Math 10

Elementary Statistics and Probability

Course Note Pack

Lisa Markus  2016
Chapter 1: Sampling and Data - Notes

Statistics:

Descriptive:

Inferential:

Probability:

Key Terms:

*Example:* We are interested in the average number of units students in this class are taking in Summer 2016.

**Population:**

**Sample:**

**Parameter:**

**Statistic:**

**Variable:**

**Data:**
Data (actual values of a variable)

Qualitative Data:

Quantitative Data:

  Discrete:

  Continuous:

Sampling:

Note: can sample with replacement or without replacement

Simple Random Sample:

Stratified Random Sample:

Cluster Random Sample:

Systematic Random Sample:

Convenience Sampling:
**Example: Survey: How many siblings did you have?**

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fill out the table.
What percent in the survey have:

- a) Exactly 2 siblings?

- b) At most 2 siblings?

- c) More than 1 sibling?

- d) At least 2 siblings?
Example: Sixty (60) De Anza students were asked how many movies (at the movie theater) they saw last month. The data is summarized in the frequency table below.

<table>
<thead>
<tr>
<th>Number of Movies</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>0.1667</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>0.2333</td>
<td>0.4000</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Put the table into your calculator, then use your calculator to complete the table.

2. The number of movies is what type of data?

3. Approximately what percent of students saw at most 4 movies?

4. Approximately what percent of students saw at least 4 movies?

5. The data was collected at De Anza College by dividing all students into 10 groups according to their majors (some groups had more than one major) and then randomly selecting 6 students from each of the 10 groups. What type of sampling is this?

6. If we took another sample by getting the alphabetical listing of all De Anza students, randomly choosing the first student in the sample from the list, and then choosing every 350th student, what kind of sample would we have?

7. The average number of movies seen by students in the sample is an example of:
   A. Parameter       B. Data       C. Statistic       D. Variable

8. Let \( X = \) the number of movies (in a theater) De Anza students saw in the last month. \( X \) is the:
   A. Parameter       B. Data       C. Statistic       D. Variable
9. Identify the type of data:
   a. Number of students enrolled in Math 10.
   c. Distance to closest grocery store.
   e. Favorite movie.

10. De Anza Security is taking a survey of the number of people arriving in each car that will be parked in the parking structure. Name the sampling method used for each of the following.
   a. Survey every 10th car that enters the parking structure.
   b. Randomly pick 10 cars from each floor of the parking structure.
   c. Randomly pick one section of the parking structure and survey every car in that section.
   d. Survey the first 50 cars entering the parking structure.

11. A lawyer is interested in the average time it takes for bills to be paid by clients who pay bills.
   a. What is the population she is interested in?
   b. $X =$ the time it takes for ONE client to pay their bill. What is $X$ an example of?
   c. The lawyer takes her sample by gathering data on 10 randomly selected clients who have paid their bills. The lawyer's sample produces an average time to pay of 2 months. What is this value (2 months) is an example of?
   d. One particular client took 4 months to pay the bill. What is this value an example of?
**Stemplots**

*Example 1:* The number of minutes 20 internet subscribers spent on the internet during their most recent session:

```
50  40  41  17  11  7  22  44  28  21
19  23  37  51  54  42  88  41  78  56
```

Draw stem-and-leaf graph (also known as stem plot) Stems here are 0 – 8.

- good for small data sets
- quick to do
- shows all data values
- look for patterns/skewness
- look for outliers/extreme values
- look for spread
**Histograms**

<table>
<thead>
<tr>
<th>interval</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Horizontal axis:

Vertical Axis:

Number of bars:

Lowest Value:

Highest Value:

Width of bar:

Sample mean:

Sample Standard deviation:
Box Plots and Histograms

Consider the following data for the number of movies 50 students watched during Spring Break:

<table>
<thead>
<tr>
<th>number of movies</th>
<th>frequency</th>
<th>Relative frequency</th>
<th>Cumulative relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Fill in the relative frequency and the cumulative relative frequency.

2. Find:
   - lowest value:
   - highest value:
   - median:
   - Q1:
   - Q3:
   - IQR:

   This means the middle 50% of the students saw between ____ and ____ movies during spring break.

   Mean:

   sample standard deviation:

   Mode:

3. Find and interpret the 40th percentile and the 70th percentile.
4. Draw a histogram.

5. Draw a box plot.

6. Find any potential outliers, using an appropriate formula.

**POTENTIAL OUTLIERS:** values that lie more than 1.5 IQR above Q3 or more than 1.5 IQR below Q1

\[
\begin{align*}
Q3 + 1.5\times\text{IQR} \\
Q1 - 1.5\times\text{IQR}
\end{align*}
\]

7. Skewness?
**Example:** Let the population mean be 5 and the population standard deviation be 2.

1. How many standard deviations is 1 from the mean?

2. Find the value that is 2 standard deviations below the mean.

3. Khong, Megan and Jabbar are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools for a one mile run are given in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Running Time</th>
<th>School Average Running Time</th>
<th>School Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khong</td>
<td>4.9</td>
<td>5.2</td>
<td>.15</td>
</tr>
<tr>
<td>Megan</td>
<td>4.2</td>
<td>4.6</td>
<td>.25</td>
</tr>
<tr>
<td>Jabbar</td>
<td>4.5</td>
<td>4.9</td>
<td>.12</td>
</tr>
</tbody>
</table>

a) Which student is the **FASTEST** when compared to the other runners at his or her school?  
HINT – calculate the z-score for each student.

b) Which student is the **SLOWEST** when compared to the other runners at his or her school?
Chapter 3: Probability Topics - Notes

Skip Venn Diagrams.

**Example: Experiment:** Flip a coin TWICE.

Event A is “at least one head”, event B is “get a double”.

