Learning Objectives:

- Construct and interpret an interval estimate for a population parameter.
- Understand the difference between point and interval estimates.
- Identify appropriate parameters to estimate based on the data.
- Understand the relationship between margin of error and sample size.

A **point estimate** is the single value of a sample statistic used to estimate a population parameter.

Examples:

Although point estimates are extremely useful, they provide no feeling of **the estimate is**.

An **interval estimate** is bounded by two values, calculated from the sample data, used to estimate a population parameter. Rather than using a single value, we will use the point estimate as the center **and create upper and lower bounds where we are confident the true parameter value will reside.**

When we compute an interval estimate, we do so in such a way that we have a **that the parameter value will be in the interval. The level of confidence is the probability that:**

The actual interval is referred to as a **.**
Chapter 11
Inferences Regarding Proportions

IMPORTANT CONCEPT
A confidence interval is ALWAYS used to estimate a population parameter. It is NEVER used to estimate a sample statistic. Sample statistics are used to construct the confidence interval.

Recall from the previous chapter that under the right conditions we know $\hat{p} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$.

Quick Review: What are those conditions?

We can use this knowledge to help build a confidence interval for $\pi$.

The general form of all confidence intervals is: Point Estimate $\pm$ Margin of Error.

The margin of error for proportions is determined by a combination of:

1. 
2. 

A confidence interval for $\pi$ should look like: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}$

What is $\sqrt{\frac{\pi(1-\pi)}{n}}$?

What is $z_{\alpha/2}$?
Using \( \hat{p} \pm \frac{z}{\sqrt{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \) creates a problem. What is it? (Hint: what is it that we are trying to estimate?)

To avoid that problem, our confidence interval will take the form:

**Assumption:** \( \hat{P} \sim Normal \).

Reasonable if both the number of success and the number of failures is at least _____ and the sample size is at least __________.

Mathematically, we are saying:

**Example:** Campus administration is proposing that classes which meet 4 days each week be changed such that they only meet two days each week, but for a longer period. The proposal suggests this could increase the number of classes available to students allowing them to complete their degrees in less time. A survey of 137 randomly selected students was asked if they were in favor of the proposal. Of the 137 students surveyed, 114 said they were in favor. Estimate, with 95% certainty, the true proportion of students on campus who are in favor of the proposal.

1. Identify the parameter of interest.

2. Identify the probability distribution to be used addressing all applicable assumptions and how they are satisfied.

3. Construct the confidence interval.
4. Interpret the interval in English.

Calculating Sample Size for Proportions

Before we go take a sample, we must first know what we will obtain.

To calculate an appropriate sample size for a proportion, you must first know two things:

1. 

2. 

The quantity $z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is known as 

The value of z is specified by the 

To calculate the sample size, simply solve $z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = ME$ for $n$.

The resulting equation for sample size has a minor problem. What is it?

To “fix” this problem we can always use a value of ______ for ________.

Why is this a “safe” value?
Sample size as a function of the sample proportion.

\[ n = \left( \frac{z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})}}{ME} \right)^2 \]

Calculate the needed sample size for a proportion given the following criteria:

<table>
<thead>
<tr>
<th>Margin of Error (ME)</th>
<th>Level of Significance ((\alpha))</th>
<th>Level of Confidence (1 - (\alpha))</th>
<th>Sample Size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What happens as the margin of error gets smaller while maintaining the same level of confidence?

What happens as the level of significance gets smaller while maintaining the same level of confidence?
Chapter 11 – CI for the Mean

IMPORTANT CONCEPT
A confidence interval is ALWAYS used to estimate a population parameter. It is NEVER used to estimate a sample statistic. Sample statistics are used to construct the confidence interval.

Estimation of the Mean when the Standard Deviation is Known

Data types applicable: interval or ratio

Consider the Speed of Light data listed in this chapter.

Consider the following data consisting of time (in minutes) reported for a pain medication to take effect and estimate the true average time, with 95% certainty, based on a believed value of 2.8 for the population standard deviation. Assume the population is normal.

| 15.2 | 18.6 | 10.2 | 15.9 | 18.1 |
| 13.2 | 14.1 | 19.7 | 17.6 | 14.8 |
| 11.5 | 14.8 | 16.6 |     |     |

Quick Review: A confidence interval has the basic form ________________________________.

