FOUNDATIONS IN STATISTICAL REASONING

Peter Kaslik
Pierce College Fort Steilacoom

LibreTexts™

Book: Foundations in Statistical Reasoning (Kaslik)

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the hundreds of other texts available within this powerful platform, it is freely available for reading, printing and "consuming." Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects.

Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.

The LibreTexts mission is to unite students, faculty and scholars in a cooperative effort to develop an easy-to-use online platform for the construction, customization, and dissemination of OER content to reduce the burdens of unreasonable textbook costs to our students and society. The LibreTexts project is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning by developing an Open Access Resource environment. The project currently consists of 14 independently operating and interconnected libraries that are constantly being optimized by students, faculty, and outside experts to supplant conventional paper-based books. These free textbook alternatives are organized within a central environment that is both vertically (from advance to basic level) and horizontally (across different fields) integrated.

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

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

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org. More information on our activities can be found via Facebook (https://facebook.com/Libretexts), Twitter (https://twitter.com/libretexts), or our blog (http://Blog.Libretexts.org).

This text was compiled on 03/18/2025

TABLE OF CONTENTS

Licensing

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

1: Statistical Reasoning

Take a moment to visualize humanity in its entirety, from the earliest humans to the present. How would you characterize the well-being of humanity? Think beyond the latest stories in the news. To help clarify, think about medical treatment, housing, transportation, education, and our knowledge. While there is no denying that we have some problems that did not exist in earlier generations, we also have considerably more knowledge.

The progress humanity has made in learning about ourselves, our world and our universe has been fueled by the desire of people to solve problems or gain an understanding. It has been financed through both public and private monies. It has been achieved through a continual process of people proposing theories and others attempting to refute the theories using evidence. Theories that are not refuted become part of our collective knowledge. No single person has accomplished this, it has been a collective effort of humankind.

As much as we know and have accomplished, there is a lot that we don't know and have not yet accomplished. There are many different organizations and institutions that contribute to humanity's gains in knowledge, however one organization stands out for challenging humanity to achieve even more. This organization is XPrize.1 On their webpage they explain that they are an innovation engine. A facilitator of exponential change. A catalyst for the benefit of humanity." This organization challenges humanity to solve bold problems by hosting competitions and providing a monetary prize to the winning team. Examples of some of their competitions include:

- 2004: Ansari XPrize ($10 million) – Private Space Travel – build a reliable, reusable, privately financed, manned spaceship capable of carrying three people to 100 kilometers above the Earth's surface twice within two weeks.
- Current: The Barbara Bush Foundation Adult Literacy XPrize ($7 million) - "challenging teams to develop mobile applications for existing smart devices that result in the greatest increase in literacy skills among participating adult learners in just 12 months."

There are an estimated 36 million American adults with a reading level below third grade level. They have difficulty reading bedtime stories, reading prescriptions, and completing job applications, among other things. Developing a good app could have huge benefits for a lot of people, which would also provide benefits for the country.

The following fictional story will introduce you to the way data and statistics are used to test theories and make decisions. The goal is for you to see that the thought processes are not algebraic and that it is necessary to develop new ways of thinking so we can validate our theories or make evidence- based decisions.

Adult Literacy Prize Story

Imagine being part of a team competing for the Adult Literacy Xprize. During the early stages of development, a goal of your team is to create an app that is engaging for the user so that they will use it frequently. You tested your first version (Version 1) of the app on some adults who lacked basic literacy and found it was used an average of 6 hours during the first month. Your team decided this was not very impressive and that you could do better, so you developed a completely new version of the software designated as Version 2. When it was time to test the software, the 10 members of your team each gave it to 8 different people with low literacy skills. This group of 80 individuals that received the software is a small subset, or sample, of all those who have low literacy skills. The objective was to determine if Version 2 is used more than an average of 6 hours per month.

While the data will ultimately be pooled together, your teammates decide to compete against each other to determine whose group of 8 does better. The results are shown in the table below. The column on the right is the mean (average) of the data in the row. The mean is found by adding the numbers in the row and dividing that sum by 8.

Team Member	Version 2 Data (hours of use in 1 month)								Mean
You, The reader	4.4	3.8	4.4	6.7	1.1	5.7	0.8	2.5	3.675
Betty	11	8.4	8.4	2.7	4.4	8.4	5.7	4.4	6.675
Joy	1.6	2.2	12.5	5.7	2.2	6.6	0.8	0.3	3.9875
Kerissa	16.1	11.1	8.7	9.1	1.4	9.1	1.2	14.4	8.8875

Team Member	Version 2 Data (hours of use in 1 month)								Mean
Crystal	0	2.1	0	3.2	0.2	1.8	9.1	3.3	2.4625
Marcin	2.2	6.3	1.3	8.8	0.8	2.7	0.9	0.8	2.975
Tisa	8.8	5.8	9.7	2.8	3.2	0.9	0.1	16.1	5.925
Tyler	11	0.9	11.3	6.6	0.3	5.9	1.7	1.9	4.95
Patrick	0.9	1.8	6.3	3.1	6.1	6.3	3.2	6.7	4.3

One way to make sense of the data is to graph it. The graph to the right is called a histogram. It shows the distribution of the amount of time the software was used by each participant. To interpret this graph, notice the scale on the horizontal (x) axis counts by 2. These numbers represent hours of use. The height of each bar shows how many usage times fall between the x values. For example, 26 people used the app between 0 and 2 hours while 2 people used the app between 16 and 18 hours.

Time of Use of Version 2 During One Month Trial
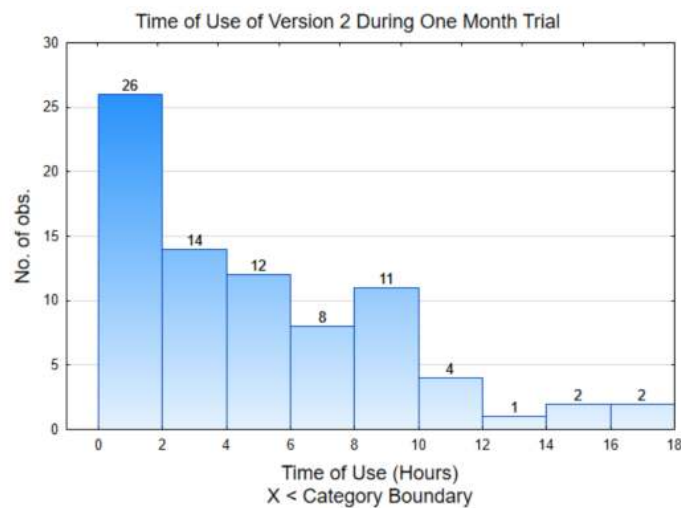
The second graph is a histogram of the mean (average) for each of the 10 groups. This is a graph of the column in the table that is shaded. A histogram of means is called a sampling distribution. The distribution to the right shows that 4 of the means are between 2 and 4 hours while only one mean was between 8 and 10 hours. Notice how the means are grouped closer together than the original data.

Sampling Distribution For Version 2
Mean Time For Each Group
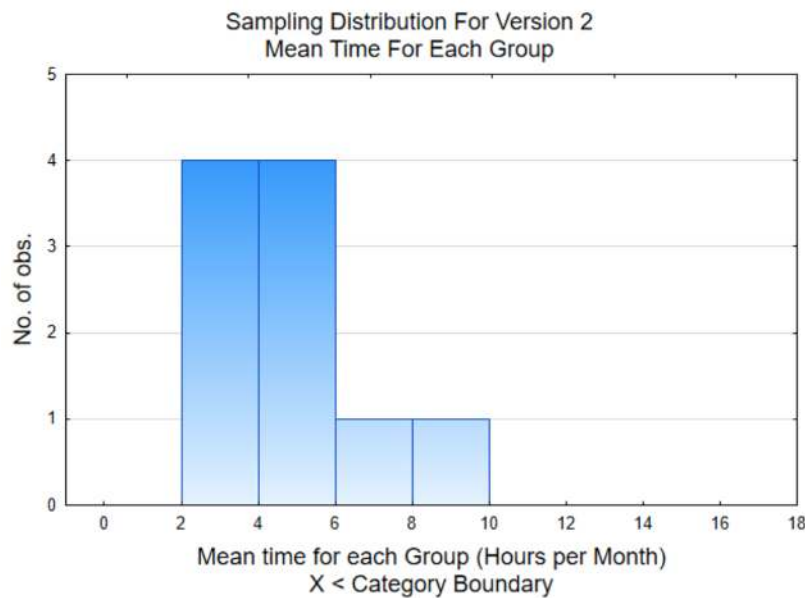
The overall mean for the 80 data values is 4.88 hours. Our task is to use the graphs and the overall mean to decide if Version 2 is used more than the Version 1 was used (6 hours per month). What is your conclusion? Answer this question before continuing your reading.

Yes Version 2 is better than Version 1 No, Version 2 is not better than Version 1

Which of the following had the biggest influence on your decision?

_____ 54 of the 80 data values were below 6

_____ The mean of the data is 4.88, which is below 6

_____ 8 of the 10 sample means are below 6.

Version 3

Version 3 was a total redesign of the software. A similar testing strategy was employed as with the prior version. When you received the data from the 8 users you gave the software to, you found that the average length of usage was 10.25 hours. Based on your results, do you feel that this version is better than version 1?

Team Member	Version 3 Data (hours of use in 1 month)								Mean
You, The reader	14	13	8	4	8	21	3	11	10.25

Yes Version 3 is better than Version 1 No, Version 3 is not better than Version 1

Your colleague Keer looked at her data, which is shown in the table below. What conclusion would Keer arrive at, based on her data?

| Team Member	Version 3 Data (hours of use in 1 month)								Mean
Keer	0	3	2	3	5	4	8	11	4.

Yes Version 3 is better than Version 1 No, Version 3 is not better than Version 1

If your interpretation of your data and Keer's data are typical, then you would have concluded that Version 3 was better than Version 1 based on your data and Version 3 was not better based on Keer's data. This illustrates how different samples can lead to different conclusions. Clearly, the conclusion based on your data and the conclusion based on Keer's data cannot both be correct. To help appreciate who might be in error, let's look at all the data for the 80 people who tested Version 3 of the software.

Team Member	Version 3 Data (hours of use in 1 month)								Mean
You, The reader	14	13	8	4	8	21	3	11	10.25
Keer	0	3	2	3	5	4	8	11	4.5
Betty	8	5	5	4	5	0	1	16	5.5
Joy	7	5	8	4	7	13	7	6	7.125
Kerissa	8	6	14	3	11	2	5	8	7.125
Crystal	6	7	4	7	6	3	7	5	5.625
Marcin	7	7	6	1	2	7	5	5	5
Tisa	3	3	5	4	14	13	3	2	5.875
Tyler	0	7	2	7	4	2	5	2	3.625
Patrick	8	3	1	14	2	6	7	2	5.375

The histogram on the right is of the data from individual users. This shows that about half the data (42 out of 80) are below 6 and the rest are above 6.
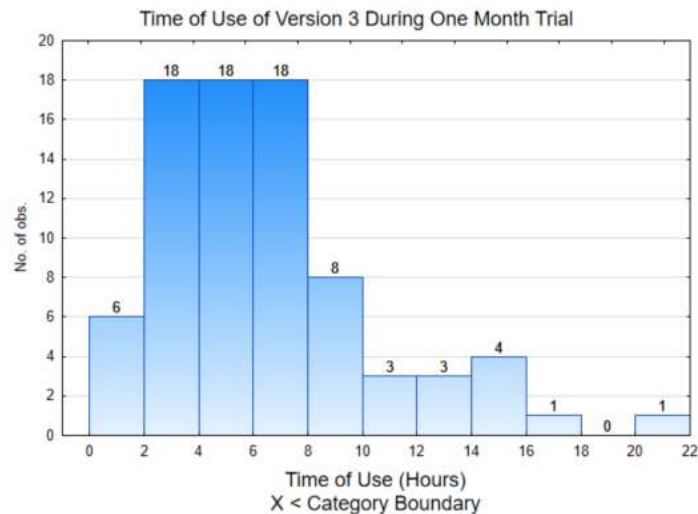
The histogram on the right is of the mean of the 8 users for each member of the team. This sampling distribution shows that 7 of the 10 sample means are below 6.

The mean of all the individual data values is 6.0. Consequently, if you concluded that Version 3 was better than Version 1 because the mean of your 8 users was 10.25 hours, you would have come to the wrong conclusion. You would have been misled by data that was selected by pure chance.

None of the first 3 versions was particularly successful but your team is not discouraged. They already have new ideas and are putting together another version of their literacy program.

Version 4.

When Version 4 is complete, each member of the team randomly selects 8 people with low literacy levels, just as was done for the prior versions. The data that is recorded is the amount of time the app is used during the month. Your data is shown below.

Team Member	Version 4 Data (hours of use in 1 month)							Mean
You, The reader	60	44	37	62	32	88	32	48.375

Based on your results, do you feel that this version is better than version 1?

Yes Version 4 is better than Version 1 No, Version 4 is not better than Version 1

The results of all 80 participants is shown in the table below.

Team Member	Version 4 Data (hours of use in 1 month)								Mean
You, The reader	60	44	37	32	62	32	88	32	48.375
Keer	48	37	24	20	82	76	67	67	52.625
Betty	88	39	67	24	71	85	81	24	59.875
Joy	23	58	21	88	81	75	84	81	63.875
Kerissa	88	24	58	53	81	57	88	24	59.125
Crystal	47	85	767	24	39	67	40	77	56.875
Marcin	61	45	75	58	87	51	37	73	60.875
Tisa	76	77	58	84	20	55	81	82	66.625

| Tyler | 82 | 47 | 48 | 60 | 88 | 21 | 50 | 24 | 52.5 |
| Patrick | 20 | 40 | 52 | 24 | 55 | 33 | 33 | 84 | 42.625 |

The histogram on the right is of the data from individual users. Notice that all these values are higher than 20.
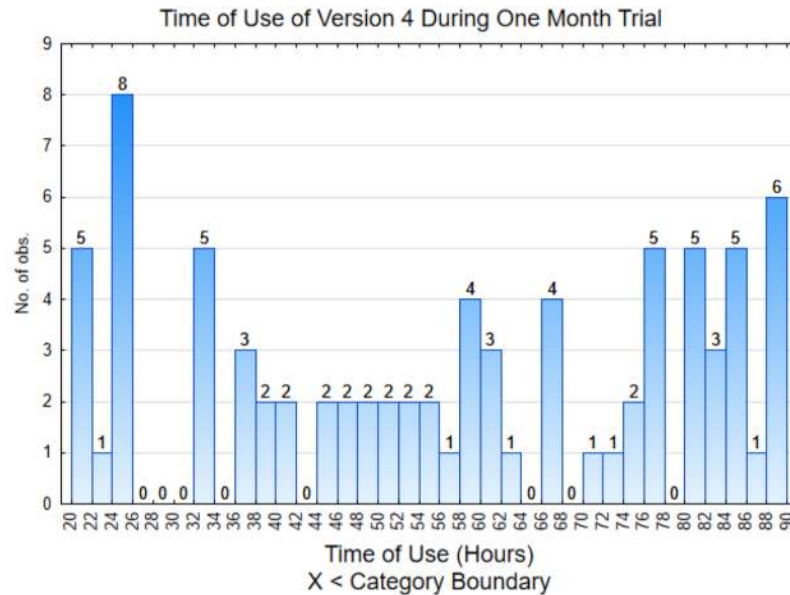
The histogram on the right is of the mean of the 8 users for each member of the team. Notice that all the sample means are significantly higher than 6.
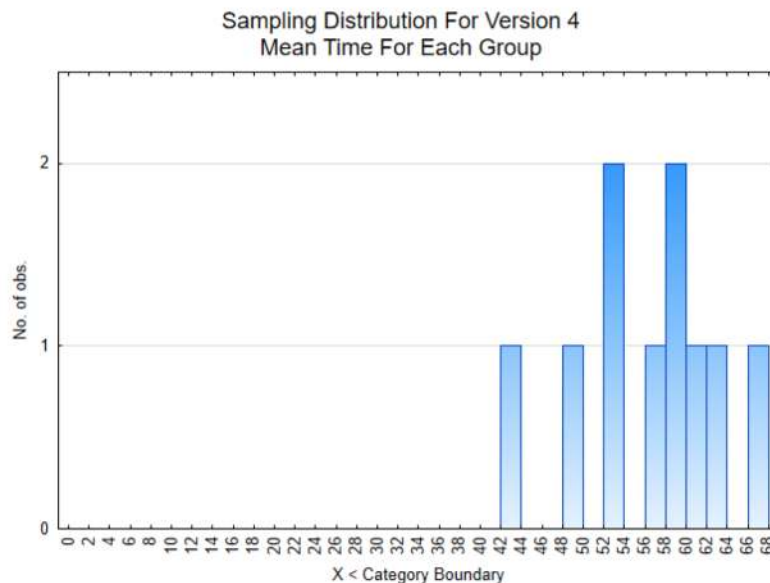
Based on the results of Version 4, all the data is much higher than 6 hours per month. The average is 56.3 hours per month which is almost 2 hours per day. This is significantly more usage of the app than the early versions and consequently will be the app that is used in the XPrize competition.

Making decisions using statistics

There were several objectives of the story you just read.

1. To give you an appreciation of the variation that can exist in sample data.
2. To introduce you to a type of data graph called a histogram, which is a good way for looking at the distribution of data.
3. To introduce you to the concept of a sampling distribution, which is a distribution of means of a sample, rather than of the original data.
4. To illustrate the various results that can occur when we try to answer questions using data. These results are summarized below in answer to the question of whether the new version is better than the first version.

   a. Version 2: This was not better. In fact, it appeared to be worse.

   b. Version 3: At first it looked better, but ultimately it was the same.

   c. Version 4: This was much better.

Because data sometimes provide clarity about a decision that should be made (Versions 2 and 4), but other times is not clear (Version 3), a more formal, statistical reasoning process will be explained in this chapter with the details being developed throughout the rest of the book.

Before beginning with this process, it is necessary to be clear about the role of statistics in helping us understand our world. There are two primary ways in which we establish confidence in our knowledge of the world, by providing analytical evidence or empirical evidence.

Analytical evidence makes use of definitions or mathematical rules. A mathematical proof is an analytical method for using established facts to prove something new. Analytical evidence is useful for proving things that are deterministic. **Deterministic** means that the same outcome will be achieved each time (if errors aren't made). Algebra and Calculus are examples of deterministic math and they can be used to provide analytical evidence.

In contrast, empirical evidence is based on observations. More specifically, someone will propose a theory and then research can be conducted to determine the validity of that theory. Most of the ideas we believe with confidence have resulted because of the rejection of theories we previously had and our current knowledge consists of those ideas we have not been able to reject with empirical evidence. Empirical evidence is gained through rigorous research. This contrasts with anecdotal evidence, which is also gained through observation, but not in a rigorous manner. Anecdotal evidence can be misleading.

The role of statistics is to objectively evaluate the evidence so a decision can be made about whether to reject, or not reject a theory. It is particularly useful for those situations in which the evidence is the result of a sample taken from a much larger population. In contrast to deterministic relationships, **stochastic** populations are ones in which there is randomness, while the evidence is gained though random sampling, thus meaning the evidence we see is the result of chance.

The scientific method that is used throughout the research community to increase our understanding of the world is based on proposing and then testing theories using empirical methods. Statistics plays a vital role in helping researchers understand the data they produce. The scientific method contains the following components.

1. Ask a question
2. Propose a hypothesis about the answer to the question
3. Design research (Chapter 2)
4. Collect data (Chapter 2)
5. Develop an understanding of the data using graphs and statistics (Chapter 3)
6. Use the data to determine if it supports, or contradicts the hypothesis (Chapters 5,7,8)
7. Draw a conclusion.

Before exploring the statistical tools used in the scientific method, it is helpful to understand the challenges we face with stochastic populations and the statistical reasoning process we use to draw conclusions.

1. When a theory is proposed about a population, it is based on every person or element of the population. A **population** is the entire set of people or things of interest.
2. Because the population contains too many people or elements from which to get information, we make a hypothesis about what the information would be, if we could get all of it.
3. Evidence is collected by taking a sample from the population.
4. The evidence is used to determine if the hypothesis should be rejected or not rejected.

These four components of the statistical reasoning process will now be developed more fully. The challenge is to determine if there is sufficient support for the hypothesis, based on partial evidence, when it is known that partial evidence varies, depending upon the sample that was selected. By analogy, it is like trying to find the right person to marry, by getting partial evidence from dating or to find the right person to hire, by getting partial evidence from interviews.

1. Theories about populations.

When someone has a theory, that theory applies to a population that should be clearly defined. For example, a population might be everyone in the country, or all senior citizens, or everyone in a political party, or everyone who is athletic, or everyone who is bilingual, etc. Populations can also be any part of the natural world including animals, plants, chemicals, water, etc. Theories that might be valid for one population are not necessarily valid for another. Examples of theories being applied to a population include the following.

- The team working on the literacy app theorizes that one version of their app will be used regularly by the entire population of adults with low literacy skills who have access to it.
- A teacher theorizes that her teaching pedagogy will lead to the greatest level of success for the entire population of all the students she will teach.
- A pharmaceutical company theorizes that a new medicine will be effective in treating the entire population of people suffering from a disease who use the medicine.
- A water resource scientist theorizes that the level of contamination in an entire body of water is at an unsafe level.

1.5 Data, Parameters, and Statistics

Before discussing hypotheses, it is necessary to talk about data, parameters and statistics.

On the largest level, there are two types of data, categorical and quantitative. **Categorical data** is data that can be put into categories. Examples include yes/no responses, or categories such as color, religion, nationality, pass/fail, win/lose, etc. **Quantitative data** is data that consists of numbers resulting from counts or measurements. Examples include, height, weight, time, amount of money, number of crimes, heart rate, etc.

The ways in which we understand the data, graphs and statistics, are dependent upon the type of data. Statistics are numbers used to summarize the data. For the moment, there are two statistics that will be important, proportions and means. Later in the book, other statistics will be introduced.

A **proportion** is the part divided by the whole. It is similar to percent, but it is not multiplied by 100. The part is the number of data values in a category. The whole is the number of data values that were collected. Thus, if 800 people were asked if they had ever visited a foreign country and 200 said they had, then the proportion of people who had visited a foreign country would be:

$$\frac{\text{part}}{\text{whole}} = \frac{x}{n} = \frac{200}{800} = 0.25$$

The part is represented by the variable x and the whole by the variable n.

A **mean**, often known as an average, is the sum of the quantitative data divided by the number of data values. If we refer back to the literacy app, version 3, the data for Marcin was:

Marcin	7	7	6	1	2	7	5	5	5

The mean is $\dfrac{7+7+6+1+2+7+5+}{8} = \dfrac{40}{8} = 5$

While statistics are numbers that are used to summarize sample data, parameters are numbers used to summarize all the data in the population. To find a parameter, however, requires getting data from every person or element in the population. This is called a **census**. Generally, it is too expensive, takes too much time, or is simply impossible to conduct a census. However, because our theory is about the population, then we have to distinguish between parameters and statistics. To do this, we use different variables.

Data Type	Summary	Population	Sample
Categorical	Proportion	p	$\hat{p}$ (p-hat)
Quantitative	Mean	μ	$bar x$ (x-bar)

To elaborate, when the data is categorical, the proportion of the entire population is represented with the variable p, while the proportion of the sample is represented with the variable $\hat{p}$. When the data is quantitative, the mean of the entire population is represented with the Greek letter μ, while the mean of the sample is represented with the variable $barx$.

In a typical situation, we will not know either p or μ and so we would make a hypothesis about them. From the data we collect we will find $\hat{p}$ or $barx$ and use that to determine if we should reject our hypothesis.

2. Hypotheses

Hypotheses are written about parameters before data is collected (*a priori*). Hypotheses are written in pairs that contain a null hypothesis (H_0) and an alternative hypothesis (H_1).

Suppose someone had a theory that the proportion of people who have attended a live sporting event in the last year was greater than 0.2. In such a case, they would write their hypotheses as:

$H_0 : p = 0.2$

$H_1 : p > 0.2$

If someone had a theory that the mean hours of watching sporting events on the TV was less than 15 hours per week, then they would write their hypotheses as:

$H_0 : \mu = 15$

$H_1 : \mu < 15$

The rules that are used to write hypotheses are:

1. There are always two hypotheses, the null and the alternative.
2. Both hypotheses are about the same parameter.
3. The null hypothesis always contains the equal sign (=).
4. The alternative contains an inequality sign ($<, >, \neq$).
5. The number will be the same for both hypotheses.

When hypotheses are used for decision making, they should be selected in such a way that if the evidence supports the null hypothesis, one decision should be made, while evidence supporting the alternative hypothesis should lead to a different decision.

The hypothesis that researchers desire is often the alternative hypothesis. The hypothesis that will be tested is the null hypothesis. If the null hypothesis is rejected because of the evidence, then the alternative hypothesis is accepted. If the evidence does not lead to a rejection of the null hypothesis, we cannot conclude the null is true, only that it was not rejected. We will use the term "supported" in this text. Thus either the null hypothesis is supported by the data or the alternative hypothesis is supported. Being supported by the data does not mean the hypothesis is true, but rather that the decision we make should be based on the hypothesis that is supported.

Two of the situations you will encounter in this text are when there is a theory about the proportion or mean for one population or when there is a theory about how the proportion or mean compares between two populations. These are summarized in the table below.

Hypothesis about one population	Notation	Hypothesis about 2 populations	Notation
The proportion is greater than 0.2	$H_0 : p = 0.2$ $H_1 : p > 0.2$	The proportion of population A is greater than the proportion of population B	$H_0 : p_A = p_B$ $H_1 : p_A > p_B$
The proportion is less than 0.2	$H_0 : p = 0.2$ $H_1 : p < 0.2$	The proportion of population A is less than the proportion of population B	$H_0 : p_A = p_B$ $H_1 : p_A < p_B$
The proportion is not equal to 0.2	$H_0 : p = 0.2$ $H_1 : p \neq 0.2$	The proportion of population A is different than the proportion of population B	$H_0 : p_A = p_B$ $H_1 : p_A \neq p_B$

The mean is greater than 15	$H_0 : \mu = 15$ $H_1 : \mu > 15$	The mean of population A is greater than the mean of population B	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A > \mu_B$
The mean is less than 15	$H_0 : \mu = 15$ $H_1 : \mu < 15$	The mean of population A is less than the mean of population B	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A < \mu_B$
The mean does not equal 15	$H_0 : \mu = 15$ $H_1 : \mu \neq 15$	The mean of population A is different than the mean of population B	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A \neq \mu_B$

3. Using evidence to determine which hypothesis is more likely correct.

From the Literacy App story, you should have seen that sometimes the evidence clearly supports one conclusion (e.g. version 2 is worse than version 1), sometimes it clearly supports the other conclusion (version 4 is better than version 1), and sometimes it is too difficult to tell (version 3). Before discussing a more formal way of testing hypotheses, let's develop some intuition about the hypotheses and the evidence.

Suppose the hypotheses are

H_0: p = 0.4

H_0: p < 0.4

If the evidence from the sample is $\hat{p} = 0.45$, would this evidence support the null or alternative? Decide before continuing.

The hypotheses contain an equal sign and a less than sign but not a greater than sign, so when the evidence is greater than, what conclusion should be drawn? Since the sample proportion does not support the alternative hypothesis because it is not less than 0.4, then we will conclude 0.45 supports the null hypothesis.

If the evidence from the sample is $\hat{p} = 0.12$, would this evidence support the null or alternative? Decide before continuing.

In this case, 0.12 is considerably less than 0.4, therefore it supports the alternative.

If the evidence from the sample is $\hat{p} = 0.38$, would this evidence support the null or alternative? Decide before continuing.

This is a situation that is more difficult to determine. While you might have decided that 0.38 is less than 0.4 and therefore supports the alternative, it is more likely that it supports the null hypothesis.

How can that be?

In arithmetic, 0.38 is always less than 0.4. However, in statistics, it is not necessarily the case. The reason is that the hypothesis is about a parameter, it is about the entire population. On the other hand, the evidence is from the sample. Different samples yield different results. A direct comparison of the statistic (0.38) to the hypothesized parameter (0.4) is not appropriate. Rather, we need a different way of making that determination. Before elaborating on the different way, let's try another one.

Suppose the hypotheses are

$H_0 : \mu = 30$

$H_1 : \mu > 30$

If the evidence from the sample is $\bar{x} = 80$, which hypothesis is supported? Null Alternative

If the evidence from the sample is $\bar{x} = 26$, which hypothesis is supported? Null Alternative

If the evidence from the sample is $\bar{x} = 32$, which hypothesis is supported? Null Alternative

If the evidence is $\bar{x} = 80$, the alternative would be supported. If the evidence is $\bar{x} = 26$, the null would be supported. If the evidence is $\bar{x} = 32$, at first glance, it appears to support the alternative, but it is close to the hypothesis, so we will conclude that we are not sure which it supports.

It might be disconcerting to you to be unable to draw a clear conclusion from the evidence. After all, how can people make a decision? What follows is an explanation of the statistical reasoning strategy that is used.

Statistical Reasoning Process

The reasoning process for deciding which hypothesis the data supports is the same for any parameter (p or μ).

1. Assume the null hypothesis is true.
2. Gather data and calculate the statistic.
3. Determine the likelihood of selecting the data that produced the statistic or could produce a more extreme statistic, assuming the null hypothesis is true.
4. If the data are likely, they support the null hypothesis. However, if they are unlikely, they support the alternative hypothesis.

To illustrate this, we will use a different research question: "What proportion of American adults believe we should transition to a society that no longer uses fossil fuels (coal, oil, natural gas)? Let's assume a researcher has a theory that the proportion of American adults who believe we should make this transition is greater than 0.6. The hypotheses that would be used for this are:

H_0 : p = 0.6

H_1 : p > 0.6

We could visualize this situation if we used a bag of marbles. Since the first step in the statistical reasoning process is to assume the null hypothesis is true, then our bag of marbles might contain 6 green marbles that represent the adults who want to stop using fossil fuels, and 4 white marbles to represent those who want to keep using fossil fuels. Sampling will be done with replacement, which means that after a marble is picked, the color is recorded and the marble is placed back in the bag.

If 100 marbles are selected from the bag (with replacement), do you expect exactly 60 of them (60%) to be green? Would this happen every time?

The results of a computer simulation of this sampling process are shown below. The simulation is of 100 marbles being selected, with the process being repeated 20 times.

0.62	0.57	0.58	0.64	0.64	0.53	0.73	0.55	0.58	0.55
0.61	0.66	0.6	0.54	0.54	0.5	0.62	0.55	0.61	0.61

Notice that some of the times, the sample proportion is greater than 0.6, some of the times it is less than 0.6 and there is only one time in which it actually equaled 0.6. From this we can infer that although the null hypothesis really was true, there are sample proportions that might make us think the alternative is true (which could lead us to making an error).

There are three items in the statistical reasoning process that need to be clarified. The first is to determine what values are likely or unlikely to occur while the second is to determine the division point between likely and unlikely. The third point of clarification is the direction of the extreme.

Likely and Unlikely values

When the evidence is gathered by taking a random sample from the population, the random sample that is actually selected is only one of many, many, many possible samples that could have been taken instead. Each random sample would produce different statistics. If you could see all the statistics, you would be able to determine if the sample you took was likely or unlikely. A graph of statistics, such as sample proportions or sample means, is called a **sampling distribution**.

While it does not make sense to take lots of different samples to find all possible statistics, a few demonstrations of what happens if someone does do that can give you some confidence that similar results would occur in other situations as well. The data used in the graphs below were done using computer simulations.

The histogram at the right is a sampling distribution of sample proportions. 100 different samples that contained 200 data values were selected from a population in which 40% favored replacing fossil fuel (green marbles). The proportion in favor of replacing fossil fuels (green marbles) was found for each sample and graphed. There are two things you should notice in the graph. The first is that most of the
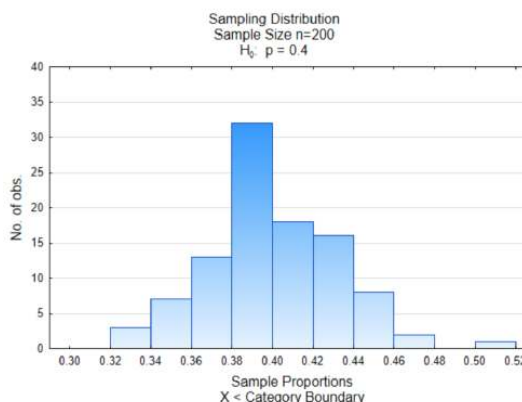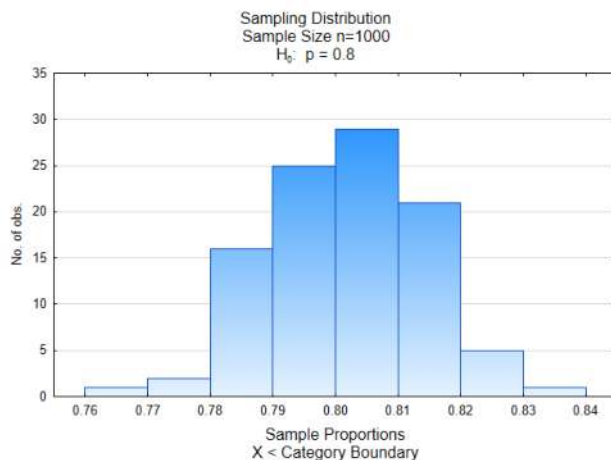
sample proportions are grouped together in the middle and the second thing is that the middle is approximately 0.40 which is equivalent to the proportion of green marbles in the container.

That may, of course, have been a coincidence. So let's look at a different sample. In this one, the original population was 60% green marbles representing those in favor of replacing fossil fuels. The sample size was 500 and the process was repeated 100 times.

Once again we see most of the sample proportions grouped in the middle and the middle is around the value of 0.60, which is the proportion of green marbles in the original population.

We will look at one more example. In this example, the proportion in favor of replacing fossil fuels is 0.80 while the proportion of those opposed is 0.20. The sample size will be 1000 and there will be 100 samples of that size. Where do you expect the center of this distribution to fall?

As you can see, the center of this distribution is near 0.80 with more values near the middle than at the edges.

One issue that has not been addressed is the effect of the sample size. Sample sizes are represented with the variable n. These three graphs all had different sample sizes. The first sample had n=200, the second had n=500 and the third had n=1000. To see the effect of these different sample sizes, all three sets of sample proportions have been graphed on the same histogram.

What this graph illustrates is that the smaller the sample size, the more variation that exists in the sample proportions. This is evident because they are spread out more. Conversely, the larger the sample size, the less variation that exists. What this means is the larger the sample size, the closer the sample result will be to the parameter. Does this seem reasonable? If there were 10,000 people in a population and you got the opinion of 9,999 of them, do you think all your possible sample proportions would be closer to the parameter (population proportion) than if you only asked 20 people?

We will return to sampling distributions in a short time, but first we need to learn about directions of extremes and probability.

Direction of Extreme

The direction of extreme is the direction (left or right) on a number line that would make you think the alternative hypothesis is true. Greater than symbols have a direction of extreme to the right, less than symbols indicate the direction is to the left and not-equal signs indicate a two-sided direction of extreme.
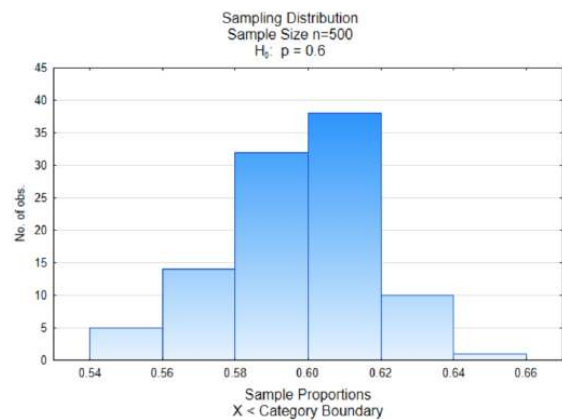
Histogram for 100 sample proportions with sample sizes of n = 200, n = 500, and n = 1000, corresponding with p=0.4, p=0.6, and p=0.8, respectively.
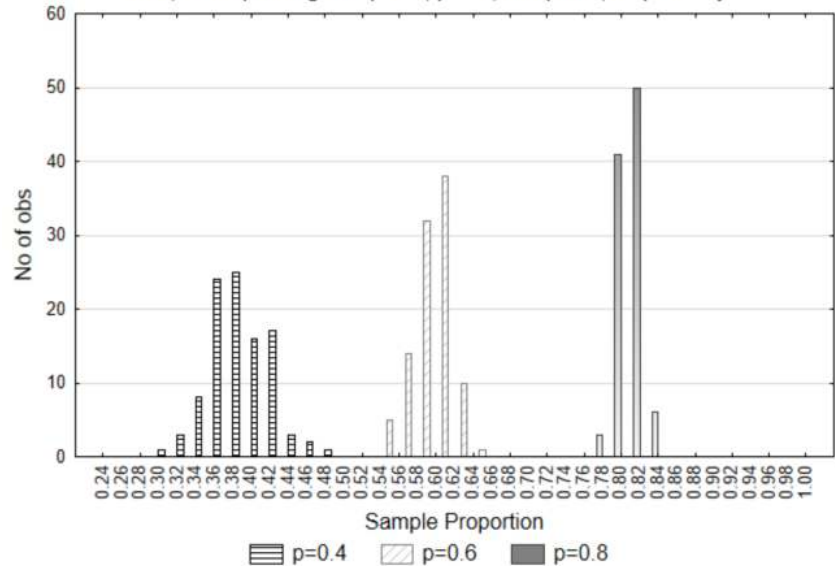
Notation	Notation	Direction of Extreme
$H_0 : p = 0.2$ $H_1 : p > 0.2$	$H_0 : p_A = p_B$ $H_1 : p_A > p_B$	Right
Left	$H_0 : p_A = p_B$ $H_1 : p_A < p_B$	Left
$H_0 : p = 0.2$ $H_1 : p \neq 0.2$	$H_0 : p_A = p_B$ $H_1 : p_A \neq p_B$	Two-sided
$H_0 : \mu = 15$ $H_1 : \mu > 15$	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A > \mu_B$	Right
$H_0 : \mu = 15$ $H_1 : \mu < 15$	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A < \mu_B$	Left
$H_0 : \mu = 15$ $H_1 : \mu \neq 15$	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A \neq \mu_B$	Two-sided

Probability

At this time it is necessary to have a brief discussion about probability. A more detailed discussion will occur in Chapter 4. When theories are tested empirically by sampling from a stochastic population, then the sample that is obtained is based on chance. When a sample is selected through a random process and the statistic is calculated, it is possible to determine the probability of obtaining that statistic or more extreme statistics if we know the sampling distribution.

By definition, probability is the number of favorable outcomes divided by the number of possible outcomes.

$$P(A) = \frac{Number\ of\ Favorable\ Outcomes}{Number\ of\ Possible\ Outcomes} \tag{1.1}$$

This formula assumes that all outcomes are equally likely as is theoretically the case in a random selection processes. It reflects the proportion of times that a result would be obtained if an experiment were done a very large number of times. Because you cannot have a negative number of outcomes or more successful outcomes than are possible, probability is always a fraction or a decimal between 0 and 1. This is shown generically as $0 \leq P(A) \leq 1$ where P(A) represents the probability of event A.

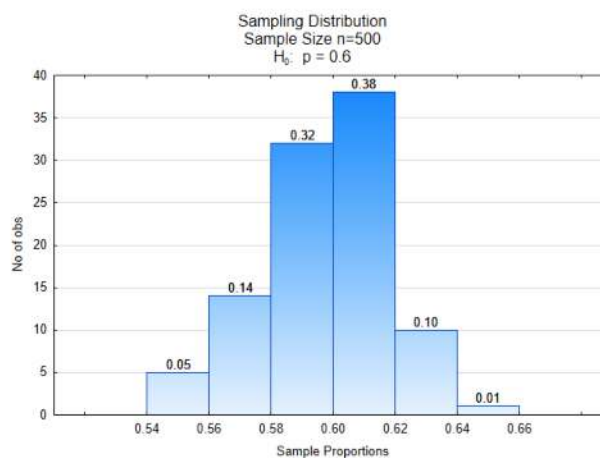**Using Sampling Distributions to Test Hypotheses**

Remember our research question, "What proportion of American adults believe we should transition to a society that no longer uses fossil fuels (coal, oil, natural gas)? The researchers theory is that the proportion of American adults who believe we should make this transition is greater than 0.6. The hypotheses that would be used for this are:

$H_0 : p = 0.6$

$H_1 : p > 0.6$

To test this hypothesis, we need two things. First, we need the sampling distribution for the null hypothesis, since we will assume that is true, as stated first in the list for the reasoning process used for testing a hypothesis. The second thing we need is data. Because this is instructional, at this point, several sample proportions will be provided so you can compare and contrast the results.

A small change has been made to the sampling distribution that was shown previously. At the top of each bar is a proportion. On the x-axis there are also proportions. The difference between these proportions is that the ones on the x-axis indicate the sample proportions while the proportions at the top of the bars indicate the proportion of sample proportions that were between the two boundary values. Thus, out of 100 sample proportions, 0.38 (or 38%) of them were between 0.60 and 0.62. The proportions at the top of the bars can also be interpreted as probabilities.

It is with this sampling distribution from the null hypotheses that we can find the likelihood, or probability of getting our data, or more extreme data. We will call this probability a **p-value**.

As a reminder, for the hypothesis we are testing, the direction of extreme is to the right.

Suppose the sample proportion we got for our data was $\hat{p}$ = 0.64. What is the probability we would have gotten that sample proportion or more extreme from this distribution? That probability is 0.01, consequently the p-value is 0.01. This number is found at the top of the right-most bar.

Suppose the sample proportion we got from our data was $\hat{p}$ = 0.62. What is the probability we would have gotten that sample proportion from this distribution? That probability is 0.11. This was calculated by adding the proportions on the top of the two right-most bars. The p-value is 0.11.

You try it. Suppose the sample proportion we got from our data was $\hat{p}$ = 0.60. What is the probability we would have gotten that sample proportion from this distribution?

Now, suppose the sample proportion we got from our data was $\hat{p}$ = 0.68. What is the probability we would have gotten that sample proportion from this distribution? In this case, there is no evidence of any sample proportions equal to 0.68 or higher, so consequently the probability, or p-value would be 0.

Testing the hypothesis

We will now try to determine which hypothesis is supported by the data. We will use the p=0.8 distribution to represent the alternative hypothesis. Both the null and alternative distributions are shown on the same graph.
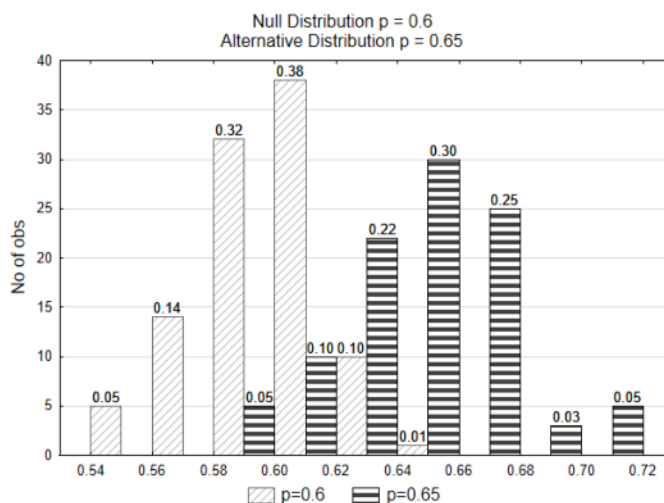
If the data that is selected had a statistic of $\hat{p} = 0.58$, what is the p-value? Which of the two distributions do you think the data came from? Which hypothesis is supported?

The p-value is 0.81 (0.32+0.38+0.10+0.01). This data came from the null distribution (p=0.6). This evidence supports the null hypothesis.

If the data that is selected was $\hat{p} = 0.78$, what is the p-value? Which of the two distributions do you think the data came from? Which hypothesis is supported?

The p-value is 0 because there are no values in the p=0.6 distribution that are 0.78 or higher. The data came from the alternative (p=0.8) distribution. The alternative hypothesis is supported.

In the prior examples, there as a clear distinction between the null and alternative distributions. In the next example, the distinction is not as clear. The alternative distribution will be represented with a proportion of 0.65

.

If the data that is selected was $\hat{p} = 0.62$, from which of the two distributions do you think the data came from? Which hypothesis is supported?

Notice that in this case, because the distributions overlap, a sample proportion of 0.62 or more extreme could have come from either distribution. It isn't clear which one it came from. Because of this lack of clarity, we could possibly make an error. We might think it came from the null distribution whereas it really came from the alternative distribution. Or perhaps we thought it came from the alternative distribution, but it really came from the null distribution. How do we decide???

Before explaining the way we decide, we need to discuss errors, as they are part of the decision- making process.

There are two types of errors we can make as a result of the sampling process. They are known as **sampling errors**. These errors are named Type I and Type II errors. A **type I error** occurs when we think the data supports the alternative hypothesis but in reality, the null hypothesis is correct. A **type II error** occurs when we think the data supports the null hypothesis, but in reality the alternative hypothesis is correct. In all cases of testing hypotheses, there is the possibility of making either a type I or type II error.

The probability of making either Type I or Type II errors is important in the decision-making process. We represent the probability of making a Type I error with the Greek letter alpha, α. It is also called the **level of significance**. The probability of making a Type II error is represented with the Greek letter Beta, β. The probability of the data supporting the alternative hypothesis, when the alternative is true is called **power**. Power is not an error. The errors are summarized in the table below.

		The True Hypothesis	
		H_0 Is True	H_1 Is True
The Evidence upon which the decision is based	The Data Supports H_0	No Error	Type II Error Probability: β
	The Data Supports H_1	Type I Error Probability: α	No Error Probability: Power

The reasoning process for deciding which hypothesis the data supports is reprinted here.

1. Assume the null hypothesis is true.
2. Gather data and calculate the statistic.
3. Determine the likelihood of selecting the data that produced the statistic or could produce a more extreme statistic, assuming the null hypothesis is true. This is called the p-value.
4. If the data are likely, they support the null hypothesis. However, if they are unlikely, they support the alternative hypothesis.

The determination of whether data are likely or not is based on a comparison between the p- value and α. Both alpha and p-values are probabilities. They must always be values between 0 and 1, inclusive. **If the p-value is less than or equal to α, the data supports the alternative hypothesis.** If the p-value is greater than α, the data supports the null hypothesis. When the data supports the alternative hypothesis, the data are said to be **significant**. When the data supports the null hypothesis, the data are**not significant**. *Reread this paragraph at least 3 times as it defines the decision making rule used throughout statistics and it is critical to understand.*

Because some values clearly support the null hypothesis, others clearly support the alternative hypothesis but some do not clearly support either, then a decision has to be made, before data is ever collected (*a priori*), as to the probability of making a type I error that is acceptable to the researcher.The most common values for α are 0.05, 0.01, and 0.10. There is not a specific reason for these choices but there is considerable historical precedence for them and they will be used routinely in this book. The choice for a level of significance should be based on several factors.

1. If the power of the test is low because of small sample sizes or weak experimental design, a larger level of significance should be used.
2. Keep in mind the ultimate objective of research – "to understand which hypotheses about the universe are correct. Ultimately these are yes and no decisions." (Scheiner, Samuel M., and Jessica Gurevitch. *Design and Analysis of Ecological Experiments*. Oxford [etc.: Oxford UP, 2001. Print.) Statistical tests should lead to one of three results. One result is that the hypothesis is almost certainly correct. The second result is that the hypothesis is almost certainly incorrect. The third result is that further research is justified. P- values within the interval (0.01,0.10) may warrant continued research, although these values are as arbitrary as the commonly used levels of significance.
3. If we are attempting to build a theory, we should use more liberal (higher) values of α, whereas if we are attempting to validate a theory, we should use more conservative (lower) values of α.

Demonstration of an elementary hypothesis test

Now, you have all the parts for deciding which hypothesis is supported by the evidence (the data). The problem will be restated here.

$H_0 : p = 0.6$

$H_1 : p > 0.6$

$\alpha = 0.01$

A vertical line was drawn on the graph so that a proportion of only 0.01 was to the right of the line in the null distribution. This is called a decision line because it is the line that determines how we will decide if the statistic supports the null or alternative hypothesis. The number at the bottom of the decision line is called the critical value.
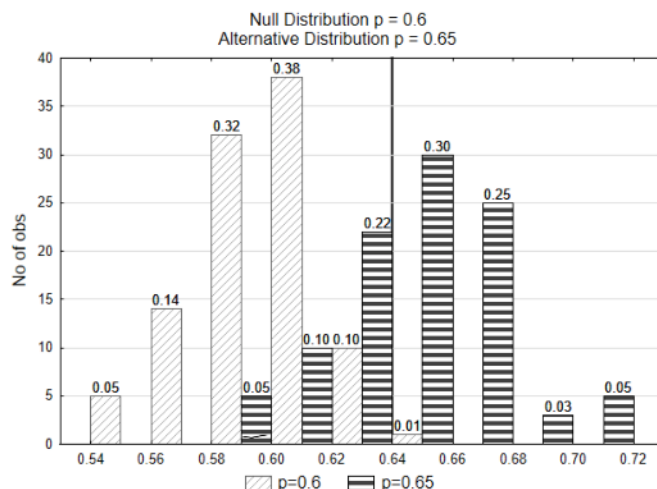
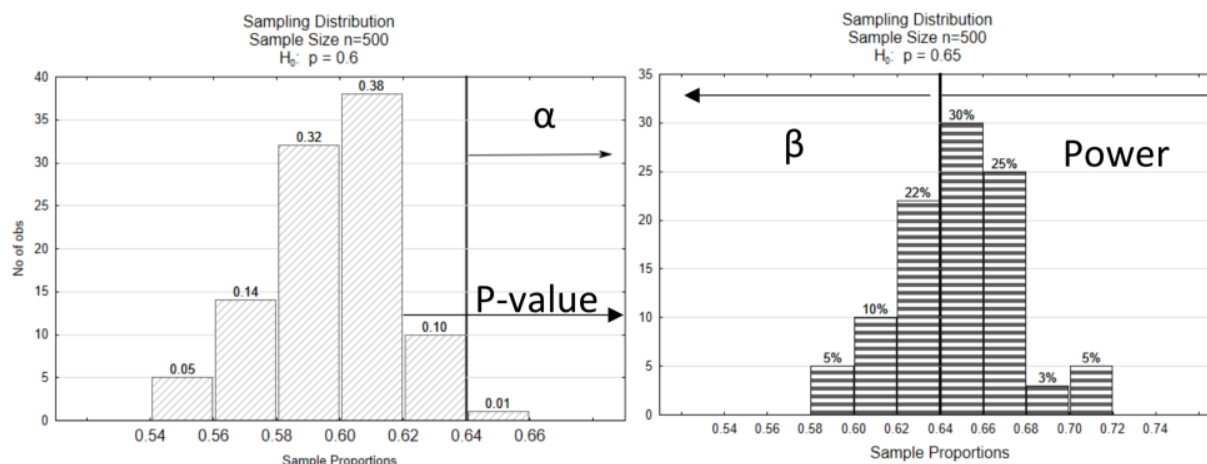Null Distribution p = 0.6
Alternative Distribution p = 0.65

If the data that is selected was $\hat{p} = 0.62$, from which of the two distributions do you think the data came from? Which hypothesis is supported?

To answer these questions, first find the p-value. The p-value is 0.11 (0.10 + 0.01).

Next, compare the p-value to α. Since 0.11 > 0.01, this evidence supports the null hypothesis.

Because showing both distributions on the same graph can make the graph a little difficult to read, this graph will be split into two graphs. The decision line is shown at the same critical value on both graphs (0.64). The level of significance, α, is shown on the null distribution. It points in the direction of the extreme. β and power are shown on the alternative distribution. Power is on the same side of the distribution as the direction of extreme while β is on the opposite side. The p-value is also shown on the null distribution, pointing in the direction of the extreme.

Another example will be demonstrated next.

Question: What is the proportion of people who have visited a different country?
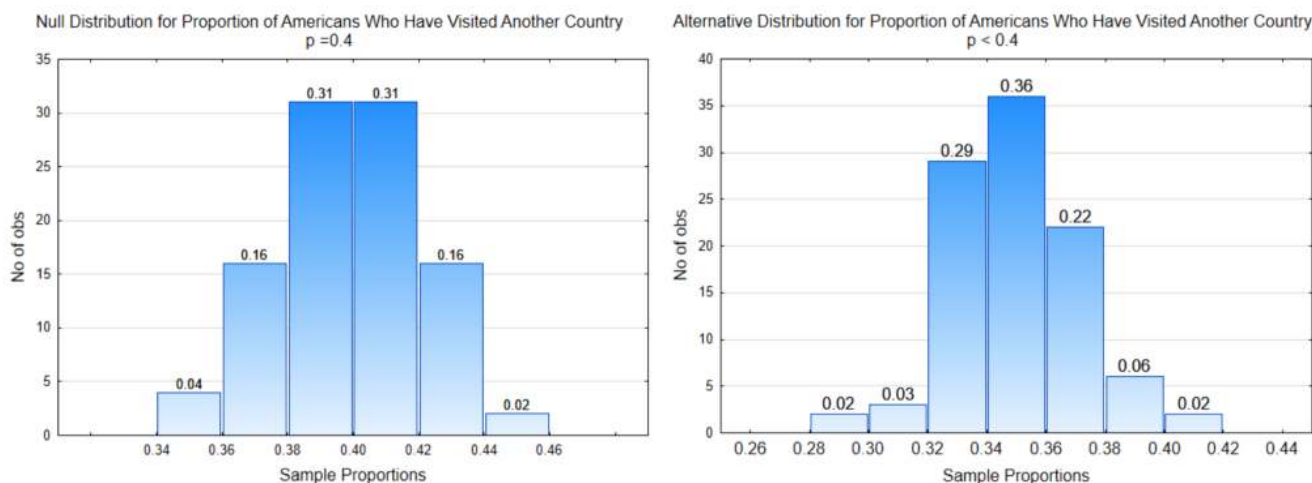
Theory: The proportion is less than 0.40

Hypotheses: $H_0 : p = 0.40$
$H_1 : p < 0.40$

$\alpha = 0.04$

The distribution on the left is the null distribution, that is, it is the distribution that was obtained by sampling from a population in which the proportion of people who have visited a different country is really 0.40. The distribution on the right is representing the alternative hypothesis.

The objective is to identify the portion of each graph associated with α, β, and power. Once the data has been provided, you will also be able to show the part of the graph that indicates the p-value.

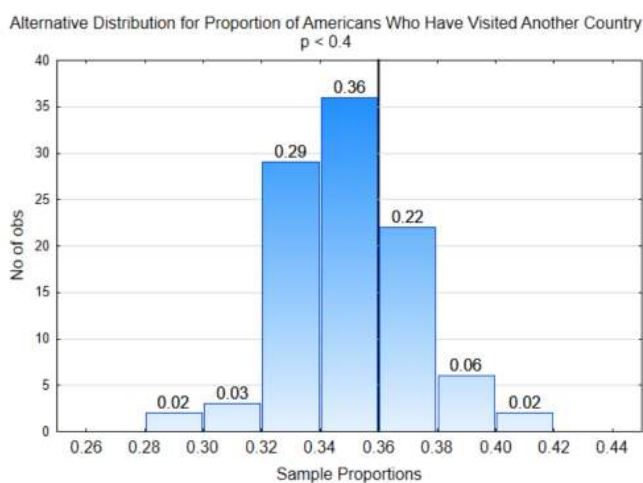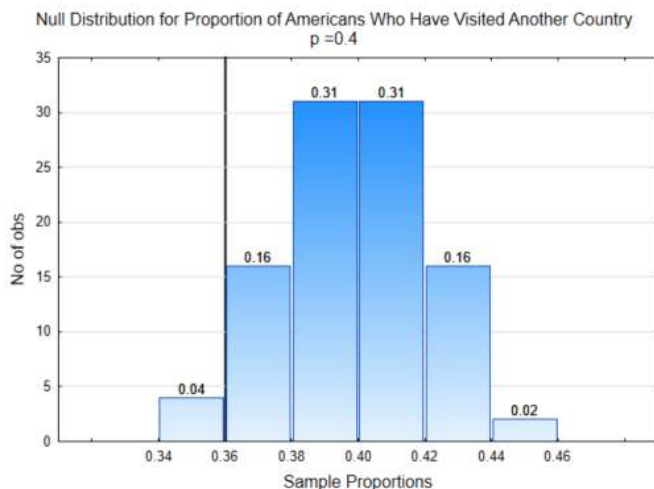The reasoning process for labeling the distributions is as follows.

1. Determine the direction of the extreme. This is done by looking at the inequality sign in the alternative hypothesis. If the sign is <, then the direction of the extreme is to the left. If the sign is >, then the direction of the extreme is to the right. If the sign is $\neq$, then the direction of extreme is to the left and right, which is called two-sided. Notice that the inequality sign points towards the direction of extreme. To keep these concepts a little easier as you are learning them, we will not do two-sided alternative hypotheses until later in the text.

In this problem the direction of extreme is to the left because smaller sample proportions support the alternative hypothesis.

2. Draw the Decision line. The direction of extreme along with α are used to determine the placement of the decision line. Alpha is the probability of making a Type I error. A Type I error can only occur if the null hypothesis is true, therefore, we always place alpha on the null distribution. Starting on the side of the direction of extreme, add the proportions at the top of the bars until they equal alpha. Draw the decision line between bars separating those that could lead to a Type I error from the rest of the distribution.

Notice the x-axis value at the bottom of the decision line. This value is called the critical value. Identify the critical value on the alternative distribution and place another decision line there.

In this problem, the direction of extreme is to the left and $\alpha = 4\%$ (0.04) so the decision line is placed so that the proportion of sample proportions to the left is 0.04. The critical value is 0.36 so the other decision line is placed at 0.36 on the alternative distribution.

Null Distribution for Proportion of Americans Who Have Visited Another Country p =0.4

Alternative Distribution for Proportion of Americans Who Have Visited Another Country p < 0.4

3. Labeling α, β, and power. α is always placed on the null distribution on the side of the decision line that is in the direction of extreme. β is always placed on the alternative distribution on the side of the decision line that is opposite of the direction of extreme. Power is always placed on the alternative distribution on the side of the decision line that is in the direction of extreme.

Null Distribution for Proportion of Americans Who Have Visited Another Country p =0.4

Alternative Distribution for Proportion of Americans Who Have Visited Another Country p < 0.4
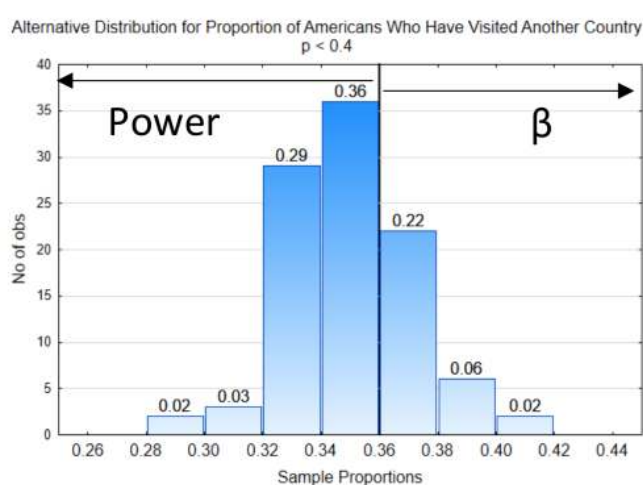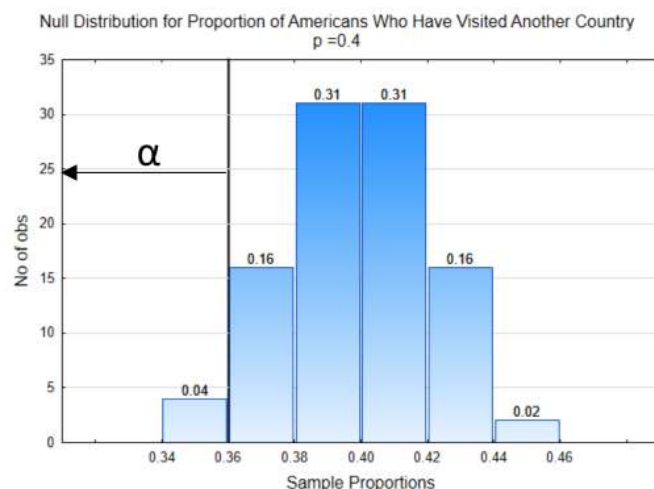
4. Identify the probabilities for α, β, and power. This is done by adding the proportions at the top of the bars.