List the outcomes in the sample space S

List the outcomes in event A

List the outcomes in event B

Find the following probabilities:

P(A)

P(B)

P(A AND B)

P(A OR B)

P(B | A)

P(A | B)

P(A’)

Independent Events:

Mutually Exclusive Events:

Multiplication and Addition Rules:
- Multiplication Rule:
- Addition Rule:

Example

Write out sample space, event A and event B.

Find P(A)

Find P(B)
Find $P(A \text{ and } B)$

Find $P(A \text{ or } B)$

Find $P(A \mid B)$

Find $P(B \mid A)$.

Are $A$ and $B$ mutually exclusive events?

Are $A$ and $B$ independent events?

Find $P(B')$
Contingency Tables:
Example: Random Sample of 100 hikers and the areas of preferred hiking.

<table>
<thead>
<tr>
<th>HIKING AREA PREFERENCE</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Coastline</td>
<td>18</td>
<td>14</td>
<td>41</td>
</tr>
<tr>
<td>Near Lakes and Streams</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Mountain Peaks</td>
<td></td>
<td>14</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

M = being male  F = being female  MP = On Mt. Peaks  LS = near lakes and streams  Find:

\( P(F) \)

\( P(M) = \)

\( P(LS) = \)

\( P(F \ and \ MP) = \)

\( P(F \ or \ MP) = \)

\( P(M|MP) = \)

\( P(LS|M) = \)
Trees:

A probability tree has branches labeled with probabilities.

*Example:* There are 2 green balls and 5 red balls in a box that you cannot see into. Draw 2 balls *WITHOUT* replacement. \(R_1 = \text{red ball on first drawing, } G_1 = \text{green ball on first drawing, etc.} \) Organize the information in a probability tree.

Calculate the following:
- \(P(G_1 \text{ then } R_2)\)
- \(P(\text{one red and one G})\)
- \(P(\text{at least 1 Green})\)
- \(P(\text{at most one Red})\)
- \(P(G_2)\)
- \(P(G_1 \mid G_2)\)
- Are \(G_1\), \(G_2\) independent?
Chapter 3: Probability Topics – More Examples

1. The following table gives the number of medical claims by type of treatment and geographic region.

<table>
<thead>
<tr>
<th>Region</th>
<th>Type of Claim</th>
<th>West</th>
<th>Mid West</th>
<th>South</th>
<th>East</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outpatient</td>
<td>99</td>
<td>65</td>
<td>326</td>
<td>100</td>
<td>590</td>
</tr>
<tr>
<td></td>
<td>Physician Visit</td>
<td>251</td>
<td>104</td>
<td>514</td>
<td>233</td>
<td>1102</td>
</tr>
<tr>
<td></td>
<td>Hospitalization</td>
<td>52</td>
<td>29</td>
<td>128</td>
<td>75</td>
<td>284</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>402</td>
<td>198</td>
<td>968</td>
<td>408</td>
<td>1976</td>
</tr>
</tbody>
</table>

If a claim is chosen at random, compute the following. **Leave your answers in unreduced fractional form.**

a. \[ P(\text{Claim is from the West}) = \frac{99}{590} \]

b. \[ P(\text{Claim is from East AND Physician Visit}) = \frac{233}{1976} \]

c. \[ P(\text{Claim is from Mid West OR Outpatient}) = \frac{402}{1976} \]

d. \[ P(\text{Claim is Hospitalization GIVEN South}) = \frac{128}{968} \]

e. \[ P(\text{West OR Midwest}) = \frac{590}{1976} \]

f. Are Hospitalization AND South MUTUALLY EXCLUSIVE events? Use numbers to justify your answer and explain.
2. Based on the results of a poll of 180 voters, out of the 80 Republican voters, 60 voted for a certain proposition and 20 voted against the proposition. Among the 100 Democrats, 30 voted in favor of the proposition and 70 voted against the proposition. 

R = Republican; D = Democrat; F = for; A = against.

a) Draw an appropriate probability tree.

b) If a voter is randomly selected from those voters polled, what is the probability that he was against the proposition?

c) If a voter who was against the proposition was randomly selected from those voters polled, what is the probability that he was a Democrat?
Chapter 4: Discrete Random Variables-Notes

Note: Skip geometric, hypergeometric, Poisson

Terms
Random variable:

Discrete random variable:

Discrete probability distribution function (pdf):

Example Suppose the pdf for the number of years it takes to complete a BS degree at Middle University is given below.

Random variable $X =$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
</tr>
<tr>
<td>7</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Expected Value: “long term” average or mean, $\mu$
On average, how many years does it take for an individual to earn a BS at Middle University?
Binomial Distribution \( X \sim B(n, p) \)

Bernoulli Trial:

Binomial Distribution:

\[ X = \]

\( X \) takes on the values

\[ n = \]

\[ p = \]

\( X \sim \)
**Example:** John comes to class totally unprepared for a 21 question math 10 Exam, so he guesses randomly on each question. There are 4 possible answers per question.

Let $X =$

$X$ takes on the values

$n =$  

$p =$  

$X \sim$  

1. Find the probability that John guesses 7 questions correctly.

2. Find the probability that John guesses at most 7 questions correctly.

3. Find the probability that John guesses at least 7 questions correctly.

4. Find the probability John guesses more than 5 questions correctly.

5. Find the probability that John guesses 6 or 7 questions correctly.
Chapter 5: Continuous Random Variables - Notes

Skip the Uniform Distribution

Probability density function \( f(X) \): \( f(X) \geq 0 \).

The area between \( f(X) \) and the x-axis is equal to a probability

Continuous random variable:

\underline{Exponential Distribution}

- Probability Density Function
- Notation: \( X \)
- decay parameter
- mean
- standard deviation
**Probability:**
Area to the left of k:

Area to the right of k:

Area between c and d:

Percentiles:
**Example:** The length of time a randomly chosen 11-year old child spends playing video games per day is approximately exponentially distributed with a **mean equal to 2.5 hours**.

