these on the graph will give a good visual picture of the situation. There are now only two situations:
  1. The test statistic is in the tail: Cannot Accept the null, the probability that this sample mean (proportion) came from the hypothesized distribution is too small to believe that it is the real home of these sample data.
  2. The test statistic is not in the tail: Cannot Reject the null, the sample data are compatible with the hypothesized population parameter.
- Reach a conclusion. It is best to articulate the conclusion two different ways. First a formal statistical conclusion such as “With a 5 % level of significance we cannot accept the null hypotheses that the population mean is equal to XX (units of measurement)”. The second statement of the conclusion is less formal and states the action, or lack of action, required. If the formal conclusion was that above, then the informal one might be, “The machine is broken and we need to shut it down and call for repairs”.

All hypotheses tested will go through this same process. The only changes are the relevant formulas and those are determined by the hypothesis required to answer the original question.

7.5: Full Hypothesis Test Examples

Tests on Means

Example 7.5.8

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using a preset $\alpha = 0.05$.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

Set the null and alternative hypothesis:

In this case there is an implied challenge or claim. This is that the goggles will reduce the swimming time. The effect of this is to set the hypothesis as a one-tailed test. The claim will always be in the alternative hypothesis because the burden of proof always lies with the alternative. Remember that the status quo must be defeated with a high degree of confidence, in this case 95 % confidence. The null and alternative hypotheses are thus:

$H_0 : \mu \geq 16.43 \qquad H_a : \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\overline{X}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test statistic:

The sample size is less than 30 and we do not know the population standard deviation so this is a t-test. and the proper formula is: $t_c = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$

$\mu_0 = 16.43$ comes from H_0 and not the data. $\overline{X} = 16$. $s = 0.8$, and $n = 15$.

Our step 2, setting the level of significance, has already been determined by the problem, .05 for a 95 % significance level. It is worth thinking about the meaning of this choice. The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.) For this case the only concern with a Type I error would seem to be that Jeffery's dad may fail to bet on his son's victory because he does not have appropriate confidence in the effect of the goggles.

To find the critical value we need to select the appropriate test statistic. We have concluded that this is a t-test on the basis of the sample size and that we are interested in a population mean. We can now draw the graph of the t-distribution and mark the critical value. For this problem the degrees of freedom are n-1, or 14. Looking up 14 degrees of freedom at the 0.05 column of the t-table we find 1.761. This is the critical value and we can put this on our graph.

Step 3 is the calculation of the test statistic using the formula we have selected. We find that the calculated test statistic is 2.08, meaning that the sample mean is 2.08 standard deviations away from the hypothesized mean of 16.43.

$$t_c = \frac{\overline{x} - \mu_0}{s / \sqrt{n}} = \frac{16 - 16.43}{.8 / \sqrt{15}} = -2.08$$

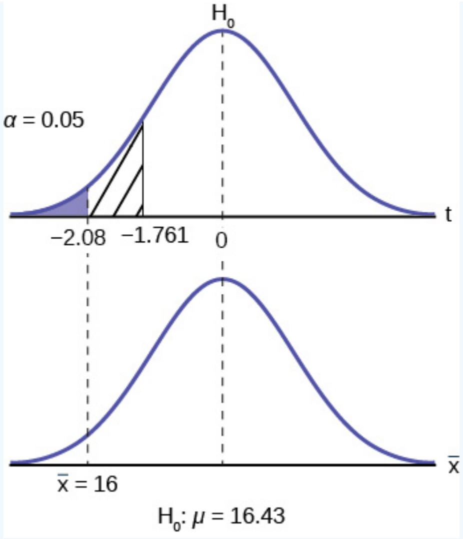


Figure 7.5.7

Step 4 has us compare the test statistic and the critical value and mark these on the graph. We see that the test statistic is in the tail and thus we move to step 4 and reach a conclusion. The probability that an average time of 16 minutes could come from a distribution with a population mean of 16.43 minutes is too unlikely for us to accept the null hypothesis. We cannot accept the null.

Step 5 has us state our conclusions first formally and then less formally. A formal conclusion would be stated as: "With a 95% level of significance we cannot accept the null hypothesis that the swimming time with goggles comes from a distribution with a population mean time of 16.43 minutes." Less formally, "With 95% significance we believe that the goggles improves swimming speed"

If we wished to use the p-value system of reaching a conclusion we would calculate the statistic and take the additional step to find the probability of being 2.08 standard deviations from the mean on a t-distribution. This value is .0187. Comparing this to the \alpha-level of .05 we see that we cannot accept the null. The p-value has been put on the graph as the shaded area beyond -2.08 and it shows that it is smaller than the hatched area which is the alpha level of 0.05. Both methods reach the same conclusion that we cannot accept the null hypothesis.

Exercise 7.5.8

The mean throwing distance of a football for Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Example 7.5.9

Jane has just begun her new job as on the sales force of a very competitive company. In a sample of 16 sales calls it was found that she closed the contract for an average value of 108 dollars with a standard deviation of 12 dollars. Test at 5% significance that the population mean is at least 100 dollars against the alternative that it is less than 100 dollars. Company policy requires that new members of the sales force must exceed an average of $100 per contract during the trial employment period. Can we conclude that Jane has met this requirement at the significance level of 95%?

Answer

1. $H_0 : \mu \leq 100$
   $H_a : \mu > 100$
   The null and alternative hypothesis are for the parameter μ because the number of dollars of the contracts is a continuous random variable. Also, this is a one-tailed test because the company has only an interested if the number of dollars per contact is below a particular number not "too high" a number. This can be thought of as making a claim that the requirement is being met and thus the claim is in the alternative hypothesis.
2. Test statistic: $t_c = \frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{108-100}{\left(\frac{12}{\sqrt{16}}\right)} = 2.67$
3. Critical value: $t_a = 1.753$ with $n-1$ degrees of freedom = 15

The test statistic is a Student's t because the sample size is below 30; therefore, we cannot use the normal distribution. Comparing the calculated value of the test statistic and the critical value of tt (ta)(ta) at a 5% significance level, we see that the calculated value is in the tail of the distribution. Thus, we conclude that 108 dollars per contract is significantly larger than the hypothesized value of 100 and thus we cannot accept the null hypothesis. There is evidence that supports Jane's performance meets company standards.

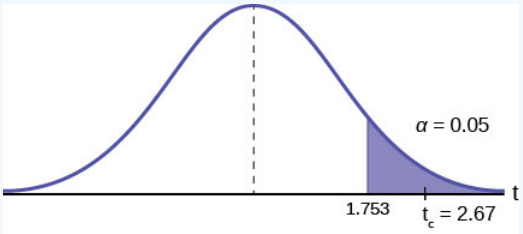


Figure 7.5.8

? Exercise 7.5.9

It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, state your conclusion, and identify the Type I errors.

? Example 7.5.10

A manufacturer of salad dressings uses machines to dispense liquid ingredients into bottles that move along a filling line. The machine that dispenses salad dressings is working properly when 8 ounces are dispensed. Suppose that the average amount dispensed in a particular sample of 35 bottles is 7.91 ounces with a variance of 0.03 ounces squared,s^2. Is there evidence that the machine should be stopped and production wait for repairs? The lost production from a shutdown is potentially so great that management feels that the level of significance in the analysis should be 99%.

Again we will follow the steps in our analysis of this problem.

Answer

STEP 1: Set the Null and Alternative Hypothesis. The random variable is the quantity of fluid placed in the bottles. This is a continuous random variable and the parameter we are interested in is the mean. Our hypothesis therefore is about the mean. In this case we are concerned that the machine is not filling properly. From what we are told it does not matter if the machine is over-filling or under-filling, both seem to be an equally bad error. This tells us that this is a two-tailed test: if the machine is malfunctioning it will be shutdown regardless if it is from over-filling or under-filling. The null and alternative hypotheses are thus:

$$H_0 : \mu = 8$$

$$Ha : \mu \neq 8$$

STEP 2: Decide the level of significance and draw the graph showing the critical value.

This problem has already set the level of significance at 99%. The decision seems an appropriate one and shows the thought process when setting the significance level. Management wants to be very certain, as certain as probability will allow, that they are not shutting down a machine that is not in need of repair. To draw the distribution and the critical value, we need to know which distribution to use. Because this is a continuous random variable and we are interested in the mean, and the sample size is greater than 30, the appropriate distribution is the normal distribution and the relevant critical value is 2.575 from the normal table or the t-table at 0.005 column and infinite degrees of freedom. We draw the graph and mark these points.

Figure 7.5.9

STEP 3: Calculate sample parameters and the test statistic. The sample parameters are provided, the sample mean is 7.91 and the sample variance is .03 and the sample size is 35. We need to note that the sample variance was provided not the sample standard deviation, which is what we need for the formula. Remembering that the standard deviation is simply the square root of the variance, we therefore know the sample standard deviation, s, is 0.173. With this information we calculate the test statistic as -3.07, and mark it on the graph.

$$Z_c = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{7.91 - 8}{\cdot 173/\sqrt{35}} = -3.07$$

STEP 4: Compare test statistic and the critical values Now we compare the test statistic and the critical value by placing the test statistic on the graph. We see that the test statistic is in the tail, decidedly greater than the critical value of 2.575. We note that even the very small difference between the hypothesized value and the sample value is still a large number of standard deviations. The sample mean is only 0.08 ounces different from the required level of 8 ounces, but it is 3 plus standard deviations away and thus we cannot accept the null hypothesis.

STEP 5: Reach a Conclusion

Three standard deviations of a test statistic will guarantee that the test will fail. The probability that anything is within three standard deviations is almost zero. Actually it is 0.0026 on the normal distribution, which is certainly almost zero in a practical sense. Our formal conclusion would be “ At a 99% level of significance we cannot accept the hypothesis that the sample mean came from a distribution with a mean of 8 ounces” Or less formally, and getting to the point, “At a 99% level of significance we conclude that the machine is under filling the bottles and is in need of repair”.

Hypothesis Test for Proportions

Just as there were confidence intervals for proportions, or more formally, the population parameter p of the binomial distribution, there is the ability to test hypotheses concerning p.

The population parameter for the binomial is p. The estimated value (point estimate) for p is p' where $p' = x/n$, x is the number of successes in the sample and n is the sample size.

When you perform a hypothesis test of a population proportion p, you take a simple random sample from the population. The conditions for a **binomial distribution** must be met, which are: there are a certain number n of independent trials meaning random sampling, the outcomes of any trial are binary, success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np' and nq' must both be greater than five ($np' > 5$ and $nq' > 5$). In this case the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. There is no distribution that can correct for this small sample bias and thus if these conditions are not met we simply cannot test the hypothesis with the data available at that time. We met this condition when we first were estimating confidence intervals for p.

Again, we begin with the standardizing formula modified because this is the distribution of a binomial.

$$Z = \frac{\mathrm{p}' - p}{\sqrt{\frac{\mathrm{pq}}{n}}}$$

Substituting p_0, the hypothesized value of p, we have:

$$Z_c = \frac{p' - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

This is the test statistic for testing hypothesized values of p, where the null and alternative hypotheses take one of the following forms:

Table 7.5.5

Two-tailed test	One-tailed test	One-tailed test
$H_0 : p = p_0$	$H_0 : p \leq p_0$	$H_0 : p \geq p_0$
$H_a : p \neq p_0$	$H_a : p > p_0$	$H_a : p < p_0$

The decision rule stated above applies here also: if the calculated value of Z_c shows that the sample proportion is "too many" standard deviations from the hypothesized proportion, the null hypothesis cannot be accepted. The decision as to what is "too many" is pre-determined by the analyst depending on the level of significance required in the test.

Example 7.5.11

The mortgage department of a large bank is interested in the nature of loans of first-time borrowers. This information will be used to tailor their marketing strategy. They believe that 50% of first-time borrowers take out smaller loans than other borrowers. They perform a hypothesis test to determine if the percentage is **the same or different from 50%**. They sample **100 first-time borrowers** and find **53** of these loans are smaller that the other borrowers. For the hypothesis test, they choose a 5% level of significance.

Answer

STEP 1: Set the null and alternative hypothesis.

$H_0 : p = 0.50 \qquad H_a : p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test. The Type I and Type II errors are as follows: The Type I error is to conclude that the proportion of borrowers is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true). The Type II error is there is not enough evidence to conclude that the proportion of first time borrowers differs from 50% when, in fact, the proportion does differ from 50%. (You fail to reject the null hypothesis when the null hypothesis is false.)

STEP 2: Decide the level of significance and draw the graph showing the critical value

The level of significance has been set by the problem at the 95% level. Because this is two-tailed test one-half of the alpha value will be in the upper tail and one-half in the lower tail as shown on the graph. The critical value for the normal distribution at the 95% level of confidence is 1.96. This can easily be found on the student’s t-table at the very bottom at infinite degrees of freedom remembering that at infinity the t-distribution is the normal distribution. Of course the value can also be found on the normal table but you have go looking for one-half of 95 (0.475) inside the body of the table and then read out to the sides and top for the number of standard deviations.

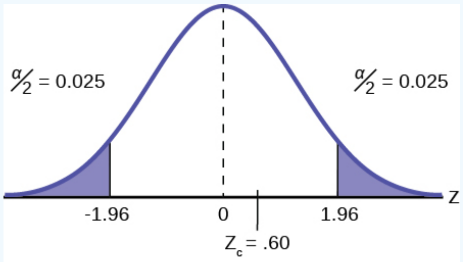


Figure 7.5.10

STEP 3: Calculate the sample parameters and critical value of the test statistic.

The test statistic is a normal distribution, Z, for testing proportions and is:

$$Z=\frac{p'-p_0}{\sqrt{\frac{p_0q_0}{n}}}=\frac{.53-.50}{\sqrt{\frac{.5(.5)}{100}}}=0.60$$

For this case, the sample of 100 found 53 first-time borrowers were different from other borrowers. The sample proportion, $p'=53/100=0.53$The test question, therefore, is : "Is 0.53 significantly different from .50?" Putting these values into the formula for the test statistic we find that 0.53 is only 0.60 standard deviations away from .50. This is barely off of the mean of the standard normal distribution of zero. There is virtually no difference from the sample proportion and the hypothesized proportion in terms of standard deviations.

STEP 4: Compare the test statistic and the critical value.

The calculated value is well within the critical values of ± 1.96 standard deviations and thus we cannot reject the null hypothesis. To reject the null hypothesis we need significant evident of difference between the hypothesized value and the sample value. In this case the sample value is very nearly the same as the hypothesized value measured in terms of standard deviations.

STEP 5: Reach a conclusion

The formal conclusion would be "At a 95% level of significance we cannot reject the null hypothesis that 50% of first-time borrowers have the same size loans as other borrowers". Less formally we would say that "There is no evidence that one-half of first-time borrowers are significantly different in loan size from other borrowers". Notice the length to which the conclusion goes to include all of the conditions that are attached to the conclusion. Statisticians for all the criticism they receive, are careful to be very specific even when this seems trivial. Statisticians cannot say more than they know and the data constrain the conclusion to be within the metes and bounds of the data.

? Exercise 7.5.11

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

? Example 7.5.12

Suppose a consumer group suspects that the proportion of households that have three or more cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three or more cell phones.

Answer

Here is an abbreviate version of the system to solve hypothesis tests applied to a test on a proportions.

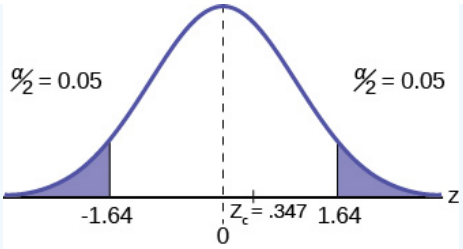
$$H_0: p=0.3$$

$$H_a: p\neq 0.3$$

$$n=150$$

$$\mathrm{p}'=\frac{x}{n}=\frac{43}{150}=0.287$$

$$Z_c=\frac{\mathrm{p}'-p_0}{\sqrt{\frac{p_0q_0}{n}}}=\frac{0.287-0.3}{\sqrt{\frac{3(7)}{150}}}=0.347$$



At a significance level of 90%
we cannot reject H_0:
the consumer group is correct.

Figure 7.5.11

Example 7.5.13

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95
Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05.

Answer

Let’s follow a four-step process to answer this statistical question.

State the Question: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be

a. $H_0 : \mu \leq 1$
b. $H_a : \mu > 1$

Plan: We are testing a sample mean without a known population standard deviation with less than 30 observations. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal. **Do the calculations and draw the graph**. **State the Conclusions**: We cannot accept the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example 7.5.14

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Answer

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
   a. $H_0 : p \leq 0.00034$
   b. $H_a : p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.

2. We will be testing a sample proportion with $x = 172$ and $n = 420,019$. The sample is sufficiently large because we have $np' = 420,019(0.00034) = 142.8$, $nq' = 420,019(0.99966) = 419,876.2$, two independent outcomes, and a fixed probability of success $p' = 0.00034$. Thus we will be able to generalize our results to the population.

7.6: Introduction to Two-Sample Tests

Figure 7.6.1 If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores. Many business applications require comparing two groups. It may be the investment returns of two different investment strategies, or the differences in production efficiency of different management styles.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**. **Independent groups** consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions of each group.

7.7: Comparing Two Independent Population Means

The comparison of two independent population means is very common and provides a way to test the hypothesis that the two groups differ from each other. Is the night shift less productive than the day shift, are the rates of return from fixed asset investments different from those from common stock investments, and so on? An observed difference between two sample means depends on both the means and the sample standard deviations. Very different means can occur by chance if there is great variation among the individual samples. The test statistic will have to account for this fact. 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 we will see later was developed by Aspin-Welch.

When we developed the hypothesis test for the mean and proportions we began with the Central Limit Theorem. We recognized that a sample mean came from a distribution of sample means, and sample proportions came from the sampling distribution of sample proportions. This made our sample parameters, the sample means and sample proportions, into random variables. It was important for us to know the distribution that these random variables came from. The Central Limit Theorem gave us the answer: the normal distribution. Our Z and t statistics came from this theorem. This provided us with the solution to our question of how to measure the probability that a sample mean came from a distribution with a particular hypothesized value of the mean or proportion. In both cases that was the question: what is the probability that the mean (or proportion) from our sample data came from a population distribution with the hypothesized value we are interested in?

Now we are interested in whether or not two samples have the same mean. Our question has not changed: Do these two samples come from the same population distribution? To approach this problem we create a new random variable. We recognize that we have two sample means, one from each set of data, and thus we have two random variables coming from two unknown distributions. To solve the problem we create a new random variable, the difference between the sample means. This new random variable also has a distribution and, again, the Central Limit Theorem tells us that this new distribution is normally distributed, regardless of the underlying distributions of the original data. A graph may help to understand this concept.

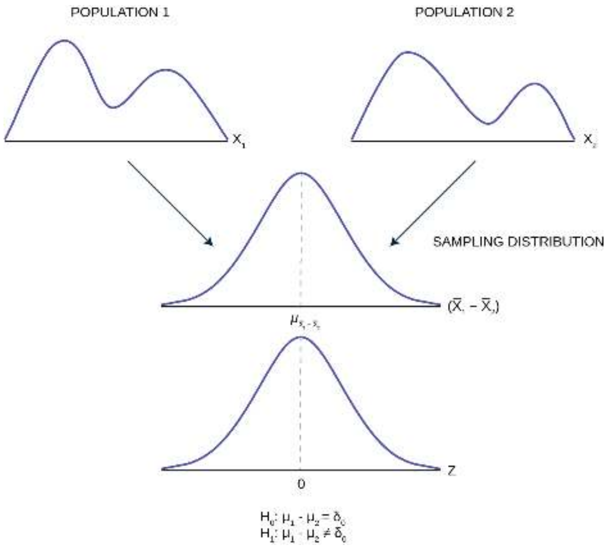

POPULATION 1
POPULATION 2
X_1
X_2
SAMPLING DISTRIBUTION
$(\bar{X}_1 - \bar{X}_2)$
$\mu_{\bar{x}_1 - \bar{x}_2}$
Z
0
$H_0: \mu_1 - \mu_2 = \delta_0$
$H_1: \mu_1 - \mu_2 \neq \delta_0$

Figure 7.7.2

Pictured are two distributions of data, X_1 and X_2, with unknown means and standard deviations. The second panel shows the sampling distribution of the newly created random variable $(\overline{X}_1 - \overline{X}_2)$. This distribution is the theoretical distribution of many many sample means from population 1 minus sample means from population 2. The Central Limit Theorem tells us that this theoretical sampling distribution of differences in sample means is normally distributed, regardless of the distribution of the actual population data shown in the top panel. Because the sampling distribution is normally distributed, we can develop a standardizing formula and calculate probabilities from the standard normal distribution in the bottom panel, the Z distribution. We have seen this same analysis before in Chapter 7 Figure 7.7.2.

The Central Limit Theorem, as before, provides us with the standard deviation of the sampling distribution, and further, that the expected value of the mean of the distribution of differences in sample means is equal to the differences in the population means. Mathematically this can be stated:

$$E\left(\mu_{\bar{x}_1}-\mu_{\bar{x}_2}\right)=\mu_1-\mu_2$$

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**, $\overline{X}_1-\overline{X}_2$.

The standard error is:

$$\sqrt{\frac{\left(s_1\right)^2}{n_1}+\frac{\left(s_2\right)^2}{n_2}}$$

We remember that substituting the sample variance for the population variance when we did not have the population variance was the technique we used when building the confidence interval and the test statistic for the test of hypothesis for a single mean back in Confidence Intervals and Hypothesis Testing with One Sample. **The test statistic (t-score) is calculated as follows:**

$$t_c=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\delta_0}{\sqrt{\frac{\left(s_1\right)^2}{n_1}+\frac{\left(s_2\right)^2}{n_2}}}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively and
- 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 unknown population means.

The number of **degrees of freedom (df)** requires a somewhat complicated calculation. The df are not always a whole number. The test statistic above is approximated by the Student's t-distribution with df as follows:

Degrees of freedom

$$df=\frac{\left(\frac{\left(s_1\right)^2}{n_1}+\frac{\left(s_2\right)^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{\left(s_1\right)^2}{n_1}\right)^2+\left(\frac{1}{n_2-1}\right)\left(\frac{\left(s_2\right)^2}{n_2}\right)^2}$$

When both sample sizes n_1 and n_2 are 30 or larger, the Student's t approximation is very good. If each sample has more than 30 observations then the degrees of freedom can be calculated as n_1+n_2-2 .

The format of the sampling distribution, differences in sample means, specifies that the format of the null and alternative hypothesis is:

$$H_0: \mu_1-\mu_2=\delta_0$$

$$H_{\mathrm{a}}: \mu_1-\mu_2 \neq \delta_0$$

where δ_0 is the hypothesized difference between the two means. If the question is simply “is there any difference between the means?” then $\delta_0=0$ and the null and alternative hypotheses becomes:

$$H_0: \mu_1=\mu_2$$

$$H_{\mathrm{a}}: \mu_1 \neq \mu_2$$

An example of when δ_0 might not be zero is when the comparison of the two groups requires a specific difference for the decision to be meaningful. Imagine that you are making a capital investment. You are considering changing from your current model machine to another. You measure the productivity of your machines by the speed they produce the product. It may be that a contender to replace the old model is faster in terms of product throughput, but is also more expensive. The second machine may also have more maintenance costs, setup costs, etc. The null hypothesis would be set up so that the new machine would have to be better than the old one by enough to cover these extra costs in terms of speed and cost of production. This form of the null and alternative hypothesis shows how valuable this particular hypothesis test can be. For most of our work we will be testing simple hypotheses asking if there is any difference between the two distribution means.

Example 7.7.1 INDEPENDENT GROUPS

The Kona Iki Corporation produces coconut milk. They take coconuts and extract the milk inside by drilling a hole and pouring the milk into a vat for processing. They have both a day shift (called the B shift) and a night shift (called the G shift) to do this part of the process. They would like to know if the day shift and the night shift are equally efficient in processing the coconuts. A study is done sampling 9 shifts of the G shift and 16 shifts of the B shift. The results of the number of hours required to process 100 pounds of coconuts is presented in Table $7.7.1$. A study is done and data are collected, resulting in the data in Table $7.7.1$.

Table 7.7.1

	Sample Size	Average Number of Hours to Process 100 Pounds of Coconuts	Sample Standard Deviation
G Shift	9	2	0.8660.866
B Shift	16	3.2	1.00

Is there a difference in the mean amount of time for each shift to process 100 pounds of coconuts? Test at the 5% level of significance.