In this example, the probability for α is 0.04. The probability for β is 0.30 (0.02 + 0.06 + 0.22). The probability for power is 0.70 (0.02 + 0.03 + 0.29 + 0.36).

5. Find the p-value. Data is needed to test the hypothesis, so here is the data: In a sample of 200 people, 72 have visited another country. The sample proportion is $\hat{p} = \dfrac{72}{200} = 0.36$. The p-value, which is the probability of getting the data, or more extreme values, assuming the null hypothesis is true, is always placed on the null distribution and always points in the direction of the extreme.

In this example, the p-value has been indicated on the null distribution.

Null Distribution for Proportion of Americans Who Have Visited Another Country p =0.4

Alternative Distribution for Proportion of Americans Who Have Visited Another Country p < 0.4
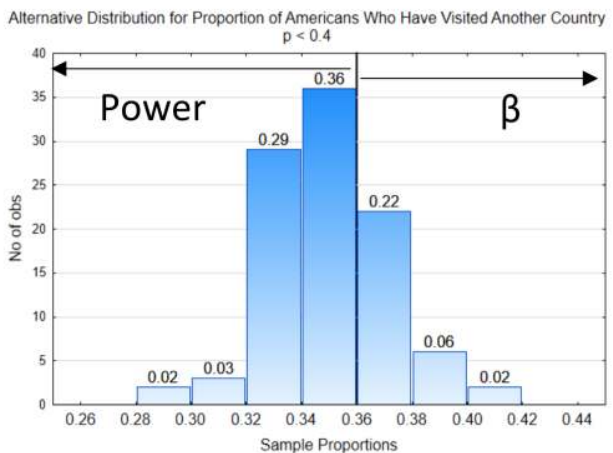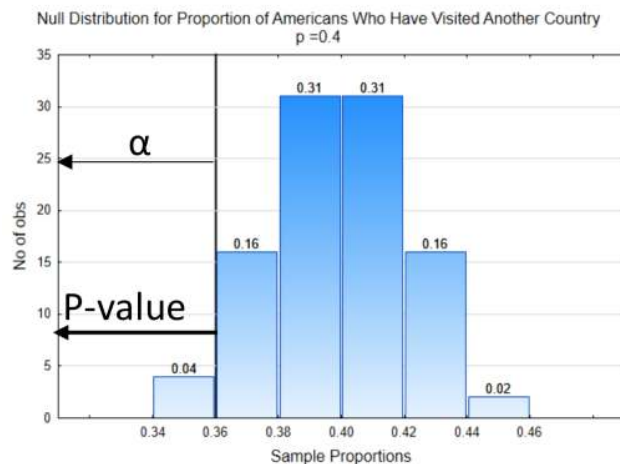
6. Make a decision. The probability for the p-value is 0.04. To determine which hypothesis is supported by the data, we compare the p-value to alpha. If the p-value is less than or equal to alpha, the evidence supports the alternative hypothesis. In this case, the p-value of 0.04 equals alpha which is also 0.04, so this evidence supports the alternative hypothesis leading to the conclusion that the proportion of people who have visited another country is less than 40%.

7. Errors and their consequence. While this problem is not serious enough to have consequences that matter, we will, nevertheless, explore the consequences of the various errors that could be made.

Because the evidence supported the alternative hypothesis, we have the possibility of making a type I error. If we did make a type I error it would mean that we think fewer than 40% of Americans have visited another country, when in fact 40% have done so.

In contrast to this, if our data had been 0.38 so that our p-value was 0.20, then our results would have supported the null hypothesis and we could be making a Type II error. This error means that we would think 40% of Americans had visited another country when, in fact, the true proportion would be less than that.

8. **Reporting results**. Statistical results are reported in a sentence that indicates whether the data are significant, the alternative hypothesis, and the supporting evidence, in parentheses, which at this point include the p-value and the sample size (n).

For the example in which $\hat{p}$ = 0.36 we would write, the proportion of Americans who have visited other countries is significantly less than 0.40 (p = 0.04, n = 200).

For the example in which $\hat{p}$ = 0.38 we would write, the proportion of Americans who have visited other countries is not significantly less than 0.40 (p= 0.20, n = 200).

At this point, a brief explanation is needed about the letter p. In the study of statistics there are several words that start with the letter p and use p as a variable. The list of words includes parameters, population, proportion, sample proportion, probability, and p-value. The words parameter and population are never represented with a p. Probability is represented with notation that is similar to function notation you learned in algebra, f(x), which is read f of x. For probability, we write P(A) which is read the probability of event A. To distinguish between the use of p for proportion and p for p-value, pay attention to the location of the p. When p is used in hypotheses, such as $H_0 : p = 0.6$, $H_1 : p > 0.6$, it means the proportion of the population. When p is used in the conclusion, such as the proportion is significantly greater than 0.6 (p = 0.01, n = 200), then the p in p = 0.01 is interpreted as a p-value. If the sample proportion is given, it is represented as $\hat{p}$ = 0.64.

We will conclude this chapter with a final thought about why we are formal in the testing of hypotheses. According to Colquhoun (1971), "Most people need all the help they can get to prevent them from making fools of themselves by claiming that their favorite theory is substantiated by observations that do nothing of the sort. And the main function of that section of statistics that deals with tests of significance is to prevent people making fools of themselves". (Green, 1979).

Chapter 1 Homework

1. Identify each of the following as a parameter or statistic.

A. p is a
B. $\bar{x}$ is a

C. $\hat{p}$ is a

D. μ is a

2. Are hypotheses written about parameters or statistics? _____

3. A sampling distribution is a histogram of which of the following?

_____original data

_____possible statistics that could be obtained when sampling from a population

4. Write the hypotheses using the appropriate notation for each of the following hypotheses. Using meaningful subscripts when comparing two population parameters. For example, comparing men to women, you might use scripts of m and w, for instance $p_m = p_w$.

4a. The mean is greater than 20. H_0: H_1:

4b. The proportion is less than 0.75. H_0: H_1:

4c. The mean for Americans is different than the mean for Canadians. H_0: H_1:

4d. The proportion for Mexicans is greater than the proportion for Americans. H_0: H_1:

4e. The proportion is different than 0.45. 4f. The mean is less than 3000. H_0: H_1:

5. If the p-value is less than α,

5a. which hypothesis is supported?

5b. are the data significant?

5c. what type error could be made?

6. For each row of the table you are given a p-value and a level of significance (α). Determine whichhypothesis is supported, if the data are significant and which type error could be made. If a given p- value is not a valid p-value (because it is greater than 1), put an x in each box in the row.

p-value	α	Hypothesis H_0 or H_1	Significant or Not Significant	Error Type I or Type II
0.043	0.05			
0.32	0.05			
0.043	0.01			
0.0035	0.01			
0.043	0.10			
0.15	0.10			
5.6×10^{-6}	0.05			
7.3256	0.01			

7. For each set of information that is provided, write the concluding sentence in the form used by researchers.

7a. $H_1 : p > 0.5, n = 350$, p - value = 0.022, $\alpha = 0.05$

7b. $H_1 : p < 0.25, n = 1400$, p - value = 0.048, $\alpha = 0.01$

7c. $H_1 : \mu > 20, n = 32$, p - value = 5.6×10^{-5}, $\alpha = 0.05$

7d. $H_1 : \mu \neq 20, n = 32$, p - value = 5.6×10^{-5}, $\alpha = 0.05$

8. Test the hypotheses:

$H_0 : p = 0.5$

$H_1 : p < 0.5$

Use a 2% level of significance.
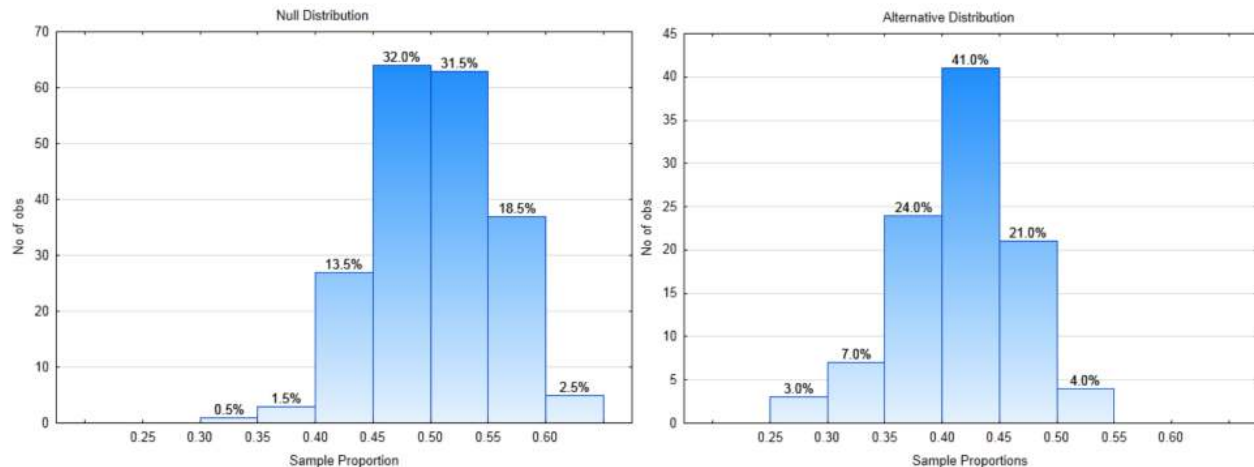
8a. What is the direction of the extreme?

8b. Label each distribution with a decision rule line. Identify α, β, and power on the appropriate distribution.

8c. What is the critical value?

8d. What is the value of α?

8e. What is the value of β?

8f. What is the value of Power?

The Data: The sample size is 80. The sample proportion is 0.45.

8g. Show the p-value on the appropriate distribution.

8h. What is the value of the p-value?

8i. Which hypothesis is supported by the data?

8j. Are the data significant?

8k. What type error could have been made?

8l. Write the concluding sentence.

9. Test the hypotheses:

$$H_0 : \mu = 300$$
$$H_0 : \mu > 300$$

Use a 3.5% level of significance.

8a. What is the direction of the extreme?

8b. Label each distribution with a decision rule line. Identify α, β, and power on the appropriate distribution.

8c. What is the critical value?

8d. What is the value of α?

8e. What is the value of β?

8f. What is the value of Power?

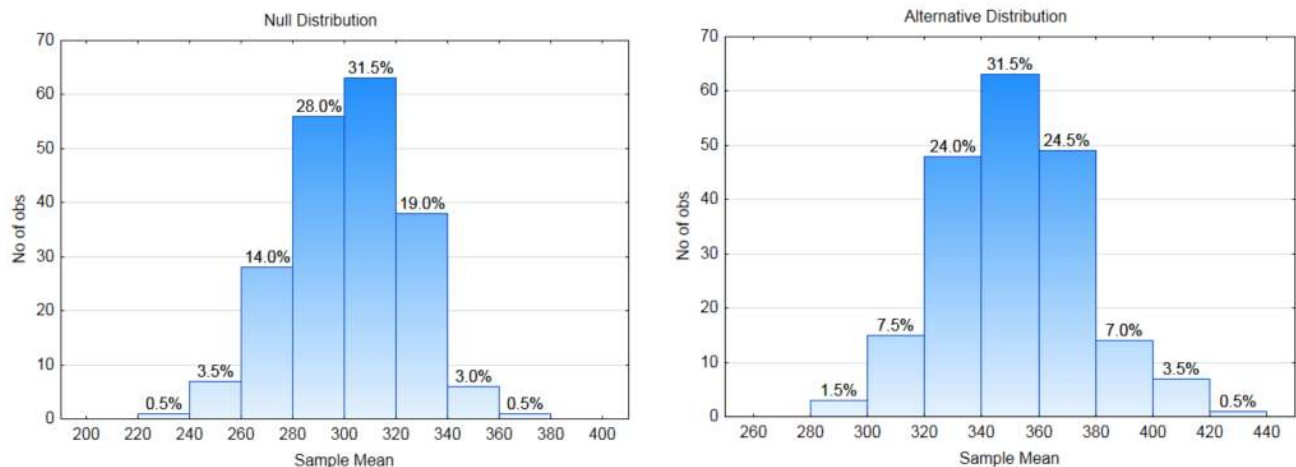The Data: The sample size is 10. The sample mean is 360.

8g. Show the p-value on the appropriate distribution.

8h. What is the value of the p-value?

8i. Which hypothesis is supported by the data?

8j. Are the data significant?

8k. What type error could have been made?

8l. Write the concluding sentence.

10. Question: Is the five-year cancer survival rate for all races improving?

> .
>
> 5 – year Cancer Survival Rate. According to the American Cancer Society, in 1974-1976 the five- year survival rate for all races was 50%. This means that 50% of the people who were diagnosed with cancer were still alive 5 years later. These people could still be undergoing treatment, could be in remission or could be disease-free. (www.cancer.org/acs/groups/con...securedpdf.pdf Viewed 5-29-13)

Study Design: To determine if the survival rates are improving, data will be gathered from people who have been diagnosed with cancer at least 5 years before the start of this study. The data that will be collected is whether the people are still alive 5 years after their diagnosis. The data will be categorical, that is the people will be put into one of two categories, survive or did not survive. Suppose the medical records of 100 people diagnosed with cancer are examined. Use a level of significance of 0.02.

10a. Write the hypotheses that would be used to show that the proportion of people who survive cancer for at least five years after diagnosis is greater than 0.5. Use the appropriate parameter.

H_0 :

H_1 :

10b. What is the direction of the extreme?

10c. Label the null and alternate sampling distributions below with the decision rule line, α, β, power.
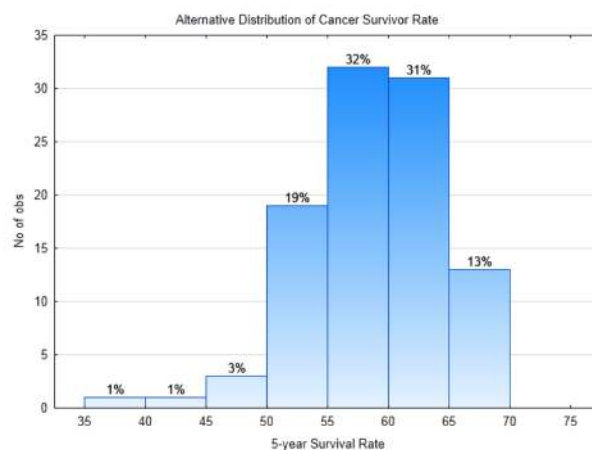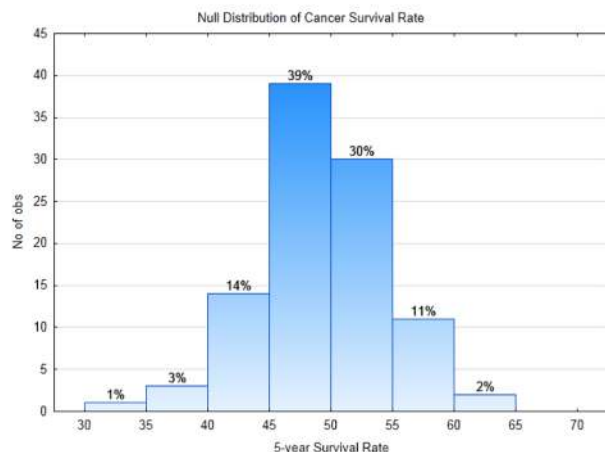
10d. What is the critical value?

10e. What is the value of α?

10f. What is the value of β?

10g. What is the value of Power?

The data: The 5-year survival rate is 65%.

10h. What is the p-value for the data?

10i. Write your conclusion in the appropriate format.

10j. What Type Error is possible?

10k. In English, explain the conclusion that can be drawn about the question.

11. Why Statistical Reasoning Is Important for a Business Student and Professional

Developed in Collaboration with Tom Phelps, Professor of Economics, Mathematics, and Statistics This topic is discussed in ECON 201, Micro Economics.

> **Briefing 1.2**
>
> Generally speaking, as the price of an item increases, there are fewer units of the item purchased. In economics terms, there is less "quantity demanded". The ratio of the percent change in quantity demanded to the percent change in price is called price elasticity of demand. The formula is $e_d = \dfrac{\%\Delta Q_d}{\%\Delta P}$. For example, if a 1% price increase resulted in a 1.5% decrease in the quantity demanded, the price elasticity is $e_d = \dfrac{-1.5\%}{1\%} = -1.5$. It is common for economists to use the absolute value of e_d since almost all e_d values are negative. Elasticity is a unit-less number called an elasticity coefficient.
>
> Food is an item that is essential, so demand will always exist, however eating out, which is more expensive than eating in, is not as essential. The average price elasticity of demand for food for the home is 0.51. This means that a 1% price increase results in a 0.51% decrease in quantity demanded. Because eating at home is less expensive than eating in restaurants, it would not be unreasonable to assume that as prices increase, people would eat out less often. If this is the case, we would expect that the price elasticity of demand for eating out would be greater than for eating at home. Test the hypothesis that the mean elasticity for food away from home is higher than for food at home, meaning that changing prices have a greater impact on eating out. (www.ncbi.nlm.nih.gov/pmc/articles/PMC2804646/) (www.ncbi.nlm.nih.gov/pmc/arti...46/table/tbl1/)

11a. Write the hypotheses that would be used to show that the mean elasticity for food away from home is greater than 0.51. Use a level of significance of 7%.

H_0:

H_1:

Null Distribution of Elasticity of Demand
Mean = 0.51

Alternative Distribution of Elasticity of Demand
Food Away From Home
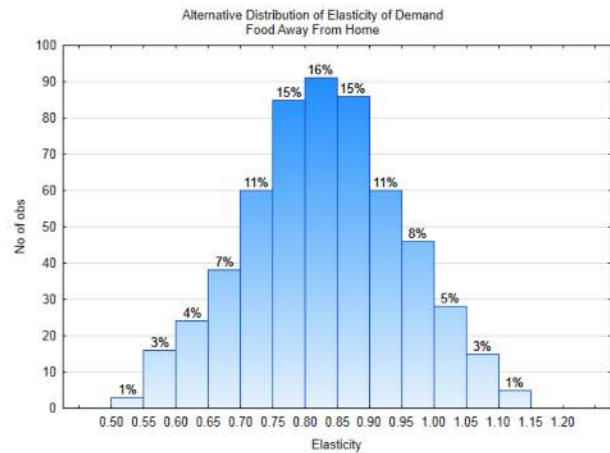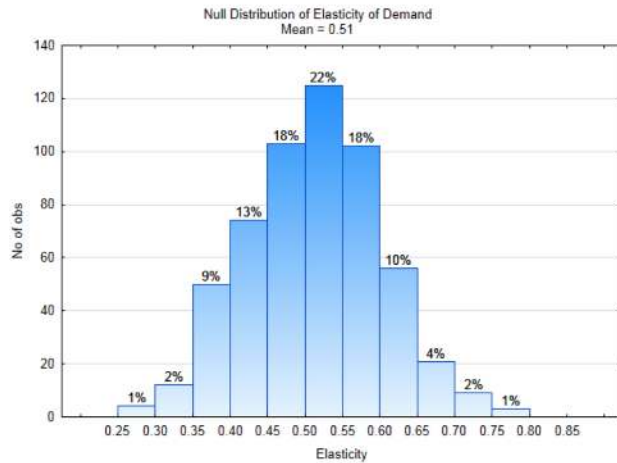
11b. Label each distribution with the decision rule line. Identify α, β, and power on the appropriate distribution.

11c. What is the direction of the extreme?

11d. What is the value of α?

11e. What is the value of β?

11f. What is the value of Power?

The Data: A sample of 13 restaurants had a mean elasticity of 0.80.

11g. Show the p-value on the appropriate distribution.

11h. What is the value of the p-value?

11i. Which hypothesis is supported by the data?

11j. Are the data significant?

11k. What type error could have been made?

11l. Write the concluding sentence.

2: Obtaining Useful Evidence

A primary role of statistics is to use evidence from stochastic populations to improve our understanding of the world. Deciding what evidence will be collected is an essential part of the process. **Research design** is that portion of the statistical process in which planning is done so that the conclusions are drawn with confidence and can be supported under scrutiny.

There are three research designs we will explore in this chapter, observational studies, observational experiments, and manipulative experiments. The type of research that is conducted is dependent upon the objective of the research. In cases where the objective is to understand a population or compare populations, an observational studies is appropriate. In cases in which we want to determine if a causal relationship exists between two variables, we conduct an experiment. A **causal relationship** (cause and effect relationships) implies the existence of two variables. The variable that is the cause must happen first. The first variable is called an explanatory variable, the variable that is affected is a response variable.

In experiments, the explanatory variable is a treatment or intervention that is imposed upon people or elements of a population. Of the two experiments, observational and manipulative, the latter is better for showing a causal relationship. In manipulative experiments the researcher can randomly assign the treatment or intervention whereas for observational experiments, the treatment or intervention is imposed by someone other than the researcher.

Before clarifying each of these research designs, a few examples might be useful.

Examples of Observational Studies

- A researcher might conduct a survey of Americans to compare the proportion of Democrats who support efforts at reducing carbon emissions to the proportion of Republicans who want to reduce carbon emissions.
- Water samples could be taken in the Puget Sound to determine the level of PCB contamination.
- Students could be given an unannounced exam on a math skill they had learned earlier in the school year to see how much they retained.

Example of Observation experiments

- Since some states have legalized the recreational use of marijuana, it is possible to determine if it really a gateway drug by seeing if there is a change in the usage of harder drugs.
- When some states increase the minimum wage, it is possible to determine if raising the minimum wage has an effect on the number of people who are employed in the state by comparing them with states that don't raise the minimum wage.
- When a natural disaster strikes an area, it is possible to determine the effect on donations to organizations such as the Red Cross.

Example of Manipulative experiments

- A coach randomly assigns some runners to a weight training program and does not allow other runners to lift weights, but otherwise all runners have the same training program, then the coach can determine the effect of weight training on running improvement.
- Loaves of homemade bread can be baked at different temperatures to determine the effect of temperature on the bread.
- A company can try different internet ads to see if there is an effect on sales of their product.

Randomness

Each research design incorporates one application of randomness. In the case of observational studies and observational experiments, a **random selection** is made from the population. This may be difficult for some observational experiments. For example, there are not enough states that have legalized marijuana to randomly select from. In manipulative experiments, the researcher **randomly assigns** participants to different groups, for example, to groups receiving a treatment and those not receiving it. The methods used for random selection and random assignment are discussed later in this chapter.

Distinguishing between research designs

It can be challenging to determine which research design is being used. The following questions can guide your decision.

1. Is the researcher looking for the effect of a treatment or intervention?

2. If the answer to the first question is yes, then can the researcher randomly assign participants to different groups?

The flow chart that follows uses these questions to determine the type of research design.

Types-of-Research Flow Chart
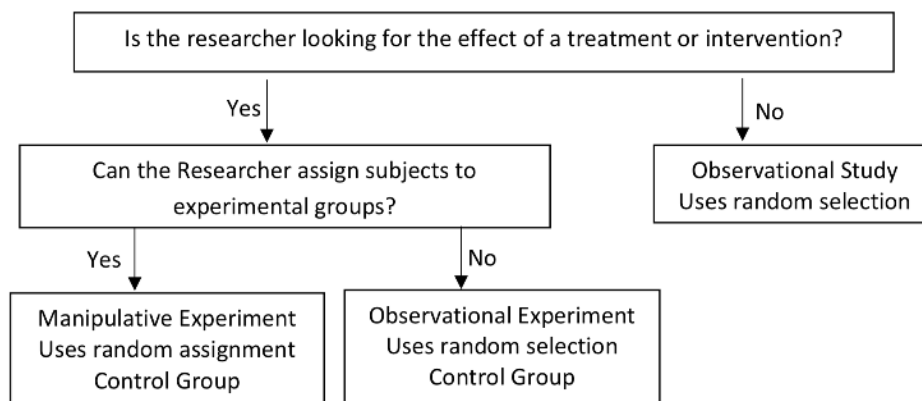
Experiments

Policy makers, businesses managers, physicians, educators, scientists, and coaches typically have an outcome they would like to achieve, but they want to make an evidence-based decision in order to achieve the outcome. That is, they want to know what variable they can change so the change has an effect on a different variable. For example,

A policy maker may wonder what variable should be changed to reduce poverty.

A business manager may wonder what advertising strategy will lead to the greatest increase in sales.

A physician may wonder which medicine will cure a person.

An educator may wonder which teaching strategy will lead to the greatest amount of learning for the students.

For causal relationships, it has already been stated that a cause must proceed an effect, but there is another criterion of importance. In a causal relationship, a treatment produces a particular outcome while not providing the treatment means that particular outcome is not produced. Thus, simply showing that a certain response occurred when a treatment was provided does not prove the treatment caused the response. There could be another factor that caused that particular response. To prove causation, the research should be designed to show that one response occurs with the treatment and does not occur without the treatment and that there is unlikely to be another variable that is causing the response. This requires having at least two groups, one (or more) which receives the treatment and one that does not receive the treatment. The group that does not receive the treatment is called a **control group**.

When experiments are conducted on non-humans, it is possible to have a control group that does not receive any treatment. An example would be agriculture researchers who might fertilize some crops but not others. However, when an experiment is conducted on humans, there can be complications. In typical experiments involving the testing of new medicines on a person with an illness, it is not sufficient to simply give some people the medicine being tested and not give it to others. Humans can have psychosomatic effects – physical changes that are a result of the expectations of a certain effect of the medicine, attributed to the mind-body interactions. To address this problem, it is customary to give an inert medicine, called a placebo, to some of the participants. It is important that the subjects do not know if they are receiving the real medicine or the placebo. It is also important that the researcher examining the subjects doesn't know either. This is achieved by doing a **double blind** experiment. In this type of experiment, subjects are randomly assigned to either the treatment group or the placebo group, but are not told which group they are in. The doctor is not told either.

A problem has been observed with this type of double blind experiment however. That problem is called **breaking blind** and is caused because subjects have to be warned about possible side effects from the medication. Consequently, those experiencing the side effects can guess they are taking the actual medicine and those that don't experience them conclude they are taking the placebo. In some experiments, more than 80% of the doctors and subjects correctly identified if a subject was in the treatment group or placebo group. Since a correct guess about the group should occur about 50% of the time, it is likely that side effects, or possibly other clues, led to the higher value of correct identification. To help minimize this problem, some researchers use an active

placebo instead of the more typical sugar pill. An active placebo produces side effects similar to the real medicine, but does not provide a cure for the medical condition.("Listening to Prozac, But Hearing Placebo." *The Emperor's New Drugs: Exploding The Antidepressant Myth*. Philadelphia: Basic Books, 2010. 7-20. Print.)

In medical studies, besides having a treatment group and a placebo group, it is appropriate to have a control group that receives no treatment at all. This is often accomplished because some people who apply to be in the experiment are not accepted. Because illnesses can go through cycles (good days, bad days) and people usually wait until they feel very bad to get treatment, then comparing the results of treatment to people who don't receive any treatment can be helpful to show if something other than the normal cycle of symptoms is occurring as a result of the treatment.

Response variables, explanatory variables, levels, and confounding

Response and explanatory variables will be explained using teachers as an example. An ideal outcome for a teacher would be for the entire class of students to be successful in the class. The teacher would like to know which teaching strategies (pedagogy) will lead to the greatest success for the students. Notice in this example, there are two variables, teaching strategy and student success. Since teaching must come before assessment of student success, then teaching strategy is the explanatory variable and student success is the response variable.

The response variable is rather vague however. What does student success mean? There are many aspects of learning, such as memorization of facts, ability to calculate, skills in the laboratory, writing skills, ability to think critically, ability to think creatively, etc. A researcher needs to be clear about the response variable. For example, since this book is used for a statistics class, then one outcome of particular interest is whether students can correctly test a hypothesis. A different outcome might be whether the students can create appropriate graphs for the data.

There are many possibilities for the explanatory variable of teaching strategies. These possibilities are called **levels**. Examples of levels include lecturing, active learning, discovery learning, computer teaching software, etc. Levels are specific examples of the explanatory variable.

While teaching pedagogy is an explanatory variable a teacher can modify, it is not the only variable that can affect the response variable of student success. Other variables include student interest and motivation, the text, study time, distractions (lack of food or shelter, deployment of a spouse, divorce, illness, etc). These other variables, which could be used as explanatory variables in different research, are called **confounding variables**. Potential confounding variables should be identified during the research design stage so they can be controlled in the experiment by making sure that they are equally distributed in the different experimental groups.

To get practice identifying the different elements of research, you will be given stories that explain a research project. From the story, your object is to identify the key elements, including the research question, the variables, parameter, and type of research. These will be organized in a research design table. When completing this table, think of the potential confounding variables yourself as they are not usually included in the story.

Research Design Table	
Research Question:	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated	Mean Proportion
List potential confounding variables.	
Grouping/Explanatory Variables 1 (if present)	Levels:

Some of the examples below contain underlined words, others do not. The purpose of underlining is to help you identify the key words in the story. Ultimately, you need to identify these parts without them being underlined.

Example 2.1 Is there a difference in the number of electronic items in the homes of people who were born and raised in the US compared to people who immigrated to the US and have lived in the US for at least 5 years?

To answer this question, the residency status of people will be classified as native or 5-year immigrant. A random samples of native residents and immigrants who have been in the US at least 5 years will be taken. All electronic items will be counted individually (e.g. cell phones, computers, TVs, radios). The objective is to determine if the mean number of electronic items is different for the two groups.

Research Design Table	
Research Question: Is there a difference in the number of electronic items in the homes of people who were born and raised in the US compared to people who immigrated to the US and have lived in the US for at least 5 years?	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	number of electronic items
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	Income, wealth, age, size of family
Grouping/explanatory Variables (if present) Residency status	Levels: native 5-year immigrant

Briefing 2.1 Bare Foot Running

In 2011, Vintage Books published the book "Born to Run: A hidden Tribe, Superathletes, and the Greatest Race the World Has Never Seen" by Christopher McDougall. One of the topics discussed was the concept of bare-foot running. The author argued that running bare foot (or with minimal protection between the sole of the foot and the ground) leads to a forefoot running style that leads to fewer injuries than those who run with padded shoes and use a heal strike.

A high school running coach would like to know if new runners using minimalist shoes that lead to a forefoot running style will have fewer injuries than new runners using padded shoes that lead to heal strikes. The coach uses a coin flip to randomly assign the type of shoe a new runner should wear. The coach will record this shoe choice and maintain an injury record for each athlete. Ultimately, the coach will determine if there is a difference in the proportion of runners from each group who are injured. Only new runners will be included because it would be difficult and perhaps inappropriate to change the running style of experienced runners.

Research Design Table	
Research Question: Does running style make a difference in injuries?	
Type of Research	Observational Study bservational Experiment Manipulative Experiment
What is the response variable?	injury
What is the parameter that will be calculated?	Mean **Proportion**
List potential confounding variables.	Prior running experience, prior injuries, general fitness
Grouping/explanatory Variables (if present) Type of shoe	Levels: minimalist shoes Padded shoes

Example 2.3 Will raising the tax rate for the wealthy solve the national debt problem?

Every time a law is changed the country conducts an experiment. One would assume that lawmakers reflect carefully about the possible consequences of any change in law they approve. The country is now faced with a large national debt that has some lawmakers concerned and which occasionally attracts the interest of investors. There is also an ideological debate that persists about the benefits or consequences of raising taxes or cutting spending. A popular recommendation of some is to raise taxes on the wealthy.

Briefing 2.2 Marginal Tax Rate

Tax brackets are used to show the amount of tax paid for each dollar earned. The marginal tax rates for 2013 are shown in the table on the next page for people who are married filing jointly. (taxfoundation.org/article/us-...usted-brackets viewed 7/08/13)

A person earning $80,000 would pay 10% tax on the first $17,488, 15% tax on the money between $17,488 and $71,030, and 25% on the amount over $71,030.

The graph below shows the change in the marginal tax rate on the wealthiest Americans and the National Debt. From this graph we see that the national debt started rising a lot in the late 1970s and 1980s. We also notice this rise was preceded by big drops in the top marginal tax rates during the Reagan administration.

Marginal Tax Rate	Married Filing Jointly Tax Brackets Over	But Not Over
10.0%	$0	$17,488
15.0%	$17,488	$71,030
25.0%	$71,030	$143,432
28.0%	$143,432	$218,528
33.0%	$218,528	$390,273
35.0%	$390,273	-
39.6%	$440,876	-

Nominal National Debt and Top Marginal Tax Rate

— Gross Public Debt-fed $billion(L)
---- Top Marginal Tax Rate(R)
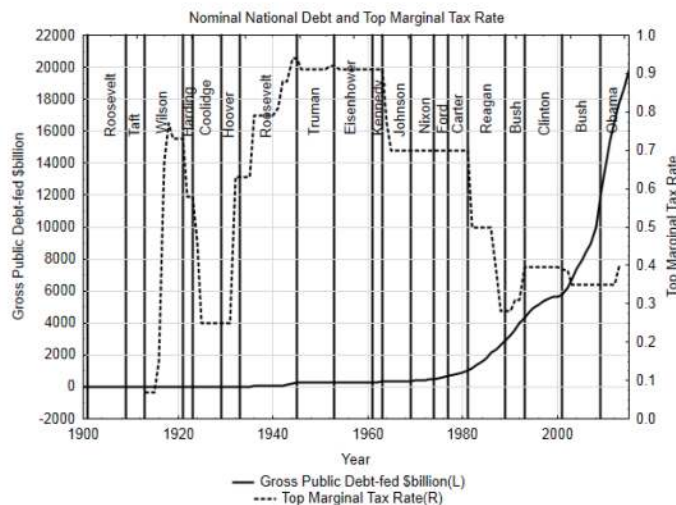
The information from this graph cannot be used to test the theory that lowering tax rates leads to increased national debt (or vice versa) because ***theories cannot be proved with the evidence that was used to create the theory in the first place***. Therefore, if an economist wanted to test the theory about the effect of marginal tax rates on national debt, they will need to get different data. It would be unrealistic to expect that any country would agree to participate in an experiment in which a research economist would make them change their tax rates. However, countries do change tax rates on their own, so a researcher could observe what happens after each such change. The national debt before the rate change and 5 years after could be determined. If the goal is to establish a cause-and -effect relationship, it will also be necessary to identify changes in national debt for countries that do not change their tax rates. A comparison of the changes in national debt could be made for both groups. Other important aspects of national debt include the amount of spending that is done as well as the amount of concern legislators have with keeping the budget balanced.

Research Design Table	
Research Question: Does lowering tax rates lead to an increase in national debt?	
Type of Research	Observational Study <u>Observational Experiment</u> Manipulative Experiment
What is the response variable?	Changes in national debt 5 years after rate change

What is the parameter that will be calculated?	<u>Mean</u> Proportion
List potential confounding variables.	State of the economy, number of people at each economic level, government budget priorities
Grouping/explanatory Variables (if present) Marginal tax rate change	Levels: Control (did not change tax rates) Impact (reduced tax rates)

Sampling

Observational studies and some observational experiments require random sampling from a population. The next step in the research design process is to determine how a sample will be taken from the population so that it is representative of the population. The objective is to avoid bias. **Bias** is systematic prejudice in one direction. Recall the sampling distributions that were discussed in Chapter 1. Half the statistics in the sampling distribution were less than the parameter and half were more, thus the probability of getting a statistic higher or lower than the parameter was the same. If sampling is not done correctly, it is easily possible to end up with biased results. That means that samples are more likely to be less than the parameter or they are more likely to be greater than the parameter. For example, if you want to determine which sport people think is the most exciting, pro football or pro soccer and you only sample people in a city with an NFL team, you are likely to get biased results in favor of pro football. On the other hand, if you conduct your survey in a city such as London, you are likely to get biased results in favor of soccer. Either way, you are getting biased results, which means any conclusion you draw is not valid.

Biased results are obtained when doing voluntary sampling and convenience sampling. Voluntary sampling occurs when people voluntarily agree to participate in a survey, such as an online survey or a TV survey where people can text their response. Convenience sampling occurs because of getting responses from people who are convenient. It is possible that these people share an opinion and consequently group together, resulting in biased results.

The best sampling is achieved using probability sampling methods. The four methods that will be discussed are:

1. Simple Random Sampling
2. Stratified Sampling
3. Systematic Sampling
4. Cluster Sampling

Simple Random Sampling

Simple random sampling meets two desirable criteria. First, every individual or unit in the population has an equal chance of being selected and second, every collection of selected units has an equal chance of being selected. The sampling distributions that underlie testing of hypotheses are based on simple random sampling with replacement. That means that once selected, a unit is put back into the pool and can be selected again. Consequently, information from the same unit can be used more than once.

The simplest example of a simple random sample is pulling names out of a hat. That is, everyone in a group can have their name written on a piece of paper and then put into a hat or other container. Someone mixes the pieces of paper and then pulls out a name. This is much like raffles that are done by organizations.

Putting names on a piece of paper quickly becomes unmanageable with larger populations and so a different strategy is needed. Instead, each person or unit is given a number and then numbers are selected. Data is then gathered from the person or unit with the selected number. Three different methods will be provided for doing a simple random sample. These methods make use of a table of random digits, the TI83 or TI84 calculator, and the website called Random.Org. The first two methods are known as pseudo-random meaning that while a random process is used to generate the numbers, it is a repeatable process. They will be explained below. The random numbers generated at Random.Org are truly random as they are based on atmospheric noise. Visit the website and select integer generator to try their selection process.

Table of Random Digits

A table of random digits consists of the digits 0 – 9 that have been randomly selected, with replacement. They are grouped with 5 digits together for visual convenience. Rows and columns are numbered.

To use the table, determine the size of the population from which a sample will be drawn. Assign a number to each person or unit in the population. The easiest way to do this is to assign a 1 to the first person (unit), a 2 to the second person (unit), etc. However,

this is not the only strategy. People or units may already have a number (e.g. student ID number, production number), which can be used. The number of digits that will be selected at the same time corresponds to the number of digits in the largest assigned number. If the selection is to be done from a population of size 89 units, then since 89 is a 2-digit number, then assigned numbers will be 01, 02, ... 89 and all selections will be 2 digits. If the size of the population is 745, since this is a 3-digit number then the assigned numbers will be 001, 002, ... 745.

Table of Random Digits

Row	Col 1-5	Col 6-10	Col 11-15	Col 16-20	Col 21-25	Col 26-30	Col 31-35	Col 36-40
1	05902	75968	00100	12330	92481	64625	83012	90763
2	53365	25560	86425	45946	67093	36638	71740	16878
3	69363	06820	49676	25363	96300	94376	65819	19636
4	37520	54955	31507	70745	41817	86606	97766	44989
5	10390	12738	54072	03238	08294	89479	03156	24217
6	98735	90798	96609	18368	74876	17403	33783	85101
7	79609	87687	77178	39784	76983	05689	84023	24804
8	00348	58777	90570	09114	99677	08126	76132	19334
9	98367	93351	08246	81492	57876	04366	21851	28620
10	34588	88493	61188	29234	32565	82010	07425	37173
11	74198	34943	64557	20118	25540	50014	29338	87231
12	00621	86824	81204	71923	03600	69080	31712	36599
13	44684	53902	86099	98640	86347	88061	60420	54118
14	43526	09310	21922	40743	64742	12780	88432	41496
15	37335	98934	61403	85336	76356	22349	31498	34136
16	25488	41567	32833	56973	04039	57733	88677	44817
17	45327	69347	85698	03248	60079	64469	71406	19478
18	47458	08093	94256	14305	42728	676159	35991	13527
19	91622	23621	91124	08233	54571	73527	29012	31534
20	77630	37356	85498	21296	14880	24981	70976	64922

For example, what will be the numbers of the first 3 people that would be selected from a population with 6890 people? People are assigned numbers such as 0001, 0002, ... 6890. The selection will begin in row 16, which is reproduced below. Four digits will be selected in a row. If they are less than or equal to 6890 they will be selected (shown with underlining). If they are larger than 6890, they will be ignored.

16	25488	41567	32833	56973	04039	57733	88677	44817

The first three numbers that are selected are 2548, 6732 and 0403.

> The Texas Instrument TI84 calculator is able to generate random integers. A process that is analogous to picking a row in a table of random digits is to seed the calculator. The calculator is seeded and then random integers are selected. For example, if the seed number is 38, then the key strokes on the calculator would be:
>
> 38 sto math prb 1 rand enter. 38 should appear on the screen.

> To generate the random number, the key strokes are:
>
> math prb 5 randint, enter. The function randint expects the input of three numbers, the low, the high and the number of values you think will fit within the screen window. If we continue with the example of 6890 people, then since this is a 4 digit number we might expect 3 such numbers to fit on the screen, so we would enter: randint(1,6890,3). If we need more than 3 numbers, then we can just push enter again as often as is necessary.

The numbers that are selected in this example are: 2283, 3612, 3884.

Stratified Sampling

There are times when parts of the population might be expected to produce different data than other parts. For example, it might be expected that the concentration of a toxic chemical in the Puget Sound would be higher near industrial areas than in locations that are far from those industrial areas. Since random sampling may result in areas being missed, then a stratified sampling can be done. In this approach, areas are defined, with each area being a stratum. A simple random sampling process is then used within each stratum.

As a separate example, a group seeking to expand public transportation in a state may wonder how much support there would be for an initiative. They might expect the support for public transportation will be substantially different for people who use public transportation than for people who never use it. Consequently, they may do a simple random sampling from each of these groups.

It should be noticed that stratifying is based on the assumption that there will be differences between the strata although this may not be something that has been proved. This is different than actually having a hypothesis about the difference between the strata, in which case each stratum is considered to be a different population, rather than different parts of the same population.

Systematic Sampling

A sampling strategy this is particularly useful for sampling time series data is systematic sampling. This is a 1 in k sampling method in which every k^{th} unit is selected. Since the value of data in one year may be influenced by the previous year (or more), the data are not independent. For example, this year's cost of tuition is closely related to last year's cost. Suppose that a sample is to be taken from time series data that is serially dependent when the data are 1,2 and 3 years separated but not when they are separated by 4 years. In this case, sampling every $4t^{th}$ year would be appropriate. Suppose that data is available from 1961 to the present and a 1 in 4 systematic sampling method is used. What will be the first year in which data are selected? Since every year has to have a chance of being selected, then it will be necessary to randomize the initial value. This will be done by randomly selecting one number between one and k. To find successive numbers, add k to the number selected. For example, if a TI84 calculator is seeded with the number 42, then randint(1961,1964,1) will produce the number 1962. To this will be added 4 repeatedly until a sample of the desired size has been selected. The table below shows the years that will be selected.

1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
1971	1972	1973	1974	1975	1976	1977	1978	1979	1980

The value of k is dependent upon the size of the population (N) and the size of the sample (n)and is found by dividing the former by the latter: $K \approx N/n$.

Cluster Sampling

Collecting data can be time consuming and expensive, neither of which is a trivial factor for any organization that needs the data. When data must be collected from different locations and there is not an assumption that the locations will cause the variation in the data, then cluster sampling can be used. For example, a community college may want to sample the student body about charging students a technology fee so that a new student computer lab can be built. Because students take many different classes and these classes are not likely to have a major impact on their preference about the fees, then different classes can be selected and all the students in those classes can be asked their preference on the fees. If a college has 450 classes, they can be numbered from 1 to 450 and a simple random sampling process can be used to select the desired number of classes. If the goal is to select 8 classes and a seed value of 16 is use, then on the TI84, the function Randint(1,450,4) will give the class numbers of 419, 313, 273, 229, 445, 162, 127, 428.

These methods are often confused. The following guidelines may help clarify the differences. Simple Random Sample – Random Sampling is done from the entire population.

Stratified Sampling – The entire population is divided into strata then simple random sampling is done from each strata. The samples from each strata are combined before being analyzed.

Systematic Sampling – One number is randomly selected from the first k numbers. The numbers of the other data are found by adding k to the last number that was selected.

Cluster Sampling – The entire population is divided into groups or clusters, which are given numbers. The groups are randomly selected and every unit within the group becomes part of the sample.

Chapter 2 Homework

Complete the design-layout tables. Use underlined words when available.

1. A student would like to know which of two possible routes is faster for the daily trips to school. Route 1 is shorter but has many traffic lights. Route 2 is a little longer but doesn't have traffic lights. Each morning, a coin flip will be used to determine the route taken to school. The time it takes for the commute will be measured with a stopwatch. After approximately 15 trials on each route, the average time for each will be compared.

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

2. Suppose researchers wanted to know if the opinion people had about the future was influenced by the amount of news they consume (watched, listened to, or read). The researchers categorized news consumption into three categories: 5-7 days/week, 1-4 days/week, less than 1 day/week. They then asked the people their opinion of the future (if they expected the future to be better or worse than the present). They will compare the proportion of optimistic people in each group.

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

3. Because so many species are becoming extinct, scientists would like to know how to increase biodiversity. There are two approaches to improve biodiversity in the world. The hands-off approach is one in which no one makes any deliberate changes to the environment with the intent of improving biodiversity. The deliberate approach is to deliberately introduce species that will reshape the environment, using surrogate species when necessary (e.g. use elephants instead of woolly mammoths, which are extinct). Examples of the first approach include the DMZ between North and South Korea. An example of the second includes the creation of a Pleistocene park in northeast Siberia by ecologist Sergei Zimov. Whether they occur by accident or design, there is no central planning organization that will randomly determine the approach that will be taken, so researchers can only look at the evidence after ecosystems have been engineered. A comparison will also be made with similar areas

(control groups) that do not receive either the hands-off or deliberate approach. The researchers might record data on the increase in the number of species and determine if the average increase in number of species is different for the two approaches and the control groups. (Brand, Stewart. *Whole Earth Discipline: An Ecopragmatist Manifesto*. New York: Viking, 2009. Print.)
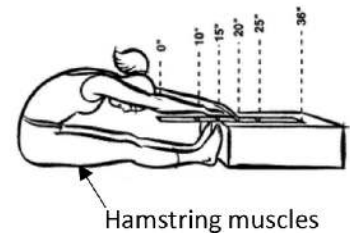
Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

4. a. It has been hypothesized that a lack of flexibility of the hamstring muscles can contribute to poor posture. To determine if that is the case, a group of adults was randomly selected. The group was divided into two, those with good posture and those with poor posture. The flexibility of their hamstrings was measured using a sit and reach test.(http://silbergen564s15.weebly.com/. Viewed 4/8/2017) The further a person can reach, the greater their hamstring flexibility.

Hamstring muscles

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

b. Two types of stretching can be done to improve flexibility, static stretching and dynamic stretching. Static stretching involves stretching a muscle and holding it in a stretched position for about 30 sec. Dynamic stretching involves stretching while moving through a range of motion. To determine which type of stretching resulted in improvement, the group of people with poor hamstring flexibility were randomly assigned to one of three groups. One group did static stretching daily for one month. Once group did dynamic stretching daily for one month. The third group was the control group, which did not do any stretching. Afterwards, the subjects were retested and categorized as improving or not improving since their first test.

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment

What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

5. Researchers want to know the proportion of acres of forest in the state that show evidence of the brown beetle infestation.

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

6. A teacher wants to know the mean amount of time community college students spend doing homework each night.

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

7. A fisheries biologist want to know the average weight of Coho Salmon returning to spawn.

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

1. In Chapter 2, you were introduced to sampling distributions. Understanding these distributions proves challenging for many but since they form the bases upon which p-values are determined and therefore conclusions are drawn, knowing how the distributions are created and what they mean is helpful for your understanding of statistics. Sampling distributions are really

theoretical in nature because they would be extremely difficult to make in reality, but having the experience of partially making one should give you greater insight into what one would really be like. In this problem, you are given a set of data, which is considered the entire population. Each data value has been numbered. You will then practice the various sampling methods multiple times, using different seed values. In each case, you will determine the statistic of the sample, which in this case will be a sample proportion. You will then fill in one box in the distribution that is provided. The first box you put in should be considered the one and only sample that you would have taken. Use a different color for shading the box. The remaining samples you take will represent other possible samples that you would have gotten with a different seed number.

The population consists of all the berths in a harbor. Each dock has room for 20 boats. In this problem, each cluster is a different dock. The two strata are the west side of the harbor and the east side. Yes means there is a boat at the berth, no means that it is vacant.

West Side of Harbor				East Side of Harbor		
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
1 Yes	21 No	41 Yes	61 No	81	101	121
2 Yes	22 No	42 No	62 No	82 Yes	102	122 Yes
3 No	23 No	43 No	63 No	83	103 Yes	123
4 Yes	24 No	44 Yes	64 No	84	104 Yes	124
5 No	25 No	45 No	65 No	85 Yes	105 Yes	125 Yes
6 Yes	26 Yes	46 Yes	66 No	86 Yes	106 No	126 No
7 No	27 No	47 No	67 Yes	87 No	107 No	127 Yes
8 Yes	28 No	48 No	68 No	88 No	108 Yes	128 No
9 No	29 No	49 No	69 No	89 Yes	109 No	129 Yes
10 No	30 No	50 No	70 Yes	90 No	110 No	130 No
11 No	31 No	51 Yes	71 No	91 No	111 No	131 No
12 Yes	32 No	52 Yes	72 Yes	92 Yes	112 Yes	132 Yes
13 Yes	33 No	53 Yes	73 Yes	93 No	113 No	133 No
14 Yes	34 No	54 Yes	74 Yes	94 Yes	114 No	134 Yes
15	35 Yes	55 No	75 Yes	95 No	115 No	135 No
16	36 Yes	56 No	76 Yes	96 No	116 No	136 No
17 Yes	37 Yes	57 No	77 No	97 Yes	117 Yes	137 Yes
18 Yes	38 Yes	58 No	78 Yes	98 No	118 Yes	138 No
19 No	39 No	59 Yes	79 No	99 No	119 Yes	139 Yes
20 No	40 No	60 No	80 No	100 No	120 Yes	140 Yes

For each sampling method, 20 samples will be taken. Sample with replacement, which means the same number can be selected more than once. Determine the proportion of samples that are Yes. On each line, write the number selected and a Y for yes or N for no (e.g. 8Y)

8. a. Use a simple random sample. The seed number for what will be considered the official sample is 5.

_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,

Sample Proportion _____

The following are alternate sample results you could get if you had used different sampling methods and seed numbers.

b. Use a stratified sample with a seed number of 10 for the West and 11 for the East.
West _____, _____, _____, _____, _____,_____, _____, _____, _____, _____,_____,
East _____, _____, _____, _____, _____,_____, _____, _____, _____

Sample Proportion _____

c. Use systematic sampling with a seed number of 15. Let k = 7.

_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
Sample Proportion _____

d. Use a cluster sampling method with a seed number of 20.
Which cluster is selected? _____ **Sample Proportion _____**8e. Use a simple random sample with a seed number of 25.

_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
Sample Proportion _____

f. Use a stratified sample with a seed number of 30 for the West and 31 for the East.
West _____, _____, _____, _____, _____,_____, _____, _____, _____, _____,_____,
East _____, _____, _____, _____, _____,_____, _____, _____, _____
Sample Proportion _____

g. Use systematic sampling with a seed number of 35. Let k = 7.

_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
Sample Proportion _____

h. Use a cluster sampling method with a seed number of 40.
Which cluster is selected? _____ **Sample Proportion _____**

i. Use a simple random sample with a seed number of 45.

_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
Sample Proportion _____

j. Use a stratified sample with a seed number of 50 for the West and 51 for the East.
West _____, _____, _____, _____, _____,_____, _____, _____, _____, _____,_____,
East _____, _____, _____, _____, _____,_____, _____, _____, _____

Sample Proportion _____

k. Use a cluster sampling method with a seed number of 55. Which cluster is selected? _____ **Sample Proportion _____**

l. Use systematic sampling with a seed number of 60. Let k = 7.

_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
_____, _____, _____, _____, _____, _____, _____, _____, _____, _____,
Sample Proportion _____

m. Fill in a square in the appropriate column, starting at the bottom row (that does not contain the numbers). The first sample proportion you get (from problem 8a) should be shaded differently than the rest of the sample proportions.

0.00					
0.05					
0.10					
0.15					
0.20					
0.25					
0.30					
0.35					
0.40					
0.45					
0.50					
0.55					
0.60					
0.65					
0.70					
0.75					
0.80					
0.85					
0.90					
0.95					
1.00					

n. Find the parameter by finding the proportion of all the 140 responses that are yes. Show this on the chart in 8m. How do the sample proportions compare to the population proportion?

9. The first graph shows the change in employment when the Federal minimum wage has been increased. This graph shows a comparison in the number of people employed 6 months after the increase, compared to six months before the increase. The numbers on the x-axis represent millions of people (e.g. 1000 x 1000) with positive numbers reflecting an increase in employment. Notice that most of the time, minimum wage went up, so did employment. However, this graph does not provide solid evidence that raising the minimum wage leads to an increase in employment. This is because there is no comparison. It could be that jobs were increasing or decreasing anyway, because of bigger economic changes, and that the minimum wage had only minor effect.

A better way to determine the effect of raising the minimum wage is to compare states that raise it with states that don't since states have the ability to raise the minimum wage above the Federal level. The average after – before change in annual unemployment can be compared between these groups of states. For example, if the minimum wage in a state is increased in 2003, then the unemployment rate in 2002 can be subtracted from the unemployment rate in 2003. If the 2003 rate is lower than the 2002 rate, it means the unemployment rate went down and the difference would be a negative number. (Note: while the graph above was about the number of employed people, the graphs that follow are about the number of unemployed people).

a. Complete the Research Design table.

Research Design Table

Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

Hypotheses and the level of significance are to be established before data is collected. The hypotheses for this question are that the average after-before difference in annual unemployment rates is different in the states that raise their minimum wage compared to states that don't.

$H_0 : \mu_{\text{Raise}} = \mu_{\text{Not Raise}}$

$H_1 : \mu_{\text{Raise}} \neq \mu_{\text{Not Raise}}$

$\alpha = 0.05$

From the table at http://www.dol.gov/whd/state/stateMinWageHis.htm, minimum wage data is available for consecutive years from 2000 to 2013, with an indication of rate changes beginning in the year 2001. Sampling from this set of data will be done by selecting 3 different years and using all the data from those years.

b. What is the name of the sampling method that is being used? _____

Which three years will be selected if your TI84 calculator is seeded with the number 42 and the years 2001 thru 2012 can be selected? These years were chosen because there is minimum wage increase data for these years and unemployment records for the year of, and the year before the unemployment rate increased, are available. Unemployment rates are found at http://www.bls.gov/lau/tables.htm.

c. Which years are selected? _____, _____, _____

The two graphs below are of the actual After-Minus-Before change in unemployment rates for the various states in the years that were randomly selected.
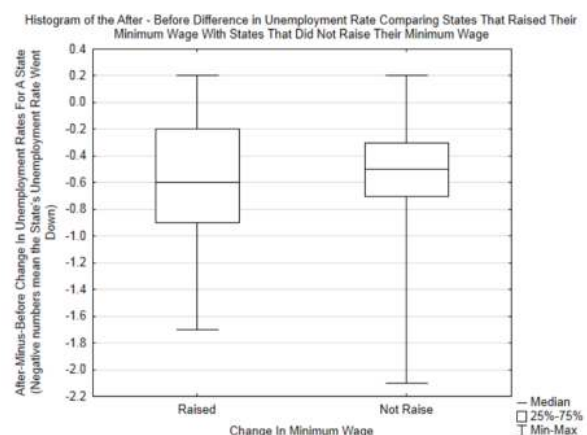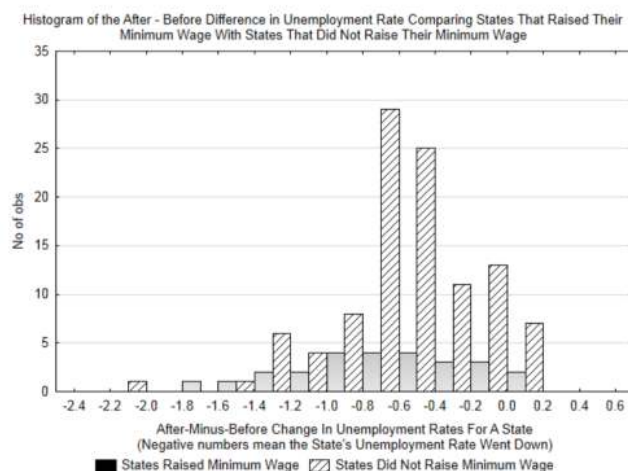
d. Do the graphs appear to support the null hypothesis or the alternate hypothesis better?

e. Both graphs are based on the same data. Which graph do you think shows the data better? Why?

One additional graph is shown to the right. It includes concepts that will be discussed near the end of the book, but because this topic is of interest to the many people working at minimum wage, the graph is being included here. Each line is for a different year. The mean for each year is in the center of the vertical bar. The vertical bars on the left show the change in unemployment for the states that raised their minimum wage and the vertical bars on the right are for the states that did not. The bars represent the confidence interval. Since decreasing unemployment is viewed as desirable, then this graph shows that in two of the years (2004 and 2012), the states that raised their minimum wage reduced their unemployment rate more than the states that didn't raise their rates. In 2006, the states that didn't raise their minimum wage reduced their unemployment rate more than the states that did raise their rates.
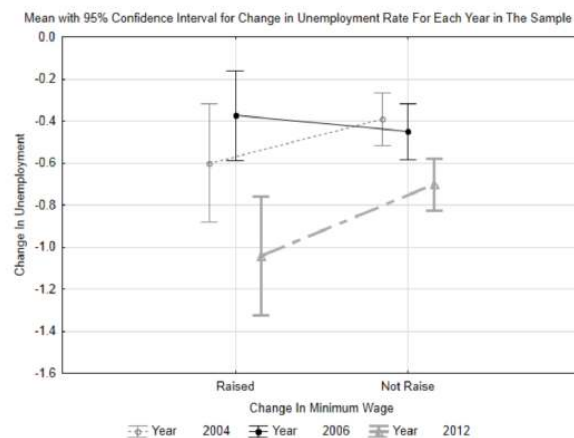
f. The table below shows the average change in unemployment rates for all the data combined. Which hypothesis do these statistics support?

	States Raised Minimum Wage	States Did Not Raise Minimum Wage
Mean	-0.615	-0.519
n	26	105

g. The p-value for a comparison of the two means is 0.286. Write a concluding sentence in the style used in scholarly journals (like you were taught in Chapter 1).

h. Suppose you were in a class in which this topic was being discussed. What would you say to a classmate who argued that the minimum wage should not be raised because it will lead to more unemployment?

What would you say to the classmate who argued that the minimum wage should be raised because it means the poorer people will have more money to spend which means businesses will do better and have to hire more people thereby causing unemployment to drop even more?

10. Why Statistical Reasoning Is Important for a Nursing Student and Professional Developed in collaboration with Becky Piper, Pierce College Puyallup Nursing Program Director This topic is discussed in NURS 112.
This problem is based on *An Analysis of Falls in the Hospital: Can We Do Without Bedrails?*

by H.C. Hanger, M.C. Ball and L.A. Wood. the American Geriatrics Society, 47:529-531.

There was a time when women who helped the sick and injured were poorly regarded. However, in 1844, Florence Nightingale, daughter of a British banker, started visiting hospitals and learning about the care of patients. She eventually provided leadership to the British field hospitals during the Crimean War of 1853-56.(http://en.Wikipedia.org/wiki/Crimean_War) While her efforts helped improve the quality of the hospitals, it was after the war that she reflected about results she considered disappointing. She sought the assistance of William Farr who had recently invented the field of medical statistics. To help Florence understand the reasons for all the deaths in the hospital, he suggest that "We do not want impressions, we want facts." One of her theories had been that many of the deaths were the result of inadequate food and supplies. The statistics lead to a rejection of this theory and instead pointed to lack of sanitation as a cause.(https://www.sciencenews.org/article/...e-statistician)

Nightingale was also known for her use of graphs as a way of showing her analysis. Because of Florence Nightingale, the profession of Nursing is inextricably linked with statistics. In the modern context, it is called "evidence-based practices".