Let $X =$

$X \sim$

$f(X) =$

1. Find the mean and standard deviation

2. Find the probability that a randomly chosen 11-year old spends more than 1.5 hours playing video games per day.

3. Find the probability that a randomly chosen 11-year old spends less than 2 hours playing video games per day.

4. Find the probability that a randomly chosen 11-year old spends between 1 and 1.8 hours playing video games per day.

5. Ninety percent of the 11-year olds spend at LEAST how long playing video games per day?

6. Find the median length of time 11-year olds spend playing video games per day.
**Uniform**

• Function is \( f(X) = \frac{1}{(b - a)} \) where \( a < X < b \) (or \( a \leq X \leq b \))

  ![Uniform Distribution Diagram]

• Notation: \( X \sim \)

• Probability = area of a rectangle = (base)(height)

**Mean**

**Standard deviation**
**Example:** Suppose the length of time it takes Lisa to grade a quiz for her class is **uniformly distributed** in the interval of 2 to 4 hours.

Let $X =$

$X \sim$

$f(X) =$

a. Find the mean and the standard deviation

b. What is the probability it takes Lisa at least 3.5 hours to grade a quiz? Draw an appropriate picture.

c. Find the probability it takes Lisa between 2.5 and 3.2 hours to grade a quiz. Draw an appropriate picture.

d. Find the probability it takes Lisa exactly 3 hours to grade a quiz. Draw an appropriate picture.

e. Find the 95th percentile for the time it takes Lisa to grade a quiz. Draw an appropriate picture. Write a sentence interpreting this percentile.
Chapter 6: Normal Distribution - Notes

- $X$ is a continuous random variable

- Parameters:

- Graph

- Total area under the curve
- A change in the standard deviation, $\sigma$:

- A change in the mean, $\mu$:

- $z$-score

Standard Normal Distribution
- A normal (bell-shaped) distribution of standardized values called $z$-scores.
- Notation: $Z \sim$
**Example:** Suppose $Z \sim N(0, 1)$. Draw pictures and find the following.

1. $P(-1.28 < Z < 1.28)$

2. $P(Z < 1.645)$

3. $P(Z > 1.645)$

4. The 85th percentile, $k$.

**Nonstandard normal:**

**Example:** Compare the GPA’s of 2 students. Student A has GPA of 3.5, at a school with $\mu = 3.7$, $\sigma = 0.1$. Student B has GPA of 3.4 at school with $\mu = 3.7$, $\sigma = 0.2$.

Calculate their z-scores. Which student has higher GPA relative to their school?
Example: The patient recovery time from a surgical procedure is normally distributed with a mean of 4.7 days and a standard deviation of 1.3 days.

a) In words, describe the random variable.

b) Give the distribution of X

c) What is the probability that a patient will take more than 6.2 days to recover?

d) What is the probability a patient will take between 4 and 6 days to recover?

f) What lengths of recovery time bound the middle 80% of patients?
Chapter 6: Normal Distribution – More Examples

1. A stamping machine produces can tops whose diameters are normally distributed with a mean of 3.124 inches and a standard deviation of 0.03 inches. Let \( X \) be the diameter of a can top produced by this stamping machine.

   a) The symbolic notation for this distribution is: \( X \sim \) _______________

   b) Find and interpret the 90\(^{th}\) percentile.

   c) Draw a graph illustrating the 90\(^{th}\) percentile.

2. The time, in minutes, to complete the daily Sudoku puzzle in the San Jose Mercury News follows a normal distribution. For Ali, the distribution is \( N(5, 1.2) \); for Binh, \( N(8, 2.1) \); for Carlos, \( N(7, 2.8) \). On Saturday, the puzzle was particularly hard. Ali took 7.5 minutes, Binh took 12 minutes and Carlos took 11 minutes.

   For each person, find their z-score for Saturday.

   Ali ________________

   Binh _____________

   Carlos______________

   Is Ali faster than the others when each is compared to his/her usual performance?
Chapter 7: Central Limit Theorem (CLT) – Notes

NOTE: We do CLT for AVERAGES.

Experiment:

<table>
<thead>
<tr>
<th>Average values for rolling 2 dice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average values for rolling 5 dice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Draw a histogram for each.

What distributions do we seem to have?
The Central Limit Theorem (CLT) for Averages:

Suppose \( X \) is a random variable with a probability distribution that may be known or unknown and suppose

- \( \mu_x \) = the mean of \( X \)
- \( \sigma_x \) = the standard deviation of \( X \)

If you draw random samples of size \( n \), then as \( n \) increases, the random variable \( \bar{X} \) which consists of averages tends to be normally distributed and

\[ \bar{X} \sim \]

The CLT for Averages says that if you keep drawing larger and larger samples and taking their averages, the averages themselves form a normal distribution.

\( n \) = the number of values that are averaged

The Law of Large Numbers:
Example 1: Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If $\overline{X}$ = average distance in feet for 49 fly balls, then $\overline{X} \sim (__, __)$

b. Define $X$ and give its distribution

c. What is the probability that the 49 balls traveled an average of less than 240 feet?

d. What is the probability that a single fly ball traveled less than 240 feet?

e. Find the 80th percentile of the distribution of the average distance of 49 fly balls.
Example 2: The monthly income of the trainees at a large corporation is approximately normally distributed. The mean monthly income of the trainees is $1500 with a standard deviation of $200. We’ll take a random sample of size 25.

a. Define $X$ and $\bar{X}$. Give their distributions.

b. What is the probability that a trainee has a monthly income in excess of $1400.$

c. If 25 trainees are selected randomly, what is the probability that their average income will be in excess of $1400?$

d. What is the probability that a trainee has income between $1400$ and $1600$?