Answer

Solution 10.1

The population standard deviations are not known and cannot be assumed to equal each other. Let g be the subscript for the G Shift and b be the subscript for the B Shift. Then, μ_g is the population mean for G Shift and μ_b is the population mean for B Shift. This is a test of two **independent groups**, two population **means**.

Random variable: $\overline{X}_g - \overline{X}_b$ = difference in the sample mean amount of time between the G Shift and the B Shift takes to process the coconuts.

\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, is either faster or slower. 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 above. Using a calculator, df is approximately 18.8462.

Graph:

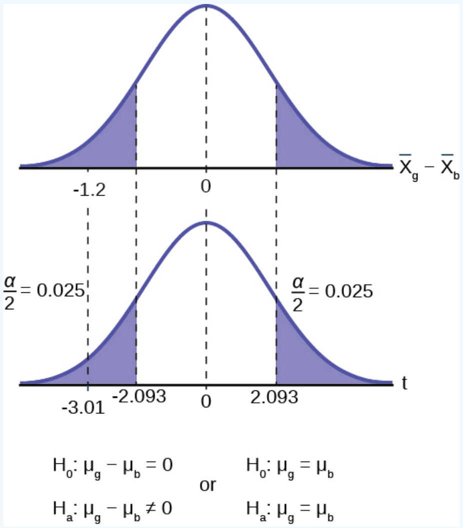


Figure 7.7.3

$$t_c = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = -3.01$$

We next find the critical value on the t-table using the degrees of freedom from above. The critical value, 2.093, is found in the .025 column, this is $\alpha/2$, at 19 degrees of freedom. (The convention is to round up the degrees of freedom to make the conclusion more conservative.) Next we calculate the test statistic and mark this on the t-distribution graph.

Make a decision: Since the calculated t-value is in the tail we cannot accept the null hypothesis that there is no difference between the two groups. The means are different.

The graph has included the sampling distribution of the differences in the sample means to show how the t-distribution aligns with the sampling distribution data. We see in the top panel that the calculated difference in the two means is -1.2 and the bottom panel shows that this is 3.01 standard deviations from the mean. Typically we do not need to show the sampling distribution graph and can rely on the graph of the test statistic, the t-distribution in this case, to reach our conclusion.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that the G Shift takes to process 100 pounds of coconuts is different from the B Shift (mean number of hours for the B Shift is greater than the mean number of hours for the G Shift).

NOTE

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 7.7.2

A study is done to determine if Company A retains its workers longer than Company B. It is believed that Company A has a higher retention than Company B. The study finds that in a sample of 11 workers at Company A their average time with the company is four years with a standard deviation of 1.5 years. A sample of 9 workers at Company B finds that the average time with the company was 3.5 years with a standard deviation of 1 year. Test this proposition at the 1% level of significance.

a. Is this a test of two means or two proportions?

Answer

Solution 10.2

a. two means because time is a continuous random variable.

b. Are the populations standard deviations known or unknown?

Answer

Solution 10.2

b. unknown

c. Which distribution do you use to perform the test?

Answer

Solution 10.2

c. Student's t

d. What is the random variable?

Answer

Solution 10.2

d. $\overline{X}_A - \overline{X}_B$

e. What are the null and alternate hypotheses?

Answer

Solution 10.2

e.

- $H_0 : \mu_A \leq \mu_B$
- $H_a : \mu_A > \mu_B$

f. Is this test right-, left-, or two-tailed?

Answer

Solution 10.2

f. right one-tailed test

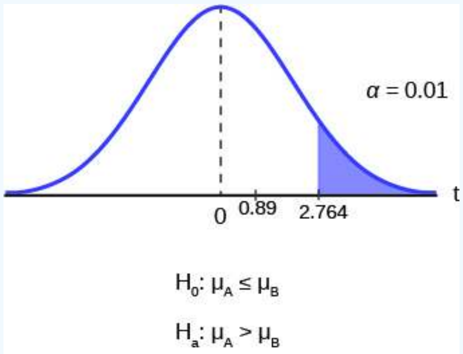


Figure 7.7.4

g. What is the value of the test statistic?

Answer

Solution 10.2

g.

$$t_c = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = 0.89$$

h. Can you accept/reject the null hypothesis?

Answer

Solution 10.2

h. Cannot reject the null hypothesis that there is no difference between the two groups. Test statistic is not in the tail. The critical value of the t distribution is 2.764 with 10 degrees of freedom. This example shows how difficult it is to reject a null hypothesis with a very small sample. The critical values require very large test statistics to reach the tail.

i. **Conclusion:**

Answer

Solution 10.2

i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the retention of workers at Company A is longer than Company B, on average.

Example 7.7.3

An interesting research question is the effect, if any, that different types of teaching formats have on the grade outcomes of students. To investigate this issue one sample of students' grades was taken from a hybrid class and another sample taken from a standard lecture format class. Both classes were for the same subject. The mean course grade in percent for the 35 hybrid students is 74 with a standard deviation of 16. The mean grades of the 40 students form the standard lecture class was 76 percent with a standard deviation of 9. Test at 5% to see if there is any significant difference in the population mean grades between standard lecture course and hybrid class.

Answer

Solution 10.3

We begin by noting that we have two groups, students from a hybrid class and students from a standard lecture format class. We also note that the random variable, what we are interested in, is students' grades, a continuous random variable. We could have asked the research question in a different way and had a binary random variable. For example, we could have studied the percentage of students with a failing grade, or with an A grade. Both of these would be binary and thus a test of proportions and not a test of means as is the case here. Finally, there is no presumption as to which format might lead to higher grades so the hypothesis is stated as a two-tailed test.

$H_0: \mu_1 = \mu_2$
$H_a: \mu_1 \neq \mu_2$

As would virtually always be the case, we do not know the population variances of the two distributions and thus our test statistic is:

$$t_c = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(74 - 76) - 0}{\sqrt{\frac{16^2}{35} + \frac{9^2}{40}}} = -0.65$$

To determine the critical value of the Student's t we need the degrees of freedom. For this case we use: $df = n_1 + n_2 - 2 = 35 + 40 - 2 = 73$. This is large enough to consider it the normal distribution thus $t_{\alpha/2} = 1.96$. Again as always we determine if the calculated value is in the tail determined by the critical value. In this case we do not even need to look up the critical value: the calculated value of the difference in these two average grades is not even one standard deviation apart. Certainly not in the tail.

Conclusion: Cannot reject the null at $\alpha = 5\%$. Therefore, evidence does not exist to prove that the grades in hybrid and standard classes differ.

7.8: Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's d is a measure of "effect size" based on the differences between two means. Cohen's d, named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Table 10.2 Cohen's Standard Effect Sizes

Size of effect	d
Small	0.2
Medium	0.5
Large	0.8

Cohen's d is the measure of the difference between two means divided by the pooled standard deviation: $d = \frac{\bar{x}_1 - \bar{x}_2}{s \text{ pooled}}$ where $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

It is important to note that Cohen's d does not provide a level of confidence as to the magnitude of the size of the effect comparable to the other tests of hypothesis we have studied. The sizes of the effects are simply indicative.

Exercise 7.8.1

Calculate Cohen's d for example 10.2.2. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Answer

$$\begin{aligned} &\bar{x}_1 = 4 s_1 = 1.5 n_1 = 11 \\ &\bar{x}_2 = 3.5 s_2 = 1 n_2 = 9 \\ &d = 0.384 \end{aligned} \tag{7.8.1}$$

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 companies is small indicating that there is not a significant difference between them.

7.9: Test for Differences in Means- Assuming Equal Population Variances

Typically we can never expect to know any of the population parameters, mean, proportion, or standard deviation. When testing hypotheses concerning differences in means we are faced with the difficulty of two unknown variances that play a critical role in the test statistic. We have been substituting the sample variances just as we did when testing hypotheses for a single mean. And as we did before, we used a Student's t to compensate for this lack of information on the population variance. There may be situations, however, when we do not know the population variances, but we can assume that the two populations have the same variance. If this is true then the pooled sample variance will be smaller than the individual sample variances. This will give more precise estimates and reduce the probability of discarding a good null. The null and alternative hypotheses remain the same, but the test statistic changes to:

$$t_c=\frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{\sqrt{S^2p\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where S_p^2 is the pooled variance given by the formula:

$$S_p^2=\frac{(n_1-1)\,s_2^1+(n_2-1)\,s_2^2}{n_1+n_2-2}$$

✓ Example 7.9.1

A drug trial is attempted using a real drug and a pill made of just sugar. 18 people are given the real drug in hopes of increasing the production of endorphins. The increase in endorphins is found to be on average 8 micrograms per person, and the sample standard deviation is 5.4 micrograms. 11 people are given the sugar pill, and their average endorphin increase is 4 micrograms with a standard deviation of 2.4. From previous research on endorphins it is determined that it can be assumed that the variances within the two samples can be assumed to be the same. Test at 5% to see if the population mean for the real drug had a significantly greater impact on the endorphins than the population mean with the sugar pill.

Solution

First we begin by designating one of the two groups Group 1 and the other Group 2. This will be needed to keep track of the null and alternative hypotheses. Let's set Group 1 as those who received the actual new medicine being tested and therefore Group 2 is those who received the sugar pill. We can now set up the null and alternative hypothesis as:

$$\begin{aligned}H_0&:\mu_1\leq\mu_2\\H_1&:\mu_1>\mu_2\end{aligned}\tag{7.9.1}$$

This is set up as a one-tailed test with the claim in the alternative hypothesis that the medicine will produce more endorphins than the sugar pill. We now calculate the test statistic which requires us to calculate the pooled variance, S_p^2 using the formula above.

$$t_c=\frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{(8-4)-0}{\sqrt{20.4933\left(\frac{1}{18}+\frac{1}{11}\right)}}=2.31\tag{7.9.2}$$

t_α, allows us to compare the test statistic and the critical value.

$$t_\alpha=1.703\text{ at }df=n_1+n_2-2=18+11-2=27\tag{7.9.3}$$

The test statistic is clearly in the tail, 2.31 is larger than the critical value of 1.703 , and therefore we cannot maintain the null hypothesis. Thus, we conclude that there is significant evidence at the 95% level of confidence that the new medicine produces the effect desired.

7.10: 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 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 even perhaps 20 times the size of the sample. This keeps each population from being over-sampled and causing biased 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 in the sampling. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the two population proportions.

Like the case of differences in sample means, we construct a sampling distribution for differences in sample proportions: $(p'_A - p'_B)$ where $p'_A = X_{\frac{A}{n_A}}$ and $p'_B = X_{\frac{B}{n_B}}$ are the sample proportions for the two sets of data in question. X_A and X_B are the number of successes in each sample group respectively, and n_A and n_B are the respective sample sizes from the two groups. Again we go the Central Figure $7.10.5$.

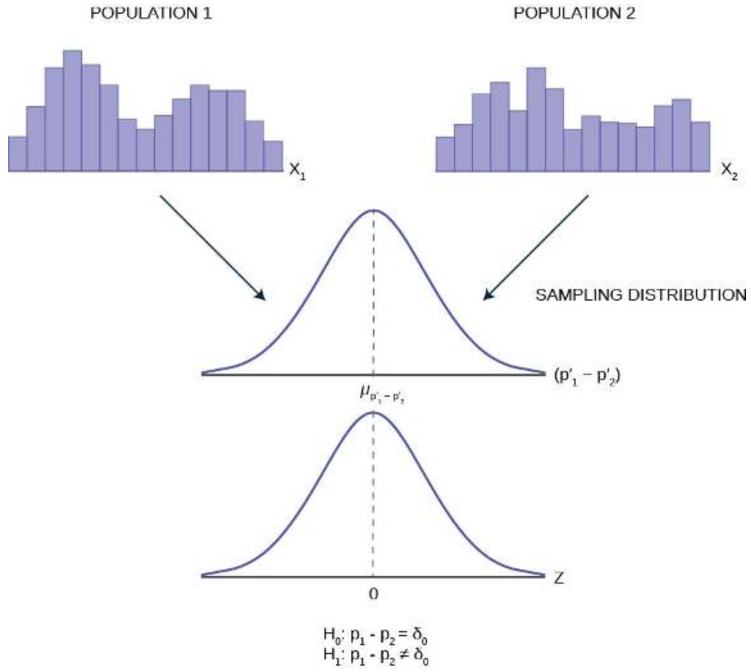


Figure 7.10.5

Generally, the null hypothesis allows for the test of a difference of a particular value, δ_0, just as we did for the case of differences in means.

$$H_0 : p_1 - p_2 = \delta_0$$

$$H_1 : p_1 - p_2 \neq \delta_0$$

Most common, however, is the test that the two proportions are the same. That is,

$$H_0 : p_{\mathrm{A}} = p_B$$

$$H_a : p_{\mathrm{A}} \neq p_B$$

To conduct the test, we use a pooled proportion, p_c.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}$$

$$\textbf{The test statistic (z-score) is:}$$

$$Z_c = \frac{\left(p'_A - p'_B\right) - \delta_0}{\sqrt{p_c\left(1 - p_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

where δ_0 is the hypothesized differences between the two proportions and p_c is the pooled variance from the formula above.

Example 7.10.1

A bank has recently acquired a new branch and thus has customers in this new territory. They are interested in the default rate in their new territory. They wish to test the hypothesis that the default rate is different from their current customer base. They sample 200 files in area A, their current customers, and find that 20 have defaulted. In area B, the new customers, another sample of 200 files shows 12 have defaulted on their loans. At a 10% level of significance can we say that the default rates are the same or different?

Answer

Solution 10.6

This is a test of proportions. We know this because the underlying random variable is binary, default or not default. Further, we know it is a test of differences in proportions because we have two sample groups, the current customer base and the newly acquired customer base. Let A and B be the subscripts for the two customer groups. Then $p_A$and p_B are the two population proportions we wish to test.

Random Variable:

$P'_A - P'_B$ = difference in the proportions of customers who defaulted in the two groups.

$H_0 : p_A = p_B$

$H_a : p_A \neq p_B$

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:

$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 0.08 \quad 1 - p_c = 0.92$

$(p'A - p'B) = 0.04$ follows an approximate normal distribution.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

The estimated difference between the two groups is : $p'_A - p'_B = 0.1 - 0.06 = 0.04$.

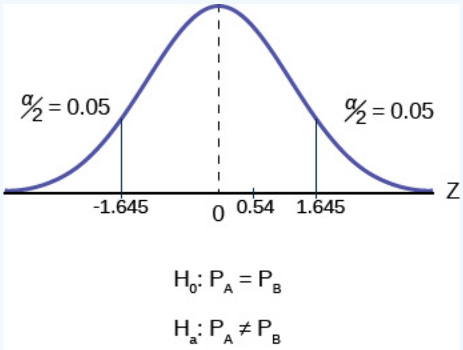


Figure 7.10.6

$$Z_c = \frac{(\mathrm{P}'_A - \mathrm{P}'_B) - \delta_0}{P_c\left(1 - P_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} = 0.54$$

The calculated test statistic is .54 and is not in the tail of the distribution.

Make a decision: Since the calculate test statistic is not in the tail of the distribution we cannot 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 between the proportions of customers who defaulted in the two groups.

? Try It 7.10.1

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.

7.11: Two Population Means with Known Standard Deviations

Even though this situation is not likely (knowing the population standard deviations is very unlikely), the following example illustrates hypothesis testing for independent means with known population standard deviations. The sampling distribution for the difference between the means is normal in accordance with the central limit theorem. The random variable is $\overline{X_1} - \overline{X_2}$. The normal distribution has the following format:

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$$

The test statistic (z-score) is:

$$Z_c = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

? Exercise 7.11.1

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. **Twenty floors** are randomly assigned **to test each wax**. Both populations have a normal distributions. The data are recorded in the table below

Wax	Sample mean number of months floor wax lasts	Population standard deviation
1	3	0.33
2	2.9	0.36

Answer

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\bar{X}_1 - \bar{X}_2$ = difference in the mean number of months the competing floor waxes last.

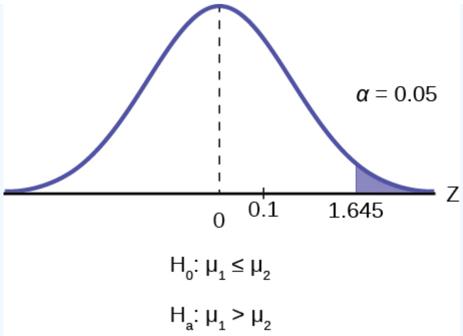
$$\begin{aligned} H_0 &: \mu_1 \leq \mu_2 \\ H_a &: \mu_1 > \mu_2 \end{aligned} \tag{7.11.1}$$

The words "is more effective" says that wax 1 lasts longer than wax 2, on average. "Longer" is a " $>$ " symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The population standard deviations are known so the distribution is normal. Using the formula for the test statistic we find the calculated value for the problem.

$$Z_c = \frac{(\mu_1 - \mu_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 0.1 \tag{7.11.2}$$

Zc=(μ1−μ2)−δ0σ21n1+σ22n2−−−−−−−√=0.1Zc=(μ1-μ2)-δ0σ12n1+σ22n2=0.1



The estimated difference between he two means is : $\bar{X}_1 - \bar{X}_2 = 3 - 2.9 = 0.1$
Compare calculated value and critical value and $\mathbf{Z}_\alpha$: We mark the calculated value on the graph and find the calculated value is not in the tail therefore we cannot reject the null hypothesis.

Make a decision: the calculated value of the test statistic is not in the tail, therefore you cannot reject H_0
Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

✓ Try It7.11.1

The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines of each type are randomly assigned to be tested. Both populations have normal distributions. in the table below shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5% level of significance.

Engine	Sample mean number of RPM	Population standard deviation
1	1,500	50
2	1,600	60

? Exercise 7.11.1

An interested citizen wanted to know if Democratic U. S. senators are older than Republican U.S. senators, on average. On May 26 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years.

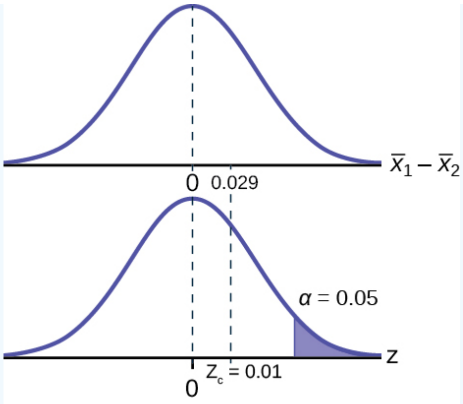
Problem
Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance.
[Show/Hide Solution]

Answer

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is $30 + 30 = 60$, which is greater than 30 , so we can use the normal approximation to the Student's-t distribution. Subscripts: 1: Democratic senators 2: Republican senators
Random variable: $\bar{X}_1 - \bar{X}_2 =$ difference in the mean age of Democratic and Republican U.S. senators.

$$\begin{gathered} H_0 : \mu_1 \leq \mu_2 H_0 : \mu_1 - \mu_2 \leq 0 \\ H_a : \mu_1 > \mu_2 H_a : \mu_1 - \mu_2 > 0 \end{gathered} \tag{7.11.3}$$

The words "older than" translates as a " $>$ " symbol and goes into H_a. Therefore, this is a right-tailed test.



Make a decision: The p-value is larger than 5%, therefore we cannot reject the null hypothesis. By calculating the test statistic we would find that the test statistic does not fall in the tail, therefore we cannot reject the null hypothesis. We reach the same conclusion using either method of a making this statistical decision.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

7.12: Matched or Paired Samples

In most cases of economic or business data we have little or no control over the process of how the data are gathered. In this sense the data are not the result of a planned controlled experiment. In some cases, however, we can develop data that are part of a controlled experiment. This situation occurs frequently in quality control situations. Imagine that the production rates of two machines built to the same design, but at different manufacturing plants, are being tested for differences in some production metric such as speed of output or meeting some production specification such as strength of the product. The test is the same in format to what we have been testing, but here we can have matched pairs for which we can test if differences exist. Each observation has its matched pair against which differences are calculated. First, the differences in the metric to be tested between the two lists of observations must be calculated, and this is typically labeled with the letter "d." Then, the average of these matched differences, $\overline{X}_d$ is calculated as is its standard deviation, S_d. We expect that the standard deviation of the differences of the matched pairs will be smaller than unmatched pairs because presumably fewer differences should exist because of the correlation between the two groups.

When using a hypothesis test for matched or paired samples, the following characteristics may 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, that is, the number of pairs not the number of observations.

The null and alternative hypotheses for this test are:

$$H_a : \mu_d \neq 0$$

the test statistic is

$$t_c = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Exercise 7.12.1

A company has developed a training program for its entering employees because they have become concerned with the results of the six-month employee review. They hope that the training program can result in better six-month reviews. Each trainee constitutes a “pair”, the entering score the employee received when first entering the firm and the score given at the six-month review. The difference in the two scores were calculated for each employee and the means for before and after the training program was calculated. The sample mean before the training program was 20.4 and the sample mean after the training program was 23.9. The standard deviation of the differences in the two scores across the 20 employees was 3.8 points. Test at the 10% significance level the null hypothesis that the two population means are equal against the alternative that the training program helps improve the employees’ scores.

Answer

The first step is to identify this as a two sample case: before the training and after the training. This differentiates this problem from simple one sample issues. Second, we determine that the two samples are "paired." Each observation in the first sample has a paired observation in the second sample. This information tells us that the null and alternative hypotheses should be:

$$H_0 : \mu_d \leq 0 \tag{7.12.1}$$

$$H_a : \mu_d > 0 \tag{7.12.2}$$

This form reflects the implied claim that the training course improves scores; the test is one-tailed and the claim is in the alternative hypothesis. Because the experiment was conducted as a matched paired sample rather than simply taking scores from people who took the training course those who didn't, we use the matched pair test statistic:

$$\text{Test Statistic: } t_c = \frac{\bar{X}_d - \mu_d}{\frac{S_d}{\sqrt{n}}} = \frac{(23.9 - 20.4) - 0}{\left(\frac{3.8}{\sqrt{20}}\right)} = 4.12 \tag{7.12.3}$$

In order to solve this equation, the individual scores, pre-training course and post-training course need to be used to calculate the individual differences. These scores are then averaged and the average difference is calculated:

$$\bar{X}_d = \bar{x}_1 - \bar{x}_2 \tag{7.12.4}$$

From these differences we can calculate the standard deviation across the individual differences:

$$S_d = \frac{\Sigma\left(d_i - \bar{X}_d\right)^2}{n-1} \text{ where } d_i = x_{1\text{i}} - x_{2\text{i}} \tag{7.12.5}$$

We can now compare the calculated value of the test statistic with the critical value. The critical value is a Student's t with degrees of freedom equal to the number of pairs, not observations, minus 1. In this case 20 pairs and at 90% confidence level $t_{a/2} = \pm 1.729$ at $df = 20 - 1 = 19$. The calculated test statistic is most certainly in the tail of the distribution and thus we cannot accept the null hypothesis that there is no difference from the training program. Evidence seems indicate that the training aids employees in gaining higher scores.

Exercise 7.12.1

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table below. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. 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

After data	Before data	Difference
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 d to denote "differences."
Random variable: $\bar{X}_d =$ the mean difference of the sensory measurements

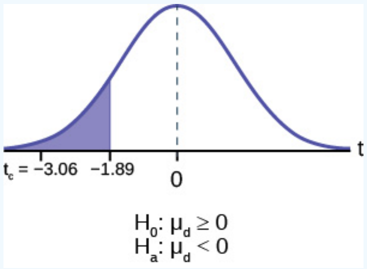
$$H_0 : \mu_d \geq 0 \tag{7.12.6}$$

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.)

$$H_a : \mu_d < 0 \tag{7.12.7}$$

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 test statistic and look up the critical value using the Student's-t distribution: The calculated value of the test statistic is 3.06 and the critical value of the t-distribution with 7 degrees of freedom at the 5% level of confidence is 1.895 with a one-tailed test.



$\bar{X}_d$ is the random variable for the differences.
The sample mean and sample standard deviation of the differences are:

$$\begin{aligned} \bar{x}_d &= -3.13 \\ \bar{s}_d &= 2.91 \end{aligned} \tag{7.12.8}$$

Compare the critical value for alpha against the calculated test statistic.