From the previous chapter we know the sample means, under the right conditions, will be distributed normally.

\[ \bar{X} \sim N(\mu, \ ) \].

What are the “right conditions”?

1.

2.

Construct a 95% confidence interval.

The confidence interval will look like:

You can also use your TI-83/84 to find this interval. Go to: Stat>Tests>7: ZInterval
Estimation of the Mean when the Standard Deviation is Unknown

Data types applicable: ________________________.

(use same data as previous example)

Realistically, it is seldom, if ever, that we actually have the value of $\sigma$ without already having the value of $\mu$. Recall the formula for finding $\sigma^2$.\[ \sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \]

If we do not know $\sigma$, which we typically will not, our best point estimate for $\sigma$ is ______, the ________________________.

The confidence interval should then have the form: $\bar{x} \pm z \frac{S}{\sqrt{n}}$. However, thanks to beer and a guy named Gossett (England 1876 – 1936) we know that the standard normal distribution ($z$) is not appropriate when estimating the mean with small sample sizes and unknown __________. The distribution will be that of a ________, not a _______ distribution.

**The Student-T Distribution**

1. Bell shaped. Looks like the normal distribution.
2. The t-distribution is actually a ______________ of curves. Each curve being defined by the degrees of freedom (df). The degrees of freedom are simply \( n-1 \). Recall the equation for the sample variance:

\[
s^2 = \sum_{i=1}^{n} \left( x_i - \bar{x} \right)^2
\]

The numerator is the degrees of freedom.

3. As the sample size increases, the student t-distribution __________________________

__________________________________________________________________________.

Our confidence interval for the mean, when the population standard deviation is unknown will then have the form: __________________________, provided the necessary assumptions can be reasonably shown.

**Assumptions for using the Student-T Distribution:** The data was sampled from a normally distributed population.

The best way to make this determination is to look at __________________________. If the sample is _______________ and the sample is representative of _______________ then the population must also be____________________.

If the sample size is large enough (approximately 30) then the CLT will apply and we need not be concerned about the normal plot.
Construct a Normal Plot for the Pain Relief Data (with and without the negative values)

What are we looking for in a normal plot?

Construct a 95% CI for _____ based on the t-distribution. Stat>Calc>8:TInterval

Calculating the needed sample size for inferences regarding ________.

\[
ME = t_{df, \alpha/2} \frac{s}{\sqrt{n}} \quad \Rightarrow \quad n = \frac{t_{df, \alpha/2} s}{ME} \quad \Rightarrow \quad n = \left( \frac{t_{df, \alpha/2} s}{ME} \right)^2
\]

But there are two problems with this calculation. What are they, and how do we fix them?
Suppose you are about to conduct research to determine the average time it takes for a pain reliever to take effect. The doctor wants to be certain within 3 minutes with 98% certainty. What sample size is needed?
Chapter 11 – CI for the Median

IMPORTANT CONCEPT
A confidence interval is ALWAYS used to estimate a population parameter. It is NEVER used to estimate a sample statistic. Sample statistics are used to construct the confidence interval.

Estimation of the Median

A confidence interval for the median is typically needed when:

1. The assumptions for the mean can not be reasonably satisfied, or
2. The research requirements dictate the median.

Consider the following data, consisting of the time minutes) reported for a pain medication to take effect ad estimate the true average time, with 95% certainty.

<table>
<thead>
<tr>
<th>15.2</th>
<th>18.6</th>
<th>10.2</th>
<th>15.9</th>
<th>18.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.2</td>
<td>14.1</td>
<td>19.7</td>
<td>17.6</td>
<td>14.8</td>
</tr>
<tr>
<td>11.5</td>
<td>14.8</td>
<td>16.6</td>
<td>31.8</td>
<td>46.5</td>
</tr>
</tbody>
</table>

Proceed with the process of completing the appropriate confidence interval.
In a case where we want to construct a confidence interval for $\mu$, but cannot due to a violation of the necessary assumptions, consider a confidence interval for the median.

Our text book presents two cases, the large sample and the small sample case. We will look only at the small sample case.

**Outline The Procedure Here**
Chapter 12
Introduction

Learning Objectives

1. Understand the logic behind a hypothesis test for a population parameter.
2. Construct and interpret a test of hypothesis for a population parameter.
3. Identify appropriate parameters to test based on the data.
4. Comprehend the types of statistical errors that can be made in the decision process for a hypothesis test.