e. If 25 trainees are selected randomly, what is the probability that their average income will be between $1400$ and $1600?$
Chapter 7: Central Limit Theorem (CLT) – More Examples

1. The amount of time that De Anza students play “Texas Hold ‘Em” each week has an unknown distribution with a mean of 6.5 hours. Suppose the standard deviation is 2 hours. Consider a random sample of 20 De Anza students.

a) In words, $X =$

b) In words, $\bar{X} =$

c) $\bar{X} \sim$

d) Find the probability that the average time the 20 De Anza students play “Texas Hold ‘Em” is between 6.2 and 6.7 hours. Draw a picture, shading the appropriate area. Label the x-axis.

e) Find the $95^{th}$ percentile of the distribution for the average time that the 20 De Anza students play “Texas Hold ‘Em” each week. Round answer to 2 decimal places.

f) Describe, in a complete sentence, what the $95^{th}$ percentile found in e) means.
Chapter 8: Confidence Intervals – Notes

point estimate

confidence interval

confidence interval has the form

Central Limit Theorem for Averages:

If you draw random samples of size \( n \), then as \( n \) increases, the random variable \( \bar{X} \) which consists of averages tends to be normally distributed.

\[
\bar{X} \sim N(\mu_x, \sigma_x/\sqrt{n})
\]
Confidence interval for single population mean, population standard deviation known – use normal distribution

**Example:** Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its average numbers of unoccupied seats per flight over the past year. To accomplish this, the records of 15 flights are randomly selected and the number of unoccupied seats is noted for each of the sample flights. The sample mean is \( \bar{x} = 11.6 \). Assume the standard deviation for the population of unoccupied seats is \( \sigma = 2.25 \). Find and interpret a 95% confidence interval for the true average number of unoccupied seats per flight.

a) Define the random variable \( X \).

b) Give the distribution of \( X \). \( X \sim \)

c) What is a point estimate for the true average number of unoccupied seats per flight?

d) Find a 95% confidence interval for the true mean number of unoccupied seats per flight.

e) What is the error bound on your confidence interval?

f) Draw an appropriate graph for this confidence interval

g) Interpret your confidence interval.
Confidence interval for single population mean, population standard deviation unknown – use student-t distribution

**Example:** With the price of ski slope passes, most skiers or snowboarders are interested in finding a reasonably priced room within the ski area. A random sample of the minimum price per night of a room in 10 ski areas is as follows: $35, $65, $65, $72, $79, $79, $79, $89, $99, $99.

a) Define the random variable X.

b) Give the distribution of X. X~

c) What is a point estimate for the true mean minimum price per night of a room in a ski area?

d) Find a 93% confidence interval for the true mean minimum price per night of a room in a ski area.

e) Draw an appropriate graph for this confidence interval

f) Give the error bound on the confidence interval.

f) Interpret your confidence interval.
Confidence interval for a population proportion – use normal approximation to the binomial

**Example:** An elementary school administrator wants to determine the true population proportion of elementary school students that are low income in his school district (K – 8). He randomly samples records from 140 students and determines that 39 are low income.

a) Define the random variable X.

b) Give the distribution of X. \( X \sim \)

c) Find a 95% confidence interval of the true population proportion of elementary school students in the administrator’s district that are low income.

d) Draw an appropriate graph for your confidence interval.

e) CHOOSE ONE: The term “95% confidence” means if we took repeated samples, then:

A. approximately 95% of the confidence intervals would be the same.
B. approximately 95% of the confidence intervals would contain the population proportion.
C. approximately 95% of the population data will be included in the confidence interval.
D. approximately 95% of the confidence intervals would contain the sample proportion.
1: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 - 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 - 12, in all beginning ice-skating classes at the Ice Chalet.

**Estimated distribution**

1. What is being counted?

2. In words, define the Random Variable X.
   \[ X = \text{ } \]  

3. Calculate the following:
   a. \[ x = \text{ } \]  
   b. \[ n = \text{ } \]  
   c. \[ p' = \text{ } \]  

4. State the estimated distribution of X. \[ X \sim \]  

5. What is \( p' \) estimating?

6. In words, define the Random Variable \( P' \).
   \[ P' = \text{ } \]  

7. State the estimated distribution of \( P' \). \[ P' \sim \]
EXPLAINING THE CONFIDENCE INTERVAL

Construct a 92% Confidence Interval for the true proportion of girls in the age 8 - 12 beginning ice-skating classes at the Ice Chalet.

8. How much area is in both tails (combined)? $\alpha = \underline{\hspace{2cm}}$

9. How much area is in each tail? $\frac{\alpha}{2} = \underline{\hspace{2cm}}$

10. Calculate the following:
    a. lower limit = \underline{\hspace{2cm}}
    b. upper limit = \underline{\hspace{2cm}}
    c. error bound = \underline{\hspace{2cm}}

11. The 92% Confidence Interval is: \underline{\hspace{2cm}}

12. Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval, and the sample proportion.

\[
\frac{\alpha}{2} = \underline{\hspace{2cm}} \quad \text{C.L.} = \underline{\hspace{2cm}} \quad \frac{\alpha}{2} = \underline{\hspace{2cm}}
\]

\[
____ \quad ____ \quad ____ \quad p' \]

13. In one complete sentence, explain what the interval means.


**Discussion Questions**

1. Using the same \( p' \) and level of confidence, suppose that \( n \) were increased to 100. Would the error bound become larger or smaller? How do you know?

2. Using the same \( p' \) and \( n = 80 \), how would the error bound change if the confidence level were increased to 98%? Why?

2: The following real data are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true average number of colors on a national flag. Let \( X \) = the number of colors on a national flag.