The conclusion from using the comparison of the calculated test statistic and the critical value will gives us the result. In this question the calculated test statistic is 3.06 and the critical value is 1.895. The test statistic is clearly in the tail and thus we cannot accept the null hypotheses that there is no difference between the two situations, hypnotized and not hypnotized.

Make a decision: Cannot accept the null hypothesis, H_0. This means that μ_d < 0 and there is a statistically significant 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.

✓ Example 7.12.1

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}.

$$\bar{x}_d = 21.3, s_d = 46.7 \tag{7.12.9}$$

Using the difference data, this becomes a test of a single mean.
Define the random variable: $\bar{X}_d$ mean difference in the maximum lift per player.
The distribution for the hypothesis test is a student's t with 3 degrees of freedom.

$$H_0 : \mu_d \leq 0, H_a : \mu_d > 0 \tag{7.12.10}$$

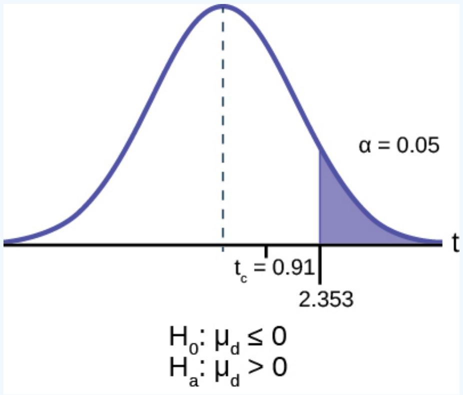


Figure **10.10**

Calculate the test statistic look up the critical value: Critical value of the test statistic is 0.91. The critical value of the student's t at 5% level of significance and 3 degrees of freedom is 2.353.

Decision: If the level of significance is 5%, we cannot reject the null hypothesis, because the calculated value of the test statistic is not in the tail.

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.

Solution

Add example text here.

7.13: Prelude to the Chi-Square Distribution

Have you ever wondered if lottery winning numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

Figure 7.13.1: The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution. In this chapter, you will learn the three major applications of the chi-square distribution:

1. the goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
2. the test of independence, which determines if events are independent, such as in the movie example
3. the test of a single variance, which tests variability, such as in the coffee example

---

7.14: Facts About the Chi-Square Distribution

The notation for the **chi-square distribution** is:

$$\chi \sim \chi^2_{df}$$

where df = degrees of freedom which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2(df)}$.

The random variable is shown as χ^2.

The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_k)^2$$

1. The curve is non-symmetrical and skewed to the right.
2. There is a different chi-square curve for each df (7.14.1).
3. The test statistic for any test is always greater than or equal to zero.
4. When $df > 90$, the chi-square curve approximates the normal distribution. For $\chi \sim \chi^2_{1,000}$ the mean, $\mu = df = 1,000$ and the standard deviation, $\sigma = \sqrt{2(1,000)} = 44.7$. Therefore, $\chi \sim N(1,000, 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.

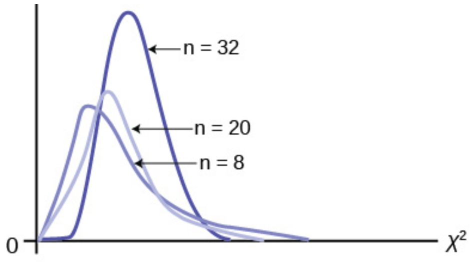


Figure 7.14.1

7.15: 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:

$$\sum_k \frac{(O-E)^2}{E}$$

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 = (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.

NOTE

The number of expected values inside each cell needs to be at least five in order to use this test.

Exercise $7.15.1$

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 Table $7.15.1$

Table 7.15.1

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 Table $7.15.2$ displays the results of that survey.

Table 7.15.2

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20

Number of absences per term	Actual number of students
9–11	1
12+	4

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

H_a: Student absenteeism **fits** faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

H_a: Student absenteeism **does not fit** faculty perception.

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Answer

Solution 11.4

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 Table 7.15.3and Table 7.15.4

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

Table 11.3

Table 7.15.4

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

b. What is the number of degrees of freedom (df)?

Answer

Solution 11.4

b. There are four "cells" or categories in each of the new tables.

$$df = \text{number of cells } - 1 = 4 - 1 = 3$$

Example 7.15.1

A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in Table 7.15.5.

Table 7.15.5

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. Table $7.15.6$ shows the results of the survey.

Table 7.15.6

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.

Exercise $7.15.2$

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 Table $7.15.7$. 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.

Table 7.15.7 Day of the Week Employees were Most Absent

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of absences	15	12	9	9	15

Answer

Solution 11.5

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 \chi2 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.

The calculated test statistics is 3 and the critical value of the χ^2 distribution at 4 degrees of freedom the 0.05 level of confidence is 9.48. This value is found in the χ^2 table at the 0.05 column on the degrees of freedom row 4.

The degrees of freedom are the number of cells $- 1 = 5 - 1 = 4$

Next, complete a graph like the following one with the proper labeling and shading. (You should shade the right tail.)

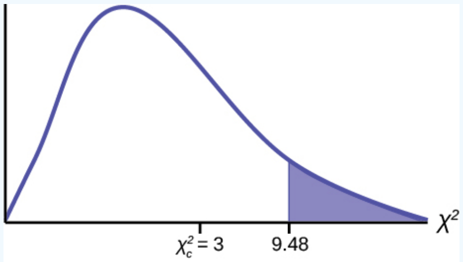


Figure 7.15.5

$$\chi^2_c = \sum_k \frac{(\mathbf{O} - \mathbf{E})^2}{\mathbf{E}} = \mathbf{3}$$

The decision is not to reject the null hypothesis because the calculated value of the test statistic is not in the tail of the distribution.

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.

? Example 7.15.2

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 56 students were asked on which night of the week they did the most homework. The results were distributed as in Table 7.15.8

Table 7.15.8

	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?

? Exercise 7.15.3

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 Table 7.15.9

Table 7.15.9

Number of Televisions	Percent
0	10
1	16
2	55

Number of Televisions	Percent
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 Table $7.15.10$

Table 7.15.10

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (O) frequency values.

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

Solution 11.6

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 Table $7.15.11$

Table 7.15.11

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.

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 $df = (\text{ the number of cells }) - 1 = 5 - 1 = 4$.

Calculate the test statistic: $\chi^2 = 29.65$

Graph:

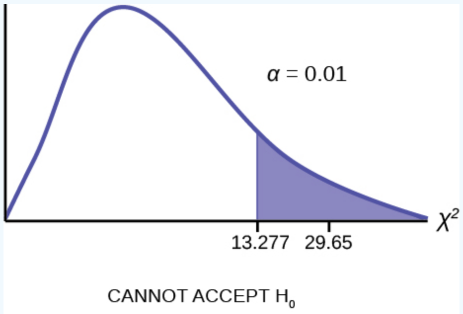


Figure 7.15.6

The graph of the Chi-square shows the distribution and marks the critical value with four degrees of freedom at 99% level of confidence, α = .01, 13.277. The graph also marks the calculated chi squared test statistic of 29.65. Comparing the test statistic with the critical value, as we have done with all other hypothesis tests, we reach the conclusion.

Make a decision: Because the test statistic is in the tail of the distribution we cannot accept the null hypothesis.

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.

? Example 7.15.3

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 Table 7.15.12

Table 7.15.12

Number of pets	Percent
0	18
1	25
2	30
3	18
4+	9

A random sample of 1,000 students from the Eastern United States resulted in the data in Table 7.15.13

Table 7.15.13

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?

Exercise 7.15.4

Suppose you flip two coins 100 times. The results are $20HH, 27HT, 30TH,$ and $23TT$. 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 $HH, 25HT, 25TH,$ and $25TT$. This is the expected distribution from the binomial probability 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.

$\mathbf{H_0}$: The coins are fair.

$\mathbf{H_a}$: 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:

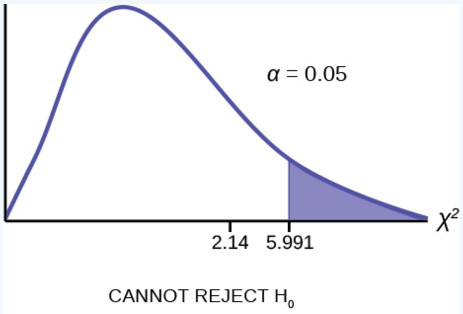


Figure 7.15.7

The graph of the Chi-square shows the distribution and marks the critical value with two degrees of freedom at 95% level of confidence, $\alpha = 0.05$, 5.991. The graph also marks the calculated χ^2 test statistic of 2.14. Comparing the test statistic with the critical value, as we have done with all other hypothesis tests, we reach the conclusion.

Conclusion: There is insufficient evidence to conclude that the coins are not fair: we cannot reject the null hypothesis that the coins are fair.

7.16: 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:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$$

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 $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Table 3.1 earlier. As a review, consider the following example.

Note

The expected value inside each cell needs to be at least five in order for you to use this test.

Example 7.16.1

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then $P(A \cap B) = P(A)P(B)$. $A \cap B$ 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 $P(A \cap B) = P(A)P(B)$. 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; in other words, they have no effect on each other.

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):

$$E=\frac{(\text{ row total })(\text{ column total })}{\text{total number surveyed}}$$

? Try It 7.16.1

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven of the 300 surveyed 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?

? Exercise 7.16.1

A volunteer group, provides from one to nine hours each week with disabled senior citizens. The program recruits among community college students, four-year college students, and nonstudents. In Table 11.14 is a **sample** of the adult volunteers and the number of hours they volunteer per week.

The table contains **observed (O)** values (data).
Table 11.14 Number of Hours Worked Per Week by Volunteer Type (Observed)

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

Solution 11.9

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 result are in Table 11.15.

The table contains **expected** (E) values (data).
Table 11.15 Number of Hours Worked Per Week by Volunteer Type (Expected)

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:

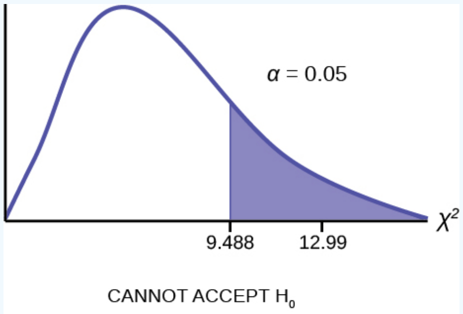


Figure 11.8

The graph of the Chi-square shows the distribution and marks the critical value with four degrees of freedom at 95% level of confidence, $\alpha = 0.05$, 9.488. The graph also marks the calculated χ^2_c test statistic of 12.99. Comparing the test statistic with the critical value, as we have done with all other hypothesis tests, we reach the conclusion.

Make a decision: Because the calculated test statistic is in the tail we cannot accept 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 11.15, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

? Try It 7.16.2

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table 11.16 shows the results:

Table 11.16

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.

? Exercise 7.16.2

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 11.17 shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Table 11.17 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?

Answer

Solution 11.10

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.

$$E=\frac{(\text{ row total })(\text{ column total })}{\text{total surveyed}}=\frac{155\cdot 57}{400}=22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

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?

Answer

Solution 11.10

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. $E=\frac{(\text{ row total })(\text{ column total })}{\text{total surveyed}}=$ ________

Answer

Solution 11.10

c. $E = \frac{(\text{ row total })(\text{ column total })}{\text{total surveyed}} = 8.19$

d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about ________.

Answer

Solution 11.10

d. 8

7.17: Chapter Key Terms

Key Terms	**Definition**
Binomial Distribution	a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n,p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X=x) = \left(\begin{array}{l} n \\ x \end{array}\right) p^x q^{n-x}$.
Central Limit Theorem	Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$. If the size n of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distribution regardless of the shape of the population. The expected value of the mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.
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.
Critical Value	The t or Z value set by the researcher that measures the probability of a Type I error, σ.
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).
Hypothesis Testing	Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.
Normal Distribution	a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.
Standard Deviation	a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Key Terms	Definition
Student's t-Distribution	investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are: • It is continuous and assumes any real values. • The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution. • It approaches the standard normal distribution as n gets larger. • There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.
Test Statistic	The formula that counts the number of standard deviations on the relevant distribution that estimated parameter is away from the hypothesized value.
Type I Error	The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.
Type II Error	The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

7.17.1: Chapter Key Terms

Key Terms	Definitions
Cohen’s d	a measure of effect size based on the differences between two means. If d is between 0 and 0.2 then the effect is small. If d approaches is 0.5, then the effect is medium, and if d approaches 0.8, then it is a large effect.
Independent Groups	two samples that are selected from two populations, and the values from one population are not related in any way to the values from the other population.
Matched Pairs	two samples that are dependent. Differences between a before and after scenario are tested by testing one population mean of differences.
Pooled Variance	a weighted average of two variances that can then be used when calculating standard error.

7.17.2: Chapter Key Terms

Key Term	Definition
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.
Goodness-of-Fit	a hypothesis test that compares expected and observed values in order to look for significant differences within one non-parametric variable. The degrees of freedom used equals the (number of categories – 1).
Test for Homogeneity	a test used to draw a conclusion about whether two populations have the same distribution. The degrees of freedom used equals the (number of columns – 1).
Test of Independence	a test used to draw a conclusion about whether two populations have the same distribution. The degrees of freedom used equals the (number of columns – 1).

7.18: Chapter References

9.2 Null and Alternative Hypotheses

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

9.5 Full Hypothesis Test Examples

Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.

Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.

Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).

Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).

Data from *Growing by Degrees* by Allen and Seaman.

Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.

Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).

Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).

Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).

Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.

Data from Toastmasters International. Available online at toastmasters.org/artisan/deta...eID=429&Page=1.

Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).

Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).

“Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.

Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).

Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at http://www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

7.18.1: Chapter References

10.2 Comparing Two Independent Population Means

Data from Graduating Engineer + Computer Careers. Available online at www.graduatingengineer.com

Data from *Microsoft Bookshelf*.

Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).

"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).

"Sectoring by Industry Groups." Nasdaq. Available online at www.nasdaq.com/markets/barcha...&base=industry (accessed June 17, 2013).

"Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).

"World Series History." Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

10.5 Comparing Two Independent Population Proportions

Data from *Educational Resources*, December catalog.

Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013).

Data from Hyatt Hotels. Available online at hyatt.com (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services.

Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).

Data from the Chancellor's Office, California Community Colleges, November 1994.

"State of the States." Gallup, 2013. Available online at http://www.gallup.com/poll/125066/St...ef=interactive (accessed June 17, 2013).

"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).

10.6 Two Population Means with Known Standard Deviations

Data from the United States Census Bureau. Available online at www.census.gov/prod/cen2010/b...c2010br-02.pdf

Hinduja, Sameer. "Sexting Research and Gender Differences." Cyberbulling Research Center, 2013. Available online at http://cyberbullying.us/blog/sexting...r-differences/ (accessed June 17, 2013).

"Smart Phone Users, By the Numbers." Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).

Smith, Aaron. "35% of American adults own a Smartphone." Pew Internet, 2013. Available online at www.pewinternet.org/~/media/F...martphones.pdf (accessed June 17, 2013).

"State-Specific Prevalence of Obesity AmongAduls—Unites States, 2007." MMWR, CDC. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm (accessed June 17, 2013).

"Texas Crime Rates 1960–1012." FBI, Uniform Crime Reports, 2013. Available online at: http://www.disastercenter.com/crime/txcrime.htm (accessed June 17, 2013).

7.18.2: Chapter References

11.2 Facts About the Chi-Square Distribution

Data from *Parade Magazine*.

“HIV/AIDS Epidemiology Santa Clara County.”Santa Clara County Public Health Department, May 2011.

11.3 Test of a Single Variance

“AppleInsider Price Guides.” Apple Insider, 2013. Available online at http://appleinsider.com/mac_price_guide (accessed May 14, 2013).

Data from the World Bank, June 5, 2012.

11.4 Goodness-of-Fit Test

Data from the U.S. Census Bureau

Data from the College Board. Available online at http://www.collegeboard.com.

Data from the U.S. Census Bureau, Current Population Reports.

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.

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).

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).

11.5 Test of Independence

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).

Harris Interactive, “Favorite Flavor of Ice Cream.” Available online at http://www.statisticbrain.com/favori...r-of-ice-cream (accessed May 24, 2013)

“Youngest Online Entrepreneurs List.” Available online at http://www.statisticbrain.com/younge...repreneur-list (accessed May 24, 2013).

11.6 Test for Homogeneity

Data from the Insurance Institute for Highway Safety, 2013. Available online at www.iihs.org/iihs/ratings (accessed May 24, 2013).

“Energy use (kg of oil equivalent per capita).” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...G.OE/countries (accessed May 24, 2013).

“Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES),” U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubsearch/pubsinf...?pubid=2009030 (accessed May 24, 2013).

“Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES),” U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubs2009/2009030_sup.pdf (accessed May 24, 2013).

7.19: Chapter Formula Review

9.4 Distribution Needed for Hypothesis Testing

Table 7.19.6 Test Statistics for Test of Means, Varying Sample Size, Population Known or Unknown

Sample size	Test statistic
< 30 (σ unknown)	$t_c = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$
< 30 (σ known)	$Z_c = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$
> 30 (σ unknown)	$Z_c = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$
> 30 (σ known)	$Z_c = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$

9.5 Full Hypothesis Test Examples

Test statistic for a hypothesis test of proportions:

$$Z_c = \frac{\mathrm{p}' - p_0}{\sqrt{\frac{p_0 p_0}{n}}}$$

7.20: Chapter Review

9.2 Null and Alternative Hypotheses

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 H0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤ or ≥)
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using not equal, less than or greater than symbols, i.e., (neq, <, 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.

9.3 Outcomes and the Type I and Type II Errors

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– β, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

9.4 Distribution Needed for Hypothesis Testing

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of successes and the mean number of failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

9.5 Full Hypothesis Test Examples

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph and calculate the test statistic.
5. Compare the calculated test statistic with the Z critical value determined by the level of significance required by the test and make a decision (cannot reject H_0 or cannot accept H_0), and write a clear conclusion using English sentences.

- **9.6: Chapter Review** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

7.H: Hypothesis Testing (Homework + Practice)

9.2 Null and Alternative Hypotheses

45. Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

46. Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

47. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

48. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. H_o: x¯$\overline{x}$ = 4.5, H_a : x¯$\overline{x}$ > 4.5
b. $H_o: \mu \geq 4.5$, $H_a: \mu < 4.5$
c. $H_o: \mu = 4.75$, $H_a: \mu > 4.75$
d. $H_o: \mu = 4.5$, $H_a: \mu > 4.5$

9.3 Outcomes and the Type I and Type II Errors

49. State the Type I and Type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

50. For statements a-j in Exercise 9.109, answer the following in complete sentences.

- a. State a consequence of committing a Type I error.
- b. State a consequence of committing a Type II error.

51. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

- a. To conclude the drug is safe when in, fact, it is unsafe.
- b. Not to conclude the drug is safe when, in fact, it is safe.
- c. To conclude the drug is safe when, in fact, it is safe.
- d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

52. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is ________.

- a. at least 20%, when in fact, it is less than 20%.
- b. 20%, when in fact, it is 20%.
- c. less than 20%, when in fact, it is at least 20%.
- d. less than 20%, when in fact, it is less than 20%.

53. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

- a. is more than seven hours.
- b. is at most seven hours.
- c. is at least seven hours.
- d. is less than seven hours.

54. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

- a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
- b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
- c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
- d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

9.4 Distribution Needed for Hypothesis Testing

55. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is X–X– ~ ________________

- a. N(7.24,1.9322√)N(7.24,1.9322)
- b. N(7.24,1.93)N(7.24,1.93)
- c. t_{22}
- d. t_{21}

9.5 Full Hypothesis Test Examples

56. A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, is the data highly inconsistent with the claim?

57. From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at

least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?
58. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?
59. An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?
60. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?
61. In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?
62. Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?
63. A Nissan Motor Corporation advertisement read, “The average man’s I.Q. is 107. The average brown trout’s I.Q. is 4. So why can’t man catch brown trout?” Suppose you believe that the brown trout’s mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.
64. Refer to Exercise 9.119. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout’s mean I.Q. is **not** four.

65. According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don’t believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

66. A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent’s survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

67. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it’s shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

68. Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California’s percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

69. According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year’s rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

70. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

71. Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

72. La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

73. Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?
After conducting the test, your decision and conclusion are

74. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.

At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

75. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

Instructions: For the following ten exercises,
Hypothesis testing: For the following ten exercises, answer each question.

a. *State the null and alternate hypothesis.*
b. *State the p-value.*
c. *State alpha.*
d. *What is your decision?*
e. *Write a conclusion.*
f. *Answer any other questions asked in the problem.*

76. According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

77. A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

78. Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

79. The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

80. For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use α = 0.01 level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

81. The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the α = 0.05 level, can it be concluded that the mean rainfall was below the reported average? What if α = 0.01? Assume the amount of summer rainfall follows a normal distribution.

82. A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the α = 0.10 level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

83. A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

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

At the α = 0.05 level can it be concluded that the sample mean is higher than 5.8 visits per year?

84. According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:

5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2

At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

85. The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

1.1: One-Sample Practice

9.2 Null and Alternative Hypotheses

1. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

2. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.
3. The American family has an average of two children. What is the random variable? Describe in words.
4. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.
5. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.
6. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.
7. In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.
8. Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0: ________
b. H_a: ________

9. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0: __________
b. H_a: __________

10. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: ________
b. H_a: ________

9.3 Outcomes and the Type I and Type II Errors

11. The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.
12. A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.
13. For 12 what are α and β in words?
14. In words, describe $1 - \beta$ For Exercise 12.

15. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

16. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

17. The power of a test is 0.981. What is the probability of a Type II error?

18. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

19. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

20. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

9.4 Distribution Needed for Hypothesis Testing

21. Which two distributions can you use for hypothesis testing for this chapter?

22. Which distribution do you use when you are testing a population mean and the population standard deviation is known? Assume sample size is large. Assume a normal distribution with $n \geq 30$.

23. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume a normal distribution, with $n \geq 30$.

24. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

25. A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

26. It is thought that 42% of respondents in a taste test would prefer Brand A. In a particular test of 100 people, 39% preferred Brand A. What distribution should you use to perform a hypothesis test?

27. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. What must you assume about the distribution of the data?

28. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

29. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

30. You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

31. You are performing a hypothesis test of a single population proportion. The data come from which distribution?

9.4 Full Hypothesis Test Examples

32. Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

33. Assume $H_0 : \mu \leq 6$ and \(H_a: \mu > 6). Is this a left-tailed, right-tailed, or two-tailed test?

34. Assume $H_0 : p = 0.25$ and $H_a : p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

35. Draw the general graph of a left-tailed test.

36. Draw the graph of a two-tailed test.

37. A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

38. Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

39. A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

40. You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

41. If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

42. Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

43. Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

44. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

1.2: Two-Sample Homework

10.2 Comparing Two Independent Population Means

64. The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

65. A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

66. At Rachel’s 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

67. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

68. Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls’ mean is higher than the boys’ mean.

Use the information from Appendix C: Data Sets to answer the next four exercises.

69. Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

70. Repeat the test in Table $1.2.16$. Test at the 1% level of significance.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

87. Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. Table $1.2.17$ details the number of tablet owners for each age group. Test at the 1% level of

significance.

Table 1.2.17

	16–29 year olds	30 years old and older
Own a tablet	69	231
Sample size	628	2,309

88. A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5% level of significance.

89. While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

90. We are interested in whether children’s educational computer software costs less, on average, than children’s entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children’s educational software costs less, on average, than children’s entertainment software.

91. Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?

92. "To Breakfast or Not to Breakfast?" by Richard Ayore

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, …, birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, “This is why I always work more hours than you guys!”

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in Table 1.2.18, solve our problem.

Table 1.2.18

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7

Work hours with breakfast	Work hours without breakfast
10	7
7	5
6	6
9	5

10.6 Two Population Means with Known Standard Deviations

NOTE

If you are using a Student's t-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

93. A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

94. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

95. A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $947 and $1,011, respectively. The population standard deviations are known to be $254 and $87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

96. Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg. Suppose that the population standard deviations are known to be six and three, respectively. Conduct a hypothesis test to evaluate the manufacturers claim.

97. A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

98. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). Table $1.2.19$ contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's score	2	2	3	3	4	2	1	1	2	4
Husband's score	2	2	1	3	2	1	1	1	2	4

10.7 Matched or Paired Samples

99. Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. The data are recorded in Table $1.2.20$. Do you think that their cholesterol levels were significantly lowered?