Logic Behind Hypothesis Testing

- Univariate hypothesis testing is a technique that allows us to compare the value of an observed sample statistic with a

- Suppose you believe a new treatment for a particular form of cancer will increase survival rate. The current average survival is 2.25 years. The average from your test of 13 subjects is 3.1 years. Is your treatment statistically better?

- A univariate hypothesis test will give us a way to make that determination, based on probabilities. The technique will help us decide if the difference between the sample statistic (3.1 years) and the believed parameter value (2.25) is simply by ____________, or if it is probably ____________.

Hypothesis Statement

- A hypothesis statement is a statement that something is believed to be true.

- A ________________________ is a process by which a decision is made between two opposing hypotheses.

- Those two opposing hypotheses are known as the _____ hypothesis and the hypothesis.
The null hypothesis, denoted by \( H_0 \), is a statement of “\( \quad \)”. 

In the cancer example we will assume the new treatment is no better than the old and look for contradicting evidence.

The null hypothesis is then:

The alternative hypothesis, denoted by \( H_A \), is the “\( \quad \)”. 

This is the hypothesis the researcher is typically trying to suggest is \( \quad \) than the null hypothesis.

For the cancer example, we would write the statement as:

Notice the value of the sample statistic \( \quad \) show up in the null or the alternative hypothesis. Nor does the symbol for any sample statistics.

Hypothesis testing is about \( \quad \) not sample statistics.

A hypothesis test is an indirect proof. Researchers gather data in hopes that data will contain sufficient evidence to suggest the null hypothesis is \( \quad \). If the null hypothesis is shown to be unlikely then the alternative hypothesis, the only other choice, is therefore \( \quad \).

Keep in mind that we cannot “prove” anything with a hypothesis test. Rather, we hope to \( \quad \) the likelihood based on probabilities.

**Types of Errors**

When you complete a hypothesis test you will always either \( \quad \) or \( \quad \) the \( \quad \) hypothesis based on the test statistic.

Bottom line is that the null hypothesis is, in reality, either \( \quad \) or \( \quad \).
<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Decision</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fail to Reject</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A _________ is the probability of observing a test statistic at _______ as extreme as observed.

A p-value can be thought of as the calculated probability of a ______________ error.

**Example:** When a parachute is inspected, the inspector is looking for evidence the chute will not open.

H₀:

Hₐ:

Type I Error:

Type II Error:

Which is more serious? Why?

**Example:** When an ambulance crew arrives at the scene of a serious auto wreck, they are looking for evidence the injured person has died. If they believe the person is alive, they will continue working on them.

H₀:

Hₐ:

Type I Error:
Type II Error:

Which is more serious? Why?

Recall our cancer example: Suppose you believe a new treatment for a particular form of cancer will increase survival rate. The current average survival is 2.25 years. The average from your test of 13 subjects is 3.1 years. Is your treatment statistically better?

This is an example of a ____________ tail test.

Left tail test:

Two tail test:
Chapter 12
Hypothesis Test for the mean, median and proportion

Consider the following data, consisting of the time it took 15 students to complete an in class quiz. Is there evidence to suggest the “typical” time it takes to complete the quiz is greater than 10 minutes?

<table>
<thead>
<tr>
<th>5.5</th>
<th>7.5</th>
<th>9.5</th>
<th>10.75</th>
<th>11.25</th>
<th>11.75</th>
<th>12.25</th>
<th>12.5</th>
<th>12.75</th>
<th>13.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.25</td>
<td>13.5</td>
<td>13.5</td>
<td>13.75</td>
<td>14.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We will use this data to complete examples of hypothesis testing for both the mean and the median; however, note that only one procedure will actually be correct.

Steps for Completing a Hypothesis Test

- **Step 1:** Identify the ________________ for which we wish to infer.
- **Step 2:** State the ________ and ____________ hypothesis.
- **Step 3:** Identify the appropriate______________.
- **Step 4:** State the ________________ and how you feel they have been reasonably satisfied.
- **Step 5:** Calculate the______________.
- **Step 6:** State your______________.
- **Step 7:** State your ____________ in English (an English statement).
Complete a hypothesis test for the quiz time data (median).

- Step 1:
- Step 2:
- Step 3:
- Step 4:
- Step 5:
- Step 6:
- Step 7:
Complete a hypothesis test for the quiz time data (mean).

• Step 1:

• Step 2:

• Step 3:

• Step 4:

• Step 5:

• Step 6:

• Step 7:
A Hypothesis Test for the Proportion

A community college is investigating the proportion of students who complete their two year degrees and continue on to attend a four year school. Historically that proportion has been 23%. Suppose a survey of 137 recent graduates showed 44 had gone on to a four year school. Do the data suggest the number of students going on to a four year school has increased?

• Step 1:
• Step 2:
• Step 3:
• Step 4:
• Step 5:
• Step 6:
• Step 7:
Chapter 13
Comparing Two Population Parameters
(Introduction)

Learning Objectives

1. Understand the difference between independent and dependent samples.
2. Construct and interpret a test of hypothesis comparing two population parameters.
3. Identify appropriate parameters to compare based on the data.
4. Comprehend the difference between practical and statistical significance.

Dependent vs. Independent Data

Before we can look at techniques used for the comparison of two samples, we must first
discuss the meaning of _______________ and _______________ samples.

In general, if the same source is used to obtain the two samples, then the data is thought
of as being _______________(related).

The classic dependent data example is that of a “___________________________”
situation or a “before” and “after” situation.

The idea is that there is a natural _______________ of the data.

If two different sources are used, then the data is thought of as being _______________,
(not related).
Dependent Data Example:

<table>
<thead>
<tr>
<th>Item</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albertsons</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Independent Data Example:

Understanding the difference between dependent and independent data is crucial for the comparison of two population parameters when we are primarily interested in the variances and/or means because the proper technique is dictated by the type of data you are working with.
Chapter 13
Comparing Two Population Parameters
(Ratio of Two Variances)

Data types applicable: *interval, ratio*.

Having reliable information regarding the mean or median can often be very important; however, information regarding ________________ can, at times, be even more important.

**Example:** Pepsi Corporation considering two different bottling techniques/machines.


<table>
<thead>
<tr>
<th>Current Technique</th>
<th>10</th>
<th>17</th>
<th>20</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>12</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>20</td>
<td>13</td>
<td>21</td>
<td>10</td>
<td>14</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>New Technique</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>14</td>
<td>15</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
<td>16</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

**Example:** A furniture manufacturer is concerned about the number of blemishes on furniture coming out of the factory. A test run is conducted where the number of blemishes on a specific piece of furniture is recorded for 15 consecutive work shifts both from the currently used assembly process and a new process being tested. Is there evidence to suggest the new process reduces both the typical number of blemishes along with the variability observed in the number of blemishes?

Before we address the means, we will first address the variances. The test we will be using is based on the ________________.
Properties of the F Distribution
1. The F-distribution is always \( \text{(zero or positive)} \).
2. The F-distribution is not symmetric, it is skewed \( \text{__________________________} \).
3. Like the t-distribution, the F-distribution is a \( \text{__________________________} \) distributions where each member is identified by the \( \text{__________________________} \).

Assumptions
Like other statistical procedures, a hypothesis test based on the F-distribution has assumptions that must be reasonably satisfied prior to employing the procedure. Before setting up the hypothesis test and calculating the p-value, we should first check the needed assumptions.

1. the sample data was obtained from two independent populations, and
2. the distributions of both populations are normal.

\[
H_0 : \frac{\sigma_0^2}{\sigma_N^2} = 1 \quad H_A : \frac{\sigma_0^2}{\sigma_N^2} > 1 \quad \alpha = \]

Calculator Commands: Stat > Tests > 2-SampFTest

p-value:

Decision:

Conclusion:
Chapter 13
Comparing Two Population Parameters
(Independent Means)

Data types applicable: *interval, ratio*.

There are two main scenarios concerning the difference between two independent means.

The first is when \( \sigma \) is known (use a \_________\).

The second is when \( \sigma \) is unknown (use a \_________\).

Because it is very rare that we truly know \( \sigma \), we will concentrate on the \_________ \_________.