<table>
<thead>
<tr>
<th>( X )</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

**Calculating the Confidence Interval**

1. Calculate the following:
   a. \( \bar{x} = \ldots \)
   b. \( s_x = \ldots \)
   c. \( n = \ldots \)

2. Define the Random Variable, \( \bar{X} \), in words.
   \[ \bar{X} = \text{______________________________} \]

3. What is \( \bar{x} \) estimating?
4. Is \( \sigma \) known?

5. As a result of your answer to (4), state the exact distribution to use when calculating the Confidence Interval.

**CONFIDENCE INTERVAL FOR THE TRUE AVERAGE NUMBER**

Construct a 95\% Confidence Interval for the true average number of colors on national flags.

6. How much area is in both tails (combined)? \( \alpha = \) ______

7. How much area is in each tail? \( \frac{\alpha}{2} = \) ______

8. lower limit = _____ upper limit = _____ error bound = _____

9. The 95\% Confidence Interval is: _________________

10. Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval and the sample mean.

\[
\frac{\alpha}{2} = \quad \text{C.L.} = \quad \frac{\alpha}{2} =
\]

11. In one complete sentence, explain what the interval means.
Chapter 9: Hypothesis Testing with One Sample – Notes

- Null hypothesis:

- Alternate/Alternative hypothesis:

**Example:** State the null hypothesis, \( H_0 \), and the alternative hypothesis, \( H_a \), in terms of the appropriate parameter (\( \mu \) or \( p \)).

At most 60% of Americans vote in presidential elections.

\( H_0: \)
\( H_a: \)

Fewer than 5% of adults ride the bus to work in New York City.

\( H_0: \)
\( H_a: \)

Europeans have an average paid vacation each year of six weeks.

\( H_0: \)
\( H_a: \)

Private universities cost, on average, more than $20,000 per year for tuition, room, and board.

\( H_0: \)
\( H_a: \)
Correct Decisions and Errors:

<table>
<thead>
<tr>
<th>Decision: Do not reject $H_0$</th>
<th>$H_0$ is true</th>
<th>$H_0$ is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision: Reject $H_0$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\alpha$ is preconceived. Its value is set before the hypothesis test starts. If there is no given preconceived $\alpha$, then use $\alpha=0.05$

$\alpha = \text{probability of a Type I error} = P(\text{Type I error}) = \text{probability of rejecting the null hypothesis when the null hypothesis is true.}$

$\beta = \text{probability of a Type II error} = P(\text{Type II error}) = \text{probability of not rejecting the null hypothesis when the null hypothesis is false.}$

Goal: Minimize $\alpha$ and $\beta$

**The Power of the Test:** $1 - \beta$ (want to be large)

*Example:* What are the Type I and Type II errors?

At most 60% of Americans vote in presidential elections.

$H_0$: 

$H_a$: 

Type I Error:

Type II Error:
To perform a hypothesis test:

- Set up hypotheses
- **sample data is gathered**
- **data typically favors one of the hypotheses**
- Calculate **p-value**

\[
p\text{-value} = \\
\]

If \( \alpha \leq p\text{-value} \), then do not reject \( H_0 \).

If \( \alpha > p\text{-value} \), then reject \( H_0 \)

- **Decisions**
  - if data favors the null hypothesis \( (H_0) \), we “do not reject the null”
  - if data favors the alternate hypothesis \( (H_a) \), we “reject the null”

- Write an appropriate conclusion

**Types of Hypothesis Tests**

- Single population mean, **known** population variance (or standard deviation):

- Single population mean, **unknown** population variance (or standard deviation):

- Single population proportion:
**Example 1:** According to an article in *The New York Times* (5/12/2004), 19.3% of New York City adults smoked in 2003. Suppose that a survey is conducted to determine this year’s rate. Twelve out of 70 randomly chosen N.Y. City residents reply that they smoke. At the 5% level, conduct a hypothesis test to determine if the rate is less than 19.3%.

a. Ho: ____________  
b. Ha: ______________

c. In words, CLEARLY state what your random variable $\bar{X}$ or $P'$ represents.

__________________________________________________________
__________________________________________________________


d. State the distribution to use for the test. ________________

e. Test Statistic: t or $z = __________$

f. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

__________________________________________________________
__________________________________________________________


g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<---------------------------------------------------------------->
h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion:__________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________

i. Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.

Confidence Interval: (___________, __________)
Example 2: The average number of sick days an employee takes per year is believed to be about 10. Members of a personnel department do not believe this figure. They randomly survey 8 employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the average number is about 10?

a. Ho: ____________  
b. Ha: ______________

c. In words, CLEARLY state what your random variable $\bar{X}$ or P’ represents.

______________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

d. State the distribution to use for the test. ______________

e. Test Statistic: t or z = __________

f. p-value = __________  In 1 – 2 complete sentences, explain what the p-value means for this problem.

______________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<------------------------------------------------------------------------------------------------>
h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion:**

________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________

i. Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.

Confidence Interval: ( __________, __________)
**Example 3:** Employees in a large firm claim that the mean annual salary of the firm’s accountants is less than that of its competitors, which is $45,000. A random sample of 30 of the firm’s accountants has a mean salary of $43,999 with a standard deviation of $3,000. At a significance level of 5%, test the employee’s claim.

a. Ho: ________________  
b. Ha: ________________  

c. In words, CLEARLY state what your random variable $\bar{X}$ or $P'$ represents.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

d. State the distribution to use for the test. ________________  

e. Test Statistic: $t$ or $z$ = ____________  

f. $p$-value = ____________  In 1 – 2 complete sentences, explain what the $p$-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the $p$-value.

<--------------------------------------------------------------------->
h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
</table>

Conclusion:  


i. Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.