Because hospital patients, particularly the elderly, have physical and possible cognitive problems that required placement in a hospital or nursing home, there is a need for nurses to keep the patients safe. One problem for these patients is falls, including falling out of bed. A standard practice for facilities has been to use bedrails so that a patient doesn't accidently roll out of bed.

The researchers who wrote the article could find no evidence that bedrails prevented falls, so they conducted their own experiment. They instituted a policy at their hospital (in Christchurch, NZ) to discontinue the use of bedrails unless there was a justifiable reason for their use that was documented and approved. Their experiment was to compare the average number of falls per 10,000 bed days after the implementation of the policy to before its implementation. If the bedrails helped reduce falls, the number of falls should increase after they are removed.

$H_0 : \mu_{\text{after}} = \mu_{\text{before}}$

$H_1 : \mu_{\text{after}} > \mu_{\text{before}}$

$\alpha = 0.05$

a. Complete the experiment design table

Research Design Table	
Research Question:	
Type of Reserach	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

b. Before implementing the policy, the average number of falls per 10,000 bed days was 164.8 (S.D. = 20.6). After the new policy was implemented, the average number of falls per 10,000 bed days was 191.7 (S.D. = 40.7). The p-value was 0.18. Write a complete concluding sentence.

c. An additional part of the experiment was to compare the severity of the falls. Falls were classified as serious injury, minor injury or no injury. The table below shows the distribution of the injuries.

	Pre-policy	Post-policy
Serious injury	33	18
Minor injury	43	60
No injury	110	154

There is a significant difference in the injuries (p = 0.008). Explain what the difference is and give a possible reason for the difference.

d. If you were a nurse, would you suggest that bedrails be required or be removed? Why?

3: Examining the Evidence using Graphs and Statistics

We live in a world in which decisions must be made without complete information. Knowing this, we intuitively seek to gather as much information as possible before making the actual decision. Consider marriage, which is a rather important decision. We can never know everything possible about a person we'd want to marry but we do seek as much information as possible by dating first. Employment is another example of an important decision, for both the employer and the potential employee. In each case, information is gained through interviews, resumes, references and research before a job offer is given or accepted.

When faced with a decision that will be based on data, it is the production of graphs and statistics that will be analogous to dating and interviews. The data that is collected must be useful to answer the questions that were asked. Chapter 2 focused on both the planning of the experiment and the random selection process that is important for producing good sample data. Chapter 3 will now focus on what to do with the data once you have it.

Types of Data

We have already classified data into two categories. Numeric data is considered quantitative while data consisting of words is called categorical or qualitative data. Quantitative data can be subdivided into discrete and continuous data.

- **Discrete** data contains a finite number of possible values because they are often based on counts. Often these values are whole numbers, but that is not a requirement. Examples of discrete data include the number of salmon migrating up a stream to spawn, the number of vehicles crossing a bridge each day, or number of homeless people in a community.
- **Continuous** data contains an infinite number of possible values because they are often based on measurements, which in theory could be measured to many decimal places if the technology existed to do so. Examples of continuous data include the weight of the salmon that are spawning, the time it takes to cross the bridge, or the number of calories a homeless person consumes in a day.

Discrete quantitative data and categorical data are often confused. Look at the actual data that would be written for each unit in the sample to determine the type of data. As an example, consider the brown beetle, which is infecting trees in the western US and Canada. If the purpose of the research was to determine the proportion of trees that are infected, then the data that would be collected for each tree is "infected" or "not infected". Ultimately, the researcher would count the number of trees marked infected or not infected, but the data itself would be those words. If the purpose of the research was to determine the average number of brown beetle on each tree, then the data that would be collected is "the number of brown beetle on a tree", which is a count. Thus, counts are involved for both categorical and discrete quantitative data. Categorical data is counted were as if categorical data is counted in multiple places or times, then the counts become discrete quantitative data. For example, in class today, students in the class roster can be marked as present or absent and this would be categorical. However, if we consider the number of students who have been present each class during the past week, then the data in which we are interested is quantitative discrete.

Examining the evidence from sample data

Since sample data are our window into the stochastic data in the population, we need ways to make the data meaningful and understandable. This is accomplished by using a combination of graphs and statistics. There is one or more graph and statistic that is appropriate for each type of data. In the following sections you will learn how to make the graphs by hand and how to find the statistics. There are many other graphs that exist besides this collection, but these are the basic ones.

Examining the evidence provided by sample categorical data

There are two graphs and two statistics that are appropriate for categorical data. The graphs most commonly used are bar graphs and pie charts. The statistics are counts and proportions. If the hypothesis being tested is about counts, then a bar graph and sample counts should be used. If the hypothesis being tested is about proportions, then a pie chart and sample proportions should be used. For categorical data, the statistics are found first and then used in the production of a graph.

Counts and Bar Graphs

Political leadership in the US is typically divided between two political parties, the Democrats and the Republicans. Only a few politicians have been elected as independents meaning they do not belong to one of these parties. The highest politically elected positions other than the President are congressmen, senators and state governors. If we want to understand the distribution of political parties in 2013, then the political party of our leaders is categorical data that can be put into a contingency table in which each cell represents a count of the number of people who fit both the leadership position category and the political party category. A bar graph can be made from these counts.

2013		Leadership Position		
		Congress	Senate	Governor
Political Party	Democrats	200	52	20
	Independents	0	2	0
	Republicans	233	46	30

Distribution of Political Parties in Major Leadership Positions, 2013
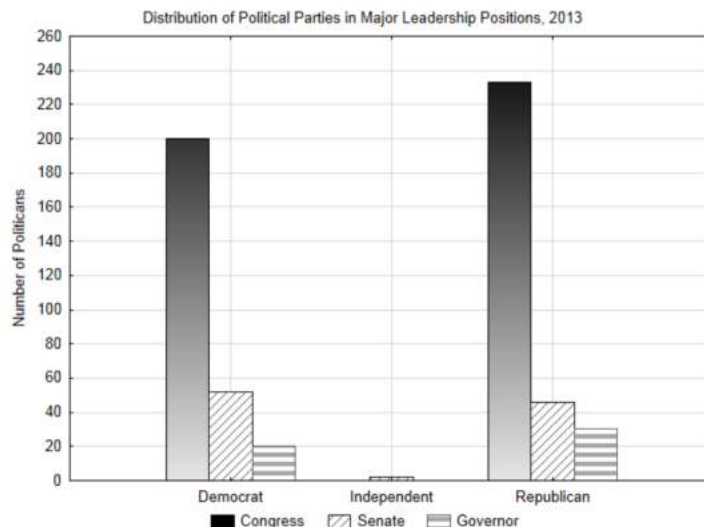
Proportions and Pie Charts

Opinion polls frequently use proportions or percentages to show support for candidates or initiatives. The difference between proportions and percentages is that percentages are obtained by multiplying the proportion by 100. Thus, a proportion of 0.25 would be equivalent to 25%. Formulas use proportions while we often communicate verbally using percentages. You should be able to move from one to the other effortlessly.

There are almost always two proportions of interest to us. The population proportion, represented with the symbol p, is the proportion we would really like to know, but which is usually unknowable. We make hypotheses about p. The sample proportion, represented with $\hat{p}$, is what we can find from sample data and is used to test the hypothesis. The formula for proportions are:

$$p = \frac{x}{N} \tag{3.1}$$

and

$$\hat{p} = \frac{x}{n} \tag{3.2}$$

where x is a count of the number of values in a category, N is the size of the population, and n is the size of the sample.

The results of two surveys discussed on a washingtonstatewire.com blog will be used for an example. Given that much of the transportation gridlock is caused by cars, and that Washington State's bridges need maintenance (there was a bridge collapse on Interstate 5 near Mount Vernon, WA in 2013) it would be natural to wonder about voter support for state funding of transportation projects. Two polls were conducted at about the same time in 2013. (washingtonstatewire.com/blog/...portation-tax- package-offer-a-measure-of-voter-mood-after-bridge-collapse/ viewed 7-25-13.)

Poll 1 used human interviewers who began a scripted interview by observing that "of course transportation projects are expensive and take a long time to complete," and concluded with, "as I said, transportation projects are expensive. The other part of the package will be how to pay for those improvements. No one likes to raise taxes, but as I read some funding options, tell me whether you would favor the proposal, be inclined to accept it, be inclined to oppose, or find it unacceptable."

Poll 2 used robo-polling which asked voters whether it is important for "the legislature to pass a statewide package this year to address congestion and safety issues, fund road and bridge maintenance and improvement, and provide additional transit funding."

As best as can be estimated from the article, the results of Poll 1 were that 160 out of 400 people who were surveyed supported raising taxes for improving the transportation system. The results of Poll 2 were that 414 out of 600 think it is important for the Legislature to pass the funding package.

From data such as this we can make a pie chart. This will be demonstrated with Poll 1 and then you should make a pie chart for Poll 2.

The first step in making a pie chart is to calculate the proportion of values in each group. In Poll 1, we will consider there are two groups. The first group is for those who supported raising taxes and the second group is for those who did not support raising taxes. Since 160 out of 400 people supported raising taxes, then the proportion is found by dividing 160 by 400. Therefore, $\hat{p} = \dfrac{160}{400} = 0.40$. As a

reminder, $\hat{p}$ is the proportion of the sample that supports raising taxes. It is a statistic, which provides insight into the population proportion, represented with the variable p. The legislators would like to know the value of p, but that would require doing a census, so they must settle for the sample proportion, $\hat{p}$. It is likely the p does not equal $\hat{p}$, but that it is close to that value. When making a pie chart, draw the line separating the slices so that 40% of the circle is in one slice, which means that 60% of the circle is in the other.

Poll 1 - WA Support of Taxes for Highway Improvements - 2013


Support Taxes, 160, 40%
Don't Support Taxes, 240, 60%
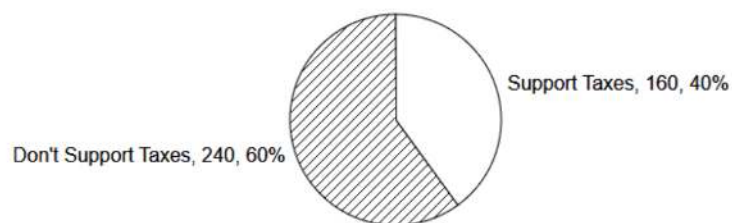
There are a few things to notice about the pie chart. First, it contains a title that describes the content of the graph. Next, each slice contains a label that briefly explains the meaning of the slice, the number of data values that contributed to the slice and the percent of all the values that are put in the slice. Why should all of this information be included?

If you are going to use any graph to show the results of your research, it is important to communicate those results clearly. The goal is to produce reader-friendly graphs. A reader looking at an unlabeled graph will not be able to gain any understanding from it, and thus you have failed to communicate something important. The percentage is included to make it easy for the reader to know the percent of values in each slice. Without the percentages, a person would need to guess at the percentage and it is likely their guess would not be precise. Including the number of people in each slice is important because it gives the reader an indication of how seriously to treat the results. A survey of 40 people of which 16 supported taxes would have a pie chart identical to the one above. Likewise, a survey of 40,000 people, of which 16,000 supported taxes, would also be identical to the above graph. The more people there are, the stronger the support. This should be obvious from the graph and therefore it is important to include the value.

A mention must be made about computer graphics since most pie charts are produced on a computer. While computers can make very fancy and colorful graphs, the colors can be indistinguishable if printed on a black and white printer or photo copied in black and white. Keep this in mind when you make graphs and pick colors that will be distinguishable when copied in black and white.

Use the results of Poll 2 to produce a completely labeled pie chart. Find the sample proportion first.
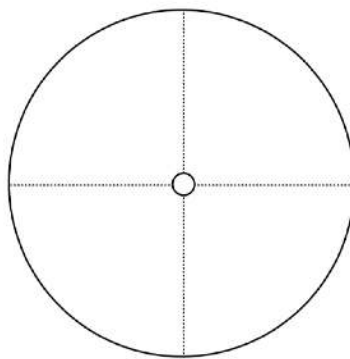
Do these two polls produce similar results or opposite results? Were the questions well worded?

Why or why not?

A final word about pie charts needs to be made. In some circles, pie charts are not considered useful graphs. There is some evidence that people do not do a good job of interpreting them. Pie charts very seldom appear in scholarly journals. However, pie charts do appear in print media and can give an indication of how the whole is divided. They may be of benefit to those who like the visual representation, rather than just the statistics.

Examining the evidence provided by sample quantitative data

The three most common types of graphs used for quantitative data are histograms, box plots and scatter plots. Histograms and box plots are used for univariate data whereas scatter plots are used for bivariate data. A variate is a single measurement or observation of a random variable obtained for each subject or unit in a sample.(Sokal, Robert R., and F. James Rohlf. *Introduction to Biostatistics*. New York: Freeman, 1987, Print.) When there is only one random variable that is being considered, the data that are collected are univariate. When two random variables are being considered simultaneously for the same unit, then the data for the two variables are considered bivariate. Examples of univariate data include the number of vehicles on a stretch of highway, the amount it costs for a student to earn their degree, or the amount of water used by a household each month. Examples of bivariate data include the pairing of number of cars on the highway and the commute time, the amount of tuition and the amount of financial aid a student uses, or the number of people in a household and the amount of water used.

The statistics used for univariate data fit one of two objectives. The first objective is to define the center of the data and the second objective is to define the variation that exists in the data. The most common ways of defining the center are with the arithmetic mean and the median, although these are not the only two measures of center. In cases where the arithmetic mean is used, variation is quantified using standard deviation. The statistic most commonly used for bivariate data is correlation, which indicates the strength of the linear relationship between the two variables.

Histograms

Chapters one and two contained numerous examples of histograms. They are used to show the distribution of the data by showing the frequency or count of data in each class. The process for creating histograms by hand includes the following steps.

1. Identify the lowest and highest data values.
2. Create reader-friendly boundaries that will be used to sort the data into 4 to 10 classes. The lowest boundary should be a number that either equals, or is a nice number below, the lowest data value. The class width, which is the difference between consecutive boundaries, should be a factor of the boundary values.
3. Make a frequency distribution to provide an organized structure to count the number of data values in each class.
4. Create the histogram by labeling the x-axis with the lower boundaries and the y-axis with the frequencies. The height of the bars reflects the number of values in each class. Adjacent bars should touch.
5. Put a title on the graph and on each axis.

There isn't a precise mathematical way to pick the starting value and the class width for a histogram. Rather, some thought is necessary to use numbers that are easy for a reader to understand. For example, if the lowest number in a set of data is 9 and the highest number is 62, then using a starting value of 0 and a class width of 10 would result in the creation of 7 classes with reader-friendly boundaries of 0,10,20,30,40,50,60, and 70. On the other hand, starting at 9 and using a class width of 10 would not

produce reader-friendly boundaries (9,19,29, ...). Numbers such as 2,4,6,8... or 5,10,15,20... or any version of these numbers if they are multiplied by a power of 10 make good class boundaries.

Once the class boundaries have been determined, a frequency distribution is created. A frequency distribution is a table that shows the classes and provides a place to tally the number of data values in each class. The frequency distribution should also help clarify which class will be given the boundary values. For example, would a value of 20 be put into a 10 – 20 class or a 20 – 30 class? While there is no universal agreement on this issue, it seems a little more logical to have all the values that begin with the same number be grouped together. Thus, 20 would be put into the 20 – 30 class which contains all the values from 20.000 up to 29.999. This can be shown in a few ways as are demonstrated in the table below.

0 up to, but not including 10	$0 \leq x < 10$	[0, 10)
10 up to, but not including 10	$10 \leq x < 20$	[10, 20)
20 up to, but not including 10	$20 \leq x < 30$	[20, 30)
30 up to, but not including 10	$30 \leq x < 40$	[30, 40)

All three columns indicate the same classes. The third column uses interval notation and because it is explicit and uses the least amount of writing, will be the method used in this text. As a reminder about interval notation, the symbol " [" indicates that the low number is included whereas the symbol ") " indicates the high number is not included.

To demonstrate the construction of a histogram, data from the US Department of Transportation, Federal Highway Administration will be used.(explore.data.gov/Transportat...3-mssz,7-28-13) The data is the estimated number of miles driven in a state in December, 2010. A stratified sample will be used since the data are already divided into regions of the country. The data in the table has units of millions miles.

4778	768	859	3816
6305	4425	789	1517
9389	3681	21264	8394
583	2958	2034	2362
712	5858	738	7861
5664	352	16256	2594
665	28695	4435	

1. The low value is 352, the high value is 28,695.
2. The lowest class boundary will be 0, the class width will be 5000. This will produce 6 classes.
3. This is the frequency distribution that includes a count of the number of values in each class.

Classes	Frequency
[0, 5000)	19
[5000, 10000)	6
[10000, 15000)	0
[15000, 20000)	1
[20000, 25000)	0
[25000, 30000)	1

4, This is the completely labeled histogram. Notice how the height of the bars corresponds with the frequencies in the frequency distribution.

Million Miles Driven in Select States, December 2010
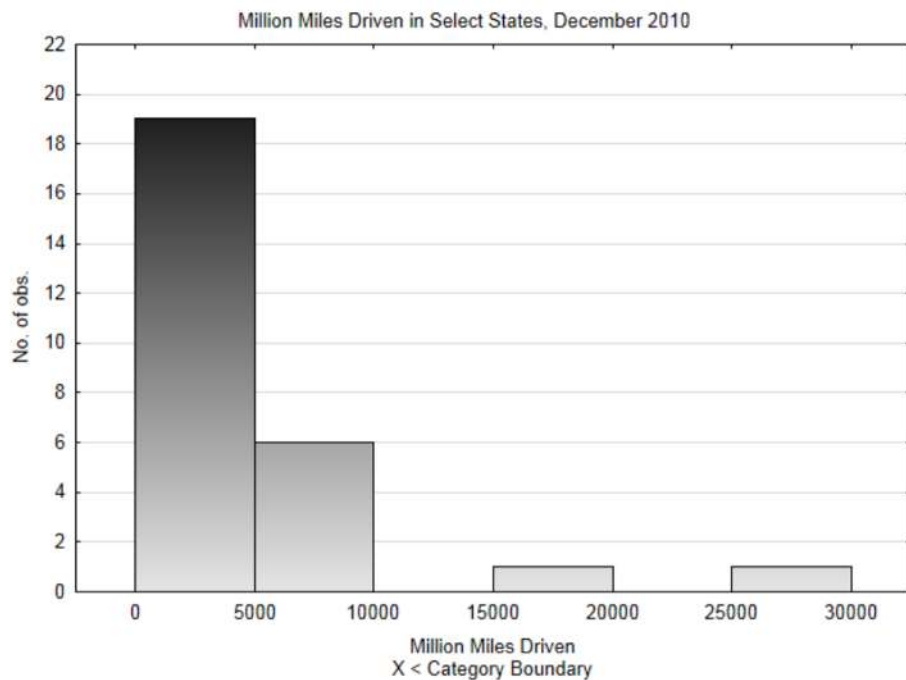
Suppose we want to compare the amount of driving in states with a large area to those with a smaller area. This could be done using a multiple bar histogram in which one set of bars will be for larger states and the other for smaller states.

Frequency Distribution and Multiple Bar Histogram:

Classes	Frequency States with Least Area	Frequency States with Most Area
[0,5000)	17	17
[5000,10000)	7	5
[10000,15000)	1	0
[15000,20000)	0	1
[20000,25000)	0	0
[25000,30000)	0	1

Comparison of the Amount of Driving in States with Small Areas and Large Areas December 2010
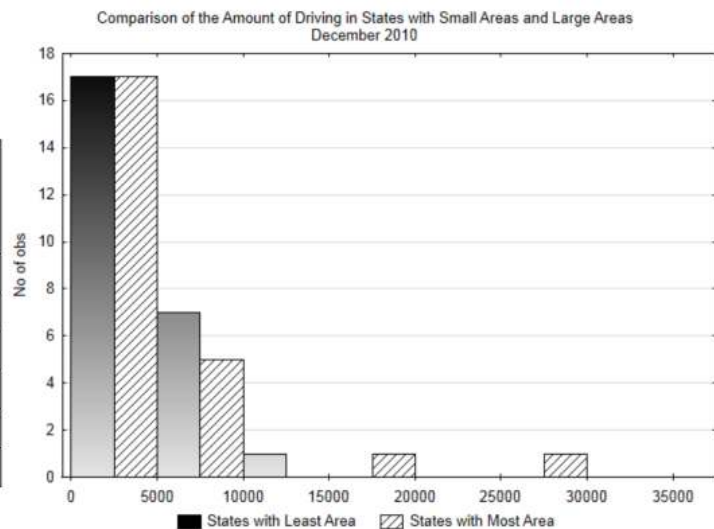
Interpretation: While it might be reasonable to assume there would be more driving in bigger states because the distance between cities is larger, it is difficult to discern from this graph if that is the case. Therefore, in addition to the use of a graph, this data can be compared using the arithmetic mean and the standard deviation.

Arithmetic Mean, Variance and Standard Deviation

The arithmetic mean and standard deviation are common statistics used in conjunction with histograms. The mean is probably the most commonly used way to identify the center of data, but it is not the only method. The mean can be thought of as the balance point for the data, much like the fulcrum on a teeter-totter. Values far from the mean have a greater impact on it than do values

closer to the mean in the same way a small child sitting at the end of a teeter-totter can balance with a larger person sitting near the fulcrum.

There are almost always two arithmetic means of interest to us. The population mean, represented with the symbol μ (mu), is the mean we would really like to know, but which is usually unknowable. We make hypotheses about μ. The sample mean, represented with $\bar{x}$ (x-bar), is what we can find from a sample and is what is used to test the hypothesis. The formula for the means, as shown in Chapter 1 are:

$$\mu = \frac{\sum x_i}{N} \text{ and } \bar{x} = \frac{\sum x_i}{n}$$

Where $\sum$ is an upper case sigma used in summation notation that means add everything that follows, x*i*is the set of data values and N is the number of values in the population and n is the number of values in the sample. These formulas say to add all the values and divide by the number of values.

There are several reasons why the arithmetic mean is commonly used and some reasons why it shouldn't be used at times. A primary reason it is commonly used is because the sample mean is an unbiased estimator of the population mean. This is because about half the sample means that could be obtained from a population will be lower than the population mean and half will be higher. An arithmetic mean is not the best measure of center when there are a few extremely high values in the data, as they will have more of an impact on the mean than the remaining values.

In addition to the mean, it is also useful to know how much variation exists in the data. Notice in the double bar histogram how the data in the states with the largest area is spread out more than the data in the states with the smallest area. The more spread out data is, the more difficult it is to obtain a significant result when testing hypotheses.

Standard deviation is the primary way in which the spread of data is quantified. It can be thought of as the approximate average distance between each data value and the mean. As with the mean, there are two values of standard deviation that interest us. The population standard deviation,represented with the symbol σ (lower case sigma), is the standard deviation we would really like to know, but which is usually unknowable. The sample standard deviation, represented with s, is what we can find from a sample. The formulas of standard deviation are:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{3.3}$$

and

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \tag{3.4}$$

North Atlantic hurricane data will be used to demonstrate the process of finding the mean and standard deviation. (Data from: www.wunderground.com/hurrican...asp?region=ep.)

Year	2005	2006	2007	2008	2009	2010	2011
Number of Hurricanes	15	5	6	8	3	12	7

Since this is a sample, the appropriate formula for finding the sample mean is $\bar{x} = \frac{\sum x_i}{n}$. The calculation is $\frac{15+5+6+8+3+12+7}{7} = \frac{56}{7} = 8$. There were an average of 8 North Atlantic hurricanes per year between 2005 and 2011. Notice that there weren't 8 hurricanes every year. This is because there is natural variation in the number of hurricanes. We can use standard deviation as one way for determining the amount of variation. To do so, we will build a 3-column table to help with the calculations.

x	$(x-\bar{x})$	$(x-\bar{x})^2$
15	15- 8 = 7	$(7)^2 = 49$
5	5 - 8 = -3	$(-3)^2 = 9$

x	$(x - \bar{x})$	$(x - \bar{x})^2$
6	6 - 8 = -2	$(-2)^2 = 4$
8	8 - 8 = 0	$(0)^2 = 0$
3	3 - 8 = -5	$(-5)^2 = 25$
12	12 - 8 = 4	$(4)^2 = 16$
7	7 - 8 = -1	$(-1)^2 = 1$
	$\sum(x - \bar{x}) = 0$	$\sum(x - \bar{x})^2 = 0$

Since this is a sample, the appropriate formula for finding the sample standard deviation is $s = \sqrt{\dfrac{\sum(x - \bar{x})^2}{n - 1}}$ which, after substitution is $\sqrt{\dfrac{104}{7 - 1}} = 4.16$. This number indicates that the average variation from the mean in each year is 4.16 hurricanes.

Variance is another measure of variation that is related to the standard deviation. Variance is the square of the standard deviation or, conversely, the standard deviation is the square root of the variance. The formulas for variance are:

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} \tag{3.5}$$

and

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \tag{3.6}$$

In the example about hurricanes, the variance is $s^2 = \dfrac{104}{7 - 1} = 17.33$.

Medians and Box Plots

Another combination of statistics and graphs are medians and box plots. A median is found before a box plot can be created. A median is the value of a variable in an ordered array that has an equal number of items on either side of it.(5 Sokal, Robert R., and F. James Rohlf. *Introduction to Biostatistics*. New York: Freeman, 1987. Print.) To find the median, put the data in order from small to large. Assign a rank to the numbers. The smallest number has a rank of 1, the second smallest has a rank of 2, etc. The rank of the median is found using formula 4.5.

$$Rank\,of\,Median = \frac{n + 1}{2} \tag{3.7}$$

If n is odd, that is, if there are an odd number of data values, then the median will be one of the data values. If n is an even number, then the median will be the average of the two middle values.

The same hurricane data will be used in the first of two demonstrations for finding the median.

Year	2005	2006	2007	2008	2009	2010	2011
Number of Hurricanes	15	5	6	8	3	12	7

The first step is to create an ordered array.

Number of Hurricanes	3	5	6	7	8	12	15
Rank	1	2	3	4	5	6	7

The second step is to find the rank of the median using the formula $Rank\ of\ Median = \dfrac{n+1}{2}$, $\dfrac{7+1}{2} = 4$.

The third step is to find the data value that corresponds with the rank of the median.

Since the rank of the median is 4 and the corresponding number is 7 hurricanes then the median number is 7 hurricanes.

The second demonstration will be with the number of East Pacific Hurricanes. Since there is no data for 2011, only the years 2005-2010 will be used.

Year	2005	2006	2007	2008	2009	2010	2011
Number of Hurricanes	5	10	2	4	7	3	

The first step is to create an ordered array.

Number of Hurricanes	2	3	4	5	7	10
Rank	1	2	3	4	5	6

The second step is to find the rank of the median using the formula $Rank\ of\ Median = \dfrac{n+1}{2}$

$\dfrac{6+1}{2} = 3.5$. This means the median is halfway between the third and fourth values.

The third step is to find the data value that corresponds with the rank of the median.

The average of the third and fourth values is $\dfrac{4+5}{2} = 4.5$. Therefore the median number of East Pacific

hurricanes between 2005 and 2010 is 4.5. Notice that in this case, 4.5 is not one of the data values and it is not even possible to have half of a hurricane, but it is still the median.

A box plot is a graph that shows the median along with the highest and lowest values and two other values called the first quartile and the third quartile. The first quartile can be thought of as the median of the lower half of the data and the third quartile can be thought of as the median of the upper half of the data.

The North Atlantic Hurricane Data will be used to produce a box plot.

The first step is to create an ordered array.

Number of Hurricanes	3	5	6	7	8	12	15
Rank	1	2	3	4	5	6	7

The second step is to identify the lowest value, the median, and the highest value.

	Lowest			Median			Highest
Number of Hurricanes	3	5	6	7	8	12	15
Rank	1	2	3	4	5	6	7

The third step is to identify the first quartile and the third quartile. This is done by finding the median of all the values below the median and above the median.

	These are the values below the median		
Number of Hurricanes	3	5	6
Rank	1	2	3
		Q1	

	These are the values above the median		
Number of Hurricanes	8	12	15
Rank	5	6	7
		Q3	

Thus, the five values of interest are:

Maximum: 15

Third Quartile (Q3): 12

Median: 7

First Quartile (Q1): 5

Minimum: 3

Number of North Atlantic Hurricanes per Year 2005-2011
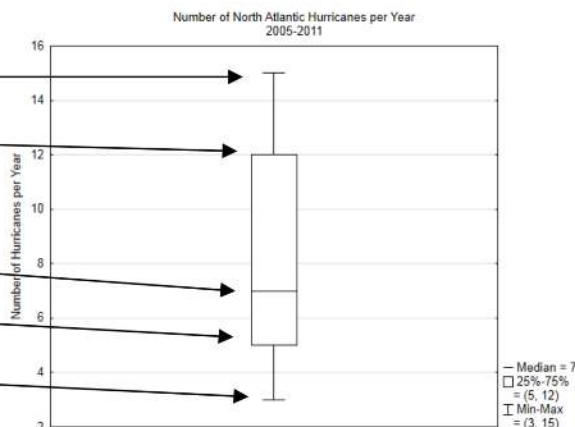
The box plot divides the data into 4 groups. It shows how the data within each group is spread out.

When graphing quantitative data, is it better to use a histogram or box plot? Compare the follow graphs that show a comparison of the number of hurricanes in four areas, North Atlantic, East Pacific, West Pacific, Indian Ocean. The data is from the years 1970 – 2010.
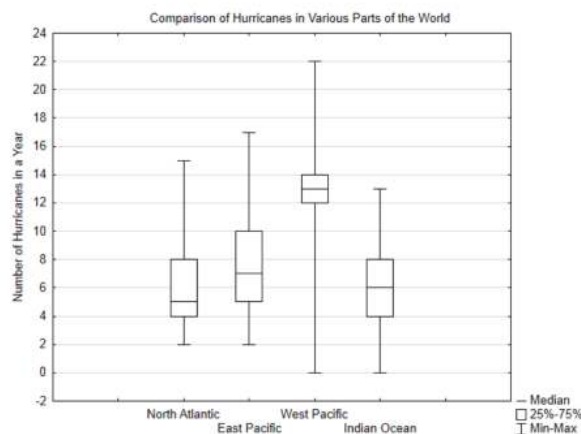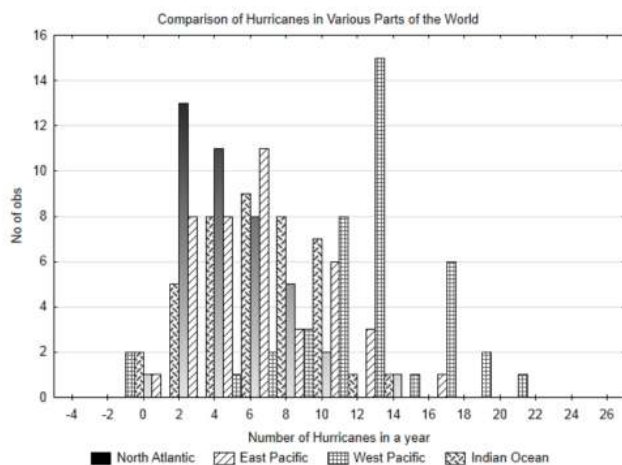
While the histogram gives a more detailed break down of the data, it is very cluttered and difficult to interpret. Therefore, in spite of the additional information it provides, the reader has to study the graph intently to understand what it shows. On the other hand, the box plot provides less information, but it is much easier to draw a comparison between the different hurricane areas. In general, if there is only one set of data being graphed, a histogram is the better choice. If there are three or more sets of data being graphed, a box plot is the better choice. If there are two sets of data being graphed, make both a histogram and a box plot and decide which is more effective for helping the reader understand the data.

Scatter Plots and Correlation

Some research questions result from the desire to find an association between two quantitative variables. Examples include wealth gap (Gini Coefficient)/poverty rates, driving speed/distance to stop, height/jumping ability. The goal is to determine the relationship between these two random variables and in many cases to see if that relationship is linear.

For demonstration purposes, we will explore the relationship between the wealth gap as measured by the Gini Coefficient and the poverty rate. The units will be randomly selected US states from the year 2010. A scatter plot will give a quick understanding of the relationship.

Scatterplot of Poverty Rate against Gini Coefficient
Based On US States, 2010
Data from US Census Bureau
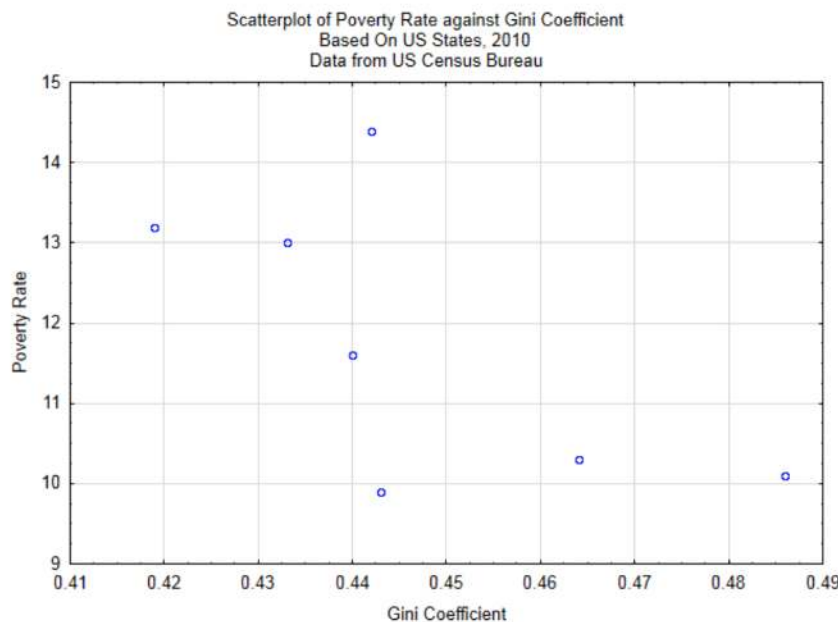
From this scatter plot it appears that the greater the wealth gap, the lower the poverty rate, although the relationship is not a strong one since the points do not appear to be grouped close together to form a straight line. To determine the strength of the linear relationship between these variables we use the Pearson Product-Moment Correlation Coefficient.

There are almost always two correlation coefficients of interest to us. The population correlation, represented with the symbol ρ (rho), is the correlation coefficient we would really like to know, but which is usually unknowable. We make hypotheses about ρ. The sample correlation, represented with r, is what we can find from a sample and is what is used to test the hypothesis. The formula for the sample correlation coefficient is:

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} \tag{3.8}$$

The numerator is the covariance between the two variables, the denominator is the product of the standard deviation of each variable.

Correlation will always be a value between -1 and 1. A correlation of 0 means no correlation. A correlation of 1 means a direct linear relationship in which y gets larger as x gets larger. A correlation of -1 means an inverse linear relationship in which y gets smaller as x gets larger.

Correlation r = 0

Correlation r = 1

Correlation r = -1

r = 0.60

r = -0.31
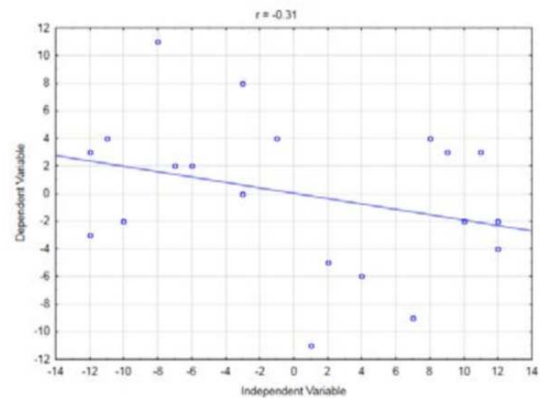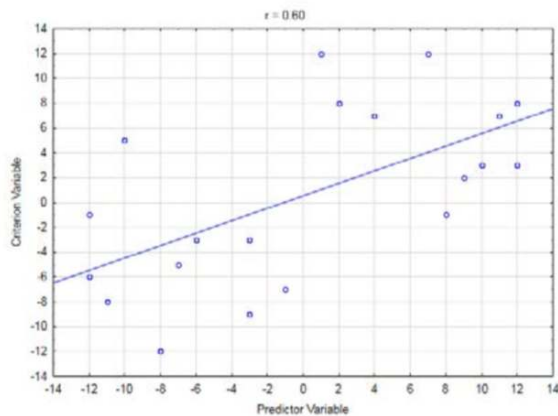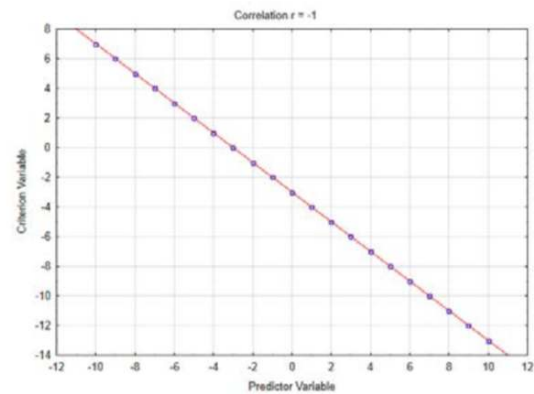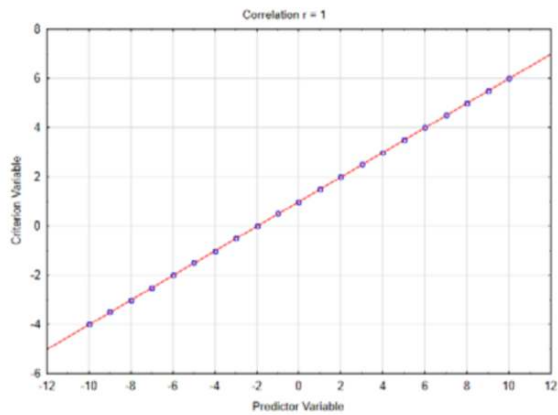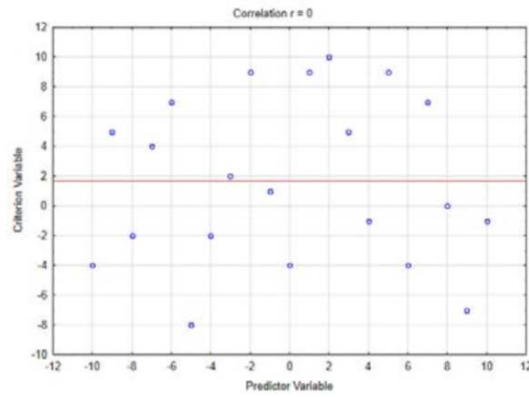
A brief explanation of the correlation formula follows. Think of bivariate data as an (x,y) ordered pair. The ordered pair $(\bar{x}, \bar{y})$ is the centroid of the data. For this data, the centroid is at (0.4467, 11.7857). This is shown in the graph below.

Scatterplot of Poverty Rate against Gini Coefficient
Based On US States, 2010
Data from US Census Bureau
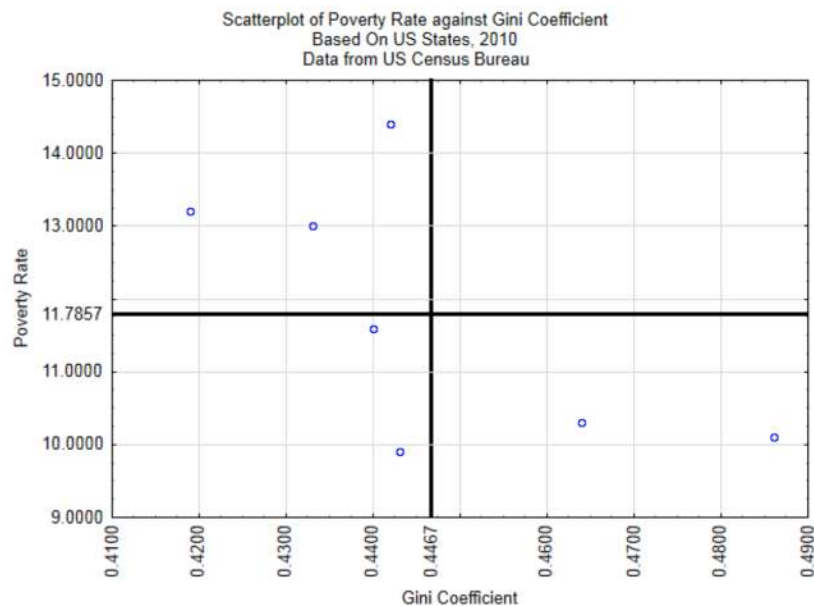
The covariance is given by the formula $\text{cov}(x, y) = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$. It shows the product of

the distance each point is away from the average x value and the average y value. Since multiplying both the x values and y values by 10 would result in a covariance that is 100 times larger than this data would produce, yet the graph would look the same, the covariance is standardized by dividing by the product of the standard deviations of x and y.

Calculate the Covariance

(x, y) or (gini, pov)	$(x - \bar{x})$ (x - 0.4467)	$(y - \bar{y})$ (y - 11.7857)	$(x - \bar{x})(y - \bar{y})$
(0.486, 10.1)	0.0393	-1.6857	-0.0662
(0.443, 9.9)	-0.0037	-1.8857	0.0070
(0.44, 11.6)	-0.0067	-0.1857	0.0012
(0.433, 13)	-0.0137	1.2143	-0.0167
(0.419, 13.2)	-0.0277	1.4143	-0.0392
(0.442, 14.4)	-0.0047	2.6143	-0.0123
(0.464, 10.3)	0.0173	-1.4857	-0.0257
Sum	0.0000	0.0000	-0.1518

$\text{cov}(x, y) = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$

$\text{cov}(x, y) = \dfrac{-0.1518}{7 - 1}$

$\text{cov}(x, y) = -0.0253$

Calculate the standard deviation of x and y

(x, y) or (gini, pov)	$(x - \bar{x})$ (x - 0.4467)	$(x - \bar{x})^2$	$(y - \bar{y})$ (y - 11.7857)	$(y - \bar{y})^2$
(0.486, 10.1)	0.0393	0.00154	-1.6857	2.84163

(x, y) or (gini, pov)	$(x - \bar{x})$ $(x - 0.4467)$	$(x - \bar{x})^2$	$(y - \bar{y})$ $(y - 11.7857)$	$(y - \bar{y})^2$
(0.443, 9.9)	-0.0037	0.00001	-1.8857	3.55592
(0.44, 11.6)	-0.0067	0.00005	-0.1857	0.03449
(0.433, 13)	-0.0137	0.00019	1.2143	1.47449
(0.419, 13.2)	-0.0277	0.00077	1.4143	2.00020
(0.442, 14.4)	-0.0047	0.00002	2.6143	6.83449
(0.464, 10.3)	0.0173	0.00030	-1.4857	2.20735
Sum	0.0000	0.0029	0.0000	18.9486

$$S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \qquad (3.9)$$

$$S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}} \qquad (3.10)$$

$$S_x = \sqrt{\frac{0.0029}{7 - 1}} \qquad (3.11)$$

$$S_y = \sqrt{\frac{19.9486}{7 - 1}} \qquad (3.12)$$

$$S_x = 0.0219 \qquad (3.13)$$

$$S_y = 1.777 \qquad (3.14)$$

Use these results to calculate the correlation.

$$= \frac{\text{cov}(x, y)}{S_x S_y}$$

$$= \frac{-0.0253}{0.0219 \cdot 1.777}$$

$$= -0.650$$

This correlation indicates that higher Gini Coefficients correspond with lower poverty levels. Whether a correlation of –0.650 indicates the data are significant or simply random results from a population without correlation, is a matter for a later chapter. (www.census.gov/prod/2012pubs/acsbr11-02.pdf)

While it is important to understand that a correlation between variables does not imply causation, a scatter plot is drawn with one of the variables being the independent x value, also known as the explanatory variable and the other being the dependent y value, also known as the response variable. The names explanatory and response are used because if a linear relationship between the two variables exists, the explanatory variable can be used to predict the response variable. For example, one would expect that driving speed would influence stopping distance rather than stopping distance influencing driving speed so that driving speed would be the explanatory variable and stopping distance would be the response variable. However, a person may choose to drive slower under certain conditions because of how long it could take them to stop (e.g. a school zone) so the choice of explanatory and response variables must be consistent with the intent of the research. The accuracy of the prediction is based on the strength of the linear relationship. (www.census.gov/prod/2012pubs/acsbr11-01.pdf Sheskin, David J. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman & Hall/CRC, 2000. Print.)

If a correlation between the explanatory variable and the response variable can be established, one of seven possibilities exists.

1. Changing the x variable will cause a change in the y variable
2. Changing the y variable will cause a change in the x variable

3. A feedback loop may exist in which a change in the x variable leads to a change in the y variable which leads to another change in the x variable, etc.
4. The changes in both variables are determined by a third variable
5. The changes in both variables are coincidental.
6. The correlation is the result of outliers, without which there would not be significant correlation.
7. The correlation is the result of confounding variables.

The best guideline is to assume that correlation is not causation, but if you think it is in a certain circumstance, additional proof will need to be provided. A causal relationship can be established easier with a manipulative experiment than an observational experiment since the later may contain hidden confounding variables that are not accounted for.

TI-84 Calculator

The TI-84 calculator has the ability to quickly find all the statistics presented in this chapter. To find the arithmetic mean, standard deviation and all 5 box plot numbers, use the Stat key on your calculator. You will be presented with three options: EDIT, CALC, TESTS. Edit is already highlighted, so press the enter key and you will find three lists labeled L1, L2 and L3. There are also three other lists labeled L4, L5, L6 that can be found by scrolling to the right. Enter your data into one of the lists. After that, press the stat key again, use your cursor arrows to scroll to the right until Calc is highlighted, then press enter. The first option is 1-Var Stats. It is already highlighted, so press enter, then press the 2ndkey and the number corresponding to the list that your data is in (1-6). You will be presented with the following information.

$\bar{x}$ - Sample Arithmetic Mean

$\sum x$

$\sum x^2$

S_x - Sample S tan dard Deviation

σ_x - Population S tan dard Deviation

n - sample size

min X - lowest value

Q1 - first quartile

Med - median

Q3 - third quartile

max X - highest value

For bivariate data, enter the x values into one list and the y values into a different list, making certain they are properly paired. Use the stat key, select Calc, then select 4:LinReg(ax + b). Use the second key to enter the list number for the x variable followed by a comma and then enter the list number for the y variable. This will provide more information than we are ready for at the moment, but the one value you will look for is labeled r. If the r is not visible, you will need to turn the calculator diagnostics on. This is done by using the 2nd key followed by 0 (which will get the catalog). Scroll down to diagnosticOn, then press enter twice.


3.E: Examining the Evidence using Graphs and Statistics (Exercises)

1. Fans of professional sports teams expect the owners of the team to spend the necessary money to get the players who will help them win a championship. The payrolls of various professional sports teams in the US were divided into thirds, and the number of championships won by teams in each third was compared. Make a complete bar graph of this data. (data from unpublished student statistics class project)

Payroll Ranking	Number of Championships
Lowest Third	2
Middle Third	7
Highest Third	11

2. According to National Geographic, in 1903 there were 307 varieties of corn seed sold by seed houses. In 1983, there were 12 varieties sold be seed houses, the rest no longer being used. Find the sample proportion of varieties of corn seed that was still available in 1983 compared to 1903. Make a complete pie chart. (ngm.nationalgeographic.com/20...ariety-graphic viewed 9/9/13)
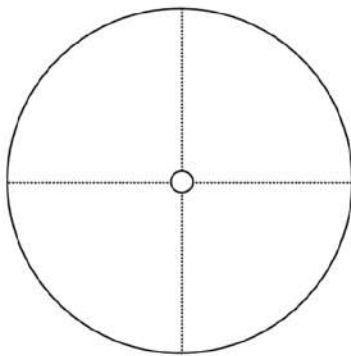
   Do you think it is good or bad that there are fewer varieties? Why?

3. Between 2010 and 2014, two dams on the Elwha River in the Olympic National Park near Port Angeles, WA were removed, allowing salmon to spawn for the first time in that river in 100 years. Assume the weights of 10 Chinook salmon that returned were recorded. These weights are shown in the table below.

41	48	40	43	45
39	35	47	41	51

   The purpose of this problem is to find all the statistics using the formulas by hand. Calculators should only be used to find the square root. Show all work.

   Find the mean, variance and standard deviation, and the 5 box plot numbers for this data.

4. Students in development mental math classes such as intermediate algebra are expected to know their math facts quickly. Automaticity, or math fact fluency, is the ability to recall math facts without having to make the calculations. The benefit of quickly knowing the math facts is that the working memory of the brain is not filled with the effort to make the calculations so that it can focus on the higher level thinking required for the algebra. Intermediate algebra students were given an automaticity test in which they had to solve as many one-step linear equations as possible in one minute. All addition, subtraction and multiplication equations used numbers between −10 and 10 while division equations had answers in that range. All answers were integers. The table below gives the number of problems completed successfully in one minute.

11	30	28	31	23
27	29	9	38	18

19	17	26	12	10
19	20	17	15	23
22	34	23	36	10

Make a frequency distribution and histogram. Using a starting value of 5 and a class width of 5. Label the graph completely.

Make a box plot.

5. The objective of the automaticity experiments is to determine if there is a relationship between a student's math fact fluency and their final grade for the quarter. The table below contains the bivariate data for 6 of the students.

	Automaticity Score	Final Grade
Student 1	19	4
Student 2	31	2.9
Student 3	16	1.4
Student 4	19	4
Student 5	20	2.3
Student 6	16	1.3

Make a scatter plot for this data.

Find the correlation using the formulas. You can use your calculator for the basic functions but not for simply finding correlation, other than to check your answer. Show all your work.

6. Automaticity is one area to investigate when a college attempts to improve success rates for developmental math classes. If it is a factor in success, then the college will develop a method for helping students improve their automaticity. An experiment was conducted in which intermediate algebra students were given a computerized automaticity test that required them to solve a mixture of one-step linear equations that required adding, subtracting or multiplying numbers between -10 and 10 and has solutions between those same values for the division problems. Examples include -3x = 12 and x + 5 = -3. A student's score was the maximum number of problems they could answer correctly in one minute. Students had to get an answer correct before moving on to a new problem. One goal was to see if the average number of problems answered correctly in one minute was greater for students who passed the class than for those who didn't pass.

The hypotheses that will be tested are:

$H_0 : \mu_{\text{pass}} = \mu_{fall}$
$H_0 : \mu_{\text{pass}} > \mu_{fall}$
$\alpha = 0.05$

a. Complete the design layout table.

Research Design Table		
Research Question:		
Type of Research		Observational Study Observational Experiment Manipulative Experiment
What is the response varialbe?		

What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables.	
Grouping/explanatory Variable 1 (if present)	Levels:

b. If two intermediate algebra classes were to be randomly selected from 12 classes being offered, with the classes being numbered 1 to 12, which two classes would be selected if the calculator was seeded with 27 or row 10 was used in the table of random digits?

c. What type of sampling method is used when a class is selected and everyone in the class participates in the research?

d. The data that will be gathered is the number of problems answered correctly in one minute. Are these data quantitative discrete, quantitative continuous or categorical?

The automaticity scores for the students who failed the class are shown in the table below.

16	15	14	15
13	16	22	30
20	8	13	14
16	16	16	16
6	27	9	

The automaticity scores for the students who passed the class are shown in the table below.

20	19	33	15	20
14	9	11	12	17
8	20	31	38	29
9	22	31	31	30
9	22	31	31	30
15	10	10	23	22
7	11	23	17	19
20	20	18	9	27
25	15	18	9	27
25	15	23	28	11
20	13	36	34	

e. Make a frequency distribution, double bar histogram and side-by-side box plots to show a graphical comparison of these two sets of scores.

f. Which graph is more effective in helping you see the difference between the data sets?

g. Find the mean, variance and standard deviation for both sets of data separately. You may use the statistical functions of your calculator.

h. The p-value of the statistical test that compares the two means is 0.0395. Write a concluding sentence in the style used in

scholarly journals (like you were taught in Chapter 1).

   i. Based on the results of this analysis and the decision rule in the story, will the college develop a program to help improve automaticity?

7. Why Statistical Reasoning Is Important for a Biology Student and Professional Developed in collaboration with Elysia Mbuja and Robert Thissen, Biology Department This topic is discussed in BIOL 160, General Biology.

To explore the scientific method, students will study the effect of alcohol on a Daphnia. Daphnia, living water fleas, are used because they are almost transparent and the beating heart can be seen. The theory to be tested is whether alcohol slows the heart rate of Daphnia. To conduct this test, a Daphnia will be placed in a drop of water on a microscope. The number of heartbeats in 15 seconds will be counted. The water will be removed from the slide and a drop of 8% alcohol will be placed on the Daphnia. After 1 minute, the heartbeats will be counted again. If the heartbeats are lower, it cannot be concluded that the reason is because of the alcohol. It could simply be the reaction to a drop of fluid being placed on the Daphnia or the effect of being on a slide under a light. Therefore, after the Daphnia is allowed to recover, it is returned to the slide following the exact same procedure except a drop of water is used instead of alcohol.

$H_0 : \mu_{\text{alcohol}} = \mu_{\text{water}}$
$H_1 : \mu_{\text{alcohol}} < \mu_{\text{water}}$
$\alpha = 0.05$

a. Complete the experiment design table.

Research Design Table	
Research Question:	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response varialbe?	
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables.	
Grouping/explanatory Variable 1 (if present)	Levels:

b. Make an appropriate graph to compare the two sets of data. The data in the shaded cells is authentic. It comes from a BIOL 160 class.

Heart Rate after 8% Alcohol – beats/15s			
36	23	40	33
37	16	26	32
40	32	17	55
33	23	22	32
51	43		

Heart Rate after 4 drops of Water - beats/15s			
45	42	54	55
70	22	50	64
42	62	67	62
54	29	62	62
44	52		

c. Show the relevant statistics for the two sets of data.

	Heart Rate after Alcohol	Heart Rate after Water
Mean		
Standard Deviation		

Median		

d. The p-value from the t-test for 2 independent populations is 1.28E-5. Write a concluding sentence.

e. What is the effect of alcohol on the heart rate of a Daphnia? Do you think it will have the same effect on a human?

4: Inferential Theory

Prior to elections, pollsters will survey approximately 1000 people and on the basis of their results, try to predict the outcome of the election. On the surface, it should seem like an absurdity that the opinions of 1000 can give any indication about the opinion of 100,000,000 voters in a national presidential election. Likewise, taking 100 or 1000 water samples from the Puget Sound, when that is a miniscule amount of water compared to the total amount of constantly changing water in the Sound, should seem insufficient for making a decision.

The objective of this chapter is to develop the theory that helps us understand why a relatively small sample size can actually lead to conclusions about a much larger population. The explanation is different for categorical and quantitative data. We will begin with categorical data.

The journey that you will take through this section has a final destination at the formula that will ultimately be used to test hypotheses. While you might be willing to accept the formula without taking the journey, it will be the journey that gives meaning to the formula. Because data are stochastic, that is, they are subject to randomness, probability plays a critical role in this journey.

Our journey begins with the concept of inference. **Inference** means that a small amount of observed information is used to draw general conclusions. For example, you may visit a business and receive outstanding customer service from which you infer that this business cares about its customers. Inference is used when testing a hypothesis. A small amount of information, the sample, is used to draw a conclusion, or infer something about the entire population.

The theory begins with finding the probability of getting a particular sample and ultimately ends with creating distributions of all the sample results that are possible if the null hypothesis is true. For each step of the 7-step journey, digressions will be made to learn about the specific rules of probability that contribute to the inferential theory.

Before starting, it is necessary to clarify some of the terminology that will be used. Regardless of the question being asked, if it produces categorical data, that data will be identified generically as a success or failure, without using those terms in their customary manner. For example, a researcher making a hypothesis about the proportion of people who are unemployed would consider being unemployed a success from the statistical point of view, and being employed as a failure, even though that contradicts the way it is viewed in the real world. Thus, success is data values about which the hypotheses are written and failure is the alternate data value.

> **Briefing 4.1 Self-driving Cars**
>
> Google, as well as most car companies, are developing self-driving cars. These autonomous cars will not need to have a driver and are considered less likely to be in an accident than cars driven by humans. Cars such as these are expected to be available to the public around the years 2020 – 2025. There are many questions that must be answered before these cars are made available. One such question is to determine who is responsible in the event of an accident. Is the owner of the car responsible, even though they were not steering the car or is the manufacturer responsible since their technology did not avoid the accident? (mashable.com/2014/07/07/drive...-cars-tipping- point/, viewed July 2014).

Step 1 – How likely is it that a particular data value is success?

Suppose a researcher wanted to determine the proportion of the public that believe the owner is responsible for the accident. The researcher has a hypothesis that the proportion is over 60%. In this case, the hypotheses will be:

- $H_0 : p = 0.60$
- $H_1 : p > 0.60$

When collecting data, the order in which the units or people are selected determines the order in which the data is collected. In this case, assigning responsibility to the owner will be considered a success and assigning responsibility to the manufacturer is considered a failure. If the first person believes the owner is responsible, the second person believes the manufacturer is responsible, the third person selects the manufacturer, the fourth, fifth and sixth people all select the owner as the responsible party, then we can convert this information to successes and failure by listing the order in which the data were obtained as SFFSSS.

The strategy that is employed to determine which of two competing hypotheses is better supported is always to assume that the null hypothesis is true. This is a critical point, because if we can assume a condition is true, then we can determine the probability of getting our particular sample result, or more extreme results. This is a p-value.

However, before we can determine the probability of obtaining a sequence such as SFFSSS, we must first find the probability of obtaining a success. For this, we need to explore the concept of probability.

Probability is the chance that a particular outcome will happen if a process is repeated a very large number of times. It is quantified by dividing the number of favorable outcomes by the number of possible outcomes. This is shown as a formula:

$$P(A) = \frac{\text{Number of Favorable Outcomes}}{\text{Number of Possible Outcomes}} \qquad (4.1)$$

where $P(A)$ means the probability of some event called A. This formula assumes that all outcomes are equally likely, which is what happens with a good random sampling process. The entire set of possible outcomes is called the sample space. The number of possible outcomes is the same as the number of elements in the sample space.

While the intent of this chapter is to focus on developing the theory to test hypotheses, a few concepts will be explained initially with easier examples.

If we wanted to know the probability of getting a tail when we flip a fair coin, then we must first consider the sample space, which would be {H, T}. Since there is one element in that sample space that is favorable and the sample space contains two elements, the probability is $p(tails) = \frac{1}{2}$.

To find the probability of getting a 4 when rolling a fair die, we create the sample space with six elements {1,2,3,4,5,6}, since these are the possible results that can be obtained when rolling the die. To find the probability or rolling a 4, we can substitute into the formula to get $P(4) = \frac{1}{6}$.

A more challenging question is to determine the probability of getting two heads when flipping two coins or flipping one coin twice. The sample space for this experiment is {HH, HT, TH, TT}. The probability is (HH) = $\frac{1}{4}$. The probability of getting one head and one tail, in any order is (1 head and 1 tail) = $\frac{2}{4}$ = $\frac{1}{2}$.

Probability will always be a number between 0 and 1, thus $0 \leq P(A) \leq 1$. A probability of 0 means something cannot happen. A probability of 1 is a certainty.

Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

If we assume that the null hypothesis is true, then the proportion of people who believe the owner is responsible is 0.60. What does that mean? It means that if there are exactly 100 people, then exactly 60 of them hold the owner responsible and 40 of them do not.

If our goal is to find the probability of SFFSSS, then we must first find the probability of getting a success (owner). If there are 100 people in the population and 60 of these select the owner, then the probabiltiy of selecting a person who chooses the owner is $P(owner) = \frac{60}{100} = 0.60$. Notice that this probability exactly equals the proportion defined in the null hypothesis. This is not a coincidence and it will happen every time because the proportion in the null hypothesis is used to generate a theoretical population, which was used to find the probability. The first important step in the process of testing a hypothesis is to realize that the probability of any data being a success is equal to the proportion defined in the null hypothesis, assuming that sampling is done with replacement, or that the population size is extremely large so that removing a unit from the population does not change the probability a significant amount.

Example 1

- If $H_0 : p = 0.35$, then the probability the 5th value is a success is 0.35.
- If $H_0 : p = 0.82$, then the probability the 20th value is a success is 0.82.

Step 2 - How likely is it that a particular data value is a failure?