Table 1.2.20

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Use the following information to answer the next two exercises. A new AIDS prevention drug was tried on a group of 224 HIV positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript t = treated patient and ut = untreated patient.

100. The appropriate hypotheses are:

Use the following information to answer the next two exercises. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a "biofeedback exercise program." Six subjects were randomly selected and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after - before) producing the following results: $\overline{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.**101**.

101. The distribution for the test is:

a. t_5
b. t_6
c. $N(-10.2, 8.4)$
d. N $\left(-10.2, \frac{8.4}{\sqrt{6}}\right)$

102. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 10.21

The correct decision is:

a. Reject H_0.
b. Do not reject the H_0.

103.

A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are in the table below

Southern states	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

104.

A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is in the table below. Test at the 1% level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

105. A politician asked his staff to determine whether the underemployment rate in the northeast decreased from 2011 to 2012. The results are in the table below.

Northeastern states	2011	2012
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

1.3: Two-Sample Practice

10.2 Comparing Two Independent Population Means

Use the following information to answer the next 15 exercises: Indicate if the hypothesis test is for

a. independent group means, population standard deviations, and/or variances known
b. independent group means, population standard deviations, and/or variances unknown
c. matched or paired samples
d. single mean
e. two proportions
f. single proportion

1. It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.
2. A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.
3. A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.
4. The known standard deviation in salary for all mid-level professionals in the financial industry is \$11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is \$80,000. The sample mean salary for mid-level professionals in Company B is \$96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.
5. The average worker in Germany gets eight weeks of paid vacation.
6. According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.
7. It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.
8. The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?
9. In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?
10. A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after.
11. It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.
12. Varsity athletes practice five times a week, on average.
13. A sample of 12 in-state graduate school programs at school A has a mean tuition of \$64,000 with a standard deviation of \$8,000. At school B, a sample of 16 in-state graduate programs has a mean of \$80,000 with a standard deviation of \$6,000. On average, are the mean tuitions different?
14. A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?
15. A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

.**16**. Are standard deviations known or unknown?

17. What is the random variable?

18. Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for White people born in 1900 and 33.0 years for non-White people. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 White people, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 non-White people, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for White and non-White people.

19. Is this a test of means or proportions?

20. State the null and alternative hypotheses.

a. H_0: __________
b. H_a: __________

21. Is this a right-tailed, left-tailed, or two-tailed test?
22. In symbols, what is the random variable of interest for this test?
23. In words, define the random variable of interest for this test.
24. Which distribution (normal or Student's *t*) would you use for this hypothesis test?
25. Explain why you chose the distribution you did for Exercise 10.24.
26. Calculate the test statistic.
27. Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.
28. At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

29. Does it appear that the means are the same? Why or why not?

10.4 Comparing Two Independent Population Proportions

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

30. Is this a test of means or proportions?
31. What is the random variable?
32. State the null and alternative hypotheses.
33. What can you conclude about the two operating systems?

Use the following information to answer the next twelve exercises. In the recent Census, three percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

34. Is this a test of means or proportions?
35. State the null and alternative hypotheses.

a. H_0: _________
b. H_a: _________

36. Is this a right-tailed, left-tailed, or two-tailed test? How do you know?
37. What is the random variable of interest for this test?
38. In words, define the random variable for this test.
39. Which distribution (normal or Student's *t*) would you use for this hypothesis test?
40. Explain why you chose the distribution you did for the Exercise 10.56.
41. Calculate the test statistic.

42. At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

43. Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

10.6 Two Population Means with Known Standard Deviations

Use the following information to answer the next five exercises. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. the table below shows the result. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample mean speed of pitches (mph)	Population standard deviation
Wesley	86	3
Rodriguez	91	7

44. What is the random variable?

45. State the null and alternative hypotheses.

46. What is the test statistic?

47. At the 1% significance level, what is your conclusion?

Plant group	Sample mean height of plants (inches)	Population standard deviation
Food	16	2.5
No food	14	1.5

48. Is the population standard deviation known or unknown?

49. State the null and alternative hypotheses.

50. At the 1% significance level, what is your conclusion?

Use the following information to answer the next five exercises. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. 15 pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample mean melting temperatures (°F)	Population standard deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

51. State the null and alternative hypotheses.

52. Is this a right-, left-, or two-tailed test?

53. At the 1% significance level, what is your conclusion?

10.7 Matched or Paired Samples

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in Table 1.3.11. The "before" value is matched to an "after" value, and the differences are calculated. The differences have a normal distribution. Test at the 1% significance level.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

54. What is the random variable?

55.State the null and alternative hypotheses.

56. What conclusion can you draw about the software patch?

Use the following information to answer next five exercises. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1% significance level.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

57. State the null and alternative hypotheses.

58. What is the sample mean difference?

59. What conclusion can you draw about the juggling class?

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1% significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

60. State the null and alternative hypotheses.

61. What is the test statistic?

62. What is the sample mean difference?

63. What is the conclusion?

- **10.10: Practice** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

1.4: Chi-Square Tests Homework

11.2 Facts About the Chi-Square Distribution

Decide whether the following statements are true or false.

63. As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.

64. The standard deviation of the chi-square distribution is twice the mean.

65. The mean and the median of the chi-square distribution are the same if df = 24.

11.3 Test of a Single Variance

Use the following information to answer the next twelve exercises: Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

66. Is the traveler disputing the claim about the average or about the variance?

67. A sample standard deviation of 15 minutes is the same as a sample variance of __________ minutes.

68. Is this a right-tailed, left-tailed, or two-tailed test?

69. H_0: __________

70. df = ________

71. chi-square test statistic = ________

72. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the area associated with the level of confidence.

73.

Let $\alpha = 0.05$
Decision: ________
Conclusion (write out in a complete sentence.): ________

74. How did you know to test the variance instead of the mean?

75. If an additional test were done on the claim of the average delay, which distribution would you use?

76. If an additional test were done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?

77. A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. The standard deviation should be at most 0.5 oz. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

78. Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $84 and a sample standard deviation of $12, test the claim that the standard deviation is greater than $15.

79. Isabella, an accomplished **Bay to Breakers** runner, claims that the standard deviation for her time to run the 7.5 mile race is at most three minutes. To test her claim, Rupinder looks up five of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

80. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.

81. The number of births per woman in China is 1.6 down from 5.91 in 1966. This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether or not the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had had. The results are shown in Table 1.4.28. Does the students' survey indicate that the standard deviation is greater than 0.75?

Table 1.4.28

# of births	Frequency
0	5
1	30
2	10
3	5

82. According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a standard deviation of two. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the standard deviation is different from two? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11

83. The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a ten-ounce order of fries is at most 1.5 oz., but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 oz. and a standard deviation of two oz.

84. You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of $1,249 with a very narrow standard deviation of $25. You find a website that has a price comparison for the same computer at a series of stores as follows: $1,299; $1,229.99; $1,193.08; $1,279; $1,224.95; $1,229.99; $1,269.95; $1,249. Can you argue that pricing has a larger standard deviation than claimed by the manufacturer? Use the 5% significance level. As a potential buyer, what would be the practical conclusion from your analysis?

85. A company packages apples by weight. One of the weight grades is Class A apples. Class A apples have a mean weight of 150 g, and there is a maximum allowed weight tolerance of 5% above or below the mean for apples in the same consumer package. A batch of apples is selected to be included in a Class A apple package. Given the following apple weights of the batch, does the fruit comply with the Class A grade weight tolerance requirements. Conduct an appropriate hypothesis test.

a. at the 5% significance level
b. at the 1% significance level

Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166; 172;

11.4 Goodness-of-Fit Test

86. A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in Table 1.4.1 are the result of the 120 rolls.

Table 1.4.1

Face value	Frequency	Expected frequency
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

87.The marital status distribution of the U.S. male population, ages 15 and older, is as shown in Table 1.4.2.

Table 1.4.2

Marital status	Percent	Expected frequency
Never married	31.3	
Married	56.1	
Widowed	2.5	
Divorced/Separated	10.1	

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in Table 1.4.3, rounding to two decimal places.

Table 1.4.3

Marital status	Frequency
Never married	140
Married	238
Widowed	2
Divorced/Separated	20

Use the following information to answer the next two exercises: The columns in Table 1.4.4 contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the result of a survey of 1,000 local students from that year who took an AP Exam.

Table 1.4.4

Race/Ethnicity	AP examinee population	Overall student population	Survey frequency
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African-American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

88. Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

89. Perform a goodness-of-fit test to determine whether the local results follow the distribution of U.S. AP examinee population, based on ethnicity.

90. The City of South Lake Tahoe, CA, has an Asian population of 1,419 people, out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the Manhattan, NY, area yielded the data in Table 1.4.5. Conduct a goodness-of-fit test to determine if the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

Table 1.4.5

Race	Lake Tahoe frequency	Manhattan frequency

Race	Lake Tahoe frequency	Manhattan frequency
Asian Indian	131	174
Chinese	118	557
Filipino	1,045	518
Japanese	80	54
Korean	12	29
Vietnamese	9	21
Other	24	66

Use the following information to answer the next two exercises: UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006)*. Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for Table 1.4.6 shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5% significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Table 1.4.6

Business type	Number in class	Observed number that recycle one commodity	Expected number that recycle one commodity
Office	35	19	17.5
Retail/Wholesale	48	27	24
Food/Restaurants	53	35	26.5
Manufacturing/Medical	52	21	26
Hotel/Mixed	24	9	12

98. Table 1.4.7 contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. Perform a hypothesis test at the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.

Table 1.4.7

Age class (years)	Obese (percentage)	Expected USA average (percentage)
20–30	75.0	32.6
31–40	26.5	32.6
41–50	13.6	36.6
51–60	21.9	36.6
61–70	21.0	39.7

11.5 Test of Independence

99. A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. ski area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Table 1.4.8

100. Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in Table 1.4.9. Conduct a test of independence.

Table 1.4.9

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

101. College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Table 1.4.10 shows the data. Conduct a test of independence.

Major	< $50,000	$50,000 – $68,999	$69,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Table 1.4.10

102. Some travel agents claim that honeymoon hot spots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in Table 1.4.11. Conduct a test of independence.

Table 1.4.11

Location	20–29	30–39	40–49	50 and over
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

103. A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence.

Sport	18 - 25	26 - 30	31 - 40	41 and over
Racquetball	42	58	30	46
Tennis	58	76	38	65
Swimming	72	60	65	33

Table $1.4.12$

104. A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in Table $1.4.13$. Conduct a test of independence.

Table $1.4.13$

Type of Fries	Northeast	South	Central	West
Skinny fries	70	50	20	25
Curly fries	100	60	15	30
Steak fries	20	40	10	10

105. According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of males	None	< $200,000	$200,000–$400,000	$401,001–$1,000,000	$1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

Table $1.4.14$

106. Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test of independence.

Table $1.4.15$

Annual salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
< $30,000	15	25	10	5
$30,000–$40,000	20	40	70	30
$40,000–$50,000	10	20	40	55
$50,000–$60,000	5	10	20	60
$60,000+	0	5	10	150

Read the statement and decide whether it is true or false.

107. The number of degrees of freedom for a test of independence is equal to the sample size minus one.

108. The test for independence uses tables of observed and expected data values.

109. The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.

110. In a test of independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

111. An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the U.S. Based on Table 1.4.16, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5% significance level.

U.S. region/Flavor	Strawberry	Chocolate	Vanilla	Rocky road	Mint chocolate chip	Pistachio	Row total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Column total	45	112	101	46	60	27	391

Table 1.4.16

112. **Table 1.4.17** provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5% significance level.

Table 1.4.17

Age group\ Net worth value (in millions of US dollars)	1–5	6–24	≥25	Row total
17–25	8	7	5	20
26–30	6	5	9	20
Column total	14	12	14	40

113. A 2013 poll in California surveyed people about taxing sugar-sweetened beverages. The results are presented in Table 1.4.18, and are classified by ethnic group and response type. Are the poll responses independent of the participants’ ethnic group? Conduct a test of independence at the 5% significance level.

Opinion/Ethnicity	Asian-American	White/Non-Hispanic	African-American	Latino	Row total
Against tax	48	433	41	160	682
In favor of tax	54	234	24	147	459
No opinion	16	43	16	19	94
Column total	118	710	81	326	1235

Table 1.4.18

114. A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown in Table 1.4.19. Conduct a test of homogeneity. Test at a 5% level of significance.

Table 1.4.19

	Open	**Conscientious**	**Extrovert**	**Agreeable**	**Neurotic**
Business	41	52	46	61	58
Social Science	72	75	63	80	65

115. Do men and women select different breakfasts? The breakfasts ordered by randomly selected men and women at a popular breakfast place is shown in Table 1.4.20. Conduct a test for homogeneity at a 5% level of significance.

	French toast	**Pancakes**	**Waffles**	**Omelettes**
Men	47	35	28	53
Women	65	59	55	60

Table 1.4.20

116. A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform a test for homogeneity at a 5% level of significance.

117. In 2007, the United States had 1.5 million homeschooled students, according to the U.S. National Center for Education Statistics. In Table 1.4.21 you can see that parents decide to homeschool their children for different reasons, and some reasons are ranked by parents as more important than others. According to the survey results shown in the table, is the distribution of applicable reasons the same as the distribution of the most important reason? Provide your assessment at the 5% significance level. Did you expect the result you obtained?

Reasons for fomeschooling	Applicable reason (in thousands of respondents)	Most important reason (in thousands of respondents)	Row total
Concern about the environment of other schools	1,321	309	1,630
Dissatisfaction with academic instruction at other schools	1,096	258	1,354
To provide religious or moral instruction	1,257	540	1,797
Child has special needs, other than physical or mental	315	55	370
Nontraditional approach to child’s education	984	99	1,083
Other reasons (e.g., finances, travel, family time, etc.)	485	216	701
Column total	5,458	1,477	6,935

Table 1.4.21

118. When looking at energy consumption, we are often interested in detecting trends over time and how they correlate among different countries. The information in Table 1.4.52 shows the average energy use (in units of kg of oil equivalent per capita) in the USA and the joint European Union countries (EU) for the six-year period 2005 to 2010. Do the energy use values in these two areas come from the same distribution? Perform the analysis at the 5% significance level.

Table 1.4.22

Year	European Union	United States	Row total
2010	3,413	7,164	10,557
2009	3,302	7,057	10,359
2008	3,505	7,488	10,993
2007	3,537	7,758	11,295
2006	3,595	7,697	11,292
2005	3,613	7,847	11,460
Column total	20,965	45,011	65,976

119. The Insurance Institute for Highway Safety collects safety information about all types of cars every year, and publishes a report of Top Safety Picks among all cars, makes, and models. Table 1.4.23 presents the number of Top Safety Picks in six car categories for the two years 2009 and 2013. Analyze the table data to conclude whether the distribution of cars that earned the Top Safety Picks safety award has remained the same between 2009 and 2013. Derive your results at the 5% significance level.

Year \ Car type	Small	Mid-size	Large	Small SUV	Mid-size SUV	Large SUV	Row total
2009	12	22	10	10	27	6	87
2013	31	30	19	11	29	4	124
Column total	43	52	29	21	56	10	211

Table 1.4.23

11.7 Comparison of the Chi-Square Tests

120. Is there a difference between the distribution of community college statistics students and the distribution of university statistics students in what technology they use on their homework? Of some randomly selected community college students, 43 used a computer, 102 used a calculator with built in statistics functions, and 65 used a table from the textbook. Of some randomly selected university students, 28 used a computer, 33 used a calculator with built in statistics functions, and 40 used a table from the textbook. Conduct an appropriate hypothesis test using a 0.05 level of significance.

Read the statement and decide whether it is true or false.

121. If $df = 2$, the chi-square distribution has a shape that reminds us of the exponential.

1.5: Chi-Square Practice

11.2 Facts About the Chi-Square Distribution

1. If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

2. If *df* > 90, the distribution is _____________. If *df* = 15, the distribution is ________________.
3. When does the chi-square curve approximate a normal distribution?
4. Where is μ located on a chi-square curve?

5. Is it more likely the *df* is 90, 20, or two in the graph?

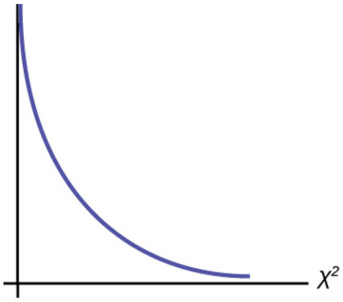


Figure 1.5.1

11.3 Test of a Single Variance

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.

6. What type of test should be used?

7. State the null and alternative hypotheses.

8. Is this a right-tailed, left-tailed, or two-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.

9. What type of test should be used?
10. State the null and alternative hypotheses.
11. *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.

12. What type of test should be used?
13. What is the test statistic?
14. What can you conclude at the 5% significance level?

11.4 Goodness-of-Fit Test

Determine the appropriate test to be used in the next three exercises.

15. An archaeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archaeologist 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.
16. 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.
17. 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 Table 1.5.1.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

Table 1.5.1

The actual distribution for a class of 20 is in Table 1.5.2

Grade	Frequency
A	7
B	7
C	5
D	1

Table 1.5.2

18. df= df = ______
19. State the null and alternative hypotheses.
20. χ^2 test statistic = ______
21. At the 5% significance level, what can you conclude?

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 Table 1.5.3

Ethnicity	Number of cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

Table 1.5.3

The percentage of each ethnic group in Santa Clara County is as in Table 1.5.4

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%	

Table 1.5.4

22. 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.

23. H_0: _______
24. H_a: _______
25. Is this a right-tailed, left-tailed, or two-tailed test?
26. degrees of freedom = _______
27. χ^2 test statistic = _______

28. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the confidence level.

Let $\alpha = 0.05$

Decision: ________________

Reason for the Decision: ________________

Conclusion (write out in complete sentences): ________________

29. Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

11.4 Test of Independence

Determine the appropriate test to be used in the next three exercises.

30. 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.

31. 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.

32. 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.

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 1.5.1 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 1.5.5

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

Traveling distance	Third class	Second class	First class	Total
Total	73	67	60	200

33. State the hypotheses.
H_0: _______
H_a: _______

34. df = _______

35. How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

36. How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

37. What is the test statistic?

38. 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.

39. Complete the table.

Table 1.5.6 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						

40. State the hypotheses.
H_0: _______
H_a: _______

41. Enter expected values in Table 1.5.26. Round to two decimal places.

Calculate the following values:

42. df = _______

43. χ^2 test statistic = ______

44. Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

45. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the confidence level.

Figure 1.5.12

State the decision and conclusion (in a complete sentence) for the following preconceived levels of *alpha*.

46.

$\alpha = 0.05$

1. Decision: ____________________
2. Reason for the decision: ____________________
3. Conclusion (write out in a complete sentence): ____________________

47.

$\alpha = 0.01$

1. Decision: ____________________
2. Reason for the decision: ____________________
3. Conclusion (write out in a complete sentence): ____________________

11.6 Test for Homogeneity

48. A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

49. What are the null and alternative hypotheses for Table 1.5.7.

Table 1.5.7

	20–30	30–40	40–50	50–60
Private practice	16	40	38	6
Hospital	8	44	59	39

53. State the null and alternative hypotheses.

54. df = _______

55. What is the test statistic?

56. What can you conclude at the 5% significance level?

11.7 Comparison of the Chi-Square Tests

57. Which test do you use to decide whether an observed distribution is the same as an expected distribution?

58. What is the null hypothesis for the type of test from **Exercise 1.5.57**?

59. Which test would you use to decide whether two factors have a relationship?

60. Which test would you use to decide if two populations have the same distribution?

61. How are tests of independence similar to tests for homogeneity?

62. How are tests of independence different from tests for homogeneity?

7.S: Hypothesis Testing (Solutions- Practice + Homework)

38. a right-tailed test

40. a left-tailed test

42. This is a left-tailed test.

44. This is a two-tailed test.

45.

a. H_0: $\mu = 34$; H_a: $\mu \neq 34$
b. H_0: $p \leq 0.60$; H_a: $p > 0.60$
c. H_0: $\mu \geq 100{,}000$; H_a: $\mu < 100{,}000$
d. H_0: $p = 0.29$; H_a: $p \neq 0.29$
e. H_0: $p = 0.05$; H_a: $p < 0.05$
f. H_0: $\mu \leq 10$; H_a: $\mu > 10$
g. H_0: $p = 0.50$; H_a: $p \neq 0.50$
h. H_0: $\mu = 6$; H_a: $\mu \neq 6$
i. H_0: $p \geq 0.11$; H_a: $p < 0.11$
j. H_0: $\mu \leq 20{,}000$; H_a: $\mu > 20{,}000$

47.c

49.

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.
f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000

51. b

55. d

56.

a. H_0: $\mu \geq 50{,}000$
b. H_a: $\mu < 50{,}000$
c. Let $\overline{\text{X}}$ $\overline{X}$ = the average lifespan of a brand of tires.
d. normal distribution
e. z = -2.315
f. p-value = 0.0103
g. Check student’s solution.
h. i. alpha: 0.05
  ii. Decision: Reject the null hypothesis.
  iii. Reason for decision: The p-value is less than 0.05.
  iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.
i. (43,537, 49,463)

58.

a. H_0: $\mu = \$1.00$
b. H_a: $\mu \neq \$1.00$
c. Let $\overline{\text{X}}$ $\overline{X}$ = the average cost of a daily newspaper.
d. normal distribution
e. z = –0.866
f. p-value = 0.3865
g. Check student’s solution.
h. i. Alpha: 0.01
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value is greater than 0.01.
  iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is \$1. The mean cost could be \$1.
i. (\$0.84,

60.

a. H_0: $\mu = 10$
b. H_a: $\mu \neq 10$
c. Let $\overline{\text{X}}$ $\overline{X}$ the mean number of sick days an employee takes per year.
d. Student’s t-distribution
e. t = –1.12
f. p-value = 0.300
g. Check student’s solution.
h. i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value is greater than 0.05.
  iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.
i. (4.9443, 11.806)

62.

a. H_0: $p \geq 0.6$
b. H_a: $p < 0.6$
c. Let P' = the proportion of students who feel more enriched as a result of taking Elementary Statistics.
d. normal for a single proportion
e. 1.12
f. p-value = 0.1308
g. Check student’s solution.
h. i. Alpha: 0.05

  - ii. Decision: Do not reject the null hypothesis.
  - iii. Reason for decision: The *p*-value is greater than 0.05.
  - iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. Confidence Interval: (0.409, 0.654)
The “plus-4s” confidence interval is (0.411, 0.648)

64.

a. H_0: $\mu = 4$
b. H_a: $\mu \neq 4$
c. Let $\overline{\text{X}}$ $\overline{X}$ the average I.Q. of a set of brown trout.
d. two-tailed Student's t-test
e. $t = 1.95$
f. *p*-value = 0.076
g. Check student’s solution.
h.
  - i. Alpha: 0.05
  - ii. Decision: Reject the null hypothesis.
  - iii. Reason for decision: The *p*-value is greater than 0.05
  - iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

i. (3.8865,5.9468)

66.

a. H_0: $p \geq 0.13$
b. H_a: $p < 0.13$
c. Let P' = the proportion of Americans who have seen or sensed angels
d. normal for a single proportion
e. –2.688
f. *p*-value = 0.0036
g. Check student’s solution.
h.
  - i. alpha: 0.05
  - ii. Decision: Reject the null hypothesis.
  - iii. Reason for decision: The *p*-value is less than 0.05.
  - iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.

i. (0, 0.0623).
The“plus-4s” confidence interval is (0.0022, 0.0978)

69.

a. H_0: $p = 0.14$
b. H_a: $p < 0.14$
c. Let P' = the proportion of NYC residents that smoke.
d. normal for a single proportion
e. –0.2756
f. *p*-value = 0.3914
g. Check student’s solution.
h.
  - i. alpha: 0.05
  - ii. Decision: Do not reject the null hypothesis.
  - iii. Reason for decision: The *p*-value is greater than 0.05.
  - iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.

i. Confidence Interval: (0.0502, 0.2070): The “plus-4s” confidence interval (see chapter 8) is (0.0676, 0.2297).