Confidence Interval: (___________, __________)
1. Suppose the percent of students expected to pass the second exam of the quarter in Calculus is expected to be 74%. A sample of 21 students results shows that 66% of the students passed the exam. Conduct a hypothesis test to determine if the survey result is different from the expected pass percentage of 74%.

a. Ho: ____________  

b. Ha: ____________  

c. In words, CLEARLY state what your random variable $\bar{X}$ or $P'$ represents.

_______________________________________________________________________________________________
_______________________________________________________________________________________________


d. State the distribution to use for the test. ______________

e. Test Statistic: $t$ or $z = ________$

f. $p$-value = _________  In 1 – 2 complete sentences, explain what the $p$-value means for this problem.

_______________________________________________________________________________________________
_______________________________________________________________________________________________


g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the $p$-value.

<----------------------------------------------------------------->


h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion: ____________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

i. Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.

Confidence Interval: ( ___________ , ___________ )
2. A student is doing a statistics project on the weight of small “fun-sized” bags of M&M’s. The product information states that the average weight of a bag is 1.75 ounces. The student weighs 18 bags of candy and finds that the mean weight is 1.7 ounces with a standard deviation of 0.0475 ounces. At a 5% significance level, is the manufacturer’s stated weight accurate? (Assume that the underlying distribution is normal)

a. Ho: ___________  
b. Ha: ______________

c. In words, CLEARLY state what your random variable $\bar{X}$ or $P'$ represents.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

d. State the distribution to use for the test. ______________

e. Test Statistic: t or z = __________

f. p-value = __________  In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<---------------------------------------------------------->
h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion:**

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

i. Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.

Confidence Interval: (__________,__________)
Chapter 10: Hypothesis Testing with Two Samples - Notes

Independent groups (samples are independent)

Two Means

Two proportions

Matched or Paired Samples
**Example 1:** The average number of English courses taken in a two-year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of 3 English courses with a standard deviation of 0.8. The females took an average of 4 English courses with a standard deviation of 1.0. Are the averages statistically the same?

a. Ho: ___________  

b. Ha: ___________

c. In words, CLEARLY state what your random variable $\overline{X}_1 - \overline{X}_2$, $P_1' - P_2'$ or $\overline{X_d}$ represents.  

__________________________________________________________

__________________________________________________________

d. State the distribution to use for the test. ____________________

e. Test Statistic: $t$ or $z$ = __________

f. $p$-value = __________  In 1 – 2 complete sentences, explain what the $p$-value means for this problem.

__________________________________________________________

__________________________________________________________

__________________________________________________________
g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion: ________________________________________________________________

____________________________________________________________________________

____________________________________________________________________________

____________________________________________________________________________
**Example 2:** Proportions
A local women’s group has claimed that men and women differ in attitudes about the environment. A group of 50 men (M) and 40 women (W) were asked if they thought that protecting the environment was an important issue. Of those persons sampled, 11 of the men and 19 of the women said they believed that protecting the environment was an important issue. Conduct a hypothesis test to test the local women’s group claim.

Pooled Proportion:

Distribution to use:

a. Ho: _____________  
b. Ha: _____________  

d. In words, CLEARLY state what your random variable $\bar{X}_1 - \bar{X}_2$, $p_1' - p_2'$ or $\bar{X}_d$ represents.

----------------------------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------------------------

d. State the distribution to use for the test. ____________________________

e. Test Statistic: t or z = ____________
f. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<--------------------------------------------------------------------------------------------->

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>_______</td>
<td>_______</td>
<td>_______</td>
</tr>
</tbody>
</table>

Conclusion:____________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________
Example 3 (Matched or paired samples)
Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. Evaluate the data below. Do you think that their cholesterol levels were significantly lowered?

<table>
<thead>
<tr>
<th>Starting cholesterol level</th>
<th>Ending cholesterol level</th>
<th>Different (end - start)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>After - before</td>
</tr>
<tr>
<td>140</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>220</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>190</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>360</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>280</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>260</td>
<td>240</td>
<td></td>
</tr>
</tbody>
</table>

a. Ho: ___________     b. Ha: ___________

c. In words, CLEARLY state what your random variable $\bar{X}_1 - \bar{X}_2$, $P_1' - P_2'$ or $\bar{X}_d$ represents.

_______________________________________________________________________________________________
_______________________________________________________________________________________________

d. State the distribution to use for the test. __________________________

e. Test Statistic: t or z = __________
f. p-value = __________  In 1 – 2 complete sentences, explain what the p-value means for this problem.

___________________________________________________________

___________________________________________________________

___________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<------------------------------------------------------------------>

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>______</td>
<td>_______</td>
<td>_________</td>
</tr>
</tbody>
</table>

Conclusion:___________________________________________________________

___________________________________________________________
Chapter 10: Hypothesis Testing with Two Samples – More Examples

1. In a controlled environment, random samples of 10 adults and 10 children were tested to determine the room temperature that each person found most comfortable. The data are summarized as follows:

<table>
<thead>
<tr>
<th></th>
<th>Sample Mean (°F)</th>
<th>Sample Standard Deviation (°F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adults (A)</td>
<td>77.5</td>
<td>2.12</td>
</tr>
<tr>
<td>Children (C)</td>
<td>74.5</td>
<td>1.58</td>
</tr>
</tbody>
</table>

Test the hypothesis that adults prefer warmer room temperatures than children.

a. Ho: ____________  
b. Ha: ______________

c. In words, CLEARLY state what your random variable $\bar{X}_1 - \bar{X}_2$, $p_1' - p_2'$ or $\bar{X}_d$ represents.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

d. State the distribution to use for the test. __________________________

e. Test Statistic: $t$ or $z = __________$
f. p-value = __________  In 1 – 2 complete sentences, explain what the p-value means for this problem.