Now that we know how to find the probability of success, we must find the probability of failure. For this, we will again digress to the rules of probability.

Digression 2 - Probability of A or B

When one unit is selected from a population, there can be several possible measures that can be taken on it. For example, a new piece of technology could be put into several brands of cars and then tested for reliability. The contingency table below shows the number of cars in each of the categories. These numbers are fictitious and we will pretend it is the entire population of cars under development.

	Hyundai	Nissan	Google	Total
Reliable	80	75	90	245
Not Reliable	25	10	20	55
Total	105	85	110	Grand Total 300

From this we can ask a variety of probability questions.

If one car is randomly selected, what is the probability that it is a Nissan?

$$P(Nissan) = \frac{85}{300} = 0.283$$

If one car is randomly selected, what is the probability that the piece of technology was not reliable?

$$P(not reliable) = \frac{55}{300} = 0.183$$

If one car is randomly selected, what is the probability that it is a Hyundai or the piece of technology was reliable?

This question introduces the word "or" which means that the car has one characteristic or the other characteristic or both. The word "or" is used when only one selection is made. The table below should help you understand how the formula will be derived.

Notice that the 2 values in the Hyundai column are circled and the three values in the Reliable row are circled, but that the value in the cell containing the number 80 that represents the Hyundai and Reliable is circled twice. We don't want to count those particular cars twice so after adding the column for Hyundai to the row for Reliable, it is necessary to subtract the cell for Hyundai and Reliable because it was counted twice but should only be counted once. Thus, the equation becomes:

P(Hyundai or Reliable) = P(Hyundai) + P(Reliable) – P(Hyundai and Reliable)

$$= \frac{105}{300} + \frac{245}{300} - \frac{80}{300} = \frac{270}{300} = 0.90$$

From this we will generalize the formula to be

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{4.2}$$

What happens if we use the formula to determine the probability of randomly selecting a Nissan or a Google car?

Because these two criteria cannot both happen at the same time, they are said to be mutually exclusive. Consequently, their intersection is 0.

P(Nissan or Google) = P(Nissan) + P(Google) – P(Nissan and Google)

$$= \frac{85}{300} + \frac{110}{300} - \frac{0}{300} = \frac{195}{300} = 0.65$$

Because the intersection is 0, this leads to a modified formula for categorical data values that are mutually exclusive.

$$P(A \text{ or } B) = P(A) + P(B) \tag{4.3}$$

This is the formula that is of primary interest to us for determining how to find the probability of failure.

If success and failure are the only two possible results, and it is not possible to simultaneously have success and failure, then they are mutually exclusive. Furthermore, if a random selection is made, than it is certain that it will be a success or failure. Things that are certain have a probability of 1. Therefore, we can write the formula using S and F as:

$$P(S \text{ or } F) = P(S) + P(F)$$

or

$$1 = P(S) + P(F)$$

with a little algebra this becomes

$$1 - P(S) = P(F) \tag{4.4}$$

What this means is that subtracting the probability of success from 1 gives the probability of failure. The probability of failure is called the complement of the probability of success.

Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

Recall that the original hypothesis for the responsibility in an accident is that: H0: p = 0.60. We have established that the probability of success is 0.60. The probability of failure is 0.40 since it is the complement of the probability of success and *1 − 0.60 = 0.40*.

> **Example 2**
>
> - If $H_0 : p = 0.35$, then the probability the 5th value is a success is 0.35. The probability the 5th value is a failure is 0.65.
> - If $H_0 : p = 0.82$, then the probability the 20th value is a success is 0.82. The probability the 20th value is a failure is 0.18.

Step 3 - How likely is it that a sample consists of a specific sequence of successes and failures?

We now know that the probability of success is identical to the proportion defined by the null hypothesis and the probability of failure is the complement. But these probabilities apply to only one selection. What happens when more than one is selected? To find that probability, we must learn the last of the probability rules:

$$P(A \text{ and } B) = P(A)P(B) \tag{4.5}$$

> **Digression 3 - P(A and B) = P(A)P(B)**
>
> If two or more selections are made, the word "and" becomes important because it indicates we are seeking one result for the first selection and one result for the second selection. This probability is found by multiplying the individual probabilities. Part of the key to choosing this formula is to identify the problem as an "and" problem. For instance, early in this chapter we found the probability of getting two heads when flipping two coins is 0.25. This problem can be viewed as an "and" problem if we ask "what is the probability of getting a head on the first flip and a head on the second flip"? Using subscripts of 1 and 2 to represent the first and second flips respectively, we can rewrite the formula to show:
>
> $$P(H_1 \text{ and } H_2) = P(H_1)P(H_2) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4} = 0.25.$$
>
> Suppose the researcher randomly selects three cars. What is the probability that there will be one car of each of the makes (Hyundai, Nissan, Google)?
>
> | | Hyundai | Nissan | Google | Total |
> |---|---|---|---|---|
> | Reliable | 80 | 75 | 90 | 245 |
> | Not Reliable | 25 | 10 | 20 | 55 |

	Hyundai	Nissan	Google	Total
Total	105	85	110	Grand Total 300

First, since there are three cars being selected, this should be recognized as an "and" problem and can be phrased P(Hyundai and Nissan and Google). Before we can determine the probability however, there is one important question that must be asked. That question is whether the researcher will select with replacement.

If the researcher is sampling with replacement, then the probability can be determined as follows.

P(Hyundai and Nissan and Google) = P(Hyundai)P(Nissan)P(Google) =

$(\frac{105}{300})(\frac{85}{300})(\frac{110}{300}) = 0.03636$.

Notice the slight change in the probability as a result of not using replacement.

Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

We are now ready to answer the question of what is the probability that we would get the exact sequence of people if the first person believes the owner is responsible, the second person believes the manufacturer is responsible, the third person selects the manufacturer, the fourth, fifth and sixth people all select the owner as the responsible party. Because there are six people selected, then this is an "and" problem and can be written as P(S and F and F and S and S and S) or more concisely, leaving out the word "and" but retaining it by implication, we write P(SFFSSS). Remember that P(S) = 0.6 and P(F) = 0.4

P(SFFSSS) = P(S)P(F)P(F)P(S)P(S)P(S) = (0.6)(0.4)(0.4)(0.6)(0.6)(0.6) = 0.0207.

To summarize, if the null hypothesis is true, then 60% of the people believe the owner is responsible for accidents. Under these conditions, if a sample of six people is taken, with replacement, then the probability of getting this exact sequence of successes and failures is 0.0207.

Step 4 - How likely is it that a sample would contain a specific number of successes?

Knowing the probability of an exact sequence of successes and failures is not particularly important by itself. It is a stepping-stone to a question of greater importance – what is the probability that four out of six randomly selected people (with replacement) will believe the owner is responsible? This is an important transition in thinking that is being made. It is the transition from thinking about a specific sequence to thinking about the number of successes in a sample.

When data are collected, researchers don't care about the order of the data, only how many successes there were in the sample. We need to find a way to transition from the probability of getting particular sequence of successes and failures such as SFFSSS to finding the probability of getting four successes from a sample of size 6. This transition will make use of the commutative property of multiplication, the P(A or B) rule and combinatorics (counting methods).

At the end of Step 3 we found that P(SFFSSS) = 0.0207.

- What do you think will be the probability of P(SSSFSF)?
- What do you think will be the probability of P(SSSSFF)?

The answer to both these questions is 0.0207 because all of these sequences contain 4 successes and two failures and since the probability is found by multiplying the probabilities of success and failure in sequence and since multiplication is commutative (order doesn't matter) then

(0.6)(0.4)(0.4)(0.6)(0.6)(0.6)= (0.6)(0.6)(0.6)(0.4)(0.6)(0.4)= (0.6)(0.6)(0.6)(0.6)(0.4)(0.4) = 0.020736.

If the question now changes from what is the probability of a sequence to what is the probability of 4 successes in a sample of size 6, then we have to consider all the different ways in which four successes can be arranged. We could get 4 successes if the sequence of our selection is SFFSSS or SSSFSF or SSSSFF or numerous other possibilities. Because we are sampling one time and because there are many possible outcomes we could have, this is an "or" problem that uses an expanded version of the formula P(A or B) = P(A) + P(B). This can be written as:
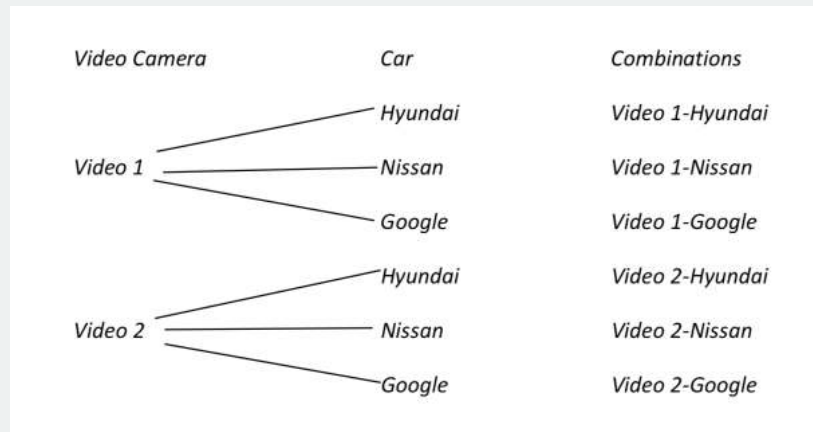
P(4 out of 6) = P(SFFSSS or SSSFSF or SSSSFF or ...) = P(SFFSSS) + P(SSSFSF) + P(SSSSFF) + ...

In other words, we can add the probability of each of these orders. However, since the probability of each of these orders is the same (0.0207) then this process would be much quicker if we simply multiplied 0.0207 by the number of orders that are possible. The question we must answer then is how many ways are there to arrange four successes and 2 failures? To answer this, we must explore the field of combinatorics, which provide various counting methods.

Researchers designing the cars will compare different technologies to see which works better. Suppose two brands of a video camera are available for a car. How many different pairs are possible?

A tree-diagram can help explain this.

Video Camera	Car	Combinations
	Hyundai	Video 1-Hyundai
Video 1	Nissan	Video 1-Nissan
	Google	Video 1-Google
	Hyundai	Video 2-Hyundai
Video 2	Nissan	Video 2-Nissan
	Google	Video 2-Google

Making a tree-diagram to answer questions such as this can be tedious, so an easier approach is to use the **fundamental counting rule** which states that if there are M options for one choice that must be made and N options for a second choice that must be made, then there are MN unique combinations. One way to show this is to draw and label a line for each choice that must be made and on the line write the number of options that are available. Multiply the numbers.

_____2_____ _____3_____ = 6

Videos Cars

This tells you there are six unique combinations for one camera and one make of car.

If researchers have 4 test vehicles that will be driving on the freeway as a convoy and the colors of the vehicles are blue, red, green, and yellow, how many ways can these cars be ordered in the convoy?

To answer this question, think of it as having to make four choices, which color of car is first, second, third, fourth. Draw a line for each choice and on the line write the number of options that are available and then multiply these numbers. There are four options for the first car. Once that choice is made there are three options remaining for the second car. When that choice is made, there are two options remaining for the third car. After that choice is made, there is only one option available for the final car.

4 3 2 1 = 24 unique orders

First Car Second Car Third Car Fourth Car

Examples of some of these unique orders include: blue, red, green, and yellow
red, blue, green, and yellow
green, red, blue, and yellow

Each unique sequence is called a permutation. Thus in this situation, there are 24 permutations.

The way to find the number of permutations when all available elements are used is called factorial. In this problem, all four cars are used, so the number of permutations is 4 factorial which is shown symbolically as 4!. 4! means (4)(3)(2)(1). To be more general,

$$n! = n(n-1)(n-2)\ldots 1 \qquad (4.6)$$

Permutations can also be found when fewer elements are used than are available. For example, suppose the researchers will only use two of the four cars. How many different orders are possible? For example, the blue car followed by the green car is a different order than the green car followed by the blue car. We can answer this question two ways (and hopefully get the same answer both ways!). The first way is to use the fundamental counting rule and draw two lines for the choices we make, putting the number of options available for each choice on the line and then multiplying.

$\underline{4}\ \underline{3}$ = 12 permutations

First Car Second Car

Examples of possible permutations include:

Blue, Green, Green, Blue, Blue, Red Yellow, Green

The second approach is to use the formula for permutations when the number selected is less than or equal to the number of available. In this formula, r represents the number of items selected, n represents the number of items available.

$$nPr = \frac{n!}{(n-r)!} \tag{4.7}$$

For this example, n is 4 and r is 2 so with the formula we get:

$${}_4P_2 = \frac{4!}{(4-2)!} = \frac{4!}{2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 4 \cdot 3 = 12 \quad \text{permutations.}$$

Notice the final product of $4 \cdot 3$ is the same as we have when using the fundamental counting rule. The denominator term of (n-r)! is used to cancel the unneeded numerator terms.

For permutations, order is important, but what if order is not important? For example, what if we wanted to know how many pairs of cars of different colors could be combined if we didn't care about the order in which they drove. In such a case, we are interested in combinations. While Blue, Green, and Green, Blue represent two permutations, they represent only one combination. There will always be more permutations than combinations. The number of permutations for each combination is r!. That is, when 2 cars are selected there are 2! permutations for each combination.

To determine the number of combinations there are if two of the four cars are selected we can divide the total number of permutations by the number of permutations per combination.

$$Number\ of\ combinations = Number\ of\ permutations(\frac{1\ Combination}{Number\ of\ Permutations})$$

Using similar notation as was used for permutations (nPr), combinations can be represented with nCr, so the equation can be rewritten as

$$
\begin{aligned}
nCr &= nPr(\frac{1}{r!})\ or \\
nCr &= \frac{n!}{(n-r)!}(\frac{1}{r!}) \\
nCr &= \frac{n!}{(n-r)!r!}
\end{aligned}
\tag{4.8}
$$

An alternate way to develop this formula that could be used for larger sample sizes that contain successes and failures is to consider that the number of permutations is n! while the number of permutations for each combination is r!(n-r)!. For example, in a sample of size 20 with 12 successes and 8 failure, there are 20! permutations of the successes and failures combined with 12! permutations of successes and 8! permutations of failures for each combination. Thus,

$$Number\ of\ combinations = 20!(\frac{1\ Combination}{12!8!}) = \frac{20!}{12!8!}\ or\ as\ a\ formula:$$

$$Number\ of\ combinations = n!(\frac{1\ Combination}{r!(n-r)!}) = \frac{n!}{r!(n-r)!}\ or\ \frac{n!}{(n-r)!r!}\ .$$

For our example about car colors we have:

$$_4C_2 = \frac{4!}{(4-2)!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6 \text{ combinations.}$$

This sequence of combinatorics concepts has reached the intended objective in that the interest is in the number of combinations of successes and failure there are for a given number of successes in a sample. We will now return to the problem of the responsibility for accidents.

Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

Recall that 6 people were selected and 4 thought the owner should be responsible. We saw that the probability of any sequence of 4 successes and 2 failures, such as SFFSSS or SSSFSF or SSSSFF is 0.0207 if the null hypothesis is p = 0.60. If we knew the number of combinations of these 4 successes and 2 failures, we could multiply that number times the probability of any specific sequence to get the probability of 4 successes in a sample of size 6.

Using the formula for nCr, we get: $_6C_4 = \frac{6!}{(6-4)!4!} = 15$ *combinations.*

Therefore, the probability of 4 successes in a sample of size 6 is 15*0.020736 = 0.31104. This means that if the null hypothesis of p=0.60 is true and six people are asked, there is a probability of 0.311 that four of those people will believe the owner is responsible.

We are now ready to make the transition to distributions. The following table summarizes our journey to this point.

Step 1	Use the null hypothesis to determine P(S) for any selection, assuming replacement.
Step 2	Use the P(A or B) rule to find the complement, which is the P(F) = 1 - P(S)
Step 3	Use the P(A or B) rule to find the probability of a specific sequence of a specific sequence of successes and failures, such as SFFSSS, by multiplying the individual probabilities.
Step 4	Recognize that all combinations of r successes out of a sample of size n have the same probability of occurring. Find the number of combinations nCr and multiply this times the probability of any of the combinations to determine the probability of getting r successes out of a sample of size n.

Step 5 – How can the exact p-value be found using the binomial distribution?

Recall that in chapter 2, we determined which hypothesis was supported by finding the p-value. If the p-value was small, less than the level of significance, we concluded that the data supported the alternative hypothesis. If the p-value was larger than the level of significance, we concluded that the data supported the null hypothesis. The p-value is the probability of getting the data, <u>or more extreme data</u>, assuming the null hypothesis is true.

In Step 4 we found the probability of getting the data (for example, four successes out of 6) but we haven't found the probability of getting more extreme data yet. To do so, we must now create distributions. A distribution shows the probability of all the possible outcomes from an experiment. For categorical data, we make a discrete distribution.

Before looking at the distribution that is relevant to the problem of responsibility for accidents, a general discussion of distributions will be provided.

In chapter 4 you learned to make histograms. Histograms show the distribution of the data. In chapter 4 you also learned about means and standard deviations. Distributions have means and standard deviations too.

To demonstrate the concepts, we will start with a simple example. Suppose that someone has two routes used for running. One route is 2 miles long and the other route is 5 miles long. Below is the running schedule for last week.

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday

A distribution of the amount run each day is shown below.
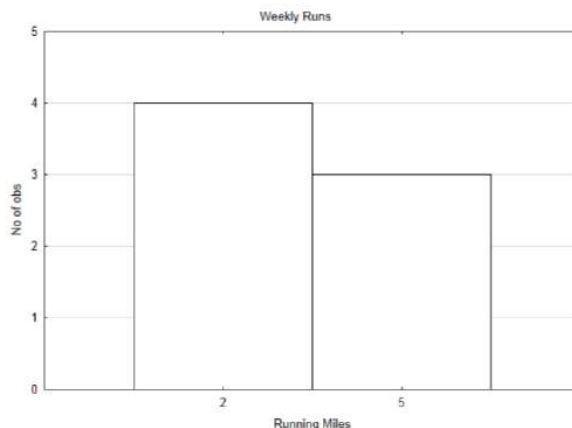
The mean can be found by adding all the daily runs and dividing by 7. The mean is 3.286 miles per day. Because the same distances are repeated on different days, a weighted mean can also be used. In this case, the weight is the number of times a particular distance was run. A weighted mean takes advantage of multiplication instead of addition. Thus, instead of calculating: $\frac{2+2+2+2+5+5+5}{7} = 3.286$. we can multiply each number by the number of times it occurs then divide by the number of occurrences: $\frac{4 \cdot 2 + 3 \cdot 5}{4 + 3} = 3.286$. The formula for a weighted mean is:

$$\frac{\sum w \cdot x}{\sum w} \tag{4.9}$$

The same graph is presented below, but this time there are percentages above the bars.

Instead of using counts as the weight, the percentages (actually the proportions) can be used as the weight. Thus 57.143% can be written as 0.57143. Likewise, 42.857% can be written 0.42857.

Substituting into the formula gives: $\frac{0.57143 \cdot 2 + 0.42857 \cdot 5}{0.57143 + 0.42857} = 3.286$. Notice that the denominator adds to 1. Therefore, if the weight is the proportion of times that a value occurs, then the mean of a distribution that uses percentages can be found using the formula:

$$\mu = \sum P(x)x \tag{4.10}$$
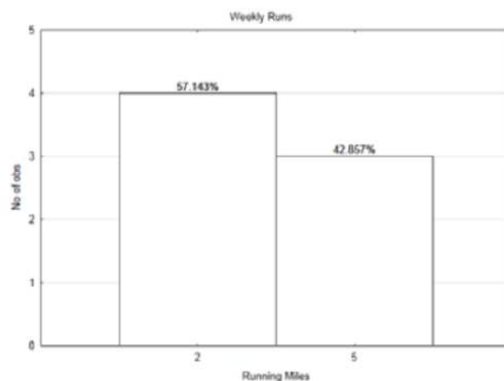
This mean, which is also known as the expected value, is the sum of the probability of a value times the value. There is no need for dividing, as is customary when finding means, because we would always just divide by 1.

Recall from chapter 4 that the standard deviation is the square root of the variance. The variance is $\sigma^2 = \sum [(x - \mu)^2 \cdot P(x)]$. The standard deviation is $\sigma = \sqrt{\sum [(x - \mu)^2 \cdot P(x)]}$.

$$\sigma = \sqrt{\sum [(x - \mu)^2 \cdot P(x)]} \, .$$
$$\sigma = \sqrt{(2 - 3.286)^2 \cdot 0.57143 + (5 - 3.286)^2 \cdot 0.42857} = 1.485$$

The self-driving car problem will show us one way in which we encounter discrete distributions. In fact, it will result in the creation of a special kind of discrete distribution called the Binomial distribution, which will be defined after exploring the concepts that lead to its creation.

When testing a hypothesis about proportions of successes, there are two random variables that are of interest to us. The first random variable is specific to the data that we will collect. For example, in research about who is responsible for accidents caused by autonomous cars, the random variable would be "responsible party". There would be two possible values for this random variable, owner or car manufacturer. We have been considering the owner to be a success and the manufacturer to be a failure. The data the researchers collect is about this random variable. However, creating distributions and finding probabilities and p-values requires a shift of our focus to a different random variable. This second random variable is about the number of successes in a sample of size n. In other words, if six people are asked, how many of them think the owner is responsible? It is possible that none of them think the owner is responsible, or one thinks the owner is responsible, or two, or three, or four, or five, or all six think the owner is responsible. Therefore, in a sample of size 6, the random variable for the number of successes can have the values of 0,1,2,3,4,5,6. We have already found that the probability of getting four successes is 0.3110. We will now find the probability of getting 0,1,2,3,5,6 successes, assuming the hypotheses are still $H_0 : p = 0.60$, $H_1 : p > 0.60$. This will allow us to create the discrete binomial distribution of all possible outcomes.

Find the probability of 0 successes

The only way to have 0 successes is to have all failures, thus we are seeking P(FFFFFF).

P(FFFFFF) = P(F)P(F)P(F)P(F)P(F)P(F) = (0.4)(0.4)(0.4)(0.4)(0.4)(0.4) = 0.004096.

Since there is only one combination for 0 successes, then the probability of 0 successes is 0.0041.

Find the probability of 1 success

We know that all combinations have the same probability, so we may as well create the simplest combination for 1 success. This would be SFFFFF.

P(SFFFFF) = P(S)P(F)P(F)P(F)P(F)P(F) = (0.6)(0.4)(0.4)(0.4)(0.4)(0.4) = $(0.6)^1 (0.4)^5$ = 0.006144.

How many combinations are there for 1 success? This can be found using $_6C_1$.

$_6C_1 = \dfrac{6!}{(6-1)!1!} = 6$ combinations. Does this answer seem reasonable? Consider there are only 6 places in which the success could happen.

The probability of 1 success is then 6*0.006144 = 0.036864 or if we round to four decimal places, 0.0369.

Find the probability of 2 successes.

Instead of doing this problem in steps as was done for the prior examples, it will be demonstrated by combining steps.

$_6C_2 P(\text{SSFFFF})$

$$\dfrac{6!}{(6-2)!2!}(0.6)(0.6)(0.4)(0.4)(0.4)(0.4) = \dfrac{6!}{(6-2)!2!}(0.6)^2(0.4)^4 =$$

15(0.009216) = 0.13824 or with rounding to four decimal places 0.1382.

Find the probability of 3 successes using the combined steps. (Now its your turn).

Find the probability of 5 successes.

Find the probability of 6 successes.

When all the probabilities have been found, we can create a table that shows the values the random variable can take and their probabilities. We will define the random variables for the number of successes as X with the possible values defined as x.

X = x	0	1	2	3	4	5	6
P(X = x)	0.0041	0.0369	0.1382	0.2765	0.3110	0.1866	0.0467

Do your values for 3,5, and 6 successes agree with the values found in the table?

A graph of this distribution can lead to a better understanding of it. This graph is called a probability mass function, which is shown using a stick graph. It is a way to graph discrete distributions, since there cannot be any values in between the numbers on the x-axis. The heights of the bar correspond to the probability of getting the number of successes.
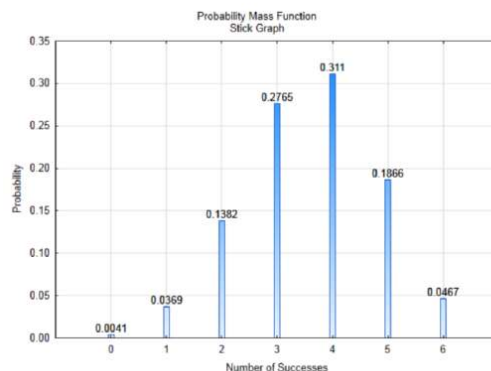
There are three things you should notice about this distribution. First, it is a complete distribution. That is, in a sample of size six, it is only possible to have 0,1,2,3,4,5, or 6 successes and all of those have been included in the graph. The second thing to notice is that all the probabilities have values between 0 and 1. This should be expected because probabilities must always be between 0 and 1. The final thing to notice, which may not be obvious at first, is that if you add all the probabilities, the sum will be 1. The sum of all complete probability distributions should equal 1, $\sum P(x) = 1$. If you add all the probabilities and they don't equal one, but are very close, it could be because of rounding, not because you did anything wrong.

Digression 5 - Binomial Distributions

The entire journey that has been taken since the beginning of this chapter has led to the creation of a very important discrete distribution called the binomial distribution, which has the following components.

1. A Bernoulli Trial is a sample that can have only two possible results, success and failure.
2. An experiment can consist of n independent Bernoulli Trials.
3. A Binomial Random Variable, X is the number of successes in an experiment
4. A Binomial Distribution shows all the values for X and the probability of each of those values occurring.

The assumptions are that:

1. All trials are independent.
2. The number of trials in an experiment is the same and defined by the variable n.
3. The probability of success remains constant for each sample. The probability of failure is the complement of the probability of success. The variable p = P(S) and the variable $q = P(F)$. $q = 1 - p$.
4. The random variable X can have values of 0, 1, 2,...n.

The probability can be found for each possible number of successes that the random variable X can have using the binomial distribution formula

$$P(X = x) =_n C_x P^x q^{n-x} \tag{4.11}$$

.

If this formula looks confusing, review the work you did when finding the probability that 3,5 or 6 people believe the owner is responsible, because you were actually using this formula. $_nC_x$, which is shown in your calculator as $_nC_r$, is the number of combinations for x successes. The x and the r represent the same thing and are used interchangeably.

p is the probability of success. It comes from the null hypothesis.

q is the probability of failure. It is the complement of p.

n is the sample size

x is the number of successes

If we use this formula for all possible values of the random variable, X, we can create the binomial distribution and graph.

$$P(X = 0) =_6 C_0 (0.60)^0 (0.40)^{6-0} = 0.0041$$
$$P(X = 1) =_6 C_1 (0.60)^1 (0.40)^{6-1} = 0.0369$$
$$P(X = 2) =_6 C_2 (0.60)^2 (0.40)^{6-2} = 0.1382 \text{You can finish the rest of them.}$$

The TI84 calculator has an easier way to create this distribution. Find and press your Y= key. The cursor should appear in the space next to Y1 = . Next, push the 2nd key, then the key with VARS on it and DISTR above it. This will take you to the collection of distributions. Scroll up until you find Binompdf. This is the binomial probability distribution function. Select it and then enter the three values n, p, x. For example, if you enter Y1=Binompdf(6,0.60,x) and then select 2nd TABLE, you should see a table that looks like the following:

X Y1
0 0.0041
1 0.03686
2 0.13824
3 0.27648
4 0.31104
5 0.18662
6 0.04666

If the table doesn't look like this, press 2nd TBLSET and make sure your settings are:

TblStart = 0
Δ TBL = 1
Indpnt: Auto
Depend: Auto.

Binomial distributions have a mean and standard deviation. The approach for finding the mean and standard deviation of a discrete distribution can be applied to a binomial distribution.

$X = x$	0	1	2	3	4	5	6
$P(x = x)$	0.0041	0.0369	0.1382	0.2765	0.3110	0.1866	0.0467
$x(P(x))$	0	0.0369	0.2764	0.8295	1.244	0.933	0.2802
$(x - \mu)^2$	$(0 - 3.6)^2 = 12.96$ 6	2.56	0.36	0.16	1.96	5.76	
$(x - \mu)^2 \cdot P(x)$	0.0531	0.2494	0.3538	0.0995	0.0498	0.3657	0.2690

$$\mu = \sum P(x)x$$
$$\mu = 0 + 0.0369 + 0.2764 + 0.8295 + 1.244 + 0.933 + 0.2802 = 3.6$$
$$\sigma = \sqrt{\sum [(x - \mu)^2 \cdot P(x)]}$$
$$\sigma = \sqrt{0.0531 + 0.2494 + 0.3538 + 0.0995 + 0.0498 + 0.3657 + 0.2690} = \sqrt{1.4403} = 1.20$$

The mean is also called the expected value of the distribution. Finding the expected value and standard deviation for using these formulas can be very tedious. Fortunately, for the binomial distribution, there is an easier way. The expected value can be found with the formula:

$$E(x) = \mu = np \tag{4.12}$$

The standard deviation is found with the formula

$$\sigma = \sqrt{npq} = \sqrt{np(1-p)} \tag{4.13}$$

To determine the mean number of people who think the owner is responsible for accidents, use the formula

$E(x) = \mu = np = 6\,(0.6) = 3.6$.

This indicates that if lots of samples of 6 people were taken the average number of people who believe the owner is responsible would be 3.6. It is acceptable for this average to not be a whole number.

The standard deviation of this distribution is: $\sigma = \sqrt{np(1-p)} = \sqrt{6(0.6)(0.4)} = 1.2$

Notice the same results were obtained with an easier process. Formulas 5.12 and 5.13 should be used to find the mean and standard deviation for all binomial distributions.

Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

Now that you have the ability to create a complete binomial distribution, you are ready to test a hypothesis. This will be demonstrated with the autonomous car example that has been used throughout this chapter.

Suppose a researcher wanted to determine the proportion of people who believe the owner is responsible. The researcher may have had a hypothesis that the proportion of people who believe the owner is responsible for accidents is over 60%. In this case, the hypotheses will be: H0: p = 0.60 and H1: p > 0.60. The level of significance will be 0.10 because only a small sample size will be used. In this case, the sample size will be 6.

With this sample size we have already seen what the binomial distribution will be like. We also know that the direction of the extreme is to the right because the alternative hypothesis uses a greater than symbol.

The researcher randomly selects 6 people. Of these, four say the owner is responsible. Which hypothesis is supported by this data?

The p-value is the probability the researcher would get four or more people claiming the owner is responsible. From the table we created earlier, we see the probability of getting 4 people who think

$X = x$	0	1	2	3	4	5	6
$P(X = x)$	0.0041	0.0369	0.1382	0.2765	0.3110	0.1866	0.0467

the owner is responsible is 0.3110, the probability of getting 5 is 0.1866 and the probability of getting 6 is 0.0467. If we add these together, we find the probability of getting 4 or more is 0.5443. Since this probability is larger than our level of significance, we conclude the data supports the null hypothesis and is therefore not significant. The conclusion that would be written is: At the 0.10 level of significance, the proportion of people who think the owner is responsible is not significantly greater than 0.60 ($x = 4$, $p = 0.5443$, $n = 6$). Remember that in statistical conclusions, the p is the p-value, not the sample proportion.

The TI84 has a quick way to add up the probabilities. It uses the function binomcdf, for binomial cumulative distribution function. It is found in the 2nd DISTR list of distributions. Binomcdf will add up all the probabilities beginning on the left, thus binomcdf(6,.6,1) will add the probabilities for 0 and 1. There are two conditions that are encountered when testing hypotheses using the binomial distribution. The way to find the p-value using binomcdf is based on the alternative hypothesis.

Condition 1. The alternative hypothesis has a less than sign (<).

Since the direction of the extreme is to the left, then using binomcdf(n,p,x) will produce the p value. The variable n represents the sample size, the variable p represents the probability of success (see null hypothesis), and x represents the specific number of successes from the data.

Condition 2. The alternative hypothesis has a greater than sign (>).

Since the direction of the extreme is to the right, it is necessary to use the complement rule and also reduce the value of x by 1, so enter $1 - \text{binomcdf}(n, p, x - 1)$ in your calculator. For example, if the data is 4, then enter $1 - \text{binomcdf}(6,0.6,3)$. Can you figure out why x – 1 is used and why binomcdf(n,p,x-1) is subtracted from 1? If not, ask in class.

In this example, the data were not significant and so the researcher could not claim the proportion of people who think the owner is responsible is greater than 0.60. A sample size of 6 is very small for categorical data and therefore it is difficult to arrive at any significant results. If the data are changed so that instead of getting 4 out of 6 people, the researcher gets 400 out of 600, does the conclusion change? Use 1 – binomcdf(600,0.6,399) to find the p-value for this situation.

1 – binomcdf(600,0.6,399) = _____

Write the concluding sentence:

Step 6 - How can the approximate p-value be found using the normal approximation to the binomial distribution?

When a hypothesis is tested using the binomial distribution, an exact p-value is found. It is exact because the binomial distribution is created from every combination of successes and failures that is possible for a sample of size n. There are other methods for determining the p-value that will give an approximate p-value. In fact, the typical method that is used to test hypotheses about proportions will give an approximate p-value. You may wonder why a method that gives an approximate p-value is used instead of the method that gives an exact p-value. This will be explained after the next two methods have been demonstrated. Before these can be demonstrated, we need to learn about a different distribution called the normal distribution.

Digression 6 – The Normal Distribution

Behind Pierce College is Waughop Lake, which is used by many students for learning scientific concepts outside of a classroom. The approximate shape of the lake is shown below. If one of the science labs required students to estimate the surface area of the water, what strategy could they use for this irregularly shaped lake?

A possible strategy is to think that this lake is almost a rectangle, and so they could draw a rectangle over it. Since a formula is known for the area of a rectangle, and if we know that each arrow below is 200 meters, can the area of the lake be estimated?
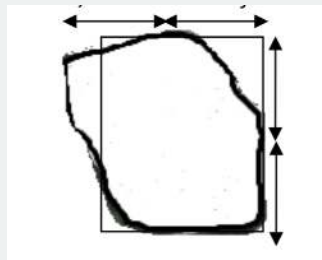
There are two important questions to consider. If this approach is taken, will the area of the lake exactly equal the area of the rectangle? Will it be close?

The answer to the first question is no, unless we happened to be extremely lucky with our drawing of the rectangle. The answer to the second question is yes it should be close.

The concept of approximating an irregular shape with a shape for which the properties are know is the strategy we will use to find new ways of determining a p-value. To the right is the irregular shape of a binomial distribution if n = 60, p = 0.60. The smooth curve that is drawn over the top of the bars is called the normal distribution. It also goes by the names bell curve and Gaussian distribution.

The formula for the normal distribution is $f(x, \mu, \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{[\frac{1}{2}(\frac{x-\mu}{\sigma})^2]}$. It is not important that you know this formula. What is important is to notice the variables in it. Both π and e are constants with the values of 3.14159 and 2.71828 respectively. The x is the independent variable, which is found along the x-axis. The important variables to notice are μ and σ, the mean and standard deviation. The implication of these two variables is that they play an important role in defining this curve. The function can be shown as $N(\mu, \sigma)$.

The binomial distribution is a discrete distribution whereas the normal distribution is a continuous distribution. It is known as a density function.

A normal distribution is contrasted with skewed distributions below.
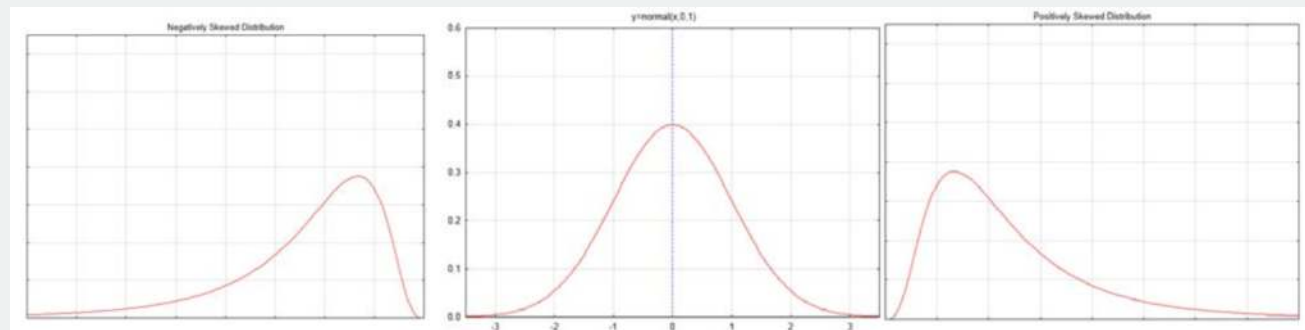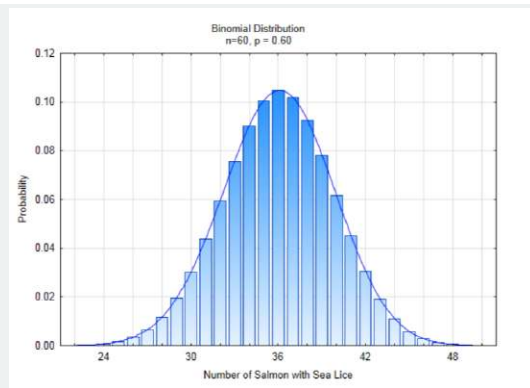
A negatively skewed distribution, such as is shown on the left, has some values that are very low causing the curve to be stretched to the left. These low values would cause the mean to be less than the median for the distribution. The positively skewed distribution, such as is shown on the right, has some values that are very high, causing the curve to be stretched to the right. These high values would cause the mean to be greater than the median for the distribution. The normal curve in the middle is symmetrical. The mean, median and mode are all in the middle. The mode is the high point of the curve.

The normal curve is called a density function, in contrast to the binomial distribution, which is a probability mass function. The space under the curve is called the area under the curve. The area is synonymous with the probability. The area under the entire curve, corresponding to the probability of selecting a value from anywhere in the distribution is 1. This curve never touches the x-axis, in spite of the fact that it looks like it does. Our ultimate objective with the normal curve is to find the area in the tail, which corresponds with finding the p-value.

We will start to think about the area (probability) under the curve by looking at the standard normal curve. The standard normal curve has a mean of 0 and a standard deviation of 1 and is shown as a function N(0,1). Notice that the x-axis of the curve is numbered with –3, -2, -1, 0, 1, 2, 3. These numbers are called z scores. They represent the number of standard deviations x is from the mean, which is in the middle of the curve.
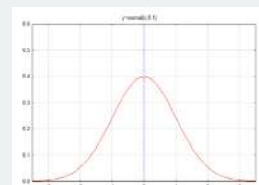
Does it seem reasonable that half of the curve is to the left of the mean and half the curve is to the right? We can label each side with this value, which is interpreted as both an area and a probability that a value would exist in that area.

Thinking about the area under the normal distribution is not as easy as thinking about the area under a uniform distribution. For example, we could create a uniform distribution for the outcome of an experiment in which one die is rolled. The probability of rolling any number is 1/6. Therefore the uniform distribution would look like this.

$\frac{1}{6}$

1	2	3	4	5	6
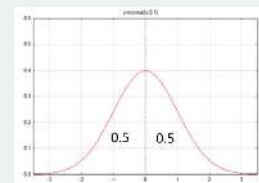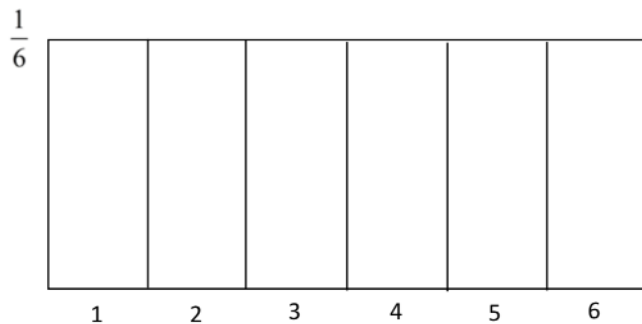
The area on this distribution can be found by multiplying the length by the width (height). Thus, to find the probability of getting a 5 or higher, we consider the length to be 2 and the width to be 1/6 so that $2 \times (\frac{1}{6} = \frac{1}{3}$. That is, there is a probability of 1/3 that a 5 or 6 would be rolled on the die.

But a normal distribution is not as familiar as a rectangle, for which the area is easier to find. The **Empirical Rule** is an approximation of the areas for different sections of the normal curve; 68% of the curve is within one standard deviation of the mean, 95% of the curve is within two standard deviations of the mean, and 99.7% of the curve is within three standard deviations of the mean.
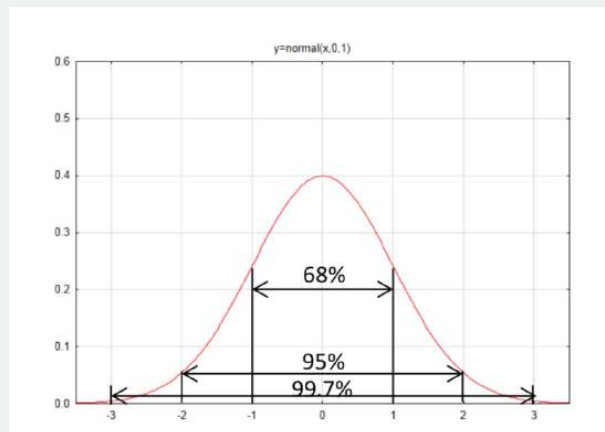
To find the area under a normal distribution was originally done using a technique called integration, which is taught in Calculus. However, these areas have already been found for the standard normal distribution N(0,1) and are provided in a table on the next page. The tables will always provide the area to the left. The area to the right is the complement of the area to the left, so to find the area to the right, subtract the area to the left from 1. A few examples should help clarify this.

Example 3. Find the areas to the left and right of z = -1.96.

Since the z value is less than 0, use the first of the two tables. Find the row with -1.9 in the left column and find the column with the 0.06 in the top row. The intersection of those rows and columns gives the area to the left, designated as A_L as 0.0250. The area to the right, designated as $A_R = 1 - 0.0250 = 0.9750$

Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287

Example 4. Find the areas to the left and right of z = 0.57.

Since the z value is greater than 0, use the second of the two tables. Find the row with 0.5 in the left column and find the column with 0.07 in the top row. The intersection of those rows and columns gives $A_L = 0.7157$, therefore $A_R = 1 - 0.7157 = 0.2843$

Standard Normal Distribution – N(0,1)

Area to the left when $z \leq 0$										
Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0446
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1334	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.2557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821

-0.2	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

Standard Normal Distribution – N(0,1)

Area to the left when $z \geq 0$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0..9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Since it is very unlikely that we will encounter authentic populations that are normally distributed with a mean of zero and a standard deviation of one, then of what use is this? The answer to this question has two parts. The first part is to answer the question about which useful populations are normally distributed. The second part is to determine how these tables can be used by other distributions with different means and standard deviations.

You have already seen that the normal curve fits very nicely over the binomial distribution. In chapters one and two you also saw distributions of sample proportions and sample means that look normally distributed. Therefore, the primary use of the normal distribution is to find probabilities when it is used to model other distributions such as the binomial distribution or the sampling distributions of $\hat{p}$ or $\bar{x}$. The following illustrate the elements of the distributions being modeled by the curve.
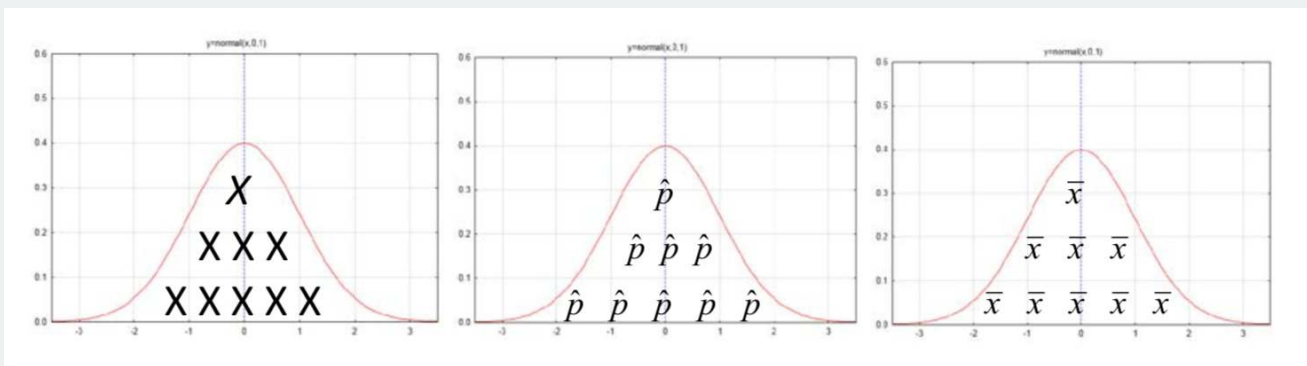
Now that some of the distributions that can be modeled with a normal curve have been established, we can address the second question, which is how to make use of the tables for the standard normal curve. Probabilities and more specifically p-values, can only be found after we have our sample results. Those sample results are part of a distribution of possible results that are approximately normally distributed. By determining the number of standard deviations our sample results are from the mean of the population, we can use the standard normal distribution tables to find the p-value. The transformation of sample results into standard deviations from the mean makes use of the z formula.

The z score is the number of standard deviations a value is from the mean. By subtracting the value from the mean and dividing by the standard deviation, we calculate the number of standard deviations. The formula is

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

(4.14)

This is the basic formula upon which many others will be built.
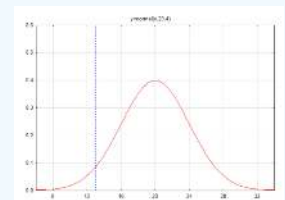

Example 5

Suppose the mean number of successes in a sample of 100 is 20 and the standard deviation is 4. Sketch and label a normal curve and find the area in the left tail for the number 13.

First find the z score: $z = \dfrac{x - \mu}{\sigma}$

$z = \dfrac{13 - 20}{4} = -1.75$

Find the area to the left in the table

$A_L = 0.0401$

Example 6

If the mean is 30 and the standard deviation is 5, then sketch and label a normal curve and find the area in the right tail for the number 44.1.

First find the z score: $z = \dfrac{x - \mu}{\sigma}$

$z = \dfrac{44.1 - 30}{5} = 2.82$

Find the area to the left in the table

$A_L = 0.9976$

Use this to find the area to the right by subtracting from 1.

$A_R = 0.0024$
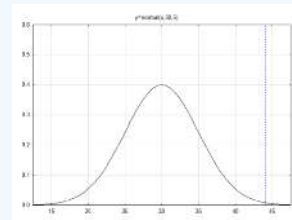
Return to Step 6: Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

Remember that the hypotheses for the autonomous car problem are: $H_0 : p = 0.60$, $H_1 : p > 0.60$. In the original problem, the researcher found that 4 out of 6 people thought the owner was responsible. Which hypothesis does this data support if the level of significance is 0.10?

This hypothesis test will be done using a method called the Normal Approximation to the Binomial Distribution.

The first step is to find the mean and standard deviation of the binomial distribution (which was done earlier but is now repeated):

$\mu = np = 6(0.6) = 3.6$

$\sigma = \sqrt{npq} = \sqrt{6(0.6)(0.4)} = 1.2$

Draw and label a normal curve with a mean of 3.6 and a standard deviation of 1.2.
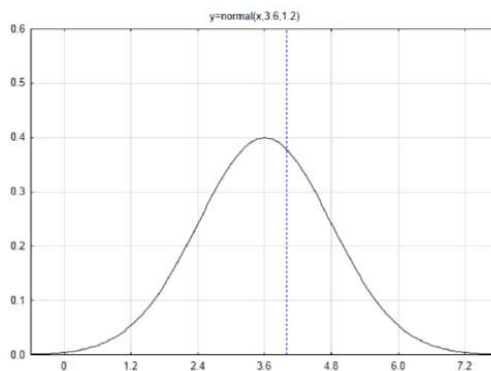
Find the z score if the data is 4.

$z = \dfrac{x - \mu}{\sigma}$ $z = \dfrac{4 - 3.6}{1.2} = 0.33$

From the table, the area to the left is $A_L = 0.6255$. Since the direction of the extreme is to the right, subtract the area to the left from 1 to get $A_R = 0.3745$. This is the p-value.

> This p-value can also be found with the calculator (2nd Distr #2: normalcdf(low, high, μ, σ)) shown as normalcdf(4, 1E99, 3.6,1.2)=0.3694.

Since this value is greater than the level of significance, if the calculator generated p-value is used, the conclusion will be written as: At the 0.10 level of significance, the proportion of people who think the owner is responsible is not significantly more than 0.60 (z = 0.33, p = 0.3694, n = 6).

Let us now take a moment to compare the p-value from the Normal Approximation to the Binomial Distribution (0.3694) to the exact p-value found using the Binomial Distribution (0.5443). While these p- values are not very close to each other, the conclusion that is drawn is the same. The reason they are not very close is because a sample size of 6 is very small and the normal approximation is not very good with a small sample size.

Test the hypothesis again if the researcher finds that 400 out of 600 of the people believe the owner is responsible for accidents.

$\mu = np = 600(0.6) = 360$ This indicates that if lots of samples of 600 people were sampled the average number of people who think the owner is responsible would be 360.

$\sigma = \sqrt{npq} = \sqrt{600(0.6)(0.4)} = 12$

Draw a label a curve with a mean of 360 and a standard deviation of 12.
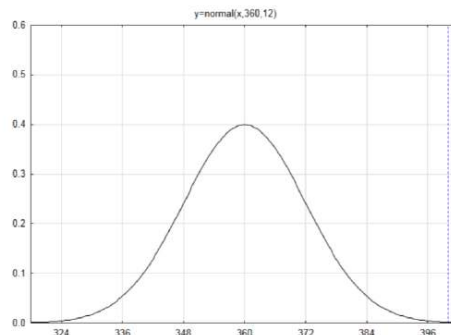
Find the z score if the data is 400.

$$z = \frac{x - \mu}{\sigma} \quad z = \frac{400 - 360}{12} = 3.33$$

Using the table, the area to the left is $A_L = 0.9996$. Since the direction of the extreme is to the right, subtract the area to the left from 1 to get $A_R = 0.0004$. More precisely, it is 0.000430.

This time when the results of the Normal Approximation to the Binomial Distribution (0.000430) are compared to the results of the binomial distribution (0.000443), they are very close. This is because the sample size is larger.

In general, if $np \geq 5$ and $nq \geq 5$, then the normal approximation makes a good, but not perfect estimate for the binomial distribution. When a sample of size 6 was used, $np = 3.6$ which is less than 5. Also, $nq = 6(0.4) = 2.4$, which is less than 5, too. Therefore, using the normal approximation for samples that small is not a good strategy.

Step 7 – Find the approximate p-value using the Sampling Distribution of Sample Proportions

Up to this point the discussion has been about the number of people. When sampling that produces categorical data is done, these numbers or counts can also be represented as proportions by dividing the number of successes by the sample size. Thus, instead of the researcher saying that 4 out of 6 people believe the owner is responsible, the researcher could say that 66.7% of the people believe the owner is responsible. This leads to the concept of looking at proportions rather than counts which means that instead of the distribution being made up of the number of successes, represented by x, it is made up of the sample proportion of successes represented by $\hat{p}$.
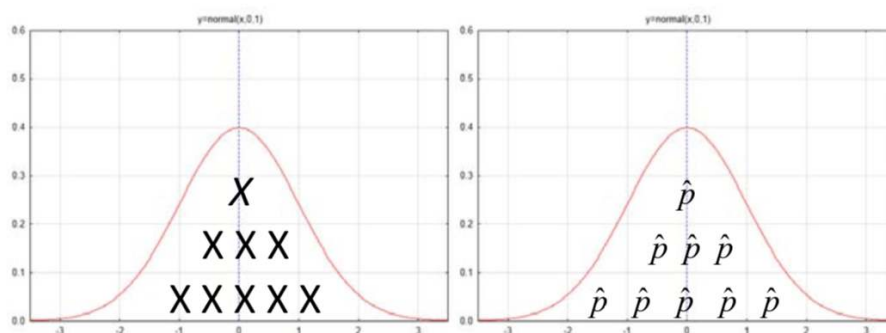
Digression 7 – Sampling Distribution of Sample Proportions

Since the binomial distribution contains all possible counts of the number of successes and it is approximately normally distributed and since all counts can be converted to proportions by dividing by the sample size, then the distribution of $\hat{p}$ is also approximately normally distributed. This distribution has a mean and standard deviation that can be found by dividing the mean and standard deviation of the binomial distribution by the sample size n.

The mean of all the sample proportions is the mean number of successes divided by n.

$\mu_{\hat{p}} = \dfrac{\mu}{n} = \dfrac{np}{n} = p$ This indicates that the mean of all possible sample proportions equals the true proportion for the population.

$$\mu_{\hat{p}} = p \qquad (4.15)$$

The standard deviation of all the sample proportions is the standard deviation of the number of successes divided by n.

$\sigma_{\hat{p}} = \dfrac{\sigma}{n} = \sqrt{\dfrac{npq}{n^2}} = \sqrt{\dfrac{pq}{n}}$ or $\sqrt{\dfrac{p(1-p)}{n}}$

$$\sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}} \quad or \quad \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}} \qquad (4.16)$$

The basic z formula $z = \dfrac{x - \mu}{\sigma}$ can now be rewritten knowing that in a distribution of sample proportions, the results of the sample that have formerly been represented with X can now be represented with $\hat{p}$. The mean, μ can now be represented with p, since $\mu_{\hat{p}} = p$ and the standard deviation σ can now be represented with $\sqrt{\dfrac{p(1-p)}{n}}$ since $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$. Therefore, for the sampling distribution of sample proportions, the z formula $z = \dfrac{x - \mu}{\sigma}$ becomes

$$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}. \qquad (4.17)$$

Apply this concept of probability to the hypothesis about the responsibility for a self-driving car in an accident.

Remember that the hypotheses for the people who think the owner is responsible are: $H_0 : p = 0.60$, $H_1 : p > 0.60$. In the original problem, the researcher found that 4 out of 6 people think the owner is responsible. Which hypothesis does this data support if the level of significance is 0.10?

Since $\mu_{\hat{p}} = p$ then the mean is 0.60 (from the null hypothesis).

Since $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.6(0.4)}{6}} = 0.2$ then the standard deviation is 0.2.

Draw a label a normal curve with a mean of 0.6 and a

standard deviation of 0.2.
If the data is 4, then the sample proportion,

$\hat{p} = \dfrac{x}{n} = \dfrac{4}{6} = 0.6667$

Find the z score if the data is 4.

$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ $z = \dfrac{0.6667 - 0.6}{0.2} = 0.33$
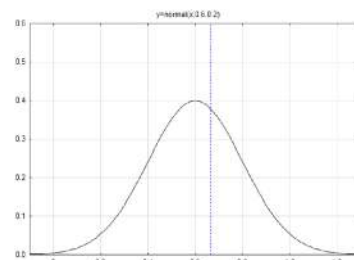
The area to the left is $A_L = 0.6304$. Since the direction of the extreme is to the right, subtract the area to the left from 1 to get $A_R = 0.3696$.

Compare this result to the result found when using the Normal Approximation to the Binomial Distribution. Notice that both results are exactly the same. This should happen every time, provided there isn't any rounding of numbers. The reason this has happened is because the number of successes can be represented as counts or proportions. The distributions are the same, although the x-axis is labeled differently. Divide the z scores for the normal approximation by the sample size 6 and you will get the z scores for the sampling distribution.

Test the hypothesis again if the researcher finds that 400 out of 600 of the people believe the owner is responsible.

Since $\mu_{\hat{p}} = p$ then the mean is 0.60 (from the null hypothesis).

Since $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.6(0.4)}{600}} = 0.02$ then the standard deviation is 0.02.

If the data is 400, then the sample proportion, $\hat{p} = \dfrac{x}{n} = \dfrac{400}{600} = 0.66667$

$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \quad z = \dfrac{0.66667 - 0.6}{0.02} = 3.33$

The area to the left is $A_L = 0.9996$. Since the direction of the extreme is to the right, subtract the area to the left from 1 to get $A_R = 0.0004$. More precisely, it is 0.000430.

Conclusion for testing hypotheses about categorical data.

By this time, many students are wondering why there are three methods and why the binomial distribution method isn't the only one that is used since it produces an exact p-value. One justification of using the last method is comparing the results of surveys or other data. Imagine if one news organization reported their results of a survey as 670 out of 1020 were in favor while another organization reported they found 630 out of 980 were in favor. A comparison between these would be difficult without converting them to proportions, therefore, the third method, which uses proportions, is the method of choice. When the sample size is sufficiently large, there is not much difference between the methods. For smaller samples, it may be more appropriate to use the binomial distribution.

Making inferences using quantitative Data

The strategy for making inferences with quantitative data uses sampling distributions in the same way that they were used for making inferences about proportions. In that case, the normal distribution was used to model the distribution of sample proportions, $\hat{p}$. With quantitative data, we find the mean, therefore the normal distribution will be used to model the distribution of sample means, $\bar{x}$.

To demonstrate this, a small population will be used. This population consists of the 50 states of the United States plus the District of Columbia and the 5 US territories of American Samoa, Guam, Northern Mariana Islands, Puerto Rico and the Virgin Islands, each of which has one representative in Congress with limited voting authority. A histogram showing the distribution of the number of representatives in a state or territory is provided. On the graph is a normal distribution based on the mean of this population being 7.875 representatives and a standard deviation of 9.487. The distribution is positively skewed and cannot be modeled by the normal curve that is on the graph.
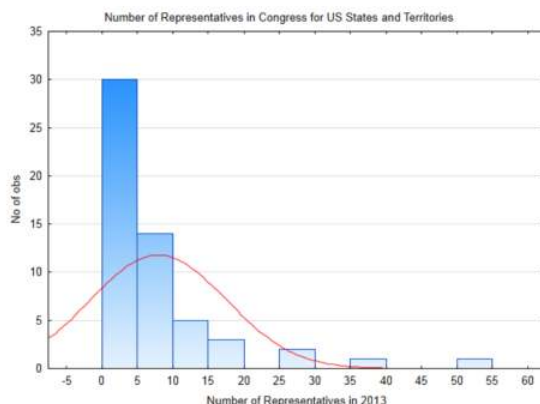
Be aware that in reality, the mean and standard deviation, which are parameters, are not known and so we would normally write hypotheses about them. However, for this demonstration, a small population with a known mean and standard deviation are necessary. With this, it is possible to illustrate what happens when repeated samples of the same size are drawn from this population, with replacement, and the means of each sample are found and becomes part of the distribution of sample means.

A sampling distribution of sample means (a distribution of x) contains all possible sample means that in theory could be obtained if a random selection process was used, with replacement. The number of possible sample means can be found using the fundamental rule of counting. Draw a line to represent each state/territory that would be selected. On the line write the number of options, so that it would look like this:

Options:	56	56	56	56	56
State:	1	2	3	4...	n

If our sample size is 40, then there are 56^{40} possible samples that could be selected which is equal to 8.46×10^{69}. That is a lot of possible samples. For this demonstration, only 10,000 samples of size 40 will be taken. The distribution of these sample means when this was done is shown in the histogram below.
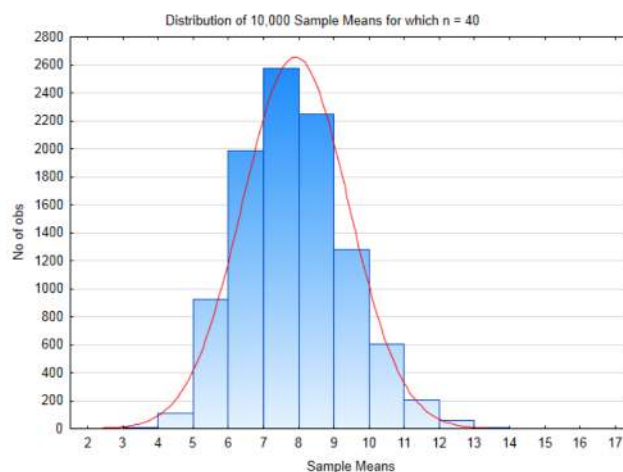
The mean of all these sample means is 7.8748 and the standard deviation is 1.502. Notice that the mean of all these sample means is almost exactly the same as the mean of the original population. Also notice that the standard deviation of all these sample means is much smaller than the standard deviation of the population. This is summarized in the table below.