**Example:** A furniture manufacturer is concerned about the number of blemishes on furniture coming out of the factory. A test run is conducted where the number of blemishes on a specific piece of furniture is recorded for 15 consecutive work shifts both from the currently used assembly process and a new process being tested. Is there evidence to suggest the new process reduces both the typical number of blemishes along with the variability observed in the number of blemishes?

<table>
<thead>
<tr>
<th>Current Technique</th>
<th>10</th>
<th>17</th>
<th>20</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>12</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>20</td>
<td>13</td>
<td>21</td>
<td>10</td>
<td>14</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>New Technique</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>14</td>
<td>15</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
<td>16</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

**Assumptions**

1. the sample data was obtained from two independent populations, and
2. the distributions of both populations are normal.

Are these two data sets independent of each other? Justify why or why not:

Normality assumption: (does the CLT apply?)
The two-sample t-test has two different forms. In order to determine which form we need to use, we must determine if the variances from each group “are ________________”. If they are similar then we will ________________ the variances. If they are not similar then we will not pool them.

To determine if they are similar, use an _____ test.

\[ H_0 : \quad H_A : \quad \alpha = \]

p-value:                   Decision:

Conclusion:

Once this has been decided, continue with the two-sample t-test.

\[ H_0 : \quad H_A : \quad \alpha = \]

**Calculator Commands:** Stat > Test > 2-SampTTest

p-value:                   Decision:

Conclusion:
Chapter 13
Comparing Two Population Parameters
(Wilcoxon Rank-Sum)

Data types applicable: *interval, ratio*.

The Wilcoxon Rank-Sum test is an example of a ____________________________ test.

Advantages of non-parametric tests:

Major disadvantage:

Let’s continue using the scenario as the previous example for the Wilcoxon Rank-Sum Test. The Current Technique data is the same, the New Technique data has been changed.

**Example:** A furniture manufacturer is concerned about the number of blemishes on furniture coming out of the factory. A test run is conducted where the number of blemishes on a specific piece of furniture is recorded for 15 consecutive work shifts both from the currently used assembly process and a new process being tested. Is there evidence to suggest the new process reduces both the typical number of blemishes along with the variability observed in the number of blemishes?

<table>
<thead>
<tr>
<th>Current Technique</th>
<th>10</th>
<th>17</th>
<th>20</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>12</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>20</td>
<td>13</td>
<td>21</td>
<td>10</td>
<td>14</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>New Technique</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>9</td>
<td>10</td>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

We would start by wanting to use the ____________________________ test.

**Assumptions**

1. the sample data was obtained from two independent populations, and
2. the distributions of both populations are normal.

Normality assumption: (does the CLT apply?)

Look at normal plots and assume, for this example, one or both **were not** normal.
\[ H_0 : \quad H_A : \quad \alpha = \]

**Calculator Commands: PRGM > RANKSUM**

p-value: \hspace{1cm} Decision:

Conclusion:

**Note:** If you can show the shapes of the distributions are \underline{___________} then the Wilcoxon Rank-Sum will allow you to continue with a hypothesis test for the \underline{_________}. If the shapes are not similar then we must switch to a hypothesis for the difference between the two \underline{__________}. 
Inferences Regarding the Difference Between Two Distribution Centers

Want to use t-distribution. Decide if independent or Dependent samples.

Independent
Check Assumptions:
Both data sets came from normal parent distributions.

Normal
Determine pool or not based on F-test

Complete 2-sample t-test. Hypothesis is about the mean.

Not Normal
Complete Wilcoxon Rank Sum Hypothesis is about the mean/median, depending on distribution shapes.

Dependent
Check differences for normality.

Normal
T-test on the differences. Hypothesis is about the mean.

Not Normal
Sign test on the differences. Hypothesis is about the median.
Chapter 13
Comparing Two Population Parameters
( Dependent Data )

Data types applicable: interval, ratio.

The dependent case is the easiest case of all.

Mechanics:

1. Determine if data is independent or dependent.
2. If dependent, subtract one from the other. The order you subtract will be important in determining the null and alternative hypothesis.
3. Complete a one sample case based on the differences where the null hypothesis will always be:

4. The alternative hypothesis will depend on the order you did your subtraction.

Example: After completing an exam, the professor was not satisfied with the results. As such he decided to spend an extra day on review and give every one a similar exam in hopes of increasing grades. The data below represents the scores both before and after. Do the data suggest the extra day of review helped to increase the exam scores?

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Exam</td>
<td>32</td>
<td>71</td>
<td>68</td>
<td>18</td>
<td>91</td>
<td>88</td>
<td>73</td>
<td>62</td>
<td>58</td>
<td>66</td>
<td>70</td>
</tr>
<tr>
<td>Second Exam</td>
<td>41</td>
<td>77</td>
<td>61</td>
<td>15</td>
<td>90</td>
<td>92</td>
<td>81</td>
<td>73</td>
<td>47</td>
<td>70</td>
<td>76</td>
</tr>
<tr>
<td>Differences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Chapter 13
Comparing Two Population Parameters
(Difference of Proportions and Odds Ratios)

Data types applicable: nominal.

Mechanically, completing a hypothesis test for the difference of proportions is very much like doing a hypothesis test for a single proportion, twice.

Example: A recent survey of 273 men and 291 women resulted in 192 men and 214 women being in favor of changing the college from a semester system to a quarter system. Is there evidence to suggest more women are in favor of the change than men?

Assumptions: _____ and _____ are both distributed normal.

The assumptions are reasonable if:

Complete the hypothesis test:
**Odds Ratio**
There are times when the assumptions for proportions are violated or the use of an odds ratio has more meaning.

Consider the same scenario, but place the data in a 2x2 table.

<table>
<thead>
<tr>
<th></th>
<th>In Favor</th>
<th>Not In Favor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Males</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Females</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 14
One-Way ANOVA
(Analysis of Variance)

Learning Objectives

1. Compare more than two population means using the ANOVA procedure.
2. Use a multiple comparison procedure to infer which population is different.
3. Identify appropriate parameters to test and compare based on the data.

If we want to compare means from two independent groups we can use a ____________.

How would you accomplish this for 3 independent groups?

How about 5 independent groups?

What would the error be (α) if we did t-tests for 5 independent groups with α = 0.05 for each group?

\[ 1 - (1 - \alpha)^5 = 1 - (1 - \alpha)^5 = \______________ \]

Is this acceptable?
ANOVA is a technique which allows us to make comparisons between _________ independent groups.

Specifically we will be working with a technique known as ________________________________.

It is called one-way ANOVA because the data has multiple levels of the same variable.

**Assumptions for ANOVA**

1. Each sample has been randomly and independently selected from the population is represents.

2. The parent distributions (populations) are all normal.

3. The variances of the $k$ samples are all equal (similar).

   We can satisfy this assumption with ________________________________.

**Example:**

Three different types of feed were tested on rates over a 30 day period. The data recorded is weight gain for each of the 30 rates in the study. Do the data suggest a difference in weight gain between the three diets?

<table>
<thead>
<tr>
<th>Diet-1</th>
<th>Diet-2</th>
<th>Diet-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>118</td>
<td>120</td>
<td>111</td>
</tr>
<tr>
<td>117</td>
<td>108</td>
<td>96</td>
</tr>
<tr>
<td>111</td>
<td>105</td>
<td>93</td>
</tr>
<tr>
<td>107</td>
<td>102</td>
<td>90</td>
</tr>
<tr>
<td>104</td>
<td>102</td>
<td>86</td>
</tr>
<tr>
<td>102</td>
<td>98</td>
<td>84</td>
</tr>
<tr>
<td>100</td>
<td>96</td>
<td>80</td>
</tr>
<tr>
<td>87</td>
<td>94</td>
<td>75</td>
</tr>
<tr>
<td>81</td>
<td>91</td>
<td>72</td>
</tr>
<tr>
<td>73</td>
<td>79</td>
<td>54</td>
</tr>
</tbody>
</table>

$H_0$:  

$H_a$:  

$\alpha$

**Address the assumptions:**

1. Independence
2. Normality Assumption

3. Similar Variances

4. Complete ANOVA

5. Decision:

6. Conclusion:
A Nonparametric Approach

In the event we have a violation of our assumptions, such as non-normality. We can complete a test known as the ________________. This test is based on ________________ of the data.

When the Kruskal-Wallis test is used, we must modify our hypothesis statement to read:

\[ H_0 : \]
\[ H_A : \]
\[ \alpha \]
Chapter 15
Categorical Data Analysis
Goodness of Fit

Learning Objectives

1. Identify and understand applications using count data.
2. Use the Chi-Squared distribution to compare two distributions.
3. Use the Chi-Squared distribution to determine if a relationship exists between two qualitative variables.

\( \chi^2 \) Goodness of Fit Test

Data types: \underline{__________________} or \underline{__________________}.

The data consists of \underline{________ data} meaning data values represent the number of times a specific category was observed.

The goodness of fit test will help us decide if the observed data fits a believed distribution.

This is based on what are known as \underline{________________} experiments.

A multinomial experiment is any experiment which satisfies the following:

1. The experiment consists of \underline{________} identical and \underline{________} trials.
2. The outcome of each trial is classified into one, and only one, of \underline{________} different categories.
3. The data consists of count data, the number of \underline{________} in each of the \( k \) categories.
4. The probabilities associated with the \( k \) outcomes, denoted by \( \pi_1, \pi_2, \ldots, \pi_k \) remain constant for all trials. \( \sum_{i=1}^{k} \pi_i = 1 \).

Properties of the \( \chi^2 \) distribution and the basis for the test statistic.

1. The \( \chi^2 \) distribution is a family of distributions determined by the degrees of freedom.
2. The \( \chi^2 \) distribution is a positive distribution (similar to the F-distribution).
\[ \chi^2 = \sum_{i=1}^{k} \]

Where \( E_i = n \pi \),
**Example:** A random sample of students on a community college campus was selected as part of a survey process. Those conducting the survey wanted to determine if the sample was representative of the various ethnicities on campus. Based on published data, the true proportion for each ethnicity was obtained. Does it seem reasonable to believe the sample was ethnically representative of the campus?

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Pacific Islander</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed</strong></td>
<td>53</td>
<td>27</td>
<td>42</td>
<td>18</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td><strong>True proportions (π_i)</strong></td>
<td>0.35</td>
<td>0.12</td>
<td>0.31</td>
<td>0.05</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Expected Values</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$\chi^2 = \sum_{i} \frac{(O_i - E_i)^2}{E_i} =$$

Assumptions:

1. The experiment satisfies the properties of a multinomial experiment.
2. All expected values are at least 1.
3. No more than 20% of the expected values are less than 5.
Chapter 15
Categorical Data Analysis
Test of Independence

The data for this example is from HANES – Health and Nutrition Examination Survey of 1976-1980 – in which the Public Health Service examined a representative cross section of 20,322 Americans ages 1 to 74. The goal was to obtain baseline data regarding

• demographic variables such as age, education and income
• physiological variables such as weight, cholesterol levels
• prevalence of diseases

From the sample of 20,322 they took a subset of 2,237 Americans specifically in the 25-34 age group. One of the variables recorded in that sub sample was handedness. The table below represents the data collected. Is there evidence to suggest there is a difference of handedness based on gender?

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Right-handed</strong></td>
<td>934</td>
<td>1070</td>
</tr>
<tr>
<td><strong>Left-handed</strong></td>
<td>113</td>
<td>92</td>
</tr>
<tr>
<td><strong>Ambidextrous</strong></td>
<td>20</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Right-handed</strong></td>
<td>87.54%</td>
<td>91.45%</td>
</tr>
<tr>
<td><strong>Left-handed</strong></td>
<td>10.59%</td>
<td>7.86%</td>
</tr>
<tr>
<td><strong>Ambidextrous</strong></td>
<td>1.87%</td>
<td>0.68%</td>
</tr>
</tbody>
</table>

$H_0 :$

$H_A :$
Address the assumptions:

Enter the data (the count frequencies) in your calculator. Observed frequencies go in Matrix A.

Stat > Tests > $\chi^2$ – Test
The p-value is generated for you along with the matrix of Expected Values needed to address the assumptions.

### Expected Value Matrix

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right-handed</td>
<td>955.86</td>
<td>1048.1</td>
</tr>
<tr>
<td>Left-handed</td>
<td>97.781</td>
<td>107.22</td>
</tr>
<tr>
<td>Ambidextrous</td>
<td>13.355</td>
<td>14.645</td>
</tr>
</tbody>
</table>
Chapter 16
(Correlation and Simple Linear Regression)

Learning Objectives

1. Visually determine if a linear or curvilinear relationship exists between two quantitative variables.
2. Measure and interpret the strength of the linear or monotonic relationship.
3. Construct and interpret the regression equation relating two quantitative variables.
4. Make inferences concerning population parameters.

**Pearson’s Correlation Coefficient** is a measurement of ________________ association.
This type of measurement applies only to data which has a measurement scale of ________________ or ________________. The data is bivariate meaning:

<table>
<thead>
<tr>
<th>English</th>
<th>Parameter</th>
<th>Statistic</th>
<th>What does it Measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson’s Correlation Coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman’s Correlation Coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The value of the correlation coefficient is always on [ , ]. It is mathematically impossible to be:
A Few Examples:
Example:
Bill Caroscio
Elmira Southside High School
Elmira, NY
1997

Fishermen in the Finger Lakes Region have been recording the dead fish they encounter while fishing in the region. The Department of Environmental Conservation monitor the pollution index for the Finger Lakes Region. The data table below show this information for the past eight years.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>Pollution Index</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>2.5</td>
<td>147</td>
</tr>
<tr>
<td>1988</td>
<td>2.6</td>
<td>130</td>
</tr>
<tr>
<td>1989</td>
<td>8.3</td>
<td>210</td>
</tr>
<tr>
<td>1990</td>
<td>3.4</td>
<td>130</td>
</tr>
<tr>
<td>1991</td>
<td>1.3</td>
<td>114</td>
</tr>
<tr>
<td>1992</td>
<td>3.8</td>
<td>162</td>
</tr>
<tr>
<td>1993</td>
<td>11.6</td>
<td>208</td>
</tr>
<tr>
<td>1994</td>
<td>6.4</td>
<td>178</td>
</tr>
</tbody>
</table>

Step #1: Produce a scatter plot and decide if there appears to be a linear trend.

Step #2: Establish a hypothesis statement and test while simultaneously calculating r.

Step #3: State your decision:
Step #4: State your conclusion:

Spearman’s Correlation: A Nonparametric Approach

Spearman’s Correlation is really nothing more than Pearson’s Correlation computed on the ________________of the data. As such, it can be applied to ordinal data, such as a likert scale.

Spearman’s Correlation is a measurement of __________________________ association.

<table>
<thead>
<tr>
<th>English</th>
<th>Parameter</th>
<th>Statistic</th>
<th>What Does It Measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson’s Correlation</td>
<td>$\rho$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman’s Correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: A dental hygienist conducts research recording the ounces of sugar consumed weekly and the number of cavities observed among 12 year old children. Is there evidence to suggest a relationship between the amount of sugar consumed weekly and the number of cavities?

Step 1: Produce a scatter plot.

<table>
<thead>
<tr>
<th>Child</th>
<th>Sugar</th>
<th>Cavities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

Scatter Plot of the Ranks

<table>
<thead>
<tr>
<th>Sugar Ranks</th>
<th>Cavities</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

Step #2: Calculate the value of Spearman’s Correlation.
**Step #3:** Establish a hypothesis statement and test.

\[ t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \]

**Step #4:** State your decision.

**Step #5:** State your conclusion.
Causation vs. Association

Calculating Confidence Intervals
**Simple Linear Regression**

Simple Linear Regression is called “Simple” because it involves only one predictor variable, such as hours studied to predict course score.

Linear regression is the technique we use to model linear data with a linear model, i.e. the equation of a line.

The processes employed is referred to as Least Squares Regression.

Equation for the population regression line: \( y = \beta_0 + \beta_1 x + \varepsilon \)

Equation for the sample regression line:

**Assumptions:**

1. The data is bivariate normal.
   
   *Satisfied by:*
2. The error term, $\epsilon$, is a random variable that is distributed normally with a mean of zero and a constant variance.

*Satisfied by:

3. Any two values of $\epsilon$ are independent of each other.

*Satisfied by:

**Simple Linear Regression Example:** Find a linear model for the Finger Lakes data.

1. Determine the appropriateness of the model.
2. Use the model to estimate the number of dead fish if the pollution index were found to be:
a. 10.00
b. 13.75

3. Discuss the value and interpretation of $r^2$, the coefficient of determination.

**Step #1**: Produce a scatter plot.

**Step #2**: If a linear trend appears to exist, continue with the model building process.

**STAT > TESTS > LinRegTTest**