71.

a. H_0: $\mu = 69{,}110$
b. H_a: $\mu > 69{,}110$
c. Let X¯¯¯¯¯ $\overline{X}$ = the mean salary in dollars for California registered nurses.
d. Student's *t*-distribution
e. $t = 1.719$
f. *p*-value: 0.0466
g. Check student’s solution.
h.
    i. Alpha: 0.05
    ii. Decision: Reject the null hypothesis.
    iii. Reason for decision: The *p*-value is less than 0.05.
    iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

i. ($68,757, $73,485)

73. c

75. c

77.

a. H_0: $p = 0.488$ H_a: $p \neq 0.488$
b. *p*-value = 0.0114
c. alpha = 0.05
d. Reject the null hypothesis.
e. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.
f. The survey does not appear to be accurate.

79.

a. H_0: $p = 0.517$ H_a: $p \neq 0.517$
b. *p*-value = 0.9203.
c. alpha = 0.05.
d. Do not reject the null hypothesis.
e. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample’s population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

81.

a. H_0: $\mu \geq 11.52$ H_a: $\mu < 11.52$
b. *p*-value = 0.000002 which is almost 0.
c. alpha = 0.05.
d. Reject the null hypothesis.
e. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1% because the *p*-value is almost 0.

83.

a. H_0: $\mu \leq 5.8$ H_a: $\mu > 5.8$
b. *p*-value = 0.9987
c. alpha = 0.05
d. Do not reject the null hypothesis.
e. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

85.

1. $H_0 : \mu \geq 150 H_a : \mu < 150$
2. p-value = 0.0622
3. alpha = 0.01
4. Do not reject the null hypothesis.
5. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
6. The student academic group's claim appears to be correct.

1.1: Two-Sample (Solutions- Practice + Homework)

1. two proportions

3. matched or paired samples

5. single mean

7. independent group means, population standard deviations and/or variances unknown

9. two proportions

11. independent group means, population standard deviations and/or variances unknown

13. independent group means, population standard deviations and/or variances unknown

15. two proportions

17. The random variable is the difference between the mean amounts of sugar in the two soft drinks.

19. means

21. two-tailed

23. the difference between the mean life spans of whites and nonwhites

25. This is a comparison of two population means with unknown population standard deviations.

27. Check student’s solution.

28.

a. Cannot accept the null hypothesis
b. *p*-value < 0.05
c. There is not enough evidence at the 5% level of significance to support the claim that life expectancy in the 1900s is different between White and non-White people.

31. $P'_{OS1} - P'_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

34. proportions

36. right-tailed

38. The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

40. Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

42.

a. Cannot accept the null hypothesis.
b. *p*-value < alpha
c. At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

44. The difference in mean speeds of the fastball pitches of the two pitchers

46–2.46

47. At the 1% significance level, we can reject the null hypothesis. There is sufficient data to conclude that the mean speed of Rodriguez’s fastball is faster than Wesley’s.

49. Subscripts: 1 = Food, 2 = No Food

$H_0 : \mu_1 \leq \mu_2$

$H_a : \mu_1 > \mu_2$

51. Subscripts: 1 = Gamma, 2 = Zeta

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

53. There is sufficient evidence so we cannot accept the null hypothesis. The data support that the melting point for Alloy Zeta is different from the melting point of Alloy Gamma.

54. the mean difference of the system failures

56. With a p-value 0.0067, we can cannot accept the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

60.

$H_0 : \mu_d \geq 0$

$H_a : \mu_d < 0$

63. We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

65. Subscripts: 1: two-year colleges; 2: four-year colleges

67.

Subscripts: 1: mechanical engineering; 2: electrical engineering

69.

1. 71.

74.

c

Test: two independent sample means, population standard deviations unknown. Random variable:

$$\overline{X}_1 - \overline{X}_2$$

Distribution: $H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 < \mu_2$. The mean age of entering prostitution in Canada is lower than the mean age in the United States.

Graph: left-tailed p-value : 0.0151

Decision: Cannot reject H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering prostitution in Canada is lower than the mean age in the United States.

78. d

80.

a. H_0: $P_W = P_B$
b. H_a: $P_W \neq P_B$
c. The random variable is the difference in the proportions of White and Black suicide victims, aged 15 to 24.
d. normal for two proportions
e. test statistic: –0.1944
f. p-value: 0.8458
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Cannot accept the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportions of White and Black female suicide victims, aged 15 to 24, are different.

82.

1 = Cabrillo College, 2 = Lake Tahoe College

a. H0:p1=p2$H0:p1=p2$
b. Ha:p1≠p2$Ha:p1\neq p2$
c. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
d. normal for two proportions
e. test statistic: 4.29
f. *p*-value: 0.00002
g. Check student’s solution.
h.
    i. Alpha: 0.05
    ii. Decision: Cannot accept the null hypothesis.
    iii. Reason for decision: *p*-value < alpha
    iv. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

84. a

85. Test: two independent sample proportions.

Random variable: $p'_1 - p'_2$

Distribution: $H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$. The proportion of eReader users is different for the 16- to 29-year-old users from that of the 30 and older users.

Graph: two-tailed

87.

Test: two independent sample proportions

Random variable: $p'_1 - p'_2$

Distribution:$H_0 : p_1 = p_2$ $H_a : p_1 > p_2$. A higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Graph: right-tailed

Do not reject the H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

89.

Subscripts: 1: men; 2: women

a. H0:p1≤p2
b. Ha:p1>p2
c. P′1−P′2 is the difference between the proportions of men and women who enjoy shopping for electronic equipment.
d. normal for two proportions
e. test statistic: 0.22
f. *p*-value: 0.4133
g. Check student’s solution.
h.
    i. Alpha: 0.05
    ii. Decision: Cannot reject the null hypothesis.
    iii. Reason for Decision: *p*-value > alpha
    iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

91.

a. H0:p1=p2
b. Ha:p1≠p2
c. P′1−P′2 is the difference between the proportions of men and women that have at least one pierced ear.

d. normal for two proportions
e. test statistic: –4.82
f. p-value: zero
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Cannot accept the null hypothesis.
   iii. Reason for Decision: p-value < alpha
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

92.

a. H_0: $\mu_d = 0$
b. H_a: $\mu_d > 0$
c. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
d. t_9
e. test statistic: 4.8963
f. p-value: 0.0004
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Cannot accept the null hypothesis.
   iii. Reason for Decision: p-value < alpha
   iv. Conclusion: At the 5% level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

94.

Subscripts: 1 = boys, 2 = girls

a. $H0:\mu1\leq\mu2$
b. $Ha:\mu1>\mu2$
c. The random variable is the difference in the mean auto insurance costs for boys and girls.
d. normal
e. test statistic: $z = 2.50$
f. p-value: 0.0062
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Cannot accept the null hypothesis.
   iii. Reason for Decision: p-value < alpha
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

96.

Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans

a. $H0:\mu1\geq\mu2$
b. $Ha:\mu1<\mu2$
c. The random variable is the difference in the mean miles per gallon of non-hybrid sedans and hybrid sedans.
d. normal
e. test statistic: 6.36
f. p-value: 0
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Cannot accept the null hypothesis.
   iii. Reason for decision: p-value < alpha

    iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

98.

a. H_0: $\mu_d = 0$
b. H_a: $\mu_d < 0$
c. The random variable X_d is the average difference between husband's and wife's satisfaction level.
d. t_9
e. test statistic: $t = –1.86$
f. p-value: 0.0479
g. Check student's solution
h.
    i. Alpha: 0.05
    ii. Decision: Cannot accept the null hypothesis, but run another test.
    iii. Reason for Decision: p-value < alpha
    iv. Conclusion: This is a weak test because alpha and the p-value are close. However, there is insufficient evidence to conclude that the mean difference is negative.

99. p-value = 0.1494

103. Test: two matched pairs or paired samples (t-test)

Random variable: $\overline{X}_d$

Distribution: t_{12}

$H_0 : \mu_d = 0 \;\; H_a : \mu_d > 0$

The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012.

Graph: right-tailed

p-value: 0.0004

Decision: Cannot accept H_0

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

105. Test: matched or paired samples (t-test)

Difference data: $\{-0.9, -3.7, -3.2, -0.5, 0.6, -1.9, -0.5, 0.2, 0.6, 0.4, 1.7, -2.4, 1.8\}$

Random Variable: $\overline{X}_d$

Distribution: $H_0 : \mu_d = 0 H_a : \mu_d < 0$

The mean of the differences of the rate of underemployment in the northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012.

Graph: left-tailed.

Decision: Cannot reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

107. e

109. d

111. f

113. e

115. f

117. a

1.2: Chi-square (Solutions- Practice + Homework)

1.

mean = 25 and standard deviation = 7.0711

3.

when the number of degrees of freedom is greater than 90

5.

$df = 2$

6.

a test of a single variance

8.

a left-tailed test

10.

$H_0 : \sigma^2 = 0.812;$

$H_a : \sigma^2 > 0.812.$

12.

a test of a single variance

16.

a goodness-of-fit test

18.

3

20.

2.04

21.

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.

23.

H_0: the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

25.

right-tailed

27.

2016.136

28.

- **30**.

  a test of independence

  a test of independence

  34.

  8

  36.

6.6

39.

Table 1.2.54

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

41.

Table 1.2.55

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

43.

10,301.8

44.

right

46.

1. **48**.

   test for homogeneity

   test for homogeneity

   52.

   All values in the table must be greater than or equal to five.

   54.

   3

   57.

   a goodness-of-fit test

   59.

   a test for independence

   61.

   Answers will vary. Sample answer: Tests of independence and tests for homogeneity both calculate the test statistic the same way $\sum_{(ij)} \frac{(O-E)^2}{E}$. In addition, all values must be greater than or equal to five.

63.

true

65.

false

67.

225

69.

$H_0 : \sigma^2 \leq 150$

71.

36

72.

Check student’s solution.

74.

The claim is that the variance is no more than 150 minutes.

76.

a Student's t- or normal distribution

78.

1. **80**.
   1. **82**.
      1. **84**.
         1. **87**.

Table 1.2.56

Marital status	Percent	Expected frequency
Never married	31.3	125.2
Married	56.1	224.4
Widowed	2.5	10
Divorced/Separated	10.1	40.4

1. **89**.
   1. **91**.
      1. **94**.

true

false

98.

1. **100**.
   1. **102**.
      1. **104**.
         1. **106**.
            1. **108**.

true

true

112.

1. **114**.
   1. **116**.
      1. **118**.
         1. **120**.
            1. **122**.
               1. The test statistic is always positive and if the expected and observed values are not close together, the test statistic is large and the null hypothesis will be rejected.
               2. Testing to see if the data fits the distribution “too well” or is too perfect.

CHAPTER OVERVIEW

8: Linear Regression and Correlation

8.1: Introduction to Linear Regression and Correlation

Figure 13.1 Linear regression and correlation can help you determine if an auto mechanic’s salary is related to his work experience. (credit: Joshua Rothhaas)

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability, or your gender or color. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

These examples may or may not be tied to a model, meaning that some theory suggested that a relationship exists. This link between a cause and an effect, often referred to as a model, is the foundation of the scientific method and is the core of how we determine what we believe about how the world works. Beginning with a theory and developing a model of the theoretical relationship should result in a prediction, what we have called a hypothesis earlier. Now the hypothesis concerns a full set of relationships. As an example, in Economics the model of consumer choice is based upon assumptions concerning human behavior: a desire to maximize something called utility, knowledge about the benefits of one product over another, likes and dislikes, referred to generally as preferences, and so on. These combined to give us the demand curve. From that we have the prediction that as prices rise the quantity demanded will fall. Economics has models concerning the relationship between what prices are charged for goods and the market structure in which the firm operates, monopoly verse competition, for example. Models for who would be most likely to be chosen for an on-the-job training position, the impacts of Federal Reserve policy changes and the growth of the economy and on and on.

Models are not unique to Economics, even within the social sciences. In political science, for example, there are models that predict behavior of bureaucrats to various changes in circumstances based upon assumptions of the goals of the bureaucrats. There are models of political behavior dealing with strategic decision making both for international relations and domestic politics.

The so-called hard sciences are, of course, the source of the scientific method as they tried through the centuries to explain the confusing world around us. Some early models today make us laugh; spontaneous generation of life for example. These early models are seen today as not much more than the foundational myths we developed to help us bring some sense of order to what seemed chaos.

The foundation of all model building is the perhaps the arrogant statement that we know what caused the result we see. This is embodied in the simple mathematical statement of the functional form that $y=f(x)$. The response, Y, is caused by the stimulus, X. Every model will eventually come to this final place and it will be here that the theory will live or die. Will the data support this hypothesis? If so then fine, we shall believe this version of the world until a better theory comes to replace it. This is the process by which we moved from flat earth to round earth, from earth-center solar system to sun-center solar system, and on and on.

The scientific method does not confirm a theory for all time: it does not prove “truth”. All theories are subject to review and may be overturned. These are lessons we learned as we first developed the concept of the hypothesis test earlier in this book. Here, as we begin this section, these concepts deserve review because the tool we will develop here is the cornerstone of the scientific method and the stakes are higher. Full theories will rise or fall because of this statistical tool; regression and the more advanced versions call econometrics.

In this chapter we will begin with correlation, the investigation of relationships among variables that may or may not be founded on a cause and effect model. The variables simply move in the same, or opposite, direction. That is to say, they do not move randomly. Correlation provides a measure of the degree to which this is true. From there we develop a tool to measure cause and effect relationships; regression analysis. We will be able to formulate models and tests to determine if they are statistically sound. If they are found to be so, then we can use them to make predictions: if as a matter of policy we changed the value of this variable what would happen to this other variable? If we imposed a gasoline tax of 50 cents per gallon how would that effect the carbon emissions, sales of Hummers/Hybrids, use of mass transit, etc.? The ability to provide answers to these types of questions is the value of regression as both a tool to help us understand our world and to make thoughtful policy decisions.

8.2: The Correlation Coefficient r

As we begin this section we note that the type of data we will be working with has changed. Perhaps unnoticed, all the data we have been using is for a single variable. It may be from two samples, but it is still a univariate variable. The type of data described in the examples above and for any model of cause and effect is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

For our work we can classify data into three broad categories, time series data, cross-section data, and panel data. We met the first two very early on. Time series data measures a single unit of observation; say a person, or a company or a country, as time passes. What are measured will be at least two characteristics, say the person's income, the quantity of a particular good they buy and the price they paid. This would be three pieces of information in one time period, say 1985. If we followed that person across time we would have those same pieces of information for 1985,1986, 1987, etc. This would constitute a times series data set. If we did this for 10 years we would have 30 pieces of information concerning this person's consumption habits of this good for the past decade and we would know their income and the price they paid.

A second type of data set is for cross-section data. Here the variation is not across time for a single unit of observation, but across units of observation during one point in time. For a particular period of time we would gather the price paid, amount purchased, and income of many individual people.

A third type of data set is panel data. Here a panel of units of observation is followed across time. If we take our example from above we might follow 500 people, the unit of observation, through time, ten years, and observe their income, price paid and quantity of the good purchased. If we had 500 people and data for ten years for price, income and quantity purchased we would have 15,000 pieces of information. These types of data sets are very expensive to construct and maintain. They do, however, provide a tremendous amount of information that can be used to answer very important questions. As an example, what is the effect on the labor force participation rate of women as their family of origin, mother and father, age? Or are there differential effects on health outcomes depending upon the age at which a person started smoking? Only panel data can give answers to these and related questions because we must follow multiple people across time. The work we do here however will not be fully appropriate for data sets such as these.

Beginning with a set of data with two independent variables we ask the question: are these related? One way to visually answer this question is to create a scatter plot of the data. We could not do that before when we were doing descriptive statistics because those data were univariate. Now we have bivariate data so we can plot in two dimensions. Three dimensions are possible on a flat piece of paper, but become very hard to fully conceptualize. Of course, more than three dimensions cannot be graphed although the relationships can be measured mathematically.

To provide mathematical precision to the measurement of what we see we use the correlation coefficient. The correlation tells us something about the co-movement of two variables, but **nothing** about why this movement occurred. Formally, correlation analysis assumes that both variables being analyzed are **independent** variables. This means that neither one causes the movement in the other. Further, it means that neither variable is dependent on the other, or for that matter, on any other variable. Even with these limitations, correlation analysis can yield some interesting results.

The correlation coefficient, ρ (pronounced rho), is the mathematical statistic for a population that provides us with a measurement of the strength of a linear relationship between the two variables. For a sample of data, the statistic, r, developed by Karl Pearson in the early 1900s, is an estimate of the population correlation and is defined mathematically as:

$$r=\frac{\frac{1}{n-1} \Sigma\left(X_{1 i}-\overline{X}_{1}\right)\left(X_{2 i}-\overline{X}_{2}\right)}{s_{x_{1}} s_{x_{2}}}$$

OR

$$r=\frac{\sum X_{1 i} X_{2 i}-n \overline{X}_{1}-\overline{X}_{2}}{\sqrt{\left(\Sigma X_{1 i}^{2}-n \overline{X}_{1}^{2}\right)\left(\Sigma X_{2 i}^{2}-n \overline{X}_{2}^{2}\right)}}$$

where sx_1 and sx_2 are the standard deviations of the two independent variables X_1 and X_2, $\overline{X}_1$ and $\overline{X}_2$ are the sample means of the two variables, and X_{1i} and X_{2i} are the individual observations of X_1 and X_2. The correlation coefficient r ranges in value from -1 to 1. The second equivalent formula is often used because it may be computationally easier. As scary as these formulas look they are really just the ratio of the covariance between the two variables and the product of their two standard deviations. That is to say, it is a measure of relative variances.

In practice all correlation and regression analysis will be provided through computer software designed for these purposes. Anything more than perhaps one-half a dozen observations creates immense computational problems. It was because of this fact that correlation, and even more so, regression, were not widely used research tools until after the advent of “computing machines”. Now the computing power required to analyze data using regression packages is deemed almost trivial by comparison to just a decade ago.

To visualize any **linear** relationship that may exist review the plot of a scatter diagrams of the standardized data. Figure $8.2.2$ presents several scatter diagrams and the calculated value of r. In panels (a) and (b) notice that the data generally trend together, (a) upward and (b) downward. Panel (a) is an example of a positive correlation and panel (b) is an example of a negative correlation, or relationship. The sign of the correlation coefficient tells us if the relationship is a positive or negative (inverse) one. If all the values of X_1 and X_2 are on a straight line the correlation coefficient will be either 1 or -1 depending on whether the line has a positive or negative slope and the closer to one or negative one the stronger the relationship between the two variables. BUT ALWAYS REMEMBER THAT THE CORRELATION COEFFICIENT DOES NOT TELL US THE SLOPE.

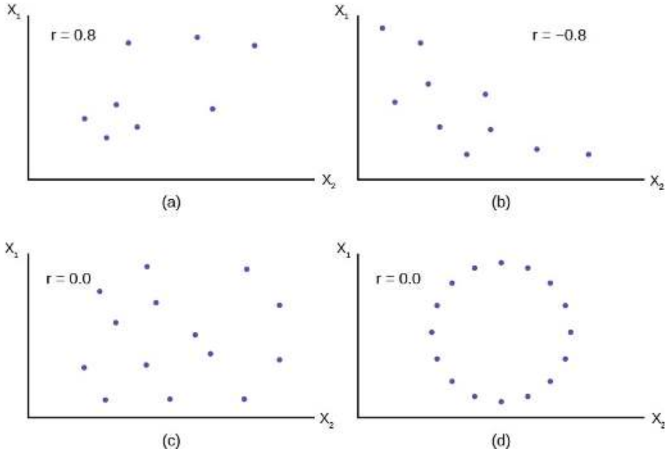


Figure 8.2.2

Remember, all the correlation coefficient tells us is whether or not the data are linearly related. In panel (d) the variables obviously have some type of very specific relationship to each other, but the correlation coefficient is zero, indicating no **linear** relationship exists.

If you suspect a linear relationship between X_1 and X_2 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: $-1 \leq r \leq 1$.
- The size of the correlation *r* indicates the strength of the **linear** relationship between X_1 and X_2. Values of *r* close to –1 or to +1 indicate a stronger linear relationship between X_1 and X_2.
- If $r = 0$ there is absolutely no linear relationship between X_1 and X_2 **(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: ANY straight line no matter what the slope. 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_1 increases, X_2 tends to increase and when X_1 decreases, X_2 tends to decrease **(positive correlation)**.
- A negative value of *r* means that when X_1 increases, X_2 tends to decrease and when X_1 decreases, X_2 tends to increase **(negative correlation)**.

Note

Strong correlation does not suggest that X_1 causes X_2 or X_2 causes X_1. We say **"correlation does not imply causation."**

8.3: Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between X_1 and X_2.

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.

- ρ = 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_1 and X_2 because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship X_1 and X_2. 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".

Performing the Hypothesis Test

- **Null Hypothesis: H_0: $\rho = 0$**
- **Alternate Hypothesis: H_a: $\rho \neq 0$**

Drawing a Conclusion There are two methods of making the decision concerning the hypothesis. The test statistic to test this hypothesis is:

$$t_c = \frac{r}{\sqrt{\left(1-r^2\right)/(n-2)}}$$

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where the second formula is an equivalent form of the test statistic, n is the sample size and the degrees of freedom are $n-2$. This is a t-statistic and operates in the same way as other t tests. Calculate the t-value and compare that with the critical value from the t-table at the appropriate degrees of freedom and the level of confidence you wish to maintain. If the calculated value is in the tail then cannot accept the null hypothesis that there is no linear relationship between these two independent random variables. If the calculated t-value is NOT in the tailed then cannot reject the null hypothesis that there is no linear relationship between the two variables.

A quick shorthand way to test correlations is the relationship between the sample size and the correlation. If:

$$|r| \geq \frac{2}{\sqrt{n}}$$

then this implies that the correlation between the two variables demonstrates that a linear relationship exists and is statistically significant at approximately the 0.05 level of significance. As the formula indicates, there is an inverse relationship between the sample size and the required correlation for significance of a linear relationship. With only 10 observations, the required correlation for significance is 0.6325, for 30 observations the required correlation for significance decreases to 0.3651 and at 100 observations the required level is only 0.2000.

Correlations may be helpful in visualizing the data, but are not appropriately used to "explain" a relationship between two variables. Perhaps no single statistic is more misused than the correlation coefficient. Citing correlations between health conditions and everything from place of residence to eye color have the effect of implying a cause and effect relationship. This simply cannot be accomplished with a correlation coefficient. The correlation coefficient is, of course, innocent of this misinterpretation. It is the duty of the analyst to use a statistic that is designed to test for cause and effect relationships and report only those results if they are intending to make such a claim. The problem is that passing this more rigorous test is difficult so lazy and/or unscrupulous "researchers" fall back on correlations when they cannot make their case legitimately.

- 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_1 and X_2 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_1 and X_2 in the population.

  Drawing a Conclusion There are two methods of making the decision concerning the hypothesis. The test statistic to test this hypothesis is:

  tc=r(1−r2)(n−2)/−−−−−−−−−√

- **13.2: Testing the Significance of the Correlation Coefficient** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

8.4: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + bx$$

where a and b are constant numbers.

The variable **x is the independent variable, and y is the dependent variable.** Another way to think about this equation is a statement of cause and effect. The X variable is the cause and the Y variable is the hypothesized effect. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

? Example 8.4.1

The following examples are linear equations.

$$(y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation

? Example 8.4.2

Graph the equation $y = -1 + 2x$.

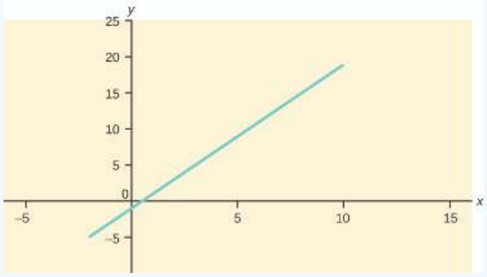


Figure 8.4.3

? Try It 8.4.1

Is the following an example of a linear equation? Why or why not?

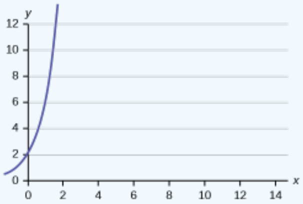


Figure 8.4.4

? Exercise 8.4.1

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Answer

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis. From calculus the slope is the first derivative of the function. For a linear function the slope is $dy/dx = b$ where we can read the mathematical expression as "the change in y (dy) that results from a change in $x(dx) = b * dx$".