	Population	Sampling Distribution
Mean	7.875	7.8748
Standard Deviation	9.487	1.502

The following graph has both the original data and the sample means on it. Notice how the two normal curves are centered at approximately the same place but the curve for the sample means is narrower. This shows that when samples of sufficient size are taken from any population, the means of those samples will be close to the means of the population.

Comparison of Population Data and Sample Means
Number of Reps = 56*5*normal(x, 7.875, 9.487)
Var9 = 10000*5*normal(x, 7.8748, 1.5021)
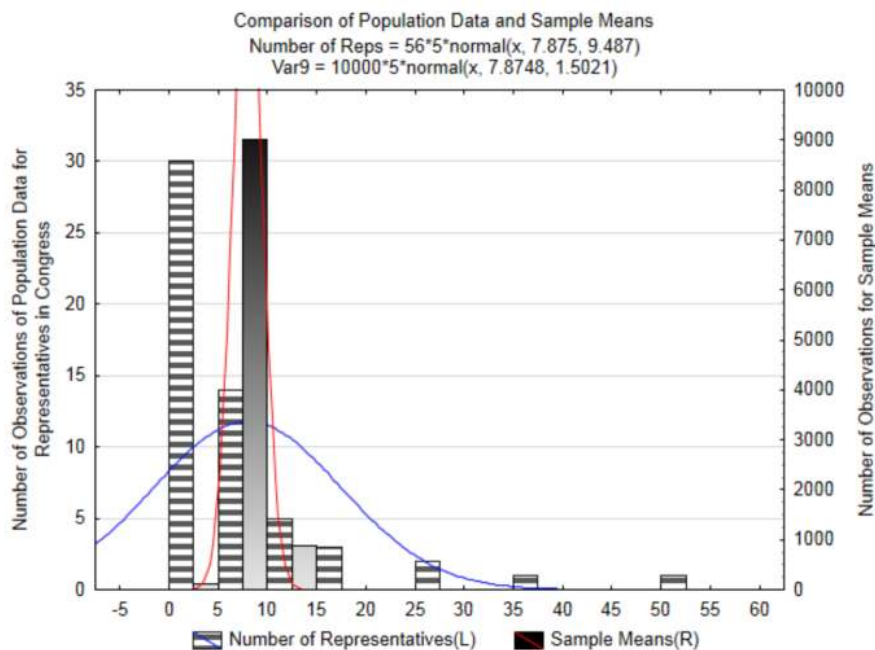
We are now ready to discuss the **Central Limit Theorem**. This theorem states that for any set of quantitative data with a mean μ and a standard deviation σ, the mean of all possible sample means will equal the mean of the population. The standard deviation of all the sample means, which is also called the standard error, will equal the standard deviation of the population divided by the square root of n. These are shown as:

$$\mu_{\bar{x}} = \mu \tag{4.18}$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{4.19}$$

It also says the distribution of sample means will be normal if the sample size is sufficiently large (generally considered to be 30 or more). If the original population is normally distributed, than the distribution of sample means will be normally distributed for any sample size.

Before doing an example, it will be important to see the effect of sample sizes. Compare the following 4 graphs that show the distribution of sample means for samples of size 40, 30, 20, and 10.
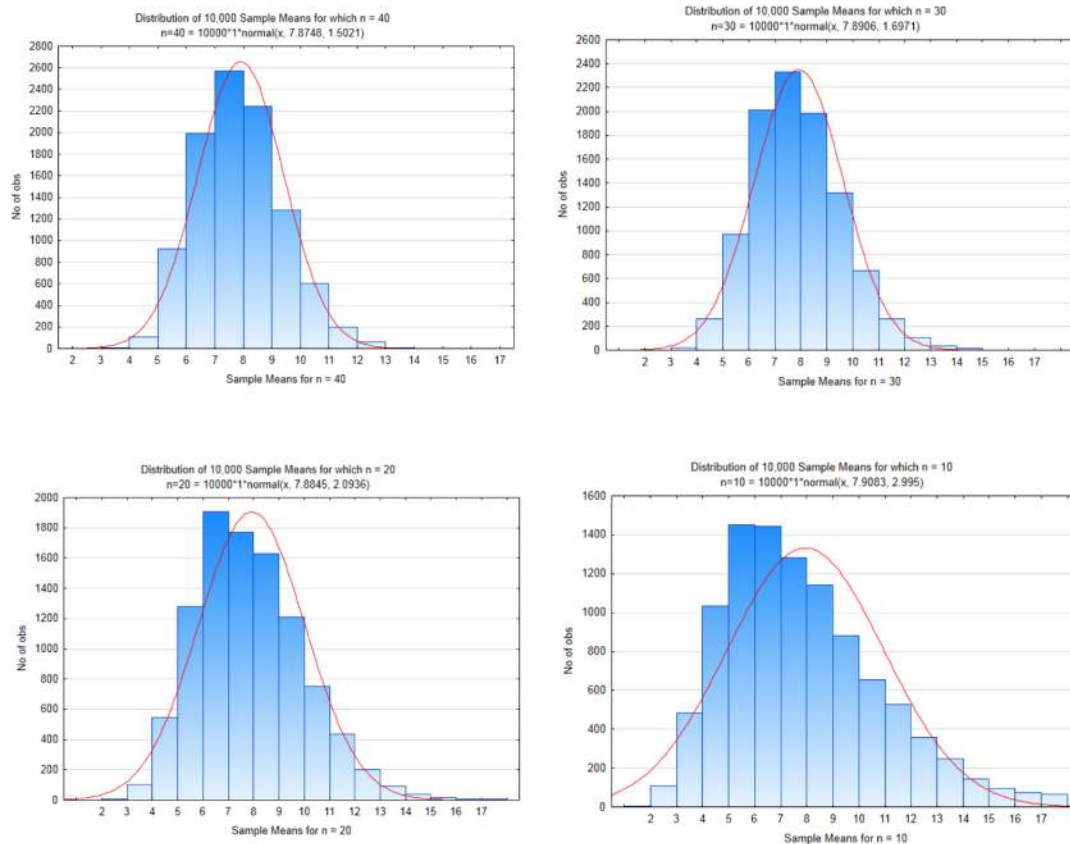
Notice how the distributions become more skewed as the sample size decreases. Notice also that the mean of the sample means are still approximately equal to the mean of the population but the standard deviations get larger as the sample size gets smaller. This implies there is more variation in sample means with small sample sizes than with large sample sizes.

	Population	$n = 40$	$n = 30$	$n = 20$	$n = 10$
Mean	7.875	7.8748	7.8906	7.8845	7.9083
Standard Deviation	9.487	1.5021	1.6971	2.0936	2.995
Calculated Standard Deviation using $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$		1.500	1.732	2.121	3.000

When making inferences about quantitative data, the basic z formula $z = \dfrac{x - \mu}{\sigma}$ can now be rewritten knowing that in a distribution of sample means, the results of the sample that have formerly been represented with x can now be represented with $\bar{x}$. The mean, μ will still be represented with μ, since $\mu_{\bar{x}} = \mu$ and the standard deviation σ can now be represented with $\dfrac{\sigma}{\sqrt{n}}$ since

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Therefore, for the sampling distribution of sample means, the z formula, $z = \dfrac{x - \mu}{\sigma}$ becomes

$$z = \frac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}. \tag{4.20}$$

It is now time to use the central limit theorem to test a hypothesis.

Example 7

Example 7 Mercury in fish is not healthy and restrictions are placed on the amount of fish that should be eaten. Suppose a researcher wanted to know if the average concentration of methylmercury per kilogram of fish tissue was greater than the maximum recommended limit of 300 μg/kg. If the average concentration is greater than 300, the fisheries will be closed, otherwise it will remain open. Suppose also that the standard deviation for the population is $\sigma = 50 \mu g/kg$. The researcher catches 36 fish. The sample mean concentration is 310.

The hypotheses to be tested are:

$H_0 : \mu = 300$

$H_1 : \mu > 300$

$\alpha = 0.1$

Since all the information that is needed is provided in the problem, the first step is to find the z score.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{310 - 300}{\frac{50}{\sqrt{36}}} = 1.2.$$

The next step is to look up 1.20 in the standard normal distribution tables. This gives an area to the left of $A_L = 0.8849$ and so the area to the right is $A_R = 0.1151$. This is a p-value.

Since the p-value is greater than the level of significance, the conclusion is that the average concentration of methylmercury in the fish tissue is not significantly greater than 300 μg/kg ($z = 1.20$, $p = 0.1151$, $n = 36$). Therefore, the fisheries will not be closed to fishing

Example 8

According to one estimate, the average wait time for subsidized housing for homeless people is 35 months. (www.stcloudstate.edu/reslife/...Statistics.pdf viewed 9/13/13) Assume the distribution of times is normal and the standard deviation is 10 months. One city evaluates its current program and to see if it is effective and justifies continued funding. If the average wait time is less than 35 months, the program will continue. Otherwise, the program will be replaced with a different program.

The hypotheses to be tested are:

- $H_0 : \mu = 35$
- $H - 1 : \mu < 35$

$\alpha = 0.01$

The wait time (in months) of twenty people who recently received subsidized housing is recorded below.

44	23	26	27	22	33	20	28	8	22
23	19	12	23	12	7	17	4	18	33

Since we are given the data, we must find the sample mean before finding the z score.

$\bar{x} = 21.05$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{21.05 - 35}{\frac{10}{\sqrt{20}}} = -6.24.$$

Because the direction of the extreme is to the left we find the area to the left on the standard normal distribution table. The lowest z score we find on that table is -3.49. The area to the left of -3.49 is 0.0002. Going even further to the left, the area will be less than that. Therefore, the p-value for this data is <0.0002. The amount of wait time before people receive subsidized housing with the current program is significantly less than 35 months (z = -6.24, p < 0.0002, n = 20). Based on the decision rule, the program is effective and will continue to be funded.

4.E: Inferential Theory (Exercises)

Chapter 4 Homework

For all parts of this problem, $H_0 : p = 0.70$, $H_1 : p < 0.70$. Show all supporting work including formulas, substitutions, and solutions as appropriate.

  a. What is the probability the first unit selected is a success?

  b. What is the probability the first unit selected is a failure?

  c. What is the probability the first five units selected will be in the order of FSSSF?

  d. If 5 values are selected, how many combinations are there for 3 successes?

  e. What is the probability three of five units will be a success?

  f. Create the entire binomial probability distribution if 5 units are selected. Record probabilities to 4 decimal places and then draw a stick graph in the provided space.

	0.40					
	0.35					
	0.30					
	0.25					
	0.20					
	0.15					
	0.10					
	0.05					
	0					
$X = x$	0	1	2	3	4	5
$P(X = x)$						
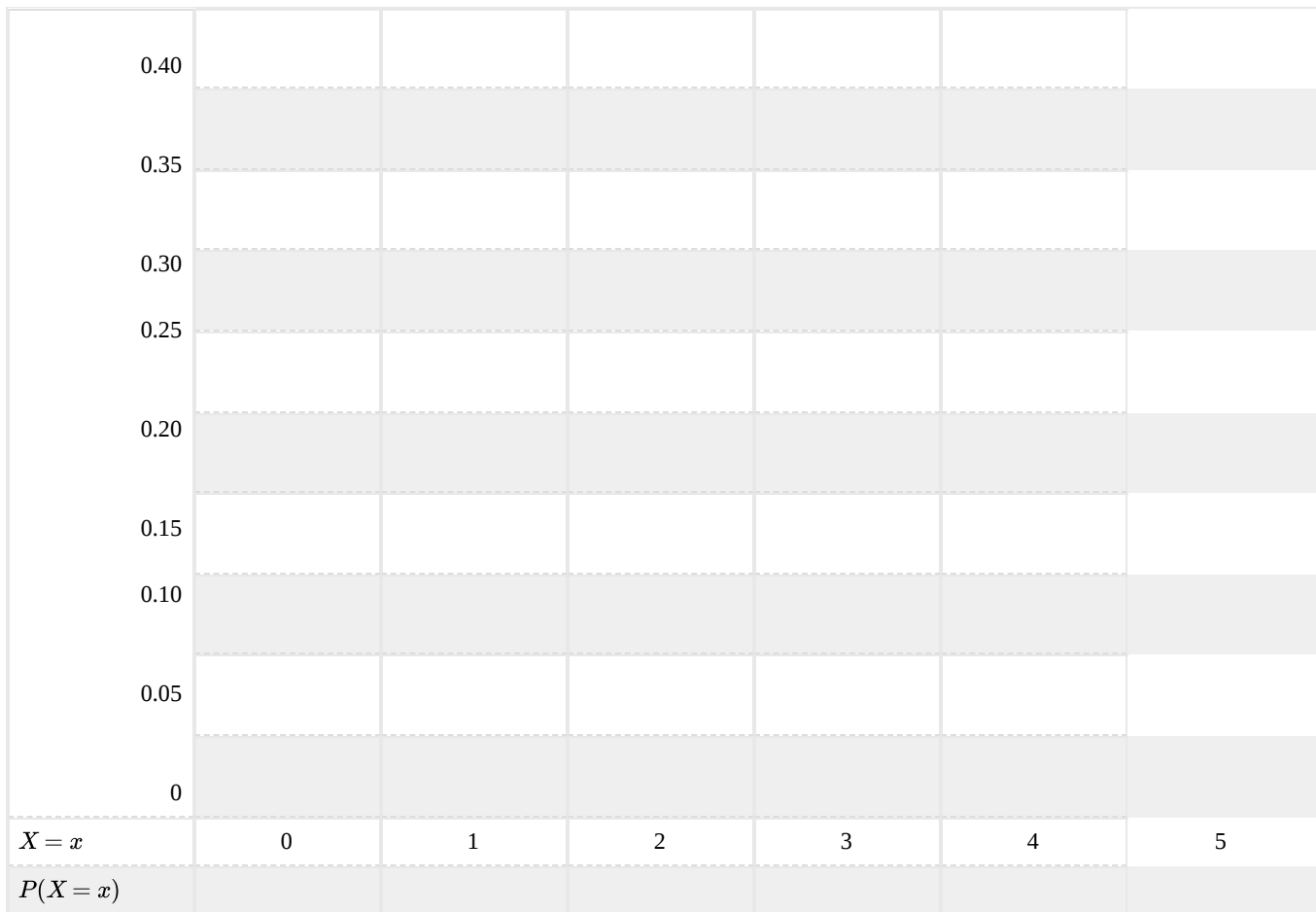
  g. What is the mean and standard deviation of this binomial distribution?

  h. Test the hypotheses if there were 3 successes in a sample of 5, what is the probability that three or fewer successes would be obtained if the null hypothesis is true? This is a p-value. At the 0.20 level of significance, which hypothesis is supported?

For all parts of this problem, $H_0 : p = 0.40$, $H_1 : p > 0.40$. Show all supporting work including formulas, substitutions, and solutions as appropriate.

a. What is the probability the first unit selected is a success?

b. What is the probability the first unit selected is a failure?

c. What is the probability the first five units selected will be in the order of SFSFSSS?

d. If 7 values are selected, how many combinations are there for 5 successes?

e. What is the probability three of five units will be a success?

f. Create the entire binomial probability distribution if 5 units are selected. Record probabilities to 4 decimal places and then draw a stick graph in the provided space.

0.40						
0.35						
0.30						
0.25						
0.20						
0.15						
0.10						
0.05						
0						
$X = x$	0	1	2	3	4	5
$P(X = x)$						
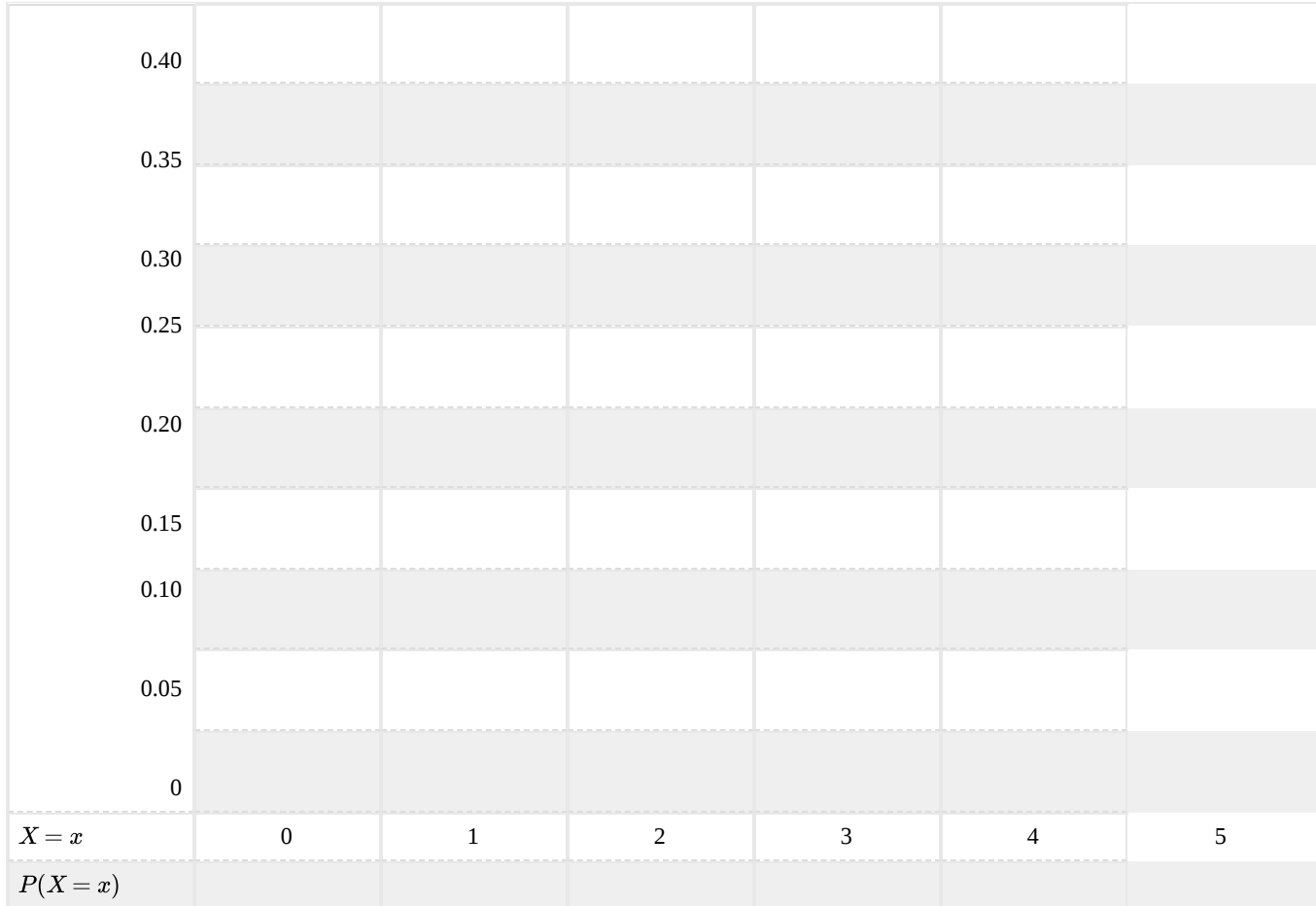
g. What is the mean and standard deviation of this binomial distribution?

h. Test the hypotheses if there were 5 successes in a sample of 7, what is the probability that three or fewer successes would be obtained if the null hypothesis is true? This is a p-value. At the 0.20 level of significance, which hypothesis is supported?

Briefing 5.1 Coal Export Terminals in the Pacific Northwest

Coal is used to produce electricity. It is also a major contributor of greenhouse gases and other pollutants to the atmosphere. A substantial amount of coal is mined in Montana and Wyoming. One goal is to export this coal to Asia. To do so means building coal terminals in Washington or Oregon. Some people want that to happen because it will bring money to the coal producers and jobs to those who work for the railroad or coal terminals. Others are opposed because of the impact to the community because of frequent long trains that will go through the towns, the pollution from the coal dust that is lost by the trains, the impact on the fishing industry from water pollution and the effect coal has on the climate.

1. Assume that in a hypothetical Pacific Northwest coast community that has been suggested as a potential coal terminal location, the mayor of the town has mixed thoughts about whether to support the project or oppose it. While it will bring more jobs to the community that needs them, the consequences are troubling. The mayor decides to have a survey conducted. 300 people will be surveyed. If a majority of the residents in the community oppose the coal terminal (success), the mayor will also oppose it; otherwise the mayor will support it. The hypotheses used to test for a majority are $H_0 : p = 0.50$ and $H_1 : p > 0.50$. The level of significance is 0.05.

a. What is the probability that the 30th person selected by the pollster opposes the coal terminal?

b. What is the probability that the 287th person selected by the pollster doesn't oppose the coal terminal?

c. What is the probability that the first ten people selected will be in this order, where S represents opposition to the terminal and F represents not being opposed: SFFFFFSFSS?

DATA: 165 out of 300 surveyed people oppose the terminal.

d. What is the probability that the pollster obtained any specific sequence of 165 successes and 135 failures?

e. How many combinations are there for 165 successes in a sample of 300? 3f. What is the probability of 165 success in a sample of 300?

g. What is the mean of the binomial distribution if n is 300?

h. What is the standard deviation of the binomial distribution if n is 300?

i. In a sample of 300, there could be between 0 and 300 successes. In this problem, you will only focus on 145 to 155 successes. Complete the partial distribution below and make a stick graph of this section of the distribution.

$X = x$	145	146	147	148	149	150	151	152	153	154	155
$P(X = x)$											

(y-axis labels: 0.046, 0.045, 0.044, 0.043, 0.042, 0.041, 0.040, 0.039)

j. Use the binomial distribution to determine which hypothesis is supported if 165 out of 300 people opposed the terminal? Show the calculator function that will be used, and your substitutions. Write a complete concluding sentence in the style of a scholarly journal that includes the p-value and sample size.

Calculator Input p-value

k. Use the normal approximation to the binomial distribution. Draw and label the normal curve. Find the z-score, and p-value then write a complete concluding sentence in the style of a scholarly journal that includes the z-score, p-value and sample size. Show the formula and substitution for the z score.

Formula Substitution z value p-value

l. What is the sample proportion of people opposed to the terminal?

m. What is the mean and standard deviation of the distribution of $\hat{p}$? Show formulas, substitutions and solutions.

3n. Use the sampling distribution of sample proportion method for testing the hypothesis. Draw and label a normal curve. Find the z-score, and p-value then write a complete concluding sentence in the style of a scholarly journal that includes the z-score, p-value and sample size. Show the formula and substitution for the z score.

Formula Substitution z value p-value

o. Based on the results of all of these hypothesis tests, will the mayor support or oppose the project?

2. In 2001, the Seattle Mariners won 116 games, which tied a record for the most number of games won by a baseball team in a season. During that year, the average attendance at home games in Safeco Field was 43,362. (http://www.baseball-almanac.com/teams/mariatte.shtml, viewed 9/13/13). Assume the standard deviation is 7,900 and that attendance is normally distributed. A sample of attendance at 10 games is taken from the 2013 season. Let α = 0.10.

10493	13000	30089	16294	13823
24701	18000	28198	15995	11656

Test the hypothesis that the average attendance in 2013 is less than it was in 2001.

a. What hypotheses would be used to test if the average attendance in 2013 is less than 43,362.

b. What is the mean of the distribution of sample means that is appropriate for testing this hypothesis?

c. What is the standard deviation of the distribution of sample means that is appropriate for testing this hypothesis?

d. Draw and label a normal curve for the sampling distribution.

e. What is the sample mean from 2013?

f. Test the hypothesis. Show all appropriate formulas, substitutions and solutions and write your complete concluding sentence.

Formula Substitution z value p-value

3. Ocean fishermen and boaters are familiar with tides and usually consult a tide table when planning a trip, but infrequent visitors to marine waters are less familiar with tides. As a first step in learning about tides, a curious person wants to determine if the time between consecutive high tides is greater than 12 hours? The hypotheses are $H_0 : \mu = 12$ and $H_1 : \mu > 12$. Assume the standard deviation is 1.4 hours. Let $\alpha = 0.05$.

A histogram of 36 times between consecutive high tides from September 2013 is shown below. (tides.mobilegeographics.com/c...onth/2152.html viewed 9/13/13 for Gig Harbor.)

Histogram of Time Between High Tides
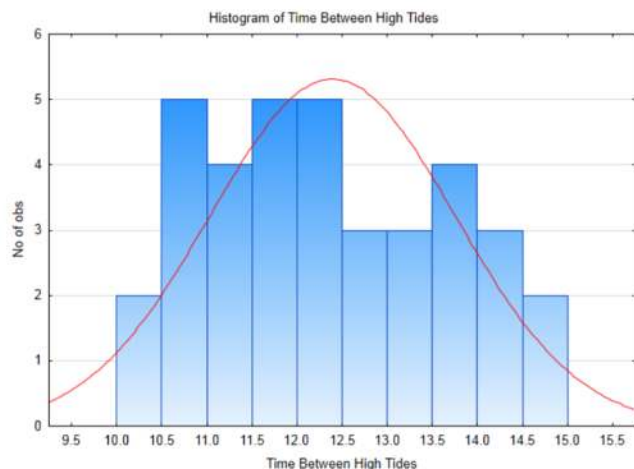
a. Assuming the null hypothesis is true, then what is the mean of the sampling distribution of sample means for 36 differences between consecutive high tides?

b. Assuming the null hypothesis is true, then what is the standard deviation of the sampling distribution of sample means for 36 differences between consecutive high tides?

c. Draw and label a normal curve for the distribution of $\bar{x}$, if n=36.

d. Test the hypothesis if the sample mean time between consecutive high tides is 12.38 hours. Show the formula, substitution and solution. Write a complete concluding sentence that includes the z score, p-value and sample size.

Formula Substitution z value p-value

4. According to the website walkscore.com, a walk score is a number between 0 and 100 that measures the walkability of any address. Scores over 90 indicate a Walker's Paradise while scores under 50 are car-dependent. Advantages of walkable neighborhoods include lower weight for residents, fewer carbon emissions and less car expenses. The objective of this experiment is to determine if smaller communities, defined for this problem as having less than 100,000 residents, have a higher average walk score than the largest cities. For the purposes of this problem, the average walk score of the largest 31 cities is 54.1. Assume the standard deviation for walk scores is 16.1. Let $\alpha = 0.05$.

a. Complete the design-layout table.

Research Design Table	
Research Question	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Porportion Correlation
List potential latent variables	
Grouping/explanatory Variables 1 (if present)	Levels:

Grouping/explanatory Variables 2 (if present)	Levels:

b. What are the hypotheses for this problem?

The walk scores of the 30 cities in the sample are provided below.

33	59	37	33	31	29
36	47	42	43	69	38
22	57	36	48	51	34
66	92	65	65	40	25
58	29	45	63	40	69

c. Make a frequency distribution and histogram for this data. Use reader-friendly class boundaries.

d. Find the sample mean and standard deviation

e. What is the mean and standard deviation of the sampling distribution of sample means that is based on the null hypothesis?

f. Draw and label the normal distribution for the sample means.

g. Test the hypothesis. Show formulas, substitutions and solutions. Write a complete conclusion including z score, p-value and sample size.

Formula Substitution z value p-value

h. Can we conclude smaller towns have a higher walk score than the largest cities?

5. Magazines about sports regularly contain predictions about who will win games or championships. One would expect that the writers who make the predictions would have considerable expertise and insight and have a high rate of success. At a minimum, one would hope that the writers are better than a coin flip for determining winners. A coin flip has a 50% chance of picking the winning team.

To test if the writers are better than a coin flip, two major sports magazines were selected and predictions from regular NFL games were compared with results. A success was if the prediction was correct, and a failure was if the prediction was wrong. Use a 5% level of significance.

b. Write the hypotheses.

The reporters picked the winning team 181 out of 322 times.
c. What is the sample proportion?

d. Make a completely labeled pie chart.

e. Test the hypothesis using the binomial distribution. Write a complete concluding sentence.

Calculator Input p-value

f. What is the mean and standard deviation of the binomial distribution for this problem?

g. Test the hypothesis using the normal approximation to the binomial distribution. Include a completely labeled drawing of the normal curve, the appropriate formulas, substitutions and solutions and then write a complete concluding sentence.

Formula Substitution z value p-value

h. Test the hypothesis using the sampling distribution method. Include a completely labeled drawing of the normal curve, the appropriate formulas, substitutions and solutions and then write a complete concluding sentence.

What is the sample proportion?

Formula Substitution z value p-value

i. Based on the statistical results, do these sports writers appear to be better than a coin flip? Are you impressed by their ability to predict NFL winners?

6. Developed in collaboration with Alan Kemp, Professor of Sociology and author of the book*Death, Dying and Bereavement in a Changing World*, published by Pearson, 2013.

This topic is discussed in SOC 212, Sociology of Death.

> **Briefing 5.2**
>
> Terror Management Theory (TMT) was developed by Ernest Becker in the 1960s and 1970s. Extensive experimentation has been done to test these theories. This problem is based on the article *Evidence for terror management theory: 1. The effects of mortality salience on reactions to those who violate or uphold cultural values*. By Rosenblatt, Greenberg, Solomon, Pyszczynski and Lyon. It was published in the Journal of Personality and Social Psychology, Vol 57(4), Oct, 1989 pp 681-690.
>
> A basic premise behind the theory is that humans are the only species who recognize their own mortality (they know they will die in the future). Consequently, humans need a way to manage the emotions related to this knowledge. The two predominate ways that humans cope are with culture (e.g. religion and other beliefs) and self-efficacy which means that we want to know that what we do matters within our cultural worldview. One such consequence of this is that following a terrorist attack or deadly natural disaster, patriotism increases (culture) as does heroism (self-efficacy).
>
> Cultures are an artificial construction and therefore the worldview they portray can be exposed to potential threats. Since a culture can provide the standards by which a person can feel that life is fair, any person or idea that threatens the cultural norms must be removed or punished. Consequently, an expected outcome of this theory is that people will respond positively toward those who support cultural values and negatively toward those who violate these values. To test the theory, the authors of the article designed an experiment to determine if a reminder of one's own mortality would lead to more negative responses for something that violates cultural values.

Municipal court judges were selected for this experiment. The purpose of the experiment was disguised. The judges were given a questionnaire. Within the questionnaire of half the judges were questions about their thoughts and feelings about the prospect of their own death. The remaining judges did not have these questions. Judges were randomly assigned the questionnaire. Following questions, the judges were given a scenario about a case of alleged prostitution and asked to set the amount of bail for the prostitute. Prostitution was used because it emphasized the moral nature of the alleged crime. No effort was made to determine the judge's opinion about prostitution, which could affect their bail amount. Judges were selected for this experiment because they have been trained to make such punishments. The objective was to determine if the reminder about their own

mortality would lead to harsher penalties when someone violated the cultural norms. The average bail amount between the two groups will be compared.

The hypotheses that will be tested are meant to show that judges who have been reminded about their own mortality (impact) will set higher bail amounts than judges who have not been reminded (control). $H_0 : \mu_{\text{impact}} = \mu_{\text{control}}$, $H_1 : \mu_{\text{impact}} > \mu_{\text{control}}$, $\alpha = 0.05$

The following data is not authentic, but it closely approximates the results obtained by the researchers. The impact group of judges was reminded of their own mortality. The control group was not.

Amount of Bail										
impact	50	50	150	200	1500	1500	1200	205	50	50	50
Control	25	25	25	50	150	50	50	25	25	25	25

a. Make an appropriate graph so that the two groups can be compared. You need to decide what is appropriate.

b. Complete the table below.

	Impact	Control
Mean		
Standard Deviation		
Median		

c. The p-value for the comparison of the mean bail amount is 0.041. The sample size is 22. Write a complete concluding sentence to show if there is a significant difference between the bail amounts set by judges reminded of their own mortality and judges who were not reminded.

d. Do the results of this experiment support the contention that contemplation of one's own death leads to increased punishment of those who violate cultural norms?

5: Testing Hypotheses

Since the beginning of the text it has been emphasized that a primary reason for doing statistics is to make a decision. Better decisions can be made if they are based on the best available evidence. While the ideal situation would be to get data from the entire population, the reality is that data will almost always come from a sample. Because sample data varies based on the random process that was used to select it, the researcher is forced to use sample data to draw a conclusion about the entire population. This is inference. It is using specific partial evidence to make a more general conclusion.

In Chapter 5, formulas were developed for testing hypotheses about proportions and means. In the former case the formula was

$$z = \frac{\hat{p} - p}{\frac{p(1-p)}{n}} \tag{5.1}$$

and in the latter case it was

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}. \tag{5.2}$$

In general, these formulas generate a test statistic, z, which is used to determine the number of standard errors a statistic is from a parameter. The normal distribution is then used to determine the probability of getting that statistic, or a more extreme statistic. That probability is called a p-value.

Every number that is needed to make use of the formula

$$z = \frac{\hat{p} - p}{\frac{p(1-p)}{n}} \tag{5.3}$$

can be found in the null hypothesis (p) or from the data ($\hat{p}$, n). The same cannot be said for the formula

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}. \tag{5.4}$$

While the value for μ comes from the null hypothesis and the value of $\bar{x}$ and n come from the sample data, there is no way to obtain the value of σ without doing a census. In the last chapter you were always told the value of σ, but this does not happen in the real world because to find σ requires first finding μ and if you knew μ, there would be no reason to test a hypothesis about it.

The resolution of this problem requires two changes to the process that was used in the previous chapter. The first change is that we will have to estimate σ. The best estimate is s, the standard deviation of the sample. Replacing σ with s means we can no longer use the standard normal distribution (z distribution). The second change therefore is to find a more appropriate distribution that can be used to model the distribution of sample means.

A set of distributions called the *t* distributions is used when the standard error of the mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is replaced with the estimated standard error of the mean $s_{\bar{x}} = \frac{s}{\sqrt{n}}$. The z formula for means,

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{5.5}$$

is then modified to become the t formula

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}. \tag{5.6}$$

Notice the only difference is the use of s instead of σ. The *t* distributions are used because they provide a better approximation of the distribution of sample means when the population standard deviation must be estimated using the sample standard deviation.

Unlike the normal distribution, there are many t distributions with each being defined by the number of degrees of freedom. Degrees of Freedom are a new concept that requires a little explanation.

The concept of degrees of freedom has to do with the number of independent values that can identify a position. This may be easier to think about if you picture a Cartesian coordinate system. With any two independently chosen values, normally called x and y, a point's position can be located somewhere on the graph. Consequently, the point that is picked has two degrees of freedom. However, if a constraint is placed on the points, such as x + y = 3, then only one of the values can be independent and the other value will depend on the independent value. Because of the constraint, one degree of freedom has been lost so now the point has only one degree of freedom. If a second constraint is placed on the system, such as x − y = 1, then another degree of freedom is lost. Degrees of freedom are lost every time a constraint is applied.

For sample data, each value represents a new piece of evidence, provided the data are independent. Dependent data would artificially inflate the sample size without providing any more information. Since a larger sample size would produce a smaller standard error, which would lead to a larger t value and therefore increase the chance of a statistically significant conclusion, then it is important to only count the number of independent data values, which are known as degrees of freedom. One degree of freedom is lost every time a parameter is replaced by a statistic. Therefore, when the standard error $\sigma_{barx} = \dfrac{\sigma}{\sqrt{n}}$ becomes the estimated standard error $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$, one degree of freedom has been lost. In this case, df = n − 1 where df is an abbreviation for degrees of freedom.

The formula for generating the test statistic, t, that is used to determine the number of standard errors a sample mean is from a hypothesized mean is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{5.7}$$

It has n-1 degrees of freedom.

In the same way that z = 1 represents 1 standard deviation above the mean for a normal distribution, $t = 1$ represents 1 standard deviation above the mean in a t distribution. Once the value of t has been determined, the p-value can be found by looking in a t table.

Student t distributions

One Tail Probability	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
Two Tail Probability	0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.001
Confidence Level	20%	50%	80%	90%	95%	98%	99%	99.9%
df								
1	0.325	1.000	3.078	6.314	12.706	31.821	63.656	636.578
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	31.600
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.437

12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.257	0.689	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.256	0.686	1.321	1.717	2.074	2.608	2.819	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.745
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.689
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.660
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.373
z^*	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.290

An assumption when using t distributions with a small sample size is that the sample is drawn from a normally distributed population. While some researchers believe this test statistic is robust enough to tolerate some violation of this assumption, at a minimum, a histogram of the data should be viewed to see if the assumption appears realistic. If it does not, other methods of analysis not discussed in this text must be pursued.

The way this t table is used to determine a p-value is to first find the row with the appropriate number of degrees of freedom. In that row, locate the range that would contain the test statistic. Move up to the first row if you are doing a one-tail test or the second row if it is a two-tail test. Next, identify the location of alpha. If your p-value is greater than alpha then use an inequality symbol to show that. If your p-value is less than alpha then show that with an inequality symbol. If greater detail can be provided, it should be. Since the t distributions are symmetric, negative t values can be found in this table by ignoring the negative signs and assuming the areas in the first row are to the left. Following are 2 examples. The sign in the alternative hypothesis, the level of significance, degrees of freedom, and the t value is provided in each example.

1. $H_1 :> \alpha = 0.05$ df = 6, t = 2.3

One Tail Probability	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
df								
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.959

t=2.3

For 6 degrees of freedom, 2.3 falls between 1.943 and 2.447, which means it has an area in the tail that is between 0.05 and 0.025. The p-value would be reported as p < 0.05.

2. $H_1 :\neq \alpha = 0.01$ df = 18, t = -1.26

Two Tail Probability	0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.001
df								
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.883

t=-1.26

For 18 degrees of freedom, -1.26 falls between 0.688 and 1.328 if the negative sign is ignored, so the area in two tails falls between 0.5 and 0.2. Since any value in this range would not be significant at the 0.01 level, then the p-value is greater than 0.01. However, greater detail can be provided by indicating the p-value is greater than 0.2. It would be incorrect to say the p-value is less than 0.5 because that does not tell us whether it is greater or less than 0.01.

There are two different inferential approaches that can be taken. Throughout most of this text the focus has been on the concept of testing hypotheses. That means there is actually a hypothesis of what would be found from a census. The alternative inferential approach occurs when there is not a hypothesis. In such cases the goal is to estimate the parameter rather than determine if the hypothesis about it is correct. Because the entire focus of the book has been on testing hypotheses, we will begin there and then address the idea of estimating the parameter in the next chapter. There are a considerable number of hypothesis test situations and formula, but we will focus on only four of them in this chapter and then add a few more in later chapters. The explanation will be provided with a discussion of exercise.

Briefing 5.1 Exercise

The US government recommends that people get 2.5 hours of moderately-intense aerobic exercise each week or 1.25 hours of vigorous-intense exercise each week along with some strength training such as weights or push-ups. Exercise helps reduce the risk of diabetes, heart disease, some types of cancer and improves mental health. (www.cbsnews.com/8301-204_162-...nded-exercise/)

The four hypothesis-test formulas that will be shown in this chapter will be illustrated with these five questions. As you read the questions, try to determine any similarities or differences between them, as that will ultimately guide you into which formula should be used.

1. Is the proportion of people who exercise enough to meet the government's recommendation less than 0.25?
2. Is the proportion of people with a health problem such as diabetes, heart disease or cancer lower for those who meet the government's exercise recommendation than it is for those who don't?
3. Is the average amount of exercise a college student does in a week greater than 2.5 hours?
4. Is the average weight of a person less after a month of new regular aerobic fitness program?
5. For those who exercise regularly, is the average amount of exercise a college graduate does in a week different than someone who does not graduate from college?

There are two different things to look for when determining similarities and differences. The first is whether the parameter that is mentioned is a mean or proportion. The second is the number of populations. The following table restates the questions, provides the

parameter of interest, the number of populations and an example of the hypotheses.

Question	Parameter	Populations	Hypotheses
Is the proportion of people who exercise enough to meet the government's recommendation less than 0.25?	proportion	1	$H_0 : P = 0.25$ $H_1 : P < 0.25$
Is the proportion of people with a health problem such as diabetes, heart disease or cancer lower for those who meet the government's exercise recommendation than it is for those who don't?	proportion	2	$H_0 : P_{\text{exercise}} = P_{\text{don't}}$ $H_1 : P_{\text{exercise}} < P_{\text{don't}}$
Is the average amount of exercise a college student does in a week greater than 2.5 hours?	mean	1	$H_0 : \mu = 2.5$ $H_1 : \mu > 2.5$
Is the average weight of a person less after a month of new regular aerobic fitness program?	mean	1	$H_0 : \mu = 0$ $H_1 : \mu < 0$
For those who exercise regularly, is the average amount of exercise a college graduate does in a week different than someone who does not graduate from college?	mean	2	$H_0 : \mu_{\text{college grad}} = \mu_{\text{not college grad}}$ $H_1 : \mu_{\text{college grad}} \neq \mu_{\text{not college grad}}$

Categorical data will be needed for questions about a proportion; quantitative data will be needed for questions about a mean. A brief explanation is needed for the fourth question. To determine the amount of change in a person after starting a fitness program, it is necessary to collect two sets of data. The person will need to be weighed prior to the fitness program and then again after one month. These data are dependent, which means they have to apply to the same person. Ultimately, the data that will be analyzed is the difference between a person's before and after weight. Therefore two data values are compressed into one value by subtraction. If the after-minus-before difference in weight is 0, then there has been no change. If it is less than 0, weight has been lost.

Since the evidence to help decide which hypothesis is supported by the data will come from a sample, and that sample is just one of the many possible sample results that form a normally distributed sampling distribution, then we can use what is known about the sampling distribution to determine the probability that we would have selected the data we got, or more extreme data (p-value).

In spite of the theoretical nature of a sampling distribution, it is the source for determining probabilities. Therefore, we will first define what the distributions contain and the important formulas related to this distribution.

The first time we encountered the normal distribution was when it was used as an approximation for the binomial distribution. In this case the data consisted of counts.
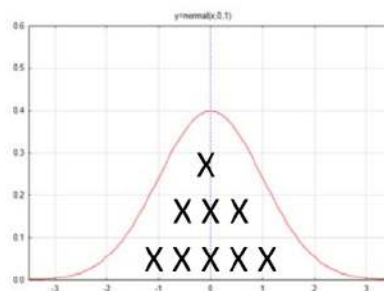
The mean of this distribution is found from $\mu = np$. The standard deviation is $\sigma = \sqrt{npq}$. The formula for determining the number of standard deviations a value is from the mean is $z = \dfrac{x - \mu}{\sigma}$.

Counts were eventually turned into proportions by dividing the counts by the sample size. The distribution consisted of all the possible sample proportions.
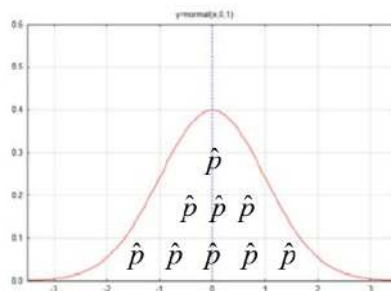
The distribution of sample proportions has a mean of $\mu_{\hat{p}} = p$ and a standard deviation of $\sigma_{\hat{p}} = \dfrac{p(1-p)}{n}$. The formula for determining the number of standard deviations a sample proportion is from the mean is $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$.

The next time we encountered the normal distribution was when we had quantitative data in which case the distribution was made up of sample means.
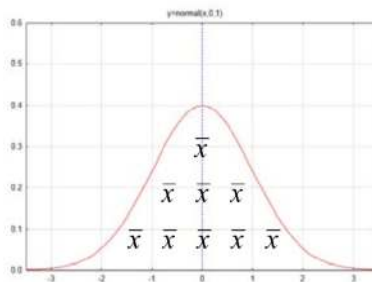
The mean of all possible sample means is $\mu_{\bar{x}} = \mu$ and the standard error is $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$. The formula for determining the number of standard deviations a sample mean is from the hypothesized population mean is $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$. Because σ is not known, it is estimated with s, so that the estimated standard error is $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$ and the Z formula is replaced by the t formula where $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$.

The distributions and formulas that were just shown are the same as, or similar to, the ones that you saw in Chapter 5 and that are appropriate for questions 1, 3, and 4. On the other hand, questions 2 and 5 have hypotheses unlike those encountered before and so some effort is needed to define the relevant distributions and their means and standard deviations. These will be based on three statistical results that will not be proven here:

1. The mean of the difference of two random variables is the difference of the means.

2. The variance of the difference of two independent random variables is the sum of the variances.

3. The difference of two independent normally distributed random variables is also normally distributed. (Aliaga, Martha, and Brenda Gunderson. *Interactive Statistics*. Upper Saddle River, NJ: Pearson Prentice Hall, 2006. Print.)

We will start with the question of whether the proportion of people with a health problem such as diabetes, heart disease or cancer is lower for those who meet the government's exercise recommendation than it is for those who don't. This means that there are two populations, the population that exercises above government recommended levels and the population that doesn't. Within each population, the proportion of people with a health problem will be found. A hypothesis test will be used to determine if people who exercise at the recommended levels have fewer health problems than people who don't. The hypotheses are:

$H_0 : P_{\text{exercise}} = P_{\text{don't}}$
$H_1 : P_{\text{exercise}} < P_{\text{don't}}$

Writing hypotheses in this manner is easy to interpret, but an algebraic manipulation of these will give us some insight into the distribution that would be used to represent the null hypothesis. $P_{\text{don't}}$ will be subtracted from both sides.

$H_0 : P_{\text{exercise}} - P_{\text{don't}} = 0$
$H_1 : P_{\text{exercise}} - P_{\text{don't}} < 0$

Since neither P_{exercise} or $P_{\text{don't}}$ is known because these are parameters, the best that can be done is estimate them using sample proportions. Therefore $\hat{p}_{exercise}$ will be used as an estimate of P_{exercise} and $\hat{p}_{don't}$ will be used as an estimate of $P_{\text{don't}}$. Then $\hat{p}_{exercise} - \hat{p}_{don't}$ as an estimate for $P_{\text{exercise}} - P_{\text{don't}}$.

The distribution of interest to us is the one consisting of the difference between sample proportions, generically shown as $\hat{p}_A - \hat{p}_B$.
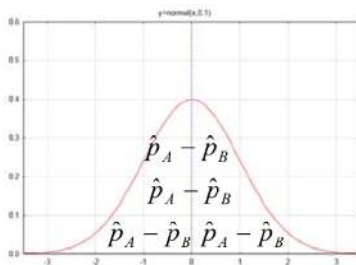
The mean of this distribution is $p_A - p_B$ and the standard deviation is $\sqrt{\dfrac{p_A(1-p_A)}{n_A} + \dfrac{p_B(1-p_B)}{n_B}}$. Since the only thing that is known about p_A and p_B is that they are equal, it is necessary to estimate their value so that the standard deviation can actually be computed. To do this, the sample proportions will be combined. The combined proportion is defined as

$$\hat{p}_c = \frac{x_A + x_B}{n_A + n_B}. \tag{5.8}$$

Replacing p_A and p_B with $\hat{p}_c$ results in the formula for estimated standard error of

$$\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_A} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_B}} \text{ or } \sqrt{\hat{p}_c(1-\hat{p}_c)(\frac{1}{n_A} + \frac{1}{n_B})} \tag{5.9}$$

We can now substitute into the z formula, $z = \dfrac{x - \mu}{\sigma}$ to get the test statistic used when testing the difference between two population proportions,

$$z = \frac{(\hat{p}_A - \hat{p}_B) - (p_A - p_B)}{\sqrt{\hat{p}_c(1-\hat{p}_c)(\dfrac{1}{n_A} + \dfrac{1}{n_B})}} \tag{5.10}$$

This can be written a little more simply in cases when the null hypothesis is $P_A = P_B$ which means that $p_A - p_B = 0$, so that term can be eliminated to give the test statistic

$$z = \frac{(\hat{p}_A - \hat{p}_B)}{\sqrt{\hat{p}_c(1-\hat{p}_c)(\frac{1}{n_A} + \frac{1}{n_B})}} \tag{5.11}$$

For this test statistic, both sample sizes should be sufficient large (n>20) with a minimum of 5 successes and 5 failures.

A similar approach will be taken with question 4, which asks if the average amount of exercise a college graduate does in a week is different than someone who does not graduate from college? There are two populations being compared, the population of college graduates and the population of non- college graduates. The average amount of exercise in each of these populations will be compared.

When the means of two populations are compared, the hypotheses are written as:

$H_0 : \mu_{\text{college grad}} = \mu_{\text{not college grad}}$
$H_1 : \mu_{\text{college grad}} \neq \mu_{\text{not college grad}}$

Writing hypotheses in this manner is easy to interpret, but an algebraic manipulation of these will give us some insight into the distribution that would be used to represent the null hypothesis.

$\mu_{\text{not college grad}}$ will be subtracted from both sides.

$H_0 : \mu_{\text{college grad}} - \mu_{\text{not college grad}} = 0$
$H_1 : \mu_{\text{college grad}} - \mu_{\text{not college grad}} \neq 0$

Since n either $\mu_{\text{college grad}}$ or $\mu_{\text{not college grad}}$ are known because these are parameters, the best that can be done is to estimate them using sample means. Therefore $\bar{x}_{college\,grad}$ will be used as an estimate of $\mu_{\text{college grad}}$ and $\bar{x}_{college\,grad}$ will be used as an estimate of $\mu_{\text{not college grad}}$. Then $\bar{x}_{college\,grad} - \bar{x}_{not\,college\,grad}$

The distribution of interest to us is the one consisting of the difference between sample means, generically shown as $\bar{x}_A - \bar{x}_B$.
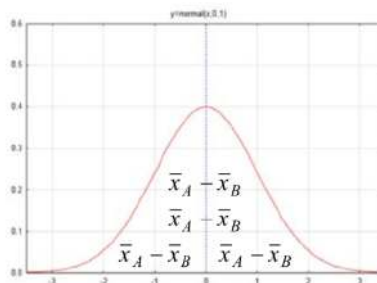
The mean of this distribution is $\mu_A - \mu_B$ and the standard deviation is $\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$. Once again we run into the problem that the standard deviation of the populations σ_A and σ_B are not known, so they must be estimated with the sample standard deviation sA and sB. An additional problem is that it is not known if the variances for the two populations are equal (homogeneous). Unequal variances (heterogeneous) increase the Type I error rate. (Sheskin, David J. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman & Hall/CRC, 2000. Print.)

The t Test for Two Independent Samples is used to test the hypothesis. This test is dependent upon the following assumptions.

1. Each sample is randomly selected from the population it represents.
2. The distribution of data in the population from which the sample was drawn is normal
3. The variances of the two populations are equal. This is the homogeneity of variance assumption. (Sheskin, David J. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman & Hall/CRC, 2000. Print.)

The test statistic follows the same basic pattern as the other tests, which involves finding the number of standard errors a statistic is away from the hypothesized parameter.

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5.12}$$

The assumption with this formula is that the two sample sizes are equal. If this formula is used when the sample sizes are not equal, there is an increased chance of making a Type I error. In such cases, an alternative formula is used which includes the weighted average of the estimated population variances of the two groups. The weighted average is based on the number of degrees of freedom in each sample. This formula can be used for both equal and non-equal sample sizes.

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{[\dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}][\dfrac{1}{n_A} + \dfrac{1}{n_B}]}} \tag{5.13}$$

Because two parameters (σ_A and σ_B) are replaced by s_A and s_B, two degrees of freedom are lost. Thus, the number of degrees of freedom for this test statistic is $n_1 + n_2 - 2$.

There are four different hypothesis tests presented in this chapter. The hypotheses and test statistics are summarized in the following table.

	Proportions (for categorical data)	Means (for quantitative data)
1 - sample	$H_0 : p = p_0$ $H_1 : p < p_0$ or $p > p_0$ or $p \neq p_0$ $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ Assumptions: $np \geq 5, n(1-p) \geq 5$	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$ or $\mu > \mu_0$ or $\mu \neq \mu_0$ $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ df = n - 1 Assumptions: If $n < 30$, population is approximately normally distributed.
2 - samples	$H_0 : p_A = p_B$ $H_1 : p_A < p_B$ or $p_A > p_B$ or $p_A \neq p_B$ $z = \dfrac{(\hat{p}_A - \hat{p}_B) - (p_A - p_B)}{\sqrt{\hat{p}_c(1 - \hat{p}_c)(\dfrac{1}{n_A} + \dfrac{1}{n_B})}}$ where $\hat{p}_c = \dfrac{x_A + x_B}{n_A + n_B}$ Assumptions: $np \geq 5, n(1-p) \geq 5$ for both populations	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A < \mu_B$ or $\mu_A > \mu_B$ or $\mu_A \neq \mu_B$ $t = \dfrac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{[\dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}][\dfrac{1}{n_A} + \dfrac{1}{n_B}]}}$ df $= n_A + n_B - 2$ Assumptions: If $n < 30$, population is approximately normally distributed.

For each hypothesis-testing situation, you will have to decide which formula and which table to use. Notice that when the hypotheses are about proportions, the standard normal z distribution is used. When the hypotheses are about means, the t distributions are used.

We will now return to our original five questions. The statistics given in this problem are fictitious.

1. Is the proportion of people who exercise enough to meet the government's recommendation less than 0.25?

   Assume that a random sample of 800 adults was taken. Of these, 184 claimed they met the government's recommendation for exercise. Can we conclude that the proportion that meets this recommendation is less than 25%? Use a level of significance of 0.05.

The hypotheses are:

$H_0 : p = 0.25$
$H_1 : p < 0.25$

The sample proportion is $\bar{p} = \dfrac{x}{n} = \dfrac{184}{800} = 0.23$

The test statistic is $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$. With substitution $z = \dfrac{0.23 - 0.25}{\sqrt{\dfrac{0.25(1 - 0.25)}{800}}} = -1.31$

Check the standard normal distribution table to find the area to the left is 0.0951. This is the p-value because the direction of the extreme is to the left. Since the p-value is greater than the level of significance, the data are consistent with the null hypothesis. We conclude that at the 0.05 level of significance, the proportion of adults who meet government recommendations for exercise is not significantly less than 25% (z = -1.31, p = 0.0951, n = 800).

2. Is the proportion of people with a health problem such as diabetes, heart disease or cancer lower for those who meet the government's exercise recommendation than it is for those who don't?

Assume a random sample is taken from both populations. For the people who meet the recommended amount of exercise, 84 out of 560 had a health problem. For the people who did not exercise enough, 204 out of 850 had a health problem.

The hypotheses are:

$H_0 : P_{\text{exercise}} = P_{\text{don't}}$
$H_1 : P_{\text{exercise}} < P_{\text{don't}}$

The sample proportions are $\hat{p}_{exercise} = \dfrac{x}{n} = \dfrac{84}{560} = 0.15$ and $\hat{p}_{don't} = \dfrac{x}{n} = \dfrac{204}{850} = 0.24$

The pooled proportion is $\hat{p}_c = \dfrac{x_A + x_B}{n_A + n_B} = \dfrac{84 + 204}{560 + 850} = 0.204$

The test statistic is $z = \dfrac{(\hat{p}_A - \hat{p}_B) - (p_A - p_B)}{\sqrt{\hat{p}_c(1 - \hat{p}_c)(\dfrac{1}{n_A} + \dfrac{1}{n_B})}}$

with substitution $z = \dfrac{(0.15 - 0.24)}{\sqrt{0.204(1 - 0.204)(\dfrac{1}{560} + \dfrac{1}{850})}} = -4.10$

Checking the standard normal distribution table, the z value of -4.10 is below the lowest value in the table (-3.49) therefore the area in the left tail is less than 0.0002. We conclude that at the 0.05 level of significance, the proportion of health problems for people meeting the government's recommendation for exercise is significant less than for people who don't exercise this much (z = -4.10, p < 0.0002, $n_{\text{exercise}} = 560$, $n_{\text{don't}} = 850$).

3. Is the average amount of exercise a college student does in a week greater than 2.5 hours?

For this question, the evidence that needs to be gathered is hours of exercise in a week. That is quantitative data. To use the *t*-test, we need to make sure the data in the sample are approximately normally distributed. The hypotheses that will be tested are:

$H_0 : \mu = 2.5$
$H_1 : \mu > 2.5$

The level of significance is 0.10.

The number of hours of exercise by 20 randomly selected students is shown in the table below.

| 3.7 | 2 | 7.1 | 1.7 | 0 | 0 | 2.1 | 2.9 | 4 | 3.2 |
| 3.4 | 1.3 | 1 | 4.2 | 0 | 1.3 | 2.9 | 5.3 | 4.4 | 2.3 |

A histogram for this data shows that it is approximately normally distributed. The biggest deviation from normality is in the left tail since it isn't possible to exercise less than 0 hours per week.
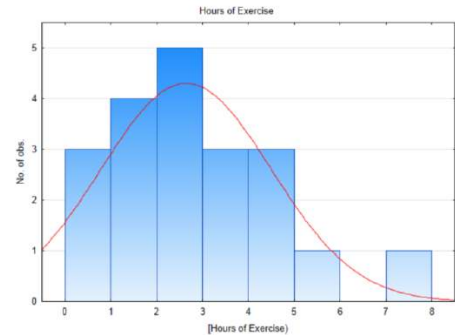
The sample mean and standard deviation are 2.64 hours and 1.855 hours, respectively. The test statistic is: $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$, with

substitution, $t = \dfrac{2.64 - 2.5}{\dfrac{1.855}{\sqrt{20}}}$. After simplification, t = 0.338. There are 19 degrees of freedom $(20 - 1)$. Use the t table, in the row with 19 degrees of freedom, find the location of 0.338. An excerpt of the table is shown below. Notice that 0.338 falls between 0.257 and 0.688 so consequently the table shows that the area in the right tail is between 0.25 and 0.40. Since the level of significance is 0.1, and since the area in the tail is greater than 0.1 and more specifically greater than 0.25, we would report that the p-value is greater than 0.25.

One Tail Probability	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
df								
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257 0.338	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.850

Conclusion: at the 0.10 level of significance, the average time that college students exercise is not significantly greater than 2.5 hours $(t = 0.338, p > 0.25, n = 20)$.

4. Is the average weight of a person less after a month of new regular aerobic fitness program?

For this question, two sets of data must be collected, the before weight and the after weight. The before weight will be subtracted from the after weight to determine the change in weight. Because ultimately there will be only one set of data, the t test for one population mean will be used.

$H_0 : \mu = 0$
$H_1 : \mu < 0$

The level of significance is 0.10.

Subject	1	2	3	4	5	6	7	8	9

Before weight	158	213	142	275	184	136	172	263	205
After weight	154	213	135	278	180	134	171	258	199
After Before	-4	0	-7	3	-4	-2	-1	-5	-6
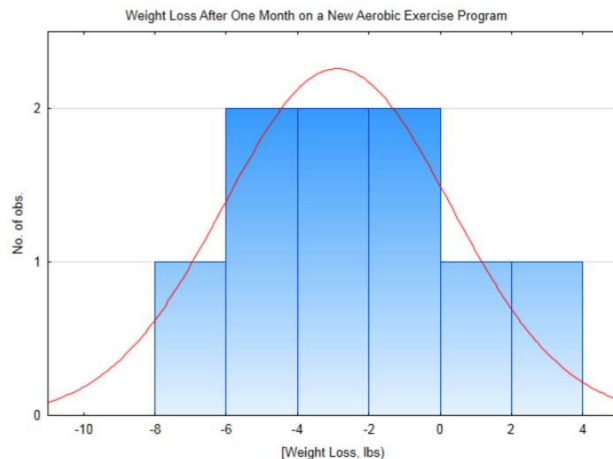
This distribution is approximately normal, so it is appropriate to use the t-test for one population mean. The sample mean is -2.89 lbs with a standard deviation of 3.18 lbs. The test statistic is: $t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$, with substitution, $t = \dfrac{-2.89 - 0}{\frac{3.18}{\sqrt{9}}}$. After simplification, t = -2.726. There are 8 degrees of freedom (9-1). Since -2.726 falls between 2.306 and 2.896 in the row for 8 degrees of freedom and since the level of significance is 0.1 but the area in the tail to the left of -2.726 is less than 0.025, then the conclusion is that the new weight is significantly less than the original weight (t = -2.726, $p < 0.025$, $n = 9$). We conclude people lost weight.

One Tail Probability	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.781

-2.726

5. For those who exercise regularly, is the average amount of exercise a college graduate does in a week different than someone who does not graduate from college?