__________________________________________________________________

__________________________________________________________________

__________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

![Diagram](image)

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>______</td>
<td>________</td>
<td>____________</td>
</tr>
</tbody>
</table>

Conclusion:_______________________________________________________________

__________________________________________________________________

__________________________________________________________________
2. A city recently launched a neighborhood watch program to control crime. The following table gives the number of crimes reported in six neighborhoods during the six months before and six months after the new neighborhood watch program was launched.

<table>
<thead>
<tr>
<th>Before</th>
<th>57</th>
<th>73</th>
<th>47</th>
<th>68</th>
<th>79</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>41</td>
<td>65</td>
<td>28</td>
<td>73</td>
<td>61</td>
<td>32</td>
</tr>
</tbody>
</table>

Test the hypothesis that crime was better controlled after the new neighborhood watch program was launched.

a. Ho: ____________

b. Ha: ______________

c. In words, CLEARLY state what your random variable $\bar{X}_1 - \bar{X}_2$, $P_1' - P_2'$ or $\bar{X}_d$ represents.

______________________________________________________________________________________________

______________________________________________________________________________________________

d. State the distribution to use for the test. __________________________

e. Test Statistic: t or z = __________
f. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<----------------------------------------------------------------------------------------------->

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>______</td>
<td>__________</td>
<td>_____________________</td>
</tr>
</tbody>
</table>

Conclusion:________________________________________________________________________________

________________________________________________________________________________

________________________________________________________________________________
Chapter 11: The Chi-Square Distribution - Notes

Test of Independence

Goodness-of-Fit Test

Chi-square or $\chi^2$

**Hypothesis Test: Goodness-of-Fit**
1. Use goodness of fit test to test whether a data set fits a particular probability distribution.
2. The degrees of freedom are number of cells or categories – 1.
3. Test statistic: $\Sigma (O-E)^2/E$ where $O =$ observed values, $E =$ expected values
4. The test is right tailed

**Hypothesis Test: Independence**
1. Use the test of independence to test whether two factors are independent or not.
2. The degrees of freedom = (# rows – 1)(# columns – 1)
3. Test statistic: $\Sigma (O-E)^2/E$ where $O =$ observed values, $E =$ expected values
4. The test is right tailed
5. If the null hypothesis is true, the expected number $E$ is:
   
   $E = (\text{row total}) \times (\text{column total}) / (\text{# surveyed})$
**Example:** The *percentage* of students who attend a local school in any given school week is as follows:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>Tuesday</td>
<td>Wednesday</td>
<td>Thursday</td>
<td>Friday</td>
</tr>
<tr>
<td>95%</td>
<td>96%</td>
<td>98%</td>
<td>97%</td>
<td>95%</td>
</tr>
</tbody>
</table>

In one given school week, the *number* of students (data) who attended school out of a student population of 500 was:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>Tuesday</td>
<td>Wednesday</td>
<td>Thursday</td>
<td>Friday</td>
</tr>
<tr>
<td>450</td>
<td>470</td>
<td>485</td>
<td>480</td>
<td>470</td>
</tr>
</tbody>
</table>

Perform a goodness-of-fit hypothesis test to determine if the numbers fit the percentages given.

a. Ho:

b. Ha:

c. What are the degrees of freedom?

d. State the distribution to use for the test.

e. What is the test statistic?
f. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

![Diagram of a normal distribution]

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>______</td>
<td>_______</td>
<td>________________</td>
</tr>
</tbody>
</table>

Conclusion:_______________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________
Example 2: Conduct a hypothesis test to determine whether there is a relationship between an employees performance in a company’s training program and his/her ultimate success on the job. Use a level of significance of 1%.

<table>
<thead>
<tr>
<th>Performance in Training program</th>
<th>Performance on Job</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below Average</td>
</tr>
<tr>
<td>Poor</td>
<td>23</td>
</tr>
<tr>
<td>Average</td>
<td>28</td>
</tr>
<tr>
<td>Very Good</td>
<td>9</td>
</tr>
<tr>
<td>TOTAL</td>
<td>60</td>
</tr>
</tbody>
</table>

a. Ho:

b. Ha:

c. What are the degrees of freedom?

e. State the distribution to use for the test.

e. What is the test statistic?

f. Fill out the table below with the Expected Values:

<table>
<thead>
<tr>
<th></th>
<th>Below Average</th>
<th>Average</th>
<th>Above Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Good</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
g. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

h. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

i. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>______</td>
<td>________</td>
<td>____________________</td>
</tr>
</tbody>
</table>

Conclusion:____________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________
Chapter 11: The Chi-Square Distribution – More Examples

Example 1: A GEICO Direct magazine had an interesting article concerning the percentage of teenage motor vehicle deaths and the time of day. The following percentages were given from a sample.

<table>
<thead>
<tr>
<th>Time of day</th>
<th>Percent of teenage motor vehicle deaths</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 – 3 am</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 – 6am</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 – 9am</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9am – noon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noon – 3pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 – 6 pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 – 9 pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9pm – midnight</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Perform a goodness-of-fit hypothesis test at the 3% level to determine if the percentage of teenage motor vehicle deaths are same for each time period?

a. Ho:

b. Ha:

c. What are the degrees of freedom?

d. State the distribution to use for the test.

e. What is the test statistic?
f. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

<--------------------------------------------------------------------------------------------->

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>_______</td>
<td>_______</td>
<td>_________________</td>
</tr>
</tbody>
</table>

Conclusion:_______________________________________________________________
**Example 2:** The table shows a random sample of 100 hikers and the area of hiking preferred. Are hiking area preference and gender independent?

<table>
<thead>
<tr>
<th>Hiking Preference Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coastline</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
</tbody>
</table>

Table shows O = observed values

Perform a hypothesis test at the 5% level to determine if hiking area preference and gender are independent.

a. Ho:

b. Ha:

c. What are the degrees of freedom?

d. State the distribution to use for the test.

e. Fill out the table below with the Expected Values

<table>
<thead>
<tr>
<th></th>
<th>Coastline</th>
<th>Lake/Stream</th>
<th>Mountains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

f. What is the test statistic?
g. p-value = __________ In 1 – 2 complete sentences, explain what the p-value means for this problem.