Three possible graphs of the equation y = a + bx. For the first graph, (a), b
0 and so the line slopes upward to the right. For the second, b = 0 and the graph of the equation is a horizontal line. In the third graph, (c), b < 0 and the line slopes downward to the right."
src="/@api/deki/files/33886/7096285f5e75a2c46961f54cfccefea4e79baaef_d2gb">

Figure 8.4.1: Three possible graphs of *y* = *a* + *bx*. (a) If *b* > 0, the line slopes upward to the right. (b) If *b* = 0, the line is horizontal. (c) If *b* < 0, the line slopes downward to the right.

? Exercise 8.4.2

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the *y*-intercept and what is the slope? Interpret them using complete sentences.

Answer

Solution 13.4

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is $25 (a = 25)$. At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is $15 (b = 15)$. For each session, Svetlana earns $15 for each hour she tutors.

8.5: The Regression Equation

Regression analysis is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further, regression analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. This last feature, of course, is all important in predicting future values.

Regression analysis is based upon a functional relationship among variables and further, assumes that the relationship is linear. This linearity assumption is required because, for the most part, the theoretical statistical properties of non-linear estimation are not well worked out yet by the mathematicians and econometricians. This presents us with some difficulties in economic analysis because many of our theoretical models are nonlinear. The marginal cost curve, for example, is decidedly nonlinear as is the total cost function, if we are to believe in the effect of specialization of labor and the Law of Diminishing Marginal Product. There are techniques for overcoming some of these difficulties, exponential and logarithmic transformation of the data for example, but at the outset we must recognize that standard ordinary least squares (OLS) regression analysis will always use a linear function to estimate what might be a nonlinear relationship.

The general linear regression model can be stated by the equation:

$$y_i=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki}+\varepsilon_i$$

where β_0 is the intercept, β_i's are the slope between Y and the appropriate X_i, and ϵ (pronounced epsilon), is the error term that captures errors in measurement of Y and the effect on Y of any variables missing from the equation that would contribute to explaining variations in Y. This equation is the theoretical population equation and therefore uses Greek letters. The equation we will estimate will have the Roman equivalent symbols. This is parallel to how we kept track of the population parameters and sample parameters before. The symbol for the population mean was μ and for the sample mean $\overline{X}$ and for the population standard deviation was σ and for the sample standard deviation was s. The equation that will be estimated with a sample of data for two independent variables will thus be:

$$y_i=b_0+b_1 x_{1i}+b_2 x_{2i}+e_i$$

As with our earlier work with probability distributions, this model works only if certain assumptions hold. These are that the Y is normally distributed, the errors are also normally distributed with a mean of zero and a constant standard deviation, and that the error terms are independent of the size of X and independent of each other.

Assumptions of the Ordinary Least Squares Regression Model

Each of these assumptions needs a bit more explanation. If one of these assumptions fails to be true, then it will have an effect on the quality of the estimates. Some of the failures of these assumptions can be fixed while others result in estimates that quite simply provide no insight into the questions the model is trying to answer or worse, give biased estimates.

1. The independent variables, x_i, are all measured without error, and are fixed numbers that are independent of the error term. This assumption is saying in effect that Y is deterministic, the result of a fixed component “X” and a random error component “ϵ.”
2. The error term is a random variable with a mean of zero and a constant variance. The meaning of this is that the variances of the independent variables are independent of the value of the variable. Consider the relationship between personal income and the quantity of a good purchased as an example of a case where the variance is dependent upon the value of the independent variable, income. It is plausible that as income increases the variation around the amount purchased will also increase simply because of the flexibility provided with higher levels of income. The assumption is for constant variance with respect to the magnitude of the independent variable called homoscedasticity. If the assumption fails, then it is called heteroscedasticity. Figure 13.6 shows the case of homoscedasticity where all three distributions have the same variance around the predicted value of Y regardless of the magnitude of X.
3. While the independent variables are all fixed values they are from a probability distribution that is normally distributed. This can be seen in Figure 13.6 by the shape of the distributions placed on the predicted line at the expected value of the relevant value of Y.
4. The independent variables are independent of Y, but are also assumed to be independent of the other X variables. The model is designed to estimate the effects of independent variables on some dependent variable in accordance with a proposed theory. The case where some or more of the independent variables are correlated is not unusual. There may be no cause and effect relationship among the independent variables, but nevertheless they move together. Take the case of a simple supply curve

where quantity supplied is theoretically related to the price of the product and the prices of inputs. There may be multiple inputs that may over time move together from general inflationary pressure. The input prices will therefore violate this assumption of regression analysis. This condition is called multicollinearity, which will be taken up in detail later.

5. The error terms are uncorrelated with each other. This situation arises from an effect on one error term from another error term. While not exclusively a time series problem, it is here that we most often see this case. An X variable in time period one has an effect on the Y variable, but this effect then has an effect in the next time period. This effect gives rise to a relationship among the error terms. This case is called autocorrelation, “self-correlated.” The error terms are now not independent of each other, but rather have their own effect on subsequent error terms.

Figure $8.5.1$ shows the case where the assumptions of the regression model are being satisfied. The estimated line is $\hat{y}=a+b\mathrm{x}$. Three values of X are shown. A normal distribution is placed at each point where X equals the estimated line and the associated error at each value of Y. Notice that the three distributions are normally distributed around the point on the line, and further, the variation, variance, around the predicted value is constant indicating homoscedasticity from assumption 2.Figure $8.5.1$ does not show all the assumptions of the regression model, but it helps visualize these important ones.

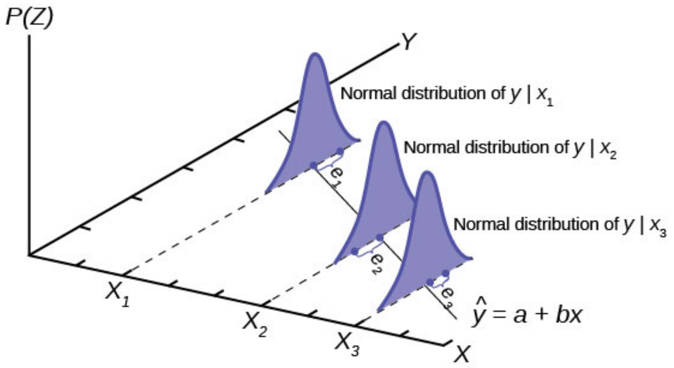
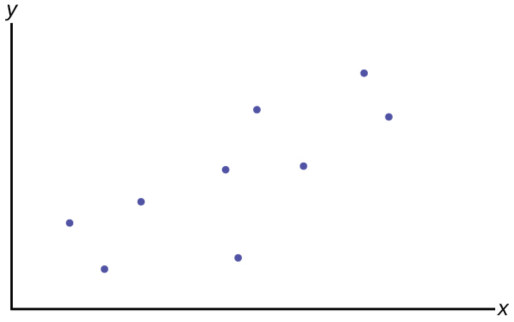

P(Z)
Y
Normal distribution of $y \mid x_1$
Normal distribution of $y \mid x_2$
Normal distribution of $y \mid x_3$
e_1
e_2
e_3
$\hat{y} = a + bx$
X_1
X_2
X_3
X

Figure 8.5.1



$$y = \beta_0 + \beta_1 X + \varepsilon$$

Figure 8.5.2

This is the general form that is most often called the multiple regression model. So-called "simple" regression analysis has only one independent (right-hand) variable rather than many independent variables. Simple regression is just a special case of multiple regression. There is some value in beginning with simple regression: it is easy to graph in two dimensions, difficult to graph in three dimensions, and impossible to graph in more than three dimensions. Consequently, our graphs will be for the simple regression case. Figure $8.5.2$ presents the regression problem in the form of a scatter plot graph of the data set where it is hypothesized that Y is dependent upon the single independent variable X.

A basic relationship from Macroeconomic Principles is the consumption function. This theoretical relationship states that as a person's income rises, their consumption rises, but by a smaller amount than the rise in income. If Y is consumption and X is income in the equation below Figure 13.7, the regression problem is, first, to establish that this relationship exists, and second, to determine the impact of a change in income on a person's consumption. The parameter β_1 was called the Marginal Propensity to Consume in Macroeconomics Principles.

Each "dot" in Figure $8.5.2$ represents the consumption and income of different individuals at some point in time. This was called cross-section data earlier; observations on variables at one point in time across different people or other units of measurement. This analysis is often done with time series data, which would be the consumption and income of one individual or country at different points in time. For macroeconomic problems it is common to use times series aggregated data for a whole country. For this particular theoretical concept these data are readily available in the annual report of the President’s Council of Economic Advisors.

The regression problem comes down to determining which straight line would best represent the data in Figure $8.5.3$. Regression analysis is sometimes called "least squares" analysis because the method of determining which line best "fits" the data is to minimize the sum of the squared residuals of a line put through the data.

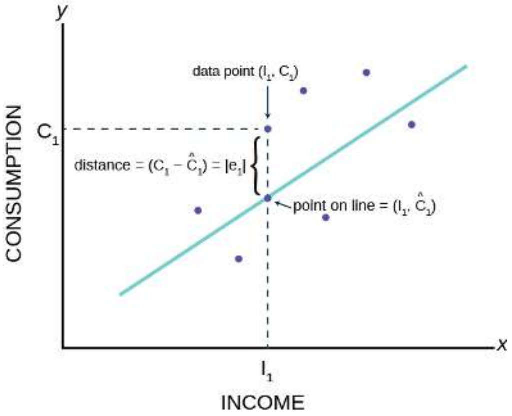


Figure $8.5.3$: Population Equation: $\mathrm{C}=\beta_0+\beta_1 \text{Income}+\varepsilon$ Estimated Equation: $C=b_0+b_1 \text{Income}+e$

This figure shows the assumed relationship between consumption and income from macroeconomic theory. Here the data are plotted as a scatter plot and an estimated straight line has been drawn. From this graph we can see an error term, e_1. Each data point also has an error term. Again, the error term is put into the equation to capture effects on consumption that are not caused by income changes. Such other effects might be a person’s savings or wealth, or periods of unemployment. We will see how by minimizing the sum of these errors we can get an estimate for the slope and intercept of this line.

Consider the graph below. The notation has returned to that for the more general model rather than the specific case of the Macroeconomic consumption function in our example.

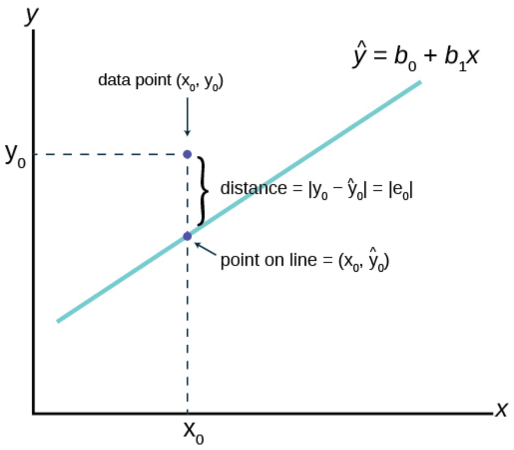

y
$\hat{y} = b_0 + b_1x$
data point (x_0, y_0)
y_0
distance = |y_0 − $\hat{y}_0$| = |e_0|
point on line = (x_0, $\hat{y}_0$)
x
x_0

Figure $8.5.4$:

The $\hat{\mathrm{y}}$ is read **"y hat"** and is the **estimated value of $\mathbf{y}$**. (In Figure 13.8 $\hat{C}$ represents the estimated value of consumption because it is on the estimated line.) It is the value of y obtained using the regression line. $\hat{\mathrm{y}}$ is not generally equal to y from the data.

The term $y_0-\hat{y}_0=e_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The error term was put into the estimating equation to capture missing variables and errors in measurement that may have occurred in the dependent variables. 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 as can be seen on the graph at point X_0.

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 graph, $y_0 - \hat{y}_0 = e_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive. For each data point the residuals, or errors, are calculated $y_i - \hat{y}_i = e_i$ for $i = 1, 2, 3, \ldots, n$ where n is the sample size. Each $|e|$ is a vertical distance.

The sum of the errors squared is the term obviously called **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the straight line that has the parameter values of b_0 and b_1 that minimizes the **SSE**. 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:

$$\hat{y} = b_0 + b_1 x$$

where $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\operatorname{cov}(x,y)}{s_x^2}$

The sample means of the x values and the y values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point ($\bar{y}$, $\bar{x}$) called the points of means.

The slope b can also be written as:

$$b_1 = r_{\mathrm{y,x}}\left(\frac{s_y}{s_x}\right)$$

where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values and r is the correlation coefficient between x and y.

These equations are called the Normal Equations and come from another very important mathematical finding called the Gauss-Markov Theorem without which we could not do regression analysis. The Gauss-Markov Theorem tells us that the estimates we get from using the ordinary least squares (OLS) regression method will result in estimates that have some very important properties. In the Gauss-Markov Theorem it was proved that a least squares line is BLUE, which is, **B**est, **L**inear, **U**nbiased, **E**stimator. Best is the statistical property that an estimator is the one with the minimum variance. Linear refers to the property of the type of line being estimated. An unbiased estimator is one whose estimating function has an expected mean equal to the mean of the population. (You will remember that the expected value of $\mu_{\bar{x}}$ was equal to the population mean μ in accordance with the Central Limit Theorem. This is exactly the same concept here).

Both Gauss and Markov were giants in the field of mathematics, and Gauss in physics too, in the 18th century and early 19th century. They barely overlapped chronologically and never in geography, but Markov's work on this theorem was based extensively on the earlier work of Carl Gauss. The extensive applied value of this theorem had to wait until the middle of this last century.

Using the OLS method we can now find the **estimate of the error variance** which is the variance of the squared errors, e^2. This is sometimes called the **standard error of the estimate**. (Grammatically this is probably best said as the estimate of the **error's**variance) The formula for the estimate of the error variance is:

$$s_e^2 = \frac{\Sigma\left(y_i - \hat{y}_i\right)^2}{n-k} = \frac{\Sigma e_i^2}{n-k}$$

where $\hat{y}$ is the predicted value of y and y is the observed value, and thus the term $\left(y_i - \hat{y}_i\right)^2$ is the squared errors that are to be minimized to find the estimates of the regression line parameters. This is really just the variance of the error terms and follows our regular variance formula. One important note is that here we are dividing by $(n-k)$, which is the degrees of freedom. The degrees of freedom of a regression equation will be the number of observations, n, reduced by the number of estimated parameters, which includes the intercept as a parameter.

The variance of the errors is fundamental in testing hypotheses for a regression. It tells us just how "tight" the dispersion is about the line. As we will see shortly, the greater the dispersion about the line, meaning the larger the variance of the errors, the less probable that the hypothesized independent variable will be found to have a significant effect on the dependent variable. In short, the theory being tested will more likely fail if the variance of the error term is high. Upon reflection this should not be a surprise. As we tested hypotheses about a mean we observed that large variances reduced the calculated test statistic and thus it failed to reach the tail of the distribution. In those cases, the null hypotheses could not be rejected. If we cannot reject the null hypothesis in a regression problem, we must conclude that the hypothesized independent variable has no effect on the dependent variable.

A way to visualize this concept is to draw two scatter plots of x and y data along a predetermined line. The first will have little variance of the errors, meaning that all the data points will move close to the line. Now do the same except the data points will have a large estimate of the error variance, meaning that the data points are scattered widely along the line. Clearly the confidence about a relationship between x and y is effected by this difference between the estimate of the error variance.

Testing the Parameters of the Line

The whole goal of the regression analysis was to test the hypothesis that the dependent variable, Y, was in fact dependent upon the values of the independent variables as asserted by some foundation theory, such as the consumption function example. Looking at the estimated equation under Figure 13.8, we see that this amounts to determining the values of b_0 and b_1. Notice that again we are using the convention of Greek letters for the population parameters and Roman letters for their estimates.

The regression analysis output provided by the computer software will produce an estimate of b_0 and b_1, and any other b's for other independent variables that were included in the estimated equation. The issue is how good are these estimates? In order to test a hypothesis concerning any estimate, we have found that we need to know the underlying sampling distribution. It should come as no surprise at his stage in the course that the answer is going to be the normal distribution. This can be seen by remembering the assumption that the error term in the population, ϵ, is normally distributed. If the error term is normally distributed and the variance of the estimates of the equation parameters, b_0 and b_1, are determined by the variance of the error term, it follows that the variances of the parameter estimates are also normally distributed. And indeed this is just the case.

We can see this by the creation of the test statistic for the test of hypothesis for the slope parameter, β_1 in our consumption function equation. To test whether or not Y does indeed depend upon X, or in our example, that consumption depends upon income, we need only test the hypothesis that β_1 equals zero. This hypothesis would be stated formally as:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

If we cannot reject the null hypothesis, we must conclude that our theory has no validity. If we cannot reject the null hypothesis that $\beta_1 = 0$ then b_1, the coefficient of Income, is zero and zero times anything is zero. Therefore the effect of Income on Consumption is zero. There is no relationship as our theory had suggested.

Notice that we have set up the presumption, the null hypothesis, as "no relationship". This puts the burden of proof on the alternative hypothesis. In other words, if we are to validate our claim of finding a relationship, we must do so with a level of significance greater than 90, 95, or 99 percent. The status quo is ignorance, no relationship exists, and to be able to make the claim that we have actually added to our body of knowledge we must do so with significant probability of being correct. John Maynard Keynes got it right and thus was born Keynesian economics starting with this basic concept in 1936.

The test statistic for this test comes directly from our old friend the standardizing formula:

$$t_c = \frac{b_1 - \beta_1}{S_{b_1}}$$

where b_1 is the estimated value of the slope of the regression line, β_1 is the hypothesized value of beta, in this case zero, and S_{b_1} is the standard deviation of the estimate of b_1. In this case we are asking how many standard deviations is the estimated slope away from the hypothesized slope. This is exactly the same question we asked before with respect to a hypothesis about a mean: how many standard deviations is the estimated mean, the sample mean, from the hypothesized mean?

The test statistic is written as a student's t distribution, but if the sample size is larger enough so that the degrees of freedom are greater than 30 we may again use the normal distribution. To see why we can use the student's t or normal distribution we have only to look at S_{b_1},the formula for the standard deviation of the estimate of b_1:

$$S_{b_1} = \frac{S_e^2}{\sqrt{\left(x_i - \bar{x}\right)^2}}$$

or

$$S_{b_1} = \frac{S_e^2}{(n-1)S_x^2} \tag{8.5.1}$$

Where S_e is the estimate of the error variance and S_x^2 is the variance of x values of the coefficient of the independent variable being tested.

We see that S_e, the **estimate of the error variance**, is part of the computation. Because the estimate of the error variance is based on the assumption of normality of the error terms, we can conclude that the sampling distribution of the b's, the coefficients of our hypothesized regression line, are also normally distributed.

One last note concerns the degrees of freedom of the test statistic, $\nu = n-k$. Previously we subtracted 1 from the sample size to determine the degrees of freedom in a student's t problem. Here we must subtract one degree of freedom for each parameter estimated in the equation. For the example of the consumption function we lose 2 degrees of freedom, one for b_0, the intercept, and one for b_1, the slope of the consumption function. The degrees of freedom would be $n-k-1$, where k is the number of independent variables and the extra one is lost because of the intercept. If we were estimating an equation with three independent variables, we would lose 4 degrees of freedom: three for the independent variables, k, and one more for the intercept.

The decision rule for acceptance or rejection of the null hypothesis follows exactly the same form as in all our previous test of hypothesis. Namely, if the calculated value of t (or Z) falls into the tails of the distribution, where the tails are defined by α, the required significance level in the test, we cannot accept the null hypothesis. If on the other hand, the calculated value of the test statistic is within the critical region, we cannot reject the null hypothesis.

If we conclude that we cannot accept the null hypothesis, we are able to state with $(1-\alpha)$ level of confidence that the slope of the line is given by b_1. This is an extremely important conclusion. Regression analysis not only allows us to test if a cause and effect relationship exists, we can also determine the magnitude of that relationship, if one is found to exist. It is this feature of regression analysis that makes it so valuable. If models can be developed that have statistical validity, we are then able to simulate the effects of changes in variables that may be under our control with some degree of probability , of course. For example, if advertising is demonstrated to effect sales, we can determine the effects of changing the advertising budget and decide if the increased sales are worth the added expense.

Multicollinearity

Our discussion earlier indicated that like all statistical models, the OLS regression model has important assumptions attached. Each assumption, if violated, has an effect on the ability of the model to provide useful and meaningful estimates. The Gauss-Markov Theorem has assured us that the OLS estimates are unbiased and minimum variance, but this is true only under the assumptions of the model. Here we will look at the effects on OLS estimates if the independent variables are correlated. The other assumptions and the methods to mitigate the difficulties they pose if they are found to be violated are examined in Econometrics courses. We take up multicollinearity because it is so often prevalent in Economic models and it often leads to frustrating results.

The OLS model assumes that all the independent variables are independent of each other. This assumption is easy to test for a particular sample of data with simple correlation coefficients. Correlation, like much in statistics, is a matter of degree: a little is not good, and a lot is terrible.

The goal of the regression technique is to tease out the independent impacts of each of a set of independent variables on some hypothesized dependent variable. If two 2 independent variables are interrelated, that is, correlated, then we cannot isolate the effects on Y of one from the other. In an extreme case where x_1 is a linear combination of x_2, correlation equal to one, both variables move in identical ways with Y. In this case it is impossible to determine the variable that is the true cause of the effect on Y. (If the two variables were actually perfectly correlated, then mathematically no regression results could actually be calculated.)

The normal equations for the coefficients show the effects of multicollinearity on the coefficients.

$$b_1 = \frac{s_y\left(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}\right)}{s_{x_1}\left(1 - r_{x_1 x_2}^2\right)}$$

$$b_2 = \frac{s_y\left(r_{x_{2y}} - r_{x_1 x_2} r_{x_1 y}\right)}{s_{x_2}\left(1 - r_{x_1 x_2}^2\right)}$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

The correlation between x_1 and x_2, $r^2_{x_1x_2}$, appears in the denominator of both the estimating formula for b_1 and b_2. If the assumption of independence holds, then this term is zero. This indicates that there is no effect of the correlation on the coefficient. On the other hand, as the correlation between the two independent variables increases the denominator decreases, and thus the estimate of the coefficient increases. The correlation has the same effect on both of the coefficients of these two variables. In essence, each variable is “taking” part of the effect on Y that should be attributed to the collinear variable. This results in biased estimates.

Multicollinearity has a further deleterious impact on the OLS estimates. The correlation between the two independent variables also shows up in the formulas for the estimate of the variance for the coefficients.

$$s^2_{b_1} = \frac{s^2_e}{(n-1)s^2_{x_1}\left(1-r^2_{x_1x_2}\right)}$$

$$s^2_{b_2} = \frac{s^2_e}{(n-1)s^2_{x_2}\left(1-r^2_{x_1x_2}\right)}$$

Here again we see the correlation between x_1 and x_2 in the denominator of the estimates of the variance for the coefficients for both variables. If the correlation is zero as assumed in the regression model, then the formula collapses to the familiar ratio of the variance of the errors to the variance of the relevant independent variable. If however the two independent variables are correlated, then the variance of the estimate of the coefficient increases. This results in a smaller t-value for the test of hypothesis of the coefficient. In short, multicollinearity results in failing to reject the null hypothesis that the X variable has no impact on Y when in fact X does have a statistically significant impact on Y. Said another way, the large standard errors of the estimated coefficient created by multicollinearity suggest statistical insignificance even when the hypothesized relationship is strong.

How Good is the Equation?

In the last section we concerned ourselves with testing the hypothesis that the dependent variable did indeed depend upon the hypothesized independent variable or variables. It may be that we find an independent variable that has some effect on the dependent variable, but it may not be the only one, and it may not even be the most important one. Remember that the error term was placed in the model to capture the effects of any missing independent variables. It follows that the error term may be used to give a measure of the "goodness of fit" of the equation taken as a whole in explaining the variation of the dependent variable, Y.

The **multiple correlation coefficient**, also called the **coefficient of multiple determination** or the **coefficient of determination**, is given by the formula:

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

where SSR is the regression sum of squares, the squared deviation of the predicted value of y from the mean value of $y(\hat{y}-\bar{y})$, and SST is the total sum Figure 13.10 shows how the total deviation of the dependent variable, y, is partitioned into these two pieces.

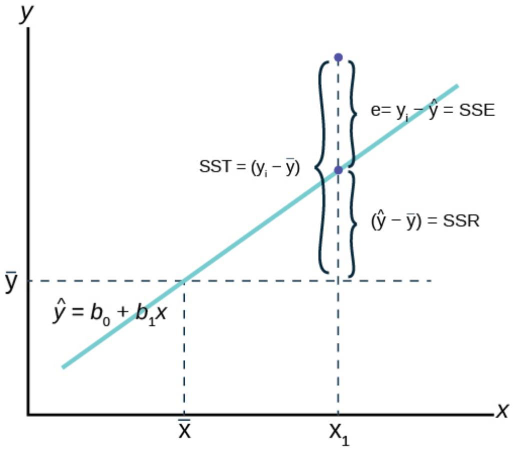


Figure 8.5.5

Figure $8.5.5$ shows the estimated regression line and a single observation, x_1. Regression analysis tries to explain the variation of the data about the mean value of the dependent variable, y. The question is, why do the observations of y vary from the average level of y? The value of y at observation x_1 varies from the mean of y by the difference $(y_i - \bar{y})$. The sum of these differences squared is SST, the sum of squares total. The actual value of y at x_1 deviates from the estimated value, $\hat{y}$, by the difference between the estimated value and the actual value, $(y_i - \hat{y})$. We recall that this is the error term, e, and the sum of these errors is SSE, sum of squared errors. The deviation of the predicted value of y, $\hat{y}$, from the mean value of y is $(\hat{y} - \bar{y})$ and is the SSR, sum of squares regression. It is called "regression" because it is the deviation explained by the regression. (Sometimes the SSR is called SSM for sum of squares mean because it measures the deviation from the mean value of the dependent variable, y, as shown on the graph.).

Because the SST = SSR + SSE we see that the multiple correlation coefficient is the percent of the variance, or deviation in y from its mean value, that is explained by the equation when taken as a whole. R^2 will vary between zero and 1, with zero indicating that none of the variation in y was explained by the equation and a value of 1 indicating that 100% of the variation in y was explained by the equation. For time series studies expect a high R^2 and for cross-section data expect low R^2.

While a high R^2 is desirable, remember that it is the tests of the hypothesis concerning the existence of a relationship between a set of independent variables and a particular dependent variable that was the motivating factor in using the regression model. It is validating a cause and effect relationship developed by some theory that is the true reason that we chose the regression analysis. Increasing the number of independent variables will have the effect of increasing R^2. To account for this effect the proper measure of the coefficient of determination is the $\overline{R}^2$, adjusted for degrees of freedom, to keep down mindless addition of independent variables.

There is no statistical test for the R^2 and thus little can be said about the model using R^2 with our characteristic confidence level. Two models that have the same size of SSE, that is sum of squared errors, may have very different R^2 if the competing models have different SST, total sum of squared deviations. The goodness of fit of the two models is the same; they both have the same sum of squares unexplained, errors squared, but because of the larger total sum of squares on one of the models the R^2 differs. Again, the real value of regression as a tool is to examine hypotheses developed from a model that predicts certain relationships among the variables. These are tests of hypotheses on the coefficients of the model and not a game of maximizing R^2.

Another way to test the general quality of the overall model is to test the coefficients as a group rather than independently. Because this is multiple regression (more than one X), we use the F-test to determine if our coefficients collectively affect Y. The hypothesis is:

$H_o : \beta_1 = \beta_2 = \ldots = \beta_i = 0$

H_a: "at least one of the β_i is not equal to 0"

If the null hypothesis cannot be rejected, then we conclude that none of the independent variables contribute to explaining the variation in Y. Reviewing Figure 13.10 we see that SSR, the explained sum of squares, is a measure of just how much of the variation in Y is explained by all the variables in the model. SSE, the sum of the errors squared, measures just how much is unexplained. It follows that the ratio of these two can provide us with a statistical test of the model as a whole. Remembering that the F distribution is a ratio of Chi squared distributions and that variances are distributed according to Chi Squared, and the sum of squared errors and the sum of squares are both variances, we have the test statistic for this hypothesis as:

$$F_c = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)}$$

where n is the number of observations and k is the number of independent variables. It can be shown that this is equivalent to:

$$F_c = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$$

Figure 13.10 where R^2 is the coefficient of determination which is also a measure of the "goodness" of the model.

As with all our tests of hypothesis, we reach a conclusion by comparing the calculated F statistic with the critical value given our desired level of confidence. If the calculated test statistic, an F statistic in this case, is in the tail of the distribution, then we cannot accept the null hypothesis. By not being able to accept the null hypotheses we conclude that this specification of this model has validity, because at least one of the estimated coefficients is significantly different from zero.

An alternative way to reach this conclusion is to use the p-value comparison rule. The p-value is the area in the tail, given the calculated F statistic. In essence, the computer is finding the F value in the table for us. The computer regression output for the calculated F statistic is typically found in the ANOVA table section labeled “significance F". How to read the output of an Excel regression is presented below. This is the probability of NOT accepting a false null hypothesis. If this probability is less than our pre-determined alpha error, then the conclusion is that we cannot accept the null hypothesis.

Dummy Variables

Thus far the analysis of the OLS regression technique assumed that the independent variables in the models tested were continuous random variables. There are, however, no restrictions in the regression model against independent variables that are binary. This opens the regression model for testing hypotheses concerning categorical variables such as gender, race, region of the country, before a certain data, after a certain date and innumerable others. These categorical variables take on only two values, 1 and 0, success or failure, from the binomial probability distribution. The form of the equation becomes:

$$\hat{y}=b_0+b_2x_2+b_1x_1$$

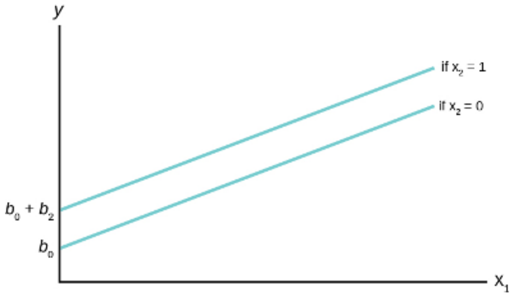


Figure 8.5.6

where $x_2=0$. X_2 is the dummy variable and X_1 is some continuous random variable. The constant, b_0, is the y-intercept, the value where the line crosses the y-axis. When the value of $X_2=0$, the estimated line crosses at b_0. When the value of $X_2=1$ then the estimated line crosses at b_0+b_2 . In effect the dummy variable causes the estimated line to shift either up or down by the size of the effect of the characteristic captured by the dummy variable. Note that this is a simple parallel shift and does not affect the impact of the other independent variable; X_1.This variable is a continuous random variable and predicts different values of y at different values of X_1 holding constant the condition of the dummy variable.

An example of the use of a dummy variable is the work estimating the impact of gender on salaries. There is a full body of literature on this topic and dummy variables are used extensively. For this example the salaries of elementary and secondary school teachers for a particular state is examined. Using a homogeneous job category, school teachers, and for a single state reduces many of the variations that naturally effect salaries such as differential physical risk, cost of living in a particular state, and other working conditions. The estimating equation in its simplest form specifies salary as a function of various teacher characteristic that economic theory would suggest could affect salary. These would include education level as a measure of potential productivity, age and/or experience to capture on-the-job training, again as a measure of productivity. Because the data are for school teachers employed in a public school districts rather than workers in a for-profit company, the school district’s average revenue per average daily student attendance is included as a measure of ability to pay. The results of the regression analysis using data on 24,916 school teachers are presented below.

Table 8.5.1: Earnings Estimate for Elementary and Secondary School Teachers

Variable	Regression Coefficients (b)	Standard Errors of the estimates for teacher's earnings function (sb)
Intercept	4269.9	
Gender (male = 1)	632.38	13.39
Total Years of Experience	52.32	1.10
Years of Experience in Current District	29.97	1.52

Variable	Regression Coefficients (b)	Standard Errors of the estimates for teacher's earnings function (sb)
Education	629.33	13.16
Total Revenue per ADA	90.24	3.76
$\overline{R}^2$	.725	
n	24,916	

The coefficients for all the independent variables are significantly different from zero as indicated by the standard errors. Dividing the standard errors of each coefficient results in a t-value greater than 1.96 which is the required level for 95% significance. The binary variable, our dummy variable of interest in this analysis, is gender where male is given a value of 1 and female given a value of 0. The coefficient is significantly different from zero with a dramatic t-statistic of 47 standard deviations. We thus cannot accept the null hypothesis that the coefficient is equal to zero. Therefore we conclude that there is a premium paid male teachers of $632 after holding constant experience, education and the wealth of the school district in which the teacher is employed. It is important to note that these data are from some time ago and the $632 represents a six percent salary premium at that time. A graph of this example of dummy variables is presented below.

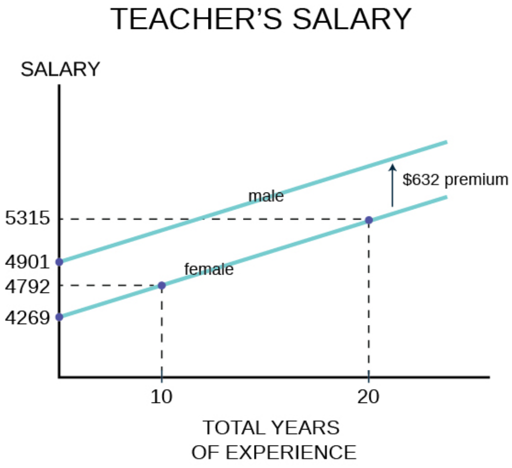
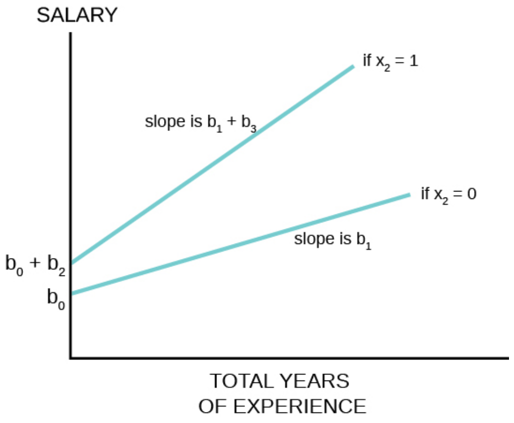


Figure 8.5.7

In two dimensions, salary is the dependent variable on the vertical axis and total years of experience was chosen for the continuous independent variable on horizontal axis. Any of the other independent variables could have been chosen to illustrate the effect of the dummy variable. The relationship between total years of experience has a slope of $52.32 per year of experience and the estimated line has an intercept of $4,269 if the gender variable is equal to zero, for female. If the gender variable is equal to 1, for male, the coefficient for the gender variable is added to the intercept and thus the relationship between total years of experience and salary is shifted upward parallel as indicated on the graph. Also marked on the graph are various points for reference. A female school teacher with 10 years of experience receives a salary of $4,792 on the basis of her experience only, but this is still $109 less than a male teacher with zero years of experience.

A more complex interaction between a dummy variable and the dependent variable can also be estimated. It may be that the dummy variable has more than a simple shift effect on the dependent variable, but also interacts with one or more of the other continuous independent variables. While not tested in the example above, it could be hypothesized that the impact of gender on salary was not a one-time shift, but impacted the value of additional years of experience on salary also. That is, female school teacher's salaries were discounted at the start, and further did not grow at the same rate from the effect of experience as for male school teachers. This would show up as a different slope for the relationship between total years of experience for males than for females. If this is so then females school teachers would not just start behind their male colleagues (as measured by the shift in the estimated regression line), but would fall further and further behind as time and experienced increased.

The graph below shows how this hypothesis can be tested with the use of dummy variables and an interaction variable.



$$\hat{y} = b_0 + b_2x_2 + b_1x_1 + b_3x_2x_1$$

Figure 8.5.8

The estimating equation shows how the slope of X_1, the continuous random variable experience, contains two parts, b_1 and b_3. This occurs because of the new variable X_2 X_1, called the interaction variable, was created to allow for an effect on the slope of X_1 from changes in X_2, the binary dummy variable. Note that when the dummy variable, $X_2 = 0$ the interaction variable has a value of 0, but when $X_2 = 1$ the interaction variable has a value of X_1. The coefficient b_3 is an estimate of the difference in the coefficient of X_1 when $X_2 = 1$ compared to when $X_2 = 0$. In the example of teacher's salaries, if there is a premium paid to male teachers that affects the rate of increase in salaries from experience, then the rate at which male teachers' salaries rises would be $b_1 + b_3$ and the rate at which female teachers' salaries rise would be simply b_1. This hypothesis can be tested with the hypothesis:

$$H_0 : \beta_3 = 0 | \beta_1 = 0, \beta_2 = 0$$

$$H_a : \beta_3 \neq 0 | \beta_1 \neq 0, \beta_2 \neq 0$$

This is a t-test using the test statistic for the parameter β_3. If we cannot accept the null hypothesis that $\beta_3 = 0$ we conclude there is a difference between the rate of increase for the group for whom the value of the binary variable is set to 1, males in this example. This estimating equation can be combined with our earlier one Figure 8.5.8 are drawn for this case with a shift in the earnings function and a difference in the slope of the function with respect to total years of experience.

? Example 8.5.1

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 randomly selected student if you know the third exam score?

Table showing the scores on the final exam based on scores from the third exam.

Table 13.2

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

x (third exam score)	y (final exam score)
69	151
69	159

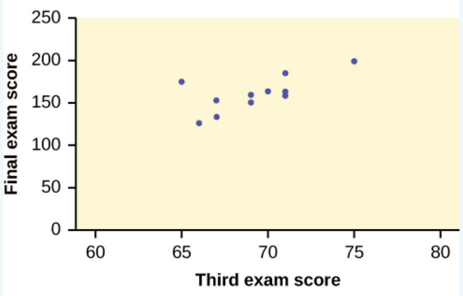


Figure 8.5.9 Scatter plot showing the scores on the final exam based on scores from the third exam.

8.6: Interpretation of Regression Coefficients- Elasticity and Logarithmic Transformation

As we have seen, the coefficient of an equation estimated using OLS regression analysis provides an estimate of the slope of a straight line that is assumed be the relationship between the dependent variable and at least one independent variable. From the calculus, the slope of the line is the first derivative and tells us the magnitude of the impact of a one unit change in the X variable upon the value of the Y variable measured in the units of the Y variable. As we saw in the case of dummy variables, this can show up as a parallel shift in the estimated line or even a change in the slope of the line through an interactive variable. Here we wish to explore the concept of elasticity and how we can use a regression analysis to estimate the various elasticities in which economists have an interest.

The concept of elasticity is borrowed from engineering and physics where it is used to measure a material's responsiveness to a force, typically a physical force such as a stretching/pulling force. It is from here that we get the term an "elastic" band. In economics, the force in question is some market force such as a change in price or income. Elasticity is measured as a percentage change/response in both engineering applications and in economics. The value of measuring in percentage terms is that the units of measurement do not play a role in the value of the measurement and thus allows direct comparison between elasticities. As an example, if the price of gasoline increased say 50 cents from an initial price of $3.00 and generated a decline in monthly consumption for a consumer from 50 gallons to 48 gallons we calculate the elasticity to be 0.25. The price elasticity is the percentage change in quantity resulting from some percentage change in price. A 16 percent increase in price has generated only a 4 percent decrease in demand: 16% price change $\rightarrow$ 4% quantity change or $.04/.16 = .25$ This is called an inelastic demand meaning a small response to the price change. This comes about because there are few if any real substitutes for gasoline; perhaps public transportation, a bicycle or walking. Technically, of course, the percentage change in demand from a price increase is a decline in demand thus price elasticity is a negative number. The common convention, however, is to talk about elasticity as the absolute value of the number. Some goods have many substitutes: pears for apples for plums, for grapes, etc. etc. The elasticity for such goods is larger than one and are called elastic in demand. Here a small percentage change in price will induce a large percentage change in quantity demanded. The consumer will easily shift the demand to the close substitute.

While this discussion has been about price changes, any of the independent variables in a demand equation will have an associated elasticity. Thus, there is an income elasticity that measures the sensitivity of demand to changes in income: not much for the demand for food, but very sensitive for yachts. If the demand equation contains a term for substitute goods, say candy bars in a demand equation for cookies, then the responsiveness of demand for cookies from changes in prices of candy bars can be measured. This is called the cross-price elasticity of demand and to an extent can be thought of as brand loyalty from a marketing view. How responsive is the demand for Coca-Cola to changes in the price of Pepsi?

Now imagine the demand for a product that is very expensive. Again, the measure of elasticity is in percentage terms thus the elasticity can be directly compared to that for gasoline: an elasticity of 0.25 for gasoline conveys the same information as an elasticity of 0.25 for $25,000 car. Both goods are considered by the consumer to have few substitutes and thus have inelastic demand curves, elasticities less than one.

The mathematical formulae for various elasticities are:

$$\text{Price elasticity: } \eta_{\mathrm{p}} = \frac{(\%\Delta \mathrm{Q})}{(\%\Delta \mathrm{P})}$$

Where η is the Greek small case letter eta used to designate elasticity. Δ is read as "change".

$$\text{Income elasticity: } \eta_{\mathrm{Y}} = \frac{(\%\Delta \mathrm{Q})}{(\%\Delta \mathrm{Y})}$$

Where Y is used as the symbol for income.

$$\text{Cross-Price elasticity: } \eta_{\mathrm{p1}} = \frac{(\%\Delta \mathrm{Q}_1)}{(\%\Delta \mathrm{P}_2)}$$

Where P2 is the price of the substitute good.

Examining closer the price elasticity we can write the formula as:

$$\eta_{\mathrm{p}}=\frac{(\%\Delta \mathrm{Q})}{(\%\Delta \mathrm{P})}=\frac{\mathrm{dQ}}{\mathrm{dP}}\left(\frac{\mathrm{P}}{\mathrm{Q}}\right)=\mathrm{b}\left(\frac{\mathrm{P}}{\mathrm{Q}}\right)$$

Where b is the estimated coefficient for price in the OLS regression.

The first form of the equation demonstrates the principle that elasticities are measured in percentage terms. Of course, the ordinary least squares coefficients provide an estimate of the impact of a unit change in the independent variable, X, on the dependent variable measured in units of Y. These coefficients are not elasticities, however, and are shown in the second way of writing the formula for elasticity as $\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)$, the derivative of the estimated demand function which is simply the slope of the regression line. Multiplying the slope times $\frac{P}{Q}$ provides an elasticity measured in percentage terms.

Along a straight-line demand curve the percentage change, thus elasticity, changes continuously as the scale changes, while the slope, the estimated regression coefficient, remains constant. Going back to the demand for gasoline. A change in price from \$3.00 to \$3.50 was a 16 percent increase in price. If the beginning price were \$5.00 then the same 50¢ increase would be only a 10 percent increase generating a different elasticity. Every straight-line demand curve has a range of elasticities starting at the top left, high prices, with large elasticity numbers, elastic demand, and decreasing as one goes down the demand curve, inelastic demand.

In order to provide a meaningful estimate of the elasticity of demand the convention is to estimate the elasticity at the point of means. Remember that all OLS regression lines will go through the point of means. At this point is the greatest weight of the data used to estimate the coefficient. The formula to estimate an elasticity when an OLS demand curve has been estimated becomes:

$$\eta_{\mathrm{p}}=\mathrm{b}\left(\frac{\overline{\mathrm{P}}}{\overline{\mathrm{Q}}}\right)$$

Where $\overline{\mathrm{P}}$ and $\overline{\mathrm{Q}}$ are the mean values of these data used to estimate b, the price coefficient.

The same method can be used to estimate the other elasticities for the demand function by using the appropriate mean values of the other variables; income and price of substitute goods for example.

Logarithmic Transformation of the Data

Ordinary least squares estimates typically assume that the population relationship among the variables is linear thus of the form presented in The Regression Equation. In this form the interpretation of the coefficients is as discussed above; quite simply the coefficient provides an estimate of the impact of a one **unit** change in X on Y measured in **units** of Y. It does not matter just where along the line one wishes to make the measurement because it is a straight line with a constant slope thus constant estimated level of impact per unit change. It may be, however, that the analyst wishes to estimate not the simple unit measured impact on the Y variable, but the magnitude of the percentage impact on Y of a one unit change in the X variable. Such a case might be how a **unit change** in experience, say one year, effects not the absolute amount of a worker's wage, but the **percentage impact** on the worker's wage. Alternatively, it may be that the question asked is the unit measured impact on Y of a specific percentage increase in X. An example may be "by how many dollars will sales increase if the firm spends X percent more on advertising?" The third possibility is the case of elasticity discussed above. Here we are interested in the percentage impact on quantity demanded for a given percentage change in price, or income or perhaps the price of a substitute good. All three of these cases can be estimated by transforming the data to logarithms before running the regression. The resulting coefficients will then provide a percentage change measurement of the relevant variable.

To summarize, there are four cases:

1. $\text{Unit } \Delta X \rightarrow \text{ Unit } \Delta Y$ (Standard OLS case)
2. $\text{Unit } \Delta X \rightarrow \% \Delta Y$
3. $\% \Delta X \rightarrow \text{ Unit } \Delta Y$
4. $\% \Delta X \rightarrow \% \Delta Y$ (elasticity case)

Case 1: The ordinary least squares case begins with the linear model developed above:

$$Y=a+bX$$

where the coefficient of the independent variable $b=\frac{dY}{dX}$ is the slope of a straight line and thus measures the impact of a unit change in X on Y measured in units of Y.

Case 2: The underlying estimated equation is:

$$\log(\mathrm{Y}) = a + bX$$

The equation is estimated by converting the Y values to logarithms and using OLS techniques to estimate the coefficient of the X variable, b. This is called a semi-log estimation. Again, differentiating both sides of the equation allows us to develop the interpretation of the X coefficient b:

$$\mathrm{d}\left(\log_{\mathrm{Y}}\right) = b\mathrm{d}X$$

$$\frac{\mathrm{d}Y}{Y} = b\mathrm{d}X$$

Multiply by 100 to covert to percentages and rearranging terms gives:

$$100b = \frac{\%\Delta Y}{\text{Unit } \Delta X}$$

$100b$ is thus the percentage change in Y resulting from a unit change in X.

Case 3: In this case the question is “what is the unit change in Y resulting from a percentage change in X?” What is the dollar loss in revenues of a five percent increase in price or what is the total dollar cost impact of a five percent increase in labor costs? The estimated equation for this case would be:

$$Y = a + B\log(X)$$

Here the calculus differential of the estimated equation is:

$$dY = bd(logX)$$

$$\mathrm{d}Y = b\frac{\mathrm{d}X}{X}$$

Divide by 100 to get percentage and rearranging terms gives:

$$\frac{b}{100} = \frac{\mathrm{d}Y}{100\frac{\mathrm{d}X}{X}} = \frac{\text{Unit } \Delta \mathrm{Y}}{\%\Delta \mathrm{X}}$$

Therefore, $\frac{b}{100}$ is the increase in Y measured in units from a one percent increase in X.

Case 4: This is the elasticity case where both the dependent and independent variables are converted to logs before the OLS estimation. This is known as the log-log case or double log case, and provides us with direct estimates of the elasticities of the independent variables. The estimated equation is:

$$logY = a + blogX$$

Differentiating we have:

$$d(logY) = bd(logX)$$

$$\mathrm{d}(\log X) = b\frac{1}{X}\mathrm{d}X$$

thus:

$$\frac{1}{Y}\mathrm{d}Y = b\frac{1}{X}\mathrm{d}X \quad \text{OR} \quad \frac{\mathrm{d}Y}{Y} = b\frac{\mathrm{d}X}{X} \quad \text{OR} \quad b = \frac{\mathrm{d}Y}{\mathrm{d}X}\left(\frac{X}{Y}\right)$$

and $b = \frac{\%\Delta Y}{\%\Delta X}$ our definition of elasticity. We conclude that we can directly estimate the elasticity of a variable through double log transformation of the data. The estimated coefficient is the elasticity. It is common to use double log transformation of all variables in the estimation of demand functions to get estimates of all the various elasticities of the demand curve.

- **13.5: Interpretation of Regression Coefficients- Elasticity and Logarithmic Transformation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

8.7: Predicting with a Regression Equation

One important value of an estimated regression equation is its ability to predict the effects on Y of a change in one or more values of the independent variables. The value of this is obvious. Careful policy cannot be made without estimates of the effects that may result. Indeed, it is the desire for particular results that drive the formation of most policy. Regression models can be, and have been, invaluable aids in forming such policies.

The Gauss-Markov theorem assures us that the point estimate of the impact on the dependent variable derived by putting in the equation the hypothetical values of the independent variables one wishes to simulate will result in an estimate of the dependent variable which is minimum variance and unbiased. That is to say that from this equation comes the best unbiased point estimate of y given the values of x.

$$\hat{y}=b_0+b, X_{1i}+\cdots+b_k X_{ki}$$

Remember that point estimates do not carry a particular level of probability, or level of confidence, because points have no "width" above which there is an area to measure. This was why we developed confidence intervals for the mean and proportion earlier. The same concern arises here also. There are actually two different approaches to the issue of developing estimates of changes in the independent variable, or variables, on the dependent variable. The first approach wishes to measure the **expected mean** value of y from a specific change in the value of x: this specific value implies the expected value. Here the question is: what is the **mean** impact on y that would result from multiple hypothetical experiments on y at this specific value of x. Remember that there is a variance around the estimated parameter of x and thus each experiment will result in a bit of a different estimate of the predicted value of y.

The second approach to estimate the effect of a specific value of x on y treats the event as a single experiment: you choose x and multiply it times the coefficient and that provides a single estimate of y. Because this approach acts as if there were a single experiment the variance that exists in the parameter estimate is larger than the variance associated with the expected value approach.

The conclusion is that we have two different ways to predict the effect of values of the independent variable(s) on the dependent variable and thus we have two different intervals. Both are correct answers to the question being asked, but there are two different questions. To avoid confusion, the first case where we are asking for the **expected value** of the mean of the estimated y, is called a **confidence interval** as we have named this concept before. The second case, where we are asking for the estimate of the impact on the dependent variable y of a single experiment using a value of x, is called the **prediction interval**. The test statistics for these two interval measures within which the estimated value of y will fall are:

$$\text{Confidence Interval for Expected Value of Mean Value of y for x} = \mathrm{x_p}$$

$$\hat{y}=\pm t_{\alpha/2} s_e\left(\sqrt{\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{s_x}}\right)$$

$$\text{Prediction Interval for an Individual y for } x=x_p$$

$$\hat{y}=\pm t_{\alpha/2} s_e\left(\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{s_x}}\right)$$

Where s_e is the standard deviation of the error term and s_x is the standard deviation of the x variable.

The mathematical computations of these two test statistics are complex. Various computer regression software packages provide programs within the regression functions to Figure $8.7.1$.

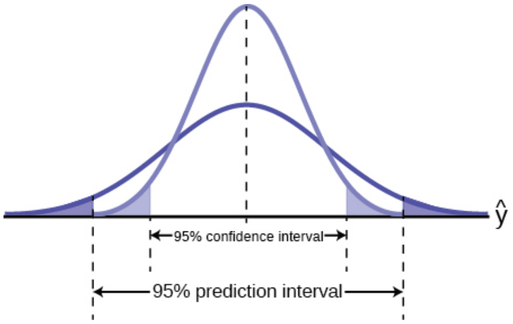


Figure 8.7.1 Prediction and confidence intervals for regression equation; 95% confidence level.

Figure 8.7.1 shows visually the difference the standard deviation makes in the size of the estimated intervals. The confidence interval, measuring the expected value of the dependent variable, is smaller than the prediction interval for the same level of confidence. The expected value method assumes that the experiment is conducted multiple times rather than just once as in the other method. The logic here is similar, although not identical, to that discussed when developing the relationship between the sample size and the confidence interval using the Central Limit Theorem. There, as the number of experiments increased, the distribution narrowed and the confidence interval became tighter around the expected value of the mean.

It is also important to note that the intervals around a point estimate are highly dependent upon the range of data used to estimate the equation regardless of which approach is being used for prediction. Remember that all regression equations go through the point of means, that is, the mean value of y and the mean values of all independent variables in the equation. As the value of x chosen to estimate the associated value of y is further from the point of means the width of the estimated interval around Figure 8.7.2 shows this relationship.

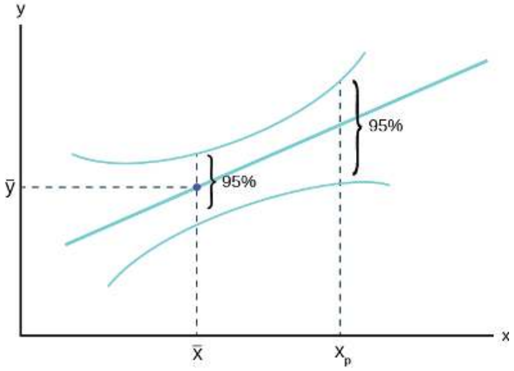


Figure 8.7.2 Confidence interval for an individual value of x, X_p, at 95% level of confidence

Figure 8.7.16 demonstrates the concern for the quality of the estimated interval whether it is a prediction interval or a confidence interval. As the value chosen to predict y, X_p in the graph, is further from the central weight of the data, $\overline{X}$, we see the interval expand in width even while holding constant the level of confidence. This shows that the precision of any estimate will diminish as one tries to predict beyond the largest weight of the data and most certainly will degrade rapidly for predictions beyond the range of the data. Unfortunately, this is just where most predictions are desired. They can be made, but the width of the confidence interval may be so large as to render the prediction useless. Only actual calculation and the particular application can determine this, however.

? Example 8.7.1

Recall the third exam/final exam example .

We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction. Assume the coefficient for X was determined to be significantly different from zero.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores **(x-values)** range from 65 to 75. Since 73 is between the x-values 65 and 75, we feel comfortable to substitute $x = 73$ into the equation. Then:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Answer

a. 145.27

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Answer

b. The x values in the data are between 65 and 75. Ninety is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will have a confidence interval that may not be meaningful.)

To understand really how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution $x = 90$ into the equation.

$\hat{y} = -173.51 + 4.83(90) = 261.19$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

8.8: Chapter Key Terms

Key Term	Definition
a is the symbol for the Y-Intercept	Sometimes written as b_0, because when writing the theoretical linear model β_0 is used to represent a coefficient for a population.
b is the symbol for Slope	The word coefficient will be used regularly for the slope, because it is a number that will always be next to the letter "x." It will be written as b_1 when a sample is used, and β_1 will be used with a population or when writing the theoretical linear model.
Bivariate	two variables are present in the model where one is the "cause" or independent variable and the other is the "effect" of dependent variable.
Linear	a model that takes data and regresses it into a straight line equation.
Multivariate	a system or model where more than one independent variable is being used to predict an outcome. There can only ever be one dependent variable, but there is no limit to the number of independent variables.
R^2 – Coefficient of Determination	This is a number between 0 and 1 that represents the percentage variation of the dependent variable that can be explained by the variation in the independent variable. Sometimes calculated by the equation $R^2 = \frac{SSR}{SST}$ where SSR is the "Sum of Squares Regression" and SST is the "Sum of Squares Total." The appropriate coefficient of determination to be reported should always be adjusted for degrees of freedom first.
Residual or "error"	the value calculated from subtracting $y_0 - \hat{y}_0 = e_0$. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y that appears on the best-fit line.
R – Correlation Coefficient	A number between -1 and 1 that represents the strength and direction of the relationship between "X" and "Y." The value for "r" will equal 1 or -1 only if all the plotted points form a perfectly straight line.
Sum of Squared Errors (SSE)	the calculated value from adding up all the squared residual terms. The hope is that this value is very small when creating a model.
X – the independent variable	This will sometimes be referred to as the "predictor" variable, because these values were measured in order to determine what possible outcomes could be predicted.
Y – the dependent variable	Also, using the letter "y" represents actual values while $\hat{y}$ represents predicted or estimated values. Predicted values will come from plugging in observed "x" values into a linear model.

8.9: Linear Regression - Income and Assets (Worksheet)

Work in groups on these problems. You should try to answer the questions without referring to your notes. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will be able to describe the relationship between two quantitative variables.
- The student will be able to compute the least-squares regression line and interpret the slope.
- The student will be able to calculate the coefficient of determination and comment on the fit of a linear model.

The Data

As an equity analyst who wants to determine how closely a firm's annual net income is related to the dollar value of the assets, you gather the data for the different companies listed on the Zimbabwe Stock Exchange (ZSE) and presented it as below:

African Financial Survey Data

Company	Assets ($1000s)	Net Income ($1000s)
A	60	20
B	60	25
C	75	25
D	80	20
E	100	30
F	145	35
G	175	40
H	190	45
I	190	55
J	200	50

The Model

You will fit a linear model to the data.

1. Create a scatter plot of the net income (y) against the assets (x) for each company.
2. Describe the relationship between net income and assets.
3. Compute the mean assets and the mean net income for the sample of companies.
4. Find the least-squares regression equation for the data and plot the line on the scatter plot you constructed.
5. Interpret the slope of the regression line you found in Question 4.
6. Calculate the residual sum of squares.
7. Calculate the coefficient of determination.
8. Comment on whether you think a linear model is appropriate for this data.

8.H: Regression (Homework)

8.S: Regression (Solution)

1.d

2. A measure of the degree to which variation of one variable is related to variation in one or more other variables. The most commonly used correlation coefficient indicates the degree to which variation in one variable is described by a straight line relation with another variable.

Suppose that sample information is available on family income and Years of schooling of the head of the household. A correlation coefficient = 0 would indicate no linear association at all between these two variables. A correlation of 1 would indicate perfect linear association (where all variation in family income could be associated with schooling and vice versa).

3. a. 81% of the variation in the money spent for repairs is explained by the age of the auto

4. b. 16

5. The coefficient of determination is $r \cdot \cdot 2$ with $0 \leq r \cdot \cdot 2 \leq 1$, since $-1 \leq r \leq 1$.

6. True

7. d. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10

8. d. there exists no linear relationship between X and Y

9. Approximately 0.9

10. d. neither of the above changes will affect r.

11. Definition: A t test is obtained by dividing a regression coefficient by its standard error and then comparing the result to critical values for Students' t with Error df. It provides a test of the claim that $\beta_i = 0$ when all other variables have been included in the relevant regression model.

Example: Suppose that 4 variables are suspected of influencing some response. Suppose that the results of fitting $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i$ include:

Table $8.S.1$

Variable	Regression coefficient	Standard error of regular coefficient
.5	1	-3
.4	2	+2
.02	3	+1
.6	4	-.5

t calculated for variables 1, 2, and 3 would be 5 or larger in absolute value while that for variable 4 would be less than 1. For most significance levels, the hypothesis $\beta_1 = 0$ would be rejected. But, notice that this is for the case when X_2, X_3, and X_4 have been included in the regression. For most significance levels, the hypothesis $\beta_4 = 0$ would be continued (retained) for the case where X_1, X_2, and X_3 are in the regression. Often this pattern of results will result in computing another regression involving only X_1, X_2, X_3, and examination of the t ratios produced for that case.

12. c. those who score low on one test tend to score low on the other.

13. False. Since $H_0 : \beta = -1$ would not be rejected at $\alpha = 0.05$, it would not be rejected at $\alpha = 0.01$.

14. True

15. d

16. Some variables seem to be related, so that knowing one variable's status allows us to predict the status of the other. This relationship can be measured and is called correlation. However, a high correlation between two variables in no way proves that a cause-and-effect relation exists between them. It is entirely possible that a third factor causes both variables to vary together.

17. True

18. $Y_j = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + b_4 \cdot X_4 + b_5 \cdot X_6 + e_j$

19. d. there is a perfect negative relationship between Y and X in the sample.

20. b. low

21. The precision of the estimate of the Y variable depends on the range of the independent (X) variable explored. If we explore a very small range of the X variable, we won't be able to make much use of the regression. Also, extrapolation is not recommended.

22. $\hat{y} = -3.6 + (3.1 \cdot 7) = 18.1$

23. Most simply, since −5 is included in the confidence interval for the slope, we can conclude that the evidence is consistent with the claim at the 95% confidence level.

Using a t test: $H_0 : B_1 = -5$ $H_A : B_1 \neq -5$ $t_{\text{calculated}} = \frac{-5-(-4)}{1} = -1$ $t_{\text{critical}} = -1.96$.

Since $t_{\text{calc}} < t_{\text{crit}}$ we retain the null hypothesis that $B_1 = -5$.

24. True.

$t_{(\text{critical, },df=23,\text{ two-tailed, }\alpha=.02)} = \pm 2.5$

$\mathrm{t}_{\text{critical ,df=23, two-tailed, }\alpha=.01} = \pm 2.8$

25.

1. $80 + 1.5 \cdot 4 = 86$
2. No. Most business statisticians would not want to extrapolate that far. If someone did, the estimate would be 110, but some other factors probably come into play with 20 years.

26. d. one quarter

27. b. $r = -.77$

28.

1. $-.72, .32$
2. the t value
3. the t value

29.

1. The population value for β_2, the change that occurs in Y with a unit change in X_2, when the other variables are held constant.
2. The population value for the standard error of the distribution of estimates of β_2.
3. $.8, .1, 16 = 20 - 4$.

CHAPTER OVERVIEW

9: Apppendices

SECTION OVERVIEW

9.1: B | Mathematical Phrases, Symbols, and Formulas

Curated and edited by Kristin Kuter | Saint Mary's College, Notre Dame, IN

SECTION OVERVIEW

9.2: A | Statistical Tables

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: BFE 1201 Statistical Methods for Finance (Kuter)

Webpages: 157

All licenses found:

- CC BY 4.0: 90.4% (142 pages)
- Undeclared: 9.6% (15 pages)

By Page

- BFE 1201 Statistical Methods for Finance (Kuter) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Sampling and Data - *CC BY 4.0*
    - 1.1: Introduction to Sampling and Data - *CC BY 4.0*
    - 1.2: Definitions of Statistics, Probability, and Key Terms - *CC BY 4.0*
    - 1.3: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 1.4: Experimental Design and Ethics - *CC BY 4.0*
    - 1.5: Sampling Experiment (Worksheet) - *CC BY 4.0*
    - 1.6: Chapter Key Terms - *CC BY 4.0*
    - 1.7: Chapter References - *CC BY 4.0*
    - 1.H: Sampling and Data (Homework) - *CC BY 4.0*
    - 1.S: Sampling and Data (Solutions) - *CC BY 4.0*
  - 2: Descriptive Statistics - *CC BY 4.0*
    - 2.1: Introduction to Descriptive Statistics - *CC BY 4.0*
    - 2.2: Display Data - *CC BY 4.0*
    - 2.3: Measures of the Location of the Data - *CC BY 4.0*
    - 2.4: Measures of the Center of the Data - *CC BY 4.0*
    - 2.5: Sigma Notation and Calculating the Arithmetic Mean - *CC BY 4.0*
    - 2.6: Geometric Mean - *CC BY 4.0*
    - 2.7: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
    - 2.8: Measures of the Spread of the Data - *CC BY 4.0*
    - 2.9: Descriptive Statistics (Worksheet) - *CC BY 4.0*
    - 2.10: Chapter Key Terms - *CC BY 4.0*
    - 2.H: Descriptive Statistics (Homework) - *CC BY 4.0*
      - 1.1: Homework - *CC BY 4.0*
    - 2.S: Descriptive Statistics (Solutions) - *CC BY 4.0*
  - 3: Probability Topics - *CC BY 4.0*
    - 3.1: Introduction to Probability - *CC BY 4.0*
    - 3.2: Probability Terminology - *CC BY 4.0*
    - 3.3: Two Basic Rules of Probability - *CC BY 4.0*
    - 3.4: Contingency Tables and Probability Trees - *CC BY 4.0*
    - 3.5: Venn Diagrams - *CC BY 4.0*
    - 3.6: Conditional Probability and Bayes' Rule - *Undeclared*
    - 3.7: Independent Events - *Undeclared*
    - 3.8: Discrete Distribution Experiment (Worksheet) - *CC BY 4.0*
    - 3.9: Chapter Reference - *CC BY 4.0*
    - 3.10: Chapter Key Terms - *CC BY 4.0*
    - 3.H: Probability (Homework) - *CC BY 4.0*
      - 1.1: Chapter Practice - *CC BY 4.0*
      - 1.2: Chapter More Practice - *CC BY 4.0*
    - 3.S: Probability (Solutions- Practice + Homework) - *CC BY 4.0*
  - 4: Random Variables - *CC BY 4.0*
    - 4.1: Introduction to Discrete Random Variables - *CC BY 4.0*
    - 4.2: Binomial Distribution - *CC BY 4.0*
    - 4.3: Poisson Distribution - *CC BY 4.0*
    - 4.4: Expected Value of Discrete Random Variables - *Undeclared*
      - 4.4.1: Variance of Discrete Random Variables - *Undeclared*
    - 4.5: Introduction to Continuous Random Variables - *CC BY 4.0*
      - 4.5.1: Properties of Continuous Probability Density Functions - *CC BY 4.0*
    - 4.6: The Uniform Distribution - *CC BY 4.0*
    - 4.7: The Exponential Distribution - *CC BY 4.0*
    - 4.8: Introduction to Normal Distribution - *CC BY 4.0*
      - 4.8.1: The Standard Normal Distribution - *CC BY 4.0*
      - 4.8.2: Using the Normal Distribution - *CC BY 4.0*
    - 4.9: Normal Distribution - Pinkie Length (Worksheet) - *CC BY 4.0*

        - 1.4: Chi-Square Tests Homework - *CC BY 4.0*
        - 1.5: Chi-Square Practice - *CC BY 4.0*
    - 7.S: Hypothesis Testing (Solutions- Practice + Homework) - *CC BY 4.0*
        - 1.1: Two-Sample (Solutions- Practice + Homework) - *CC BY 4.0*
        - 1.2: Chi-square (Solutions- Practice + Homework) - *CC BY 4.0*
- 8: Linear Regression and Correlation - *CC BY 4.0*
    - 8.1: Introduction to Linear Regression and Correlation - *CC BY 4.0*
    - 8.2: The Correlation Coefficient r - *CC BY 4.0*
    - 8.3: Testing the Significance of the Correlation Coefficient - *CC BY 4.0*
    - 8.4: Linear Equations - *CC BY 4.0*
    - 8.5: The Regression Equation - *CC BY 4.0*
    - 8.6: Interpretation of Regression Coefficients- Elasticity and Logarithmic Transformation - *CC BY 4.0*
    - 8.7: Predicting with a Regression Equation - *CC BY 4.0*
    - 8.8: Chapter Key Terms - *CC BY 4.0*
    - 8.9: Linear Regression - Income and Assets (Worksheet) - *CC BY 4.0*
    - 8.H: Regression (Homework) - *CC BY 4.0*
    - 8.S: Regression (Solution) - *CC BY 4.0*
- 9: Apppendices - *CC BY 4.0*
    - 9.1: B | Mathematical Phrases, Symbols, and Formulas - *CC BY 4.0*
    - 9.2: A | Statistical Tables - *CC BY 4.0*
- Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*
    - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: BFE 1201 Statistical Methods for Finance (Kuter)

Webpages: 157

All licenses found:

- CC BY 4.0: 90.4% (142 pages)
- Undeclared: 9.6% (15 pages)

By Page

- BFE 1201 Statistical Methods for Finance (Kuter) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Sampling and Data - *CC BY 4.0*
    - 1.1: Introduction to Sampling and Data - *CC BY 4.0*
    - 1.2: Definitions of Statistics, Probability, and Key Terms - *CC BY 4.0*
    - 1.3: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 1.4: Experimental Design and Ethics - *CC BY 4.0*
    - 1.5: Sampling Experiment (Worksheet) - *CC BY 4.0*
    - 1.6: Chapter Key Terms - *CC BY 4.0*
    - 1.7: Chapter References - *CC BY 4.0*
    - 1.H: Sampling and Data (Homework) - *CC BY 4.0*
    - 1.S: Sampling and Data (Solutions) - *CC BY 4.0*
  - 2: Descriptive Statistics - *CC BY 4.0*
    - 2.1: Introduction to Descriptive Statistics - *CC BY 4.0*
    - 2.2: Display Data - *CC BY 4.0*
    - 2.3: Measures of the Location of the Data - *CC BY 4.0*
    - 2.4: Measures of the Center of the Data - *CC BY 4.0*
    - 2.5: Sigma Notation and Calculating the Arithmetic Mean - *CC BY 4.0*
    - 2.6: Geometric Mean - *CC BY 4.0*
    - 2.7: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
    - 2.8: Measures of the Spread of the Data - *CC BY 4.0*
    - 2.9: Descriptive Statistics (Worksheet) - *CC BY 4.0*
    - 2.10: Chapter Key Terms - *CC BY 4.0*
    - 2.H: Descriptive Statistics (Homework) - *CC BY 4.0*
      - 1.1: Homework - *CC BY 4.0*
    - 2.S: Descriptive Statistics (Solutions) - *CC BY 4.0*
  - 3: Probability Topics - *CC BY 4.0*
    - 3.1: Introduction to Probability - *CC BY 4.0*
    - 3.2: Probability Terminology - *CC BY 4.0*
    - 3.3: Two Basic Rules of Probability - *CC BY 4.0*
    - 3.4: Contingency Tables and Probability Trees - *CC BY 4.0*
    - 3.5: Venn Diagrams - *CC BY 4.0*
    - 3.6: Conditional Probability and Bayes' Rule - *Undeclared*
    - 3.7: Independent Events - *Undeclared*
    - 3.8: Discrete Distribution Experiment (Worksheet) - *CC BY 4.0*
    - 3.9: Chapter Reference - *CC BY 4.0*
    - 3.10: Chapter Key Terms - *CC BY 4.0*
    - 3.H: Probability (Homework) - *CC BY 4.0*
      - 1.1: Chapter Practice - *CC BY 4.0*
      - 1.2: Chapter More Practice - *CC BY 4.0*
    - 3.S: Probability (Solutions- Practice + Homework) - *CC BY 4.0*
  - 4: Random Variables - *CC BY 4.0*
    - 4.1: Introduction to Discrete Random Variables - *CC BY 4.0*
    - 4.2: Binomial Distribution - *CC BY 4.0*
    - 4.3: Poisson Distribution - *CC BY 4.0*
    - 4.4: Expected Value of Discrete Random Variables - *Undeclared*
      - 4.4.1: Variance of Discrete Random Variables - *Undeclared*
    - 4.5: Introduction to Continuous Random Variables - *CC BY 4.0*
      - 4.5.1: Properties of Continuous Probability Density Functions - *CC BY 4.0*
    - 4.6: The Uniform Distribution - *CC BY 4.0*
    - 4.7: The Exponential Distribution - *CC BY 4.0*
    - 4.8: Introduction to Normal Distribution - *CC BY 4.0*
      - 4.8.1: The Standard Normal Distribution - *CC BY 4.0*
      - 4.8.2: Using the Normal Distribution - *CC BY 4.0*
    - 4.9: Normal Distribution - Pinkie Length (Worksheet) - *CC BY 4.0*

    - 1.4: Chi-Square Tests Homework - *CC BY 4.0*
    - 1.5: Chi-Square Practice - *CC BY 4.0*
  - 7.S: Hypothesis Testing (Solutions- Practice + Homework) - *CC BY 4.0*
    - 1.1: Two-Sample (Solutions- Practice + Homework) - *CC BY 4.0*
    - 1.2: Chi-square (Solutions- Practice + Homework) - *CC BY 4.0*
- 8: Linear Regression and Correlation - *CC BY 4.0*
  - 8.1: Introduction to Linear Regression and Correlation - *CC BY 4.0*
  - 8.2: The Correlation Coefficient r - *CC BY 4.0*
  - 8.3: Testing the Significance of the Correlation Coefficient - *CC BY 4.0*
  - 8.4: Linear Equations - *CC BY 4.0*
  - 8.5: The Regression Equation - *CC BY 4.0*
  - 8.6: Interpretation of Regression Coefficients- Elasticity and Logarithmic Transformation - *CC BY 4.0*
  - 8.7: Predicting with a Regression Equation - *CC BY 4.0*
  - 8.8: Chapter Key Terms - *CC BY 4.0*
  - 8.9: Linear Regression - Income and Assets (Worksheet) - *CC BY 4.0*
  - 8.H: Regression (Homework) - *CC BY 4.0*
  - 8.S: Regression (Solution) - *CC BY 4.0*
- 9: Apppendices - *CC BY 4.0*
  - 9.1: B | Mathematical Phrases, Symbols, and Formulas - *CC BY 4.0*
  - 9.2: A | Statistical Tables - *CC BY 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*