Assume a random sample is taken for the population of college graduates who exercise regularly and a different random sample is taken from the population of non-graduates who exercise regularly. Also assume that the amount of exercise is normally distributed for both groups and that the variance is homogeneous. The hypotheses are shown below. Use a level of significance of 0.05.

$H_0 : \mu_{\text{college grad}} = \mu_{\text{not college grad}}$
$H_1 : \mu_{\text{college grad}} \neq \mu_{\text{not college grad}}$

The table below shows the mean, standard deviation and sample size for the two samples.

Units: hours/week	College Graduates	Not College Graduations
Mean	4.2	3.8

| Standard Deviation | 1.3 | 1.2 |
| Sample size, n | 12 | 16 |

The difference in sample size means that we need the test statistic formula:

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{[\dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}][\dfrac{1}{n_A} + \dfrac{1}{n_B}]}}$$

which is used for independent populations. Substituting into the formula gives:

$$t = \frac{(4.2 - 3.8) - (0)}{\sqrt{[\dfrac{(12 - 1)1.3^2 + (16 - 1)1.2^2}{12 + 16 - 2}][\dfrac{1}{12} + \dfrac{1}{16}]}} = 0.842$$

Because of the inequality sign in the alternative hypothesis, this is a two-tailed test. The test statistic of 0.842 produces a p-value between 0.5 and 0.8. Since this is clearly higher than the level of significance, the conclusion is that at the 0.05 level of significance, the amount of exercise for college graduates is not significantly different than the amount for non-graduates ($t = 0.842$, $p > 0.5$, $n_{\text{college grads}} = 12$, $n_{\text{not college grads}} = 16$).

Two Tail Probability	0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.001
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.689

0.842

6.

The *t*-test for two independent samples has been based on the assumption of homogeneity of variance. There are tests to determine if the variance is homogeneous and modifications that can be made to the degrees of freedom if it isn't. These are not included in this text.

All of these tests can be done using the TI84 calculator. The tests are found by selecting the STAT key and then using the cursor arrows to move to the right to TESTS.

	Proportions (for categorical data)	Means (for quantitative data)
1 - sample	$H_0 : p = p_0$ $H_1 : p < p_0$ or $p > p_0$ or $p \neq p_0$ Test 5: 1- PropZTest	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$ or $mu > \mu_0$ or $mu \neq \mu_0$ Test 2: T- Test
2 - samples	$H_0 : p_A = p_B$ $H_1 : p_A < p_B$ or $p_A > p_B$ or $p_A \neq p_B$ Test 6: 2- PropZTest	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A < \mu_B$ or $\mu_A > \mu_B$ or $\mu_A \neq \mu_B$ Test 4: 2-SampTTest

5.E: Testing Hypotheses (Exercises)

Chapter 5 Homework

1. For each of the following questions, write the hypotheses that would be tested and then determine which hypothesis test should be used. Select from the following four choices.

   - 1 proportion Z test
   - 2 proportion Z test
   - 1 sample t test
   - 2 independent samples t test

   a. Is the average commute time different when people use transit compared to if they drove?

   H_0:_____ H_1:_____ Test: _____

   b. Do a majority of people eat raw cookie dough?

   H_0:_____ H_1:_____ Test: _____

   c. In a statistics class, is the proportion of STEM students different than the proportion of social science student?

   H_0:_____ H_1:_____ Test: _____

   d. Is the average income of a self-employed person greater than a person working for a large company?

   H_0:_____ H_1:_____ Test: _____

   e. Do you average more than 7 hours of sleep a night?

   H_0:_____ H_1:_____ Test: _____

   f. Is the proportion of students with student loans less than 0.60?

   H_0:_____ H_1:_____ Test: _____

   g. In a long race, such as a marathon, is the difference between the second half split time faster than the first half split time. This can be phrased as: if the first split is subtracted from the second split, will the difference be less than 0? (For example, first half split: 1:06.44, second half split 1:06.01, so 1:06.01 - 1:06.44 = -0.43. The second half was faster than the first half. This is called negative splitting).

   H_0:_____ H_1:_____ Test: _____

   Test the following hypotheses. Assume all assumptions for the tests have been met. Show the formulas, show the substitution and simplification, and write an appropriate concluding sentence.

2. When a young adult leaves home and lives on their own for the first time, they become responsible for feeding themselves. In general, there are two options, eat out or cook for themselves. Suppose someone hypothesized that the average the spent on eating out, including tax and tip, is less than $15 per day. For 30 days, they keep all their receipts and find the mean amount spent per day is $14.28 with a standard deviation of 4.6. Test the hypothesis that the mean they spend per day is less than $15.

   $\alpha = 0.05$

Write the hypotheses: H_0: _____ , H_1: _____ ,

_____ _____ _____ _____

Formula Substitution Test Statistic p-value

Fill in the blanks for the concluding sentence. At the _____ level of significance, the mean money spent per day _____ significantly less than \$15 (t = _____, p = _____, n = _____).

3. When a child has been adopted, they may grow up wondering why their birth parents did not keep them. Technology and some changes in the laws now allow these children to find their birth parents. However, there is no guarantee the birth parent will be happy to be found. The child runs an enormous risk of being rejected. On the other hand, some birth parents are delighted to be found and fully appreciate the reunification. It has been hypothesized that a majority of the reunions have a favorable outcome. Test this hypothesis if a survey of children who found their birth parents resulted in 118 outof 179 having a favorable result. Let $\alpha = 0.05$

Write the hypotheses: H_0: _____ , H_1: _____ ,

_____ _____ _____ _____

Formula Substitution Test Statistic p-value

Fill in the blanks for the concluding sentence. At the _____ level of significance, the proportion of reunions that are favorable _____ significantly greater than 0.50 (z = _____, p = _____, $n =$_____)

4. US News and World Report has stated what many think. The cheapest time to buy airline tickets is on Tuesdays after 3 pm Eastern time. money.usnews.com/money/personal- finance/articles/2012/04/18/8-insider-secrets-to-booking-cheap-airfare viewed 4-30-17. Use the data in the table below to determine if the mean difference between the Tuesday price and the price the rest of the week is less than 0.

Data was collected from Travelocity.com in May 2017 for round-trip flights on Sept 10-17 2017.

Destination	Airlines	Day	Price	Tuesday	Tuesday - Day
Seattle to Boston	Alaska	Sunday	420.40	420.40	0
Chicago to Dallas	United	Thursday	446.40	470.40	24
San Francisco to Orlando	Delta	Thursday	399.60	362.60	-37
Memphis to Phoenix	American	Saturday	277.90	277.90	0
Denver to Las Vegas	Frontier	Thursday	205.95	195.98	-9.97
New York to LA	Jet Blue	Saturday	411.40	411.40	0
Albuquerque to Philadelphia	American	Sunday	513.20	513.20	0

H_0: _____ , H_1: _____ , $\alpha = 0.05$

_____ _____ _____ _____

Formula Substitution Test Statistic p-value

Write the concluding sentence:

5. The concept of anchoring is one that makes us question our own rationality. An example of anchoring is when you see something you want to buy, but can't afford and are then told the item is on sale. Suddenly, it looks like a good price. On the other hand, if someone told you they had bought the same item for half the price you paid, you would feel cheated. Daniel Kahneman and Amos Tversky were researchers who investigated anchoring. This concept was tested in an experiment with the Spring 2017 statistics class. The class was asked to write down the last two numbers of their phone number. Then they were asked to write down their guess for the population of Morocco. The objective is to determine if the people who wrote down numbers that were greater than or equal to 60 (last 2 digits of phone number) had a higher estimate of Morocco's population than those with numbers less than or equal to 40. Use a level of significance of 0.10.

Write the hypotheses:

H_0: _____ , H_1: _____ ,
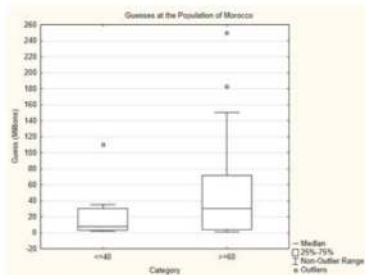
Last 2 digits of phone number	n	Mean	Minimum	Maximum	Standard Deviation
<=40	15	19.6	1.5	110	27.67
>=60	15	58.1	1.0	250	76.38

Formula Substitution

Test Statistic p-value significant?

Write the concluding sentence:

6. Since more people are aware of the problem of waste and are attempting to do their own part in reducing wasted bags by bringing their own reusable bags to a grocery store or not using a bag at all, is it possible that we now have fewer than half of shoppers using bags provided by the grocery store? Data: Out of 750 shoppers, 282 used a paper bag provided by the store (plastic bags are illegal in Seattle where this study by statistics students was conducted). $\alpha = 0.05$

H_0: _____ , H_1: _____ , What is the sample proportion? _____

_____ _____ _____ _____

Formula Substitution Test Statistic p-value

7. The Tacoma-Pierce County Health Department conducts a Healthy Youth Survey to assess the health related behaviors of Pierce County.(www.tpchd.org/resources/publi...-health-risks/) The survey is given to students in grades 6,8,10 and 12. The data from the 2002 and 2012 reports will be used to determine if there has been an increase in the use of marijuana or hashish in 12th grade students. Use a 5% level of significance.

a. Write the hypotheses that will be tested.

The Data: In 2012, $x_{2012} = 165, n_{2012} = 630, x_{2002} = 537, n_{2002} = 2184$

b. Find the sample proportions for each year. 2002_____ 2012 _____

c. Test the hypotheses.

_____ _____ _____ _____

Formula Substitution Test Statistic p-value

d. Write a concluding sentence. At the 5% level of significance

_____

_____

e. Washington just legalized recreational use of marijuana for adults. Do you expect the use of marijuana by 12th graders to increase, remain the same, or decrease because of this new law? Why? Circle one: increase remain the same decrease Why?

_____

_____

8.

Now that women are allowed to compete on an equal basis with men, we can explore the differences in performance in longer athletic events. One of the most grueling competitions is the Ironman Triathlon in which participants swim 2.4 miles, bicycle 112 miles and then run a full 26.2-mile marathon. Results from the 2013 Canadian Ironman Triathlon in Whistler, BC will be used to compare the times of the non-professional participants. The question is whether there is a significant difference in the mean times of the men and women who finished the course. Use a 5% level of significance.

a. Complete the design layout table.

Research Design Table	
Research Question:	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential confounding variables.	
Grouping/explanatory Variables 1 (if present)	Levels:

b. Write the hypotheses. H_0: _____, H_1: _____, $\alpha = 0.05$

Data

			Men									Women				
12.4	13.2	11.8	13.2	11.1	12.6	13.1	10.0	12.2	11.9		13.6	14.2	13.1	12.7	14.0	14.2
11.3	9.9	16.0	14.5	11.2	9.7	16.1	13.7	13.5	14.4		10.9	16.4	15.9	12.8	15.5	12.0
11.1	9.7	14.0	12.0	11.7	11.0	14.9	10.2	12.6	13.5		10.7	15.0	13.7	13.5	11.7	11.5
16.3	13.4	12.7	12.6	10.7	10.5	15.0	14.1	12.9	16.6		12.9	15.7	12.9	11.3		
13.5	15.1	13.3	12.1	14.5	11.9	12.9	12.3	11.8	9.9							
10.2	11.2	13.0														

Enter the data into your calculator to find the statistics needed for 7e and 7f.

c. Make a side-by-side box plot.

d. Find the mean and standard deviation for the men and women.

	n	mean	standard deviation
Men			
Women			

e. Test the hypotheses (you may use your calculator, you don't need to write the formula).

_____ _____

Test Statistic p-value

f. Write a concluding sentence.

g Based on the results of this study, what can be concluded about the physical capabilities of women compared to men in endurance activities? _____ _____

9. Why Statistical Reasoning is Important for Psychology Students and Professionals In collaboration with Tom Link, Professor of Psychology Based on the research article "How Does Stigma "Get Under the Skin"?"(Hatzenbuehler, M. L., Nolen-Hoeksema, S., & Dovidio, J. (2009). How does stigma "get under the skin"? The mediating role of emotion regulation. *Psychological Science*, *20*(10), 1282-1289.) by Mark Hatzenbuehler, Susan Nolen-Hoeksema, and John Dovidio, published in the Jorunal of the Association for Psychological Science in 2009.

This topic is taught in Psyc 100 (General Psychology) and Psyc 210 (Social Psychology)

What strategies are most effective for avoiding adverse mental health issues for victims of discrimination?

Briefing 6.3

Various groups of people feel stigma-related stress. The concept of social stigma was originally discussed by psychologist Erving Goffman in the 1963 publication Stigma: Notes on the Management of Spoiled Identity. Stigma is defined as "a set of negative and often unfair beliefs that a society or group of people have about something". (www.merriam-webster.com 11-14-13) These groups include African Americans, LGB (lesbian, gay, bisexual), women, criminals, and obese individuals. There is some research that links stigma-related stressors to adverse mental and behavioral health. Stigma can be concealed for LGB individuals or criminals but not for the other groups. This means that people cannot tell if another

person is gay unless they have been told, so the stigma is concealed from them (also called discreditable stigma). In contrast, skin color or gender is evident to another person instantly (also called discredited stigma).

Individuals subjected to discrimination have a variety of ways to respond. Two of the ways include rumination and distraction. Rumination is the process of focusing on yourself and reflecting on why you received such treatment. Distraction is to have your thoughts focused on something other than how you feel and the discrimination situation. Psychological distress, the dependent variable, will be used as a measure of mental health. It will be measured with a commonly used test.

There are three questions that will be asked in this problem. To answer these three questions, the researchers used undergraduate students and community members. The data to answer the first question will come from a daily diary the subjects maintain. The data for the second question will come from surveys given to the subjects after they wrote about a time when they were victims of discrimination. The data for the third question will be based on another survey the subjects take after they are directed towards one of the two coping mechanisms, rumination or distraction, using a random process (e.g. coin flip).

You may use your calculator to test these hypotheses. You do not need to write the formulas or show substitution.

a. Is the proportion of days with a discriminatory incident different for African Americans than it is for LGB individuals? $\alpha = 0.05$

African American and LGB subjects maintained a journal for 20 days.

	n	At least one stigma- related stressor	days of data
African Americans	19	139	190
LGB	31	226	310

H_0: _____, H_1: _____,

Test Statistic p-value

Conclusion:

b. Is there a significant difference between the mean psychological distress score of African Americans and LGB individuals following an experiment in which the individual had to recall a discrimination issue they faced? $\alpha = 0.05$

	n	Average Psychological Distress	Standard Deviation
African Americans	19	4.71	3.96
LGB	31	4.51	4.52

H_0: _____, H_1: _____,

Test Statistic p-value

Conclusion:

c. Is there a significant difference between the mean psychological distress score of those who used rumination to cope with the discrimination and those who used distraction?

Is there a significant difference between the psychological distress score of those who used rumination to cope with the discrimination and those who used distraction?

	n	Mean	Standard Deviation
Rumination	26	13.24	6.14
Distraction	26	10.07	4.10

H_0: _____, H_1: _____,

Test Statistic p-value

Conclusion:

d. The subjects in this research were undergraduate students and community members living near Yale University in Connecticut. Does this affect the conclusions that have been drawn? Why or why not?

6: Confidence Intervals and Sample Size

The inferences that were discussed in chapters 5 and 6 were based on the assumption of an *a priori* hypothesis that the researcher had about a population. However, there are times when the researchers do not have a hypothesis. In such cases they would simply like a good estimate of the parameter.

By now you should realize that the statistic (which comes from the sample) will most likely not equal the parameter of the population, but it will be relatively close since it is part of the normally distributed collection of possible statistics. Consequently, the best that can be claimed is that the statistic is a **point estimate** of the parameter. Because half the statistics that could be selected are higher than the parameter and half are lower, and because the variation that can be expected for statistics is dependent, in part, upon sample size, then the knowledge of the statistic is insufficient for determining the degree to which it is a good estimate for the parameter. For this reason, estimates are provided with confidence intervals instead of point estimates.

You are probably most familiar with the concept of confidence intervals from polling results preceding elections. A reporter might say that 48% of the people in a survey plan to vote for candidate A, with a margin of error of plus or minus 3%. The interpretation is that between 45% and 51% of the population of voters will vote for candidate A. The size of the margin of error provides information about the potential gap between the point estimate (statistic) and the parameter. The interval gives the range of values that is most likely to contain the true parameter. For a confidence interval of (0.45,0.51) the possibility exists that the candidate could have a majority of the support. The margin of error, and consequently the interval, is dependent upon the degree of confidence that is desired, the sample size, and the standard error of the sampling distribution.

The logic behind the creation of confidence intervals can be demonstrated using the empirical rule, otherwise known as the 68-95-99.7 rule that you learned in Chapter 5. We know that of all the possible statistics that comprise a sampling distribution, 95% of them are within approximately 2 standard errors of the mean of the distribution. From this we can deduce that the mean of the distribution is within 2 standard errors of 95% of the possible statistics. By analogy, this is equivalent to saying that if you are less than two meters from the student who is seated next to you, then that student is less than two meters from you. Consequently, by taking the statistic and adding and subtracting two standard errors, an interval is created that should contain the parameter for 95% of the statistics we could get using a good random sampling process.

When using the empirical rule, the number 2, in the phrase "2 standard errors", is called a **critical value**. However, a good confidence interval requires a critical value with more precision than is provided by the empirical rule. Furthermore, there may be a desire to have the degree of confidence be something besides 95%. Common alternatives include 90% and 99% confidence intervals. If the degree of confidence is 95%, then the critical values separate the middle 95% of the possible statistics from the rest of the distribution. If the degree of confidence is 99%, then the critical values separate the middle 99% of the possible statistics from the rest of the distribution. Whether the critical value is found in the standard normal distribution (a z value) or in the t distributions (a t value) is based on the whether the confidence interval is for a proportion or a mean.

The critical value and the standard error of the sampling distribution must be determined in order to calculate the margin of error.

The critical value is found by first determining the area in one tail. The area in the left tail (AL) is found by subtracting the degree of confidence from 1 and then dividing this by 2.

$$A_L = \frac{1 - \text{degree of confidence}}{2}. \tag{6.1}$$

For example, substituting into the formula for a 95% confidence interval produces

$$A_L = \frac{1 - 0.95}{2} = 0.025 \tag{6.2}$$

The critical Z value for an area to the left of 0.025 is -1.96. Because of symmetry, the critical value of an area to the right of 0.025 is +1.96. This means that if we find the critical values corresponding to an area in the left tail of 0.025, that we will find the lines that separate the group of statistics with a 95% chance of being selected from the group that has a 5% chance of being selected.
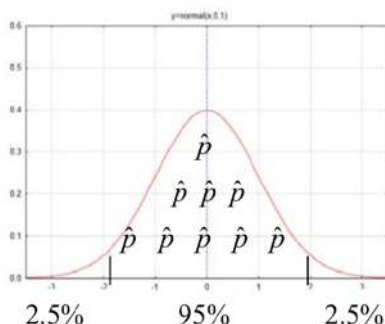
An area in the left tail of 0.025, which is found in the body of the z distribution table, corresponds with a z^* value of -1.96. This is shown in the section of the Z table shown below.

Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446

The critical z value of -1.96 is also called the 2.5th percentile. That means that 2.5% of all possible statistics are below that value.

Critical values can also be found using a TI 84 calculator. Use 2nd Distr, #3 invnorm (percentile, μ, σ). For example invnorm(0.025,0,1) gives -1.95996 which rounds to -1.96.

Confidence intervals for proportions always have a critical value found on the standard normal distribution. The z value that is found is given the notation z^*. These critical values vary based on the degree of confidence. The other most common confidence intervals are 90% and 99%. Complete the following table below to find these commonly used critical values.

Degree of Confidence	Area in Left Tail	z^*
0.90		
0.95	0.025	1.96
0.99		
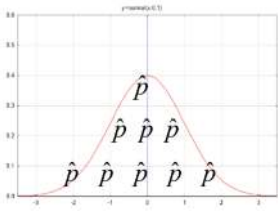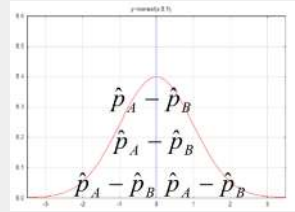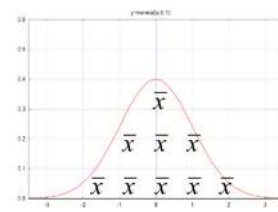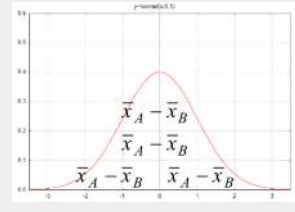
Confidence intervals for means require a critical value, t^*, which is found on the t tables. These critical values are dependent upon both the degree of confidence and the sample size, or more precisely, the degrees of freedom. The top of the t-table provides a variety of confidence levels along with the area in one or both tails. The easiest approach to finding the critical t^* value is to find the column with the appropriate confidence level then find where that column intersects with the row containing the appropriate degrees of freedom. For example, the t^* value for a 95% confidence interval with 7 degrees of freedom is 2.365.

One Tail Probability	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
Two Tail Probability	0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.001
Confidence Level	20%	50%	80%	90%	95%	98%	99%	99.9%
df								
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	5.041

The second component of the margin of error, which is the standard error for the sampling distribution, assumes knowledge of the mean of the distribution (e.g. $\mu_{\hat{p}} = p$ and $\mu_{\bar{x}} = \mu$). When testing hypotheses about the mean of the distribution, we assume these values because we assume the null hypothesis is true. However, when creating confidence intervals, we admit to not knowing these values and so consequently we cannot use the standard error. For example, the standard error for the $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$. Since we don't know p, we can't use this formula. Likewise, the standard error for the distribution of sample means is $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$. To find σ we need to know the population mean, μ, but once again we don't know it, and we don't even have a hypothesis about it, so consequently we can't find σ. The strategy in both these cases is to find an estimate of the standard error by using a statistic to estimate the missing parameter. Thus, $\hat{p}$ is used to estimate p and s is used to estimate σ. The estimated standard errors then become: $s_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ and $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$.

The groundwork has now been laid to develop the confidence interval formulas for the situations for which we tested hypotheses in the preceding chapter, namely p, $p_A - p_B$, μ, and $\mu_A - \mu_B$. The table below summarizes these four parameters, their distributions and estimated standard errors.

Parameter	Distribution	Estimated Standard Error

Parameter	Distribution	Estimated Standard Error
Proportion for one population, p	$\hat{p}$ ... $\hat{p}\ \hat{p}\ \hat{p}$ ... $\hat{p}\ \hat{p}\ \hat{p}\ \hat{p}\ \hat{p}$	$s_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
Difference between proportions for two populations, $p_A - p_B$	$\hat{p}_A - \hat{p}_B$ ... $\hat{p}_A - \hat{p}_B$ ... $\hat{p}_A - \hat{p}_B\ \ \hat{p}_A - \hat{p}_B$	$s_{\bar{p}_A - \bar{p}_B} = \sqrt{\dfrac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \dfrac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$
Mean for one population or mean difference for dependent data, μ	$\bar{x}$ ... $\bar{x}\ \bar{x}\ \bar{x}$ ... $\bar{x}\ \bar{x}\ \bar{x}\ \bar{x}\ \bar{x}$	$s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$
Difference between means of two independent populations, $\mu_A - \mu_B$	$\bar{x}_A - \bar{x}_B$ ... $\bar{x}_A - \bar{x}_B$ ... $\bar{x}_A - \bar{x}_B\ \ \bar{x}_A - \bar{x}_B$	$s_{\bar{x}_A - \bar{x}_B} = \sqrt{\left[\dfrac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}\right]\left[\dfrac{1}{n_A} + \dfrac{1}{n}\right.}$

The reasoning process for determining the formulas for the confidence intervals is the same in all cases.

1. Determine the degree of confidence. The most common are 95%, 99% and 90%.
2. Use the degree of confidence along with the appropriate table (z* or t*) to find the critical value.
3. Multiply the critical value times the standard error to find the margin of error.
4. The confidence interval is the statistic plus or minus the margin of error.

Notice that all the confidence intervals have the same format, even though some look more difficult than others.

$$\text{statistic} \pm \text{margin of error}$$
$$\text{statistic} \pm \text{critical value} \times \text{estimated standard error}$$

Confidence intervals about the proportion for one population:

$$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})n} \tag{6.3}$$

Confidence intervals for the difference in proportions between two populations:

$$(\hat{p}_A - \hat{p}_B \pm z^* \sqrt{\dfrac{\hat{p}_A \hat{q}_A}{n_A} + \dfrac{\hat{p}_B \hat{q}_B}{n_B}} \tag{6.4}$$

Remember that $q = 1 - p$.

Confidence intervals for the mean for one population:

$$\bar{x} \pm t^* \dfrac{s}{\sqrt{n}} \tag{6.5}$$

Confidence interval for the difference between two independent mean:

$$\left(\bar{x}_A + \bar{x}_B \pm t^* \left(\sqrt{\left[\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}\right]\left[\frac{1}{n_A} + \frac{1}{n_B}\right]}\right)\right) \tag{6.6}$$

where t^* is the appropriate percentile from the t(n_A + n_B - 2) distribution.

The confidence interval formulas are organized below in the same way the hypothesis test formulas were organized in Chapter 6. You should see a similarity between corresponding formulas.

	Proportions (for categorical data)	Means (for quantitative data)
1 - sample	$\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})}\, n$ Assumptions: $np \geq 5, n(1 - p) \geq 5$	$\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$ df = n - 1 Assumptions: If $n < 30$, population is approximately normally distributed.
2 - samples	$(\hat{p}_A - \hat{p}_B \pm z^* \sqrt{\dfrac{\hat{p}_A \hat{q}_A}{n_A} + \dfrac{\hat{p}_B \hat{q}_B}{n_B}}$ Assumption: $np \geq 5, n(1 - p) \geq 5$ for both population	$\left(\bar{x}_A + \bar{x}_B \pm t^*\left(\sqrt{\left[\dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}\right]\left[\dfrac{1}{n_A}\right.}\right.\right.$ df = n_A + n_B - 2 Assumptions: If $n < 30$, population is approximately normally distributed.

What does a confidence interval mean? For a 95% confidence interval, 95% of all possible statistics are within z* (or t*) standard errors of the mean of the distribution. Therefore, there is a 95% probability that the data that is randomly selected will produce one of those statistics and the confidence interval that is created will contain the parameter. Whether the interval ultimately does include the parameter or not is unknown. We only know that if the sampling processes was repeated a large number of times producing many confidence intervals, about 95% of them would contain the parameter.

> **Example 1**
>
> In an automaticity experiment community college students were given two opportunities to go to the computer lab to test their automaticity skills (math fact fluency). Students were randomly assigned to use one of two practice programs to determine if one program leads to greater improvement than the other. These programs will be called program A and program B.
>
> a. What is the 95% confidence interval for the proportion of students who improve from their first attempt to their second attempt if 99 out of 113 students improved?
>
> To help pick the correct confidence interval formula, notice this problem is about proportions and there is only one group of student.
>
> The formula that meets these criteria is: $\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})}\, n$. Before substituting, it is necessary to calculate $\hat{p}$. Since $\hat{p} = \dfrac{x}{n} = \dfrac{99}{113} = 0.876$ (round to 3 decimal places), then the confidence interval formula becomes $0.876 \pm 1.96 \sqrt{\dfrac{0.876(1 - 0.876)}{113}}$. This simplifies to 0.876 ± 0.061. The margin of error is 6.1%. The confidence interval is (0.815,0.937). The conclusion is that we are 95% confident that the true proportion of students who would improve from one test to the next is between 0.815 (81.5%) and 0.937 (93.7%).
>
> > To find this interval on the TI 84 calculator, select Stat, Tests, A 1-PropZInt.
>
> b. What is the 90% confidence interval for the difference in the proportion of students who improved from their first attempt to their second attempt using Program A (37/45) and Program B (61/67)?
>
> To help pick the correct confidence interval formula, notice this problem is about proportions and there are two different populations – one using Program A and the other using Program B.
>
> The formula that meets these criteria is: $(\hat{p}_A - \hat{p}_B \pm z^* \sqrt{\dfrac{\hat{p}_A \hat{q}_A}{n_A} + \dfrac{\hat{p}_B \hat{q}_B}{n_B}}$. Since $\hat{p}_A = \dfrac{x}{n} = \dfrac{37}{45} = 0.822$ and $\hat{p}_B = \dfrac{x}{n} = \dfrac{61}{67} = 0.919$, then the confidence interval formula becomes
>
> $0.822 - 0.910) \pm 1.645 \sqrt{\dfrac{0.822(1 - 0.822)}{45} + \dfrac{0.910(1 - 0.910)}{67}}$
>
> After simplification the confidence interval can be written as a statistic ± margin of error: -0.088 ± 0.110. The margin of error is 11.0%. The confidence interval is (-0.198,0.022). The conclusion is that we are 90% confident that the true difference in proportions between those using

Program A and those using Program B is between -0.198 and 0.022. Notice that 0 falls within that range, which indicates there is potentially no difference between these two proportions.

> To find this interval on the TI 84 calculator, select Stat, Tests, B 2-PropZInt.

c. What is the 99% confidence interval for the average improvement from Introductory Algebra students using program B (mean = 5.0, SD = 3.18, n = 19).

To help pick the correct confidence interval formula, notice this problem is about means and there is only one population (Introductory Algebra students using program B).

The formula that meets these criteria is: $\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$. There are 18 degrees of freedom (df = n-1) so the t^* value is 2.878. After substituting for all the variables the formula becomes $5.0 \pm 2.878 \dfrac{3.18}{\sqrt{19}}$. This simplifies to 5.0 ± 2.1. The confidence interval is (2.9, 7.1).

> To find this interval on the TI 84 calculator, select Stat, Tests, #8 Tinterval.

d. What is the 95% confidence for the difference in improvement between introductory algebra and intermediate algebra students using program A. The statistics for introductory algebra are mean = 2.4, SD = 3.53, n = 16. The statistics for intermediate algebra are mean = 4, SD = 4.89, n = 21.

To help pick the correct confidence interval formula, notice this problem is about means but there are two populations (introductory algebra students and intermediate algebra students). The formula that meets these criteria is:

$$(\bar{x}_A + \bar{x}_B \pm t^* (\sqrt{[\dfrac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}][\dfrac{1}{n_A} + \dfrac{1}{n_B}]})$$

There are 35 degrees of freedom ($n_A + n_B$ - 2). Unfortunately, this value does not exist on the t table in Chapter 6, so it will be necessary to estimate it. One approach is to use the critical value for 30 degrees of freedom (2.042) which is larger than the critical value for 40 degrees of freedom (2.021) as this will ensure that the confidence interval is at least as large as necessary. After substituting for all the variables, the formula becomes $(2.4 - 4) \pm 2.042(4.359\sqrt{\dfrac{1}{16} + \dfrac{1}{21}})$ and with simplification -1.6 ± 2.95. The interval is (-4.55,1.35). Because the critical t^* value is slightly larger than it should be, the interval is slightly wider than it would be calculated using the functions on a TI 84 calculator (-4.537,1.3368).

> The second approach is to find this interval on the TI 84 calculator. Select Stat, Tests, #0 2-SampTInt.

Sample Size Estimation

The margin of error portion of a confidence interval formula can also be used to estimate the sample size that needed. Let E represent the desired margin of error. If sampling of categorical data for one population is done, then $E = z^* \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$. Solve this for n using algebra. Since the goal is to make sure the sample size is large enough, and since $\hat{p}$ is not known in advance, then it is necessary to make sure that $\hat{p}(1-\hat{p})$ is the largest possible value. That will happen when $\hat{p} = 0.5$.

$$E = z^* \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$$

$$\dfrac{E}{z^*} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$$

$$\dfrac{E^2}{z^{*2}} = \dfrac{\hat{p}(1-\hat{p})}{n}$$

$$n = \dfrac{z^{*2}\hat{p}(1-\hat{p})}{E^2}$$

$$n = \dfrac{z^{*2}0.5(0.5)}{E^2}$$

$$n = \dfrac{0.25 z^{*2}}{E^2} \text{ or } n = \dfrac{z^{*2}}{4E^2}$$

> **Example 2**
>
> Estimate the sample size needed for a national presidential poll if the desired margin of error is 3%. Assume 95% degree of confidence.
>
> $$n = \frac{1.96^2}{4(0.03)^2} = 1067.1 \text{ or } 1068 \text{ (round up to get enough in the ample).}$$

6.E: Confidence Intervals and Sample Size (Exercises)

Chapter 6 Homework

> ### Briefing 6.1 Gender gap in Science
>
> A variety of explanations have been provided for why males are more likely to study science and have a profession in the field of science than are females. One explanation is that teachers are more likely to encourage boys to ask questions and integrate concepts. Kevin Crowley and other researchers sought to answer questions about the role of parents in contributing to the gender gap in science.(Crowley, K., Callanan, M. A., Tenenbaum, H. R., & Allen, E. (2001). Parents explain more often to boys than to girls during shared scientific thinking. *Psychological Science, 12*(3), 258-261.) Their research was published in Psychological Science, May 2001.
>
> The research was conducted at a children's museum using video cameras and wireless microphones. It forms the basis for the first four questions.

1. Find the 95% confidence interval for the proportion of times a boy chose to interact with an exhibit at the museum if 144 out of 185 boys initiated this interaction? This means the child chose to interact without parental encouragement.

   Formula Substitution Margin of Error Confidence Interval

   Calculator confidence interval_____

2. Find the 99% confidence interval for the difference in the proportion of times a boy initiated interaction with the exhibit and a girl initiated interaction with the exhibit. Out of 185 boys, 144 initiated interaction. Out of 113 girls, 84 initiated interaction.

   Formula Substitution Margin of Error Confidence Interval

   Calculator confidence interval_____

3. Find the 90% confidence interval for the mean length of time girls remained engaged with the exhibit if the sample mean time is 88 seconds, the standard deviation is 93 seconds and there were 113 girls.

Formula Substitution Margin of Error Confidence Interval

Calculator confidence interval_____

4. Find the 95% confidence interval for the difference in the mean length of time boys remained engaged with an exhibit (mean = 107 sec, SD = 117 sec, n = 185) and girls remained engaged (mean = 88 sec, SD = 93 sec, n = 113) with the exhibit.

Formula Substitution Margin of Error Confidence Interval

Calculator confidence interval_____

5. What is the 90% confidence interval for the difference in the mean weight of hatchery and wild Coho salmon that have returned to spawn? What is the point estimate for the difference? (Student project, Summer 2002)

	Hatchery	Wild
Mean	2434 grams	2278 grams
Median	2234 grams	2048 grams
Standard Deviation	1066 grams	1000 grams
Sample Size	602	745

Point estimate _____

Formula Substitution Margin of Error Confidence Interval

Calculator confidence interval_____

6. If a person cannot afford to pay for heat, how much warmer will their home be than the outside temperature? Outside and inside temperatures were recorded for a vacant log cabin. Find the point estimate for the difference between outside and inside air temperature. Find the 95% confidence interval for the difference between outside air temperature and inside air temperature (inside – outside). Temperatures are recorded in degrees Celsius. (student project Winter 2002)

Outside	Inside	Inside - Outside
2.2	10.5	
6.1	10.5	
8.3	12.2	

6.7	13.3	
13.3	11,7	
15.5	12.8	
3.9	11.1	
2.2	10.0	
7.8	9.4	
0.5	8.9	
-3.3	10	

Point estimate _____

Formula Substitution Margin of Error Confidence Interval

Calculator confidence interval_____

Can it be concluded that the inside temperature is warmer than the outside temperature?

7. An experiment was conducted at a photo copy store in which coupons were given to customers. Half of the coupons were black and white while the other half were printed on bright yellow paper. The printing on both was identical as was the amount of discount the customers received (10%). What is the point estimate for the difference in the proportion of color and of black and white coupons that were returned? What is the 95% confidence interval for the difference in the proportion of color and of black and white coupons that were returned? (student project)

	Color	Black and White
Number returned (used)	129	87
Number distributed	250	250

Point estimate _____

Formula Substitution Margin of Error Confidence Interval

Calculator confidence interval_____

Can it be concluded that color coupons have a better return (use) rate than black and white coupons?

8. In the early 1900s, males accounted for approximately 10% of all nurses. By 1960, this percentage had fallen to about 2% but since that time it has increased to over 12%. Data was collected from colleges that offered a BS degree in nursing to determine the proportion of the students who are male, as this might give some insight into potential changes within the profession. Out of 2352 nursing students, 273 are male. What is the point estimate? What is the 99% confidence interval for the proportion of male nursing students in a nursing degree program? (based on student project, Brian Walsh Fall 2013)

Point estimate _____

Formula Substitution Margin of Error Confidence Interval

Calculator confidence interval_____

9. Determine the effect of the desired margin of error on the size of the samples that must be taken for 1 population categorical data. Complete the chart. Show formula and substitutions. Use a 95% degree of confidence.

Margin of Error	1%	5%	10%	20%
Sample Size				

What do you conclude?

10. Determine the effect of the degree of confidence on the size of the samples that must be taken for 1 population categorical data. Use a margin of error of 3%.

Degree of Confidence	99%	95%	90%	80%
Sample Size				

What do you conclude?

11. Why Statistical Reasoning Is Important for a Diagnostic Health and Fitness Technician (DHFT) Student and Professional

Developed in collaboration with
Lisa Murray, Professor of HSCI, Nutrition and Physical Education.
This topic is discussed in Nutrition 101.

The FDA recommends that daily sodium intake should not exceed 2300 mg per day. High sodium consumption has been shown to have a negative effect on blood pressure and other health problems.

One of the more popular treats for moviegoers is popcorn. Popcorn by itself is considered a healthy snack, but adding oil, butter, and salt to it can decrease its nutritional value. To estimate the salt content of movie theater popcorn, popcorn of various sizes will be purchased from randomly selected theaters and then sent to a lab for analysis. The final results will be presented as mg of sodium per cup of popcorn. In this case, we don't have a hypothesis about the amount, so the objective will be to create a confidence interval.

Because different theater chains may use different amounts of salt, a random sample will be taken from each of three large theater companies.

a. What sampling method is being used?

b. If one of the chains has 389 theaters, which 3 theaters would be selected if the calculator is seeded with the number 21?

_____, _____, _____

The data (mg sodium per cup of popcorn):

50	49	49	35	37
36	86	103	88	53
48	54	38	33	33
80	98	95	55	70

(These numbers are based on data from a study in the Nutrition Action Healthletter).

c. Make a frequency distribution and histogram for this data.

d. Find the mean and standard deviation for this sample.

e. Show the 95% confidence interval for the amount of sodium per cup. Include formula, substitution and the interval.

f. If the size of bags of popcorn range from 6 cups to 20 cups, what is the range of sodium that could be consumed by buying popcorn at a theater?

g. How will knowledge of this influence your next purchase of movie theater popcorn?

7: Analysis of Bivariate Quantitative Data

For the past three chapters you have been learning about making inferences for **univariate** data. For each research question that could be asked, only one random variable was needed for the answer. That random variable could be either categorical or quantitative. In some cases, the same random variable could be sampled and compared for two different populations, but that still makes it univariate data. In this chapter, we will explore bivariate quantitative data. This means that for each unit in our sample, two quantitative variables will be determined. The purpose of collecting two quantitative variables is to determine if there is a relationship between them.

The last time the analysis of two quantitative variables was discussed was in Chapter 4 when you learned to make a scatter plot and find the correlation. At the time, it was emphasized that even if a correlation exists, that fact alone is insufficient to prove causation. There are a variety of possible explanations that could be provided for an observed correlation. These were listed in Chapter 4 and provided again here.
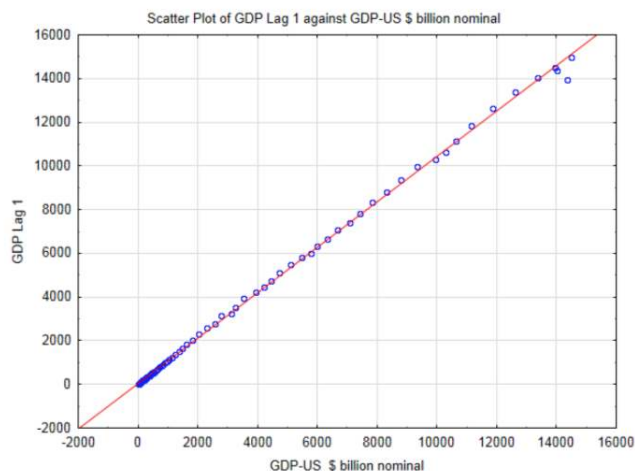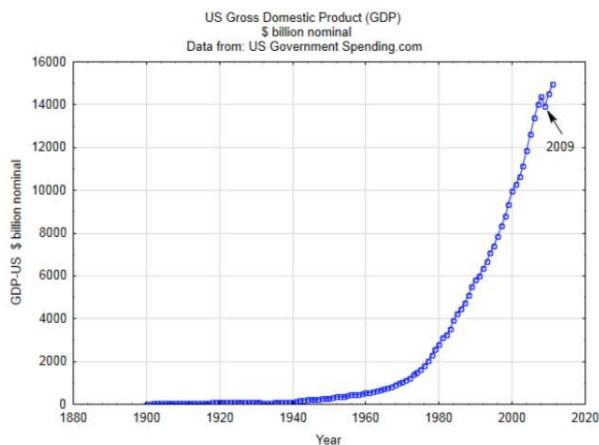
1. Changing the x variable will cause a change in the y variable
2. Changing the y variable will cause a change in the x variable
3. A feedback loop may exist in which a change in the x variable leads to a change in the y variable which leads to another change in the x variable, etc.
4. The changes in both variables are determined by a third variable
5. The changes in both variables are coincidental.
6. The correlation is the result of outliers, without which there would not be significant correlation.
7. The correlation is the result of confounding variables.

Causation is easier to prove with a manipulative experiment than an observational experiment. In a manipulative experiment, the researcher will randomly assign subjects to different groups, thereby diminishing any possible effect from confounding variables. In observational experiments, confounding variables cannot be distributed equitably throughout the population being studied. Manipulative experiments cannot always be done because of ethical reasons. For example, the earth is currently undergoing an observational experiment in which the explanatory variable is the amount of fossil fuels being converted to carbon dioxide and the response variable is the mean global temperature. It would have been considered unethical if a scientist had proposed in the 1800s that we should burn as many fossil fuels as possible to see how it affects the global temperature. Likewise, experiments that would force someone to smoke, text while driving, or do other hazardous actions would not be considered ethical and so correlations must be sought using observational experiments.

There are several reasons why it is appropriate to collect and analyze bivariate data. One such reason is that the dependent or response variable is of greater interest but the independent or explanatory variable is easier to measure. Therefore, if there is a strong relationship between the explanatory and response variable, that relationship can be used to calculate the response variable using data from the explanatory variable. For example, a physician would really like to know the degree to which a patient's coronary arteries are blocked, but blood pressure is easier data to obtain. Therefore, since there is a strong relationship between blood pressure and the degree to which arteries are blocked, then blood pressure can be used as a predictive tool.

Another reason for collecting and analyzing bivariate data is to establish norms for a population. As an example, infants are both weighed and measured at birth and there should be a correlation between their weight and length (height?). A baby that is substantially underweight compared to babies of the same length would raise concerns for the doctor.

In order to use the methods described in this chapter, the data must be independent, quantitative, continuous, and have a bivariate normal distribution. The use of discrete quantitative data exceeds the scope of this chapter. Independence means that the magnitude of one data value does not affect the magnitude of another data value. This is often violated when time series data are used. For example, annual GDP (gross domestic product) data should not be used as one of the random variables for bivariate data analysis because the size of the economy in one year has a tremendous influence on the size of it the next year. This is shown in the two graphs below. The graph on the left is a time series graph of the actual GDP for the US. The graph on the right is a scatter plot that uses the GDP for the US as the x variable and the GDP for the US one year later (lag 1) for the y value. The fact that these points are in such a straight line indicates that the data are not independent. Consequently, this data should not be used in the type of the analyses that will be discussed in this chapter.

US Gross Domestic Product (GDP)
$ billion nominal
Data from: US Government Spending.com

Scatter Plot of GDP Lag 1 against GDP-US $ billion nominal

A bivariate normal distribution is one in which y values are normally distributed for each x value and x values are normally distributed for each y value. If this could be graphed in three dimensions, the surface would look like a mountain with a rounded peak.

We will now return to the example in chapter 4 in which the relationship between the wealth gap, as measured by the Gini Coefficient, and poverty were explored. Life can be more difficult for those in poverty and certainly the influence they can have in the country is far more limited than those who are affluent. Since people in poverty must channel their energies into survival, they have less time and energy to put towards things that would benefit humanity as a whole. Therefore, it is in the interest of all people to find a way to reduce poverty and thereby increase the number of people who can help the world improve.

There are a lot of possible variables that could contribute to poverty. A partial list is shown below. Not all of these are quantitative variables and some can be difficult to measure, but they can still have an impact on poverty levels

1. Education
2. Parent's income level
3. Community's income level
4. Job availability
5. Mental Health
6. Knowledge
7. Motivation and determination
8. Physically disabilities or illness
9. Wealth gap
10. Race/ethnicity/immigration status/gender
11. Percent of population that is employed

In Chapter 4, only the relationship between wealth gap and poverty level was explored. Data was gathered from seven states to determine if there is a correlation between these two variables. The scatter plot is reproduced below. The correlation is -0.65.

Scatterplot of Poverty Rate against Gini Coefficient
Based On US States, 2010
Data from US Census Bureau
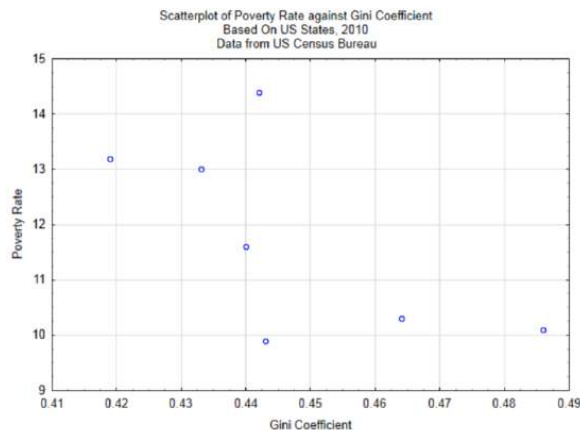
As a reminder, correlation is a number between -1 and 1. The population correlation is represented with the Greek letter ρ, while the sample correlation coefficient is represented with the letter r. A correlation of 0 indicates no correlation, whereas a correlation of 1 or -1 indicates a perfect correlation. The question is whether the underlying population has a significant linear relationship. The evidence for this comes from the sample. The hypotheses that are typically tested are:

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

This is a two-tailed test for a non-directional alternative hypothesis. A significant result indicates only that the correlation is not 0, it does not indicate the direction of the correlation.

The logic behind this hypothesis test is based on the assumption the null hypothesis is true which means there is no correlation in the population. An example is shown in the scatter plot on the left. From this distribution, the probability of getting the sample data (shown in solid circles in the graph at the right), or more extreme data (forming a straighter line), is calculated.
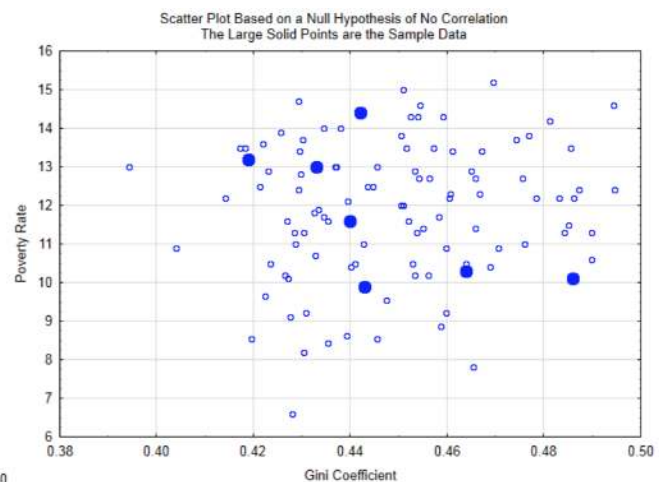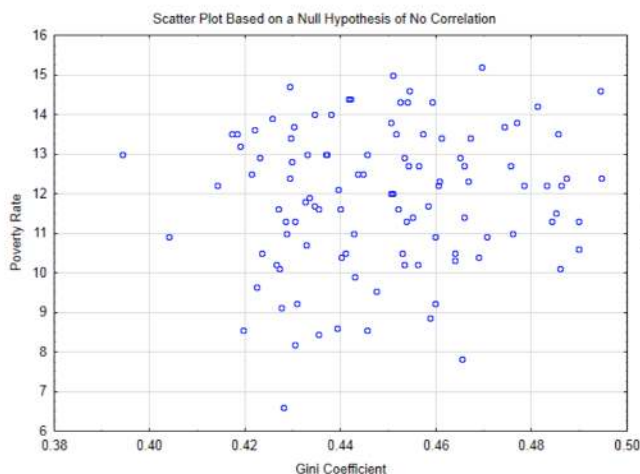
The test used to determine if the correlation is significant is a *t* test. The formula is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \tag{7.1}$$

There are n - 2 degrees of freedom.

This can be demonstrated with the example of Gini coefficients and poverty rates as provided in Chapter 4 and using a level of significance of 0.05. The correlation is -0.650. The sample size is 7, so there are 5 degrees of freedom. After substituting into the test statistic, $t = \dfrac{-0.650\sqrt{7-2}}{\sqrt{1-(-0.650)^2}}$, the value of the test statistic is -1.91. Based on the t-table with 5 degrees of freedom, the two-

sided p-value is greater than 0.10 (actual 0.1140). Consequently, there is not a significant correlation between Gini coefficient and poverty rates.

Another explanatory variable that can be investigated for its correlation with poverty rates is the employment-population ratio (percent). This is the percent of the population that is employed at least one hour in the month

Scatterplot of Poverty Rate against Employment-Population Ratio

Employment-Population Ratio:Poverty Rate: r = -0.6445, p = 0.0174

.
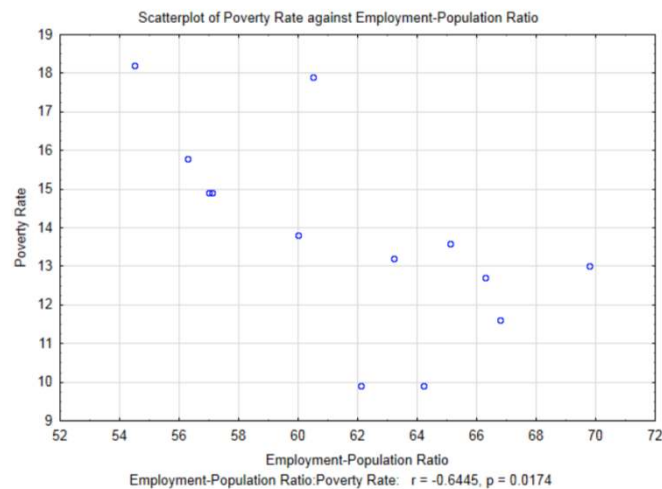
The correlation for this data is -0.6445, $t = -2.80$ and $p = 0.0174$. Notice at the 0.05 level of significance, this correlation is significant. Before exploring the meaning of a significant correlation, compare the results of the correlation between Gini Coefficient and poverty rate which was -0.650 and the results of the correlation between Employment-Population Ratio and poverty rates which is -0.6445. The former correlation was not significant while the later was significant even though it is less than the former. This is a good example of why the knowledge of a correlation coefficient is not sufficient information to determine if the correlation is significant. The other factor that influences the determination of significance is the sample size. The Employment-Population Ratio/poverty rates data was determined from a larger sample size (13 compared with 7). Sample size plays an important role in determining if the alternative is supported. With very large samples, very small sample correlations can be shown to be significant. The question is whether significant corresponds with important.

The effect of sample size on possible correlations is shown in the four distributions below. These distributions were created by starting with a population that had a correlation of $\rho = 0.000$. 10,000 samples of size 5,15,35, and 300 were drawn from this population, with replacement.

Histogram of 10,000 Sample Correlations
From a Population with Correlation = 0
n = 5

Histogram of 10,000 Sample Correlations
From a Population with Correlation = 0
n = 15

Histogram of 10,000 Sample Correlations
From a Population with Correlation = 0
n = 35

Histogram of 10,000 Sample Correlations
From a Population with Correlation = 0
n = 300
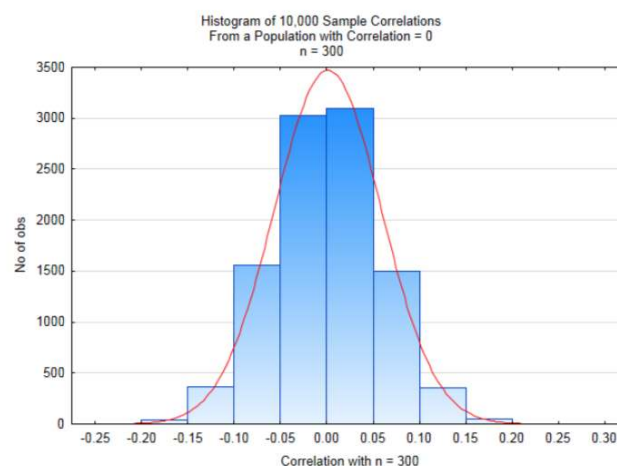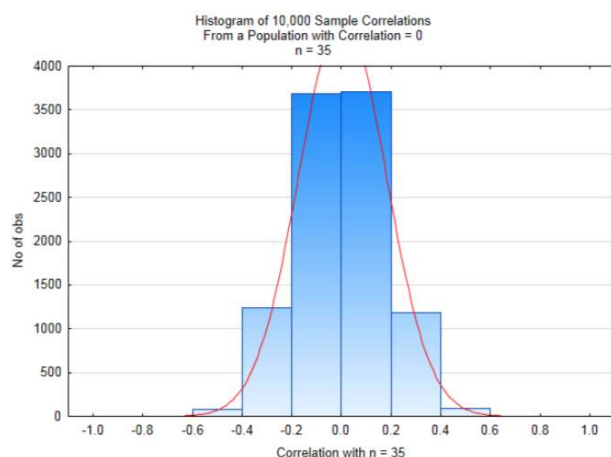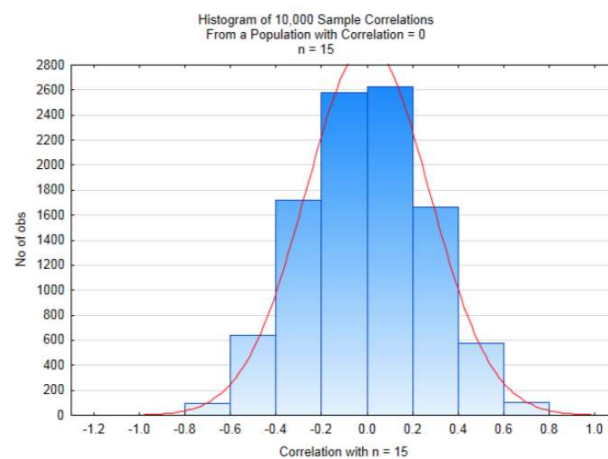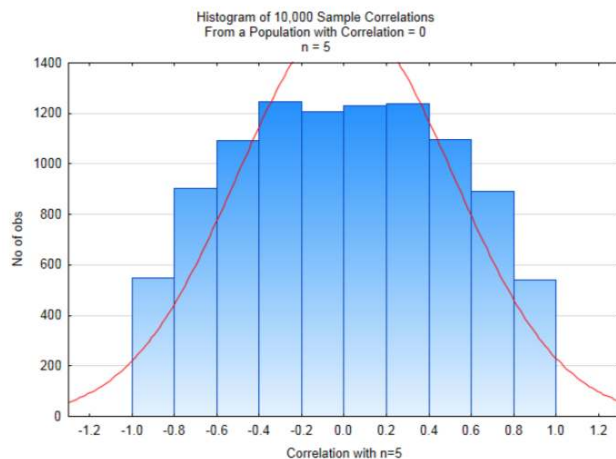
Look carefully at the x-axis scales and the heights of the bars. Values near the middle of the graphs are likely values while values on the far left and right of the graph are unlikely values which, when testing a hypothesis, would possibly lead to a significant conclusion. With small sample sizes, the magnitude of the correlation must be very large to conclude there is significant correlation. As the sample size increases, the magnitude of the correlation can be much smaller to conclude there is significant correlation. The critical values for each of these are shown in the table below and are based on a two-tailed test with a level of significance of 5%.

n	5	15	35	300
t	2.776	2.145	2.032	1.968
$|r|$	0.848	0.511	0.334	0.113

In the histogram in the bottom right in which the sample size was 300, a correlation that exceeds 0.113 would lead to a conclusion of significant correlation, yet there is the question of whether a correlation that small is very meaningful, even if it is significant. It might be meaningful or it might not. The researcher must determine that for each situation.

Returning to the analysis of Gini coefficients and poverty rates, since there was not a significant correlation between these two variables, then there is no point in trying to use Gini Coefficients to estimate poverty rates or focusing on changes to the wealth gap as a way of improving the poverty rate. There might be other reasons for wanting to change the wealth gap, but its impact on poverty rates does not appear to be one of the reasons. On the other hand, because there is a significant correlation between Employment-Population Ratio and poverty rates, then it is reasonable to use the relationship between them as a model for estimating poverty rates for specific Employment-Population Ratios. If this relationship can be determined to be causal, then it

justifies improving the employment-population ratio to help reduce poverty rates. In other words, people need jobs to get out of poverty.

Since the **Pearson Product Moment Correlation Coefficient** measures the strength of the linear relationship between the two variables, then it is reasonable to find the equation of the line that best fits the data. This line is called the least squares regression line or the line of best fit. A regression line has been added to the graph for Employment-Population Ratio and Poverty Rates. Notice that there is a negative slope to the line. This corresponds to the sign of the correlation coefficient.

Scatterplot of Poverty Rate against Employment-Population Ratio
Poverty Rate = 35.8284-0.3567*x

Employment-Population Ratio:Poverty Rate: r = -0.6445, p = 0.0174
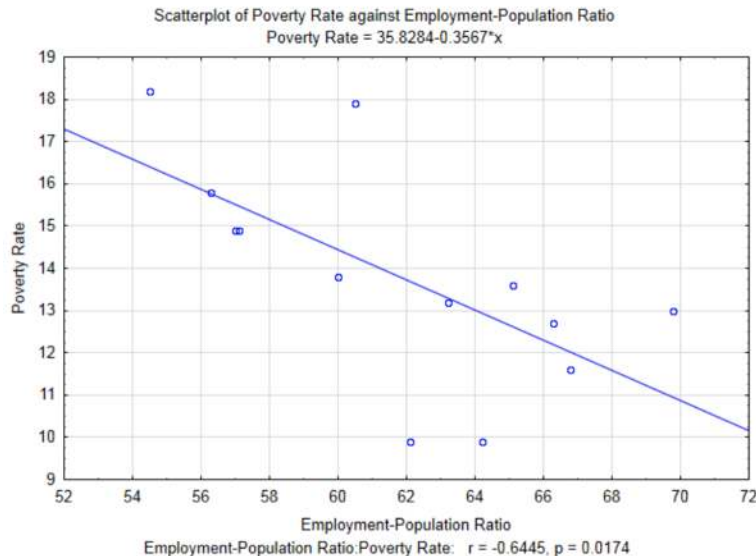
The equation of the line, as it appears in the subtitle of the graph is $y = 35.8284 - 0.3567x$ where x is the Employment-Population Ratio and y is the poverty rate. As an algebra student, you were taught that a linear equation can be written in the form of $y = mx + b$. In statistics, linear regression equations are written in the form $y = b + mx$ except that they traditionally are shown as $y' = a + bx$ where y' represents the y value predicted by the line, a represents the y intercept and b represents the slope.

To calculate the values of a and b, 5 other values are needed first. These are the correlation (r), the mean and standard deviation for x ($\bar{x}$ and s_x) and the mean and standard deviation for y ($\bar{y}$ and s_y). First find b using the formula: $b = r(\frac{s_y}{s_x})$. Next, substitute $\bar{y}$, $\bar{x}$, and b into the basic linear equation $\bar{y} = a + b\bar{x}$ and solve for a.

For this example, $r = -0.6445$, $barx = 61.76$, $s_x = 4.67$, $bary = 13.80$, and $s_y = 2.58$.

$b = r(\frac{s_y}{s_x})$

$b = -0.6445(\frac{2.58}{4.67}) = -0.3561$

$\bar{y} = a + b\bar{x}$

$1380 = a + -0.3561(61.76)$

$a = 35.79$

Therefore, the final regression equation is $y' = 35.79 - 0.3561x$. The difference between this equation and the one in the graph is the result of rounding errors used for these calculations.

The regression equation allows us to estimate the y value, but does not provide an indication of the accuracy of the estimate. In other words, what is the effect of the relationship between x and y on the y value?

Scatterplot of Poverty Rate against Employment-Population Ratio
Poverty Rate = 35.8284-0.3567*x

Employment-Population Ratio:Poverty Rate.   r = -0.6445, p = 0.0174
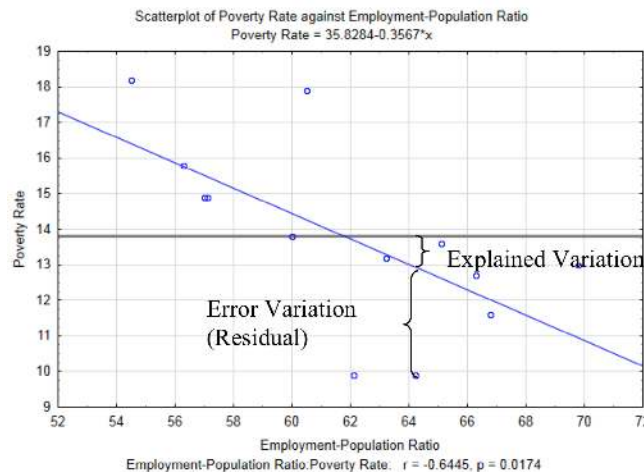
To determine the influence of the relationship between x and y begins with the idea that there is variation between the y value and the mean of all the y values ($\bar{y}$). This is something that you have seen with univariate quantitative data. There are two reasons why the y values are not equivalent to the mean. These are called explained variation and error variation. Explained variation is the variation that is a consequence of the relationship y has with x. In other words, y does not equal the mean of all the y values because the relationship shown by the regression line influences it. The error variation is the variation between an actual point and the y value predicted by the regression line that is a consequence of all the other factors that impact the response random variable. This vertical distance between each actual data point and the predicted y value (y') is called the residual. The explained variation and error variation is shown in the graph below. The horizontal line at 13.8 is the mean of all the y values.

The total variation is given by the sum of the squared distance each value is from the average y value. This is shown as $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

The explained variation is given by the sum of the squared distances the y value predicted by the regression equation (y') is from the average y value, $\bar{y}$. This is shown as

$$\sum_{i=1}^{n}(y'_i - \bar{y})^2. \tag{7.2}$$

The error variation is given by the sum of the squared distances the actual y data value is from the predicted y value (y'). This is shown as $\sum_{i=1}^{n}(y_i - y'_i)^2$.

The relationship between these can be shown with a word equation and an algebraic equation.

Total Variation = Explained Variation + Error Variation

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y'_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - y'_i)^2 \tag{7.3}$$

The primary reason for this discussion is to lead us to an understanding of the mathematical (though not necessarily causal) influence of the x variable on the y variable. Since this influence is the explained variation, then we can find the ratio of the explained variation to the total variation. We define this ratio as the coefficient of determination. The ratio is represented by r^2.

$$r^2 = \frac{\sum_{i=1}^{n}(y'_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The coefficient of determination is the square of the correlation coefficient. What it represents is the proportion of the variance of one variable that results from the mathematical influence of the variance of the other variable. The coefficient of determination will always be a value between 0 and 1, that is $0 \le r^2 \le 1$. While r^2 is presented in this way, it is often spoken of in terms of percent, which results by multiplying the r^2 value by 100.

In the scatter plot of poverty rate against employment-population ratio, the correlation is $r = -0.6445$, so $r^2 = 0.4153$. Therefore, we conclude that 41.53% of the influence on the variance in poverty rate is from the variance in the employment-population ratio.

The remaining influence that is considered error variation comes from some of the other items in the list of possible variables that could affect poverty.

There is no definitive scale for determining desirable levels for r^2. While values close to 1 show a strong mathematical relationship and values close to 0 show a weak relationship, the researcher must contemplate the actual meaning of the r^2 value in the context of their research.

Technology

Calculating correlation and regression equations by hand can be very tedious and subject to rounding errors. Consequently, technology is routinely employed to in regression analysis. The data that was used when comparing the Gini Coefficients to poverty rates will be used here.

Gini Coefficient	Poverty Rate
0.486	10.1
0.443	9.9
0.44	11.6
0.433	13
0.419	13.2
0.442	14.4
0.464	10.3

---

ti 84 Calculator

To enter the data, use Stat – Edit – Enter to get to the lists that were used in Chapter 4. Clear lists one and two by moving the cursor up to L1, pushing the clear button and then moving the cursor down. Do the same for L2.

Enter the Gini Coefficients into L1, the Poverty Rate into L2. They must remain paired in the same way they are in the table.

To determine the value of t, the p-value, the r and r2 values and the numeric values in the regression equation, use Stat – Tests – E: LinRegTTest. Enter the Xlist as L1 and the Ylist as L2. The alternate hypothesis is shown as β & $\rho: \neq 0$. Put cursor over Calculate and press enter.

The output is:

LinRegTTest
$y = a + bx$
$\beta \neq 0$ and $\rho \neq 0$
t = -1.912582657
p = 0.1140079665
df = 5
b = -52.72871602
$s = 1.479381344$(standard error)
$r^2 = 0.4224975727$
$r = -0.6499981406$

Microsoft's Excel contains an add-in that must be installed in order to complete the regression analysis. In more recent versions of Excel (2010), this addin can be installed by

- Select the file tab
- Select Options
- On the left side, select Add-Ins
- At the bottom, next to where it says Excel Add-ins, click on Go Check the first box, which says Analysis ToolPak then click ok. You may need your Excel disk at this point.

To do the actual Analysis:

- Select the data tab
- Select the data analysis option (near the top right side of the screen)
- Select Regression
- Fill in the spaces for the y and x data ranges.
- Click ok.

A new worksheet will be created that contains a summary output. Some of the numbers are shown in gray to help you know which numbers to look for. Notice how they correspond to the output from the TI 84 and the calculations done earlier in this chapter.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.649998141	(absolute value of the correlation coefficient r)
R Square	0.422497583	(r square)
Adjusted R	0.306997099	
Standard E	1.479381344	
Observatio	7	

ANOVA

	df	SS	MS	F	ignificance F
Regression	1	8.005726	8.005726	3.657972	0.114008
Residual	5	10.94285	2.188569		
Total	6	18.94857			

	Coefficients	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	35.340385	12.32832	2.866601	0.035134	3.649476	67.03129	3.649476	67.03129
Gini Coeffi	-52.728716	27.56938	-1.91258	0.114008	-123.598	18.14051	-123.598	18.14051

This page titled 7: Analysis of Bivariate Quantitative Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Peter Kaslik via source content that was edited to the style and standards of the LibreTexts platform.

7.E: Analysis of Bivariate Quantitative Data (Exercises)

Chapter 7 Homework

In the first problem, all calculations, except finding the correlation, should be done using the formulas and tables. For the remaining problems you may use either the calculator or Excel.

1. In the game of baseball the objective is to win games by scoring more runs than the opposing team. Runs can only be scored if someone gets on base. Traditionally, batting average (which is actually a proportion of hits to at bats) has been used as one of the primary measures of player success. An alternative is slugging percent which is the ratio of total number of bases reached during an at bat to the number of at bats. A walk or single counts as one base, a double counts as two bases, etc. The table below contains the batting average, slugging percentage, and runs scored from 10 Major League Baseball teams randomly selected from the 2012 and 2013 seasons.(http://www.fangraphs.com, 12-12-13)

Team Batting Average	Team Slugging Percentage	Team Runs Scored
0.242	0.380	614
0.231	0.335	513
0.283	0.434	796
0.240	0.375	610
0.252	0.398	640
0.268	0.422	726
0.245	0.407	716
0.260	0.390	701
0.240	0.394	697
0.255	0.422	748

a. Make a scatter plot of team batting average and team runs scored. Label the graph completely.
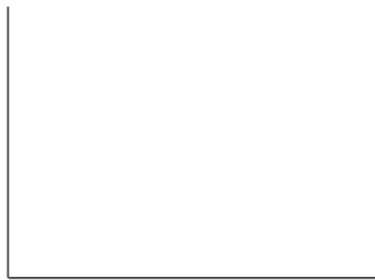
b. Use your calculator to find the mean and standard deviation for batting average and runs scored. The correlation between these is 0.805.

Mean batting average _____ Standard deviation for batting average _____

Mean runs scored _____ Standard deviation for runs scored_____

c. Use the appropriate t test statistic to determine if the correlation is significant at the 0.05 level of significance. Show the formula, substitution and the results in a complete concluding sentence.

Formula Substitution

Concluding sentence:

d. Find the equation of the regression line.

$$b = r(\frac{s_y}{s_x})$$

$$\bar{y} = a + b\bar{x}$$

Regression equation:

Draw this line on your scatter plot. (Hint: pick two different x values, one near either side of the x-axis, substitute into the regression equation to find y. Then plot the two (x, y) ordered pairs that you produced. This is how you learned to graph in Algebra using a table of values).

e. What is the r^2 value and what does it mean?

f. Predict the number of runs scored for a team with a batting average of 0.250.

g. Repeat this entire problem for slugging percent and runs scored, only this time use the LinRegTTest function on your calculator.
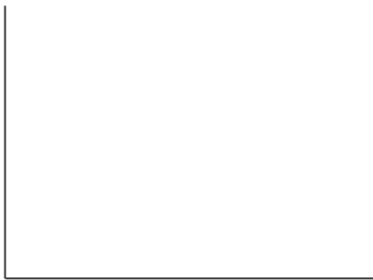
Correlation _____

Hypothesis test concluding sentence

Regression equation _____

Coefficient of determination (R^2) _____

Predict the number of runs scored for a team with a slugging percentage of 0.400. _____

Compare and contrast the results from the analysis of batting average and slugging percentage and their relationship to runs scored.

2. In an ideal society, crime would seldom happen and consequently the population's financial resources could be spent on other things that benefit society. The primary categories for state spending are k-12 education, higher education, public assistance, Medicaid, transportation and corrections. Many of us in the field of education believe that it is critical for the country and holds the possibility of reducing both crime and public assistance. Is there a significant correlation between the percent of state budgets spent on k-12 and higher education and the percent spent on public assistance? Is there a significant correlation between the percent of state budgets spent on education and corrections? Data is from 2011. (www.nasbo.org/sites/default/f...20Report_1.pdf 12-12-13.)

Percent of State Budget		
Education	Public Assistance	Corrections
24.4	1.6	2.6
33.9	2.2	2.9
22.7	2.1	3.2
31	0.3	3
29.1	0.6	2.3
30	0.3	4.4
30.8	0.5	2.8
33.9	0.2	3.5
33.1	0.2	3.3
27.5	4.7	4.5

a. Make a scatter plot, use your calculator to test the hypothesis that there is a correlation between education spending and public assistance spending. Show calculator outputs including the correlation, r^2 value and equation of the regression line. Write a statistical conclusion then interpret the results. Use a level of significance of 0.05.

Correlation _____

Coefficient of determination (r^2 value) _____

Regression equation _____

What does x represent in this equation? _____

What does y represent in this equation? _____

Hypothesis test concluding sentence:

b. Make a scatter plot, use your calculator to test the hypothesis that there is a correlation between education spending and corrections spending. Show calculator outputs including the correlation, r^2 value and equation of the regression line. Write a statistical conclusion then interpret the results. Use a level of significance of 0.05.

Correlation _____

Coefficient of determination (r^2 value) _____

Regression equation _____

What does x represent in this equation? _____

What does y represent in this equation? _____

Hypothesis test concluding sentence:

3. Is there a correlation between the population of a state and the median income in the state? (Data from http://www.city-data.com/ 12-12-13.)

State Population (millions)	Median income ($)
2.7	55764
2.8	49444
2	43569
7.9	61090
9.6	48448
1.3	46405
11.5	46563
2.7	54065
4.5	43362

Make a scatter plot, use your calculator to test the hypothesis that there is a correlation between population and median income. Show calculator outputs including the correlation, r2 value and equation of the regression line. Write a statistical conclusion then interpret the results. Use a level of significance of 0.05.

Correlation _____

Coefficient of determination (r^2 value) _____

Regression equation _____

What does x represent in this equation? _____

What does y represent in this equation? _____

Hypothesis test concluding sentence:

4. One theory about the benefit of large cities is that they serve as a hub for creativity due to the frequent interactions between people. One measure of creative problem solving is the number of patents that are granted. Is there a correlation between the size of a metropolitan or micropolitan area and the number of patents that were granted to someone in that area? (www.uspto.gov/web/offices/ac/...allcbsa_gd.htm andwww.census.gov/popfinder/ (12-12-13))

Metropolitan Area	Population	Total Patents 2000-2011
Las Vegas-Paradise, NV	806,923	2160
Fresno, CA	494,665	468
Decatur, AL	55,683	130
Guayama, PR	22,691	4
Taylorville, IL	11,246	97
Harriman, TN	6,350	53
Kapaa, HI	10,699	32
Minot, ND	40,888	19
Lewisburg, TN	11,100	10

a. Make a scatter plot, use your calculator to test the hypothesis that there is a correlation between population and total patents 2000-2011. Show calculator outputs including the correlation, r^2 value and equation of the regression line. Write a statistical conclusion then interpret the results. Use a level of significance of 0.05.

Correlation _____

Coefficient of determination (r^2 value) _____

Regression equation _____

What does x represent in this equation? _____

What does y represent in this equation? _____

Hypothesis test concluding sentence:

b. There are two outliers in this data. Do you think they have too great an influence on the correlation and therefore should be removed or do you think they are relevant and should be kept with the data?

c. Use the regression line to predict the number of patents for a city with 60,000 people.

5. Why Statistical Reasoning is Important for Anatomy and Physiology Students and Professionals In collaboration with Barry Putman, Professor of Biology, Natural Science Coordinator, JBLM

This topic is discussed in the following Pierce College Course: Biol 241

> **briefing 8.1**
>
> Near Point of Accommodation (NPA) is the nearest point at which the eyes can comfortably focus. In the lab conducted in the anatomy and physiology class, students will hold a meter stick against their forehead, close their left eye and with their right eye they will focus on a small ruler held against the meter stick. With the ruler starting at arm's length they will slowly move it toward their eye. When they reach the point where the ruler has the greatest focus (NPA), a partner will record the distance, in centimeters, from their eye.
>
> Since people often need glasses later in life, it would be reasonable to determine if there is a correlation between a person's age and their NPA. Consequently, students in the study record both their age and NPA.

a. Of the two variables, NPA and age, which should be the explanatory variable? Why?

b. Of the two variables, NPA and age, which should be the response variable? Why?

c. There were 103 data values made available for this problem. This number will be reduced using a random process to save you time. If a systematic sampling method is used with every 10th value selected, what are the 10 or 11 numbers that would be

selected if the calculator is seeded with a 31?

, , , , , , , , , ,

The table below contains the data.

Age	26	28	30	26	36	19	20	20	27	25	24
NPA	31	13	36	22	34	8	8	10	24	14	11

d. Make a scatter plot. Write a complete sentence explaining your interpretation of the graph.

e. Use your calculator to find the sample correlation.

f. Write and test the hypotheses to determine if there is a significant correlation in the population. Use a 0.05 level of significance. Write a concluding sentence.

g. What type error could have been made?

h. What do you conclude about the relationship between age and NPA?

8: Chi Square

In chapter 5, the inferential theory for categorical data was developed based upon the binomial distribution. Recall that the binomial distribution shows the probability of the possible number of successes in a sample of size n when there were only two possible independent outcomes, success and failure. What happens if there are more than two possible outcomes however?

Consider the following three questions.

1. Does the TI 84 calculator generate equal numbers of 0-9 when using the random integer generator?
2. Doing something about climate change has been a challenge for humanity. The website Edge.Org had one proposal put forth by Lee Smolin, a physicist with the Perimeter Institute and author of book Time Reborn. (www.edge.org/conversation/del...ooperation/#rc Nov 30, 2013.) The essence of the proposal is that a carbon tax should be placed on all carbon that is used but instead of the money going to the government it goes to individual climate retirement accounts. Each person would have such an account. Each account would have two categories of possible investments that an individual could choose. Category A investments would be in things that will mitigate climate change (e.g. solar, wind, etc). Category B investments would be in things that might do well if climate change does not happen (e.g. utilities that burn coal, coastal real estate developments and car companies that do not produce fuel efficient or electric cars). Is there a correlation between a person's opinion about climate change and their choice of investment?
3. Hurricanes are classified as category 1,2,3,4,5. Is the distribution of hurricanes in the years 1951- 2000 different than it was in 1901-1950?

Before an analysis can be done, it is necessary to understand the type of data that is gathered for each of these questions.

In question 1, the data that will be gathered are the numbers 0 though 9. While numbers are typically considered quantitative, in this case we simply want to know if the calculator produces each specific number. Therefore, this is actually about the frequency with which these numbers are produced. If the process used by the calculator is sufficiently random, then the frequencies for all the numbers should be equal if a large enough sample is taken. So, in spite of appearing to be quantitative data, this is actually categorical data, with 10 different categories and the data being that a number was selected.

In question 2, imagine a two-question survey in which people are asked:

1. Do you believe climate change is happening because humans have been using carbon sources that lead to an increase in greenhouse gases? Yes No
2. Which of the following most closely represents the choice you would make for your individual climate retirement account investments? Category A Category B

For this question, there is one population. Each person that takes the survey would provide two answers. The objective is to determine if there is a correlation between their climate change opinion and their investment choice. An alternate way of saying this is that the two variables are either independent of each other, which means that one response does not affect the other, or they are not independent which means that climate change opinion and investment strategy are related.

In question 3, there are two populations. The first population is hurricanes in 1901-50 and the second population is hurricanes in 1951-2000. There are 5 categories of hurricanes and the goal is to see if the distribution of hurricanes in these categories is the same or different.

The problems fit one of the following classes of problems, in order: goodness of fit, test for independence, and test for homogeneity. The use of these problems and their hypotheses are shown below.

1. Goodness of Fit

   The goodness of fit test is used when a categorical random variable with more than two levels has an expected distribution.

   H_0: The distribution is the same as expected

   H_1: The distribution is different than expected
2. Test for Independence

   The test for independence is used when there are two categorical random variables for the same unit (or person) and the

objective is to determine a correlation between them.

H_0: The two random variables are independent (no correlation)

H_1: The two random variables are not independent (correlation)

If the data are significant, than knowledge of the value of one of the random variables increases the probability of knowing the value of the other random variable compared to chance.

3. Test for Homogeneity

The test for homogeneity is used when there are samples taken from two (or more) populations with the objective of determining if the distribution of one random variable is similar or different in the two populations.

H_0: The two populations are homogeneous

H_1: The two populations are not homogeneous

Since all of the problems have data that can be counted exactly one time, the strategy is to determine how the distribution of counts differs from the expected distribution. The analysis of all these problems uses the same test statistic formula called chi^2 (Chi Square).

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
(8.1)

The distribution that is used for testing the hypotheses is the set of chi^2 distributions. These distributions are positively skewed. They cannot be negative. Each distribution is based on the number of degrees of freedom. Unlike the t distributions in which degrees of freedom were based on the sample size, in the case of chi^2, the degrees of freedom are based on the number of levels of the random variable(s).

The following distributions show 10,000 samples of size n = 100 in which the chi^2 test statistics calculated and graphed. The numbers of degrees of freedom in these four graphs are 1,2,5, and 9.

Distribution of Chi Square
Original Population: {0,1}, DF = 1
Sample size: n = 100

df = 1

Distribution of Chi Square
Original Population: {0,1,2}, DF = 2
Sample size: n = 100

df = 2

Distribution of Chi Square
Original Population: {0,1,2,3,4,5}, DF = 5
Sample size: n = 100

df = 5

Distribution of Chi Square
Original Population: {0,1,2,3,4,5,6,7,8,9}, DF = 9
Sample size: n = 100

df = 9
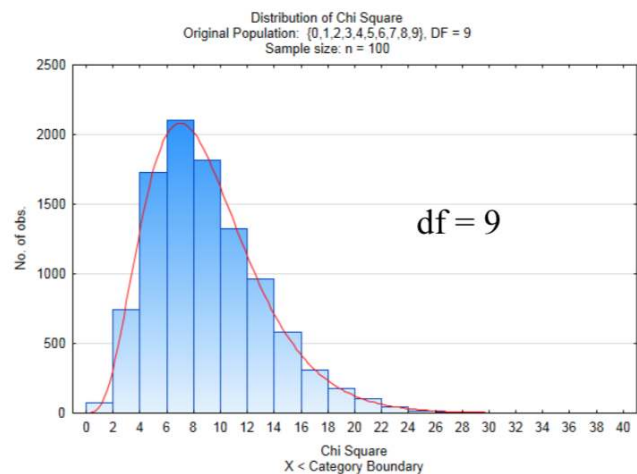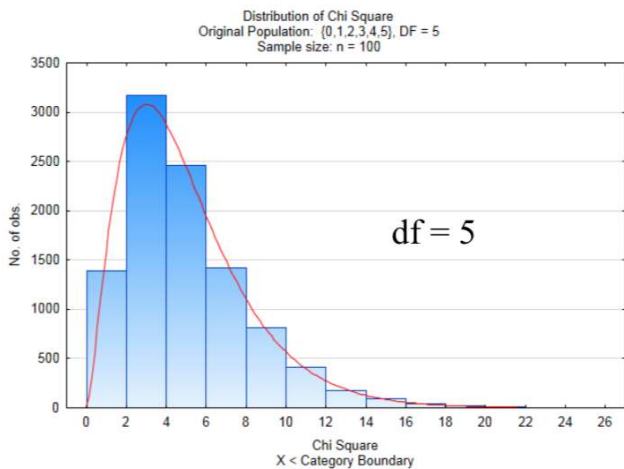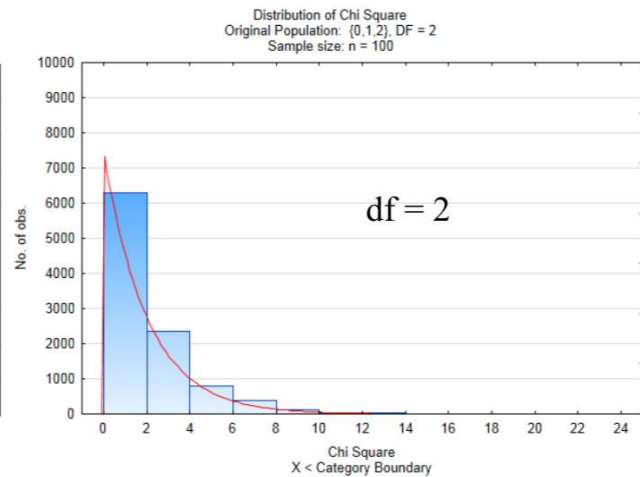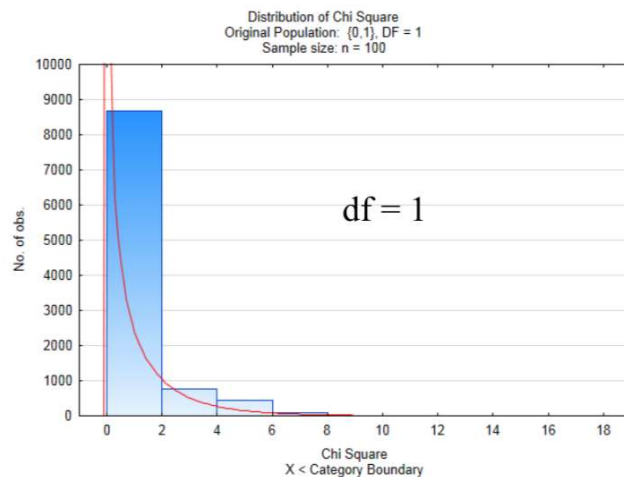
Notice how the Chi Square distribution becomes less skewed and is approaching a normal distribution as the number of degrees of freedom increase. An increase in the number of degrees of freedom corresponds to an increase in the number of levels of the explanatory factor. The way in which degrees of freedom are found is different for the goodness of fit test compared to the test for independence and test for homogeneity. Each method will be explained in turn.

Goodness of Fit Test

1. Does the TI 84 calculator generate equal numbers of 0-9 when using the random integer generator?

In this experiment, 12 numbers between 1 and 100 were randomly generated by the TI 84 calculator. These 12 numbers were used as seed values. After seeding the calculator with each number, 10 new numbers between 0 and 9 were randomly generated using the randint function on the calculator. Thus, a total of 120 numbers between 0 and 9 were produced. The frequency of these numbers is shown in the table below.

0	1	2	3	4	5	6	7	8	9
15	11	12	14	10	14	10	11	14	9

The hypotheses to be tested are:

H_0: The observed cell frequency equals the expected cell frequency for all cells

H_1: The observed cell frequency does not equal the expected cell frequency for at least one cell. Use a 0.05 level of significance

This can be represented symbolically as

$H_0: o_1 = \epsilon_1$ for all cells

$H_1: o_1 \neq \epsilon_1$ for at least one cell

where o is the lower case Greek letter omicron that represents the observed cell frequency in the underlying population and ε is the lower case Greek letter epsilon that represents the expected cell frequency. The expected cell frequency should always be 5 or higher. If it isn't, cells should be regrouped.

The table above shows the observed frequencies, but what are the expected frequencies? In theory, if the process is truly random, then each number would occur with the same frequency if the sampling were to be done a very large number of times. If this is the case, then in a sample of size 120, with 10 possible alternatives, the expected number of frequencies for each alternative should be 12. From the table, we see that most frequencies are not 12, but what is needed is a way to determine if the amount of variation that exists is enough to suggest that the observed frequencies do not equal the expected frequencies. Such a conclusion would imply the calculator does not produce a truly random set of numbers. The strategy is to find χ^2 and then use the appropriate χ^2 distribution to find the p-value. One way to find $\chi^2 = \sum \dfrac{(O - E)^2}{E}$ is with a table.

Observed	Expected	O - E	$(O - E)^2$	$\dfrac{(O - E)^2}{E}$
15	12	3	9	$\dfrac{9}{12}$
11	12	-1	1	$\dfrac{1}{12}$
12	12	0	0	$\dfrac{0}{12}$
14	12	2	4	$\dfrac{4}{12}$
10	12	-2	4	$\dfrac{4}{12}$
14	12	2	4	$\dfrac{4}{12}$
10	12	-2	4	$\dfrac{4}{12}$
11	12	-1	1	$\dfrac{1}{12}$
14	12	2	4	$\dfrac{4}{12}$
9	12	-3	9	$\dfrac{9}{12}$
				$\chi^2 = \dfrac{40}{12} = 3.33$

If r represents the number of rows, then the number of degrees of freedom in a goodness of fit test is:

$$df = r - 1.$$

For this Goodness of fit test, there are 10 rows of data. Consequently there are 9 degrees of freedom.

The p-value for χ^2 can be found using the table of the Chi Square Distributions at the end of this chapter or your calculator.

The Chi-Square Distributions can also be used to find the p-value. Using the table below, find the degrees of freedom in the left column, locate the χ^2 value in the row, then move to the row that shows the area to the right and use an inequality sign to show the p-value. If the p-value is greater than α, then use the greater than symbol. If it is less than α, use the less than symbol, but in either case, use as much precision as possible. For example, if α is 0.05 but the area to the right is less than 0.025, then p < 0.025 is preferred over p < 0.05.

In this example, $\chi^2 = 3.33$, there are 9 degrees of freedom, so the p-value > 0.9.

Area Left	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
Area Right	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589

$$\chi^2 = 3.33$$

Using χ^2 cdf (low, high, df) in the TI 84 calculator results in χ^2 cdf (3.33, 1E99,9) = 0.9496.

Since this p-value is clearly higher than 0.05, the conclusion can be written:

At the 5% level of significance, the observed cell values are not significantly different than the expected cell values (χ^2 = 3.33, p = 0.9496, df=9). The TI84 calculator appears to produce a good set of random integers.

In the case of the calculator, if it is random in generating numbers, we would expect the same number of values in each category. That is, we would expect to get the same number of 0s, 1s, 2s, etc. Since the sample consisted of 120 trials with 10 possibilities for each outcome, the expected value is 12 because 120 divided by 10 is 12. But what happens if the expected outcome is not the same in all cases?

In the fall of 2013, our college was made up of 54% Caucasian, 14% Hispanic/Latino, 11% African American, 10% Asian/Pacific Islander, 1% Native American, 3% international, and 7% other. If we wanted to determine if the racial/ethnic distribution of statistics students is different than of the entire school, we could take a survey of statistics students to obtain the observed data. The table below contains hypothetical observed data. Since there are 300 students in the sample and based on college enrollment, 54% of the student body is white, then the expected number of students in the class who are white is found by multiplying 300 times 0.54. The same approach is taken for each race. This is shown in the table. Notice the total in the expected column is the same as in the observed column.

Race/Ethnicity	Observed	Expected
Caucasian/white (54%)	154	0.54(300) = 162
Hispanic/Latino (14%)	48	0.14(300) = 42
African American/Black (11%)	36	0.11(300) = 33
Asian/Pacific Islander (10%)	35	0.10(300) = 30
Native American (1%)	6	0.01(300) = 3
International (3%)	9	0.03(300) = 9
Other (7%)	12	0.07(300) = 21
Total	300	Total 300

The remainder of the goodness of fit test is done the same as with the calculator example and will not be demonstrated here.

Chi Square Test for Independence

The Chi Square Test for Independence is used when a researcher wants to determine a relationship between two categorical random variables collected on the same unit (or person). Sample questions include:

1. Is there a relationship between a person's religious affiliation and their political party preference?
2. Is there a relationship between a person's willingness to eat genetically engineered food and their willingness to use genetically engineered medicine?
3. Is there a relationship between the field of study for a college graduate and their ability to think critically?
4. Is there a relationship between the quality of sleep a person gets and their attitude during the next day?

As an example, we will learn the mechanics of the test for independence using the hypothetical example of responses to the two questions about climate change and investments.

1. Do you believe climate change is happening because humans have been using carbon sources that lead to an increase in greenhouse gases? Yes No
2. Which of the following most closely represents the choice you would make for your individual climate retirement account investments? Category A Category B

Category A – solar, wind Category B – Coal, ocean side development

> H_0: The two random variables are independent (no correlation)
> H_1: The two random variables are not independent (correlation)

This can also be represented symbolically as

$H_0 : o_1 = \epsilon_1$ for all cells
$H_1 : o_1 \neq \epsilon_1$ for at least one cell
where o is the lower case Greek letter omicron that represents the observed cell frequency in the underlying population and ϵ is the lower case Greek letter epsilon that represents the expected cell frequency. The expected cell frequency should always be 5 or higher. If it isn't, cells should be regrouped.

Use a level of significance of 0.05.

Because this will be done with pretend data, it will be useful to do it twice, producing opposite conclusions each time.

The data will be presented in a 2 x 2 contingency table.

Version 1 Observed	Yes - humans contribute to clime change	No - humans do not contribute to climate change	Totals
Category A Investments (wind, solar)	56	54	
Category B Investments (coal, ocean shore developments)	47	43	
Total			

The test for independence uses the same formula as the goodness of fit test. $\chi^2 = \sum \dfrac{(O-E)^2}{E}$. Unlike that test however, there is no clear indication of what the expected values are. Instead they must be calculated, which is a four-step process.

Step 1, Find the row and column totals and the grand total.

Version 1 Observed	Yes - humans contribute to clime change	No - humans do not contribute to climate change	Totals
Category A Investments (wind, solar)	56	54	110
Category B Investments (coal, ocean shore developments)	47	43	90
Total	103	97	200

Step 2. Create a new table for the expected values. The reasoning process for calculating the expected values is to first consider the proportion of all the values that fall in each column. In the first column there are 103 values out of 200 which is $\cdot = 0.515$. In the second column there are 97 out of 200 values (0.485). Since 51.5% of the values are in the first column, then it would be expected that 51.5% of the first row's values would also be in the first column. Thus, 0.515(110) gives an expected value of 56.65. Likewise, 0.485(90) will produce the expected value of 43.65 for the last cell. As a formula, this can be expressed as

$$\frac{Column\ Total}{Grand\ Total} \cdot Row\ Total \tag{8.2}$$

Version 1 Observed	Yes - humans contribute to clime change	No - humans do not contribute to climate change	Totals
Category A Investments (wind, solar)	$\frac{103}{200} \cdot 110 = 56.65$	$\frac{97}{200} \cdot 110 = 53.35$	110
Category B Investments (coal, ocean shore developments)	$\frac{103}{200} \cdot 90 = 46.35$	$\frac{97}{200} \cdot 110 = 43.65$	90
Total	103	97	200

Step 3. Use a table similar to the one used in the Goodness of Fit test to calculate Chi Square.

Observed	Expected	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
56	56.65	-0.65	0.4225	0.0075
54	53.35	0.65	0.4225	0.0079
47	46.35	0.65	0.4225	0.0091
43	43.65	-0.65	0.4225	0.0097
				$\chi^2 = 0.0342$

Step 4. Determine the Degrees of Freedom and find the p-value

If R is the number of Rows in the contingency Table and C is the number of columns in the contingency table, then the number of degrees of freedom for the test for independence is found as

$$df = (R - 1)(C - 1).$$

For a 2 x 2 contingency table such as in this problem, there is only 1 degree of freedom because (2-1)(2-1) = 1.

The p-value for χ^2 can be found using the table or your calculator.

In the table we locate 0.034 in the row with 1 degree of freedom, then move up to the row for the area to the right. Since the area to the right is greater than 0.05, but more specifically it is greater than 0.1, the p-value is written as p > 0.1.

Area Left	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
Area Right	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
df										
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879

$$\chi^2 = 0.034$$

On your calculator, use χ^2 cdf (low, high, df) . In this case, χ^2 cdf (0.0342, 1E99, 1) = 0.853.

Since the data are not significant, we conclude that people's investment strategy is independent of their opinion about human contributions to climate change.

Version 2 of this problem uses the following contingency table.

Version 2 Observed	Yes - humans contribute to clime change	No - humans do not contribute to climate change	Totals
Category A Investments (wind, solar)	80	30	
Category B Investments (coal, ocean shore developments)	30	60	
Total			

This time, the entire problem will be calculated using the TI 84 calculator instead of building the tables that were used in Version 1.

Step 1. Matrix
Step 2. Make 1:[A] into a 2 x 2 matrix by selecting Edit Enter then modify the R x C as necessary. Step 3. Enter the frequencies as they are shown in the table.
Step 4. STAT TESTS χ^2 – Test
Observed:[A]
Expected:[B] (you do not need to create the Expected matrix, the calculator will for you.)
Select Calculate to see the results:
χ^2 = 31.03764922
p=2.5307155E-8
df=1

In this case, the data are significant. This means that there is a correlation between each person's opinion about human contributions to climate change and their choice of investments. Remember that correlation is not causation.

Chi Square Test for Homogeneity

The third and final problem is about the classification of hurricanes in two different decades, 1901-50 and 1951-2000. One theory about climate change is that hurricanes could get worse. be worked using tables.

Hurricanes are classified by the Saffir-Simpson Hurricane Wind Scale.2

Category 1 Sustained Winds 74-95 mph
Category 2 Sustained Winds 96-110 mph
Category 3 Sustained Winds 111-129 mph
Category 4 Sustained Winds 130-156 mph
Category 5 Sustained Winds 157 or higher.

Category 3, 4, and 5 hurricanes are considered major.

This problem will

The population of interest is the distribution of hurricanes for the prevailing climate conditions at the time. The hypotheses being tested are

H_0: The distributions are homogeneous

H_1: The distributions are not homogeneous

This can also be represented symbolically as

$H_0 : o_1 = \epsilon_1$ for all cells
$H_1 : o_1 \neq \epsilon_1$ for at least one cell

where o is the lower case Greek letter omicron that represents the observed cell frequency in the underlying population and ϵ is the lower case Greek letter epsilon that represents the expected cell frequency. The expected cell frequency should always be 5 or higher. If it isn't, cells should be regrouped.

A 5 x 2 contingency table will be used to show the frequencies that were observed. The expected frequencies were calculated in the same way as in the test of independence. (http://www.nhc.noaa.gov/pastdec.shtml viewed 12/7/13)

Observed	1901 - 1950	1951 - 2000	Totals
Category 1	37	29	66
Category 2	24	15	39
Category 3	26	21	47
Category 4	7	5	12
Category 5	1	2	3
Totals	95	72	167

Expected	1901 - 1950	1951 - 2000	Totals
Category 1	37.54	28.46	66
Category 2	22.19	16.81	39
Category 3	26.74	20.26	47
Category 4	6.83	5.17	12
Category 5	1.71	1.29	3
Totals	95	72	167

Notice that the expected cell frequencies for category 5 hurricanes are less than 5, therefore it will be necessary for us to redo this problem by combining groups. Group 5 will be combined with group 4 and the modified tables will be provided.

Observed	1901 - 1950	1951 - 2000	Total
Category 1	37	29	66
Category 2	24	15	39
Category 3	26	21	47
Category 4 & 5	8	7	15
Total	95	72	167

Observed	1901 - 1950	1951 - 2000	Total
Category 1	37.54	28.46	66
Category 2	22.19	16.81	39
Category 3	26.74	20.26	47
Category 4 & 5	8.53	6.47	15
Total	95	72	167

	Observed	Expected	$O - E$	$(O - E)^2$	$\dfrac{(O - E)^2}{E}$

1901 - 50					
Category 1	37	37.54	-0.54	0.30	0.008
Category 2	24	22.19	1.81	3.29	0.148
Category 3	26	26.74	-0.74	0.54	0.020
Category 4 & 5	8	8.53	-0.53	0.28	0.033
1951 - 2000					
Category 1	29	28.46	0.54	0.30	0.010
Category 2	15	16.81	-1.81	3.29	0.196
Category 3	21	20.26	0.74	0.54	0.027
Category 4 & 5	7	6.47	0.53	0.28	0.044
					$\chi^2 = 0.487$

If R is the number of Rows in the contingency Table and C is the number of columns in the contingency table, then the number of degrees of freedom for the test for homogeneity is found as

$$df = (R-1)(C-1).$$

For a 4×2 contingency table such as in this problem, there are 3 degrees of freedom because (4-1)(2-1) = 3 degrees of freedom.

$$\chi^2 = 0.487$$

The table shows the p-value is less than 0.05. The calculator confirms this because χ^2 cdf (0.486, 1E99, 3) = 0.9218. Consequently the conclusion is that there is not a significant difference between the distribution of hurricanes in 1951-2000 and 1901-50.

Distinguishing between the use of the test of independence and homogeneity

While the mathematics behind both the test of independence and the test of homogeneity are the same, the intent behind their usage and interpretation of the results is different.

The test for independence is used when two random variables, both of which are considered to be response variables, are determined for each unit. The test for homogeneity is used when one of the random variables is the explanatory variable and subjects are selected based on their level of this variable. The other random variable is the response variable.

The determination of which test to used is established by the sampling approach. If two populations are clearly defined beforehand and a random selection is made from each population, then the populations will be compared using the test of homogeneity. If no effort is made to distinguish populations beforehand, and a random selection is made from this population and then the values of the two random variables are determined, the test of independence is appropriate.

An example may clarify the subtle difference between the two tests. Consider one random variable to be a person's preference between running and swimming for exercise and the other random variable to be a person's preference between watching TV or reading a book. If the researcher randomly selects some runners and some swimmers and asks each group about their preference for TV or reading a book, the test for homogeneity would be appropriate. On the other hand, if the researcher survey's randomly selected people and asks if they prefer running or swimming and if they prefer TV or reading, then the objective will be to determine if there is a correlation between these two random variables by using the test of independence.

Area Left	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
Area Right	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
df										
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.287	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3,490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21/064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5,812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.365	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.398	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766

50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170
110	75.550	78.458	82.867	86.792	91.471	129.385	135.480	140.916	147.414	151.948

8.E: Chi Square (Exercises)

Q8.1

For each of the following questions, determine the appropriate test that should be used. Pick from the following three tests.

A. Goodness of Fit
B. Test for Independence
C. Test for Homogeneity

a. The tutor center maintains a list of student who use their services. These students are classified as drop-in students or appointment students. At the end of the term, the director of the tutor center randomly select students from each of the groups then looks up the grade they received in the class for which they were being tutored. The objective is to determine if there is a difference in the distribution of grades for the two groups. Grades are classified as A,B,C,F.
b. Historically, a teacher found that 33% of the students in a class earned an A, 47% a B, 15% a C, and 5% a D or F. After modifying the way she teaches, she wants to know if her most recent class of students was consistent with past students.
c. Students are given a math assessment and a musical assessment with the objective of determining if there is a correlation between mathematical ability and musical ability.
d. Quantitative data are grouped by the number of standard deviations they are from the mean (e.g. z intervals of [-3, -2), [-2, -1),... [2, 3)). The objective is to determine if the distribution is normal and is based on the probability that a value would fall within each of those ranges.
e. A researcher with the Department of Social and Health Services reviewed records of families who were receiving government assistance two years early. The researcher recorded if it was a one-parent household or a two-parent household. The researcher also recorded if the family was currently receiving government assistance. The objective was to determine if there is a correlation between then number of the parents in the household and whether the household was still receiving government assistance.
f. A researcher with the Department of Social and Health Services wants to know if the number of parents in a household affects the length of time a family receives government assistance. The researcher identifies one-parent families and two-parent families then randomly selects from each of these two groups to determine the number of years in which they receive government assistance. A comparison will be made between the distribution of one-parent families and two-parent families.

Q8.2

For each of the following problems, identify the test that should be done, then write the hypotheses, conduct the test to find chi square and the p-value, and then write a concluding sentence.

1. Bunko is a dice game that serves as the motivation for a group of people to get together for an evening of socializing and eating. One regular Bunko player called it a mindless dice game, because it doesn't require much thinking and players can talk (or eat!) while playing. A normal game of Bunko involves 12 players, but other multiples of 4 can work nicely. If there are three tables, with the head table being number 1, then after each round of play, winning players move up one table with the goal of being at the head table (the one closest to the food!). The losers from the head table go to table 3. Three dice are used at each table. On each turn a player will roll all three dice. The first round the objective is to get ones, the second round the objective is to get twos, etc. If one of the desired numbers is obtained, the player gets a point. If two of the desired numbers are obtained on the same roll, the player gets 2 points. If all three of the dice are the desired number, the player yells bunko and gets 21 points. If none of the dice show the desired number, the dice are passed to the next person. When the head table reaches 21 points, the round is over for everyone.

   As with any game of chance, there is an expected probability distribution. The expected distribution for the probability of having 0,1,2 or 3 successes can be found using the binomial distribution. Complete the probability distribution table. Refer to Chapter 4 if you can't remember the process.

$X = x$	0	1	2	3
$P(X = x)$				

Three dice were rolled 158 times and the number of ones was recorded for each turn. If the dice are fair, the sample distribution should be a good fit with the expected distribution. The sample data is shown in the table below.

$X = x$	0	1	2	3
Sample Results	100	44	13	1

Which test is appropriate for this problem?

Conduct the test then write a concluding sentence.

2. In Major League Soccer, is there a correlation between the number of shots a forward attempts and the number of goals he scores? A systematic random sample was taken from the 2013 MLS season results for all players classified as forward. The number of shots the player took was categorized as high (20 or more) and low (less than 20). The number of goals he scored was categorized as high (5 or more) and low (less than 5). The contingency table shows the results of the sample.

	High Shots	Low Shots	
High Goals	13	0	
Low Goals	15	21	

Which test is appropriate?

Complete the test using tables. Test the hypothesis at the 0.1 level of significance. Write a complete concluding sentence.

3. Lower back pain can be treated with a variety of approaches including using drugs and non-drug therapies. Data from a clinic that specializes in pain management was used to determine if there was a difference in the change in pain level for the patients being treated with a combination of drugs (local anesthetic, anti-inflammatory and a muscle relaxer) and those receiving physical therapy (lumbar traction, heat and ultrasound therapy and transcutaneous electrical nerve stimulation). Pain levels, on a scale of $1 - 5$, were determined during the initial visit. The change in pain level was assessed at the 4- week period. If pain improved by 4 or 5 levels it was classified as substantial improvement. If pain improved by 1,2 or 3 levels it was classified as moderate improvement. If pain was unchanged or got worse, it was classified as no improvement. The table below shows the changes. Use this data to determine if there is a difference in pain reduction using drugs vs non-drug therapy. (data from unpublished student statistics class project)

Observed	Drug	Non-Drug	
Substantial improvement	9	6	
moderately improvement	22	23	
no improvement	9	11	

Which test is appropriate?

Complete the test using tables. Test the hypothesis at the 0.1 level of significance. Write a complete concluding sentence.

Expected	Drug	Non-Drug	
Substantial improvement			
moderately improvement			
no improvement			

Observed	Expected	$O - E$	$(O - E)^2$	$\dfrac{(O - E)^2}{E}$

				$\chi^2 =$

4. Nationwide, for Native American tribal members with college degrees, 37% are associate degrees, 48% are bachelor degrees and 15% are Masters or PhDs. The distribution of degrees in one of the Puget Sound area tribes is 36 associate degrees, 22 bachelor degrees and 7 masters or PhDs. Is the distribution of degrees in the Puget Sound area tribe different than the national distribution?(data from unpublished student statistics class project)

Which test is appropriate?

Complete the test using tables or calculator. Test the hypothesis at the 0.1 level of significance. Write a complete concluding sentence. Show work (either tables or calculator inputs).

5. Why Statistical Reasoning Is Important for a Criminal Justice Student and Professional Developed in collaboration with Teresa Carlo, Professor of Criminal Justice
This topic is discussed in CJ 200 and others (Conflict view of Injustice).

The table below shows the racial distribution for Washington State. The data is from the WA State Government, Office of Financial Management. These percentages include those of Hispanic origin. (http://www.ofm.wa.gov/pop/census2010/data.asp)

White	Asian	Black	Native	Other
77.3%	7.2%	3.6%	1.5%	10.4%

In theory, the racial distribution of prisoners in WA state prisons should be consistent with this distribution. To determine if this is the case, a sample of prisoners can be taken. The random variable that will be measured is race. The hypotheses to be tested are:

H_0: The racial distribution in WA prisons is the same as the racial distribution of the WA population

H_1: The racial distribution in WA prisons is not the same as the racial distribution of the WA population.

Use a 5% level of significance. If the data are not significant then we will consider that society and justice are blind to race. If the data are significant, then we will seek a solution to this injustice.

There are 12 prison facilities in WA of which eight are major prisons and four are minimum- security. There is the possibility that the racial distribution varies based on location and security level and because of this random samples will be taken from each prison.

a. What type of sampling method is being used? _____

b. One prison has 2156 prisoners. If thirty prisoners will be selected from this prison, what are the first three random numbers that would be selected if the calculator were seeded with the number 12?

_____, _____, _____

Suppose the entire sample included prisoners from all the prisons. In total, 300 prisoners were selected. The number of prisoners of each race in this sample is shown in the table below. (This distribution is based on the actual distribution in WA prisons.) (www.doc.wa.gov/facilities/prison/)

White	Asian	Black	Native	Other
216	11	56	12	5

c. Which test is appropriate for this problem?

d. Make a double bar graph that shows a comparison between the observed and expected number of prisoners for each race.

e. Make a table to find the χ2 value. Use the χ2 table to find the p-value.

f. Write a concluding sentence.

g. Explain this conclusion in English. What do you think is the reason for this result?

h. If a solution is needed, what solution would you suggest?

Chapter 1 Data and Statistics

1. A survey question asked whether you were looking forward to the time when most of the cars on the road were self-driving (autonomous) cars, and the choice of answers was yes or no.

   a. Is the data from the responses to this question categorical or quantitative?

   b. Is the appropriate statistic $\hat{p}$ or $\bar{x}$?

   c. The table below gives the responses to 20 questions. Calculate the value of the appropriate statistic used for the answer yes.

no	yes	yes	yes	no	no	yes	no	no	yes
yes	no	yes	no	yes	no	yes	yes	yes	no

2. In the school's cafeteria, an employee counted the number of people sitting at each table.

   a. Is the data from the responses to this question categorical or quantitative?

   b. Is the appropriate statistic $\hat{p}$ or $\bar{x}$?

   c. The table below gives the number at 10 different tables. Calculate the value of the appropriate statistic.

5	6	8	7	4
1	7	8	3	1

Chapter 1 Writing Hypotheses

Name_____ Effort _____/4 Attendance _____/1 Total _____/5

1. The equal sign must always go in the null hypothesis (H_0)
2. The equal sign may never appear in the alternate hypothesis (H_1)
3. The alternate hypothesis uses one of the following: $<, >, \neq$
4. Both hypotheses must be about the same parameter (mean (μ) or proportion (p)). If the hypothesis is about a proportion then use $H_0 : p = a$ number between 0 and 1. If the hypothesis is about a mean, use $H_0 : \mu = a$ number.
5. The number in the null and alternate hypothesis must be the same.

Example: What proportion of students ate breakfast today?
$H_0 : p = 0.60$
$H_1 : p < 0.60$

Example: What is the average number of calories consumed for breakfast today by students?
$H_0 : \mu = 200$
$H_1 : \mu > 200$

Write your hypotheses for each question. Use each of the three inequalities at least once.

1. What is the average heart rate of college students?

   $H_0 :$
   $H_1 :$

2. Given the choice between humanity creating a fantastic future with technology or suffering a collapse of society due to resource depletion and other environmental problems, what proportion of college students do you hypothesize believes the future will be fantastic?

   $H_0 :$
   $H_1 :$

3. What is the average time, in minutes, that it takes students to get to school in the morning?

   $H_0 :$
   $H_1 :$

4. What proportion of students eat raw cookie dough?

   $H_0 :$
   $H_1 :$

Chapter 1 Sampling Distributions

1. In the distribution to the right:

What proportion of sample means will be between 150 and 170?

What proportion of sample means will be between 200 and 230?

What proportion of sample means will be between 150 and 230?

2. In the distribution to the right:

What proportion of sample proportions will be between 0.70 and 0.74?

What proportion of sample proportions will be between 0.84 and 0.90?

What proportion of sample proportions will be less than 0.70?

Chapter 2 p-values and levels of significance

1. For each row of the table you are given a p-value and a level of significance (α). Determine which hypothesis is supported, if the data are significant and which type error could be made. If a given p-value is not a valid p-value, put an x in each box in the row.
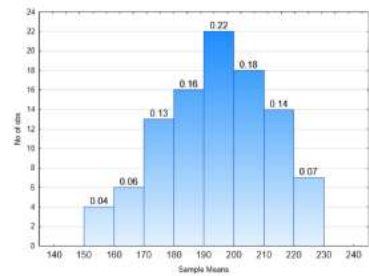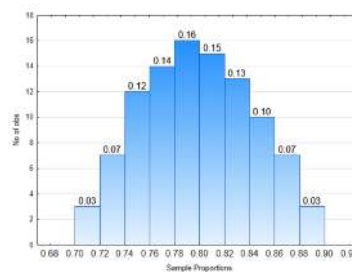
p - value	α	Hypothesis H_0 or H_1	Significant or Not Significant	Error Type I or Type II
0.48	0.05			
0.023	0.10			
6.7E-6	0.01			

Identify each as true or false if data are not significant
_____ The null hypothesis is definitely true
_____ The alternative hypothesis is definitely true
_____ The alternative hypothesis is rejected
_____ The null hypothesis was not rejected
_____The p-value is larger than α

2. For each row of the table you are given a p-value and a level of significance (α). Determine which hypothesis is supported, if the data are significant and which type error could be made. If a given p-value is not a valid p-value, put an x in each box in the row.

p - value	α	Hypothesis H_0 or H_1	Significant or Not Significant	Error Type I or Type II
0.048	0.05			
0.0023	0.10			
6.70	0.01			

Identify each as true or false if data are not significant
_____ The null hypothesis is definitely true
_____ The alternative hypothesis is definitely true
_____ The alternative hypothesis is rejected
_____ The null hypothesis was not rejected
_____The p-value is larger than α

Elementary Hypothesis Test, Example 1 Arsenic

Briefing: Arsenic is a naturally occurring element and also a human produced element (e.g. fracking, combustion of coal) that can be found in ground water. It causes a variety of health problems and can lead to death. The EPA limit is 10 ppb, meaning 10 ppb or higher is unsafe. Problem: Fracking was started in your community. A year later, sickness in the community leads health department officials to test your water to determine if it is contaminated with arsenic. The official will take 5 samples of water over the next 2 months and decide whether you have safe water or unsafe water based on the average of these samples. The hypotheses to be tested are: $H_0 : \mu = 10$ (Not safe) $H_1 : \mu < 10$ (Safe). The level of significance is: $\alpha = 0.12$.

Assume these are the two possible distributions that exist.
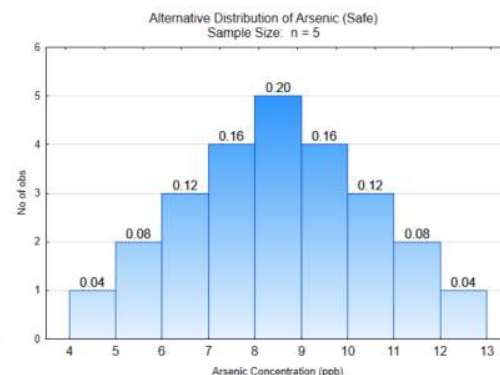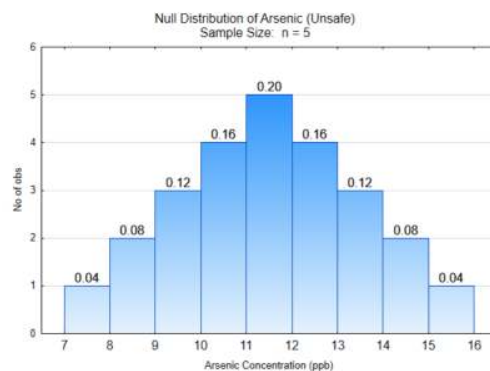
What is the direction of the extreme?
Show the decision line on both distributions.

What is the critical value?

Label α, β, and power

What is the probability of α?

What is the probability of β?

What is the power?

What is the consequence of a Type I error?

What is the consequence of a Type II error?

Data: What you select from the container that was passed around the classroom

Write a concluding sentence:

What decision do you make about your house and water supply?

<center>Elementary Hypothesis Test, Example 2: Do a majority of people in the US believe it is time for a new voting system?</center>

Briefing: The plurality voting system has been used in this, and other countries, since the democracies were formed. However, this system has led to the domination of two parties which don't necessarily reflect the opinions of the citizens. Some countries, such as New Zealand, and some states and communities in the US have adopted other voting systems which allow for better representation. Imagine a survey in which people were asked if they think it is time to change the voting system as a solution to the decisive partisanship that currently exists in the US. The objective is to determine if a majority of voters are ready to explore alternative voting systems. The hypotheses are: $H_0 : p = 0.50$, $H_1 : p > 0.50$, $\alpha = 0.07$.
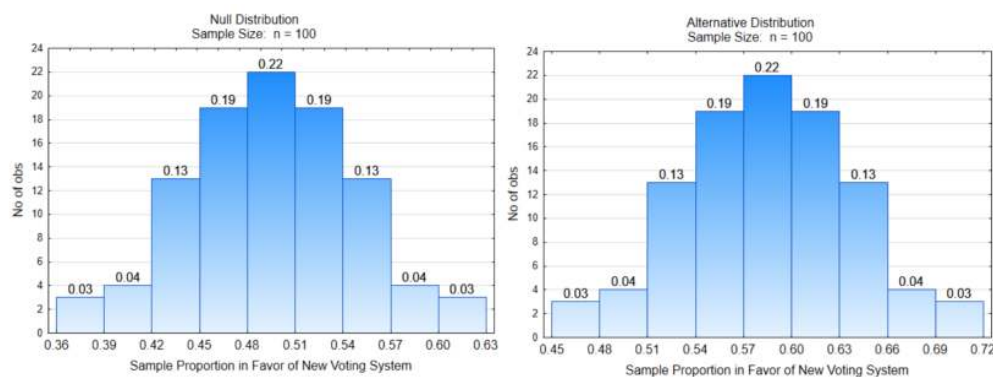
What is the direction of the extreme?
Show the decision line on both distributions.
What is the critical value?

Label α, β, and power

What is the probability of α?

What is the probability of β?

What is the power?

What is the consequence of a Type I error?

What is the consequence of a Type II error?

Data: 54 out of 100 voters wanted to explore alternative voting systems.

What is the sample proportion?

Write a concluding sentence:

<center>Chapter 2 Design Tables</center>

1. In an effort to determine which strategy is most effective for losing weight, a researcher randomly assigns subjects to one of four groups. One group (exercise) will become involved in a regular exercise program, a second group will be fed a balanced diet (food) but with appropriate size portions, a third group (exercise and food) will use both the exercise program and the balanced diet, while the fourth group (no change) will not change their diet or exercise.

Research Design Table	
Research Question:	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variables 1 (if present)	Levels:

2. People get excited when a young athlete achieves great success but there is always the question of whether the best college athletes were actually among the best young athletes. If interviews of starting varsity athletes from Division 1 schools were done and they were asked if they were considered a superior athlete as a 10 year old in their sport, would the proportion that were successful as a young child be different for males and females?

Research Design Table	
Research Question:	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variables 1 (if present)	Levels:

Chapter 2 Random Numbers

1. A survey at our college will be done. The administration expects different responses from running start students, traditional students, returning students and veterans. Sampling will be done from each of these groups.

What sampling method is being used?

If there are 1320 veterans (1-1320), what are the numbers of the first 3 randomly selected veterans if a seed value of 3 is used?

2. Time series data will be selected 5 years apart so that the data are independent. What are the numbers of the first 3 randomly selected years of data if the first year of data is 1960? Use a seed value of 4.

Chapter 2 Compare and Contrast Sampling Methods

Name_____ Effort_____/5 Attendance ____/1 Total ___/6

A current debate in Washington is whether to build coal export terminals so that coal mined in Montana and Wyoming can be sent by train to the Washington, Oregon or British Columbia coast and then exported to Asia. Some concerns include long trains that will be a constant disruption to traffic, coal dust from the trains will pollute the air near the rail lines, water pollution that will destroy the fisheries and fishing industry, and the concern that coal will contribute to climate change. Suppose a task force of 100 people from Idaho, Washington, Oregon and British Columbia gather to determine a regional policy for this situation. The task force is made up of government officials (G) and public citizens (C). They have all been assigned a number from 1 to 100. All sampling will be done with replacement. That means you can use the same number twice within one sampling method. This activity is meant to allow you to compare and contrast the 4 sampling methods.

Group 1 Idaho		Group 2 Washington		Group 3 Oregon		Group 4 British Columbia	
1 -C	No Coal	23 -G	No Coal	49 -G	Terminals	71 -C	Terminals
2 -C	Terminals	24 -C	Terminals	50 -G	No Coal	72 -G	Terminals
3 -C	Terminals	25 -G	No Coal	51 -G	No Coal	73 -C	No Coal
4 -C	Terminals	26 -G	No Coal	52 -G	No Coal	74 -G	Terminals
5 -C	No Coal	27 -C	Terminals	53 -C	No Coal	75 -C	Terminals
6 -C	Terminals	28 -G	No Coal	54 -C	Terminals	76 -C	Terminals
7 -C	Terminals	29 -G	No Coal	55 -G	No Coal	77 -C	Terminals
8 -G	No Coal	30 -G	No Coal	56 -C	No Coal	78 -G	Terminals
9 -G	Terminals	31 -C	No Coal	57 -G	No Coal	79 -G	Terminals
10 -G	No Coal	32 -C	Terminals	58 -G	No Coal	80 -C	Terminals
11 -C	Terminals	33 -G	Terminals	59 -G	No Coal	81 -C	No Coal
12 -G	No Coal	34 -G	Terminals	60 -C	No Coal	82 -G	Terminals
13 -G	Terminals	35 -G	Terminals	61 -C	Terminals	83 -G	Terminals
14 -G	No Coal	36 -G	Terminals	62 -G	No Coal	84 -C	No Coal
15 -G	Terminals	37 -C	Terminals	63 -C	No Coal	85 -C	No Coal
16 -G	No Coal	38 -G	Terminals	64 -C	Terminals	86 -G	Terminals
17 -C	Terminals	39 -C	Terminals	65 -C	Terminals	87 -C	No Coal
18 -G	No Coal	40 -G	No Coal	66 -G	No Coal	88 -C	Terminals
19 -G	Terminals	41 -G	No Coal	67 -G	Terminals	89 -G	No Coal
20 -G	Terminals	42 -G	Terminals	68 -G	No Coal	90 -G	No Coal
19 -C	Terminals	43 -G	No Coal	69 -C	Terminals	91 -G	No Coal
22 -G	No Coal	44 -C	No Coal	70 -C	Terminals	92 -C	Terminals
		45 -C	No Coal			93 -G	No Coal

		46 -C	No Coal			94 -C	No Coal
		47 -G	No Coal			95 -G	No Coal
		48 -G	No Coal			96 -G	Terminals
						97 -G	Terminals
						98 -C	Terminals
						99 -C	Terminals
						100 -G	Terminals

1. Simple Random Sample

Use your calculator with a seed of 23 to randomly select a sample of size 10. The lowest number is 1 and the highest is 100. List the selected numbers then determine the proportion of the sample that is against the coal terminals (No Coal).

Number: _____, _____, _____, _____, _____, _____, _____, _____, _____, _____,

N or T _____, _____, _____, _____, _____, _____, _____, _____, _____, _____,

Proportion that is against the coal terminals: $\hat{p} = $ _____

2. Stratified Random Sample

Use your calculator with a seed of 13. The low is 1 and the high is 100. Put the random numbers in the appropriate strata. When a stratum is filled, ignore other numbers that belong in it.

Citizens: Number _____, _____, _____, _____, _____,

N or T _____, _____, _____, _____, _____,

Government: Number _____, _____, _____, _____, _____,

N or T _____, _____, _____, _____, _____,

Proportion (use citizens and government officials combined) that is against coal terminals: $\hat{p} = $ _____

3. Systematic Random Sample

Use a 1 in K sampling method, with k = 10 to randomly select a sample of size 10. To determine the first number selected, use your calculator with a seed of 18, a low of 1 and a high of 10. Determine the proportion of the sample that is against coal terminals.

Number: _____, _____, _____, _____, _____, _____, _____, _____, _____, _____,

N or T _____, _____, _____, _____, _____, _____, _____, _____, _____, _____,

Proportion that is against coal terminals: $\hat{p} = $ _____

4. Cluster Sampling

Use your calculator with a seed value of 33 to randomly select one of the groups (1-4). Which group is selected? _____. What is the sample proportion of the selected group that is against coal terminals? $\hat{p} = $ _____

Chapter 3 Histograms and Box Plots

Name_____ Effort_____/5 Attendance ____/1 Total ___/6

The results of an exam on Chapters 2 and 3 from one statistics class are shown in the table below. The numbers represent the percent of possible points the student earned.

76.8	91.5	98.8	97.6	76.8	93.9	57.3	86.6	90.2
93.9	93.9	82.9	92.7	89.0	72.0	57.3	93.9	92.7
93.9	81.7	63.4	68.3	85.4	50.0	84.1	90.2	86.6
97.6	84.1	81.7	95.1	87.8	75.6	92.7	73.2	91.5

Low value _____ High value _____

Make a frequency distribution. Use interval notation for the boundaries [lower,upper).

Classes	

Make a histogram. Label completely.

Use your calculator to complete the table below by entering the original data into the lists.

	page266image8740176
Mean	
Standard Deviation Sx	page266image156002816
Minimum	page266image156003584
Q1	

Median	
Q3	
Maximum	

Make a box plot. Label completely.

Chapter 4 Inferential Theory

Question 2: Do more than 70% of Americans drink tea (either hot or iced)?

a. Write your null and alternate hypothesis:

b. Find P(S): c. Find P(F):

d. If you took a sample of 7 people, what is the probability the exact order would be SFSSFSS? That is, find P(SFSSFSS).

e. How many combinations are there for 5 successes in a sample of 7 people?

f. What is the probability you would get 5 successes in a sample of 7 people?

g. Make a binomial distribution for the number of successes in a sample of 7 people.

h. What is the mean and standard deviation for this distribution?

i. Finish the concluding sentence if there were 5 successes in a sample of 7 people. At the 5% level of significance, the proportion of Americans who drink tea
_____

Chapter 4 Inferential Theory – Testing Hypotheses

Pacific Northwest residents are often concerned with the issue of sustainability. If a survey of 400 Pacific Northwest individuals resulted in 296 who said they make choices based on being sustainable, then test the hypothesis that over 67% of individuals in this region make choices based on being sustainable.

Test the hypotheses ($H_0 : p = 0.67$ $H_1 : p > 0.67$) using three different methods and a level of significance of 0.05. For each method, you will be asked which hypothesis is supported.

1a. Binomial Distribution: Use the binomial distribution to calculate the exact p-value based on the data (296 out of 400).

_____ _____

Calculator input p-value

Which hypothesis is supported by the data? Choose 1: H_0 H_1

1b. Normal Approximation: Use the normal approximation to the binomial distribution to calculate the approximate p-value based on the data (296 out of 400). Provide the requested information.

$\mu = np =$, $\sigma = \sqrt{npq} =$

Formula Substitution z value p-value

Which hypothesis is supported by the data? Choose 1: H_0 H_1

1c. Sampling Distribution for Sample Proportions: Find the p-value using sample proportions for the data (296 out of 400). Provide the requested information.

Sample proportion

Formula Substitution z value p-value

Which hypothesis is supported by the data? Choose 1: H_0 H_1

A student at UC Santa Barbara(http://www.culturechange.org/cms/content/view/704/62/) did some research on the plastic red cups that people use for drinks at parties. These cups are made of Polystyrene, which cannot be recycled in Santa Barbara. Many of the cups end up in the landfill, but some end up in the ocean. In the nearby college town of Isla Vista, the researcher estimated that the average number of cups used per person per year was 58. Assume the standard deviation is 8.

In an effort to change the culture, suppose an education campaign was used to reduce the number of red cups by encouraging the purchase of beverages in cans (since they can be recycled). To determine if this is effective, a random sample of 16 students will keep track of the number of red cups they use throughout the year. The hypotheses that will be tested are: $H_0 : \mu = 58$ $H_1 : \mu < 58$, $\alpha = 0.05$

2a. What is the mean of the sampling distribution of sample means? $\mu_{\bar{x}}$ _____

2b. What is the standard deviation of the sampling distribution of sample means? $\sigma_{\bar{x}}$ _____

2c. Draw and label a normal distribution showing the mean and first three standard deviations (standard errors) on each side of the mean for the distribution of sample means of 16 students.

2d, Test the hypothesis if the sample mean of the 16 students is 55 using a level of significance of $\alpha = 0.05$.

Formula Substitution z value p-value

2e. Based on the results in this experiment, has there been a reduction in the use of red cups? Choose 1: Yes No

Chapters 5 and 6 Mixed Practice with Hypothesis Testing and Confidence Intervals

For each problem, provide the hypotheses and test the hypotheses by calculating the test statistic and p-value. Fill in all the blanks in the following sentence. Also, give calculator answer in parentheses for the test statistic and p-value. This will not be corrected or graded but will help prepare you for the exam.

1. A student read that in the bay area of California, the average person produces 2 pounds of garbage per day. The student believed that she produced less than that but wanted to test her hypothesis statistically. She collected data on 10 randomly selected days. Use $\alpha = 0.05$.

2.0	2.3	1.9	1.9	2.3
1.2	2.3	2.1	1.7	1.8

H_0 :

H_1 :

What is the sample mean? Sample Mean _____
What is the sample standard deviation? Sample Standard Deviation_____

Formula Substitution Test Statistic value p-value

Calculator:

Test Statistic value p-value

The average amount of garbage produced daily by the student _____ significantly less than 2 pounds (t = _____, p = _____, n=_____).

What is the 95% confidence interval for the amount of garbage she produces?

Formula Substitution Margin of Error Confidence Interval

Calculator confidence Interval: _____

2. A living wage is the hourly rate that an individual must earn to support their family, if they are the sole provider and are working full-time. In 2005, it was estimated that 33% of the job openings had wages that were inadequate (below the living wage). A researcher wishes to determine if that is still the case. In a sample of 460 jobs, 207 had wages that were inadequate.Test the claim that the proportion of jobs with inadequate wages is greater than 0.33. Let $\alpha = 0.01$.

H_0 H_1

Formula Substitution Test Statistic value p-value

Calculator:

Test Statistic value p-value

What is the 90% confidence interval for the proportion of jobs with inadequate wages?

Formula Substitution Margin of Error Confidence Interval

Calculator confidence Interval: _____

3. Suppose you had two different ways to get to school. One way was on main roads with a lot of traffic lights, the other way was on back roads with few traffic lights. You would like to know which way is faster. You randomly select 6 days to use the main road and 6 days to use the back roads. Your objective is to determine if the mean time it takes on the back road μb is different than the mean time on the main road μm. The data is presented in the table below. The units are minutes. Assume population variances are equal. Because the sample size is small, you decide touse a significance level of $\alpha = 0.1$.

Back Road	14.5	15.0	16.2	18.9	21.3	17.4

Main Road	19.5	17.3	21.2	20.9	21.1	17.7

Write the appropriate null and alternate hypotheses: H0: _____ H1:_____
What is the sample mean for each route? Back Road_____ Main Road _____
What is the sample standard deviation for each route? Back Road_____ Main Road _____
Test this using your calculator
Test Statistic value p-value

There _____ a significant difference between taking the back road and the main road (t = _____, p = _____, n=_____).

What is the 99% confidence interval for the difference in the mean times?

Calculator confidence Interval: _____

Use your calculator generated confidence interval to calculate the margin of error _____

4. Some parents of age group athletes believe their child will be better if they pay them a financial reward for being successful. For example they may pay $5 for scoring a goal in soccer or $1 for a best time at a swim meet. The argument against paying is that it is counterproductive and destroys the child's self-motivation. Is the dropout rate of children that have been paid different than of children who have not been paid? Let $\alpha = 0.05$.

Dropout rate of children who have been paid: 450 out of 510
Dropout rate of children who have not been paid: 780 out of 930

H_0 H_1

Test this using your calculator

Test Statistic value p-value

What is the 95% confidence interval for the difference between the dropout rate of children that have been paid and children who have not been paid? Let α = 0.05.

Calculator confidence interval: _____

Use your calculator generated confidence interval to calculate the margin of error _____

Chapter 7 – Linear Regression Analysis

Homework problem 4 looks at the relationship between the population of a metropolitan area and the number of patents produced in that area. Below is an expand sample. It includes more of the large metropolitan areas. Make a new scatter plot. Use a different color marker to Indicate Las Vegas and Fresno on this scatter plot. In the homework, these two communities looked like outliers. Do they still?

Use a 5% level of significance.

Metropolitan Area	Population	Total Patents 2000-2011
Las Vegas-Paradise, NV	806,923	2160
Fresno, CA	494,665	468
Decatur, AL	55,683	130
Guayama, PR	22,691	4
Taylorville, IL	11,246	97
Harriman, TN	6,350	53
Kapaa, HI	10,699	32
Minot, ND	40,888	19
Lewisburg, TN	11,100	10
Austin-Round Rock-San Marcos, TX	935,171	22916
Oxnard-Thousand Oaks-Ventura, CA	234,417	5245
Colorado Springs, CO	416,427	2840
Virginia Beach-Norfolk-Newport News, VA-NC	861,516	1638
Omaha-Council Bluffs, NE-IA	471,188	1072
Ames, IA	58,965	722
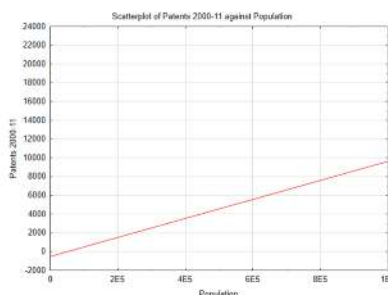
Show calculator outputs including the correlation, r^2 value and equation of the regression line (which has been conveniently placed on the graph for you). Write a statistical conclusion then interpret the results. Use a level of significance of 0.10.

Correlation _____

Coefficient of determination (r^2 value) _____

Regression equation _____

Hypothesis test concluding sentence:

Chapter 7 – χ^2

If a teacher changes the way a course is taught or uses a new book, how does the teacher know if the changes resulted in better success for the students? One way is to compare the distribution of grades (A, B, C, below C) to what has happened in past classes, assuming that assessments and grading were similar.

The distribution of grades for past classes that used the first edition of Foundations in Statistical Reasoning is shown in the middle column of the table below. The number of students who received each grade when using the second edition is shown below.

Grade	Proportion	Count from the second edition
A	0.349	16
B	0.287	11
C	0.204	7
Below C	0.160	6

Test the hypothesis that the distribution of grades from the second edition is different than the distribution from the first edition.
Write the hypotheses:

H_0:
H_1:

Which test is appropriate for this problem?

A. _____ Goodness of Fit B. _____ Test for Independence C. _____ Test for Homogeneity

Test the hypothesis using the table below.

Observed	Expected	$O - E$	$(O - E)^2$	$\dfrac{(O - E)^2}{E}$
				$\chi^2 =$

Write a concluding sentence:

Which of the following conclusions does the evidence support?

_____The second edition resulted in a significantly improved distribution of grades

_____The second edition resulted in a significantly worsening of the distribution of grades

_____The second edition did not appear to affect the distribution of grades

This problem could be done a different way if you were told the number of people who got each grade using the first edition.

Grade	Count from the first edition	Count from the second edition
A	174	16
B	143	11
C	102	7
Below C	80	6

Test the hypothesis that the distribution of grades from the second edition is different than the distribution from the first edition.
Write the hypotheses:

H_0 :
H_1 :

Which test is appropriate for this problem?

A. _____ Goodness of Fit B. _____ Test for Independence C. _____ Test for Homogeneity

Use the matrix and χ^2 test on your calculator to test the hypothesis.

10: Communication of Statistical Results

In Algebra or other deterministic math, if you substitute numbers into a formula and calculate the answer, then the results can be reported without too much additional thought. However, with statistics, there is not necessarily one simple answer. Rather, it is necessary to consider all the evidence that can be understood from the sample. This means a careful interpretation of the graphs, and evaluation of the statistics, and a consideration of the test of significance. To simply rely on a p-value, or conversely ignore it all together, are both problematic. A p-value is important, but it is not sacred.

In this chapter you will be given graphs, statistics, and p-values. Your task will be to give a written explanation that is justified by the results. You should provide context as well as reference to the evidence. Before writing, you should answer the following questions in your mind.

1. What is the context? What is the story about and what is the purpose?

2. What does the graph show? Think about the distribution. Do you see any patterns or outliers?

3. Look at the statistics. Do they do a good job of representing the distribution shown in the graph?

4. Identify the hypothesis test. Is it appropriate?

5. Does the p-value indicate the data are significant? Keep in mind that significant and important are not synonymous. Is the evidence strong or weak? Use a 5% level of significance for all problems.
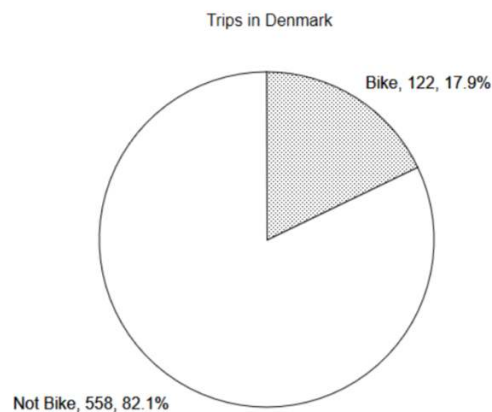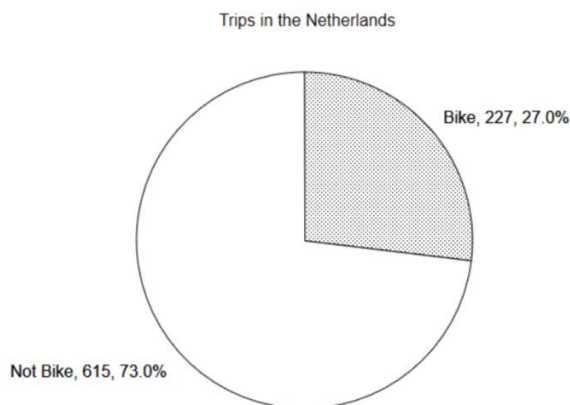
There should be a flow to your writing.

- It should begin with background information to provide context. This should include a statement of the objective or the question to be answered.
- Once the context has been provided, write about the evidence that you can gather from the graphs and the statistics.
- Since you have only been given sample data, it is necessary to make an inference. This is where you will write a concluding sentence such as you have been practicing throughout this text.
- Write a conclusion that directly answers the question and is consistent with the evidence and inference. If some of the evidence is contradictory, address the contradictions.

There are three communication activities to do. More direction is given with the first than the others. They increase in point values as your writing should improve each time. The first should be submitted after the exam on Chapter 1. The second should be submitted after the exam on Chapter 3, and the third should be submitted after the exam on Chapter 6.

<div align="center">Effective Communication 1 (due the day after the exam on Chapter 1)</div>

Name _____ Points _____/4 (-1 per day for late)

The information presented below is about a comparison of the proportion of trips made by bicycle between The Netherlands and Denmark, the two top countries in the world for bicycling. A sample was taken of people in the Netherlands and Denmark. They were asked about the mode of transportation used during the last trip they made from their home to another destination. The data is whether the trips were made by bike or other mode of transportation. The objective is to determine if the proportion of trips by bicycle is higher in the Netherlands than Denmark.(http://top10hell.com/top-10-countrie...es-per-capita/ Viewed 6/21/17) On the back of this page, write your analysis legibly, or type it. There are guides about what should be written.

Trips in the Netherlands

Bike, 227, 27.0%

Not Bike, 615, 73.0%

Trips in Denmark

Bike, 122, 17.9%

Not Bike, 558, 82.1%

Results of hypothesis test: p-value = 0.000016.

Section 1. Write about the context. What is this information about? Why might it be of interest? What is the question that is being asked? Try to engage the reader.

Section 2. Give evidence. This is where you explain the distribution of the data and give the statistics. Make use of all relevant statistics and evidence from the graph.

Section 3. Make an inference. This is where you write the concluding sentence to extend the results from the sample to the population. Make sure it is phrased in a way that is consistent with the question being asked in Section 1. Include p-value and sample sizes.

Section 4. Conclusion. Provide a direct answer to the question, being consistent with the evidence and inference.

Effective Communication 2 (due the day after the exam on Chapter 3)

Name _____ Points _____/8 (-2 per day for late)

The current national debt of the United States is about 20 trillion dollars. While this number sounds high (okay, it is high), there is a difference if a country of the size of the US has a 20 trillion dollar national debt compared with a country the size of Bermuda (for example). Therefore, one thing economists will do is to find the ratio of the national debt of a country to the Gross Domestic Product (GDP) for the country. The current ratio for the US is 106. This means that the debt is 106% of the GDP. The information below will allow you to determine if the average debt to GDP ratio for other randomly selected countries is less than the US.2 The US is not included in the distribution below.

Debt to GDP Ratio of Randomly Selected Countries

No of obs

Debt to GDP Ratio (Percent)
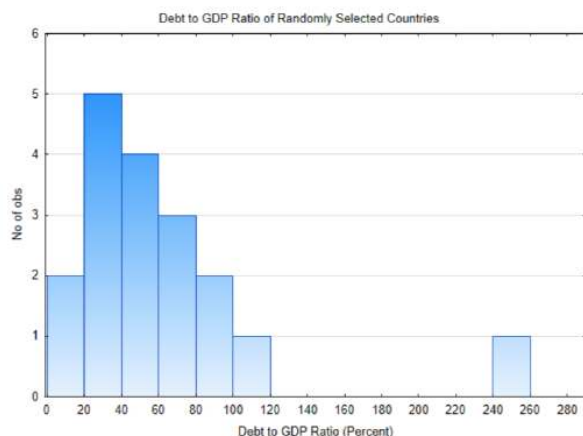
Mean: 60.8, Median: 45.2, Minimum: 3.1, Max: 250.4 (Japan), Standard Deviation: 54.1

Hypothesis Test Results: P-value = 0.0012, n = 18.

Write an analysis that compares the debt to GDP ratio of other countries to the US. Include context, statistics and a statement of significance. Use good grammar and spelling. Write legibly or type.

Organize your writing in the order that was done in the first effective communication activity, give background and context followed by evidence, inference, and then conclusion.

Effective Communication 3 (due the day after the exam on Chapter 6)

Name _____ Points _____/12 (-3 per day for late)

The CEO of an investment company is trying to fill a position of Senior Investment Manager. Two investment advisors are applying for the same position. You have been given the task of analyzing the success of the investments they managed for their clients. The data that you will compare is the change in the value of the investment for each $10,000 that is invested. For example, if the investment grew to $12,400, then the change would be $2,400. However, if the investment shrank to $6500, the change would be -$3,500. Below you will find graphs, statistics and the results of a hypothesis test that compares the two investment managers, Andy and Bobbie (note, these names can be used for males and females, so you can use whichever pronouns you want, e.g. him or her). Your objective is to write a report that compares the two and make a recommendation. You must justify your recommendation.
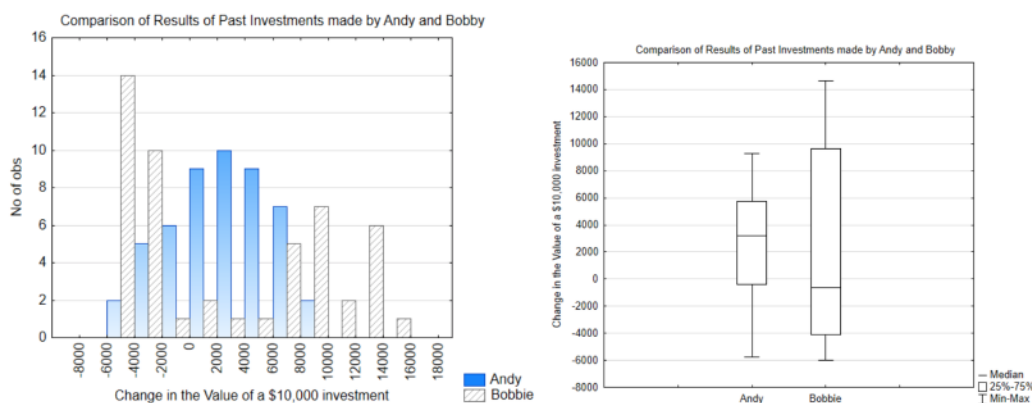
	Andy	Bobbie
Mean	2537.74	2550.76
Median	3220.50	-635.50
Standard Deviation	3736.56	7112.00
n	50	50

The results of a t-test for 2 independent means to test for a difference between means are t = -0.011, p = 0.991.

Write an analysis to help the CEO make a choice between Andy and Bobbie. Include context, statistics and a statement of significance. Use good grammar and spelling. Write legibly or type. Organize your writing in a similar way to the other effective communication activities.

Index

Answers to most problems

Answers are provided for most problems so you can immediately check your answers to see if you are doing it correctly. This should facilitate learning. Answers are not provided for some problems to simulate real-world conditions and tests, since answers are not known in either case.

Chapter 1

1a. parameter

1b. statistic

2. parameter

4a. $H_0 : \mu = 20$ $H_1 : \mu > 20$

4c. $H_0 : \mu_A = \mu_C$ $H_1 : \mu_A \neq \mu_C$

4d. $H_0 : p_m = p_A$ $H_1 : p_m \neq p_A$

6.

p-value	α	Hypothesis H_0 or H_1	Significant or Not Significant	Error Type I or Type II
0.043	0.05	H_1	Significant	Type I
0.32	0.05	H_0	Not Significant	Type II
5.6×10^{-6}	0.05	H_1	Significant	Type I
7.3256	0.01	x	x	x

7a. At the 5% level of significance, the proportion is significantly greater than 0.5 (p = 0.022, n = 350).

7b. At the 1% level of significance, the proportion is not significantly less than 0.25 (p = 0.048, n = 1400).

7d. At the 5% level of significance, the mean is different than 20 (5.6×10^{-5}, n = 32).

8.
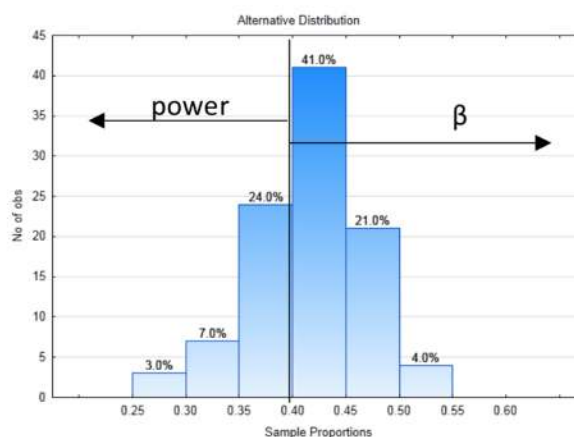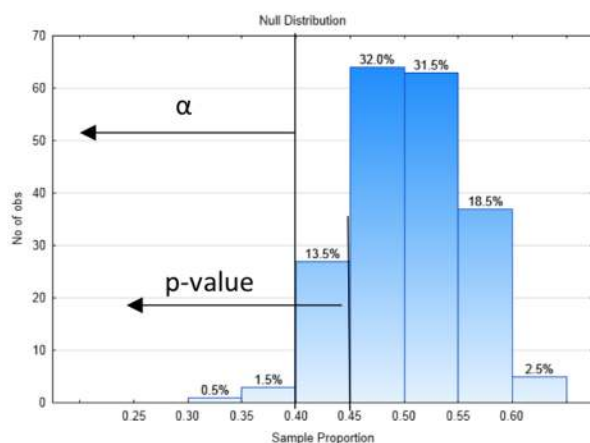
8a. Left

8c. 0.40

8e. 0.66

8h. 0.155

8i. H_0

8j. No

8k. type II

8l. At the 2% level of significance, the proportion is not significantly less than 0.5 (p = 0.155, n = 80).

9.

9a. Right

9c. 340

9d. 0.035

9f. 0.67

9h. 0.005

9i. H1

9j. Yes

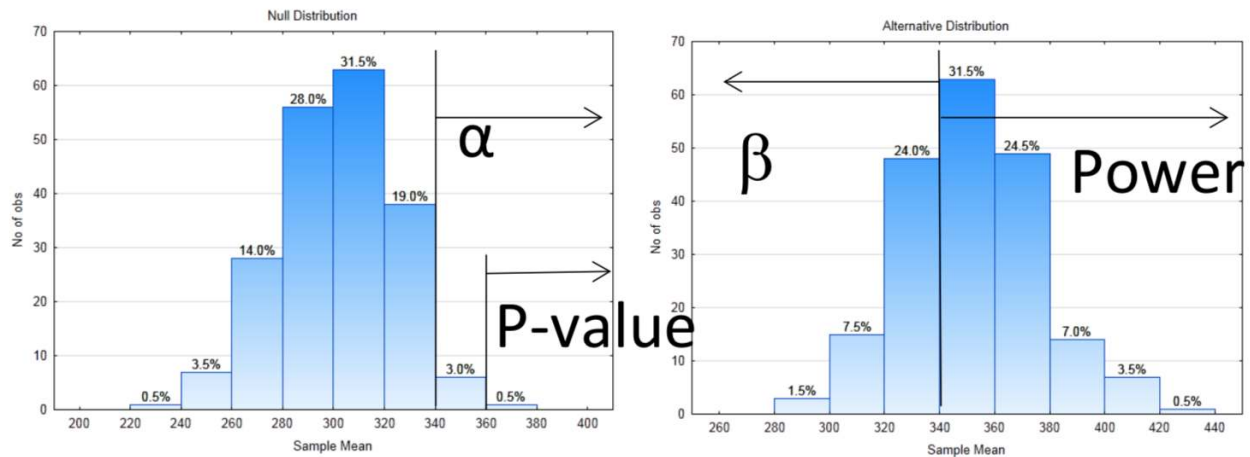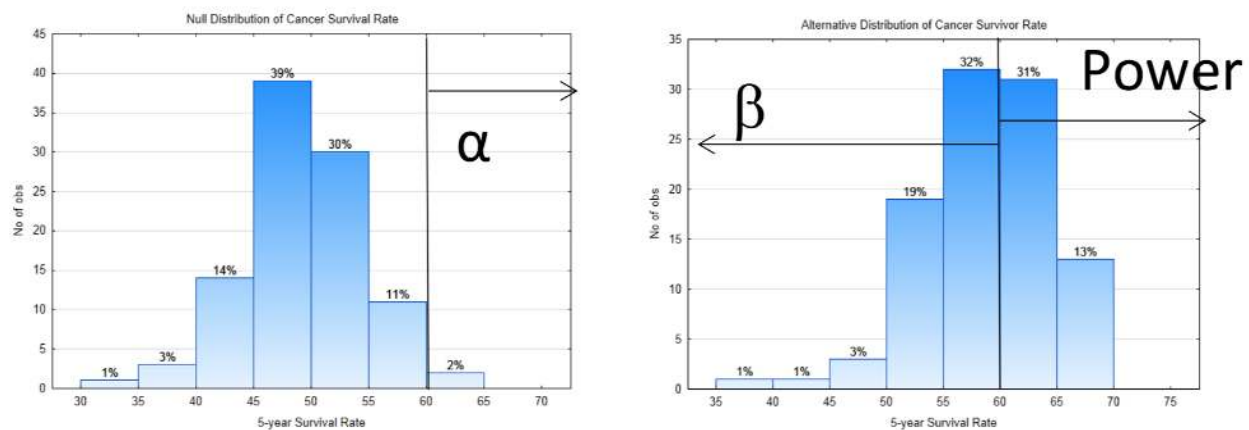9l. At the 0.035 level of significance, the mean is significantly greater than 300 (p = 0.005, n = 10).

10a. $H_0 : p = 0.5$ $H_1 : p > 0.5$

10b. Right

10d. 60

10f. 0.56

10g. 0.44

10h. 0

10i. At the 2% level of significance, the proportion who survive cancer at least 5 years is significantly greater than 0.5 (p = 0, n = 100).

Chapter 2

1.

Research Design Table	
Research Question: which route has a faster average time	
Type of Research	<u>Observational Study</u> Observational Experiment Manipulative Experiment
What is the response variable?	Time it takes for the commute
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables	Think of at least 2 yourself
Grouping/explanatory Variables 1 (if present) routes	Levels: Route 1 and Route 2

3.

Research Design Table	
Research Question: Which is more effective at increasing biodiversity, the hands-off approach or the deliberate approach?	
Type of Research	Observational Study <u>Observational Experiment</u> Manipulative Experiment
What is the response variable?	Number of species
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables	Think of at least 2 yourself
Grouping/explanatory Variables 1 (if present) approaches	Levels: hands-off deliberate control

4b.

Research Design Table	
Research Question: Does static or dynamic stretching result in improvement in flexibility in the largest proportion of people?	
Type of Research	Observational Study Observational Experiment <u>Manipulative Experiment</u>
What is the response variable?	Improvement in sit and reach test
What is the parameter that will be calculated?	Mean <u>Proportion</u>
List potential latent variables	Think of at least 2 yourself
Grouping/explanatory Variables 1 (if present) Stretching method	Levels: static dynamic

8a. 102 N, 40 N, 18 Y, 49 N, 61 N, 60 N, 57 N, 16 N, 90 N, 46 Y,

135 N, 105 Y, 83 N, 102 N, 3 N, 70 Y, 47 N, 42 N, 5 N, 68 N,

Sample Proportion __4/20 = 0.2

8b. West 37 Y,45 N, 21 N, 56 N, 70 Y, 68 N, 65 N, 18 Y, 22 N, 52 Y, 75 Y,

East 93 N, 105 Y, 109 N, 90 N, 114 N, 137 Y, 133 N, 131 N, 104 Y

Sample Proportion _8/20 = 0.4

8c. 2 Y, 9 N, 16 N, 23 N, 30 N, 37 Y, 44 Y, 51 Y, 58 N, 65 N,
72 Y, 79 N, 86 Y, 93 N, 100 N, 107 N, 114 N, 121 N, 128 N, 135 N,

Sample Proportion _6/20 = .30

8d. Which cluster is selected? ___7____ Sample Proportion _9/20=0.45

8m

				8a		**8c**		**8b**	**8d**	**8e**										
0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00

9a.

Research Design Table	
Research Question: Does raising the minimum wage cause unemployment to increase?	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	Change in Unemployment rate
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables	Think of at least 2 yourself
Grouping/explanatory Variables 1 (if present) State minimum wage change	Levels: Increase minimum wage No change in minimum wage

9b. Cluster

9c. 2004, 2012, 2006

9d. Provide your own thoughtful answer.

9e. Provide your own thoughtful answer.

9f. Provide your own thoughtful answer.

9g. At the 5% level of significance, there is not a significant difference in the change in unemployment rate between states that raised their minimum wage and those that didn't (p = 0.286).

10a.

Research Design Table	
Research Question: Will the number of falls increase after bedrails are removed?	
Type of Research	Observational Study Observational Experiment <u>Manipulative Experiment</u>
What is the response variable?	Falls
What is the parameter that will be calculated?	<u>Mean</u> Proportion Correlation
List potential latent variables	Think of at least 2 yourself
Grouping/explanatory Variables 1 (if present) Bedrails	Levels: Present Not present

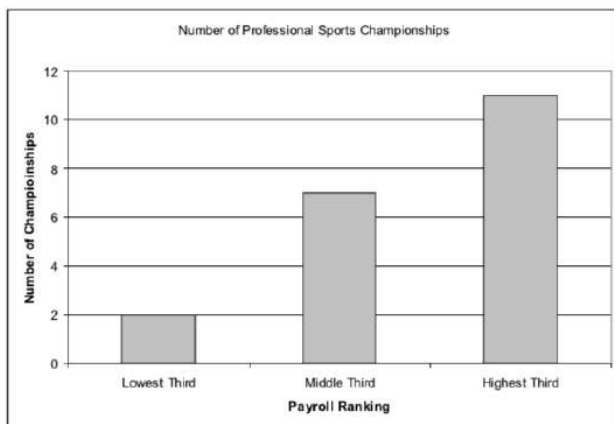10b. At the 5% level of significance, there was not a significant increase in the number of falls per 10,000 bed days after the implementation of the new policy (p = 0.18).

10c. There were fewer serious falls, more minor and no-injury falls. A possible reason is that the falls are from a lower height since the patient isn't crawling over the top of the rails.

10d. Provide your own thoughtful response.

Chapter 3

1.

2.

Corn Seed Availablity In 1983 Compared with 1903
Available in 1983, 12, 4%

Not Available in 1983, 295, 96%

Corn Seeds
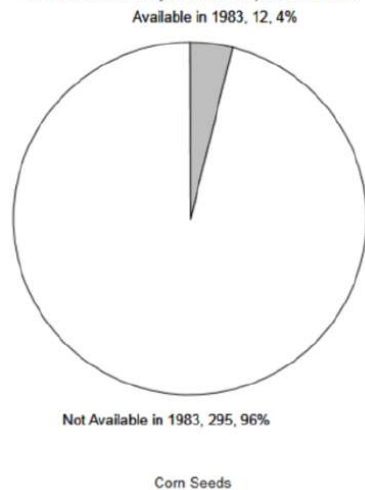
3. mean = 43, Standard deviation = 4.78, variance = 22.89

4.

Automaticity Scores

Automaticity Scores for 1 Minute Test of Mixed Arithmetic Operations

— Median = 22
□ 25%-75%
= (17, 28)
⊥ Min-Max
= (9, 38)
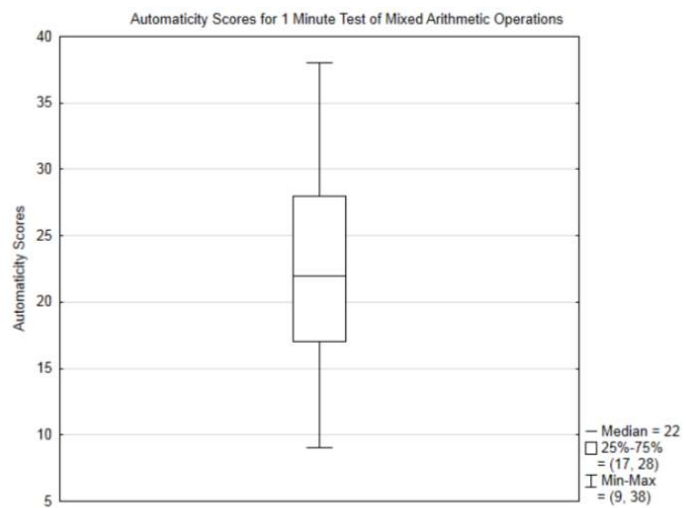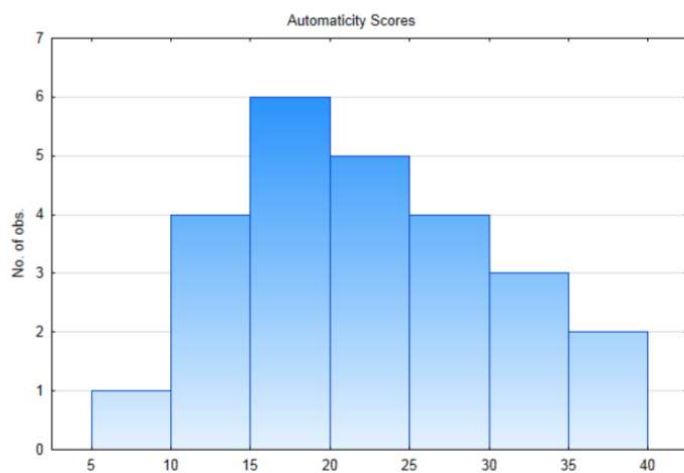
5. $r = \dfrac{\text{cov}(x, y)}{s_x s_y} = \dfrac{2.09}{5.56 \cdot 1.20} = 0.313$

6a.

Research Design Table	
Research Question: Is average number of problems answered correctly in one minute was greater for students who passed the class than for those who didn't pass?	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	Number of correctly answered problems
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables	
Grouping/explanatory Variables 1 (if present) Success in course	Levels: Pass Fail
Grouping/explanatory Variables 2 (if present)	Levels:

6b. Calc: 2,9

6c. Cluster

6d. Quantitative discrete

6e.

6f. Provide your own thoughtful response.

6g.

	Mean	Variance	Standard Deviation
Failed	15.89	33.88	5.82
Passed			

6h. At the 5% level of significance, the average automaticity score of those who pass the class is significantly more than the score of those who fail the class (p = 0.0395, $n_{\text{fail}} = 19$, $n_{\text{pass}} = 49$).

6i. Yes

7a.

Research Design Table

Research Question: Is the average heart rate lower with alcohol than with water?	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	heart rate
What is the parameter that will be calculated?	Mean Proportion Correlation
List potential latent variables	List you own ideas
Grouping/explanatory Variables 1 (if present) drop	Levels: water (first time) alcohol, water (second time)
Grouping/explanatory Variables 2 (if present)	Levels:

7b.
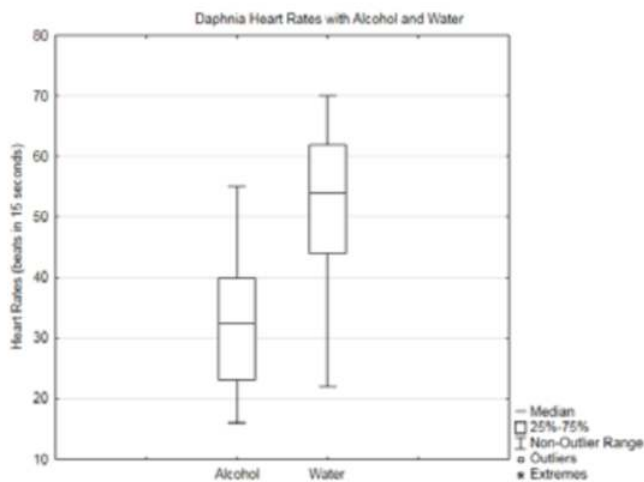
7c.

	Heart Rate after Alcohol	Heart Rate after Water
mean	32.8	52.1
Standard Deviation	10.7	13.0
Median	32.5	54.0

7d. At the 5% level of significance, the average daphnia heart rate with alcohol is significantly less than the average daphnia heart rate with water (p = 1.28×10^{-5}, $n_{\text{alcohol}} = 18$, $n_{\text{water}} = 18$).

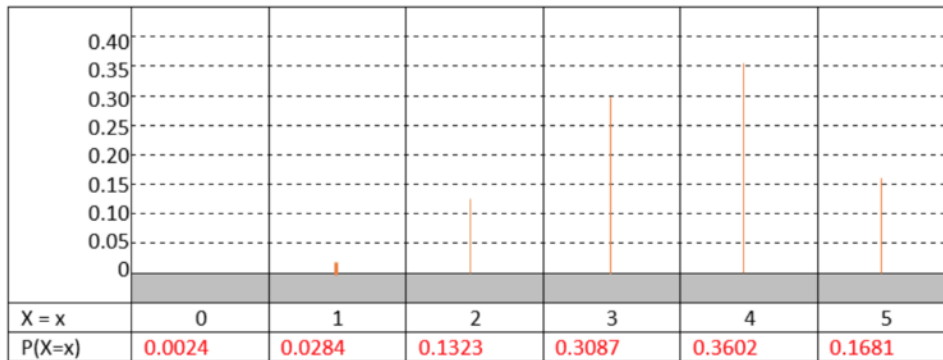7e. Provide your own thoughtful answer.

Chapter 4

1.
1a. P(S) = 0.70

1b. P(F) = 0.30

1c. P(FSSSF) = P(F)P(S) P(S) P(S) P(F) = (0.3) (0.7) (0.7) (0.7) (0.3) = 0.03087

1d. 10

1e. 0.3087

1f.

X = x	0	1	2	3	4	5
P(X=x)	0.0024	0.0284	0.1323	0.3087	0.3602	0.1681

1g. $\mu = np = 5(0.7) = 3.5$ $\sigma = \sqrt{npq} = \sqrt{5(0.7)(0.3)} = 1.025$

1h. $P(X \leq 3)$ = binomcdf(n, p, x) = binomcdf (5, 0.70, 3) = 0.4718. The null is supported. The data are not significant.

2.

2a. P(S) = 0.40

2b. P(F) = 0.60

2c. 0.0036864

2d. 21

2e. 0.0774

2f.

$X = x$	0	1	2	3	4	5	6	7
$P(X = x)$	0.0280	0.1306	0.2613	0.2903	0.1935	0.0774	0.0172	0.0016

2g. 2.8, 1.296

2h. P-value = 0.0963 alternative

3.

3j. P-value = 0.047

At the 5% level of significance, the proportion of residents opposed to the terminals is significantly greater than 0.5 (p = 0.047, n = 300)

3k. z = 1.73 p-value 0.0418

At the 5% level of significance, the proportion of residents opposed to the terminals is significantly greater than 0.5 (z = 1.73, p = 0.0418, n = 300).
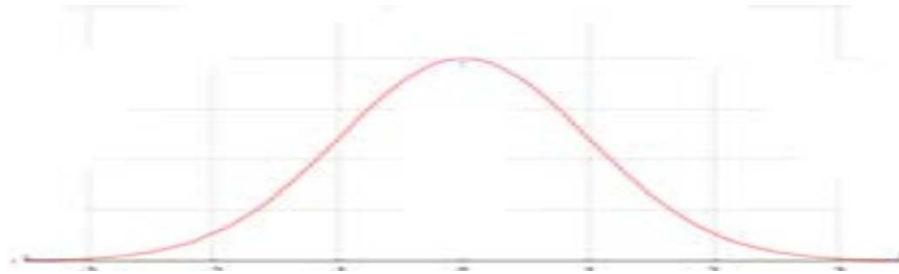
3l. 0.55

3m. 0.02887

3n. z = 1.73 p-value 0.0418

At the 5% level of significance, the proportion of residents opposed to the terminals is significantly greater than 0.5 (z = 1.73, p = 0.0418, n = 300).

4a. $H_0 : \mu = 43,362$ and $H_1 : \mu < 43,362$

4b. $\mu = 43,362$

4c. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{7900}{\sqrt{10}} = 2498$

4d.

X axis: 35868 38366 40864 43362 45860 48358 50856

4e. 18,225 (use stat-edit and then stat-calc-1-var stats)

4f. z = -10.06 p-value < 0.0002

6b. $H_0 : \mu = 54.1$ and $H_1 : \mu > 54.1$

6d. $barx = 46.73,\ s = 16.377$

6e. $\mu = 54.1\ \sigma_{\bar{x}} = 2.939$

6f.

45.4 48.3 51.2 54.1 57 59.9 62.8

6g. z = -2.51 p-value 0.9940

At the 5% level of significance, the average walk score of small cities is not significantly greater than big cities (z = -2.51, p = 0.994, n = 30)

7e. p-value = 0.3865 0.0148

7g p-value = 0.0129

7h z = 2.229, p= 0.0129

8c.

	Impact	Control
Mean	455	
Standard Deviation	614.8	
Median	150	

Chapter 5

1a. $H_0 : \mu_T = \mu_D$ $H_1 : \mu_T \neq \mu_D$ Test: 2 independent samples t test

1b. $H_0 : p = 0.5$ $H_1 : p > 0.5$ Test: 1 proportion z test

1c. $H_0 : p_{STEM} = p_{SS}$ $H_1 : p_{STEM} \neq p_{SS}$ Test: 2 proportion Z test

1e. $H_0 : \mu = 7$ $H_1 : \mu > 7$ Test: 1 sample t test

1g. $H_0 : \mu = 0$ $H_1 : \mu < 0$ Test: 1 sample t test

2a. $H_0 : \mu = 15$, $H_1 : \mu < 15$

2b. Test the bypothesis:

$t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ $\quad t = \dfrac{14.28 - 15}{\frac{4.6}{\sqrt{30}}}$ $\quad$ t = -0.857 p > 0.1 or p = 0.1991

Formula Substitution Test Statistic p-value

2c. Fill in the blanks for the concluding sentence. At the _5%__ level of significance, the mean money spent per day _is not_ significantly less than \$15 (t = _____, p _____, df = 29).

3a. Write the hypotheses: $H_0 : p = 0.5$, $H_1 : p > 0.5$, Sample proportion $(\hat{p})$ = $\dfrac{118}{179} = 0.659$

3b. Test the hypothesis:

$z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1 - p)}{n}}}$ $\quad z = \dfrac{0.659 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{179}}}$ $\quad$ p < 0.0002 or 1.02×10^{-5}

Formula Substitution Test Statistic p-value

Write the concluding sentence: At the 5% level of significance, the average price on Tuesday is not significantly less than other days (t = -0.479, p > 0.25, n = 7).

5a. Write the hypotheses: $H_0 :$ _ $\mu_{40} = \mu_{60}$ ___, $H_1 :$ __ $\mu_{40} < \mu_{60}$ __,

5c. Write the concluding sentence: At the 10% level of significance, the mean guess at Morocco's population by people with low phone digits is significantly less than the mean guess of those with high phone digits (t = -1.835, $p < 0.05$, n_{40} = 15, n_{60} = 15).

6c. Test the hypothesis: Test Statistic = *-12.41,* p-value = *p =1*

7e. Write a concluding sentence. At the 5% level of significance, the proportion of 12th grade students using drugs in 2012 is not significantly greater than in 2002 (z = 0.819, p = 0.2061, n_{2012} = 630, n_{2002} = 2184).

8a.

Research Design Table		
Research Question: Is there a significant difference in the mean times of the men and women who finish the triathlon course?		
Type of Research	Observational Study Observational Experiment Manipulative Experiment	
What is the response variable?	heart rate	
What is the parameter that will be calculated?	Mean Proportion Correlation	
List potential confounding variables		
Grouping/explanatory Variables 1 (if present) Gender	Levels: Men, Women	

8b. Write the hypotheses. $H_0 : \mu_{\text{men}} = \mu_{\text{women}}$, $H_1 : \mu_{\text{men}} \neq \mu_{\text{women}}$

9a. Conclusion: At the 5% level of significance, there is not a significant difference in the proportion of days that African Americans and LGB individuals record at least one stigma-related stressor (z = 0.062, $p = 0.950$, n_{AA} = 190, n_{LGB} = 310)

9b. Conclusion: At the 5% level of significance, the mean psychological distress score for those using rumination is significantly different than for those using distraction ($t = 2.189$, $p = 0.033$, $n_R = 26$, $n_D = 26$)

Chapter 6

1. 0.778

Margin of Error 0.060 Confidence Interval (0.718,0.838) Calculator confidence interval (0.71853,0.83823)

2. Girls: 0.743, Boys 0.778

Margin of Error 0.132, Confidence Interval (9-0.167,0.097)

3. (Note: There are 112 degrees of freedom. This df does not appear in your tables. It falls between 60 df and 120 df. To make sure that the interval is sufficiently large, the critical t value for 60 df will be used. The actual value, as found using the Excel function T.INV.2T(0.1,112) is 1.6586).

Margin of error 14.6 (73.4,102.6)

Calculator confidence interval (73.49,102.51)

5. Point estimate 156

Margin of error 92.9 Confidence Interval (63.1,248.9)

Calculator confidence interval (63.087,248.91)

6. Point Estimate 5.2 degrees

Margin of error 3.2 confidence interval (2,8.4)

Calculator confidence interval (2.0338,8.3662)

8. Point estimate 0.116

Margin of Error 0.017, Confidence interval (0.099,0.133)

9.

Margin of Error	1%	5%	10%	20%
Sample Size	9604			

10.

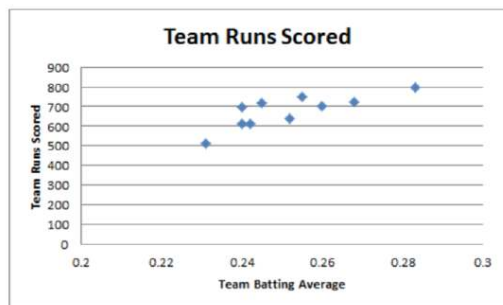| Degree of Confidence	99%	95%	90%	80%
Sample Size	1844			

11a. Stratified

11b. 256 , 379 ,

11d. Mean = 59.5, SD = 23.78

11e. (48.4, 70.6)

11f. Lowest: 290.4

Chapter 7

1b Mean batting average <u>0.2516</u> Standard deviation for batting average <u>0.0155</u>

Mean runs scored <u>676.1</u> Standard deviation for runs scored <u>82.20</u>

1c. At the 5% level of significance, there is a significant correlation between batting average and runs scored ($t = 3.84$, $p < 0.01$, $n = 10$).

1d. Regression equation: y = -397.98 + 4269x

1e. $r^2 = 0.6479$. It means 64.8% of the total variation from the mean for team runs scored is attributed to the variation in the batting average.

1f. 669.285

1g.
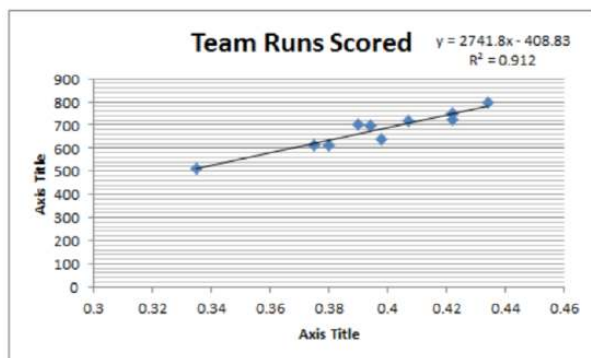
Correlation: 0.955

Hypothesis test concluding sentence: At the 5% level of significance, there is a significant correlation between slugging percentage and runs scored ($t = 9.10$, $p = 1.699\text{E-}5$, $n = 10$).

Regression equation: $y = -408.83 + 2741.80x$

Coefficient of determination (r^2): 0.912

Predict the number of runs scored for a team with a slugging percentage of 0.400. 687.89

2a.

State Budget Spending $y = -0.1615x + 6.0561$ $R^2 = 0.1847$

2b.

State Budget Spending $y = 0.0067x + 3.0503$ $R^2 = 0.0013$
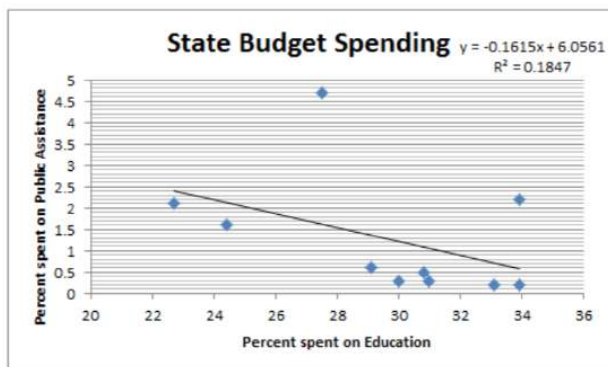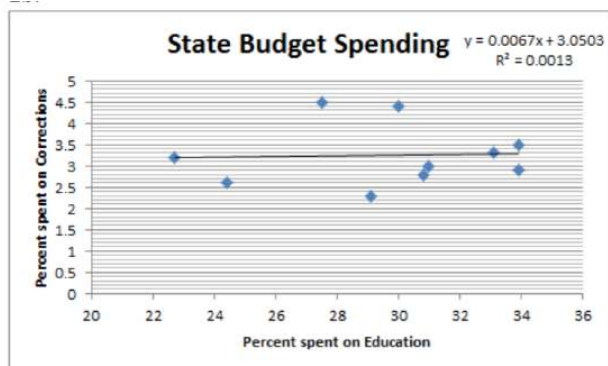
$R = 0.0359$, $r^2 = 0.0013$, $y = 3.05 + 0.0067x$

At the 5% level of significance, there is not a significant correlation between spending on education and spending on public assistance (t = -1.35, $p = 0.215$, $n = 10$).

3.

State Population and Median Income

$y = 156.21x + 49076$
$R^2 = 0.0094$

$Y = 49075.6 + 156.2x$ $r = 0.097$, $r^2 = 0.0094$, $t = 0.258$, $p = 0.8038$.

4a.

Total Patents 2000-2011 $y = 0.0023x - 36.378$
$R^2 = 0.8632$

$Y = -36.38 + 0.0023x$ $r = 0.929$, $r^2 = 0.863$, $t = 6.64$, $p = 2.92 \times 10^{-4}$

4b. Provide your own thoughtful response.

4c. 101.62

5a. Age

5b. NPA

5c. 2, 12, 22, 32, 42, 52, 62, 72, 82, 92, 102

5e. r = 0.814

5f. At the 5% level of significance, there is a significant correlation between age and NPA ($t = 4.2$, $p = 0.002$, $n = 11$).

Chapter 8

1a. Test for Homogeneity

1b. Goodness of Fit
1c. Test for Independence

1d. Goodness of Fit

2.

$X = x$	0	1	2	3
$P(X = x)$	0.5787	0.34722	0.06944	0.00463

Goodness of Fit

$\chi^2 = 3.43$

At the 5% level of significance, there is not a significant difference between the observed and expected distributions ($\chi^2 = 3.43$, $p > 0.1$, $n = 158$). Calculator p-value is 0.33.
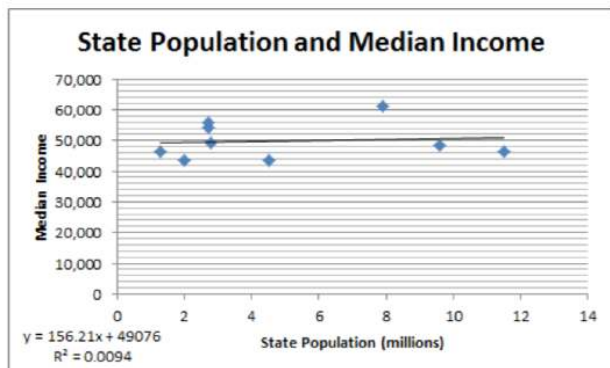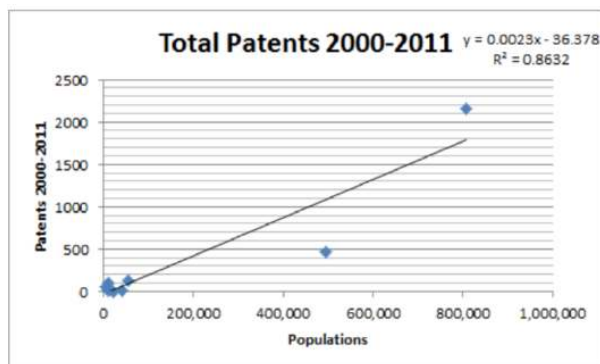
3. Test for Independence
$\chi^2 = 13.27$
At the 0.1 level of significance there is a correlation between shots and goals ($\chi^2 = 13.27$, $p < 0.005$, $n = 49$) (calculator p-value $= 2.696 \times 10^{-4}$).

4. Test for Homogeneity

At the 0.1 level of significance, the distributions for improvement from drug and non-drug treatments are homogeneous ($\chi^2 = 0.8222$, $p > 0.1$, $n = 80$) (Calculator: $p = 0.6629$, df=2)
5. $\chi^2 = 9.426$

6a. stratified

6b. 2042, 584, _____

6c. Goodness of Fit

6d.

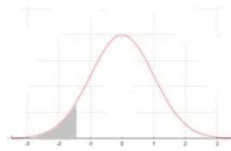Comparison of Observed and Expected Racial Distribution in WA Prisons
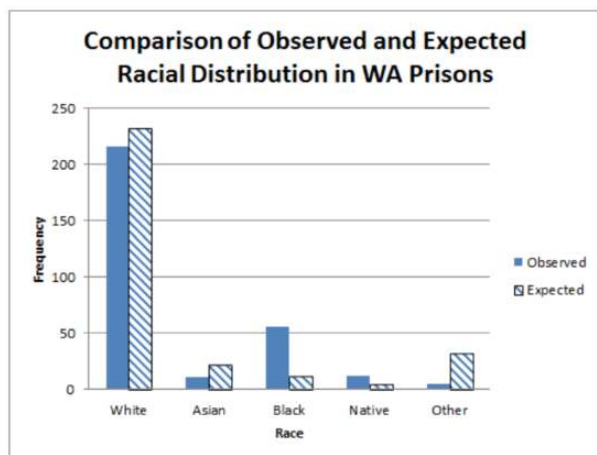
6f. At the 5% level of significance, the racial distribution of WA prisons is significantly different than what would be expected ($\chi^2 = 229.96$, $p < 0.005$, $n = 300$) (Calculator: $p = 1.3 \times 10^{-48}$, df = 4).

Tables

Standard Normal Distribution - N(0,1)

Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0391	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119

-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3493	0.3520	0.3557	0.3594	0.3532	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.24960	0.5000

Standard Normal Distribution – N(0,1)
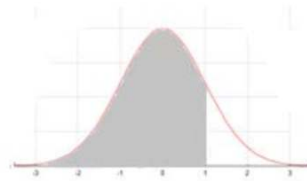
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5439	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.82=340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8505	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9541	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9727	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Student t distributions

	One Tail Probability	Two Tail Probability	Confidence Level						
One Tail Probability	0.04	0.025	0.01	0.05	0.025	0.01	0.005	0.0005
Two Tail Probability	0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.001
Confidence Level	20%	50%	80%	90%	95%	98%	99%	99.9%
df								
1	0.325	1.000	3.078	6.314	12.706	31.821	63.656	636.578
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	31.600
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	4.140

15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	-/257	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.2565	0.685	1.319	1.714	2.069	2.500	2.807	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.689
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.660
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.373
z^*	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.290

Chi-Square Distributions

Area Left	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
Area Right	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
df										
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	14.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	32.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37,689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.258	118.498	124.342	129.561	135.807	140.170
110	75.550	78.458	82.867	86.792	91.471	129.385	135.480	140.916	147.414	151.948

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Foundations in Statistical Reasoning (Kaslik)

Webpages: 29

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 86.2% (25 pages)
- Undeclared: 13.8% (4 pages)

By Page