_______________________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________

h. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

i. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using COMPLETE SENTENCES.

<table>
<thead>
<tr>
<th>alpha</th>
<th>decision</th>
<th>reason for decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>_______</td>
<td>_______</td>
<td>__________</td>
</tr>
</tbody>
</table>

Conclusion:____________________________________________________________________________________

_______________________________________________________________________________________________

_______________________________________________________________________________________________
Chapter 12: Linear Regression and Correlation - Notes

Bivariate data:

Linear Regression:

Correlation Coefficient:

Coefficient of Determination:
Determine **degrees of freedom**:

- If $r <$ negative critical value:

- If $r >$ positive critical value:

**TABLE**

**95% CRITICAL VALUES OF THE SAMPLE CORRELATION COEFFICIENT**

<table>
<thead>
<tr>
<th>Degrees of Freedom: n - 2 Critical Values: (+ and -)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

**OR** use $p$-value from LinRegTTest ($p$-value < alpha means correlation coefficient is significant)

**Regression Line as Predictor**

**Outliers!**

**NOTE:** NEED Diagnostic ON (under 2nd catalogue)
**Example:** Table shows quiz score (out of 20) and the grades on a midterm exam (out of 100) for a sample of 8 students who took this course last quarter.

<table>
<thead>
<tr>
<th>quiz</th>
<th>20</th>
<th>15</th>
<th>13</th>
<th>18</th>
<th>18</th>
<th>20</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>midterm</td>
<td>92</td>
<td>72</td>
<td>72</td>
<td>95</td>
<td>88</td>
<td>98</td>
<td>65</td>
<td>77</td>
</tr>
</tbody>
</table>

1. Draw the scatter plot. Does it look like a straight line will fit the data?

2. Calculate the equation of least squares line / regression line / line of best fit.

3. Draw the regression line over the scatterplot.

4. Give the p-value and the degrees of freedom.

5. Find the correlation coefficient. Is it significant?

6. Write an interpretation of the slope of the line.
7. What is the predicted score on the midterm for a student who got a score of 17 on the quiz? Do NOT predict outside the domain!

8. Find $s$ – the standard deviation of the residuals.

9. Are the any outliers? If so, what are they?

**TABLE for calculating outliers:**

| Quiz | Midterm | Predicted $y$ | $|y-\hat{y}|$ | $|y-\hat{y}|^2$ |
|------|---------|--------------|--------------|----------------|
| 20   | 92      |              |              |                |
| 15   | 72      |              |              |                |
| 13   | 72      |              |              |                |
| 18   | 95      |              |              |                |
| 18   | 88      |              |              |                |
| 20   | 98      |              |              |                |
| 14   | 65      |              |              |                |
| 16   | 77      |              |              |                |
Chapter 12: Linear Regression and Correlation – More Examples

The following table shows the calories and sugar content for one cup of each of 7 different breakfast cereals.

<table>
<thead>
<tr>
<th>Cereal Name</th>
<th>Apple Jacks</th>
<th>Corn Flakes</th>
<th>Corn Pops</th>
<th>Fruit Loops</th>
<th>Honey Nut Cheerios</th>
<th>Multi Grain Cheerios</th>
<th>Wheaties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calories, x</td>
<td>130</td>
<td>100</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>Sugar (in gms), y</td>
<td>16</td>
<td>2</td>
<td>14</td>
<td>15</td>
<td>11</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

1. Draw the scatter plot. Does it look like a straight line will fit the data?
2. Calculate the least squares line / regression line. line of best fit. Put the equation in the form yhat = a + b x.
3. Draw the regression line over the scatterplot.
4. Find the correlation coefficient. Is it significant? Give the p-value and the degrees of freedom.
5. Write an interpretation of the slope of the line
6. Find the estimated grams of sugar for a cereal containing 115 calories in one cup.
7. Find s - the standard deviation of the residuals
8. Are the any outliers? If so, what are they?
Table for calculating outliers:

| Calories, x | Sugar (g) y | Predicted y ̂ | y - ̂y | |y-̂y|^2 |
|-------------|-------------|---------------|--------|----------------|
| 130         | 16          |               |        |                |
| 100         | 2           |               |        |                |
| 120         | 14          |               |        |                |
| 120         | 15          |               |        |                |
| 120         | 11          |               |        |                |
| 110         | 6           |               |        |                |
| 110         | 4           |               |        |                |
Chapter 13: F Distribution and One Way ANOVA – Notes

ANOVA

Basic Assumptions for ANOVA

F-Distribution
ANOVA summary table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares (SS)</th>
<th>Degrees of Freedom (df)</th>
<th>Mean Square (MS)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Example:** Are the means for the final exam scores the same for all statistics class delivery types? The table below shows the scores on final exams from several randomly selected classes that used the different delivery types.

<table>
<thead>
<tr>
<th>Sample mean</th>
<th>Online</th>
<th>Hybrid</th>
<th>Face to Face</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>83</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>73</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>84</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>81</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assume that all distributions are normal, the 3 population standard deviations are approximately the same, and the data were collected independently and randomly. Use the 0.05 level of significance.

For each delivery type, complete the following table:

<table>
<thead>
<tr>
<th>Sample mean</th>
<th>Online</th>
<th>Hybrid</th>
<th>Face to Face</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Complete the ANOVA table:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares (SS)</th>
<th>Degrees of Freedom (df)</th>
<th>Mean Square (MS)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Distribution to use for the test:

Test Statistic:

Graph and probability statement:

Decision:

Conclusion: