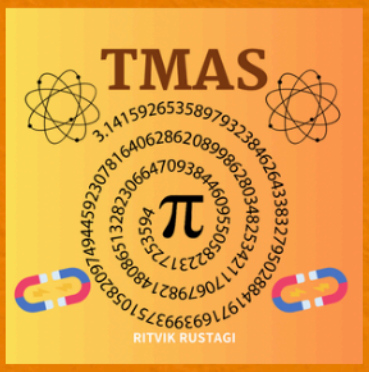


TMAS Academy

ACE AP Statistics 2025

- 100+ Problems
- All Topics
- Detailed Solutions

Gurshan Bhalrhu and Caden Wang

Contents

§0.1 About TMAS Academy

TMAS Academy, previously known as Explore Math, was started by Ritvik Rustagi in 2020 to spread competition math. In full, it is **The Math and Science Academy**. Ritvik expanded the academy in October 2023 by releasing his AMC 10/12 prep book. After that, in March 2024, he released his free AP Physics 1, AP Calculus AB/BC, and AP Physics C: Mechanics books. Now, TMAS Academy has evolved into a large team of hard-working, passionate, and dedicated students who enjoy STEM and helping others. We believe that everyone should be able to achieve their full potential with learning, so we have channeled our efforts into making educational resources accessible to all.

You can learn more by visiting the website linked below.
Website: https://www.tmasacademy.com/

§0.2 Opportunities For You To Contribute To TMAS Academy

TMAS Academy is very inclusive and you can help support its cause in many ways.
You can **join the team** by filling out the form below:
https://forms.gle/VXGvj27UvcZPGhiJ8

Donations: If you want to assist us in our monthly payments to run TMAS Academy, which includes website costs, Overleaf (the platform used to write these books) costs, and filming/editing costs, then please consider donating! For those willing to contribute, we have listed a few ways below. **Don't forget to write a message so we know who you are which will allow us to send you a thank you note!**

- You can donate through PayPal to the email: ritvikrustagi7@gmail.com
- If you want to donate and the above method doesn't work for you, then you can send an email to ritvikrustagi7@gmail.com

You can also contribute by **subscribing** to the YouTube channel:
https://www.youtube.com/@tmasacademy
Also, don't forget to join the Discord server to connect with other hardworking students preparing for AP exams and math competitions such as AMC 10/12 and AIME.
https://discord.gg/tmas-academy-1019082642794229870

There are occasional group study sessions and other review sessions led by Ritvik Rustagi and others from the server!

You can also follow all of our socials such as the LinkedIn page and the Instagram account that are run by the media team. Also, please join the mailing list to learn about all updates and our upcoming books and videos. All of that can be found at the bottom of the website: https://www.tmasacademy.com/

Finally, you can spread our efforts and initiative to anyone you know who may benefit from or support us, be it your classmates, teachers, or other nonprofit organizations focused on education.

§0.3 About the Author: Gurshan Bhalrhu

My name is Gurshan Bhalrhu. I am a junior at Antelope High School with a strong interest in STEM. I earned a score of 5 on the AP Statistics exam and authored all of Units 1-4 of this book as well as parts of Unit 5.

I am grateful for the opportunity provided by Ritvik Rustagi to join the TMAS Academy team and help write this book. I hope the book helps students succeed in their classes and earn a 5 on the AP exam.

§0.4 About the Author: Caden Wang

My name is Caden Wang. I am a high school senior in Ontario, Canada, with a strong interest in Mathematics and Business. I scored a 5 on the AP Statistics and AP Calculus BC exams and authored all of units 6-9 and most of unit 5 of this book.

I joined TMAS Academy in my junior year, and I was deeply impressed with the resources that Ritvik Rustagi has made. Inspired by this, I quickly joined the team and began working on this textbook. I have been tutoring mathematics for several years now and am also involved in math competitions, including the AMC series. I hope this book will help students achieve their academic goals and gain a better overall understanding of Statistics.

§0.5 Benefits of Taking AP Statistics

AP Statistics teaches concepts that can be applied to the real-world, such as understanding studies, probability, and polling. This broad application makes the course useful for various fields, not simply limited to mathematics. As an AP course, there is opportunity to gain college credit, enabling a person who passes the exam to skip prerequisite courses and save money.

§0.6 Credits

We would like to thank College Board for providing us with problems that were used in this book.

We also thank Ritvik Rustagi for providing us with the opportunity to give back by authoring this book.

I (Caden Wang) would like to thank the founder of Khan Academy, Sal Khan, for creating an AP Statistics course which has taught me the content of this course and helped me earn a 5 on the exam.

I (Gurshan Bhalrhu) would like to thank my AP Statistics teacher, Mr. Wood, for teaching me the content of this course and helping me earn a 5 on the exam.

1 Unit 1: (Exploring One-Variable Data)

§1.1 What Can We Accomplish Using Statistics?

Unit 1 provides an overview of this course, where you will learn various graphical displays, how to use statistics to prove and investigate claims, the normal distribution, etc.

Note 1.1.1

Statistics can be utilized to solve real-world problems by collecting data, analyzing that data, and interpreting results. You will learn various sampling methods to collect data, create graphical or numerical representations of that data, and the correct way of writing interpretations.

§1.2 Quantitative vs Categorical Variables

A **variable** is a characteristic that can take on different values or categories.

Note 1.2.1

Categorical Variables:

A **categorical variable** is a type of variable that represents categories or groups. A few examples of this are college majors or grade level. While a categorical variable can be a number, it has to represent a category. For example, a zip code is a categorical variable even though it's a numerical value.

Note 1.2.2

Quantitative Variables:

Quantitative variables are types of variable that represents numerical measurements or amounts of things. They can be either **discrete** or **continuous**. Age is a continuous variable, and number of siblings is discrete because there are gaps between values, one cannot have 5.3 siblings. Quantitative variables typically have units, for instance, time is measured in a unit like seconds, minutes, or hours. If it makes sense to find an average of a variable, then it is likely quantitative.

Problem 1.2.3 — Identify which of the following variables are categorical and which are quantitative:

- Height
- Education Level
- Music Genre Preference
- Blood Pressure
- Marital Status
- Employment Status

Solution:

Height and Blood Pressure are **quantitative** variables because they measure specific quantities, where height is measured in inches, feet, or centimeters, and blood pressure is measured in mmHg (millimeters of mercury).

Education level is a **categorical** variable because it represents the category of grade level, rather than a specific unit being measured. Music genre preference, marital status, and employment status are categories, and do not have any numeric values, making them **categorical** variables.

§1.3 Visual Representations of Categorical Variables

Categorical Variables can be represented in both **frequency** and **relative frequency tables.**

Note 1.3.1

A classroom survey was conducted where students were asked about their favorite fruit. The results were put in frequency and relative frequency tables here:

Category Name	Frequency	Relative Frequency
Apple	8	21.1%
Orange	2	5.3%
Banana	15	39.5%
Grape	3	7.9%
Strawberry	10	26.3%
Total	38	100%

The relative frequency shows the proportion of each category, whereas the frequency just shows the amount. Relative frequencies, proportions, and percentages all provide the same information, just written in different ways. Expect to be able to draw conclusions from these types of tables.

Problem 1.3.2 — Use the table to answer the following questions:

a) What proportion of students selected apples or grapes as their favorite fruit?

b) Out of the students who didn't pick strawberry as their favorite, what fraction picked banana?

c) What fraction of students picked strawberry, orange, or apple as their favorite fruit?

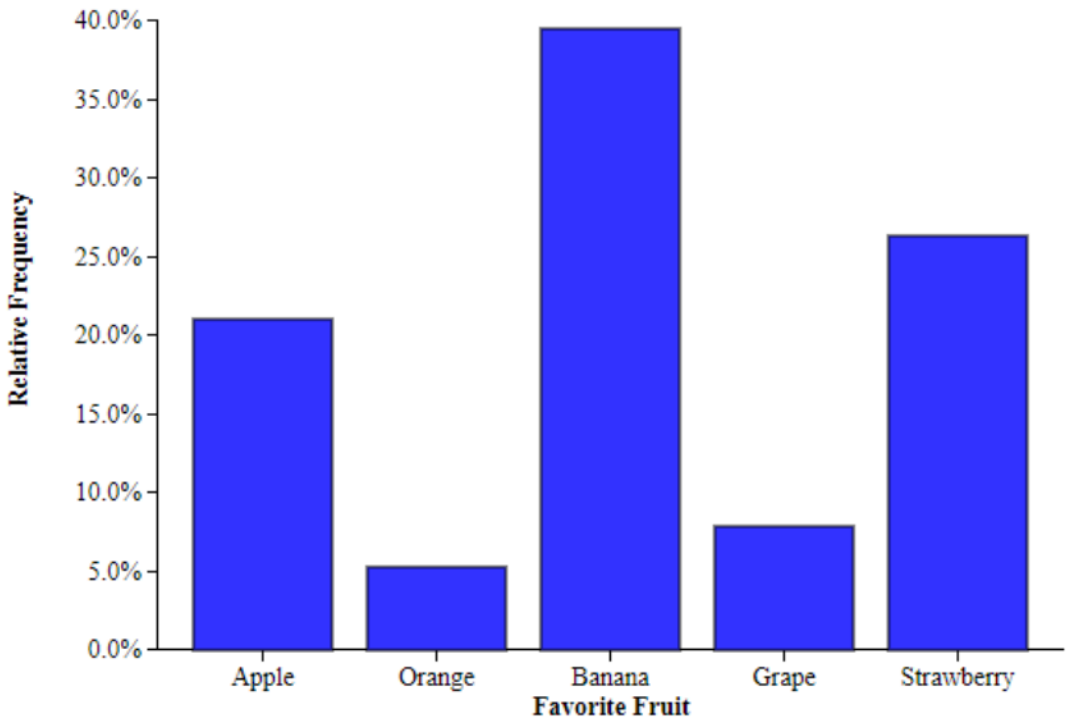
Solution to part a: 8 students selected apples and 3 selected grapes, meaning that $8+3=11$ is the numerator. We know that there are 38 students in total, thus, $\frac{11}{38} = \boxed{.289}$

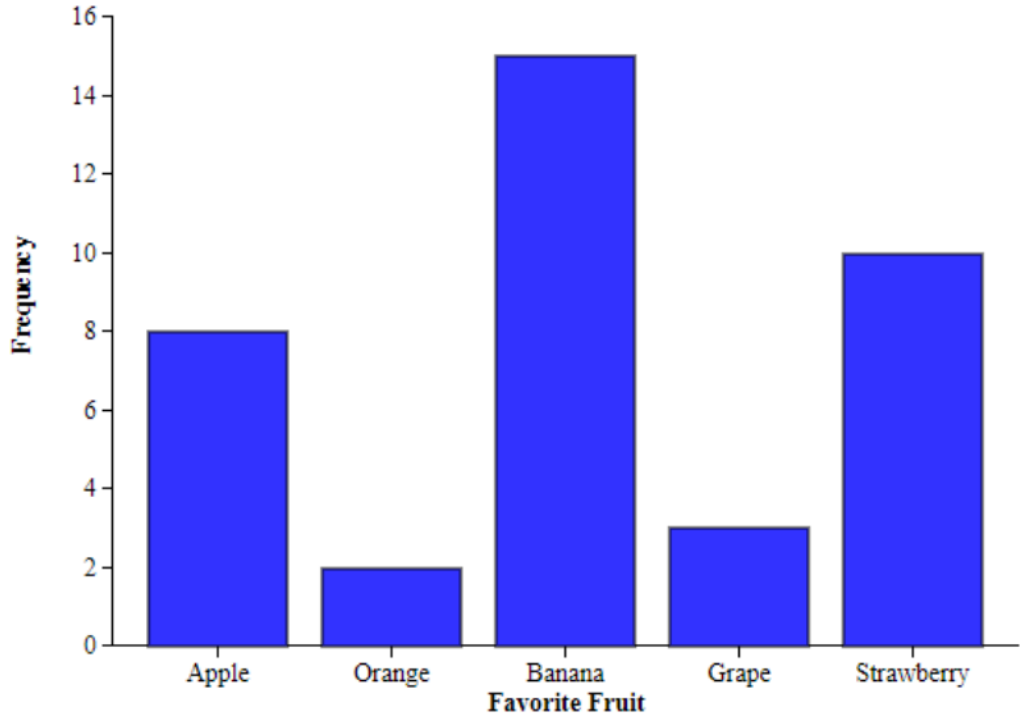
Solution to part b: $38-10=28$ yields the total students who did not select strawberries. 15 students in total picked banana, thus, the fraction of students that picked banana out of those who didn't pick strawberry is $\boxed{\frac{15}{28}}$.

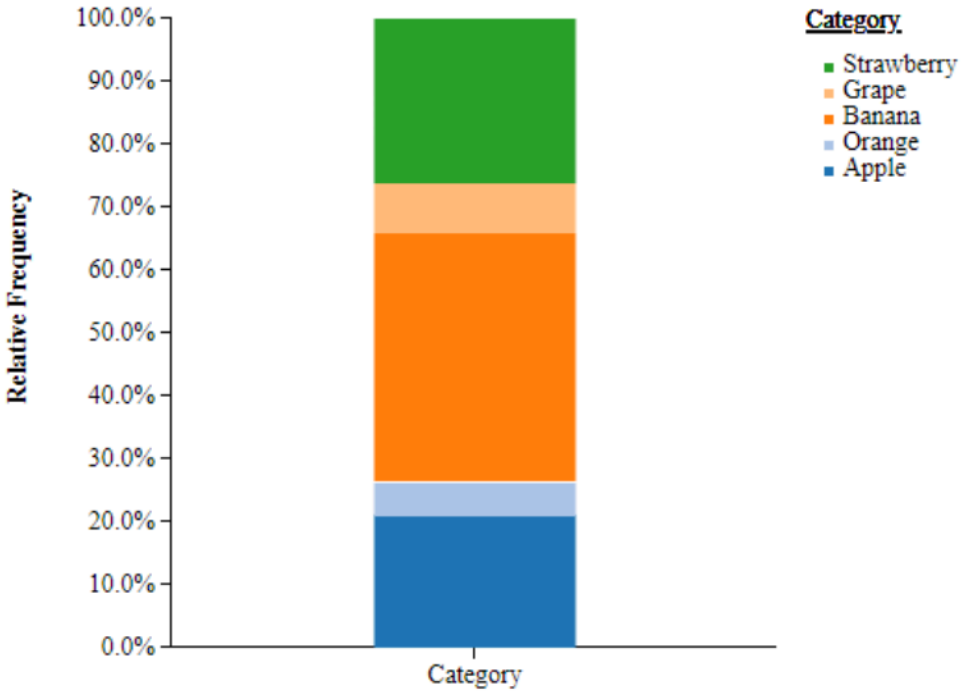
Solution to part c: 10 students picked strawberry, 2 picked orange, and 8 picked apple. Thus, $\frac{10+2+8}{38} = \frac{20}{38}$, which simplifies down to $\boxed{\frac{10}{19}}$.

Note 1.3.3

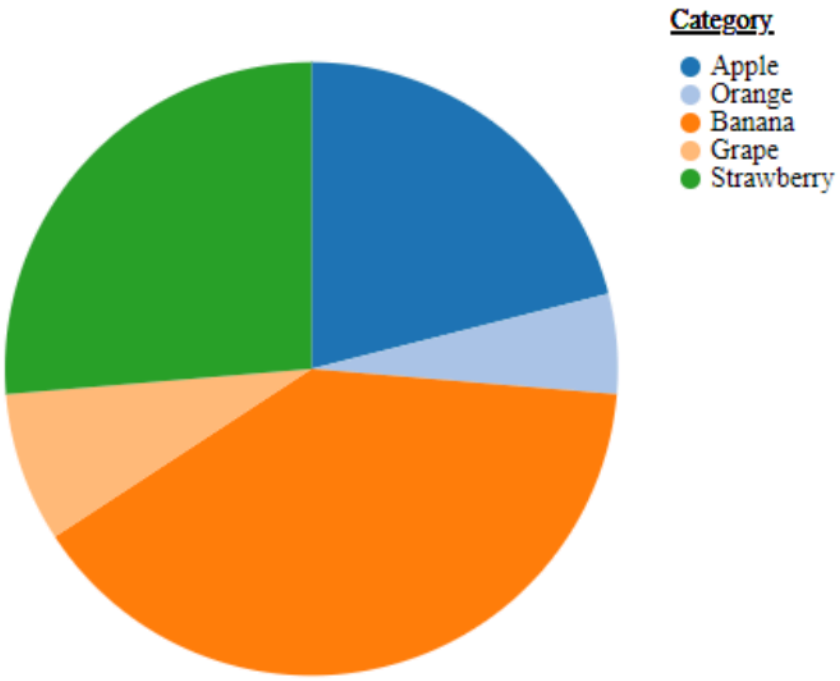
Categorical data can also be put into **bar graphs**. These can be relative frequency, frequency, or segmented graphs (graphs are shown below in that order). The height or length of each bar indicates the proportion or number of items in that category.







Notice that the relative frequency and frequency charts have the same shapes and sizes. The segmented bar chart shows all the data in one bar, with each segment representing a percentage. **Pie charts** are similar to segmented bar graphs, but rather than bars, it uses a circle, with each "piece" representing a percentage.

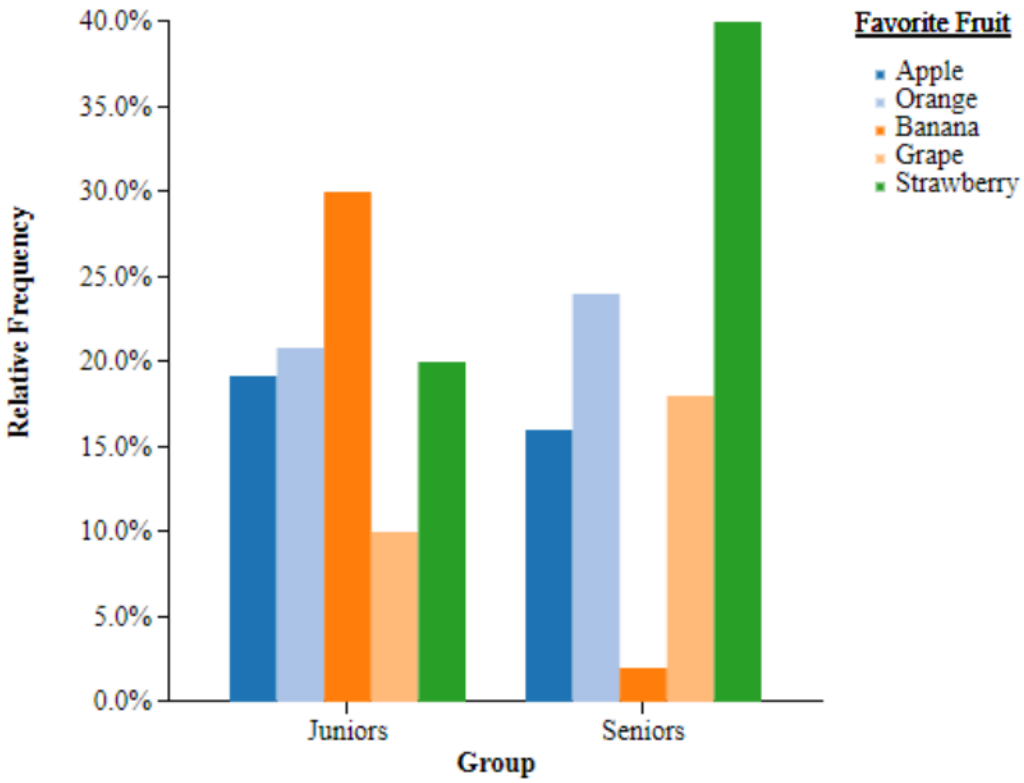


Note 1.3.4

You will also encounter categorical data with multiple groups. These will be in either tables or side-by-side bar graphs.

A sample of 120 Juniors and 100 Seniors was conducted where each participant was asked about their favorite fruit. The results are displayed here:

		Group		
		Juniors	Seniors	**Total**
Favorite Fruit	Apple	23 (19.2%)	16 (16%)	39 (17.7%)
	Orange	25 (20.8%)	24 (24%)	49 (22.3%)
	Banana	36 (30%)	2 (2%)	38 (17.3%)
	Grape	12 (10%)	18 (18%)	30 (13.6%)
	Strawberry	24 (20%)	40 (40%)	64 (29.1%)
	Total	120 (100%)	100 (100%)	**220 (100%)**



Problem 1.3.5 — From the bar graph, do a higher number of juniors or seniors prefer grapes as their favorite fruit?

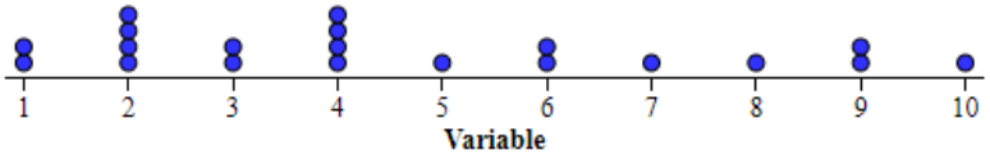
Solution: Since this is a bar graph showing relative frequency, we can not tell the specific **number** of juniors or seniors picking a particular fruit. Since there were a different amount of juniors and seniors selected, this data is not shown in the graph. Remember that graphs with relative frequencies do not show absolute numbers, only percentages.

§1.4 Visual Representations of Quantitative Variables

You will be asked more questions about **quantitative variables** than categorical on the AP test. Recall that quantitative variables can be either discrete or continuous, with discrete having a countable number of values and gaps, and continuous variables having no gaps and an infinite amount of values. For example, a continuous variable can take on all values from 0-1, which has an infinite number of decimals.

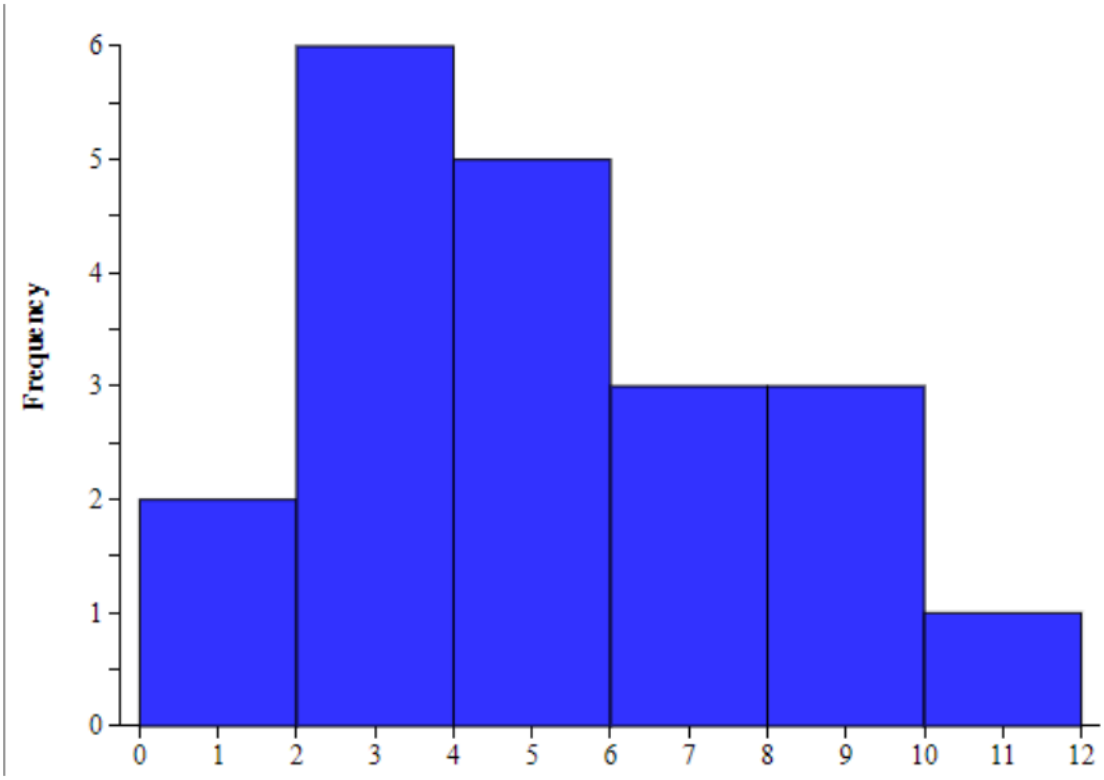
Note 1.4.1

Quantitative variables can also be represented in dotplots, stem and leaf plots, histograms, boxplots, or cumulative relative frequency (ogive) plots. Each of these have purposes and come with their own pros and cons.



Note 1.4.2

Each dot in a dotplot represents a data point. In a dotplot, the data is clearly displayed, allowing you to easily see the shape and each individual data point. However, for a large data set, it is hard to display the data in a dotplot. These are the best for discrete variables. The amount of dots shows the frequency.



Note 1.4.3

A histogram doesn't show each individual data point, rather it shows everything within a range. This histogram shows numbers from 0-2, 2-4, 4-6, 6-8, 8-10, and 10-12. You **can't** get exact data points, only a range. Histograms are good for larger data sets, and clearly display the shape. These are the best for continuous variables. The height of each bar displays the frequency.

0	1 1 2 2 2 2 3 3 4 4 4 4 5 6 6 7 8 9 9
1	0

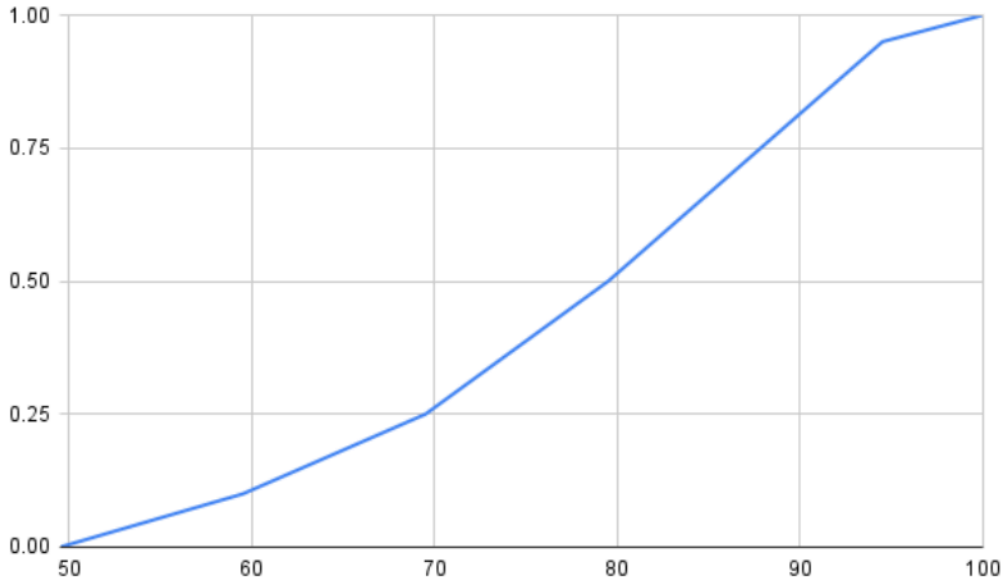
Note 1.4.4

In a stem and leaf plot, the right shows the ones place, and the left shows the numbers in the tens place. In the stem and leaf plot shown, the 1 on the left and 0 on the right represent a 10. It is easy to see each individual data point, gaps, and the shape, but it is difficult to create a stem and leaf plot for larger data sets.

Problem 1.4.5 — How many times does 9 appear in the stem and leaf plot?

Solution: Since 9 is a one digit number, we look to see where there is 0 on the left, which represents the tens place. From there, we can see that 09, which is 9, appears **twice**.

A teacher recorded the test scores of 40 students and plotted a cumulative relative frequency graph:



Note 1.4.6

A cumulative relative frequency graph (or ogive graph) shows the proportion less than or equal to a certain number. For example, the test score of approximately 80 has 50% of the data below that value. The y-axis shows the proportion.

Problem 1.4.7 — Use the graph to answer the following questions:

a) Which test score has 75% of the data below it?

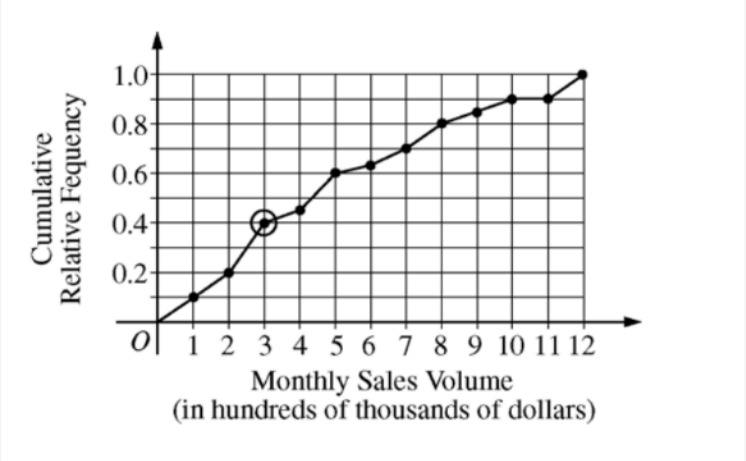
b) Which test score has 25% of the data below it?

Solution to part a: Recall that the y-axis shows the proportion below a certain value. Find what value is at 0.75. Thus, the test score of approximately $\boxed{87.5}$ has 75% of the data below it.

Solution to part b: The test score of approximately $\boxed{70}$ has 25% of the data below it.

Problem 1.4.8 — 2006 AP Statistics Form B

A large regional real estate company keeps records of home sales for each of its sales agents. Each month, the company publishes the sales volume for each agent. Monthly sales volume is defined as the total sales price of all homes sold by the agent during a month. The figure below displays the cumulative relative frequency plot of the most recent monthly sales volume (in hundreds of thousands of dollars) for these agents.



a) In the context of this question, explain what information is conveyed by the circled point.

b) What proportion of sales agents achieved monthly sales volumes between \$700,000 and \$800,000?

c) For values between 10 and 11 on the horizontal axis, the cumulative relative frequency plot is flat. In the context of this question, explain what this means.

d) A bonus is to be given to 20 percent of the sales agents. Those who achieved the highest monthly sales volume during the preceding month will receive a bonus. What is the minimum monthly sales volume an agent must have achieved to qualify for the bonus?

Solution to part a: This point means that 40% of agents have sales less than or equal to \$300,000.

Solution to part b: 80% of sales agents had sales volumes lower than or equal to \$800,000. 70% of sales agents had sales volumes less than or equal to \$700,000. Thus, 80% - 70% = 10% or $\boxed{0.1}$.

Solution to part c: The cumulative relative frequency plot being flat means that no agents had sales volumes between \$1,000,000 and \$1,100,000.

Solution to part d: The top 20 percent monthly sales volume is at the 80th percentile, which is \$800,000, thus, an agent needs a minimum monthly sales volume of $\boxed{\$800,000}$.

§1.5 Describing Distributions of Quantitative Variables

Distributions of quantitative variables can be described by describing the:

Note 1.5.1

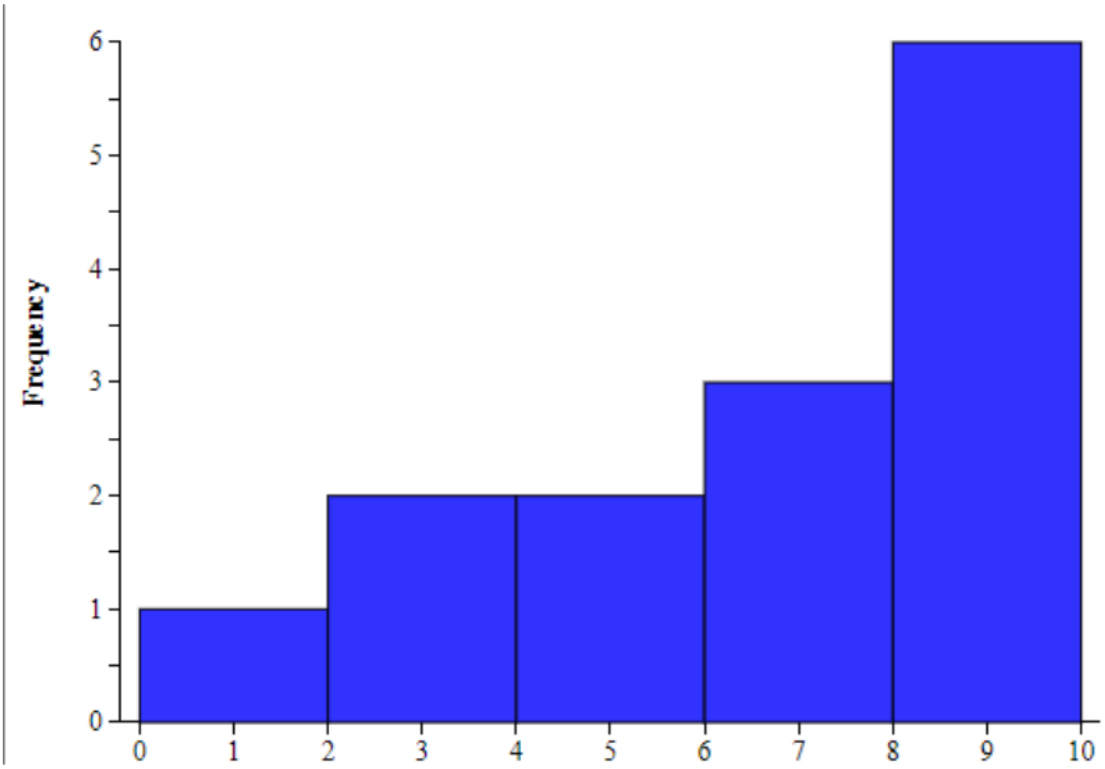
S.hape: Skewed left, skewed right, uniform, symmetric, bimodal, unimodal (describe any clusters as well)
O.utliers: Data points far away from most of the data, gaps
C.enter: Mean or median
V.ariability: Range, standard deviation, or IQR

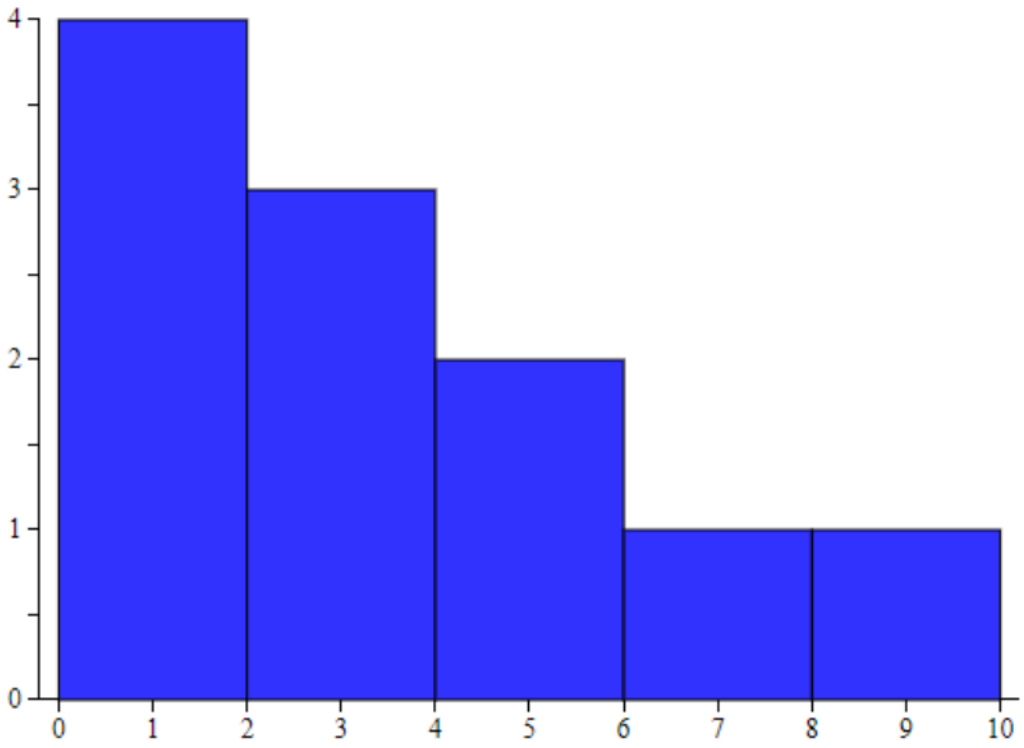
Remember this by writing out the abbreviation S.O.C.V.

Graphs of **quantitative variables** are in various shapes. It's important to know that **categorical variable** graphs do **not** have any shapes, and you can't describe the distribution of them. Only quantitative variables have distributions that take on shapes.

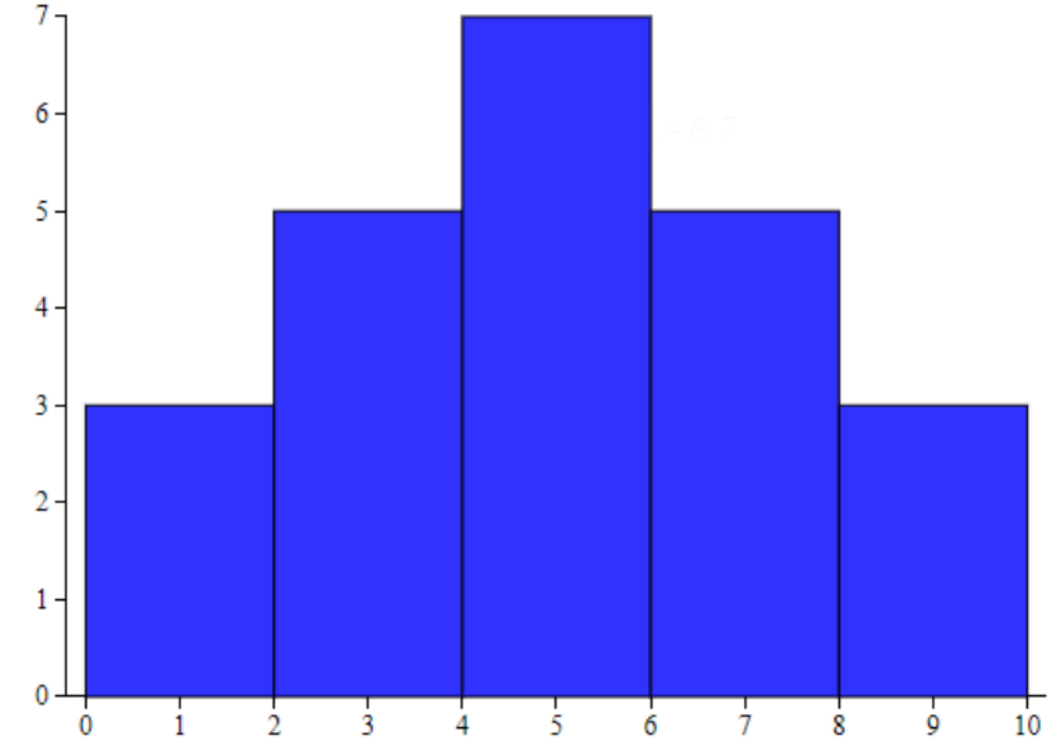
Note 1.5.2

Here are examples of graphs that are skewed left and right, respectively.



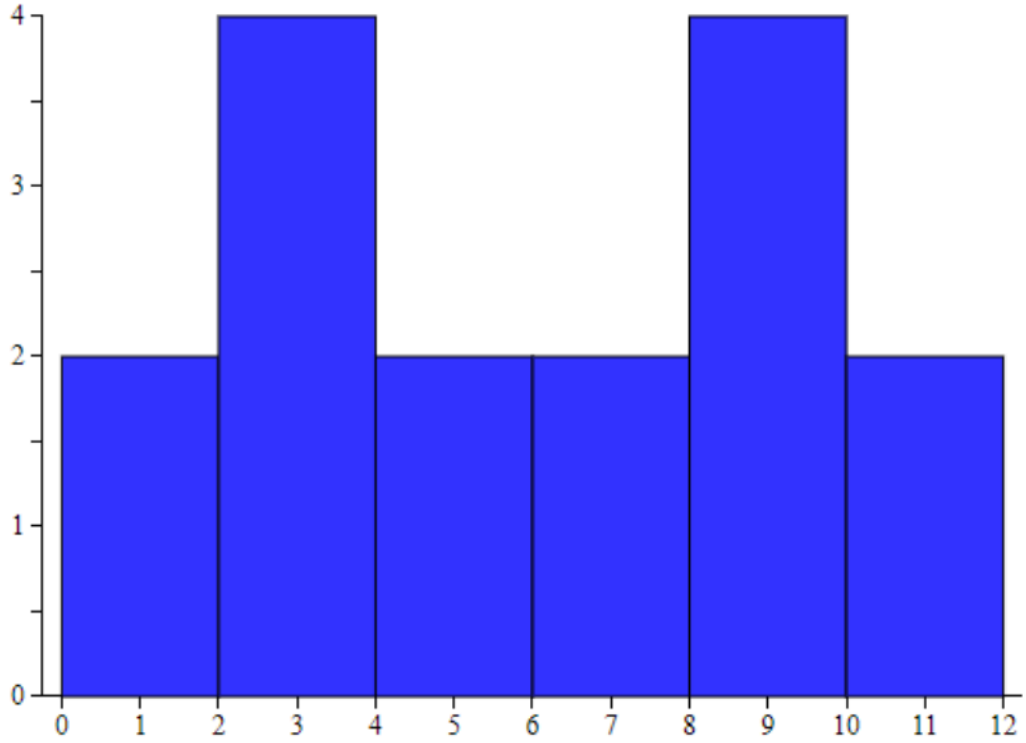


You can tell which direction a graph is skewed depending on the "tail" of it. The first graph, which is skewed left, has a "tail" of sorts created by infrequent data points further away from other values, which can sometimes be classified as **outliers** to the left. The second graph has a "tail" to the right, meaning it is skewed right.



Symmetric data is characterized by being roughly equal on both sides. There can be outliers in a symmetric graph, they just need to be on both sides. This is also an example of unimodal data. The mode is a a value that appears the most, and we can see that there is one peak, meaning this graph is unimodal. When describing the distribution of

a graph, never say that it's symmetric, say that it's approximately symmetric because perfectly symmetric data is rare.



Bimodal data is characterized by two peaks, meaning there are two values separate from each other that appear a significant amount of times. These do not have to be equal in size, but both clearly need to be peaks that are separated. This graph is also an example of symmetric data, but not all bimodal graphs are symmetric.

Uniform data is the same, or uniform throughout. The graph typically appears as a flat shape. You likely won't encounter this on FRQs, but knowing the definition is important for multiple choice questions.

Note 1.5.3

Always provide **context** when using S.O.C.V. Here are a few sentence frames to help you:
Shape: The shape of this distribution of [context] is [roughly skewed left/right, roughly symmetric, or roughly uniform] and [bimodal or unimodal].
Outliers:
For histograms: There are **potential outliers** at [ranges of outliers]
For other graphical displays: There are **potential outliers** at [data points]
If there are no outliers: There are no potential outliers.
Center:
The [mean or median] is approximately [value].
Variability:
The [IQR, standard deviation, or range] is approximately [value].

Remember that if you are able to find exact values, writing approximately is unneces-

sary.

Note 1.5.4

Outliers need to be listed in a distribution as well. The formal way of finding outliers will be explained in the next chapter, but these are characterized by being far away from the rest of the data. When describing a distribution, unless you can use the $1.5 \times$ IQR rule (will be explained in the next section), always call them potential outliers. In a histogram, give a range of where the outliers might be, since the data can't be exact. Gaps in the data also need to be described.

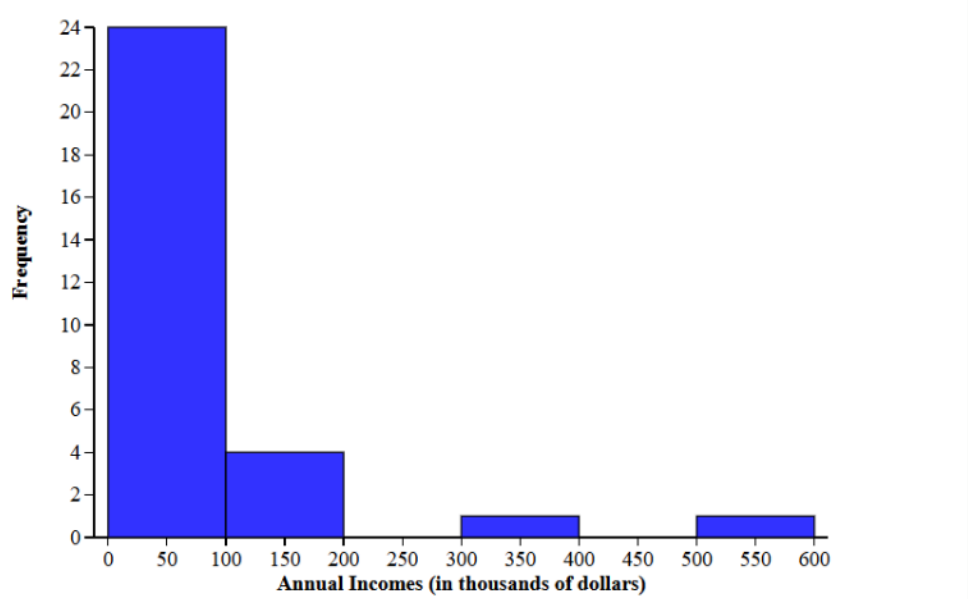
Note 1.5.5

The center of a distribution also needs to be explained. When using roughly symmetric data, the **mean** should be used as the center. The mean is heavily affected by outliers, since it is the average of all points in the data. The **median** should be used when describing skewed data, since it ends up being resistant to it, and not heavily affected.

Note 1.5.6

The last feature of a distribution you need to describe is the **variability**. This is described with either the standard deviation, range, or IQR. Standard deviation will be covered in the next chapter, but it's the average amount each value differs from the mean. The IQR is the 75th percentile value minus the 25th percentile value. The range is largest $-$ smallest value. When describing a histogram, say the range is **approximately** (the range), since you can't find exact values. IQR will be covered in the next section.

Problem 1.5.7 — 30 adults were sampled and asked about their annual incomes and the data was put into a histogram. Describe the distribution of income:



Solution: The shape of the distribution of annual income for adults is skewed right and unimodal. There are potential outliers at \$300,000-400,000 and \$500,000-600,000. The median income is from \$0-100,000, and the mean is greater than the median. The range is approximately \$600,000.

§1.6 Five Number Summary and Boxplots

The **mean** is calculated by adding up all the values, then dividing by the amount of values. It is the average.
The **median** is the middle of a data set put in order from least to greatest. When there is an odd amount of numbers, the middle number is the median. When there is an even number of values, the average of the two middle numbers is the median.

Problem 1.6.1 — Calculate the mean and median of these data sets:

a) 160, 165, 170, 172, 175, 180

b) 18, 22, 24, 26, 30, 35, 40

Solution to part a: There is an even amount of numbers, so the median is the average of the middle two values. The middle two values are 170 and 172, so the median is $\boxed{171}$. The mean is

$$\frac{160+165+170+172+175+180}{6} = \boxed{170.333}$$

Solution to part b: There is an odd amount of numbers, therefore the median is simply the middle number, which is $\boxed{26}$. The mean is

$$\frac{18+22+24+26+30+35+40}{7} = \boxed{27.857}$$

Percentile is the percentage of values less than or greater than that value. Q1 of a dataset is the 25th percentile. It is the median of the lower half of the data. Q3 is the 75th percentile, and median of the upper half of the data. IQR is Q3 − Q1, which represents the middle 50% of the data.

Problem 1.6.2 — Find the minimum, Q1, median, Q3, maximum, and IQR of these datasets:

a) 3, 7, 6, 10, 4, 8, 5, 12

b) 52, 68, 45, 74, 81, 63, 70, 49, 89

Solution to part a: The minimum value is 3 and the maximum value is 12. The data is out of order, so we need to put it from smallest to largest to find the Q1, median, and Q3. Rearranging, the data becomes

$$3, 4, 5, 6, 7, 8, 10, 12$$

The median is the average of 6 and 7, which is **6.5**. Q1 is the average of 4 and 5, which is **4.5**. Q3 is the average of 8 and 10, which is **9**. Thus, the IQR is $9 - 4.5$, which is **4.5**.

Solution to part b: Rearranging, the data becomes

$$45, 49, 52, 63, 68, 70, 74, 81, 89$$

The minimum is 45, and the maximum is 89. The median is simply the middle since there is an odd amount of values, which is **68**. When finding Q1, we look at the lower half excluding the median. Thus, we are looking for the average of 49 and 52, which is **50.5**. When you have an odd number of data points, the median is not part of the halves used for calculating Q1 and Q3. The same goes for Q3, but with the upper half. We are looking for the average of 74 and 81, which is **77.5**. Thus, the IQR is $77.5 - 50.5$, which is **27**.

Standard deviation is found with this formula:
$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$.
Where N is the amount of numbers in the data set, x_i is the data point, $\overline{x}$ is the mean. To calculate the standard deviation, you can calculate the sum by plugging in each value (as shown in the next example), or use the 1-variable statistics function on your calculator. Standard deviation measures the spread of the data, it is the average amount a number differs from the mean.

Example 1.6.3

Find the standard deviation of the dataset:
3, 7, 6, 10, 4, 8, 5, 12

Solution:

$$\mu = \frac{3+7+6+10+4+8+5+12}{8} = 6.875$$

$$(3-6.875)^2 = 15.016, \quad (7-6.875)^2 = 0.016, \quad (6-6.875)^2 = 0.766, \quad (10-6.875)^2 = 9.766$$

$$(4-6.875)^2 = 8.266, \quad (8-6.875)^2 = 1.266, \quad (5-6.875)^2 = 3.516, \quad (12-6.875)^2 = 26.266$$

$$\sqrt{\frac{15.016 + 0.016 + 0.766 + 9.766 + 8.266 + 1.266 + 3.516 + 26.266}{7}} = \sqrt{9.265} \approx 3.04$$

It is better to use the 1-variable statistics function on your calculator by entering the data into a table even if it is possible to find the standard deviation by hand. When calculating the standard deviation, show the data plugged into the formula, even if using the 1-variable function.

One way to find an **outlier** is by calculating if there is any data point greater than $1.5 \times IQR + Q3$, or $1.5 \times IQR$ less than Q1. The "fences" for outliers can be found by doing $Q3 + 1.5 \times IQR$ (upper), and $Q1 - 1.5 \times IQR$ (lower). Any value outside of this is an outlier. This method of finding outliers works well for skewed distributions because IQR and median are **resistant** to outliers, however, the next method should be used for roughly symmetric data.

Problem 1.6.4 — Find the fence for outliers in these datasets:

a) 3, 5, 7, 10, 6, 8, 4

b) 1.2, 0.8, 1.5, 1.0, 2.3, 1.7, 0.9, 2.0

Solution to part a: Rearranging, the data becomes

$$3, 4, 5, 6, 7, 8, 10$$

The median is 6, so we look for the median of the lower half excluding the median, meaning 4 is Q1. Q3 is the median of the upper half of the data, which is 8. $8 - 4 = 4$ which gives us our IQR. $1.5 \times$ IQR becomes 1.5×4, which is 6. The lower boundary is $4 - 6$ which is -2, and the upper boundary is $8 + 6$ which is 14. All values are contained within these boundaries, so there are no outliers.

Solution to part b: After rearranging, the data becomes

$$0.8, 0.9, 1.0, 1.2, 1.5, 1.7, 2.0, 2.3$$

Q1 is the average of 0.9 and 1, which is .95. Q3 is the average of 1.7 and 2.0, which is 1.85. The IQR is $1.85 - .95$, which is 0.9. $1.5 \times$ IQR is 1.35. $1.85 + 1.35$ is 3.2, which is the upper boundary. $.95 - 1.35$ is -0.4, which is the lower boundary. All values are contained within -0.4 and 3.2, so there are no outliers.

The other way of finding outliers is by using the two standard deviations rule. If a value is two standard deviations above or below the mean, it is an outlier. This works for roughly symmetric data and not skewed data because means and standard deviations are heavily affected by outliers.

Problem 1.6.5 — Find the upper and lower fence using the two standard deviations rule and calculate any outliers:

a) 5, 7, 8, 9, 10, 12, 25

b) 50, 52, 53, 54, 55, 56, 57, 70

Solution to part a: The mean of the data is

$$\frac{5 + 7 + 8 + 9 + 10 + 12 + 25}{7} = 10.857$$

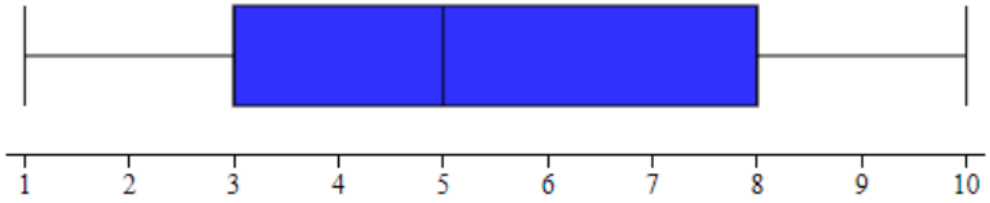
The standard deviation is approximately 6.619. The upper boundary is $10.857 + 2 \times 6.619$ = 24.095. The lower boundary is $10.857 - 2 \times 6.619$ = -2.381. 25 is outside these boundaries, therefore, it is an outlier.

Solution to part b: The mean of the data is

$$\frac{50 + 52 + 53 + 54 + 55 + 56 + 57 + 70}{8} = 55.875$$

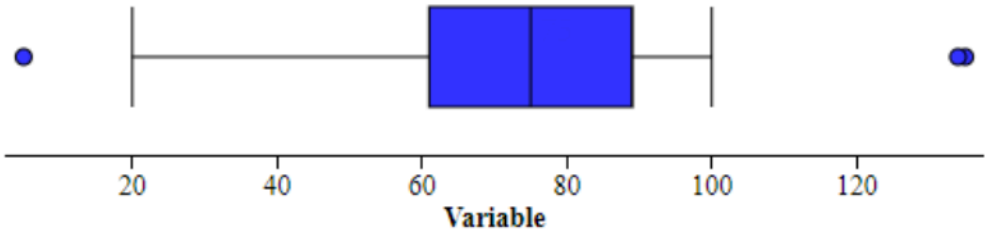
The standard deviation is roughly 6.1281. Using the two standard deviations rule, $55.875 + 2 \times 6.1281 = 68.1312$ is the upper boundary, and $55.875 - 2 \times 6.1281 = 43.6188$ is the lower boundary. 70 is outside these boundaries, so it is an outlier.

The minimum, Q_1, median, Q_3, and maximum together form a 5 number summary. This can be represented in a boxplot as shown here:



n	mean	SD	min	Q_1	med	Q_3	max
10	5.4	3.026	1	3	5	8	10

The box shown in a boxplot represents the IQR, with Q1 being the start of the box, median being the middle, and Q3 being the end. The five number summary is easily visible with the boxplot. The leftmost line is the minimum, then Q_1, then median, then Q_3, then the maximum. In a boxplot, the leftmost and rightmost lines (the whiskers) represent the smallest or largest value where there isn't an outlier. Here is an example of a boxplot with outliers:



Summary Statistics

n	mean	SD	min	Q_1	med	Q_3	max
55	74.364	22.903	5	61	75	89	135

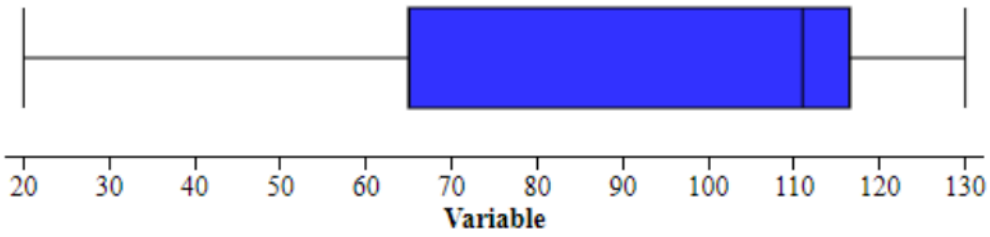
Notice that the whiskers don't extend to outliers, they instead go to the furthest point that isn't an outlier.

Note 1.6.6

Advantages of a boxplot: Easy to see outliers and the five number summary, can see if the data is skewed or symmetric.

Disadvantages: Unable to see gaps and clusters, or if the data is unimodal or bimodal, and individual values aren't shown.

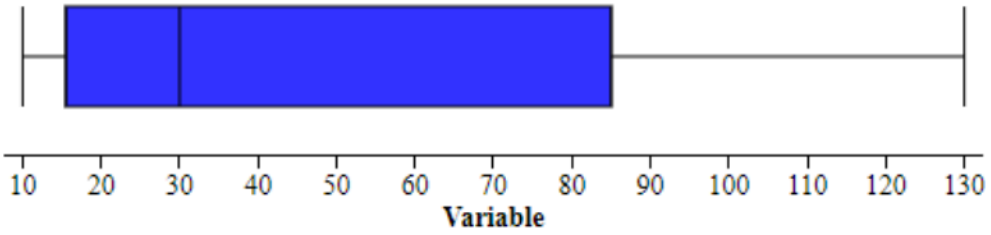
A skewed right distribution has a mean that is to the right of the median, and a skewed left distribution has a mean that is to the left of the median. Here are two graphs to illustrate:



Summary Statistics

n	mean	SD	min	Q_1	med	Q_3	max
21	92.143	33.607	20	65	111	116.5	130

Notice that the mean of 92.143 is much smaller than the median of 111 in this skewed left boxplot. A skewed right boxplot that follows the same rule is displayed here:



Summary Statistics

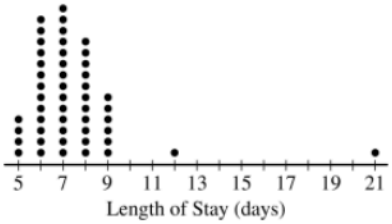
n	mean	SD	min	Q_1	med	Q_3	max
21	49.238	40.591	10	15.5	30	85	130

The mean of 49.238 is significantly higher than the median of 30. Boxplots can be skewed without outliers, as shown above.

Problem 1.6.7 — 2021 AP Statistics

The length of stay in a hospital after receiving a particular treatment is of interest to the patient, the hospital, and insurance providers. Of particular interest are unusually short or long lengths of stay. A random sample of 50 patients who received the treatment was selected, and the length of stay, in number of days, was recorded for each patient. The results are summarized in the following table and are shown in the dotplot.

Length of stay (days)	5	6	7	8	9	12	21
Number of patients	4	13	14	11	6	1	1



a) Determine the five-number summary of the distribution of length of stay.

b) Consider two rules for identifying outliers, method A and method B. Let method A represent the 1.5 * IQR rule, and let method B represent the 2 standard deviations rule.

   i) Using method A, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

   ii) The mean length of stay for the sample is 7.42 days with a standard deviation of 2.37 days. Using method B, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

c) Explain why method A might identify more data points as potential outliers than method B for a distribution that is strongly skewed to the right.

Solution to part a: Since there are 50 patients in total, Q1 can be found by looking for the middle of the 12th and 13th numbers, which is 6. The median is the 25th number, which is 7. Q3 is the the middle of the 37th and 38th number, which is 8. The minimum is clearly 5 and the maximum is clearly 21. Thus, the five number summary is:

$$\text{Minimum: } 5,\ Q1\text{: } 6,\ \text{Median: } 7,\ Q3\text{: } 8,\ \text{Maximum: } 21.$$

Solution to part bi: $8-6$ is 2, which is the IQR. IQR multiplied by 1.5 is $2 \times 1.5 = 3$. The lower boundary is $6-3=3$, and the upper boundary is $8+3=11$. Since 12 and 21 are outside these boundaries, those days are outliers.

Solution to part bii: The upper fence for outliers is $7.42 + 2 \times 2.37 = 12.16$ days. The lower fence for outliers is $7.42 - 2 \times 2.37 = 2.68$ days. Since 21 days is outside the boundaries, it is an outlier.

Solution to part c: Median and IQR are more resistant to outliers compared to mean and standard deviation. When there is an outlier, the mean is strongly pulled in that direction. Since the mean is heavily pulled to the right, the effectiveness of method B is lessened, because it is less likely to detect outliers. The mean is increased, meaning the

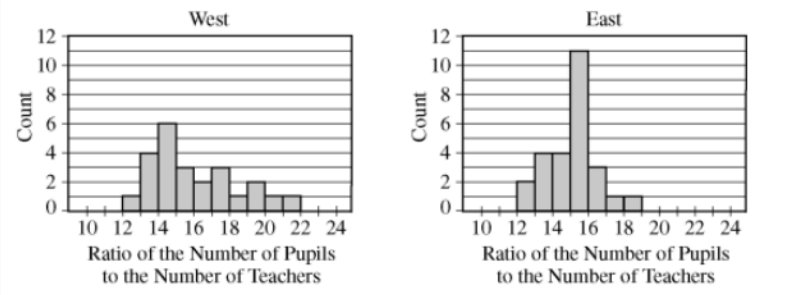
standard deviation is also increased, making method B even less effective. The standard deviation changes more than the IQR from outliers, therefore, the mean and standard deviation method is less likely to identify outliers when there is skewed data.

§1.7 Comparing Distributions of Quantitative Variables

When **comparing distributions**, use S.O.C.V, but with comparative language. This involves directly stating and comparing each part of S.O.C.V and listing which characteristics are greater than, less than, or the same as the other distribution.

Problem 1.7.1 — 2011 AP Statistics

Records are kept by each state in the United States on the number of pupils enrolled in public schools and the number of teachers employed by public schools for each school year. From these records, the ratio of the number of pupils to the number of teachers (P-T ratio) can be calculated for each state. The histograms below show the P-T ratio for every state during the 2001–2002 school year. The histogram on the left displays the ratios for the 24 states that are west of the Mississippi River, and the histogram on the right displays the ratios for the 26 states that are east of the Mississippi River.



a) Describe how you would use the histograms to estimate the median P-T ratio for each group (west and east) of states. Then use this procedure to estimate the median of the west group and the median of the east group.

b) Write a few sentences comparing the distributions of P-T ratios for states in the two groups (west and east) during the 2001–2002 school year.

c) Using your answers in parts (a) and (b), explain how you think the mean P-T ratio during the 2001–2002 school year will compare for the two groups (west and east).

Solution to part a: To find the median P-T Ratio of the states west of the Mississippi River, we find what ratio is between the 12th and 13th values, since there is an even amount of numbers. The number between the 12th and 13th values is a ratio of 15 to 16. For the eastern states, we find the ratio that is between the 13th and 14th values. The number between the 13th and 14th values is a ratio of 15 to 16. Since both graphs are histograms, exact values cannot be found. The median P-T ratio of both groups is from $\boxed{\text{15 to 16}}$.

Solution to part b: The shape of the western states is unimodal and skewed to the right, whereas the shape of the eastern states is unimodal and roughly symmetric. There do not appear to be any potential outliers in both histograms. The medians of both histograms are approximately the same, with a median ratio from 15 to 16. The range of

the western states is $22 - 12 =$ roughly 10 and the range of the eastern states is $19 - 12 =$ approximately 7.

Solution to part c: The western states have a histogram that is skewed to the right, whereas the eastern states are approximately symmetric. The medians are approximately equal, but since the western states have a skewed right distribution, the mean is likely greater than the eastern states. In a roughly symmetric distribution, the median is close to the mean, but in data skewed to the right, the mean is greater than the median.

§1.8 The Normal Distribution

When a distribution is unimodal, approximately symmetric, and bell-shaped we can use the normal distribution. This can be used to describe results of various sampling procedures you will see later on in the course, and easily shows probability, percentiles, and number of standard deviations away a value is. The mean and standard deviation are used to describe a normal distribution.

Note 1.8.1

Percentile:
Percentile is the relative position of a value in a data set, it's the percentage of data less than or equal to a value.

Note 1.8.2

Standardized Score:
In statistics, the standardized score is the number of standard deviations a data value is. This is found with the general formula: standardized score $= \frac{\text{data}-\text{mean}}{\text{standard deviation}}$. When we specifically use the **z-score** as our standardized score, we use the formula: $z = \frac{x-\mu}{\sigma}$. We take a data point, x, then subtract that by the mean, μ, then dividing the difference of that by the standard deviation, σ. This gives us how many standard deviations a value x is.

Problem 1.8.3 — Solve each question:

a) A student takes a test where the class average is 75 and has a standard deviation of 10. The student scores 90, what is their z-score?

b) The daily temperature of a city is degree 70 degrees Fahrenheit with a standard deviation of 4. On a specific day, the temperature is 72, what is the z-score of that day's temperature?

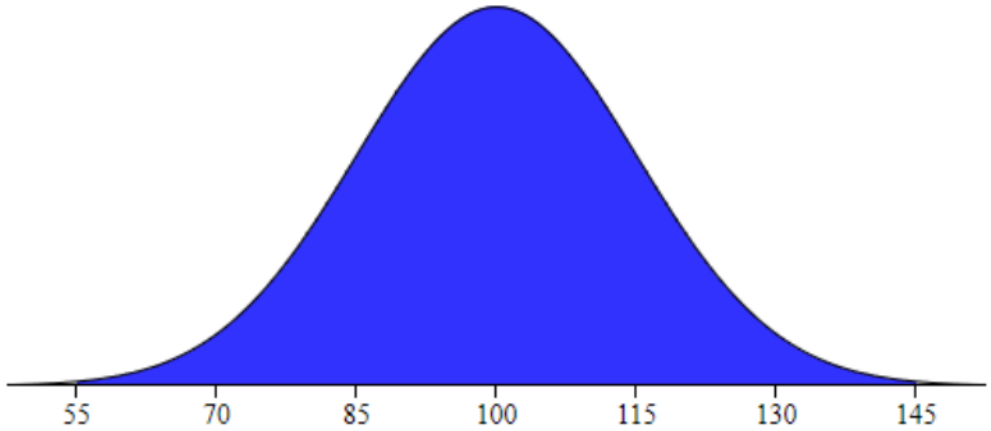
c) The average race time for a 5k is 110 minutes and the standard deviation is 20 minutes. If a runner completes the race in 130 minutes, what is their z-score?

Solution to part a: The z-score is $\frac{90-75}{10} = \boxed{1.5}$

Solution to part b: The z-score is $\frac{72-70}{4} = \boxed{.5}$

Solution to part c: The z-score is $\frac{130-110}{20} = \boxed{1.0}$

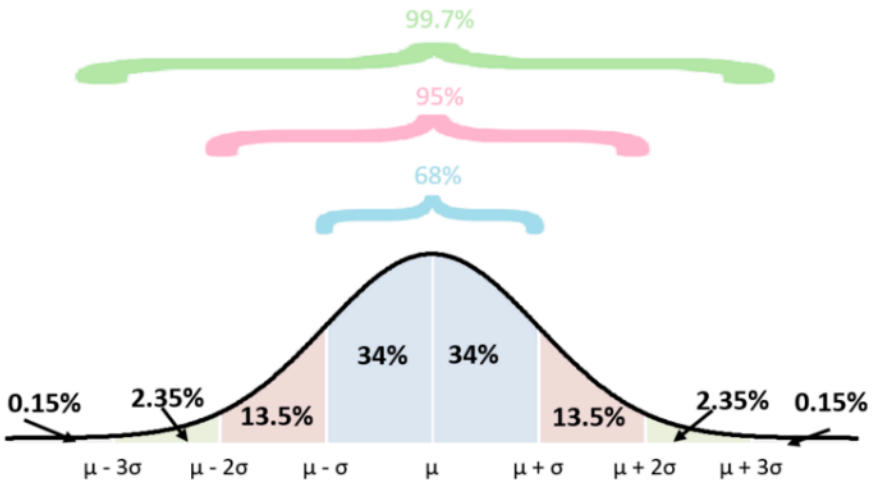
Normal distributions are determined by the mean and standard deviation. There are various applications to this, including IQ, weight, and height. The **area** under the curve represents percentage/probability.
A normal distribution for IQ is modeled here:



The mean in a bell-curve is the center, which here is **100**, and the standard deviation is 15.

Note 1.8.4

Within **one standard deviation of the data**, approximately **68%** of the data is captured. In this graph, **68%** of the data is from 85-115 IQ.
Within **two standard deviations of the data**, approximately **95%** of the data is captured. This is from 70-130 IQ in this graph.
Within **three standard deviations of the data**, approximately **99.7%** of the data is captured.
This is called the **empirical rule**, or the 68-95-99.7 rule.



Problem 1.8.5 — Solve each of these questions using the normal distribution:

a) What percentage of values are between 1 standard deviation below and 2 standard deviations above the mean?

b) What percentage of values are between 3 standard deviations below and 1 standard deviations above the mean?

c) What percentage of values are between 2 standard deviations below and 2 standard deviations above the mean?

Solution to part a: One standard deviation below has 34% of the data and two standard deviations above has $34 + 13.5 = 47.5\%$ of the data. $34\% + 47.5\% = \boxed{81.5\%}$.

Solution to part b: Three standard deviations below the mean contain $34 + 13.5 + 2.35 = 49.85\%$ of the data. One standard deviation above is 34% of the data. $49.85 + 34 = \boxed{83.85\%}$.

Solution to part c: Using the two standard deviations rule, we already know that approximately 95% of the data is contained within two standard deviations below and above the mean. However, this can be computed using $34 + 13.5 = 47.5$ to find what percent is below, and $34 + 13.5 = 47.5$ to find the percent above. Thus, $47.5 + 47.5 = \boxed{95\%}$

Proportions can also be found through the normal distribution. This can be accomplished by using the z-score table provided on your equation sheet, or a calculator. Both sides of Table A on the equation sheet are for z-scores. The p-value attached to these scores tell you the area under that specific value, which is the proportion. Read the table by looking at the z-score values on the left, then matching up where the hundredths place is. To use a calculator to calculate p-value, use the normalCDF function. Use the data point you desire as the upper value, and -99999 as the lower value to emulate negative infinity. If you are looking for the proportion **higher** than a certain point, do 1 minus the percentage yielded from the z-score. If using a calculator, set the point as the lower value and 99999 as the upper value.

Problem 1.8.6 — Answer the following questions:

a) In a school, the average height is 165 cm with a standard deviation of 10. A student is 180 cm tall, what percentage of students are shorter than this student?

b) The average commute time in a city is 30 minutes with a standard deviation of 5 minutes. If a person takes 22 minutes to commute to work, what percentage of commutes are longer than this commute?

c) The average price of a used car is 18,000 dollars with a standard deviation of 3000 dollars. A car is listed for 10000 dollars, what percentage of used cars are more expensive than it?

Solution to part a: We can use normalCDF to find the percentage. When using your calculator to find percentages, always show what you input on paper.

$$\text{normalCDF}\ (\text{lower} = -\infty, \text{upper} = 180, \mu = 165, \sigma = 10) = .9332 \text{ or } \boxed{93.32\%}$$

If using z-score, plug the values into the formula. $\frac{180-165}{10} = 1.5$. From there, look at the z-score table to find the percentage.

Solution to part b: Using normalCDF, we have

$$\text{normalCDF}(\text{lower} = 22, \text{upper} = \infty, \mu = 30, \sigma = 5) = .9452 \text{ or } \boxed{94.52\%}$$

Using z-scores, we have $\frac{22-30}{5}$ = -1.6. The z-score chart gives us the percentage **below** a certain value, so we need to do 1 minus the percentage in order to find the percentage **higher**.

Solution to part c: Using normalCDF, we have

$$\text{normalCDF}(\text{lower} = 10{,}000, \text{upper} = \infty, \mu = 18{,}000, \sigma = 3000) = .9962 \text{ or } \boxed{99.62\%}$$

Using z-scores, we have $\frac{10000-18000}{3000}$ = -2.67. Using the z-score chart, we do 1 minus the proportion yielded from the z-score.

You will often calculate z-score values for specific data points, then be asked to find the proportion/area between those same points. Take the proportion from the higher z-score value and subtract it from the proportion from the lower z-score value to find what is in between. If using a calculator, just set the lower value as lower and the upper value as upper.

Problem 1.8.7 — Answer the following questions:

a) The heights of adult women in a city have a mean of 160 cm and standard deviation of 8 cm. What proportion of women have a height between 155 and 170 cm?

b) The exam scores for a class have a mean of 75 and standard deviation of 10. What proportion of students scored between 70 and 85 on the exam?

Solution to part a: Using normalCDF, we have

$$\text{normalCDF}(\text{lower} = 155, \text{upper} = 170, \mu = 160, \sigma = 8) = \boxed{.6284}$$

We need to find the z-scores of both 155 and 170. The z-score of 155 is $\frac{155-160}{8}$ = -.625, and the z-score of 170 is $\frac{170-160}{8} = 1.25$ The proportion given by a z-score of 1.25 is .8944, and the proportion from -.625 is .2676. $.8944 - .2676 = .6268$. Don't worry about the proportions being slightly different from using the calculator, the z-score tables are not exact.

Solution to part b: Using normalCDF, we have

$$\text{normalCDF}(\text{lower} = 70, \text{upper} = 85, \mu = 75, \sigma = 10) = \boxed{.5328}$$

The z-score of 70 is $\frac{70-75}{10}$ = -.5. The z-score of 85 is $\frac{85-75}{10} = 1$. The proportion yielded from a z-score of 1 is .8413 and the proportion yielded from a z-score of -.5 is .3085. $.8413 - .3085 = .5328$.

If you are given a z-score and missing one value, you will need to use algebraic methods in order to solve for a value. Sometimes, you will not be given a z-score, and instead a proportion. You can either look at the z-score table and find the proportion, or use the **invNorm** function on your calculator. Set μ to 0 and σ to 1 and enter the proportion that is to the left or right of a value.

Problem 1.8.8 — Solve each question:

a) A class has a mean score of 80 and a standard deviation of 10. A student has a score that has a z-score of 2.5. What is the student's test score?

b) The distribution of exam scores has a mean of 50 and standard deviation of 5. If a student has a score that is greater than 15.87% of the class, what did they score?

Solution to part a: From the information given, we have $2.5 = \frac{x-80}{10}$. Multiplying by 10 on each side, we have $25 = x - 80$. After adding 80 on each side, we have $\boxed{105 = x}$.

Solution to part b: First, we need to find the z-score that has a proportion of .1587. Using invNorm(.1587) (there is no need to label the $\mu = 0$ and $\sigma = 1$), we have a z-score of approximately -1. From there, we have $-1 = \frac{x-50}{5}$. Multiplying by 5 on each side, we get $-5 = x - 50$. Adding by 50 on each side, we get $\boxed{45 = x}$.

Problem 1.8.9 — 2011 AP Statistics

A professional sports team evaluates potential players for a certain position based on two main characteristics, speed and strength.

a) Speed is measured by the time required to run a distance of 40 yards, with smaller times indicating more desirable (faster) speeds. From previous speed data for all players in this position, the times to run 40 yards have a mean of 4.60 seconds and a standard deviation of 0.15 seconds, with a minimum time of 4.40 seconds, as shown in the table below.

	Mean	Standard Deviation	Minimum
Time to run 40 yards	4.60 seconds	0.15 seconds	4.40 seconds

Based on the relationship between the mean, standard deviation, and minimum time, is it reasonable to believe that the distribution of 40-yard running times is approximately normal? Explain.

b) Strength is measured by the amount of weight lifted, with more weight indicating more desirable (greater) strength. From previous strength data for all players in this position, the amount of weight lifted has a mean of 310 pounds and a standard deviation of 25 pounds, as shown in the table below.

	Mean	Standard Deviation
Amount of weight lifted	310 pounds	25 pounds

Calculate and interpret the z-score for a player in this position who can lift a weight of 370 pounds.

c) The characteristics of speed and strength are considered to be of equal importance to the team in selecting a player for the position. Based on the information about the means and standard deviations of the speed and strength data for all players and the measurements listed in the table below for Players A and B, which player should the team select if the team can only select one of the two players? Justify your answer.

	Player A	Player B
Time to run 40 yards	4.42 seconds	4.57 seconds
Amount of weight lifted	370 pounds	375 pounds

Solution to part a: It is not reasonable to believe that the distribution of 40-yard running times is approximately normal because the minimum is $\frac{4.40-4.60}{0.15}$ = -1.333 standard deviations away from the mean. This z-score has 9.12% of the data below it in a normal distribution, but it is the minimum here, meaning none of the data is below it and the distribution is not approximately normal.

Solution to part b: The z-score for a player who can lift a weight of 370 pounds is $\frac{370-310}{25} = 2.4$. This z-score means that a player lifting a weight of 370 pounds is 2.4 standard deviations above the average.

Solution to part c: To find out which player the team should select, we need to compare how the speeds and strengths of both players are relative to the mean. To find this, we

find the z-scores. For Player A, the z-score of their speed is $\frac{4.42-4.60}{0.15}$ = -1.2. The z-score of their strength is $\frac{370-310}{25}$ = 2.4. For Player B, the z-score of their speed is $\frac{4.57-4.60}{0.15}$ = -0.2. The z-score of their strength is $\frac{375-310}{25}$ = 2.6. Player B is much slower (remember, the smaller the z-score for speed, the faster they are) but only has marginally larger strength, therefore, Player A should be selected.

2 Unit 2: (Exploring Two-Variable Data)

§2.1 Relationships Between Two Categorical Variables

Data involving two categorical variables can be represented as side-by-side bar graphs, segmented bar graphs, mosaic plots, or two-way tables.

Note 2.1.1

126 people were sampled and then asked about their gender and mode of transportation. The results are displayed as a two-way table, side-by-side bar graph, segmented bar graph, and mosaic plot here:

	Male	Female	Other
Car	15	20	5
Bicycle	10	25	7
Bus	12	18	8
Train	3	2	1

Problem 2.1.2 — Answer the following questions:

a) What proportion of people were female and used a bicycle or bus as their mode of transportation?

b) What is the relative frequency of males sampled?

c) What is the fraction of people who said cars were their mode of transportation?

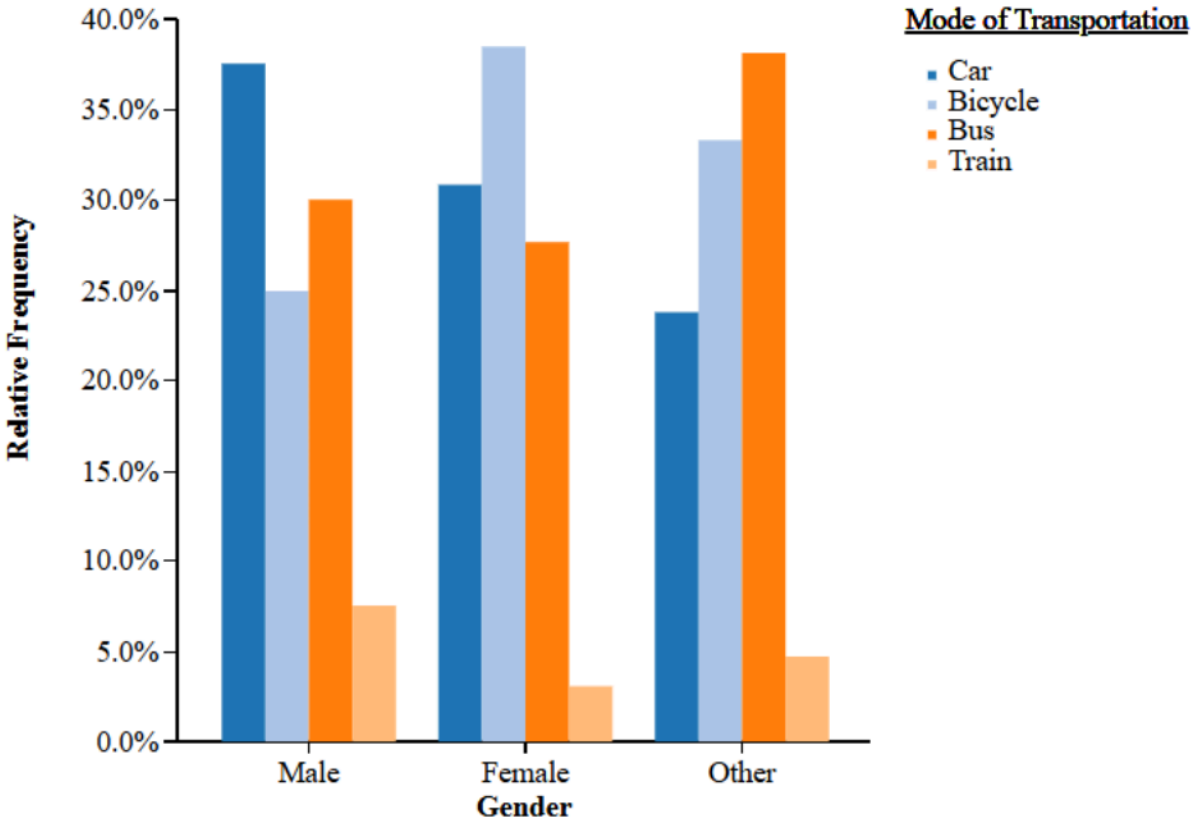
d) What percentage of those who picked bus are men?

Solution to part a: 25 females chose bicycle and 18 chose bus as their preferred mode of transportation. Thus, we have $\frac{25+18}{126} = \boxed{0.341}$.

Solution to part b: There are $15 + 10 + 12 + 3 = 40$ males in total sampled. Since there were 126 people in total, the relative frequency is $\frac{40}{126} = \boxed{0.317}$.

Solution to part c: $15 + 20 + 5 = 40$, so $\frac{40}{126}$ represents the fraction of people who said cars were their mode of transportation. This can be simplified down to $\boxed{\frac{20}{63}}$.

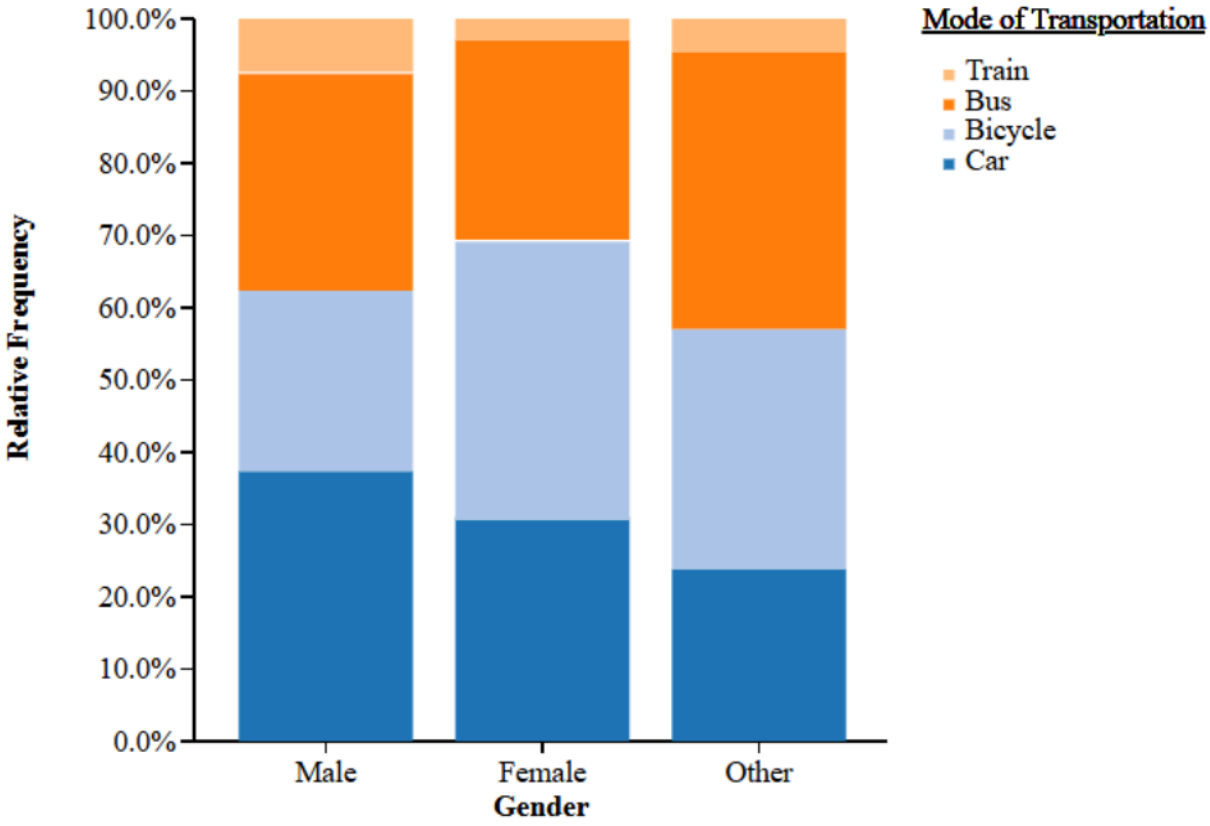
Solution to part d: $12 + 18 + 8 = 38$, which is the total number of people who picked bus. Out of those, 12 were men, thus $\frac{12}{38} = 0.316$, which is $\boxed{31.6\%}$.



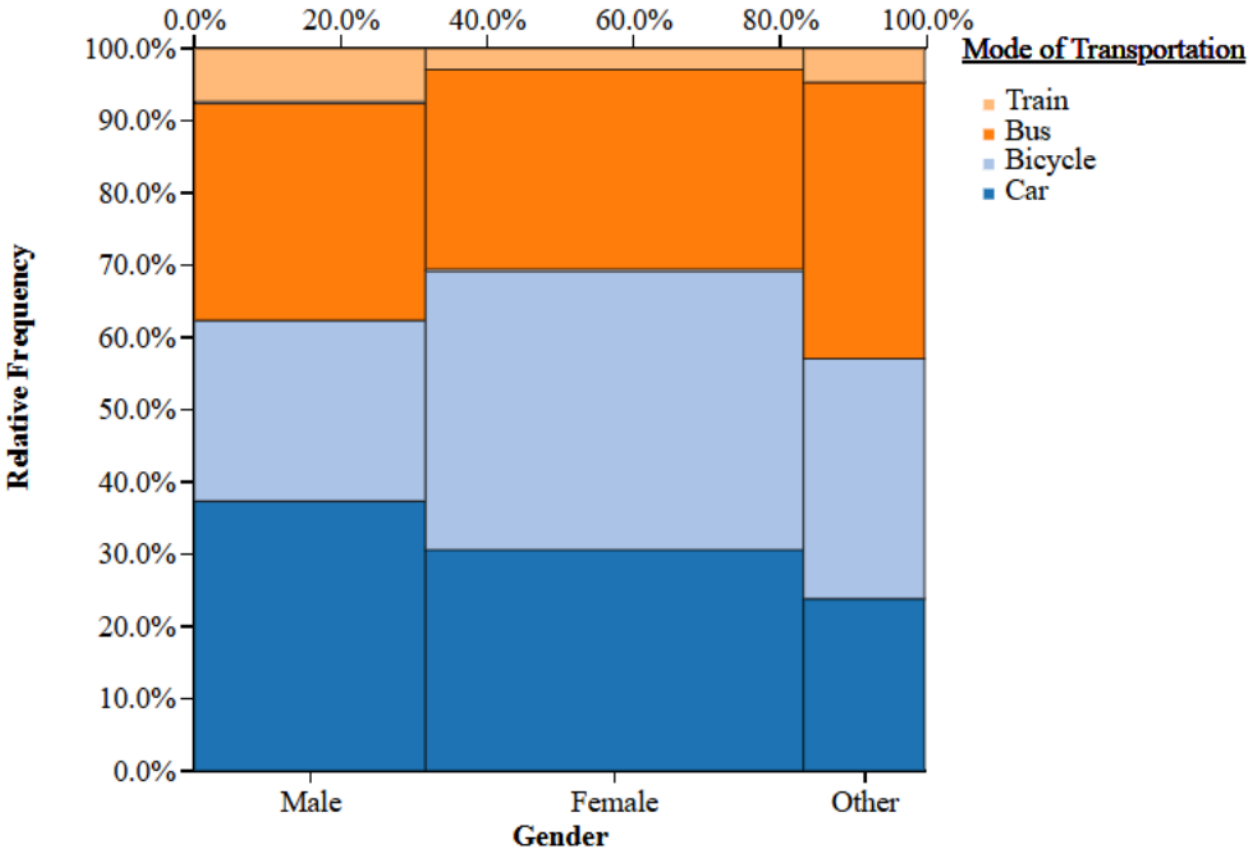
A side-by-side bar graph displays the data in bar graphs next to each other.

Problem 2.1.3 — From the bar graph, which gender had the most **amount** of people who chose car as their mode of transportation?

Solution: Recall that relative frequency graphs do **not** show the amount, only the percentages. Therefore, the answer cannot be determined.



A segmented bar graph shows the breakdown of a categorical variable into another. Here, the mode of transportation is broken down into each part proportionally, and the graphs are shown by gender.



A mosaic plot is a combination of segmented bar graphs, but with the size of each category displayed. To compare the sizes of categories, look at the area.

Problem 2.1.4 — From the mosaic plot, which gender had the most amount of people that picked bicycle as their preferred mode of transportation?

Solution: Since the largest area for bicycles is from females, they had the most amount of people picking bicycle. In a mosaic plot, the area shows the amount.

Note 2.1.5

Association:

Two variables are associated if the outcome of a variable affects the outcome of the other.

Let's look back at the table:

	Male	Female	Other
Car	15	20	5
Bicycle	10	25	7
Bus	12	18	8
Train	3	2	1

Example 2.1.6

Is there evidence of an association between gender and having a car as their mode of transportation?

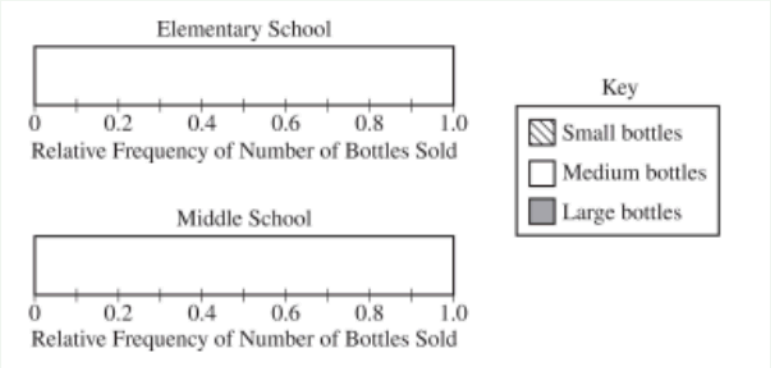
Solution: To find out if there is evidence of an association between gender and having car as their mode of transportation, we find the relative frequencies of each gender for picking car and compare it to the overall proportion picking car. The overall proportion of those surveyed choosing car is $\frac{15+20+5}{126} = 0.317$. For males, it is $\frac{15}{40} = 0.375$. For females, it is $\frac{20}{65} = 0.308$. For those who chose "other", it is $\frac{5}{21} = 0.238$. Since every gender group is different from .317, there is evidence that there is an association between gender and having car as mode of transportation. If every gender group had the same proportion, there would be no association.

Problem 2.1.7 — 2024 AP Statistics

A local elementary school decided to sell bottles printed with the school district's logo as a fund-raiser. The students in the elementary school were asked to sell bottles in three different sizes (small, medium, and large). The relative frequencies of the number of bottles sold for each size by the elementary school were 0.5 for small bottles, 0.3 for medium bottles, and 0.2 for large bottles.

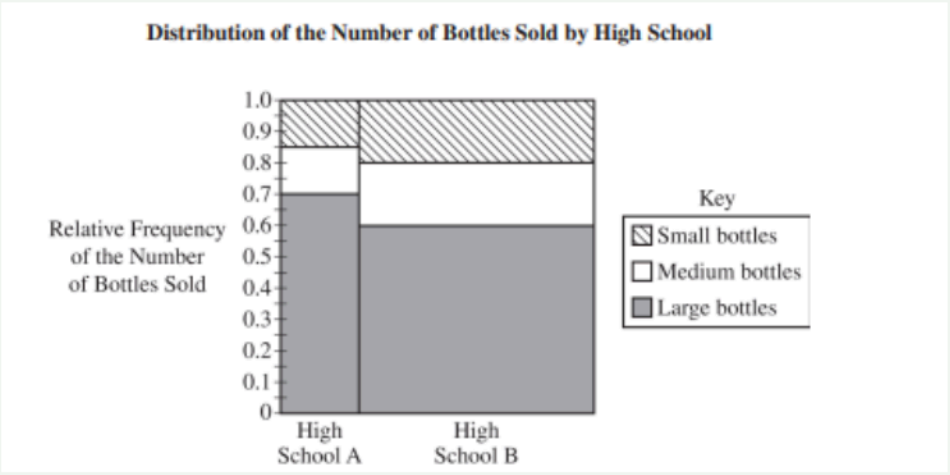
A local middle school also decided to sell bottles as a fund-raiser, using the same three sizes (small, medium, and large). The middle school students sold three times the number of bottles that the elementary school students sold. For the middle school students, the proportion of bottles sold was equal for all three sizes.

a) Complete the segmented bar graphs representing the relative frequencies of the number of bottles sold for each size by students at each school.



b) An administrator at the elementary school concluded that the elementary school students sold more small bottles than the middle school students did. Is the elementary school administrator's conclusion correct? Explain your response.

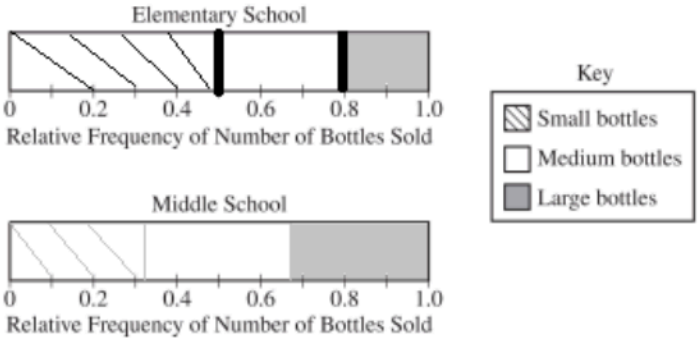
c) A mosaic plot for the distribution of the number of bottles sold by each of the high schools is shown here.



i) Which of the two high schools sold a greater proportion of large bottles? Justify your answer.

ii) Which of the two high schools sold a greater number of large bottles? Justify your answer.

Solution to part a: The relative frequencies for the elementary school are 0.5 for small, 0.3 for medium, and 0.2 for large. Since the middle school has the proportions equal for all three sizes, it is distributed as 0.33 for all of them. Thus:



Solution to part b: The elementary school's administrator is incorrect. Although the elementary school sold a higher proportion of small bottles, the **amount** of small bottles is not larger. Call the amount of total bottles that the elementary school sold x. Since they sold 50% small bottles, their small bottle total is 0.5x. The middle school sold three times as many bottles, so their total bottles is 3x. $\frac{1}{3}$ of these are small bottles, so the middle school has a total of $3x \times \frac{1}{3} = x$ small bottles. Since x is greater than 0.5x, the middle school sold more small bottles.

Solution to part ci: The proportion of large bottles sold by High School A was 0.7, whereas the proportion of large bottles sold by High School B was 0.6. Thus, High School A sold a greater proportion of large bottles.

Solution to part cii: High School B sold a greater number of large bottles because the area shown in the mosaic plot is greater than the area shown for High School A for large bottles.

§2.2 Relationships Between Two Quantitative Variables, Correlation, and Linear Regression Models

Relationships between quantitative variables make up most of this unit. When you have two quantitative variables, or bi-variate data, they are plotted as points on a graph. X is the explanatory variable and Y is the response variable.

Note 2.2.1

Explanatory and Response Variables:

The explanatory variable explains or predicts the response variable. It does not cause changes in the response variable, rather, it explains the changes. X is explanatory and Y is response.

To describe a scatter plot, use the acronym DUFS, and look at the

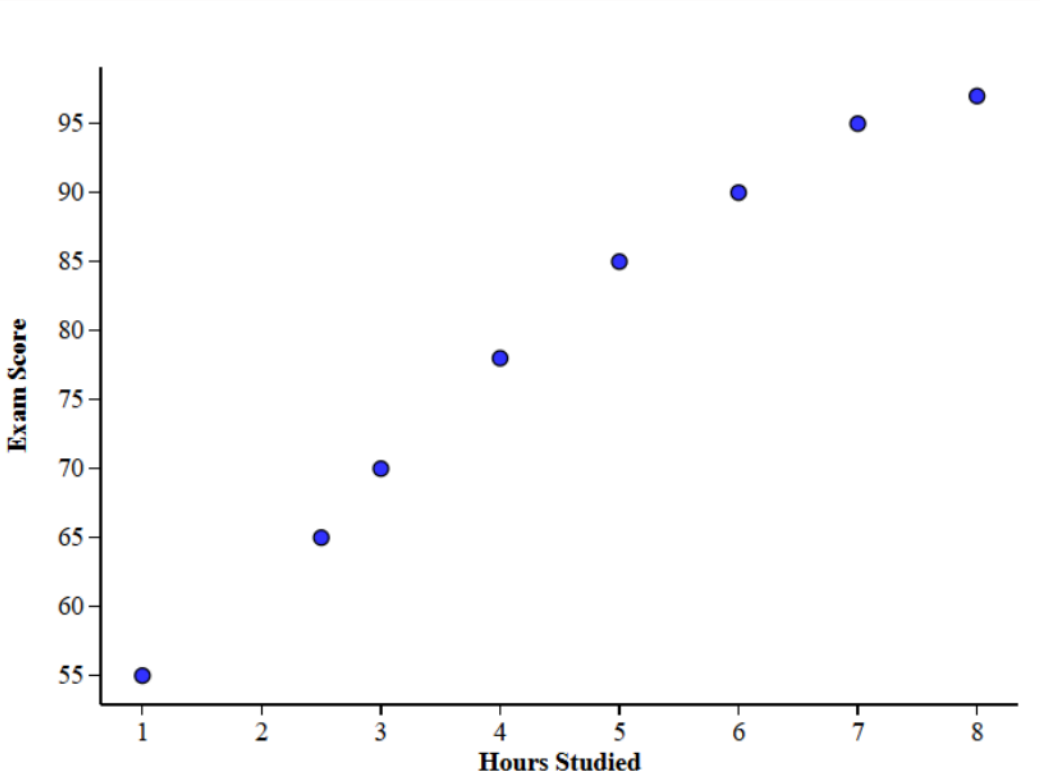
- D.irection: If the relationship between x and y is negative or positive. The relationship is negative if the graph is decreasing and positive if the graph is increasing.
- U.nusual features: If there are outliers in the scatterplot or if there are large clusters.

- F.orm: If the scatterplot is linear or nonlinear.
- S.trength: How close the points follow a pattern. If the scatterplot is very linear, then there is a strong linear relationship. Describe strength as either strong, moderate, or weak.

Remember to add context in your description.

Example 2.2.2

Describe the relationship between hours studied and exam score:



Solution: There is a strong, positive linear relationship between hours studied and exam score. There is a potential outlier at one hour studied. As hours studied increases, so does exam score.

The formula for correlation is $r = \frac{\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)}{n-1}$. On the exam, the r value will be provided to you. However, important conclusions can be drawn from the formula. Switching the x and y values does **not** affect the r value, and neither does changing units.

An r value closer to 0 means that there is a weak linear association, but one close to -1 or 1 indicates there is a strong correlation. A negative correlation means that the scatterplot is decreasing, whereas a positive one means that it is increasing. Correlation does not imply causation, so the variables having high correlation does not mean one causes the other to change.

When describing the strength of a relationship and given the correlation value, use these guidelines for your description:

- Any value between 0.8 and 1 or -1 and -0.8 is a strong correlation.

- Any value from 0.5 to 0.8 or -0.8 to -0.5 is a moderate correlation.
- Any value from -0.5 to 0.5 is a weak correlation.

The equation formed from a linear regression model is $\hat{y} = a + bx$, where $\hat{y}$ is the predicted value of the response value, a is the y-intercept (value when the x is 0), b is the slope of the model, and x is the explanatory variable value. This equation only provides an estimate. To use the equation, plug in the explanatory variable you desire to see the predicted response variable for. Never plug in the response variable and solve for the explanatory, it will not work because the equation is a prediction.

Extrapolation is when an x value is used that is beyond the x values used to construct the regression line. If a scatterplot uses x values from 1 to 9, anything beyond 9 is extrapolation. The further you get, the less accurate the predicted value is.

Problem 2.2.3 — If the regression line for hours studied (x) and exam score (y) is $\hat{y} = 55 + 5x$, find the predicted exam score for 3, 6, and 7 hours studied.

Solution: For 3 hours studied, the predicted exam score is

$$55 + 5(3) = \boxed{70}$$

For 6 hours studied, the predicted exam score is

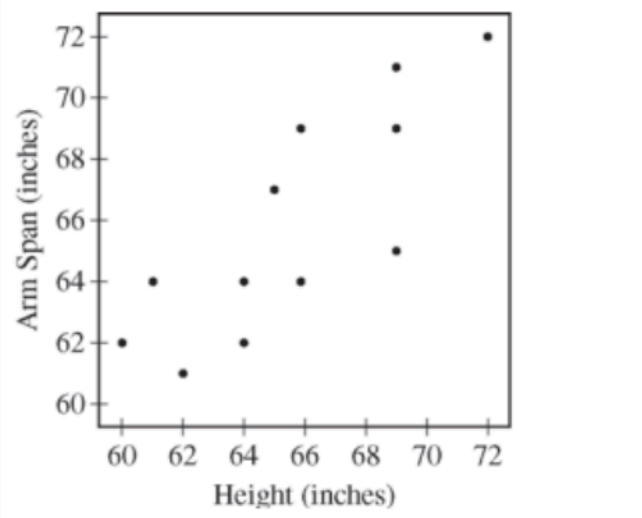
$$55 + 5(6) = \boxed{85}$$

For 7 hours studied, the predicted exam score is

$$55 + 5(7) = \boxed{90}$$

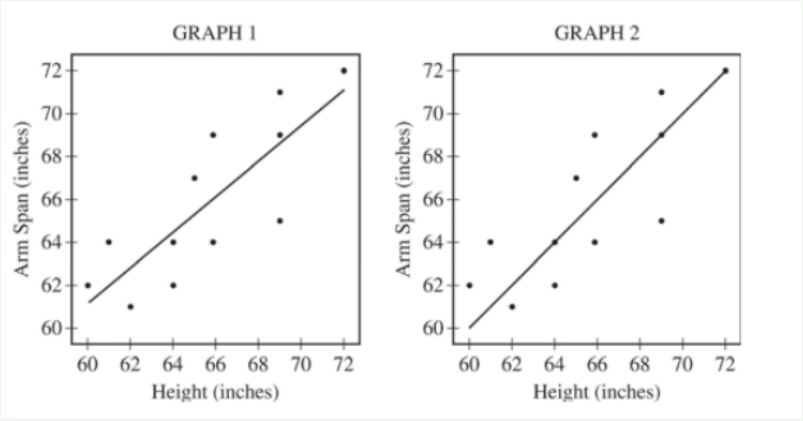
Problem 2.2.4 — 2015 AP Statistics

A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.



a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors.

b) Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.



The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

i) For which graph, 1 or 2, is the line helpful in classifying a student's body shape as square, tall rectangle, or short rectangle? Explain.

ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency			

c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.

Solution to part a: There is a moderate, positive, and linear relationship between height and arm span. Taller students typically have arm spans that are longer.

Solution to part bi: Graph 2 is helpful in classifying a student's body as square, tall rectangle, or short rectangle because it contains students on the line (height equal to arm span), students above the line (arm span greater than height), and students below the line (arm span less than height).

Solution to part bii: There are three students on the line, so three in the box for square. There are 4 students below the line, so there are four in the box for tall rectangle. 5 students are above the line, meaning that there are five in the box for short rectangle.

Solution to part c: The least squares regression line is the best model for prediction. The predicted arm length for a senior with height 61 inches is

$$11.74 + 0.8247(61) = \boxed{62.0467 \text{ inches}}$$

.

§2.3 Residuals, Least Squares Regression, Nonlinear Representations

A residual is the difference between the actual value on the scatterplot and the predicted one given by an equation. It is written as: Residual $= y - \hat{y}$. The sum and mean of residuals is always equal to 0.

Problem 2.3.1 — A linear regression model for hours studied and exam score has an equation of $\hat{y} = 50 + 3.5x$. Calculate the residuals for the coordinates:

- (5,68)
- (6,72)
- (9,82)

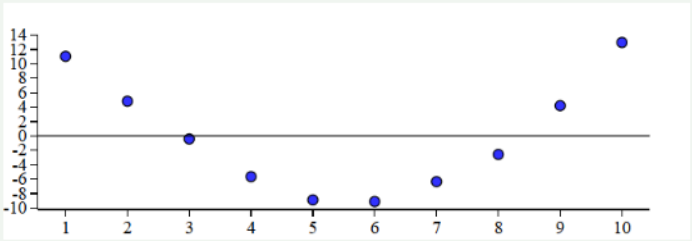
Solution: For (5,68), the predicted value is $50 + 3.5(5) = 67.5$. The actual value is 68, so the residual is $68 - 67.5 = \boxed{0.5}$.

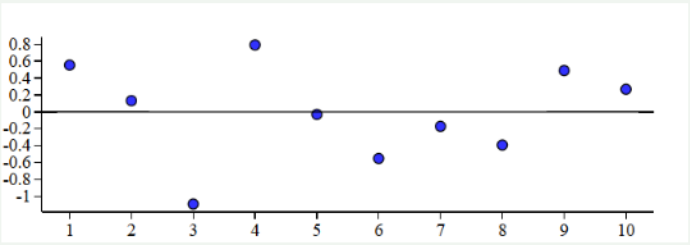
For (6,72), the predicted value is $50 + 3.5(6) = 71$. The actual value is 72, thus, the residual is $72 - 71 = \boxed{1}$.

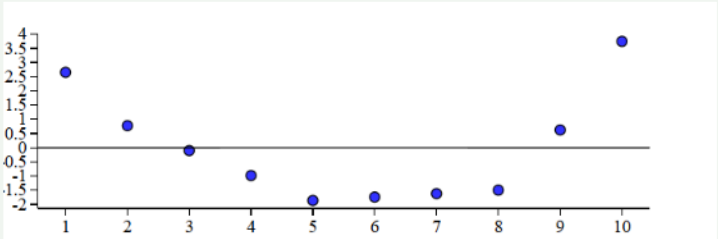
The predicted value for (9,82) is $50 + 3.5(9) = 81.5$. The actual value is 82, meaning that the residual is $82 - 81.5 = \boxed{0.5}$.

If all the residuals are calculated for a scatterplot and then plotted on a graph, randomness in the graph means a linear model is appropriate, however, if there is a clear pattern in the plot, it means that a linear model is not appropriate.

Problem 2.3.2 — Identify which of the graphs are appropriate and which are inappropriate for a linear model:







Solution: The first and third graphs have clear patterns, meaning they are inappropriate for a linear model. The second graph does not have any clear pattern, meaning it is appropriate for a linear model.

The least squares regression line given by $\hat{y} = a + bx$ minimizes the residuals as much as possible. This is why the sum of the residuals ends up equaling 0.
The slope of the regression line is given by $b = r\frac{s_y}{s_x}$, where r is the correlation coefficient, s_y is the standard deviation of the response variable, and s_x is the standard deviation of the explanatory variable.

Problem 2.3.3 — Solve the following questions:

a) If the correlation coefficient is .8, $s_y = 3$ and $s_x = 2.5$, solve for the slope.

b) If the slope $= 4$, correlation is .75, $s_y = 2.5$, solve for s_x.

Solution to a: Substituting the values into the slope equation, we have $b = .8\frac{3}{2.5}$, which equals $\boxed{.96}$.

Solution to b: Substituting the values into the slope equation, we have $4 = .75\frac{2.5}{s_x}$. Dividing by .75 on both sides, we get $5.333 = \frac{2.5}{s_x}$. Multiplying by s_x on both sides gives us $5.333 \times s_x = 2.5$. Dividing by 5.333 on both sides yields us $\frac{2.5}{5.333}$, or $\boxed{0.46875}$.

The slope is how much y is predicted to change for every one unit increase in x.

Problem 2.3.4 — Interpret the slopes for this regression models:

a) The slope for a regression line of hours worked (x) and salary (y) is equal to $500 dollars.

b) The slope for a regression line of hours exercised (x) and pounds lost (y) is equal to 2.5.

Solution to a: For every hour worked, the salary is predicted to increase by $500 dollars.

Solution to b: For every hour exercised, the predicted weight loss is 2.5 pounds.

The y-intercept of the regression line is given by $a = \bar{y} - b\bar{x}$, meaning that the y-intercept is equal to the mean of the y-values subtracted by the slope times the mean of the x-values. It is simply found by rearranging the equation of a regression line. The y-intercept does not always make sense in context, however, it is necessary for the line to work.

To find out how appropriate a linear regression model is, look at the coefficient of determination, or r^2. It provides the proportion of variation in the response variable (y) that is explained by the explanatory variable (x) in a linear regression model. To find the value of the proportion **not** explained, do 1 minus r^2.

Note 2.3.5

To interpret r^2, say that (r^2 as a percentage) of the variation in (y) is explained by (x).

Problem 2.3.6 — Interpret the coefficient of determination for the following:

a) The coefficient of determination for temperature (x) and ice cream sales (y) is .75.

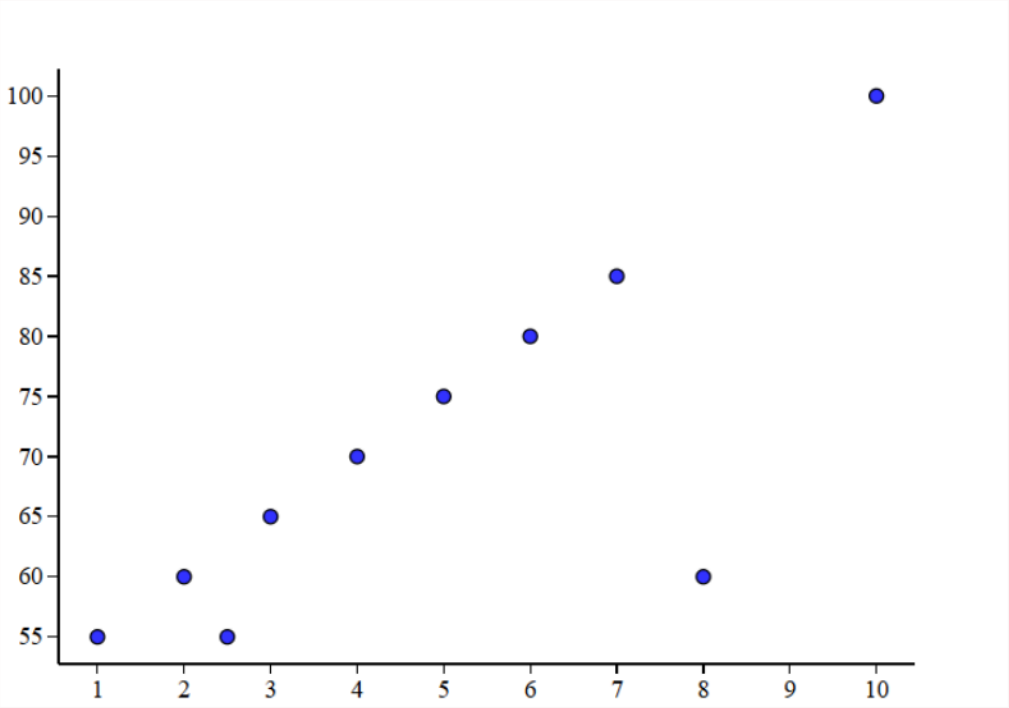
b) The coefficient of determination for square footage (x) and house price (y) is .9.

Solution to a: 75% of the variation in ice cream sales is explained by temperature.

Solution to b: 90% of the variation in house prices is explained by the square footage of the house.

Outliers are values that do not follow the trend of the data and have residuals which indicate large discrepancies. These decrease the correlation. An example of a scatterplot with outliers is listed here:

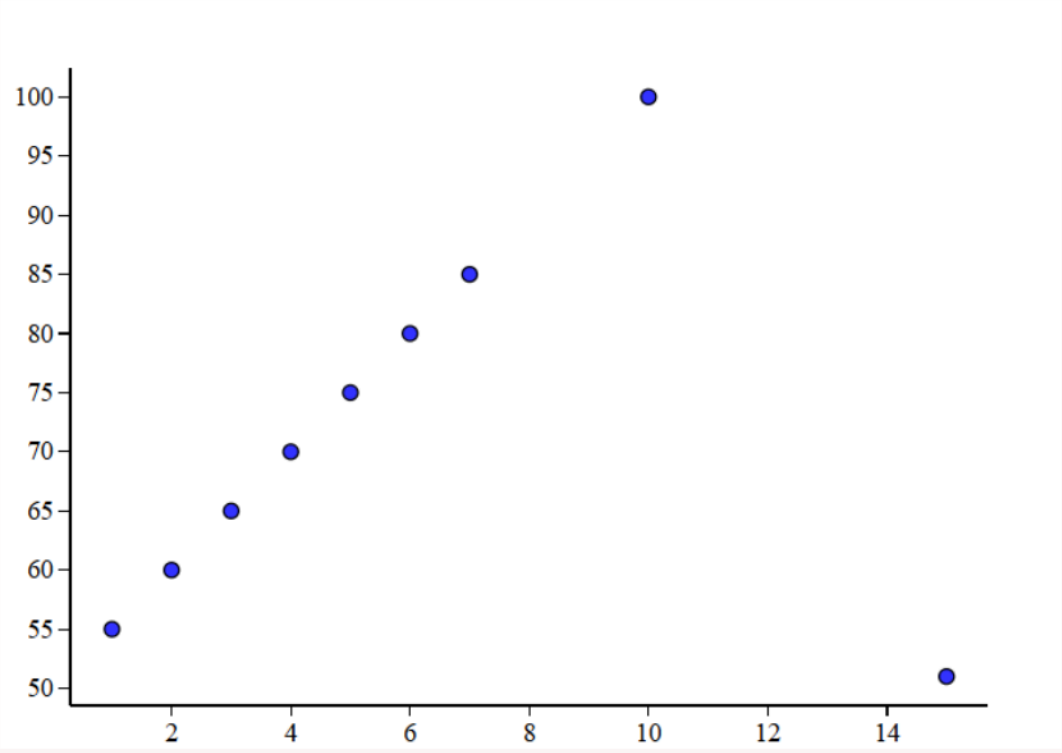
Example 2.3.7



There are obvious outliers at 8 and 2.5, which do not follow the rest of the trend. High-leverage points are points that have much different x-values than the rest of the data.

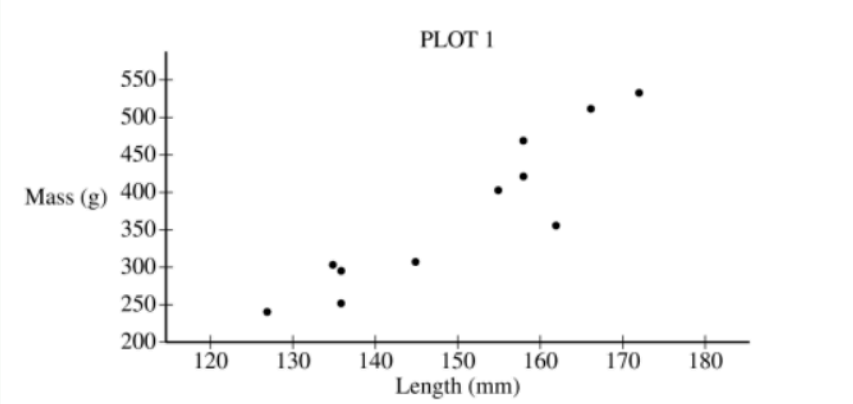
Example 2.3.8

Here is an example of a high leverage point:



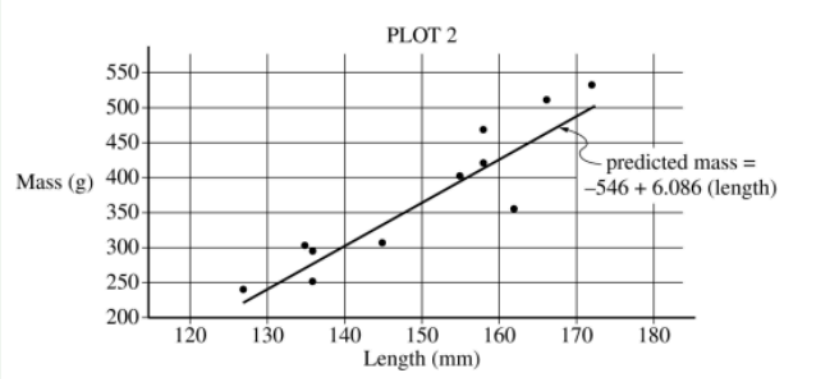
The point at 15 is a high leverage point because it is an outlier in the x direction and significantly changes the slope.
An influential point is one that significantly changes the regression line if removed. Both outliers and high leverage points are often influential points.

Problem 2.3.9 — 2022 AP Statistics
A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.



a) Based on the scatterplot, describe the relationship between mass and length, in context.

b) From the data, the biologist calculated the least-squares regression line for predicting mass from length. The least-squares regression line is shown in Plot 2.



Identify and interpret the slope of the least-squares regression line in context.

c) Interpret the coefficient of determination of the least-squares regression line, $r^2 \approx 0.819$, in context.

d) From Plot 2, consider the residuals of the 11 bullfrogs.

i) Based on the plot, approximately what is the length and mass of the bullfrog with the largest absolute value residual?

ii) Does the least-squares regression line overestimate or underestimate the mass of the bullfrog identified in part (d-i)? Explain your answer.

Solution to part a: There is a strong, positive, and linear relationship between the length and mass of bullfrogs. There do not appear to be any unusual features or outliers

in the graph.

Solution to part b: The slope of the least-squares regression line of the length and mass of bullfrogs is 6.086. As length increases by a millimeter, mass is predicted to increase by 6.086 grams.

Solution to part c: 81.9% of the variation in the mass of bullfrogs is explained by the length of the bullfrogs.

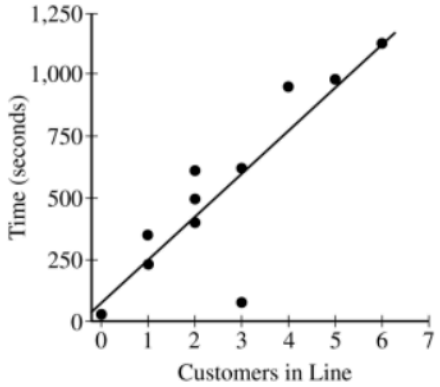
Solution to part di: The largest absolute value residual is at a length of approximately 162 millimeters and a mass of roughly 351 grams. This point is the furthest away from the least-squares regression line, making it the largest residual.

Note: Do not worry if the coordinates for the point you found is slightly off, as long as it has a length from 160 to 165 millimeters and a mass between 350 and 375 grams.

Solution to part dii: The least-squares regression line overestimates the mass of the bullfrog in at the point (162, 351) because the line is above the point.

Problem 2.3.10 — 2018 AP Statistics

The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%	R-Sq (adj) = 70.37%	

a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

b) Identify and interpret in context the coefficient of determination, r^2.

c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

Solution to part a: The estimate of the intercept for the least-squares regression life is

72.95 seconds. This means that the average time to checkout with zero customers in line is 72.95 seconds. Note: To find the slope from computer output, look at the coefficient for the x-axis. In this case, the coefficient for customers in line is 174.40, which is the slope. The y-intercept is the coefficient for the constant, which is 72.95 seconds.

Solution to part b: The coefficient of determination is 73.33%, meaning that 73.33% of the variation in time to checkout is explained by the customers in line.

Solution to part c: The data point at x = 3 and close to y = 0 is considered to be an outlier because the y-value is much lower than the other ones at a close x-value.

3 Unit 3: Collecting Data

Unit 3 is about studies, sampling, and experiments.

§3.1 Planning a Study

Note 3.1.1

Population:

A **population** is the entirety of the group being analyzed. For instance, a population could be all potatoes in Idaho.

Note 3.1.2

Sample:

A **sample** is a part of the group being analyzed. Rather than all potatoes in Idaho, a sample would be 1,000 selected potatoes.

When we collect data, it is only possible to generalize to the entire population if the samples are randomly selected, representative of the original population, and only selected from that population. It isn't correct to generalize data from a different population to another.

Problem 3.1.3 — Identify which of the following are samples and which are examples of populations:

1. All high school students in the United States.
2. 150 customers in a coffee shop surveyed over the course of a week.
3. 50 randomly selected trees measured for their height
4. All students in a university.
5. Everyone with high blood pressure in the United States.

Solution: All high school students in the United States, all students in a university, and everyone with high blood pressure in the United States are examples of populations. 2 and 3 are examples of samples because they are only a part of populations.

Problem 3.1.4 — Explain why each of the following are or aren't generalizable to their population:

1. A survey of 50 employees in a company being generalized to all employees in that city.
2. A random sample of 200 randomly chosen residents in a city to all residents in that city.
3. A random sample of 50 students from the honors program being generalized to all students in a high school.

Solution: The survey of 50 employees from a company cannot be generalized to all employees in that city because the participants were not randomly chosen, and the sample is only from one company.

A random sample of 200 residents in a city is generalizable to all residents in that city because the sample was randomly conducted, and the sample is being generalized to its population.

A random sample of 50 students from an honors program is not generalizable to all students in a high school because the sample is only from the honors program, not the population of all students.

Note 3.1.5

Observational Study:

An observational study is a study that is conducted without imposing any treatments or random assignment on individuals. This entails looking at existing data or conducting a survey. From an observational study, no causal relationship can be concluded, meaning that there can only be a correlation found. Experiments are how cause and effect relationship can be inferred.

Observational studies can either be retrospective or prospective.

Note 3.1.6

Retrospective Study:

Retrospective Studies use existing data and look at the past to draw conclusions. These are usually easier to conduct and inexpensive. However, a cause and effect relationship can't be formed, and retrospective studies are **always** observational.

Note 3.1.7

Prospective Study:

Prospective Studies follow individuals over time and collect data as circumstances change and various outcomes occur. These are also observational, and only show a correlation, rather than a causal relationship. Moreover, prospective studies are more expensive and harder to conduct that retrospective studies, as researchers need to record data and actually conduct a study. However, prospective studies are typically more accurate.

Problem 3.1.8 — Which are retrospective and which are prospective studies?:

1. Looking at 500 adults with lung cancer and examining past smoking habits.
2. Hiring new people at a company and looking at the job retention rate over the next five years.
3. Looking at 300 students over a course of a year to determine how study habits affect their test scores.
4. Asking college alumni about their career success and exaiming their past college major.

Solution: 1 and 4 are examples of retrospective studies because they analyze past behavior. 2 and 3 follow individuals over time, making them prospective studies.

Note 3.1.9

Experiment:

In experiments, researchers impose treatments onto an experimental group in order to show a causal relationship. These entail experimental units, and control groups. If designed properly with random assignment, experiments show cause and effect relationships. Experiments will be explored in more detail later on in this chapter.

Problem 3.1.10 — Which are experiments are which are observational studies?:

1. 100 students are randomly assigned into two different groups with different study techniques. The test scores are compared after the test.
2. Researchers monitor 10 intersections and record the number of cars that run red lights.
3. Survey 200 students about how much their study and compare the grades for the students.
4. Randomly assign 50 plants into two different groups, one with fertilizer, one with not. Plant growth is compared 6 weeks later.

Solution: 1 and 4 are examples of experiments because researchers are imposing treatments. 2 and 3 are observational studies because there are no treatments or random assignment, behavior is simply observed.

§3.2 Sampling

You will need to learn of multiple sampling methods and the advantages and disadvantages of those methods. Sampling is necessary because the majority of the time we cannot analyze an entire population, so we need a sample that is representative of the population with limited bias.

Note 3.2.1

Replacement:

Sampling without replacement is when each member of some population can only be selected one time. If repeats are allowed and each part of the population can be sampled more than once, it is called sampling with replacement.

Note 3.2.2

Simple Random Sample:

In a simple random sample, every possible sample of a desired size has an equal likelihood of being selected. An example of a simple random simple is numbering each individual in a population and using random number generators to select them while ignoring repeats. It is important to understand how to create a step-wise simple random sample for various scenarios.

Although simple random sampling is very effective, it comes with different advantages and disadvantages.

Note 3.2.3

Advantages: easy to implement, unbiased, and data collected is generalizable to the population.
Disadvantages: expensive, under-represents smaller groups, and some people may not respond.

Note 3.2.4

Stratified Random Sample:

In a stratified random sample, the population is divided into groups that share similar characteristics, called strata, and a simple random sample is conducted from each of the groups. For instance, a survey for a school may divide into strata of grade levels. Another type of stratified sampling is proportional sampling, where the size of the samples taken from each strata is proportional to each group. For example, if there are half as many people in group B than group A, then half as many people in group B will be sampled in a proportional sample.

Note 3.2.5

The main advantage of stratified sampling is that each group is properly represented. Some disadvantages of stratified sampling are that it may be hard to implement, and separating into different strata is time-consuming.

Note 3.2.6

Cluster Sample:

In a cluster sample, a population is divided into groups called clusters, and everything from that cluster is picked through random selection. For example, if we were to conduct a study about California, we can make each city a cluster, and we randomly select a few of those cities and sample everyone in those cities.

Note 3.2.7

Some advantages of cluster sampling are that it is cost-effective and allows for a convenient way to get a large sample size.
Some disadvantages of cluster sampling are that it may not be as representative as other sampling methods and there can be higher sampling error.

Note 3.2.8

Systematic Random Sample:

In a systematic random sample, there is a random starting point, and then every nth (where n is any positive nonzero number) person from that list is chosen. An example is picking the 5th person who enters a school, and then surveying every 10th person afterwards. The starting point and nth term vary depending on the sample.

Note 3.2.9

Some advantages of systematic random sampling are that it is easy to implement, good for large populations, and has reduced bias.
A disadvantage of systematic random sampling is that if there is a pattern that occurs in the sampling interval, the data can be very biased (for example, there may be a defect every 10th car manufactured).

Note 3.2.10

Census:

In a census, everything in a population is sampled.

Problem 3.2.11 — Explain what type of sample each statement represents:

1. Every student in a high school is assigned a random number and there is a random generator to select 50 students.
2. Five high schools are randomly selected and all students within those high schools are surveyed.
3. Students in a college are divided by grade level and 25 students from each grade level are selected.
4. All products from a factory are numbered then 100 products are selected using a random number generator.
5. A random starting point is made on a library shelf and every 10th book on the shelf afterwards is selected.
6. Employees in a company are grouped by department and proportionally selected.
7. Three apartment buildings in a city are randomly selected and all residents within those buildings are sampled.
8. All residents of a small town are selected for a survey.
9. A random person in the ticket line for a concert is selected and every 15th person is surveyed.

Solution: 1 and 4 are examples of simple random samples because everyone has an equal chance of being selected. There is no clustering or stratifying, all students or products are simply randomly chosen.

3 and 6 are examples of stratified sampling. Grouping (stratifying) occurs in both of these.

5 and 9 are examples of systematic sampling. A certain point is randomly chosen, then every nth afterward is sampled.

2 and 7 are examples of cluster sampling, all students or residents inside randomly selected

schools or apartment building are chosen.

8 is an example of a census because every resident is sampled.

Samples may come with various biases depending on how they were executed.

Note 3.2.12

Bias:

Bias is when a sample favors some responses over others, making the sample not representative of the population.

Any sample that isn't random comes with inherent bias, and increasing sample size from a biased sample does not reduce bias.

Note 3.2.13

Voluntary Response Bias:

Voluntary response bias occurs from a sample that is conducted from volunteers, not randomly selected people. This type of bias overrepresents people with strong opinions. An example is online product reviews, people who strongly dislike or strongly like a product are more likely to leave a review, and people who were just satisfied with the product are less likely to leave a review. Individuals who do not have a strong opinion on something are underrepresented in a voluntary response sample.

Note 3.2.14

Undercoverage Bias:

Undercoverage bias happens when some parts of the population have a lower chance of being in a sample. This means that the sample is not representative of the entire population. An example of undercoverage bias is online polling, since those who don't go online often or don't have access to the internet will not be represented. These people are left out of the sample, leading to results that may not accurately reflect the entire population.

Note 3.2.15

Nonresponse Bias:

Nonresponse bias happens when individuals selected for a sample do not respond, overrepresenting the groups who do respond. For example, many people do not respond to mail-in surveys, and those who choose to respond likely have an extreme opinion.

Note 3.2.16

Response Bias:

Response bias occurs when participants provide false or misleading responses or if there are problems with the way data is collected. This can occur because people will give answers that are more socially acceptable instead of their actual beliefs, or the question itself leads to a different response. If a participant in a study smokes they may underreport how much they smoke to avoid judgement or seem healthier. A question like "should the government do more to help the poor?" will lead to an answer with response bias because the question implies that the government already isn't doing enough, and people don't want to say that they believe poorer individuals shouldn't be helped.

Note 3.2.17

Convenience Sample:

Convenience sampling bias occurs when individuals aren't randomly selected and instead chosen in a way that is convenient to the researchers. Since there is no randomness, this sample will be biased and unrepresentative of the population. An example is interviewing people passing by in a park.

Problem 3.2.18 — Explain what type of bias each statement represents:

1. Only homeowners are surveyed about local issues.
2. A radio stations asks listeners to share their opinions on a controversial topic.
3. An option for reviewing a product is on a website.
4. Many people are emailed for a survey, but only 20% answer.
5. A phone survey is conducted using landlines.
6. A student is asked by their teacher about how much they study.
7. Survey respondents are asked if they support "essential needs for children".
8. Households are surveyed, but many are at work and unable to answer.
9. A teacher conducts a survey by asking only their students in their 3rd period class.
10. A researcher stops people walking by to gather opinions.

Solution: 2 and 3 are examples of voluntary response bias because they depend on people going out of their way and voluntarily participating.

1 and 5 are examples of undercoverage bias because not everyone is included in the surveys.

4 and 8 are examples of non-response bias because many people do not respond.

6 and 7 are examples of response bias or question wording bias because 6 may lead to a student feeling compelled to give a socially acceptable answer, and 7 has a question that influences respondents to agree.

9 and 10 are examples of convenience sampling bias because the surveys are based on what is near and easy to sample, making the data not representative.

§3.3 Experimental Design

The next section of Unit 3 focuses primarily on experiments. We will first delve into what makes up an experiment.

Note 3.3.1

Experimental Units:

Experimental units are what the experiments are performed on, they are assigned treatments. If these experimental units are people, they are referred to as subjects or participants.

Note 3.3.2

Explanatory Variable:

Explanatory variables are the cause of the response variables, these are manipulated intentionally by the person conducting the experiment. These predict or explain the changes in the response variable.

Note 3.3.3

Response Variable:

The response variable is the outcome of the experiment, it's what is measured. These change as a result of the explanatory variable.

Note 3.3.4

Confounding Variable:

Confounding variables are associated with the explanatory variable and have an effect on the response variable. These lead to a cause-and-effect relationship not being able to be established. If an experiment measures weight loss from taking a medication, the amount of exercise or diet are confounding variables that need to be controlled for a proper experiment.

For conclusions to be drawn from experiments, the experiments have to be well-designed. The components of a well-designed experiment are:

- Multiple treatment groups

- Random assignment of treatments
- Replication (there are multiple experimental units, experiment shouldn't just be done on one person)
- Controlling confounding variables

On the exam, you will need to be able to describe how experiments are conducted and identify them. There are various types of experiments, the first of which is a **completely randomized design**.

Note 3.3.5

Completely Randomized Design:

A completely randomized design has randomly assigned treatments and control for other confounding variables. Each subject has an equal chance of being placed in any group, and random assignments ensures that differences in outcomes are solely because of treatments. Here's an example on how a completely randomized design would be conducted:

Example 3.3.6

A company samples 60 workers. Describe an appropriate method with a completely randomized design that the company can use to randomly assign workers to equal groups where they listen to classical music, pop music, or no music at all.

Solution: To randomly assign workers to equal groups using a completely randomized design,

1. Number each worker from 1 to 60
2. Use a random number generator from 1-60 to assign workers. Assign the first 20 workers randomly selected with unique numbers to classical music, the next 20 to pop music, and the last 20 to no music.
3. Ignore any repeats for numbers already selected.

This is the template to follow when conducting a completely randomized design. First number each participant, then use a random generator to select and put in groups while ignoring repeats. Always include context.

Note 3.3.7

Single-blind Experiment:

In a single-blind experiment, either the researchers do not know what treatments the subjects receive, or the subjects do not know what treatments they receive. This type of experiment limits bias and the placebo effect.

Note 3.3.8

Double-blind Experiment:

In a double-blind experiment, both the researchers and the subjects do not know which treatment subjects are receiving. Similar to a single-blind experiment, this type of experiment limits bias and the placebo effect.

Note 3.3.9

Control Groups & Placebo Effect:

A control group is not given the treatment being tested, rather, it is given nothing, or a placebo, which allows the researchers to test if the treatments actually have an effect. The placebo effect is when the subjects show some response to the placebo they're given. Having a control group allows researchers to test and control for a placebo effect.

Note 3.3.10

Randomized Block Design:

In a randomized block design, treatments are randomly assigned to different blocks. These blocks are different groups of subjects based on similar characteristics. First, a variable to blocked is identified, then blocks are created, then there is random assignment within each block. This allows for comparison between the different blocks and less variability. Here is an example of a randomized block design:

Example 3.3.11

In a garden, half the plants have full sun exposure, and half have partial sun exposure. There are two different fertilizers being used, fertilizer A and fertilizer B. Explain how a randomized block design can be used to compare the different plants and fertilizers.

Solution: To use a randomized block design to compare the different plants and fertilizers:

1. Create two different blocks, block 1 being the plants with full sun exposure, and block 2 being the plants with partial sun exposure.
2. In each block, randomly assign the two different fertilizers where half of each block has fertilizer A and the other half has fertilizer B.
3. Measure the outcomes of plant growth and compare the results.

The template to use a randomized block design is to create different blocks, conduct random assignment in those blocks for treatments, then compare results.

Note 3.3.12

Matched Pairs Design:

A matched pairs design is a more specific type of block design. Two similar subjects are matched into pairs based on similar characteristics, and different treatments are given. One of the subjects acts as a type of control group. By using a matched pairs design, variability is reduced and precision is improved. Factors such as initial starting point become less important. Here is an example of a matched pairs design:

Example 3.3.13

20 pairs of identical twins are asked to partake in different reading programs, program A and program B. Describe how a matched-pairs design can be conducted to find the effect of the reading program.

Solution: To use a matched-pairs design to find the effect of the reading program,

1. For every pair of twins, label a twin as 1 and the other as 2
2. Use a random number generator to generate a number from 1-2 and give the twin selected Program A and the other Program B
3. Repeat for all 20 pairs of twins.
4. Compare the results

To conduct a matched-pairs design, first group participants into matched-pairs, then label each participant 1 or 2. Then use random number generation for assignment.

By utilizing **randomization**, confounding variables are minimized and allow researchers to conclude that any changes they observe are statistically significant, meaning they did not occur by chance. Results being significant show a causal (cause and effect) relationship. These results can be generalized to a population if the subjects from that population were picked through random sampling.

Note 3.3.14

If volunteers are used for an experiment, causal relationships can still be shown, however, the results will only be generalizable to groups similar to the volunteers, not the whole population like random sampling does.

If there are statistically significant results, it means that there is clear evidence that the treatments caused an effect.

§3.4 Practice Problems

Problem 3.4.1 — 2014 AP Statistics
As part of its twenty-fifth reunion celebration, the class of 1988 (students who graduated in 1988) at a state university held a reception on campus. In an informal survey, the director of alumni development asked 50 of the attendees about their incomes. The director computed the mean income of the 50 attendees to be \$189,952. In a news release, the director announced, "The members of our class of 1988 enjoyed resounding success. Last year's mean income of its members was \$189,952!"

a) What would be a statistical advantage of using the median of the reported incomes, rather than the mean, as the estimate of the typical income?

b) The director felt the members who attended the reception may be different from the class as a whole. A more detailed survey of the class was planned to find a better estimate of the income as well as other facts about the alumni. The staff developed two methods based on the available funds to carry out the survey.
Method 1: Send out an e-mail to all 6,826 members of the class asking them to complete an online form. The staff estimates that at least 600 members will respond.
Method 2: Select a simple random sample of members of the class and contact the selected members directly by phone. Follow up to ensure that all responses are obtained. Because method 2 will require more time than method 1, the staff estimates that only 100 members of the class could be contacted using method 2.
Which of the two methods would you select for estimating the average yearly income of all 6,826 members of the class of 1988? Explain your reasoning by comparing the two methods and the effect of each method on the estimate.

Solution to part a: A statistical advantage of using the median of reported incomes rather than the mean as the estimate of the typical income is that the median is less affected by outliers. High-income individuals could dramatically skew the data if using the mean, but medians are resistant to outliers. For instance, a billionaire dramatically skews the data if using mean as the estimate of the typical income.

Solution to part b: Method 2 is the better method for estimating the yearly income of all 6,826 members of the class of 1988. Method 1 has voluntary response bias, which gives data that is not representative of the students. Method 2 uses unbiased sampling and ensures that there is a 100% response rate. Although Method 2 samples fewer total people, the people sampled are much more representative of the total population.

Problem 3.4.2 — 2016 AP Statistics

Alzheimer's disease results in a loss of cognitive ability beyond what is expected with typical aging. A local newspaper published an article with the following headline.

> Study Finds Strong Association Between Smoking and Alzheimer's

The article reported that a study tracked the medical histories of 21,123 men and women for 23 years. The article stated that, for those who smoked at least two packs of cigarettes a day, the risk of developing Alzheimer's disease was 2.57 times the risk for those who did not smoke.

a) Identify the explanatory and response variables in the study.

b) Is the study described in the article an observational study or an experiment? Explain.

c) Exercise status (regular weekly exercise versus no regular weekly exercise) was mentioned in the article as a possible confounding variable. Explain how exercise status could be a confounding variable in the study.

Solution to part a: The explanatory variable in this study is if participants smoke or do not smoke, and the response variable is Alzheimer's status, if they do or do not develop Alzheimer's disease.

Solution to part b: This is an observational study because there was no assignment of treatments, the frequency that people smoked was simply observed.

Solution to part c: Exercise status could be a confounding variable in the study because people who exercise often may be more health-conscious and have better overall health. Excercising can also decrease the risk of Alzheimer's disease. This would interfere with the results in the study because only the explanatory variable is smoking frequency and exercise is not accounted for.

Problem 3.4.3 — 2021 AP Statistics

Researchers will conduct a year-long investigation of walking and cholesterol levels in adults. They will select a random sample of 100 adults from the target population to participate as subjects in the study.

a) One aspect of the study is to record the number of miles each subject walks per day. The researchers are deciding whether to have subjects wear an activity tracker to record the data or to have subjects keep a daily journal of the miles they walk each day. Describe what bias could be introduced by keeping the daily journal instead of wearing the activity tracker.

b) During the course of the study, the subjects will have their cholesterol levels measured each month by a doctor. The researchers will perform a significance test at the end of the study to determine whether the average cholesterol level for subjects who walk fewer miles each day is greater than for those who walk more miles each day.
Selecting a random sample creates a reasonable representative sample of the target population. Explain the benefit of using a representative sample from the population.

c) Suppose the researchers conduct the test and find a statistically significant result. Would it be valid to claim that increased walking causes a decrease in average cholesterol levels for adults in the target population? Explain your reasoning.

Solution to part a: By keeping daily journals, response bias could be introduced because people may not remember how much they walk. As a result, their responses are going to be inaccurate. Responses may either underreport or overreport the daily miles walked, leading to inaccurate, biased responses that are not representative of total miles walked by adults. Activity trackers are more accurate then daily journals.

Solution to part b: Using a representative sample from the population allows results to be generalized to that population. Moreover, a representative sample allows for accurate, unbiased conclusions to be drawn about miles walked and average cholesterol levels.

Solution to part c: It would not be valid to claim that increased walking causes a decrease in average cholesterol levels for adults in the target population because this is only an observational study and there is no random assignment of treatment. Since this is not an experiment, no cause and effect relationship can be formed. There is also the possibility of confounding variables that are not taken into account.

Problem 3.4.4 — 2023 AP Statistics

A developer wants to know whether adding fibers to concrete used in paving driveways will reduce the severity of cracking, because any driveway with severe cracks will have to be repaired by the developer. The developer conducts a completely randomized experiment with 60 new homes that need driveways. Thirty of the driveways will be randomly assigned to receive concrete that contains fibers, and the other 30 driveways will receive concrete that does not contain fibers. After one year, the developer will record the severity of cracks in each driveway on a scale of 0 to 10, with 0 representing not cracked at all and 10 representing severely cracked.

a) Based on the information provided about the developer's experiment, identify experimental units, treatments, and the response variable.

b) Describe an appropriate method the developer could use to randomly assign concrete that contains fibers and concrete that does not contain fibers to the 60 driveways.

c) Suppose the developer finds that there is a statistically significant reduction in the mean severity of cracks in driveways using the concrete that contains fibers compared to the driveways using concrete that does not contain fibers. In terms of the developer's conclusion, what is the benefit of randomly assigning the driveways to either the concrete that contains fibers or the concrete that does not contain fibers?

Solution to part a: The experimental units are the 60 driveways, the treatments are the concrete with fibers or no fibers, and the response variable is the severity of cracks on a scale of 0-10.

Solution to part b: Number each driveway from 1-60 and use a random number generator to select 30 unique numbers from 1-60, ignoring repeats. Assign the 30 selected to the concrete with fibers and the 30 not selected to concrete with no fibers.

Solution to part c: The benefit of randomly assigning the driveways to either the concrete that contains fibers or the concrete that does not contain fibers is that random assignment allows for a causal relationship to be formed that the type of concrete fibers affects severity of cracks.

4 Unit 4: (Probability, Random Variables, and Probability Distributions

§4.1 Probability

Probability is the measure of the chance, or likelihood, of an event occurring. It is from 0 to 1 inclusive, with 0 meaning that the event will never occur, and 1 meaning that event will always occur.

Note 4.1.1

Outcomes & Events:

An outcome is the result of a trial, whereas an event is one or more outcomes. For instance, if you are drawing a single card from a deck, an outcome is drawing an Ace of Spades, whereas an event is drawing an Ace in general. The event is the collection of outcomes. An event can be an outcome if there is only one possible result.

Simulations model random events. You will likely not be asked to conduct simulations on the exam, but you will need to understand them.

Note 4.1.2

Relative Frequency & Law of Large Numbers:

The relative frequency is the amount of times an event occurred divided by the amount of trials. Similar to probability, it is 0-1. Relative frequencies can be used to predict or estimate the probability of an event or outcome.

The more simulations occur, the closer the relative frequency of an event is to the true probability. If a basketball player has a 50 percent true free throw percentage and a simulation is conducted where the player shoots 1000 free throws, the relative frequency of making a shot will be closer to .50. However, if the player only shoots four free throws, the relative frequency of making a shot can either be 0, .25, .50, .75, or 1. The fewer trials conducted, the more fluctuations occur, the relative frequency will be highly volatile. This is called the law of large numbers, where the more simulations conducted, the closer the percentage gets to the true probability.

Problem 4.1.3 — A bag contains 100 marbles, 40 red, and 60 blue. The probability of drawing a red marble is 0.4. If you want to draw 70% red marbles, would you rather draw 10 or 50 marbles in total?

Solution: Drawing 10 marbles is preferable because if you draw 50 marbles, it is likely that there will be close to 40% red marbles due to the law of large numbers. However, if you draw less marbles, there is more variability, increasing the likelihood of 70% red marbles being drawn, thus, drawing 10 marbles is preferable in this instance.

Note 4.1.4

Random Processes:

Random processes create results that are determined by probability or chance. These can be things like dice rolls or coin flips.

Note 4.1.5

Sample Space:

The sample space is all of the possible outcomes in a random process. The sample space of a coin toss is {Heads, Tails} and the sample space of a die roll is {1, 2, 3, 4, 5, 6}.

Note 4.1.6

Calculating Probability:

Probability of an Event $= \frac{\text{number of outcomes in event}}{\text{number of outcomes in sample space}}$. For example, the outcome could be flipping a coin and getting tails. This would mean that there is one outcome in the event. There are then two outcomes in the sample space, heads or tails, so the probability is $\frac{1}{2} = .5$.

Since the probability is a fraction, and the number of outcomes in the sample space can't exceed the number of outcomes in the event, probability is between 0 and 1.

The probability of something not happening is 1 minus the probability of it happening. The formal way of writing this is $P(E^{\complement}) = 1 - P(E)$.

Problem 4.1.7 — Solve each question:

a) If a jar of cookies has 200 chocolate chip cookies and 100 Oreos, what is the probability of selecting an Oreo cookie?

b) If you roll a six-sided die, what is the probability of not rolling a 2?

Solution to part a: We are looking for P(Oreo). Since there are 200 chocolate chip and 100 Oreo cookies, there is a total of $200 + 100 = 300$ outcomes in the sample space. Thus, the probability of selecting an Oreo cookie is $\frac{100}{300} = \boxed{.333}$.

Solution to part b: We are looking for $P(2^{\complement})$. Since there are six sides, there are six outcomes in the sample space. The probability of rolling a 2 is $\frac{1}{6}$. Thus, the probability of **not** rolling a 2 is $1 - \frac{1}{6} = \frac{5}{6} = \boxed{.833}$.

The probability that two events occur simultaneously is called the joint probability. It is written in the notation of: $P(A \cap B)$ or $P(A \text{ and } B)$. If this joint probability is zero, then the two events are mutually exclusive, or disjoint, and $P(A \text{ and } B) = 0$. If events are mutually exclusive, they can not occur simultaneously.

Problem 4.1.8 — Are the events of flipping a head and flipping a tail mutually exclusive?

Solution: The events are mutually exclusive because P(Heads and Tails) = 0, both cannot occur at the same time.

Conditional probability is the probability of event A given that event B has happened. It is written in the notion: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ which breaks down to: $\frac{P(A \text{ and } B)}{P(B)}$. This is the probability of A and B occurring over the probability of B. Questions are sometimes asked using "if" rather than "given".

Example 4.1.9

Use these questions as examples to differentiate between "if" and "given".

a) A class has 50 students, 30 male and 20 female. 15 females received an A, and 10 males received an A. If a student received an A, what is the probability that the student is male?

b) A school has 200 students, with 120 taking science, and 80 taking math. 40 are taking both science and math. What is the probability that a student is taking science given that they are taking math?

Solution to part a: We are looking for $P(M \mid A)$, which breaks down to $\frac{P(M \cap A)}{P(A)}$. The probability that a student is a male and received an A is $\frac{10}{50} = .2$, and the probability that a student received an A is $\frac{15+10}{50} = .5$. Thus,

$$P(M \mid A) = \frac{P(M \cap A)}{P(A)} = \frac{.2}{.5} = \boxed{0.4}$$

Solution to part b: We are looking for $P(S \mid M)$, which breaks down to $\frac{P(S \cap M)}{P(M)}$. The probability of a student taking science and math is $\frac{40}{200} = .2$. The probability of a student taking math is $\frac{80}{200} = 0.4$.

$$P(S \mid M) = \frac{P(S \cap M)}{P(M)} = \frac{.2}{.4} = \boxed{0.5}$$

Problem 4.1.10 — Solve each problem:

a) In a bag with 10 marbles, 4 red, 3 blue, and 3 green, what is the probability of drawing a blue marble given that the marble is not red?

b) There are 35 students in a grade, 20 students are taking math, 15 are taking physics, and 5 are taking both. If a student is taking math, what is the probability that they are taking physics?

Solution to part a: We are looking for $P(B \mid R^{\complement})$, which breaks down to $\frac{P(B \cap R^{\complement})}{P(R^{\complement})}$. $P(B \cap R^{\complement})$ is simply the probability of drawing a blue marble (since the marble not being

red is already implied if we are drawing a blue one), which is $\frac{3}{10} = .3$. The probability of the marbles not being red is $\frac{10-4}{10} = .6$. Thus, we have

$$P(B \mid R^{\complement}) = \frac{P(B \cap R^{\complement})}{P(R^{\complement})} = \frac{.3}{.6} = \boxed{0.5}$$

Solution to part b: We are looking for $P(P \mid M)$, which breaks down to $\frac{(P \cap M)}{P(M)}$. The probability of a student taking physics and math is $\frac{5}{35} = .143$. The probability of a student taking math is $\frac{20}{35} = .571$. Thus, we have

$$P(P \mid M) = \frac{P(P \cap M)}{P(M)} = \frac{.143}{.571} = \boxed{0.25}$$

The multiplication rule is as follows: $P(A \cap B) = P(A) \cdot P(B \mid A)$ meaning that the probability of both events A and B occurring is equal to the probability of event A times the probability of event B given A.

Problem 4.1.11 — In a bag of 10 marbles, 4 red, 3 blue, 3 green, two marbles are drawn without replacement. What is the probability that the first marble drawn is red, and the second is blue? (the probability that red and blue marbles are selected)

Solution: The probability that the first marble is red is $\frac{4}{10} = .4$. The probability of drawing a blue marble after this is $P(B \mid R)$. After the first marble is drawn, there are 9 marbles left. Thus, the probability of drawing a blue marble given a red one was drawn first is $\frac{3}{9} = .333$. We have

$$P(R \cap B) = P(R) \cdot P(B \mid R) = 0.4 \times 0.333 = \boxed{0.133}$$

Independence is another important concept in this unit.

Note 4.1.12

Independence:

Two events are independent if event A does not affect the probability of event B. If the two events are independent, then

1. $P(A \cap B) = P(A) \cdot P(B)$
2. $P(A \mid B) = P(A)$
3. $P(B \mid A) = P(B)$

Make sure to check all of these if asked for independence. Moreover, if you know that two events are independent, you can use $P(A \cap B) = P(A) \cdot P(B)$ to get the probability of both (joint probability) occurring.

Problem 4.1.13 — In a school with 200 students, there are 120 girls and 80 boys. 50 students in total are in chess club, with 30 of them being girls. The rest of the 20 chess club members are boys. Are the events "The student is a girl" and "The student is in the chess club" independent?

Solution: If the two events are independent,

$$\text{P(Girl} \cap \text{Chess Club)} = \text{P(Girl)} \cdot \text{P(Chess Club)}$$

The probability of a student being a girl is $\frac{120}{200} = 0.6$, and the probability of a student being in the chess club is $\frac{50}{200} = 0.25$

Since there are 30 girls in the chess club, the probability of being a girl and being in the chess club, or P(Girl $\cap$ Chess Club), is $\frac{30}{200} = 0.15$.

Now we verify if P(Girl $\cap$ Chess Club) = P(Girl) $\cdot$ P(Chess Club)

$$0.6 \cdot 0.25 = 0.15$$

Thus, the events are independent.

Make sure to check the other conditions just in case a calculation was performed incorrectly. Thus, we need to find out if these two conditions apply:

$$\text{P(Chess Club} \mid \text{Girl)} = \text{P(Chess Club)}$$

$$\text{P(Girl} \mid \text{Chess Club)} = \text{P(Girl)}$$

We already have that P(Girl $\cap$ Chess Club) = .15, P(Girl) = .6, and P(Chess Club) = .25. Thus, we can easily check the conditions for independence. $\frac{.15}{.6} = .25$ and $\frac{.15}{.25} = .6$, all conditions are satisfied.

The probability that event A or event B will occur is written as $P(A \cup B)$. If you know the probability of A and B individually, and the probability of both A and B occurring, then you can use: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Try to understand this conceptually, the probability that event A or B occurs is the probability that each event happens individually minus the probability that both occur.

Use this next problem as an example for larger-scale probability calculations:

Example 4.1.14

A math class has a final. 50% of students study 1-2 hours, 35% study for more than 2 hours, and 15% study for 0-1 hours. Of the students that study for 1-2 hours, 60% receive an A, for students that study for more than 2 hours, 85% receive an A, and for those that study 0-1 hours, 10% receive an A. What is the probability that a student studied for 0-1 hours given that they received an A?

Solution: We are looking for P(0-1 hours | A). To find the value, we need to find P(0-1 hours $\cap$ A) and P(A).

For students that study 0-1 hours, 10% receive an A. Since they are 15% of the students, $0.15 \cdot 0.10 = 0.015$ represents the total proportion of students who study 0-1 hours and receive an A, in other words, P(0-1 hours $\cap$ A) = 0.015.

Now we check it for the other hour totals. P(1-2 hours $\cap$ A) $= 0.5 \cdot 0.6 = 0.3$. P(More than 2 hours $\cap$ A) $= 0.35 \cdot 0.85 = 0.2975$.

P(0-1 hours $\cap$ A) + P(1-2 hours $\cap$ A) + P(More than 2 hours $\cap$ A) = P(A), since these represent all the hour totals.

Thus, $0.015 + 0.3 + 0.2975 = 0.6125 = \text{P(A)}$.

Since we already know P(0-1 hours $\cap$ A) = .015, we have P(0-1 hours | A) $= \frac{P(\text{0-1 hours} \cap A)}{P(A)}$ $= \frac{.015}{.6125} = \boxed{0.0245}$.

Problem 4.1.15 — A test for a disease has a 95% accuracy rate for successfully detecting the disease when it is present, and a 90% accuracy rate for identifying that a disease is not present. 2% of individuals have the disease in a population. If a person tests positive, what is the probability that they actually have the disease?

Solution: We are looking for P(Disease | Test Positive). We need to find P(Disease $\cap$ Test Positive) and P(Test Positive).

We know that the P(Disease $\cap$ Test Positive), the proportion of people having the disease

and testing positive, is $.02 \cdot .95 = .019$.

The next probability we need to find is the probability that a person tests positive even though they do not have the disease. Since the test is 90% accurate at detecting that the disease is not present, that means it has a false positive rate of 10%.
Thus, P(No Disease | Test Positive) $= .98 \cdot .1 = .098$.

P(Test Positive) = P(No Disease | Test Positive) + P(Disease | Test Positive) = .019+.098 = .117.

P(Disease | Test Positive) $= \frac{\text{P(Disease} \cap \text{Test Positive)}}{\text{P(Test Positive)}} = \frac{.019}{.117} = \boxed{0.162}$.

Problem 4.1.16 — 2014 AP Statistics
Nine sales representatives, 6 men and 3 women, at a small company wanted to attend a national convention. There were only enough travel funds to send 3 people. The manager selected 3 people to attend and stated that the people were selected at random. The 3 people selected were women. There were concerns that no men were selected to attend the convention.

a) Calculate the probability that randomly selecting 3 people from a group of 6 men and 3 women will result in selecting 3 women.

b) Based on your answer to part (a), is there reason to doubt the manager's claim that the 3 people were selected at random? Explain.

c) An alternative to calculating the exact probability is to conduct a simulation to estimate the probability. A proposed simulation process is described below.

> Each trial in the simulation consists of rolling three fair, six-sided dice, one die for each of the convention attendees. For each die, rolling a 1, 2, 3, or 4 represents selecting a man; rolling a 5 or 6 represents selecting a woman. After 1,000 trials, the number of times the dice indicate selecting 3 women is recorded.

Does the proposed process correctly simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women? Explain why or why not.

Solution to part a: To find the probability, we need to use the multiplication rule. The probability that three women are selected is P(All three are women), which equals

P(First is a woman) × P(Second is a woman | First is a woman) × P(Third is a woman | Second is a woman).

The probability that the first person selected is a woman is $\frac{3}{9}$, the probability that the second is a woman given that the first is a woman is $\frac{2}{8}$, and the probability that the third is a woman given that the second is a woman is $\frac{1}{7}$. Substituting values in, we have

$$\frac{3}{9} \times \frac{2}{8} \times \frac{1}{7} = \frac{6}{504} = \frac{1}{84}$$

$\frac{1}{84} = \boxed{0.012}$.

Solution to part b: There is reason to doubt the manager's claim that the 3 people were selected at random because there is only a 1.2% of selecting 3 women at random. Since the probability is so small, we have reason to doubt the manager.

Solution to part c: The proposed process does **NOT** correct simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women because

the probability of selecting a man or a woman always stays the same with the dice rolls. In the original random selection, the probability of selecting a man or a woman changes, as seen with $\frac{3}{9}$,$\frac{2}{8}$, and$\frac{1}{7}$. With the dice rolls, the probability remains at $\frac{1}{3}$ the whole time, meaning that it is independent, whereas the original sample is **dependent**.

§4.2 Random Variables

Note 4.2.1

Random Variables:

A random variable has numerical values that explain an outcome. For example, a random variable could represent the number of games that a team would win out of four games. The probability could be 0.6 that the team wins no games, 0.2, that the team wins one game, 0.1 that they win two games, 0.075 that they win three games, and 0.025 that they win four games. The probability **always** sums to 1.

Note 4.2.2

Discrete Random Variables:

Discrete random variables are a type of random variable that take on a certain amount of values. Continuous random variables are able to take on all values from a certain range, for instance, all values from 0-999999, but discrete random variables can only take on certain values in a set. The probabilities together need to sum up to 1 in a discrete random variable. The example shown in note 4.2.1 is an example of a discrete random variable.

Note 4.2.3

Expected Value (Mean):

To calculate the expected value, or mean of a discrete random variable, the formula is $E(X) = \mu = \sum_x x \cdot P(x)$, meaning that the sum of all values x that the random variable can take multiplied by the probability of those values occurring is the expected value. Here's an example:

Example 4.2.4

A table representing the number of customers entering a store in an hour is shown here, calculate the expected value of customers entering:

Customers	Probability
0	0.05
1	0.1
2	0.2
3	0.25
4	0.2
5	0.1
6	0.1

Solution: Using $E(X) = \mu = \sum_x x \cdot P(x)$, we have

$$0(0.05) + 1(0.1) + 2(0.2) + 3(0.25) + 4(0.2) + 5(0.1) + 6(0.1) = \boxed{3.15\text{customers}}$$

On average, the store is expected to have 3.15 customers entering in an hour.

Note 4.2.5

Standard Deviation and Variance:

The formula to calculate standard deviation of a discrete random variable is $\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\sum_x (x-\mu)^2 P(x)}$, meaning that the standard deviation is equal to the square root of the sum of the values x of the random variables minus the mean (expected value) squared, multiplied by the probability of that value x occurring. The variance is simply the standard deviation squared, but has uses when combining data sets. Here's an example on how to calculate standard deviation:

Example 4.2.6

A table representing the number of customers entering a store in an hour is shown here, calculate the standard deviation of customers entering:

Customers	Probability
0	0.05
1	0.1
2	0.2
3	0.25
4	0.2
5	0.1
6	0.1

Solution: First, we will calculate what is inside the square root, and then square root the value. The mean of 3.15 was calculated earlier. The formula is given by $\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\sum_x (x-\mu)^2 P(x)}$, thus, we have

$$\begin{aligned}(0-3.15)^2 \times 0.05 &= 0.496\\ (1-3.15)^2 \times 0.1 &= 0.462\\ (2-3.15)^2 \times 0.2 &= 0.265\\ (3-3.15)^2 \times 0.25 &= 0.006\\ (4-3.15)^2 \times 0.2 &= 0.145\\ (5-3.15)^2 \times 0.1 &= 0.342\\ (6-3.15)^2 \times 0.1 &= 0.812\end{aligned}$$

Now we sum these and take the square root.

$$0.496 + 0.462 + 0.265 + 0.006 + 0.145 + 0.342 + 0.812 = 2.527$$

$\sqrt{2.527} = \boxed{1.589}$. The number of customers entering in an hour is expected to vary by about 1.589 customers from the average of 3.15 customers.

Problem 4.2.7 — Calculate and interpret the expected value and standard deviation of this table of number of goals scored in a soccer match:

Goals	Probability
0	0.3
1	0.35
2	0.2
3	0.1
4	0.05

Solution: The expected value is

$$0(0.3) + 1(0.35) + 2(0.2) + 3(0.1) + 4(0.05) = 1.25.$$

On average, a soccer team is expected to score 1.25 goals in a match.

Since we know the mean, we can use $\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\sum_x (x-\mu)^2 P(x)}$. Thus, we have

$$(0 - 1.25)^2 \times 0.3 = 0.46875$$
$$(1 - 1.25)^2 \times 0.35 = 0.021875$$
$$(2 - 1.25)^2 \times 0.2 = 0.1125$$
$$(3 - 1.25)^2 \times 0.1 = 0.30625$$
$$(4 - 1.25)^2 \times 0.05 = 0.378125$$

$$0.46875 + 0.021875 + 0.1125 + 0.30625 + 0.378125 = 1.2875$$

$\sqrt{1.2875} = \boxed{1.135}$

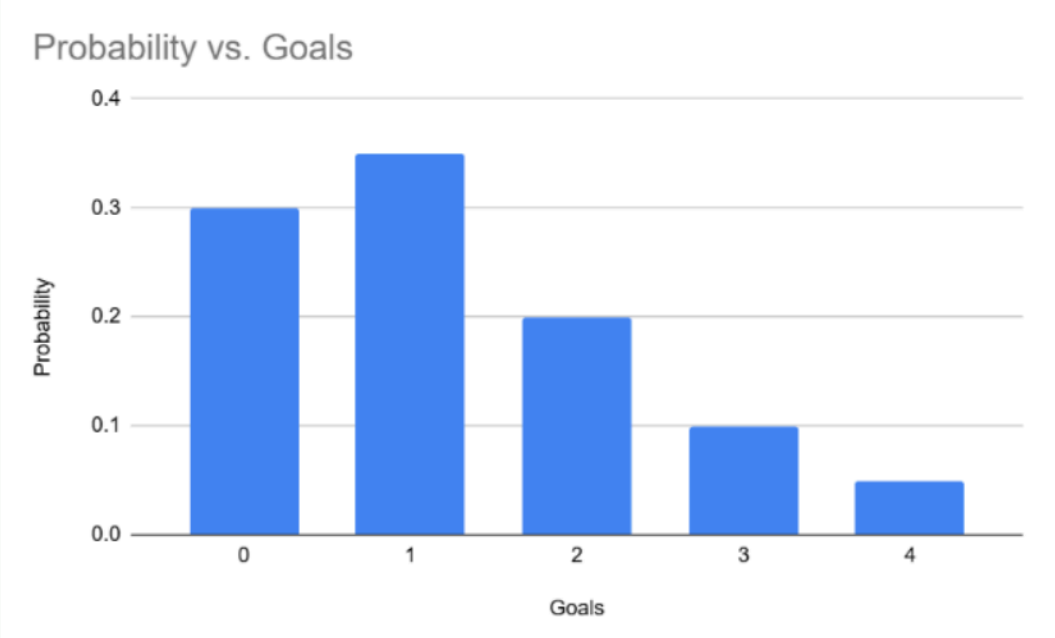
The number of goals scored in a match is expected to vary by 1.135 goals from the average of 1.25 goals.

You will be asked how to describe probability distributions. For these, you will need to identify if it is discrete or continuous, and the

1. Shape
2. Center
3. Spread

Don't forget to add context in your description. Identifying the shape is the same as Unit 1, you look to see if values are skewed left or right, uniform, symmetric, bimodal, or unimodal. For the center, find median or mean. Spread is determined from the standard deviation.

Problem 4.2.8 — Describe the shape of this probability distribution (1 goal has a probability of 0.35 and 4 goals has a probability of 0.05):



Solution: The shape of the probability distribution of goals is skewed right and unimodal. The median is at 1 goal (since that is where .5 probability is) and the mean is the expected value calculated earlier of 1.25 goals. The standard deviation was calculated to be 1.135.

When given two sets of data, you will be expected to know how to combine both.

Note 4.2.9

Mean:

To combine a sum of two different variables, you use the formula $\mu_{X+Y} = \mu_X + \mu_Y$. Add the two means of the data sets. If you are combining a difference of two variables, you use the formula $\mu_{X-Y} = \mu_X - \mu_Y$, meaning that you subtract the means of the two datasets.

Note 4.2.10

Two random variables are independent if they do not affect the probability of each other. For example, if there is a variable X, which is the hours a student studies for an exam, and a variable Y, which is the score the student achieves on the exam, these variables are not independent, meaning the standard deviation can't be combined, only the mean (expected value).

Note 4.2.11

Variance and Standard Deviation:

To find the standard deviation of the combination of two data sets, the data sets will have to be independent. The standard deviations cannot be combined directly, instead, the variances, which are the standard deviations squared, have to be added, then the sum of that is square rooted. Whether you are combining a sum or difference of two datasets, the procedure is the same. Thus, $\sigma_{X\pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

Problem 4.2.12 — Assume independence for the problems and calculate the mean and standard deviation:

a)

Exam Scores	Probability
80	0.4
85	0.5
90	0.1

Homework Scores	Probability
88	0.3
92	0.6
95	0.1

Calculate the mean and standard deviation for differences between homework and exam scores.

b)

Car A Distance	Probability
100	0.2
120	0.5
140	0.3

Car B Distance	Probability
90	0.4
110	0.4
130	0.2

Calculate the mean and standard deviation for the sum of distances for Car A and Car B in miles.

Solution to part a: We are looking for μ_{X-Y}, which is equal to $E(\text{Difference}) = E(\text{Homework}) - E(\text{Exam})$. The mean of the exam scores is:

$$E(\text{Exam}) = 80(0.4) + 85(0.5) + 90(0.1) = 32 + 42.5 + 9 = 83.5$$

The mean of the homework scores is:

$$E(\text{Homework}) = 88(0.3) + 92(0.6) + 95(0.1) = 26.4 + 55.2 + 9.5 = 91.1$$

Thus, the mean difference is:

$$E(\text{Difference}) = E(\text{Homework}) - E(\text{Exam}) = 91.1 - 83.5 = 7.6$$

On average, the difference between the homework and exam scores is 7.6.

Since we know the mean, we can proceed to calculating the standard deviation. $\sigma_{\text{Exam+Homework}} = \sqrt{\sigma^2_{\text{Exam}} + \sigma^2_{\text{Homework}}}$, so we need to find the standard deviation of exam and homework individually. For the exam scores:

$$(80 - 83.5)^2 \times 0.4 = 12.25 \times 0.4 = 4.9,$$
$$(85 - 83.5)^2 \times 0.5 = 2.25 \times 0.5 = 1.125,$$
$$(90 - 83.5)^2 \times 0.1 = 42.25 \times 0.1 = 4.225.$$

$$4.9 + 1.125 + 4.225 = 10.25 = \sigma^2_{\text{Exam}}$$

For the homework scores:

$$(88 - 91.1)^2 \times 0.3 = 9.61 \times 0.3 = 2.883,$$
$$(92 - 91.1)^2 \times 0.6 = 0.81 \times 0.6 = 0.486,$$
$$(95 - 91.1)^2 \times 0.1 = 15.21 \times 0.1 = 1.521.$$

$$2.883 + 0.486 + 1.521 = 4.89 = \sigma^2_{\text{Homework}}$$

$\sqrt{\sigma^2_{\text{Exam}} + \sigma^2_{\text{Homework}}} = \sqrt{10.25 + 4.89} = 3.89$. The difference between homework and exam scores is expected to differ by 3.89 from the mean of 7.6.

Solution to part b: We are looking for μ_{X+Y}, which is equal to $E(\text{Sum}) = E(\text{Car A}) + E(\text{Car B})$.

The mean distance for Car A is:

$$E(\text{Car A}) = 100(0.2) + 120(0.5) + 140(0.3) = 20 + 60 + 42 = 122$$

The mean distance for Car B is:

$$E(\text{Car B}) = 90(0.4) + 110(0.4) + 130(0.2) = 36 + 44 + 26 = 106$$

Thus, the mean of the sum of distances is:

$$E(\text{Sum}) = E(\text{Car A}) + E(\text{Car B}) = 122 + 106 = 228$$

On average, the sum of distances for Car A and Car B is 228 miles.

Since we know the mean, we can proceed to calculating the standard deviation. $\sigma_{\text{Car A+Car B}} = \sqrt{\sigma^2_{\text{Car A}} + \sigma^2_{\text{Car B}}}$, so we need to find the standard deviation of Car A and Car B individually.

For Car A:

$$(100 - 122)^2 \times 0.2 = 484 \times 0.2 = 96.8,$$
$$(120 - 122)^2 \times 0.5 = 4 \times 0.5 = 2,$$
$$(140 - 122)^2 \times 0.3 = 324 \times 0.3 = 97.2.$$

$$96.8 + 2 + 97.2 = 196 = \sigma^2_{\text{Car A}}$$

For Car B:

$$(90-106)^2 \times 0.4 = 256 \times 0.4 = 102.4,$$
$$(110-106)^2 \times 0.4 = 16 \times 0.4 = 6.4,$$
$$(130-106)^2 \times 0.2 = 576 \times 0.2 = 115.2.$$
$$102.4 + 6.4 + 115.2 = 224 = \sigma^2_{\text{Car B}}$$

Now, we calculate the standard deviation of the sum:

$$\sqrt{\sigma^2_{\text{Car A}} + \sigma^2_{\text{Car B}}} = \sqrt{196 + 224} = \sqrt{420} = 20.49$$

The sum of distances for Car A and Car B is expected to differ by 20.49 miles from the mean of 228 miles.

Note 4.2.13

Transformations:

When transforming a set of random variables, adding a number to every value in that set will change the mean by that number. If the numbers in a random variable are 1, 2, and 3, adding 1 to each value increases the mean by 1. However, the standard deviation is unaffected, because there is no change in how spread out the numbers are. When multiplying a random variable by a certain value, both the mean and standard deviation change by that factor. If a random variable has a mean of 4 and standard deviation of 1, multiplying the random variable by 5 will yield a mean of 20 and standard deviation of 5.

Problem 4.2.14 — Answer each question:

a) A random variable X has a mean of 10 and a standard deviation of 3. The random variable Y is equal to $3X + 4$. Find the standard deviation and mean of Y.

b) The score for a math test in a class has a mean of 25 and standard deviation of 5. The teacher decides to triple the score for each student. What is the mean and standard deviation of the new score?

c) The daily high temperatures in a city are recorded in Celsius. The mean is 23 and the standard deviation is 2. In order to convert to Fahrenheit, the formula is $F = 1.8X + 32$. What is the mean and standard deviation of the temperatures in Fahrenheit?

Solution to part a: To find the mean of Y, we simply substitute the mean of X into the random variable. Thus, we have

$$3(10) + 4 = \boxed{34} = \mu_Y$$

For the standard deviation, we substitute the standard deviation, but do **NOT** add, only substitute the value and multiply because adding does not change the variability.

$$3(3) = \boxed{9} = \sigma_Y$$

Solution to part b: If the scores are tripled, the mean and standard deviation are also tripled, thus

$$25 \times 3 = \boxed{75} = \mu$$

$$5 \times 3 = \boxed{15} = \sigma$$

Solution to part c: To find the mean of the temperature in Fahrenheit, substitute the mean in Celsius into the equation for Fahrenheit. For the standard deviation, only do the multiplication, because adding has no effect on the variability.

$$1.8(23) + 32 = \boxed{73.4 \text{ degrees Fahrenheit}} = \mu_F.$$

$$1.8(2) = \boxed{3.6 \text{ degrees Fahrenheit}} = \sigma_F$$

Example 4.2.15

A bakery sells loaves of bread with a mean profit of \$3.25 for each loaf and a standard deviation of \$0.50 per loaf. If the bakery sells 120 loaves in a week, what is the expected profit and standard deviation of the profit?

Solution: To find the mean, we multiply the mean profit per loaf (\$3.25) by the number of loaves sold (120), thus

$$3.25 \times 120 = \boxed{\$390}$$

For the standard deviation, we multiply .50 by $\sqrt{120}$. Remember that when combining independent random variables, the standard deviation is combined with variances (the standard deviations squared). Since 120 loaves are being sold, think of it as 120 random variables being combined. Thus, we have

$$.50 \times \sqrt{120} = \boxed{\$5.48}$$

For understanding **why** this works, understand that the variance of the profit is $.50^2 = .25$. Since there are 120 loaves, the total variance for selling 120 loaves is $.25 \times 120 = 30$. To find the standard deviation, the variance is square rooted, so

$$\sqrt{30} = \boxed{\$5.48}.$$

If there is a problem that involves a certain amount of quantities and calculations of the expected profit and standard deviation, calculate the standard deviation by multiplying by the square root of the amount of quantities.

Problem 4.2.16 — 2023 AP Statistics

Bath fizzies are mineral tablets that dissolve and create bubbles when added to bathwater. In order to increase sales, the Fizzy Bath Company has produced a new line of bath fizzies that have a cash prize in every bath fizzy. Let the random variable, X, represent the dollar value of the cash prize in a bath fizzy. The probability distribution of X is shown in the table.

Cash prize, x	\$1	\$5	\$10	\$20	\$50	\$100
Probability of cash prize, $P(X = x)$	$P(X = \$1)$	0.2	0.05	0.05	0.01	0.01

a) Based on the probability distribution of X, answer the following. Show your work.

   i. Calculate the proportion of bath fizzies that contain \$1.

   ii. Calculate the proportion of bath fizzies that contain at least \$10.

b) Based on the probability distribution of X, calculate the probability that a randomly selected bath fizzy contains \$100, given that it contains at least \$10. Show your work.

c) Based on the probability distribution of X, calculate and interpret the expected value of the distribution of the cash prize in the bath fizzies. Show your work.

d) The Fizzy Bath Company would like to sell the bath fizzies in France, where the currency is euros. Suppose the conversion rate for dollars to euros is 1 dollar = 0.89 euros. Using your expected value from part (c), calculate the expected value, in euros, of the distribution of the cash prize in the bath fizzies. Show your work.

Solution to part ai: We know that the probability values in the table need to sum to 1. We know all probability values except for the one at a cash price of \$1, so we add those up and subtract it from one to find the unknown probability.

$$P(X = \$1) = 1 - (0.2 + 0.05 + 0.05 + 0.01 + 0.01) = 1 - .32 = \boxed{.68}$$

Solution to part aii: We are looking for $P(X \geq \$10)$, which breaks down to

$$P(X = \$10) + P(X = \$20) + P(X = \$50) + P(X = \$100)$$

$$0.05 + 0.05 + 0.01 + 0.01 = \boxed{0.12}$$

Solution to part b: We are looking for $P(\$100 \mid \text{at least } \$10)$, which breaks down to

$$\frac{P(\$100 \cap \text{at least} \$10)}{P(\text{at least } \$10)}$$

$$\frac{P(\$100 \cap \text{at least } \$10}{P(X = \$10) + P(X = \$20) + P(X = \$50) + P(X = \$100)}$$

$$\frac{0.01}{0.12} = \boxed{0.0833}$$

Solution to part c: The expected value of the distribution of the cash prize in the bath fizzies is the probability of a cash prize multiplied by that cash prize, thus

$$1(0.68) + 5(0.2) + 10(0.05) + 20(0.05) + 50(0.01) + 100(0.01) = \boxed{\$4.68}$$

On average, the expected value of the distribution of the cash prize in the bath fizzies is \$4.68.

Solution to part d: To convert the expected value from dollars to euros, we need to do a transformation. Since 1 dollar = 0.89 euros, we multiply the mean in dollars by 0.89.

$$4.68 \times 0.89 = \boxed{4.17 \text{ euros}}$$

Problem 4.2.17 — 2015 AP Statistics

A shopping mall has three automated teller machines (ATMs). Because the machines receive heavy use, they sometimes stop working and need to be repaired. Let the random variable X represent the number of ATMs that are working when the mall opens on a randomly selected day. The table shows the probability distribution of X.

Number of ATMs working when the mall opens	0	1	2	3
Probability	0.15	0.21	0.40	0.24

a) What is the probability that at least one ATM is working when the mall opens?

b) What is the expected value of the number of ATMs that are working when the mall opens?

c) What is the probability that all three ATMs are working when the mall opens, given that at least one ATM is working?

d) Given that at least one ATM is working when the mall opens, would the expected value of the number of ATMs that are working be less than, equal to, or greater than the expected value from part (b) ? Explain.

Solution to part a: We are looking for P(At least 1 ATM).

$$\text{P(At least 1 ATM)} = \text{P(1 ATM)} + \text{P(2 ATM)} + \text{P(3 ATM)}$$

$$0.21 + 0.40 + 0.24 = \boxed{0.85}.$$

Solution to part b: The expected value of the number of ATMs that are working when the mall open is equal to

$$0(0.15) + 1(0.21) + 2(0.40) + 3(0.24) = \boxed{1.73 \text{ ATMs}}$$

Solution to part c: We are looking for $P(\text{All three ATMs} \mid \text{at least 1 ATM})$

$$\frac{P(\text{All 3} \cap \text{At least 1})}{P(\text{At least 1})}$$

$$\frac{P(\text{All 3})}{P(1\text{ ATM}) + P(2\text{ ATM}) + P(3\text{ ATM})}$$

$$\frac{0.24}{0.21 + 0.40 + 0.24} = \frac{0.24}{0.85} = \boxed{0.282}.$$

Solution to part d: Given that at least one ATM is working when the mall opens, the expected value of the number of ATMs that are working is greater than the expected value from part b because the possibility of 0 ATMs is eliminated. Without the possibility of no ATMs working, the expected value increases.

§4.3 Probability Distributions

A probability distribution is a list that provides the probability of each outcome in that distribution.

Note 4.3.1

Binomial Random Variable:

A binomial distribution is a type of probability distribution that has only two outcomes. It has a fixed number of trials and each attempt is independent of each other. The probability of success does not change from each trial to another. The probability of success is p and the probability of failure is 1-p.

To check if a distribution is binomial, use the acronym:

1. B: Binary, there are two different outcomes.
2. I: Independent, the result of one trial does not affect the result of another. Either subjects of each trial are chosen with replacement or 10% or less of the total population is chosen.
3. N: Number of trials is determined beforehand.
4. S: Success probability stays the same for each trial.

Problem 4.3.2 — Check if these represent binomial distributions:

a) A coin is flipped 10 times and we are interested in how many times heads comes up. Is this a binomial distribution?

b) A factory creates light bulbs, and 50 light bulbs are randomly selected every day. The probability of a bulb being defective is 0.05. We are interested in the number of defective light bulbs in a sample, is this a binomial distribution?

Solution to part a: The only two outcomes are heads or tails, one trial does not affect another, there is a predetermined number of 10 flips, and the probability remains 0.5 for heads and 0.5 for tails, thus, this is a binomial distribution.

Solution to part b: The bulbs can either be defective or not defective. 50 light bulbs is a small number for a factory to produce, so we can assume that the trials are independent.

There are 50 predetermined trials, and the probability is always 0.05, thus, this is a binomial distribution.

You will be expected to know how to calculate probabilities given binomial distributions on the exam.

The binomial probability function is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $x = 0, 1, 2, \ldots, n$ and is the number of successes, n is the number of trials, and p is the probability of success.

This may look complicated, but just follow the next examples closely:

Example 4.3.3

Use these questions as examples for probability calculations:

a) A coin is flipped 10 times, what is the probability of getting 6 heads?

b) A machine creates 90% normal parts and 10% defective parts. In a batch of 12 parts, what is the probability that at most 2 parts are defective?

c) A university has an acceptance rate of 70%. If there is a group of 15 applicants, what is the probability that at least 12 are accepted?

Solution to part a: We are looking for $P(X = 6)$. There are many ways this could happen; 6 heads could be flipped consecutively, or 5 heads could be flipped, the next 4 are tails, and the last is heads. Thinking of all the possible ways 6 heads could be flipped is incredibly long and tedious, and because of that, it can be simplified with either an equation or a calculator. $\binom{n}{x}$ is the amount of ways that a 6 can be flipped, which is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

n is the number of trials and x is the number of successes, thus

$$\binom{10}{6} = \frac{10!}{6!(10-6)!} = 210$$

Alternatively on your calculator can be used. Locate where the nCr command is, and put the number of trials to the left of C, and number of successes to the right. The calculated value will also be 210.

Using the binomial probability function, we have

$$P(X = 6) = 210(0.5)^6(0.5)^4 = \boxed{0.205}$$

Although the above work should be understood for calculating probability, using the binomialPDF function is much more efficient. Since this question is asking for the probability of getting an exact value (not at least or at most, asks for a certain point), we use the binomialPDF featured on graphing calculators.

$$\text{binomialPDF(x} = 6, \text{ n} = 10, \text{ p} = 0.5) = \boxed{0.205}.$$

Note 4.3.4

If you are ever asked a question where you are asked a probability and there is only one way to get that outcome, there is no need to use binomialPDF or the $\binom{n}{x}$ or the nCr portion of the calculation. Compute the probability directly. For example, if you are looking for the probability that five free throws are made in a row, there is only one way for that to happen, meaning $\binom{n}{x}$ is unnecessary.

Solution to part b: If we were to compute this value using the binomial probability function, we would have to find

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

While this is certainly doable, it can be streamlined by simply using the binomialCDF function.

$$\text{binomialCDF(x} = 2, \text{ n} = 12, \text{ p} = 0.1) = \boxed{.889}$$

Solution to part c: The binomialCDF function yields us the probability up to a certain value. We are looking for the probability that at least 12 applicants are accepted, therefore, we do 1 minus the probability that at most 11 applicants are accepted.

$$1 - \text{binomial(x} = 11, \text{ n} = 15, \text{ p} = 0.7) = 1 - 0.703 = \boxed{0.297}.$$

Problem 4.3.5 — A quiz has 8 multiple choice questions, with 4 possible answers on each question.

a) If a student guesses on all 8 questions, what is the probability that they answer exactly 3 correct?

b) If a student guesses on all 8 questions, what is the probability that they get at least two correct?

c) If a student guesses on all 8 questions, what is the probability that they get at most 6 correct?

Solution to part a: The probability of guessing one question correct is $\frac{1}{4} = .25$. Using binomialPDF, we have

$$\text{P(X} = 3) \ = \text{binomialPDF(x} = 3, \text{ n} = 8, \text{ p} = .25) = \boxed{0.208}$$

Solution to part b: Since this question involves getting at least two correct, we do 1 minus the probability of getting at most one correct.

$$P(X \geq 2) = 1 - \text{binomialCDF(x} = 1, \text{ n} = 8, \text{ p} =. \ 25) = 1 - 0.367 = \boxed{0.633}$$

Solution to part c: Using binomialCDF, we have

$$P(X \leq 6) = \text{binomialCDF(x} = 6, \text{ n} = 8, \text{ p} = .25) = \boxed{.999}$$

Note 4.3.6

Mean and Standard Deviation:

If a random variable has a binomial distribution, you can calculate the mean using the formula $\mu = np$, where n is the number of trials, and p is the probability of success for a trial. To calculate the standard deviation, the formula is $\sigma = \sqrt{np(1-p)}$.

Problem 4.3.7 — Answer each question:

a) A factory creates light bulbs and the probability that a light bulb is defective is 0.05. If there are 100 light bulbs randomly selected, what is the mean and standard deviation of number of defective light bulbs?

b) A basketball player has a free throw success rate of 0.8. If this player attempts 10 free throws, what is the mean and standard deviation of number of successful free throws?

Solution to part a: Bulbs are either defective or not defective, there are enough light bulbs in a factory that bulbs selected are independent, there is a fixed number of 100 light bulbs, and the probability remains 0.05 of a bulb being defective. This a binomial distribution, so we can apply the mean and standard deviation formulas. Thus, the mean is

$$\mu = 100(0.05) = \boxed{5 \text{ bulbs}}$$

The standard deviation is

$$\sigma = \sqrt{100 \times 0.05(1-0.05)} = \boxed{2.18 \text{ bulbs}}$$

Solution to part b: Either a free is made or missed, each attempt is independent of each other, there is a fixed number of 10 free throws, and the probability remains 0.8. Since this is a binomial distribution, we can apply the mean and standard deviation formulas.
The mean is

$$\mu = 10(0.8) = \boxed{8 \text{ free throws}}$$

The standard deviation is

$$\sigma = \sqrt{10 \times 0.8(1-0.8)} = \boxed{1.265 \text{ free throws}}$$

Note 4.3.8

Geometric Random Variable:

A geometric random variable has two possible outcomes, however, the number of trials is counted until the first success. There is no fixed number of trials. It is also independent and the probability of success must remain the same, however, it is different from a binomial random variable because there is no fixed number of trials. The probability of success is p and the probability of failure is 1-p.

The geometric probability function is:

$$P(X = x) = (1 - p)^{x-1} \cdot p, \quad x = 1, 2, 3, \ldots$$

Where x is the number of trials, and p is the probability.

Example 4.3.9

A factory creates light bulbs, and the probability that a light bulb is defective is 0.1:

a) Find the probability that the first defective light bulb occurs on the 5th trial.

b) Find the probability that the first defective light bulb occurs before the 3rd trial.

c) Find the probability that the first defective light bulb occurs after the 4th trial.

Solution to part a: Using the geometric probability function, we have

$$P(X = 5) = (0.9)^4 \times .1 = \boxed{0.066}.$$

Alternatively, geometricPDF can be used.

$$P(X = 5) = \text{geometricPDF}(\text{x} = 5, \text{ p} = 0.1) = \boxed{0.066}$$

Solution to part b: We are looking for $P(X < 3)$, which is $P(X = 1) + P(X = 2)$. This is certainly doable with the geometric probability function, but geometricCDF simplifies the calculation.

$$P(X < 3) = \text{geometricCDF}(\text{x} = 2, \text{ p} = 0.1) = \boxed{.19}.$$

Solution to part c: Since we are looking for the probability after the 4th certain trial, we do 1 minus the probability that the defective light bulb occurs on the 4th or below.

$$P(X > 4) = 1 - \text{geometricCDF}(\text{x} = 4, \text{ p} = 0.1 = 1 - 0.3439 = \boxed{0.656}.$$

Problem 4.3.10 — A restaurant wants to know how many customers they expect to serve before a customer complains about food quality. The probability of a customer complaining is 0.15.

a) Find the probability that the first complaint occurs on the 2nd customer.

b) Find the probability that the first complaint occurs after the 3rd customer.

c) Find the probability that the first complaint occurs within the first 5 customers.

Solution to part a: The probability that the first complaint occurs on the 2nd customer is

$$P(X = 2) = \text{geometricPDF}(\text{x} = 2, \text{ p} = 0.15) = 0.1275$$

Solution to part b: The probability that the first complaint occurs after the 3rd customers is 1 minus the probability that it occurred on or before the 3rd customer. Thus,

$$P(X > 3) = 1 - \text{geometricCDF}(\text{x} = 3,\ \text{p} = 0.15) = 1 - 0.386 = \boxed{0.614}$$

Solution to part c: Using geometricCDF, the probability that the first complaint occurs within the first 5 customers is

$$P(X \leq 5) = \text{geometricCDF}(\text{x} = 5,\ \text{p} = 0.15) = \boxed{0.556}$$

Note 4.3.11

Mean and Standard Deviation:

The mean of a geometric random variable is $\mu = \frac{1}{p}$, where p is the probability of success. The standard deviation is $\sigma = \sqrt{\frac{1-p}{p^2}}$.

Note 4.3.12

Remember that in binomial or geometric distributions, the probability of success or failure remains the same for all trials. If there are two successes, the probability of a success happening on the third trial is still the original probability. Past outcomes do not affect future ones. When we are calculating the probability of multiple successes in a row, the probability is different than the original one because we are finding the probability of multiple events occurring.

Problem 4.3.13 — 2016 AP Statistics

A company manufactures model rockets that require igniters to launch. Once an igniter is used to launch a rocket, the igniter cannot be reused. Sometimes an igniter fails to operate correctly, and the rocket does not launch. The company estimates that the overall failure rate, defined as the percent of all igniters that fail to operate correctly, is 15 percent.

A company engineer develops a new igniter, called the super igniter, with the intent of lowering the failure rate. To test the performance of the super igniters, the engineer uses the following process.

Step 1: One super igniter is selected at random and used in a rocket
Step 2: If the rocket launches, another super igniter is selected at random and used in a rocket.

Step 2 is repeated until the process stops. The process stops when a super igniter fails to operate correctly or 32 super igniters have successfully launched rockets, whichever comes first. Assume that super igniter failures are independent.

a) If the failure rate of the super igniters is 15 percent, what is the probability that the first 30 super igniters selected using the testing process successfully launch rockets?

b) Given that the first 30 super igniters successfully launch rockets, what is the probability that the first failure occurs on the thirty-first or the thirty-second super igniter tested if the failure rate of the super igniters is 15 percent?

c) Given that the first 30 super igniters successfully launch rockets, is it reasonable to believe that the failure rate of the super igniters is less than 15 percent? Explain.

Solution to part a: This represents a geometric distribution. The failure rate is 15 percent, meaning that the success rate is $1 - 0.15 = .85$, or 85 percent. Thus, the probability that the first 30 super igniters selected successfully launch rockets is

$$(.85)^{30} = \boxed{0.0076}$$

Solution to part b: In this geometric distribution, the probability that the first failure occurs on the 31st trial is 0.15, and the probability that the first failure does not occur on the 31st trial, but instead on the 32nd trial is

$$(0.85)(0.15) = 0.1275$$

Thus, the probability that the first failure occurs on the thirty-first or the thirty-second super igniter tested is

$$0.15 + 0.1275 = \boxed{0.2775}$$

Solution to part c: If the first 30 super igniters successfuly launch rockets, it is reasonable to believe that the failure rate of the super igniters is less than 15 percent because having 30 igniters successfully launch consecutively is 0.0076, which is incredibly low.

Problem 4.3.14 — 2021 AP Statistics

To increase morale among employees, a company began a program in which one employee is randomly selected each week to receive a gift card. Each of the company's 200 employees is equally likely to be selected each week, and the same employee could be selected more than once. Each week's selection is independent from every other week.

a) Consider the probability that a particular employee receives at least one gift card in a 52-week year.

   i. Define the random variable of interest and state how the random variable is distributed.

   ii. Determine the probability that a particular employee receives at least one gift card in a 52-week year. Show your work.

b) Calculate and interpret the expected value for the number of gift cards a particular employee will receive in a 52-week year. Show your work.

c) Suppose that Agatha, an employee at the company, never receives a gift card for an entire 52-week year. Based on her experience, does Agatha have a strong argument that the selection process was not truly random? Explain your answer.

Solution to part ai: The random variable of interest is the number of gift cards that an employee receives in a 52-week year. An employee either receives or does not receive a gift card, each selection is indpendent from every other week, there are 52 trials, and the probability remains $\frac{1}{200} = 0.005$ for each trial, meaning that this is a binomial distribution.

Solution to part aii: The probability that an employee receives at least one gift card in a 52-week year is 1 minus the probability of receiving no gift cards, thus

$$P(X \geq 1) = 1 - \text{binomialCDF(x} = 0, \text{ n} = 52, \text{ p} = 0.005) = 1 - 0.7705 = \boxed{0.2295}$$

Solution to part b: The expected value, or mean for the number of gift cards a particular employee will receive in a 52-week year is

$$\mu = np = 0.005(52) = 0.26 \text{gift cards}$$

On average, an employee can expect to receive 0.26 gift cards.

Solution to part c: Agatha does not have a strong argument that the selection process was not truly random. The probability of never receiving a gift card for an entire 52-week year is

$$(.995)^{52} = 0.771$$

Since there is a strong likelihood that an employee does not receive a gift card with a random selection process, Agatha does not have a strong argument.

Problem 4.3.15 — 2018 AP Statistics
Approximately 3.5 percent of all children born in a certain region are from multiple births (that is, twins, triplets, etc.). Of the children born in the region who are from multiple births, 22 percent are left-handed. Of the children born in the region who are from single births, 11 percent are left-handed.

a) What is the probability that a randomly selected child born in the region is left-handed?

b) What is the probability that a randomly selected child born in the region is a child from a multiple birth, given that the child selected is left-handed?

c) A random sample of 20 children born in the region will be selected. What is the probability that the sample will have at least 3 children who are left-handed?

Solution to part a: 3.5 percent of children are born from multiple births, thus 1 - 0.035 = 0.965 = 96.5 percent of children are born from single births.

$$P(L) = P(M \cap L) + P(S \cap L)$$

$$P(L) = 0.035(0.22) + 0.965(0.11) = \boxed{0.1139}$$

Solution to part b: We know that P(L) = 0.1139 and $P(M \cap L) = 0.035(0.22) = 0.0077$, and we are looking for $P(M \mid L)$. Thus

$$P(M \mid L) = \frac{P(M \cap L}{P(L)}$$

$$P(M \mid L) = \frac{0.0077}{0.1139} = \boxed{0.068}$$

Solution to part c: The children are either right or left-handed, each sample is independent of each other since a region is a large population, there are 20 children, and the probability remains the same, this is a binomial distribution. Thus, the probability that the sample will have at least 3 children who are left-handed is 1 minus the probability that the sample will have 2 children or less who are left-handed.

$$P(X \geq 3) = 1 - \text{binomialCDF(x = 2, n = 20, p = 0.1139)} = 1 - 0.597 = \boxed{0.403}$$

5 Unit 5: Sampling Distributions

§5.1 Introducing Statistics: Why Is My Sample Not Like Yours?

Let's start off by introducing what a sampling distribution is.

Definition 5.1.1

A **sampling distribution** is the distribution of a specific statistic for all possible samples of size n of a population.

Note 5.1.2

Importantly, this differs from the *sample distribution*, which is the distribution of the chosen individuals in the single sample, as well as the population distribution, which represents the distribution of all individuals in the population.

The statistic mentioned for each sample will be either:

1. Means of some characteristic of each subject for quantitative data
2. Proportions (amount of yes/no) for the subjects for qualitative/categorical data
3. Differences in means/proportions if considering population pairs

These statistics are meant to provide estimates for the true corresponding values of the entire population. Statistics of different samples of a population will naturally vary. This phenomenon is called the *sampling variation* between samples. This variation is nearly unavoidable, and are either randomly or non-randomly caused due to bias in samples or other factors.
Additionally, the data for these variables can either be *discrete*, unable to be broken down into smaller parts at some point and can be counted, or *continuous*, able to take on all real values within some range. For example, discrete variables could be age, number of apples a container can carry, or how many times a dice rolls a 1. Examples of continuous variables include the weight of a ball, the volume of a container, or the distance that a ball is thrown out.
Usually, when looking at random variables with several outcomes and their probabilities, we are given the mean and standard deviation. However, if we aren't given them for discrete variables, then we can use the following:

Theorem 5.1.3 (Mean and Standard Deviation for Discrete Random Variables)

For a discrete random variable X with possible values $x_1, x_2, x_3, \ldots$ with corresponding probabilities $P(x_1), P(x_2), P(x_3), \ldots$, then:

- The mean is
$$\mu = \sum [x_i \cdot P(x_i)]$$
- The standard deviation is
$$\sigma = \sqrt{\sum [(x_i - \mu)^2 \cdot P(x_i)]}$$

We will learn how to apply these to sampling distribution's mean and standard deviation later in the chapter.
To find the standard deviation of a variable Z, the sum or difference of two random independent variables X and Y, then we use the fact that the variances of X and Y add to Z, so

Theorem 5.1.4 (Standard Deviation of the Sum of Two Random Variables)

For two random variables X and Y with standard deviations of σ_X and σ_Y, the standard deviation of their sum $(X + Y)$ or difference $(X - Y)$ is
$$\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

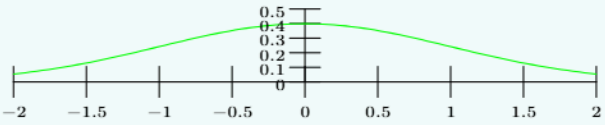
We will take differences more often than sums of these variables, as we will test the similarity of the variables through comparing how close the difference between the variables is to zero.

Although it is unfeasible to take all possible samples of a specific size from a population, we can often get a good picture of what it looks like by drawing enough samples of the size independently from each other.
In statistics from here, we will learn how these distributions can help to show whether specific events are likely or unlikely. (put before, with explanation of the fact that variation could be random or not)

§5.2 The Normal Distribution, Revisited

Definition 5.2.1 (Normal Curve)

As we have seen before, the normal curve is a symmetric bell-shaped, uni-modal, continuous *probability distribution*. Having a medium amount of outliers, the curve depends on its center (the mean), and its spread/variation of the values (the standard deviation).



The Normal Curve, here with an area of 1 and representing a probability distribution.

The normal distribution is the most important model for variation, as we will see in the next section.

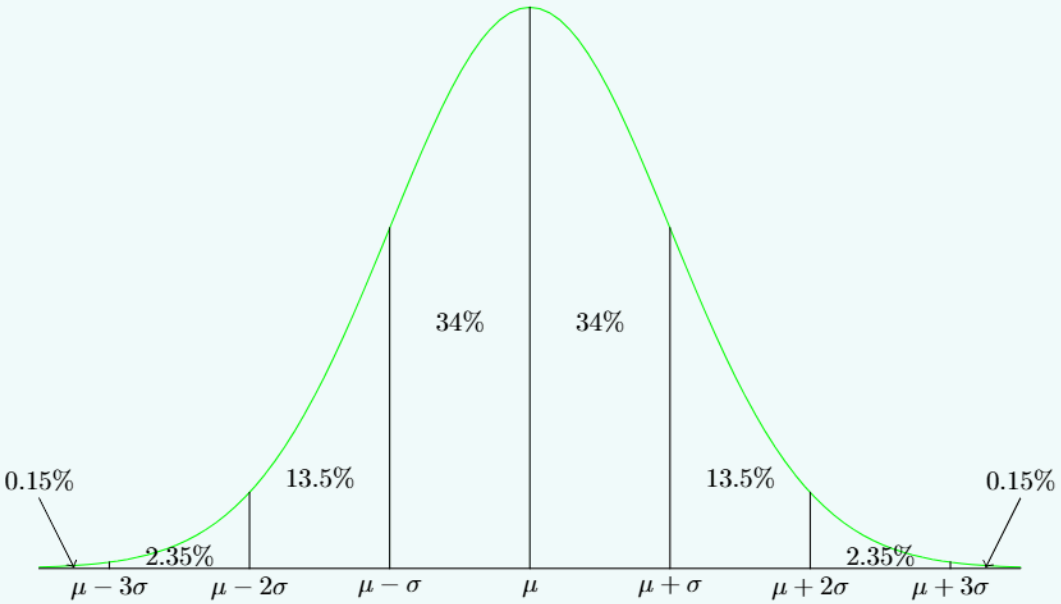
To describe the normal curve, we need to have its center, being the mean, median, or mode of the function, the standard deviation.

In the last chapter, you learned how to calculate the z-score of a specific value in a distribution, by using the formula $z = \dfrac{x - \mu}{\sigma}$. Now, let's learn how to calculate probabilities of specific ranges of values occurring with it.

The normal distribution can model many population's characteristics quite well, because if they are not simply random (such as dice) and are influenced by many factors, then those random factors (such as environment, genetics, etc.) will be all averaged into a normal-like distribution.

In particular, when we look at the normal curve, we see a mean value, and fixed areas from it to each of its standard deviation distances, as shown below.

Note 5.2.2 (Empirical Rule)



The probabilities of values in the population to fall in each region of the normal curve.

The curve above is the normal curve, and these percentages, marking the area of the curve on each interval, represent the probability that a specific value is between the two values of the curve.
Specifically,

- 68% of the values are within one standard deviation ($\pm\sigma$) from the mean
- 95% are within two standard deviations ($\pm 2\sigma$) from the mean
- 99.7% are within three standard deviations ($\pm 3\sigma$) from the mean.

The probabilities shown above are based on this 68-95-99.7 rule.
It's important to note that these values aren't exactly the proportion of values between the ranges; they are approximate; indeed, 68.4% and 95.2% are more accurate values.

The Normal Distribution can only perfectly model variables that are perfectly continuous. However, that doesn't mean that it can't be applied usefully discrete variables. Especially when taking larger sample sizes, the values of the sample means will begin to differ less and come closer to each other, which will produce a distribution that is almost continuous throughout the domain. Additionally, when looking at bins of discrete distributions, they also follow the shape of the normal curve closely.
Now, when we are asked the probabilities involving the ranges between integer z-scores between -3 and 3, then we can simply add the percentage values of the regions covered and say that as the probability.
However, more often than not, we are asked to calculate probabilities with z-scores different than $-3, -2, -1, 0, 1, 2$, or 3. To do this, we can use the calculator. To find the probability that a z-score is below the specified value, or on the contrary, find the z-score for a specific proportion below it, we can use the specified steps on the TI-84 calculator

as shown below:

Note 5.2.3 (Calculator: Probabilities of the Normal Curve)

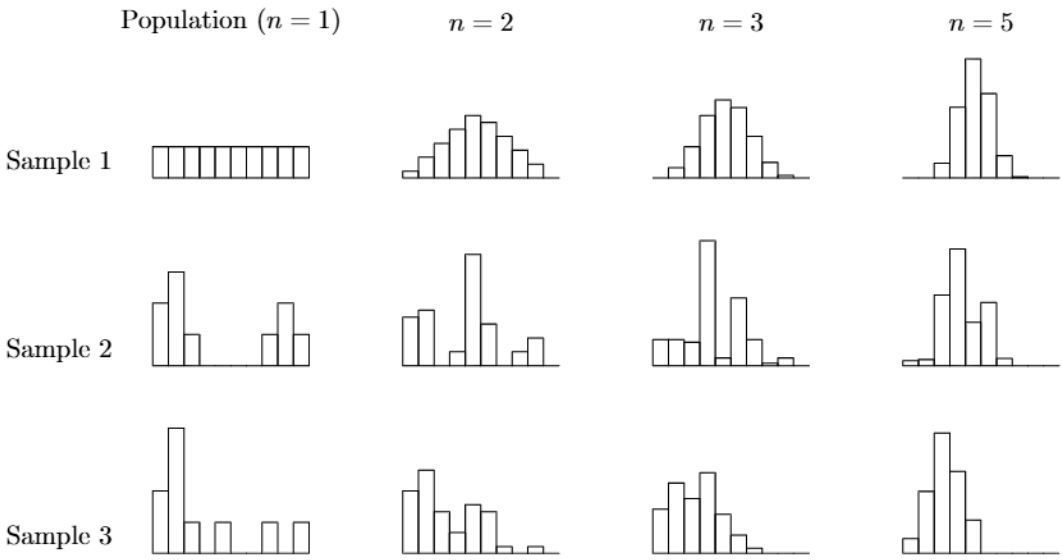
1. First, after pressing the On button, press `2nd`, then `VARS`. This takes you to the `DISTR` menu (short for distribution).
2. Once in the `DISTR` menu, there are two things we can do:
   a) Use the `normalcdf` button. This function takes four inputs, and appears as `normalcdf(lower bound, upper bound, mean, standard deviation)`. This finds the probability a specific variable described by a normal distribution with the inputted mean and standard deviation is between the lower bound and upper bound, inclusive.
   b) Use the `invNorm` button. This function takes three inputs, and appears as `invNorm(probability, mean, standard deviation)`. This finds the exact value of the variable that gives the specified probability of any value in the normal distribution with the mean and standard deviation given to be lower than that value.

   In this course, the function `normalpdf`, will give us the probability of getting exactly a specific value in the distribution, and won't be important to us.
   Also, if there is no lower or upper bound on the value, then we input $-\infty$ or ∞ in the corresponding areas, and then put in the real value of upper limit or lower limit of the variable in the remaining input.

[insert figure here of calculator]

§5.3 The Central Limit Theorem

As we recall, taking a sample is much more feasible and efficient than taking a sample, mainly for reasonably large populations. However, taking these samples has a glaring trade-off; especially with smaller sample sizes, the sample is highly unlikely to be exactly equal to the population parameter, and the sample statistic could be way off from the desired parameter.
However, especially after using larger sample sizes, the chosen sample will be less likely to accrue outliers and on average, will be closer to the true parameter. Therefore, when we look at a sampling distribution of sufficiently large sample sizes, we should see few ones far away from the center, and a gradual rise to a peak in the middle.
In fact, this is exactly what happens for these large sample sizes. Below are several population distributions of size 10 with sample sizes of 2, 3, and 5 to demonstrate this.



Sampling Distributions of populations with size 10, and samples of size n taken (distribution of means shown). Notice how for each distribution, the values get closer to the center, and it appears more normal.

As we can see, for any initial shape of the distribution, we seem to approach a normal distribution as sample sizes grows larger. Furthermore, the farther off a specific distribution is from the normal, the larger the sample size would need to be to achieve an approximately normal distribution. This could be characteristics of the original distribution such as skew, or bi-modality. This leads us to the following observation:

Theorem 5.3.1 (Central Limit Theorem)

Consider the distribution of sample means of a population. As the sample size n taken increases, the sampling distribution of such samples approaches a normal distribution.

This is a powerful result that allows us to accurately look at how samples are distributed when we choose large enough samples, as we see in later sections.

§5.4 Biased and Unbiased Point Estimates

To start, let's look at the different notations of the point estimates of each type of population parameter.

Definition 5.4.1 (Point Estimates)

In a sample, the calculated:

- mean $\overline{x}$ estimates the true mean μ
- proportion $\hat{p}$ estimates the true proportion p
- standard deviation s estimates the true standard deviation σ

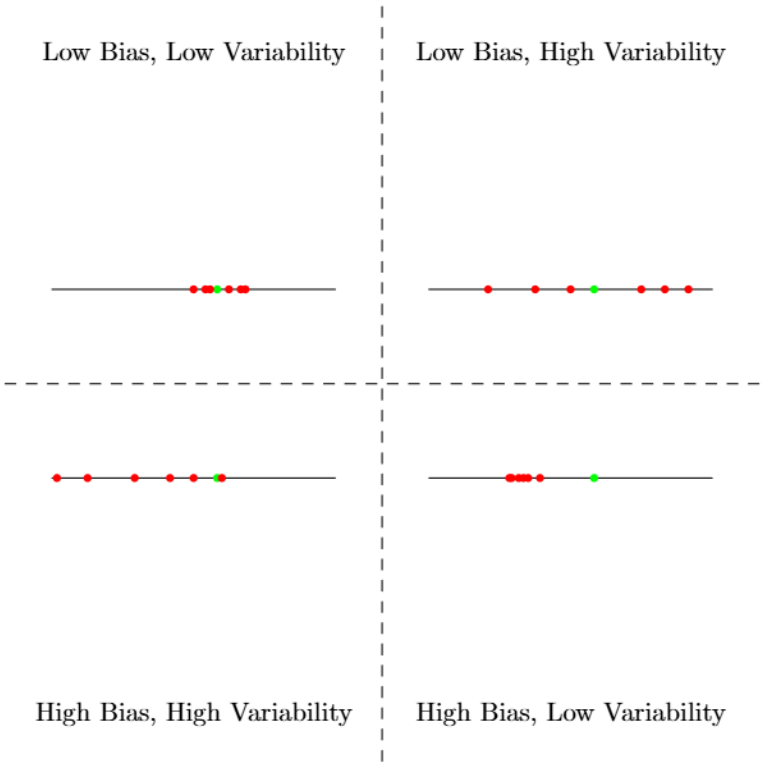
These all are sample statistics, estimating population parameters.

These are the most common examples of estimates for the true population parameters. Estimates are produced through using a process called the *estimator*.

Definition 5.4.2 (Estimators)

An estimator is *unbiased* if on average, the expected value of the statistic will be equal to the parameter.
On the other hand, an estimator is *biased* if on average, the expected value of the resulting statistic will **not** be equal to the parameter.

Notably, the most prominent example of an unbiased estimator is the mean from a random sample of a population as each member of the population is represented equally as much.
However, samples being wildly different from the true mean could just as well be due to the unavoidable *sampling error* when it comes to taking samples. This causes a specific *variability*, the variation between samples' statistics, among the samples.
Therefore, two factors; namely bias and variability affect the sampling distribution and therefore how a sample will compare to the population as well as other samples.



Examples of groups of sample statistics (red) showing each type of data approximating the true population parameter (green).

Biased estimators will not be balanced around the mean (a roughly equal occurrence of values above and below the desired parameter), but most values obtained through this will be only below/above the true value.
The biased estimators that tend to underestimate the true parameter, such as in the figure above, are called *negatively biased estimators*, while *positively biased estimators* tend to overestimate the true parameter.
The estimators that produce highly skewed distributions are much more likely to be biased estimators as there already is a tendency of overestimation or underestimation within the produced statistics.
On the other hand, several statistics, such as the sample mean or sample proportion, are expected to become more and more accurate to the true mean/proportion, as the

sample size increases. These estimators are called *consistent estimators*, but the expected reduction in variability doesn't necessarily mean that it isn't biased.
Especially when we take larger sample sizes, variability will eventually reduce as there are more and more values that represent the population more accurately, but the biased estimators will still be unable to accurately represent the true population parameter as their bias still remains.
In general, when given the actual population parameter (i.e. mean, median, range, maximum, etc.), we can see how many of the sample statistics of our chosen samples were less than or greater than it. If it is very unbalanced, we can conclude it is a biased estimator, and if it is approximately equal, we can possibly trust it to be an unbiased estimator.

§5.5 Sampling Distributions for Sample Proportions

Let us turn our attention towards actually figuring out distributions for our sample proportions *given that we know the proportion of our population.*

We know that our true proportion is $p = \dfrac{S}{N}$, where S is the number of individuals satisfying a certain characteristic of the population, and N is the population size.
If we let each individual in the population have a 1 if they have the trait, and a 0 if they don't, then the proportion is just the mean of all of the values for the people.[insert figure] This realization leads us to the following results:

Lemma 5.5.1 (Mean of the Distribution of Sampling Proportions)

Suppose that a sample of size n is chosen uniformly random from the population. If the true proportion of the population is p, then the mean of the sampling distribution is

$$\mu_{\hat{p}} = p$$

.

Proof. Take an arbitrary individual of the population, and let its value (0 or 1) be equal to ν. Then, the individual appears in $\binom{N-1}{n-1}$ of the $\binom{N}{n}$ samples.
Thus, in the entire sample distribution mean, the individual gives

$$\frac{\nu}{n} \cdot \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\nu}{n} \cdot \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{\nu}{n} \cdot \frac{n!/(n-1)!}{N!/(N-1)!} = \frac{\nu}{n} \cdot \frac{n}{N} = \frac{\nu}{N}$$

which is exactly what it would give in the actual population proportion. Summing over all individuals gives the desired. □

Essentially, we can reason that each individual matters equally in either situation, causing the two means to be equal.

Lemma 5.5.2 (Standard Deviation of the Distribution of Sample Proportions)

Suppose that we are taking uniformly random samples of size $n < N(10\%)$, where N is the total population size. Then, if the true proportion of the population satisfying a certain characteristic is p, then the standard deviation of the sampling distribution is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

.

Proof. We first find the standard deviation of the population itself. Recall that there are Np individuals with a 1, and $N(1-p)$ individuals with a 0, so we find it is $\sigma = \sqrt{\frac{Np \cdot (1-p)^2 + N(1-p) \cdot p^2}{N}} = \sqrt{p(1-p)^2 + (1-p)p^2} = \sqrt{p(1-p)[p+1-p]} = \sqrt{p(1-p)}$.

Now, when $n < N(10\%)$, we can safely equate the samples to be chosen with replacement, as opposed to without replacement without giving too much error. Therefore, when we view this with replacement, each sample mean is an average of n random variables P_1, P_2, ..., P_n, each with standard deviation $\sqrt{p(1-p)}$, and thus the standard deviation of the sampling distribution is

$$\sigma_{\hat{p}} = \frac{\sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2 + \ldots + \sigma_{P_2}^2}}{n} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}$$

. □

Note 5.5.3

Although the standard deviation was calculated by using a model based on replacement, we use the samples taken without replacement in the future units much more often! Of course, if we have repeated events that allow for samples with replacement, the sampling distribution will also be like the above.

Additionally, the Central Limit Theorem applies to proportions here as well, with larger sample values producing a more and more normal distribution:

Lemma 5.5.4 (Normality of the Distribution of Sample Proportions)

When considering a distribution of uniformly random samples, it will be approximately normal if $np \geq 10$ and $n(1-p) \geq 10$, where n is the sample size, and p is the true population proportion.
In other words, the expected amount of individuals who satisfy or don't satisfy a certain condition must each be greater than 10.

Putting this all together, we obtain a very nice model to describe sampling distributions on:

Theorem 5.5.5 (Distribution of Sample Proportions)

Let N be the size of the population, p be the true proportion of the population, and n be the sample size. If:

1. the samples are chosen randomly
2. $n < N(10\%)$
3. $np \geq 10$ and $n(1-p) \geq 10$

then the sampling distribution of the sample proportions

1. has a mean of $\mu_{\hat{p}} = p$
2. has a standard deviation of $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$
3. is approximately normal

It is important to remember that the values of the proportions aren't continuous, and are limited to values between 0 and 1, inclusive, and are only approximately normal, unlike the distribution outlined above. The distribution above however, is a good model for us to calculate probabilities with it.
Additionally, note that the sampling distributions will vary less as n gets larger, just as outlined.

Example 5.5.6

65% of people prefer dogs to cats. If a simple random sample of 40 people in the United States is conducted, what is the probability that less than 55% of people prefer dogs to cats?

Solution: First, we need to check conditions to see if we can use a normal distribution and the formulas. A simple random sample was used, so the randomness condition is met. 40 people is less than 10% of the population of the United States, so the independence condition is satisfied. The proportion of .65 is given, and we can check large counts with $40(.65) = 26$ and $40(1-.65) = 14$. Both are greater than 10, so the large counts condition is met. All conditions are met.

We can find the standard deviation with $\sqrt{\frac{.65(1-.65)}{40}} = .075$. Thus, we have a sampling distribution with a mean of .65 and standard deviation of .075. We can then use normalCDF to find the probability. We have normalCDF(lower $= -\infty$, upper $= .55$, $\mu = .65$, $\sigma = .075$) $= .0921$. Thus, there is a 9.21% chance that less than 55% of people prefer dogs to cats in a sample.

Problem 5.5.7 — 25% of people report getting eight or more hours of sleep everyday. The sampling distribution of adults getting eight or more hours of sleep is roughly symmetric and unimodal. In a simple random sample of 40 people, what is the probability that 35% or more people get over eight hours of sleep?

Solution: First, conditions need to be checked. It is a simple random sample, so the randomness condition is met. 40 adults is less than 10% of the total adult population,

so the independence condition is met. The proportion of .25 is given, thus we have $40(.25) = 10$ and $40(.75) = 30$. Both are greater than or equal to 10, so the large counts condition is met. All conditions are met.

The standard deviation is $\sqrt{\frac{.25(1-.25)}{40}} = .068$. Using normalCDF, we have normalCDF(lower $= .35$, upper $= \infty$, $\mu = .25$, $\sigma = .068$) $= .0707$. The probability that 35% or more people get over eight hours of sleep in a simple random sample of adults is $\boxed{.0707}$.

§5.6 Sampling Distributions for Differences in Sample Proportions

Similarly as looking at the sampling distribution for the proportion of one population can help us draw conclusions on the likelihood of specific samples occurring, we can also find what the sampling distribution of the difference on two population's proportions would look like.

To compare one population to another by using samples, we take one random sample from each population, and add their difference to the sampling distribution. The sampling distribution appears as such:

Theorem 5.6.1 (Distribution of Difference of Sample Proportions)

Let N_1 and N_2 be two population sizes, n_1 and n_2 be the sizes of the samples taken, and p_1 and p_2 be the true proportions of each of their populations with a certain characteristic. Then if

1. each sample is chosen randomly from its population and independently of the other population
2. $n_1 < N_1(10\%)$ and $n_2 < N_2(10\%)$
3. $n_1p_1 \geq 10$, $n_1(1-p_1) \geq 10$, and $n_2p_2 \geq 10$, $n_2(1-p_2) \geq 10$

then the sampling distribution of the sample proportions

1. has a mean of $\mu_{\hat{p_1}-\hat{p_2}} = p_1 - p_2$
2. has a standard deviation of $\sigma_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
3. is approximately normal

Proof. The initial conditions imply the populations are each normally distributed, and have means of p_1, p_2, and standard deviations of $\sqrt{\frac{p_1(1-p_1)}{n_1}}$ and $\sqrt{\frac{p_2(1-p_2)}{n_2}}$, respectively. Then, by using the rules on the difference between two random variables and that the difference of two normal variables is also normal, we get the desired. □

Note that like single proportions, this is only an approximate description of the real sampling distribution, not the actual sampling distribution (because of limited range, discontinuity, etc.).

One application of this is finding out the probability of a sample of one population being greater than the other. By using z-scores in a similar fashion to how we used it in

earlier sections, we can find the z-score of 0, and use our calculator to find the resulting probability above/below the specified value. Other calculations using the normalCDF function can also be used.

Problem 5.6.2 — Two different surveys are randomly conducted for two different states regarding how many people prefer online shopping. In the first state, 220 out of 300 respondents prefer online shopping. In the second state, 150 out of 225 respondents prefer online shopping. What is the probability that the difference in proportions for preferring online shopping between the first state and the second state is greater than 7%?

Solution: Both samples are randomly conducted, so the randomness condition is satisfied. We can assume that 300 is less than 10% of State 1's population and 225 is less than 10% of State 2's population. Moreover, since these are two separate states, the samples are independent of each other. The independence condition is satisfied for both states. For state 1, the proportion of respondents who prefer online shopping is $\frac{220}{300} = .733$. Checking large counts, we have $300(.733) = 220 \text{and} 300(1 - .733) = 80$. Both are greater than or equal to 10, so the large counts condition for State 1 is satisfied. For State 2, the proportion is $\frac{150}{225} = .666$. Checking large counts, we have 225(.666) = 150, and 225(1-.666) = 75. Both are greater than or equal to 10, so the large counts condition is met. All conditions are satisfied.

Using the data provided, the mean difference in proportions is $\mu_{\hat{p}_1 - \hat{p}_2} = .733 - .666 = .067$. The standard deviation is $\text{SD}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{.733(1-733)}{300} + \frac{.666(1-.666)}{225}} = .041$. Using normalCDF, we have (lower = .07, upper = ∞, $\mu = .067$, $\sigma = .041$) = $\boxed{.471}$.

§5.7 Sampling Distributions for Sample Means

Sample means and sample proportions come from quantitative and categorical data, respectively. This causes several differences between them; specifically, the values in the population are much more varied in quantitative data as opposed to 0s and 1s in categorical data. This leads to more varied means and standard deviations (and the standard deviation isn't based solely on proportion (and sample size) in sample means). Now we describe the sampling distribution for sample means. It is very similar as in sample proportions, which should make sense, as both of them are represented by groups of real numbers.

Theorem 5.7.1

Let N be the population size, n be the sample size taken, and μ be the real population mean of a specific trait. Then if

1. each sample is chosen randomly
2. $n < N(10\%)$
3. $n \geq 30$ **OR** the population is normally distributed

them the sampling distribution of the sample mean

1. has a mean of $\mu_{\overline{x}} = \mu$
2. has a standard deviation of $\sigma_{\hat{x}} = \dfrac{\sigma}{\sqrt{n}}$
3. is approximately normal

The sample size taken being at least 30 or being approximately normal is due to the Central Limit Theorem as discussed earlier. If the population is already normal, then the distribution has already approached the normal and doesn't need the minimum of 30. Additionally, the distribution described above is still only an approximate representation of the true sampling distribution, but measures it very closely, so we take that distribution when we perform probability calculations.
We don't have an alternative 'normal' condition in the distribution of proportions because the two values of 0 and 1 possible will lead to a bimodal distribution, which is unable to be normal.

Example 5.7.2

A large corporation wants to estimate the average number of hours that employees work every week. 50 employees are randomly sampled, and the sample mean is 42 hours with a standard deviation of 8 hours. What is the probability that a random sample shows less than 40 hours worked a week?

Solution: Since the sample was a random sample, the randomness condition is satisfied. The company is a large corporation, meaning that we can assume that 50 employees is less than 10% of the total employee population, meaning the independence condition is satisfied. 50 employees are sampled, which is greater than 30, so the central limit theorem applies, and the normality condition is met. All conditions are met.

The mean of the sampling distribution is 42. The standard deviation is $\frac{8}{\sqrt{50}} = 1.131$ apples. Using normalCDF, we have normalCDF(lower $= -\infty$, upper $= 40$, $\mu = 42$, $\sigma = 1.131$) $= \boxed{.039}$.

Problem 5.7.3 — An average American eats approximately 25 apples per year with a standard deviation of 5 apples. A random sample of 30 Americans is conducted. What is the probability that the random sample will show an average of less than 23 apples a year?

Solution: The sample is a random sample, therefore, the randomness condition is satisfied. 30 Americans is less than 10% of the total population of Americans, so the

independence condition is satisfied. 30 Americans are sampled, which is greater than or equal to 30, meaning that the central limit theorem applies, and the normality condition is satisfied.

The standard deviation of the sampling distribution is $\frac{5}{\sqrt{30}} = .913$ apples. The mean is 25 apples. Using normalCDF, we have normalCDF(lower $= -\infty$, upper $= 23$, $\mu = 25$, $\sigma = .913$) $= \boxed{.014}$.

§5.8 Sampling Distributions for Differences in Sample Means

The difference between the two means of two populations is calculated by finding the difference between two randomly chosen sample's means from each population, just like in differences between two proportions. The distribution is also very similar:

Theorem 5.8.1 (Distribution of Difference of Sample Means)

Let N_1 and N_2 be two population sizes, n_1 and n_2 be the sizes of the samples taken, μ_1 and μ_2 be the means of the populations, and σ_1 and σ_2 be the standard deviations. Then if

1. each sample is chosen randomly from its population and independently of the other population
2. $n_1 < N_1(10\%)$ and $n_2 < N_2(10\%)$
3. each of the populations are either normally distributed or the sample size of the population is at least 30

then the sampling distribution of the sample means

1. has a mean of $\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$
2. has a standard deviation of $\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
3. is approximately normal

Proof. Almost the exact same reasoning as in the difference of two proportions. Remember that the variances of the two distributions are summed, hence the reason for the squared σs in the standard deviation. □

Just as in differences in two sampling proportions, we can use the distribution to calculate the likelihood of one sample being larger than the other, via looking at the z-score of 0 in the sampling distribution generated by the differences between the samples, and using the calculator to find the probability of a greater or smaller value than 0.

Problem 5.8.2 — A researcher wants to study the difference in average hours spent sleeping between college students who work part-time and those who don't. Two random samples of 60 students each independent of each other were conducted. The sample mean hours of those who work part-time is 6.5 hours with a standard deviation of 1.2 hours. The sample mean hours of sleep for the group that doesn't work part-time is 7.2 hours with a standard deviation of 1.5 hours. What is the probability that the difference in means of sleep hours between students who do not work part-time and students who do work part-time is greater than .3?

Solution: Both samples are random samples, so the randomness condition is satisifed. We can assume that both samples are less than 10% of the total population of college students. The samples are both independent of each other, satisfying the independence condition. Both samples have over 30 students sampled, meaning that the central limit theorem applies, and the normality condition is satisfied. All conditions are met.

The mean of the sampling distribution is $7.2-6.5=.7$. The standard deviation of the sampling distribution is $\sqrt{\frac{1.5^2}{60}+\frac{1.2^2}{60}}=.248$. Using normalCDF, we have normalCDF(lower $=.3$, upper $=\infty$, $\mu=.7$, $\sigma=.248$) $=\boxed{.947}$.

Unit 5 Practice Problems

Problem 5.8.1 (1998 AP Statistics FRQ Problem 1) — Consider the sampling distribution of a sample mean obtained by random sampling from an infinite population. This population has a distribution that is highly skewed toward the larger values.

(a) How is the mean of the sampling distribution related to the mean of the population?

(b) How is the standard deviation of the sampling distribution related to the standard deviation of the population?

(c) How is the shape of the sampling distribution affected by the sample size?

(a): The mean of the sampling distribution is equal to the mean of the population.
(b): The standard deviation of the sampling distribution is equal to the standard deviation of the population divided by the square root of the sample size, $\frac{\sigma}{\sqrt{n}}$, which is decreasing as the sample size n grows larger.
(c): As we increase sample size, the sampling distribution approaches a normally distributed distribution, despite initial skew or other conditions of non-normality because of the Central Limit Theorem. However, at smaller sample sizes, the distribution may not necessarily be approximately normal.

Problem 5.8.2 (2004 AP Statistics FRQ Problem 3, Form B) — Trains carry bauxite ore from a mine in Canada to an aluminum processing plant in northern New York state in hopper cars. Filling equipment is used to load ore into the hopper cars. When functioning properly, the actual weights of ore loaded into each car by the filling equipment at the mine are approximately normally distributed with a mean of 70 tons and a standard deviation of 0.9 ton. If the mean is greater than 70 tons, the loading mechanism is overfilling.
(a) If the filling equipment is functioning properly, what is the probability that the weight of the ore in a randomly selected car will be 70.7 tons or more? Show your work.
(c) If the filling equipment is functioning properly, what is the probability that a random sample of 10 cars will have a mean ore weight of 70.7 tons or more? Show your work.

(a): Let X be the random variable representing the weight of an ore in a randomly selected car. Using the given mean and standard deviation,

$$P(X > 70.7) = P(z > \frac{70.7 - 70}{0.9}) = P(z > 0.78) = \boxed{0.2177}$$

.

(c): The mean of the described sampling distribution is 70, and there are likely more than $10 \cdot 10 = 100$ bauxite trains, so the standard deviation is $\frac{0.9}{\sqrt{10}} = 0.285$. The sampling distribution is also normal because of the population distribution being normal, thus the probability is

$$P(X > 70.7) = P(z > \frac{70.7 - 70}{0.285}) = P(z > 2.46) = \boxed{0.0069}$$

.

Problem 5.8.3 (Reworded 2007 AP Statistics FRQ Problem 2, Form B) — A random variable X represents the total number of dogs and cats owned per household, for the households in a large suburban area.
(c) The mean and standard deviation of X are 1.65 and 1.851, respectively. Suppose 150 households in this area are to be selected at random and X, the mean number of dogs and cats per household, is to be computed. Describe the sampling distribution of X, including its shape, center, and spread.

(c):
As the sample size is $150 > 30$, and samples are chosen randomly, the shape of the sampling distribution distribution is normal. The mean of the distribution is the same of that for X, which is 1.65 dog/cat. Finally, as there are likely more than $150 \cdot 10 = 1500$ households in the large suburban area, the standard deviation is $\dfrac{1.851}{\sqrt{150}} = 0.1151$ dog/cat.

Problem 5.8.4 (2007 AP Statistics FRQ Problem 3, Form A) — Big Town Fisheries recently stocked a new lake in a city park with 2,000 fish of various sizes. The distribution of the lengths of these fish is approximately normal.

(a) Big Town Fisheries claims that the mean length of the fish is 8 inches. If the claim is true, which of the following would be more likely?

- A random sample of 15 fish having a mean length that is greater than 10 inches

or

- A random sample of 50 fish having a mean length that is greater than 10 inches

Justify your answer.

(b) Suppose the standard deviation of the sampling distribution of the sample mean for random samples of size 50 is 0.3 inch. If the mean length of the fish is 8 inches, use the normal distribution to compute the probability that a random sample of 50 fish will have a mean length less than 7.5 inches.

(c) Suppose the distribution of fish lengths in this lake was non-normal but had the same mean and standard deviation. Would it still be appropriate to use the normal distribution to compute the probability in part (b)? Justify your answer.

(a): The sampling distribution for sample size $n = 15$ will have more variability than that of sample size $n = 50$ because the standard deviation of the former distribution $\sigma/\sqrt{15} > \sigma/\sqrt{50}$, the distribution of the other.
Furthermore, both distributions are normal with a mean of 8. Therefore, the z-score of 10 in the distribution will be $\dfrac{10-8}{\sigma_{\overline{x}}} = \dfrac{2}{\sigma_{\overline{x}}}$, which will taken on larger and larger positive values as $\sigma_{\overline{x}}$, the standard deviation of the sampling distribution, decreases. Thus, the area of $\overline{x} > 10$ on the tail will be larger for the former distribution with smaller sampling distribution standard deviation, and so the first is more likely to occur.
(b): Using the distribution outlined in the problem statement,
$P(\overline{x} < 7.5) = P(z < \dfrac{7.5-8}{0.3}) = P(x < -1.67) = \boxed{0.0475}$.
(c): As the sample size taken is $50 > 30$, even if the distribution is non-normal, by the Central Limit Theorem, the sampling distribution would still be normal, and would have the same mean and standard deviation because they are kept constant in the population.

Problem 5.8.5 (2008 AP Statistics FRQ Problem 2, Form B) — Four different statistics have been proposed as estimators of a population parameter. To investigate the behavior of these estimators, 500 random samples are selected from a known population and each statistic is calculated for each sample. The true value of the population parameter is 75. The graphs below show the distribution of values for each statistic. [insert data]

(a) Which of the statistics appear to be unbiased estimators of the population parameter? How can you tell?

(b) Which of statistics A or B would be a better estimator of the population parameter? Explain your choice.

(c) Which of statistics C or D would be a better estimator of the population parameter? Explain your choice.

(a): The unbiased estimators are A, C, and D, as they are all approximately symmetrically distributed around the true value of 75 and seem to average out there. We can specifically note the close to equal amount of bins having values to the left and right of the mean in each distribution (5 to 6 for A, 3 to 3 for C, and 7 to 8 for D). On the other hand, B has most values to the right of 75 and won't average out there, instead overestimating the true parameter of the population.
(b): A is, because unbiased estimators are always better than biased ones, and A seems less biased than B is.
(c): C is, because although both are unbiased estimators, C has less variability compared to D, due to the smaller range of values C covers as well as the values being more dense around the mean with taller boxes, making it better to use.

Problem 5.8.6 (2008 AP Statistics FRQ Problem 5, Form B) — Flooding has washed out one of the tracks of the Snake Gulch Railroad. The railroad has two parallel tracks from Bullsnake to Copperhead, but only one usable track from Copperhead to Diamondback, as shown in the figure below. Having only one usable track disrupts the usual schedule. Until it is repaired, the washed-out track will remain unusable. If the train leaving Bullsnake arrives at Copperhead first, it has to wait until the train leaving Diamondback arrives at Copperhead.

Every day at noon a train leaves Bullsnake heading for Diamondback and another leaves Diamondback heading for Bullsnake.

Assume that the length of time, X, it takes the train leaving Bullsnake to get to Copperhead is normally distributed with a mean of 170 minutes and a standard deviation of 20 minutes.

Assume that the length of time, Y, it takes the train leaving Diamondback to get to Copperhead is normally distributed with a mean of 200 minutes and a standard deviation of 10 minutes.

These two travel times are independent.

(a) What is the distribution of $Y - X$?

(b) Over the long run, what proportion of the days will the train from Bullsnake have to wait at Copperhead for the train from Diamondback to arrive?

(c) How long should the Snake Gulch Railroad delay the departure of the train from Bullsnake so that the probability that it has to wait is only 0.01 ?

(a): As X and Y are normally distributed, $Y - X$ will also be normally distributed. The distribution also has mean $\mu_{Y-X} = \mu_Y - \mu_X = 300 - 270 = 30$ minutes, and standard deviation $\sigma_{Y-X} = \sqrt{\sigma_Y^2 + \sigma_X^2} = \sqrt{20^2 + 10^2} = \sqrt{500} = 22.36$ minutes.
(b): Over time, as noted in Unit 4, this proportion approaches the true probability that the Bullsnake train will take less time than the Copperhead train, or equivalently, that $Y - X > 0$.
Using the normal distribution found in part (a), the probability is equal to $P(Y - X > 0) = P(z > \dfrac{0 - 30}{22.36}) = P(z > -1.34) = \boxed{0.9099}$.
(c): When we delay the Bullsnake Train by D minutes, the standard deviation remains unaffected, but as the mean of $X + D$ is $30 + D$, then the mean of the new difference is $Y - (X + D) = Y - X - D = 30 - D$.
The event of the railroad waiting for the Bullsnake train is still $P(Y - (X + D) > 0)$, which needs to be 0.01. To do this, the z-score of a 0 must be around 2.33 (by using the `invNorm` function with solving for $P(z > z_0) = 0.01$, and thus $\dfrac{0 - (30 - D)}{22.36} = 2.33$, so $D = \boxed{82.099}$ minutes.

Problem 5.8.7 (2010 AP Statistics FRQ Problem 2, Form A) — A local radio station plays 40 rock-and-roll songs during each 4-hour show. The program director at the station needs to know the total amount of airtime for the 40 songs so that time can also be programmed during the show for news and advertisements. The distribution of the lengths of rock-and-roll songs, in minutes, is roughly symmetric with a mean length of 3.9 minutes and a standard deviation of 1.1 minutes.

(a) Describe the sampling distribution of the sample mean song lengths for random samples of 40 rock-and-roll songs.

(b) If the program manager schedules 80 minutes of news and advertisements for the 4-hour (240-minute) show, only 160 minutes are available for music. Approximately what is the probability that the total amount of time needed to play 40 randomly selected rock-and-roll songs exceeds the available airtime?

(a): Let X be the random variable representing the average time of a randomly chosen sample of 40 songs. The sampling distribution of the sample mean song length has mean $\mu_{\overline{X}} = \mu = 3.9$ minutes and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{1.1}{\sqrt{40}} = 0.174$ minutes. The Central Limit Theorem , as $n = 40 > 30$, allows the sampling distribution to be approximately normally distributed due to its sufficiently large sample size.

(b) The probability that the total airtime of 40 randomly selected songs exceeds the available time is equivalent to the probability that the sample mean length of the 40 songs is greater than $\dfrac{160}{40} = 4$ minutes.

According to part (a), the distribution of the sample mean length $\overline{X}$ is approximately normal. Therefore, $P(\overline{X} > 0) = P(z > \dfrac{4 - 3.9}{0.174}) = P(z > 0.57) = \boxed{0.2843}$.

Problem 5.8.8 (2012 AP Statistics FRQ Problem 6) — Two students at a large high school, Peter and Rania, wanted to estimate m, the mean number of soft drinks that a student at their school consumes in a week. A complete roster of the names and genders for the 2,000 students at their school was available. Peter selected a simple random sample of 100 students. Rania, knowing that 60 percent of the students at the school are female, selected a simple random sample of 60 females and an independent simple random sample of 40 males. Both asked all of the students in their samples how many soft drinks they typically consume in a week.
Peter used the sample mean $\overline{X}$ as a point estimator for μ. Rania used $\overline{X}_{\text{overall}} = (0.6)\overline{X}_{\text{female}} + (0.4)\overline{X}_{\text{male}}$ a point estimator for μ, where $\overline{X}_{\text{female}}$ is the mean of the sample of 60 females and $\overline{X}_{\text{male}}$ is the mean of the sample of 40 males. Summary statistics for Peter's data are shown in the table below.
table

(b) Based on the summary statistics, calculate the estimated standard deviation of the sampling distribution (sometimes called the standard error) of Peter's point estimator $\overline{X}$.

Summary statistics for Raina's data are shown in the table below.
(c) Based on the summary statistics, calculate the estimated standard deviation of the sampling distribution of Raina's point estimate $\overline{X}_{\text{overall}} = (0.6)\overline{X}_{\text{female}} + (0.4)\overline{X}_{\text{male}}$.

(b): The estimated standard deviation of the sampling distribution will be based on the estimated standard deviation of the sample taken, which is $s = 4.13$ drinks.
To get the sampling distribution's standard deviation, we divide by $\sqrt{n}$ like usual, and get the estimate as $\dfrac{s}{\sqrt{n}} = \dfrac{4.13}{\sqrt{100}} = 0.413$ drinks.
(c): Using the formula for the standard deviation of $aX+bY$ being $\sqrt{a^2\text{Var}(X) + b^2\text{Var}(Y)}$, and using the estimated standard deviation values similarly in part (b), we compute the standard deviation to be estimated by

$$\sqrt{(0.6)^2 \cdot \frac{s_{\text{f}}^2}{n_{\text{f}}} + (0.4)^2 \cdot \frac{s_{\text{m}}^2}{n_{\text{m}}}} = \sqrt{(0.6)^2 \cdot \frac{1.8^2}{60} + (0.4)^2 \cdot \frac{2.22^2}{40}} = \sqrt{0.03916} = \boxed{0.198}.$$

Problem 5.8.9 (2015 AP Statistics FRQ Problem 6) — Corn tortillas are made at a large facility that produces 100,000 tortillas per day on each of its two production lines. The distribution of the diameters of the tortillas produced on production line A is approximately normal with mean 5.9 inches, and the distribution of the diameters of the tortillas produced on production line B is approximately normal with mean 6.1 inches. The figure below shows the distributions of diameters for the two production lines.
The tortillas produced at the factory are advertised as having a diameter of 6 inches. For the purpose of quality control, a sample of 200 tortillas is selected and the diameters are measured. From the sample of 200 tortillas, the manager of the facility wants to estimate the mean diameter, in inches, of the 200,000 tortillas produced on a given day. Two sampling methods have been proposed.

Method 1: Take a random sample of 200 tortillas from the 200,000 tortillas produced on a given day. Measure the diameter of each selected tortilla.

Method 2: Randomly select one of the two production lines on a given day. Take a random sample of 200 tortillas from the 100,000 tortillas produced by the selected production line. Measure the diameter of each selected tortilla.

(a) Will a sample obtained using Method 2 be representative of the population of all tortillas made that day, with respect to the diameters of the tortillas? Explain why or why not.

(b) The figure below is a histogram of 200 diameters obtained by using one of the two sampling methods described. Considering the shape of the histogram, explain which method, Method 1 or Method 2, was most likely used to obtain a such a sample.

(c) Which of the two sampling methods, Method 1 or Method 2, will result in less variability in the diameters of the 200 tortillas in the sample on a given day? Explain.

Each day, the distribution of the 200,000 tortillas made that day has mean diameter 6 inches with standard deviation 0.11 inches.

(d) For samples of size 200 taken from one day's production, describe the sampling distribution of the sample mean diameter for samples that are obtained using Method 1.

(e) Suppose that one of the two sampling methods will be selected and used every day for one year (365 days). The sample mean of the 200 diameters will be recorded each day. Which of the two methods will result in less variability in the distribution of the 365 sample means? Explain.

(f) A government inspector will visit the facility on June 22 to observe the sampling and to determine if the factory is in compliance with the advertised mean diameter of 6 inches. The manager knows that, with both sampling methods, the sample mean is an unbiased estimator of the population mean. However, the manager is unsure which method is more likely to produce a sample mean that is close to 6 inches on the day of sampling. Based on your previous answers, which of the two sampling methods, Method 1 or Method 2, is more likely to produce a sample mean close to 6 inches? Explain.

(a): No, a sample obtained using Method 2 will not be representative of all tortillas made that day. The sample obtained using Method 2 will only represent the tortillas from one production line, not from the entire population because the distributions of diameters for the two production lines have very different means and aren't similar to each other there. One way they would differ is that Method 2's samples would all be closer to 6.1 rather than 6, the mean of all of the tortilla diameters, and therefore wouldn't represent all tortillas.
(b): Method 2 was most likely used for this sample because the bimodal shape has each of its peaks roughly equal in height and corresponding to the means of each production line, indicating that samples were selected solely from one production line. Method 1, by the Central Limit Theorem, would produce a more unimodal and closer to normal distribution.
(c): Method 1 would result in less variability in the sample of 200 tortillas on a given day because the sample comes from only one production line. Because the distributions of diameters are not the same for the two production lines, selecting tortillas from both lines as in Method 1 would result in more variable sample data centered around two largely different peaks rather than one.
(d): The sampling distribution of the sample mean diameter for samples obtained using Method 1 would be approximately normal (as $200 > 30$) with mean $\mu_{\overline{x}} = \mu = 6$ inches and standard deviation $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.11}{\sqrt{200}} = 0.0078$ inch.
(e): Method 1 would result in less variability in the sample means over the 365 days, because with Method 2, roughly half of the sample means will be around 5.9 inches and the other half will be around 6.1 inches, all around 0.1 inches from the expected mean of 6 inches. With Method 1, however, the sample means will all be very close to 6 inches, as indicated by the small standard deviation and normality and therefore have less variability.
(f): Both methods, on average, give a mean of 6 inches as in either method, the group of 5.9 inches and 6.1 inches are expected to be represented equally. Thus, both estimators are unbiased and produce a mean of 6 inches on average. However, Method 1 has a smaller variability, making it be much more likely to be closer to 6 inches than Method 2 did.

Problem 5.8.10 (2023 AP Statistics FRQ Problem 6) — A jewelry company uses a machine to apply a coating of gold on a certain style of necklace. The amount of gold applied to a necklace is approximately normally distributed. When the machine is working properly, the amount of gold applied to a necklace has a mean of 300 milligrams (mg) and standard deviation of 5 mg.

(a) A necklace is randomly selected by the necklaces produced by the machine. Assuming that the machine is working properly, calculate the probability that the amount of gold applied to the necklace is between 296 mg and 304 mg.

Cleo, a statistician at the jewelry company, will take a random sample of the necklaces produced that day. Each selected necklace will be melted down and the amount of gold applied to that necklace will be determined. Because a necklace must be destroyed to determine the amount of gold that was applied, Cleo will use random samples of size $n = 2$ necklaces.

Cleo starts by considering the mean amount of gold applied to the necklaces. After Cleo takes a random sample of $n = 2$ necklaces, she computes the sample mean amount of gold applied to the two necklaces.
(b) Suppose the machine is working properly with a population mean amount of gold being applied of 300 mg and a population standard deviation of 5 mg.
(i) Calculate the probability that the sample mean amount of gold applied to a random sample of $n = 2$ necklaces will be greater than 303 mg.

(a): Let X represent the amount of gold applied to a necklace randomly selected from necklaces produced with this machine. The random variable X has an approximately normal distribution with mean 300 mg and standard deviation 5 mg.
Therefore, $P(296 < \overline{X} < 304) = P(\frac{296 - 300}{5} < z < \frac{304 - 300}{5}) = \boxed{0.5763}$.
(b): The original population is normal, has mean 300 mg, and standard deviation 5 mg, thus the sampling distribution is normal, has mean 300 mg, and standard deviation $\frac{5}{\sqrt{2}} = 3.5355$ mg, thus $P(\overline{X} > 303) = P(z > 0.8485) = \boxed{0.198}$.

6 Unit 6: Inference for Categorical Data: Proportions

§6.1 Introducing Statistics: Why Be Normal?

Recall from chapter 5 the conditions for normality in a sampling distribution of proportions. A sampling distribution is approximately normal when the sample satisfies the Large Counts condition: $np > 10$ and $n(1-p) > 10$, where n is the sample size for the study, and p is the population proportion for the statistic. For instance if you knew that the proportion of people who liked vanilla ice cream was 0.18, and you took a random sample of 57 people then the sampling distribution of the proportion would be approximately normal because $(57)(0.18) = 10.26 > 10$ and $(57)(0.82) = 46.74 > 10$.

While the sampling size n of a study is almost always known, the population proportion for the study is not. In this case the following theorem arises:

Theorem 6.1.1 (Large Counts condition when p is unknown but $\hat{p}$ is.)
For a sampling distribution of size n and sample proportion $\hat{p}$, the sampling distribution is approximately normal when both the following are true:

- $n\hat{p} > 10$
- $n(1-\hat{p}) > 10$

It is important to note that this theorem is still called the Large Counts condition, it is just modified to be as accurate as possible for the scenario because $\hat{p}$ is our best estimator for p. This best estimator is also sometimes called the *point estimate*.

Problem 6.1.2 (Source: Original) — Your wife Susanne is investigating the large number of recent paranormal activity reported in your hometown of Boston, Massachusetts. She conducts a random sample of 42 houses, and finds that the number of houses that have recently reported experiencing some form of paranormal activity is 17/42. She is trying to create a sampling distribution of the data, and is wondering if the sampling distribution for her study is approximately normal. Which of the following is correct?

(A) The sampling distribution is approximately normal because $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$ is true.

(B) The sampling distribution is not approximately normal because $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$ is not true.

(C) The sampling distribution is approximately normal because she sampled more than 30 houses satisfying the Central Limit Theorem.

(D) The sampling distribution is not approximately normal because she did not sample more than 30 hours failing the Central Limit Theorem.

(E) Not enough information is provided to answer Susanne's question.

Solution

(A) This answer is correct because $42(17/42) = 17 > 10$ and $42(1 - 17/42) = 25 > 10$. Since the number of "successes" and "failures" is both larger than 10, the Normal condition is satisfied.

(B) This answer is not correct because the Large Counts condition is met as shown in answer choice (A).

(C) This answer is not correct because the Central Limit Theorem only applies for means, not proportions.

(D) This answer is not correct because the Central Limit Theorem only applies for means, not proportions.

(E) This answer is not correct because the sample size of 42 and sample proportion of 17/42 give just enough information to answer Susanne's question.

Note 6.1.3 (Conditions for inference for a population proportion)

ALWAYS check that these conditions are met before conducting any sort of inference procedure.

- Random - Individual observations are selected randomly from the population of interest
- Normal - There must be at least 10 successes and 10 failures.
- Independent - Individual observations are independent of each other. This either means that sampling is done with replacement, or that the sample size is less than 10% the size of the population.

§6.2 Constructing a Confidence Interval for a Population Proportion

The purpose of the aforementioned new Large Counts condition was to determine if the sampling distribution is approximately normal based on the $\hat{p}$ value. With this $\hat{p}$ value and the sample size n, and a constructed sampling distribution for $\hat{p}$, it becomes possible to estimate a window, or interval of values for which the true population proportion is inside of based on the normal distribution formed from the sampling distribution. This interval is called a *confidence interval*, and the purpose of it is to estimate a window that the population is in x percent of the time. For instance if we are 99 percent confident, that means that if many, many intervals were constructed using the same methodology, then 99 percent of these confidence intervals would contain the true population proportion. This level of confidence is called the *confidence level*.

A confidence interval, is one of many *inference procedures* that we will be exploring within AP Statistics. All inference procedures contain the following four steps: *state*, *plan*, *do* and *conclude*. The explanation for each step is shown below:

State: State the confidence level and the population proportion to be estimated.

Plan: State the type of inference procedure to be conducted (in this section the type is a 1-sample z-interval for proportions). Verify that the appropriate conditions for inference for the procedure are met

Do: Construct the confidence interval using the Confidence Interval Formula

Theorem 6.2.1 (Confidence Interval Formula)

The formula for a 1-sample z-interval for proportions is:

$$\text{Confidence Interval: } \hat{p} \pm (z^*)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where $\hat{p}$ is the proportion found from the sample, z* is the *critical value*, and n is the sample size.

Conclude: Interpret the confidence level.

Problem 6.2.2 (Source: Original) — After a long Halloween, Daniel attempts to find the true proportion of chocolate candy in his bag. After counting for several minutes, Daniel realizes that his bag is too large to determine the number of chocolate located inside of the bag. Recalling knowledge from his statistics class, Daniel takes a random sample of 32 candies from his bag and notices that 12 of them are chocolate.

Construct and interpret a 90 percent confidence interval for the true proportion of chocolate in Daniel's bag.

Solution

State: We will construct a 95% confidence interval for the true proportion of candy p in Daniel's bag.

Plan: We will conduct a 1-sample z-test for proportions.
✓Random: The candy was randomly selected from Daniel's bag.
✓Normal: $(32)(12/32) = 12 > 10$ and $(32)(1-12/32) = 20 > 10$. Therefore the sampling distribution is approximately normal.
✓Independent: We may reasonably assume that there are more than 320 candies in Daniel's bag.

Do:

$$\text{Confidence Interval: } \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \ \hat{p} = \frac{12}{32} = 0.375$$

$$\text{Confidence Interval: } 0.375 \pm 1.645 \sqrt{\frac{0.375(1-0.375)}{32}} = 0.375 \pm 0.141$$

$$\text{Confidence Interval: } (0.234, \ 0.516)$$

Conclude: We are 90% confident that the interval $(0.234, 0.516)$ contains the true proportion of chocolate in Daniel's bag.

The final part of this section will be used to explain a two key vocabulary terms. The **margin of error** of a confidence interval is half of the total length of the confidence interval. For the previous example the margin of error would be $(0.2349 - 0.5151)/2 = 0.140$. The **point estimate** is the center of the interval and is given as $(0.2349 + 0.5151)/2 = 0.375$.

Problem 6.2.3 (Source: Original) — In physics the gravitational constant g represents the rate at which an object accelerates towards the ground on Earth. Due to physical uncertainty and changes in the Earth's radius, the gravitational constant g does not have an exact value but actually changes ever ever-so-slightly from location to location. Your research team experimentally determines with 95% confidence that the value of g within your area falls between $(9.802, 9.8142) m/s^2$. Which of the following give the confidence level, point estimate, and margin of error for g?
(A) Confidence Level: 95%, Point Estimate: 9.8081, Margin of Error: 0.0061
(B) Confidence Level: 90%, Point Estimate: 9.8081, Margin of Error: 0.0061
(C) Confidence Level: 95%, Point Estimate: 0.0061, Margin of Error: 9.8081
(D) Confidence Level: 90%, Point Estimate: 0.0061, Margin of Error: 9.8081
(E) None of the above.

Solution

Confidence Level: The passage describes a 95% interval, so the confidence level is 95%.

Point Estimate: The point estimate is the average of the maximum of the minimum of the interval, or $\frac{9.802+9.8142}{2} = 9.8081$.

Margin of Error: The margin of error is one-half the length of the confidence interval. This gives $\frac{9.802-9.8142}{2} = 0.0061$.

The only option with all three of these items being correct is **(B)**.

§6.3 Justifying a Claim Based on a Confidence Interval for a Population Proportion

A confidence interval for a population proportion either contains the population proportion or it doesn't, because each interval is constructed from random sample data, which varies from sample to sample. That is why we cannot say for certainty whether a sample will have the population percentage, so we express it as a percentage as follows:

We are C% confident that the confidence interval for a population proportion captures the population proportion.

Another way to understand this concept is that if we repeat random sampling with the same sample size, approximately C% of the confidence intervals created will capture the population proportion. Whenever we interpret a confidence interval for a one-sample proportion, we need to include a reference to the sample taken and details about the population it represents.

Problem 6.3.1 (Source: Original) — Interpret a 95% confidence interval of $(0.258, 0.272)$ based on the proportion of a nationally representative sample of eleventh-grade students who answered a particular multiple-choice question correctly.

Solution
We are 95 percent confident that the interval from 0.258 and 0.272 contains the population proportion of all United States eleventh-grade students who would answer that question correctly.

Another important concept is that when all other things remain the same, the width of the confidence interval for a population proportion decreases as sample size increases. This is because the width of the interval is proportional to $\frac{1}{\sqrt{n}}$ (to see why this is true, look at the formula again!). Additionally, the width of the confidence interval for a population proportion increases as confidence level increases (this makes intuitive sense, as to be more certain you need a greater margin of error).

§6.4 Setting Up a Test for a Population Proportion

Before we learn how to set up a test for a population proportion, we first need to learn about the types of tests that we can conduct. There are only two: one-sided and two-sided

Definition 6.4.1 (One-sided and Two-sided hypothesis test)

A one-sided test always has an alternative hypothesis that claims that the true parameter is either strictly greater or less than the hypothesized parameter.
(ex. $H_0 : p = 3$, $H_a : p > 3$)

A two-sided test always has an alternative hypothesis that claims that the true parameter is different (either greater OR less than) from the assumed population mean.
(ex. $H_0 : p = 3$, $H_a : p \neq 3$)

In a one-sided test, we reject the null hypothesis if and only if our test statistic is a certain number of standard error's more **OR** less than the observed statistic. An example is shown below: In a two-sided test, we reject the null hypothesis if and only if our test

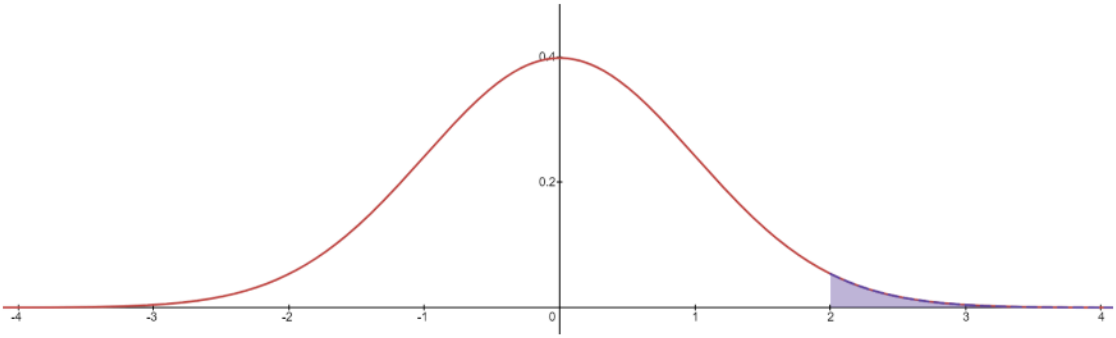


Figure 6.1: Sampling Distribution for a one-sided test

statistic is a certain number of standard error's **AWAY** (either above of below) from the observed mean. An example is shown below:

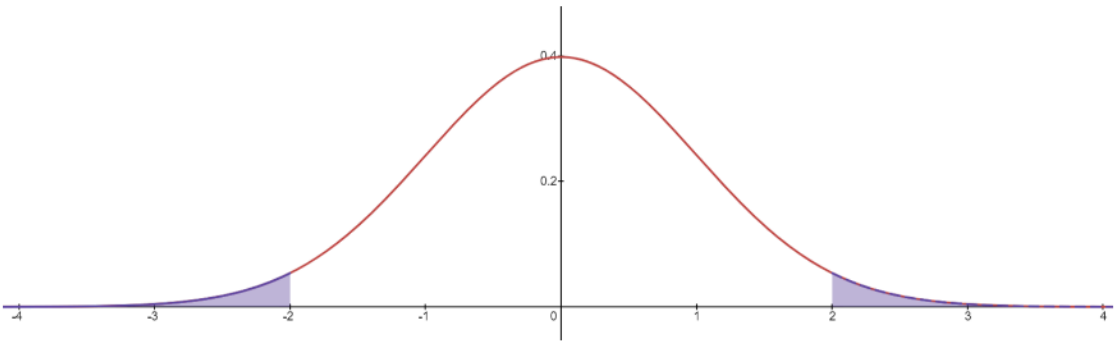


Figure 6.2: Sampling Distribution for a two-sided test

Note 6.4.2 (Steps for Setting up a Test for a Population proportion)

Follow these steps:

1. Setting up the Hypotheses
   - Name the test being used
   - Build the null and alternative hypotheses for the parameter of interest

   Note 6.4.3 (Setting up hypotheses)

   Our null and alternative hypotheses should look something like this:
   $H_0 : p = 0.7$
   $H_a : p < 0.7$
   The null and alternative hypotheses should look **EXACTLY** the same, except for the comparison operator. In other words, the number and symbol in the hypotheses should be the same.

2. Checking the conditions for inference for a population proportion

Problem 6.4.4 (Source: 2005 AP Statistics FRQ Problem 4) — Some boxes of a certain brand of breakfast cereal include a voucher for a free video rental inside the box. The company that makes the cereal claims that a voucher can be found in 20 percent of the boxes. However, based on their experiences eating this cereal at home, a group of students believes that the proportion of boxes with vouchers is less than 0.2. This group of students purchased 65 boxes of the cereal to investigate the company's claim. The students found a total of 11 vouchers for free video rentals in the 65 boxes.

Suppose it is reasonable to assume that the 65 boxes purchased by the students are a random sample of all boxes of this cereal. Based on this sample, is there support for the students' belief that the proportion of boxes with vouchers is less than 0.2? Provide statistical evidence to support your answer.

Solution

State:
$H_0 : p = 0.2$
$H_0 : p < 0.2$

Plan: We will conduct a one-sample z-test for a population proportion
✓Random - The problem states that the boxes purchased by the students are a random sample of all the boxes of this cereal.
✓Large Counts - The number of successes and failures within the sample are both greater than 10 (11 successes, $65 - 11 = 54$ failures).
✓Independent - We may assume that there are at least 650 boxes of cereal produced by this brand.

§6.5 Interpreting p-values

A p-value is a key concept in hypothesis testing. It is a measure of the probability of obtaining a sample statistic at least as extreme as the one observed, assuming the null hypothesis is true. A p-value takes on any value from the interval $(0, 1)$, with smaller p-values indicating a smaller chance of obtaining a certain test statistic given that the null hypothesis is true. The purpose of a p-value is to help us question whether or not the null hypothesis is valid, as if the p-value is too small, we may end up rejecting it. p-values work hand in hand with something known as a significance level, denoted as α.

Note 6.5.1 (Significance levels)

The significance level, α, is a threshold chosen before the study, to help us determine whether or not to reject our null-hypothesis. If our p-value is less than our pre-determined significance level, we conclude that the chances of obtaining a sample statistic as extreme as ours assuming the null hypothesis is true is too small, and thus we must reject our original hypothesis. If our p-value is greater than our significance level, we are unable to reject the null hypothesis, and cannot make any further conclusions. Most of the time, you will see significance levels of $\alpha = 0.01\%,\ 0.05\%,\ 0.1\%$ with $\alpha = 0.05\%$ being the most common.

To help us understand the concept of a p-value better, let's look at an arbitrary sampling distribution for sample proportions.

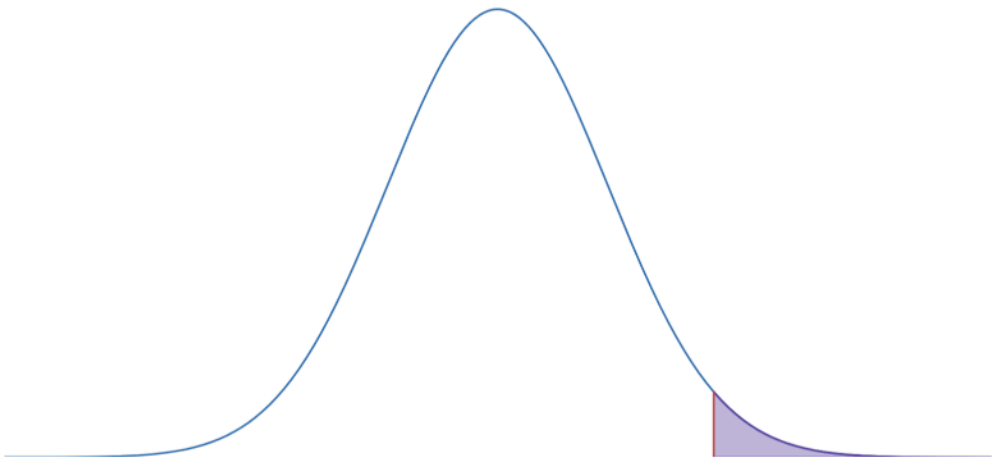


Figure 6.3: Sampling distribution with p-value

Note 6.5.2 (Critical region)

The purple region in the graph is known as the *critical* or *rejection* region, since it represents the set of values of the test-statistic for which the null hypothesis is rejected

§6.6 Concluding a Test for a Population Proportion

Now that we have set up our hypothesis test, it's time to find our test statistic and use it to calculate the associated p-value. Recall that the test statistic measures how far our observed sample proportion is from the hypothesized population proportion, in terms of standard errors. In this case, since we are dealing with proportions, we use a z-statistic rather than a t-statistic.

Theorem 6.6.1 (Test Statistic for Population Proportions)

The formula for the test statistic for population proportions is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where $\hat{p}$ is the sample proportion, p_0 is the hypothesized population proportion, and n is the sample size.

Here, the numerator represents the difference between the observed sample proportion and the hypothesized proportion, while the denominator is the standard error of the sample proportion under the null hypothesis. If you recall from unit 5, this formula should resemble that of the formula for a z-score, since that is essentially what we are calculating.

Note 6.6.2 (Steps for Carrying Out a Test for a Population Proportion)

Follow these steps:

1. Obtain the corresponding test statistic by using the formula above.
2. Use a statistical calculator or table to obtain the associated p-value from the test statistic. This tells us the likelihood of obtaining a sample proportion further from the hypothesized proportion than the one we obtained.
3. Draw a conclusion based on the p-value. If it is lower than the significance level, you may reject the null hypothesis. Otherwise, you may not reject the null hypothesis.

Finally, we continue with the example problem from 2 sections ago.

Problem 6.6.3 (Source: 2005 AP Statistics FRQ Problem 4) — Some boxes of a certain brand of breakfast cereal include a voucher for a free video rental inside the box. The company that makes the cereal claims that a voucher can be found in 20 percent of the boxes. However, based on their experiences eating this cereal at home, a group of students believes that the proportion of boxes with vouchers is less than 0.2. This group of students purchased 65 boxes of the cereal to investigate the company's claim. The students found a total of 11 vouchers for free video rentals in the 65 boxes.

Suppose it is reasonable to assume that the 65 boxes purchased by the students are a random sample of all boxes of this cereal. Based on this sample, is there support for the students' belief that the proportion of boxes with vouchers is less than 0.2? Provide statistical evidence to support your answer.

Do:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},\ \hat{p} = \frac{11}{65} = 0.1692$$

$$z = \frac{0.1692 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{65}}} = \frac{-0.0308}{\sqrt{0.00246}} = -0.6210$$

Using our calculator, we obtain a corresponding p-value of about 0.2673

Conclude: Since our P-value of 0.2673 is relatively high, we fail to reject the null hypothesis. There is no statistically significant evidence proving that the proportion of boxes within vouchers is less than 0.2.

§6.7 Potential Errors When Performing Tests

In statistics, we must always consider the possibility of making an error. That is when we make an incorrect conclusion about our parameter based on our inference test. Since we make our conclusions based on p-values, there is always a chance that our conclusion is wrong, and thus must understand the types of errors that might arise. In this section, we will discuss the two main types of errors that can occur in hypothesis testing: **Type I error** and **Type II error**.

Definition 6.7.1 (Types of errors)

A **Type I error** occurs when we reject the null hypothesis (H_0) even though it is true. In other words, we conclude that there is a significant effect or difference when, in reality, there is none. The probability of making a Type I error is denoted by α, which is also the significance level of the test!

A **Type II error** occurs when we fail to reject the null hypothesis (H_0) even though it is false. In this case, we miss a real effect or difference that actually exists. The probability of making a Type II error is denoted by β.

Definition 6.7.2 (Power)

The **power** of a test is the probability that the test correctly rejects a false null hypothesis. In other words, power is the probability of avoiding a Type II error. Mathematically, it is expressed as $1 - \beta$.

For the AP exam, you will not need to know how to calculate power, however you will need to understand how to interpret it in context with the problem. You must also understand the factors that may increase or decrease the probabilities of committing these errors. Since we already know that the probability of committing a type 1 error is α, we will focus on power. The diagram above illustrates a distribution with a false null

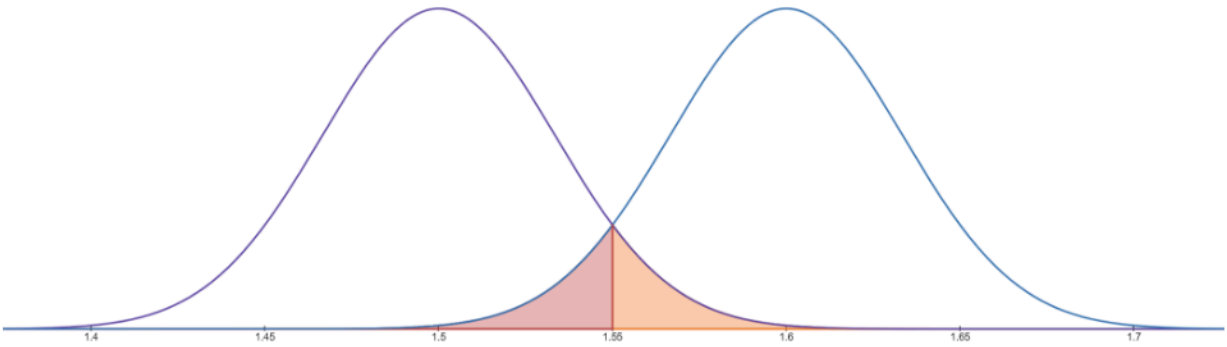


Figure 6.4: Distribution with false null hypothesis

hypothesis of $\mu_0 = 1.5$ and a true population mean of $\mu = 1.6$. We already know that the orange region represents the probability of making a type 1 error. The probability of making a type 2 error is in fact simply the red region. To see why this is true, note that a type two error occurs when we should reject the null hypothesis, however, we don't. When our test statistic lies in the red region, we don't reject the null hypothesis since it is not above our significance level. However, technically, since the red region lies in the true distribution (different from our hypothesized one), it would be correct to reject the null hypothesis. Thus, anytime our test statistic lies in the red region, a type 2 error is caused.

Note 6.7.3 (Factors affecting Power)

On the AP exam, you will need to know how to increase/decrease the power of a significance test.

1. **Sample size** (n): Increasing the sample size reduces the standard error of the estimate, making it easier to detect small differences between the null hypothesis and the true population parameter. Visually, increasing the sample size makes the sampling distribution more narrow, which reduces the red region in Figure 6.4.
2. **Significance level** (α): A higher significance level means we have a lower threshold for rejecting H_0. Increasing α reduces the probability of making a Type II error but increases the likelihood of making a Type I error (To see why this is true, look at figure 6.4 and see what happens when the significance level changes).
3. **Difference between true and hypothesized parameter**: The larger the difference between the true and hypothesized parameters, the easier it is to detect this difference, thus increasing the power of the test. In figure 6.4, if the sampling distributions were further apart, the red region would be much smaller.
4. **Variability in the data**: Higher variability (i.e., larger standard deviation) makes it harder to detect differences between the null hypothesis and the true population mean. Reducing variability increases power by making the sampling distribution more narrow, decreasing β.

Problem 6.7.4 (Source: Original) — A researcher conducts a hypothesis test at the $\alpha = 0.05$ significance level and rejects the null hypothesis. Which of the following is true?

- (A) There is a 95% chance that the null hypothesis is true.
- (B) The probability of making a Type I error is 0.05.
- (C) The probability of making a Type II error is 0.05.
- (D) The probability of correctly rejecting the null hypothesis is 95%.

Solution

The correct answer is (B). The significance level α represents the probability of making a Type I error, which is rejecting a true null hypothesis.

Problem 6.7.5 (Source: Original) — Which of the following would increase the power of a hypothesis test?

- (A) Decreasing the sample size.
- (B) Increasing the variability in the population.
- (C) Decreasing the significance level.
- (D) Increasing the effect size.

Solution
The correct answer is (D). Increasing the effect size makes it easier to detect differences, thus increasing the power of the test.

Problem 6.7.6 — If the sample size is increased while keeping the significance level constant, how is the probability of making a Type II error affected?

- (A) It increases.
- (B) It decreases.
- (C) It stays the same.
- (D) It becomes equal to the significance level.

Solution
The correct answer is (B). Increasing the sample size decreases the probability of making a Type II error by increasing the test's power.

§6.8 Confidence Intervals for the Difference of Two Proportions

When comparing two independent populations, constructing a confidence interval for the difference in their proportions allows us to estimate the true difference between these populations. The confidence interval provides a range of plausible values for the difference between the population proportions. If the interval includes zero, it suggests no significant difference between the two population proportions. On the other hand, if the interval does not include zero, we can infer that a difference does exist.

Theorem 6.8.1 (Confidence Interval Formula for the Difference of Two Proportions)
The formula is as follows:

$$(\hat{p_1} - \hat{p_2}) \pm z^* \times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

where $\hat{p_1}$ and $\hat{p_2}$ are the respective sample proportions, n_1 and n_2 are the respective sample sizes, and z^* is the critical value for the desired confidence level.

The standard error for the difference in proportions combines the variability from both samples, taking the square root of the sum of their variances. This is analogous to the process for the difference of means, but we use proportions instead of means.

The steps for conducting a confidence interval for the difference between population proportions are identical to those of a Confidence Interval for a population proportion.

Problem 6.8.2 (Source: 2006 AP Statistics FRQ Problem 2, Form B) — A large company has two shifts - a day shift and a night shift. Parts produced by the two shifts must meet the same specifications. The manager of the company believes that there is a difference in the proportions of parts produced within specifications by the two shifts. To investigate this belief, random samples of parts that were produced on each of these shifts were selected. For the day shift, 188 of its 200 selected parts met specifications. For the night shift, 180 of its 200 selected parts met specifications.

(a) Used a 96 percent confidence interval to estimate the difference in the proportions of parts produced within specifications by the two shifts.

(b) Based only on this confidence interval, do you think that the difference in the proportions of parts produced within specifications by the two shifts is significantly different from 0? Justify your answer.

Solution
(a) **State:** We will construct a 96% Confidence Interval for the true difference in the proportion of parts produced within specifications by the two shifts.

Plan: We will conduct a two-sample z-interval for the difference between population proportions.
✓Random: The problem states that the data was chosen through random samples of the population of interest.
✓Large Counts: For both samples, the number of successes and failures is greater than 10 (Day shift: 188 successes $200-188 = 12$ failures, Night shift: 180 successes, $200-180 = 20$ failures)
✓Independent: We must assume that each worker either only works a day shift OR a night shift. Furthermore, we must assume that there at least 2000 day and night workers.

Do:

$$\hat{p_1} = 188/200 = 0.94,\ \hat{p_2} = 180/200 = 0.9$$

$$\text{Confidence Interval: } (\hat{p_1} - \hat{p_2}) \pm z^* \times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

$$\text{Confidence interval: } (0.94-0.9) \pm 2.0537 \times \sqrt{\frac{\hat{0.94}(1-\hat{0.94})}{200} + \frac{\hat{0.9}(1-\hat{0.9})}{200}} = 0.04 \pm 0.0556$$

$$\text{Confidence Interval: } (-0.0156, 0.0956)$$

Conclusion: Based on these samples, we are 96% confident that the true difference between the proportion of parts meeting specifications in the day shift and the proportion of parts meeting specifications in the night shift lies within the interval $(-0.0156, 0.0956)$.

(b) More on this in the next section

§6.9 Justifying a Claim Based on a Confidence Interval for a Difference of Population Proportions

Note 6.9.1 (Interpreting Confidence Intervals)

When interpreting a confidence interval for the difference between two proportions, consider the following:

- If the confidence interval includes zero, it suggests that there is no significant difference between the two population proportions. The observed difference could be due to random sampling variability.
- If the entire confidence interval lies above or below zero, it suggests a statistically significant difference between the two proportions. In this case, we have evidence to support a claim that the two proportions are different.
- The width of the confidence interval indicates the precision of the estimate. A narrower interval means a more precise estimate of the difference, while a wider interval indicates greater uncertainty.
- Recall, your confidence is not a probability statement. Never say that there is a $C\%$ chance that the true population parameter lies within some interval. It either does, or does not.
- ALWAYS tie the conclusion back into the context of the problem

Problem 6.9.2 (Source: Original) — A political researcher is studying the voting preferences of two different age groups. A survey of 200 voters aged 18-30 shows that 130 plan to vote for Candidate A, while a survey of 250 voters aged 31-50 shows that 145 plan to vote for Candidate A. Construct and interpret a 95% confidence interval for the difference in proportions of voters who prefer Candidate A between the two age groups.

Solution: The 95% confidence interval for the difference in proportions of voters who prefer Candidate A between the two age groups is $(-0.02, 0.16)$. Since this interval includes zero, we do not have sufficient evidence to conclude that there is a significant

difference in the proportion of voters who prefer Candidate A between the two age groups. The difference observed in the sample could be due to random variation in sampling.

Problem 6.9.3 (Source) — A fitness app company investigates the difference in the proportions of users who complete a 30-day workout challenge between two different workout programs. A 99% confidence interval for the difference in proportions is $(0.12, 0.25)$. Can the company claim that one program leads to significantly higher completion rates?

Solution: Since the confidence interval $(0.12, 0.25)$ does not include zero, the company can claim that there is a statistically significant difference in the completion rates between the two programs. Program A leads to a higher completion rate than Program B.

§6.10 Setting Up a Test for the Difference of Two Population Proportions

The process of setting up and carrying out a hypothesis test for the difference of two population proportions is very similar to that for a single population proportion. In this section, we will outline the key concepts and steps involved in conducting a test for the difference of two proportions.

Note 6.10.1 (Null and Alternative Hypotheses for Two Proportions)

The null hypothesis for comparing two population proportions states that there is no difference between the proportions of the two populations. This can be written as: $H_0 : p_1 - p_2 = 0$

The alternative hypothesis will depend on the context of the problem. It can take one of three forms:

- Right-tailed: $H_a : p_1 - p_2 > 0$ (if you believe that the first population proportion is greater than the second)
- Left-tailed: $H_a : p_1 - p_2 < 0$ (if you believe that the first population proportion is less than the second)
- Two-tailed: $H_a : p_1 - p_2 \neq 0$ (if you are testing whether the two proportions are simply different)

In the context of comparing two proportions, we often ask whether the observed difference between the sample proportions, $\hat{p}_1 - \hat{p}_2$, could have occurred by random chance if the true difference in population proportions is zero. If the observed difference is large enough, we may reject the null hypothesis in favor of the alternative hypothesis.

Note 6.10.2 (Setting Up a Test for the Difference Between Population Proportions)

Here are the steps for setting up a test for the difference between population proportions:

1. State the Hypotheses: Clearly define the null and alternative hypotheses.
2. Name the inference test: Specify that you are performing a two-sample z-test for the difference in population proportions.
3. Check Conditions for Inference:
   - Random: Both samples must be randomly selected from their respective populations.
   - Large Counts: Both sample sizes should be large enough to satisfy the condition that the number of successes and failures in each sample is at least 10:
     $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$.
   - Independent: The two samples must be independent, and each sample size should be less than 10% of the population if sampling without replacement.

Problem 6.10.3 (Source: Original) — A school principal is investigating whether a new tutoring program has a significant effect on the proportion of students passing their final exams. Two groups of randomly sampled students are compared: one group that participated in the tutoring program and another group that did not. The principal finds that 70 out of 100 students who participated in the tutoring program passed, while 60 out of 100 students who did not participate passed. Do these data provide convincing evidence that the tutoring program improves the proportion of students passing their exams?

State: Let p_t be the true proportion of students passing who participated in the tutoring program, and let p_c be the true proportion of students passing who did not participate in the program.
$H_0 : p_t = p_c$
$H_a : p_t > p_c$

Plan: We conduct a 2-sample z-test for the difference in population proportions.
✓Random - Students were randomly sampled, and randomly assigned to participate in the tutoring program or not.
✓Large Counts - The number of successes and failures in each sample is greater than 10.
✓Independent - The samples are independent of each other, and each sample size is less than 10% of the total population of students in the school.

§6.11 Carrying Out a Test for the Difference of Two Population Proportions

Once we have set up a test for the difference of two population proportions, the next step is to carry out the test. This involves calculating the test statistic, determining the p-value, and making a decision about the null hypothesis based on the results.

Theorem 6.11.1 (Test statistic for the difference between two proportions)

The test statistic for comparing two population proportions measures how far the observed difference in sample proportions is from the null hypothesis value (usually zero), relative to the standard error. It is calculated as follows:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions, n_1 and n_2 are the sample sizes, and $\hat{p}$ is the pooled sample proportion, calculated as:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Where x_1 and x_2 are the number of successes in each sample.

Note 6.11.2 (Why care about pooled proportion?)

Note that our test statistic is calculated relative to the standard error of the of the difference of the sample proportion, assuming the null hypothesis is true. Assuming that $p_1 = p_2 = p$, we derive our formula for pooled population below:

$$p_1 = \frac{x_1}{n_1} = p_2 = \frac{x_2}{n_2} = p$$

$$x_1 = n_1 \times p_1,\ x_2 = n_2 \times p_2$$

$$\frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \times p_1 + n_2 \times p_2}{n_1 + n_2} = \frac{p(n_1 + n_2)}{n_1 + n_2} = p$$

Note 6.11.3 (Carrying out a test for the difference between population proportions)

Here are the steps to follow for carrying out this inference test:

1. Calculate the Pooled Proportion: Since the null hypothesis assumes that the two population proportions are equal, we pool the data from both samples to compute a single estimated proportion. The pooled proportion, $\hat{p}$, is used to calculate the standard error.
2. Compute the Test Statistic: Using the formula for the test statistic, calculate the z-value, which represents how many standard errors the observed difference in sample proportions is away from the null hypothesis value.
3. Determine the P-value: The p-value is the probability of observing a difference as extreme or more extreme than the one in the sample data, assuming the null hypothesis is true. For a two-tailed test, the p-value is the area in both tails of the standard normal distribution beyond $\pm z$. For a one-tailed test, it is the area in one tail beyond z or $-z$, depending on the direction of the alternative hypothesis.
4. Compare the P-value to the Significance Level: If the p-value is less than the significance level α, we reject the null hypothesis in favor of the alternative hypothesis. Otherwise, we fail to reject the null hypothesis.

Now, we continue with the example problem from the previous section

Problem 6.11.4 — Source: Original A school principal is investigating whether a new tutoring program has a significant effect on the proportion of students passing their final exams. Two groups of students are compared: one group that participated in the tutoring program and another group that did not. The principal finds that 70 out of 100 students who participated in the tutoring program passed, while 60 out of 100 students who did not participate passed. Do these data provide convincing evidence that the tutoring program improves the proportion of students passing their exams?

Do:

$$\hat{p_1} = \frac{70}{100} = 0.7,\ \hat{p_2} = \frac{60}{100} = 0.6,\ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{70 + 60}{100 + 100} = 0.65$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{(0.7 - 0.6) - (0)}{\sqrt{0.65(1-0.65)\left(\frac{1}{100} + \frac{1}{100}\right)}} = \frac{0.1}{\sqrt{0.0675}} = 1.48$$

Using our calculator, we obtain an associated p-value of about 0.069

Conclude: Since our P-value of 0.069 is above the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. There is no statistically significant evidence proving that for students similar to those in the study, the tutoring program improves the proportion of students passing their exams.

Problem 6.11.5 (Source: 2019 AP Statistics FRQ Problem 4) — Tumbleweed, commonly found in the western United States, is the dried structure of certain plants that are blown by the wind. Kochia, a type of plant that turns into tumbleweed at the end of the summer, is a problem for farmers because it takes nutrients away from soil that would otherwise go to more beneficial plants. Scientists are concerned that kochia plants are becoming resistant to the most commonly used herbicide, glyphosate. In 2014, 19.7 percent of 61 randomly selected kochia plants were resistant to glyphosate. In 2017, 38.5 percent of 52 randomly selected kochia plants were resistant to glyphosate. Do the data provide convincing statistical 2019 evidence, at the level of $\alpha = 0.05$, that there has been an increase in the proportion of all kochia plants that are resistant to glyphosate?

Solution

State: Let p_7 be the true proportion of the population of kochia plants in the western United States that were resistant to glyphosate in 2017. Let p_4 be the true proportion of the population of kochia plants in the western United States that were resistant to glyphosate in 2014.
$H_0 : p_7 = p_4$
$H_a : p_7 > p_4$

Plan: We conduct a 2-sample z-test for the difference in population proportions.
✓Random - All kochia plants were randomly selected.
✓Normal - The sampling distribution of $\hat{p_7} - \hat{p_4}$ is approximately normal, since $n_4\hat{p_4} = 61 \times 0.197 = 12.017 \geq 10$, $n_4(1 - \hat{p_4}) = 61 \times (1 - 0.197) = 48.983 \geq 10$, $n_7\hat{p_7} = 52 \times 0.385 = 20.02 \geq 10$, $n_7(1 - \hat{p_7}) = 52 \times (1 - 0.385) = 31.98 \geq 10$.
✓Independent - The two samples are independent of each other, and furthermore, we may assume that there were more than 610 kochia plants in 2014 and more than 520 kochia plants in 2017.

Do:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{12.017 + 20.02}{61 + 52} = 0.284$$

$$z = \frac{(0.385 - 0.197) - (0)}{\sqrt{0.284(1 - 0.2845)\left(\frac{1}{61} + \frac{1}{52}\right)}}$$

$$z = 2.21$$

Using our calculator, we obtain an associated p-value of about 0.0135

Conclude: Since our P-value of 0.0135 is below the significance level of $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence of an increase in the proportion of all kochia plants resistant to glyphosate from 2014 to 2017.

Unit 6 Practice Problems

Problem 6.11.1 (Source: 2024 AP Statistics FRQ Problem 1) — A large exercise center has several thousand members from age 18 to 55 years and several thousand members age 56 and older. The manager of the center is considering offering online fitness classes. The manager is investigating whether members' opinions of taking online fitness classes differ by age. The manager selected a random sample of 170 exercise center members ages 18 to 55 years and a second random sample of 230 exercise center members ages 56 years and older. Each sampled member was asked whether they would be interested in taking online fitness classes. The manager found that 51 of the 170 sampled members ages 18 to 55 years and that 79 of the 230 sampled members ages 56 years and older said they would be interested in taking online fitness classes. At a significance level of $\alpha = 0.05$, do the data provide convincing statistical evidence of a difference in the proportion of all exercise center members ages 18 to 55 years who would be interested in taking online fitness classes and the proportion of all exercise center members ages 56 years and older who would be interested in taking online fitness classes? Complete the appropriate inference procedure to justify your response.

Solution

State: Let p_1 represent the proportion of all members ages 18 to 55 years who would be interested in taking online fitness classes, and let p_2 represent the proportion of all members ages 56 years and older who would be interested in taking online fitness classes.
$H_0 : p_1 - p_2 = 0$ (There is no difference in proportions)
$H_a : p_1 - p_2 \neq 0$ (There is a difference in proportions)

Plan: We will conduct a two-sample z-test for the difference in proportions.
✓Random - The samples were randomly selected from the populations of interest.
✓Normal - We check if the sampling distributions are approximately normal using the conditions: $170(51/170) = 51 \geq 10$, $170(119/170) = 119 \geq 10$, $230(79/230) = 79 \geq 10$, $230(151/230) = 151 \geq 10$.
✓Independent - We may safely assume that our sample sizes are at most 10% of their respective populations.

Do:

$$\hat{p}_1 = \frac{51}{170} = 0.3,\ \hat{p}_2 = \frac{79}{230} = 0.343,\ \hat{p} = \frac{51+79}{170+230} = 0.325$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{(0.3 - 0.343) - (0)}{\sqrt{0.325(1-0.325)\left(\frac{1}{170} + \frac{1}{230}\right)}} = \frac{-0.043}{\sqrt{0.00224}} = -0.91$$

Using our calculator, we obtain an associated p-value of about 0.362.

Conclude: Since our p-value of 0.362 is above the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. There is no statistically significant evidence proving that the proportion of members interested in online fitness classes differs between members ages 18 to 55 years and members ages 56 years and older.

Problem 6.11.2 (Source: 2022 AP Statistics FRQ Problem 4) — A survey conducted by a national research center asked a random sample of 920 teenagers in the United States how often they use a video streaming service. From the sample, 59% answered that they use a video streaming service every day.

(a) Construct and interpret a 95% confidence interval for the proportion of all teenagers in the United States who would respond that they use a video streaming service every day.

(b) Based on the confidence interval in part (a), do the sample data provide convincing statistical evidence that the proportion of all teenagers in the United States who would respond that they use a video streaming service every day is not 0.5? Justify your answer.

Solution

(a) **State:** We want to construct a 95% confidence interval for the proportion of all teenagers in the United States who would respond that they use a video streaming service every day.

Plan: We will use a one-sample z-interval for proportions.
✓Random - The sample of 920 teenagers was randomly selected.
✓Normal - The sampling distribution of $\hat{p}$ is approximately normal since $n\hat{p} = 920(0.59) = 542.8 \geq 10,\ n(1-\hat{p}) = 920(0.41) = 377.2 \geq 10$
✓Independent - We can assume that there are more than 9200 U.S. teenagers.

Do:

$$\text{Confidence Interval: } \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Confidence Interval: } 0.59 \pm 1.96\sqrt{\frac{0.59(1-0.59)}{920}} = 0.59 \pm 1.96 \times 0.0161$$

$$\text{Confidence Interval: } 0.59 \pm 0.0315$$

The confidence interval is $(0.5585, 0.6215)$.

Conclude: We are 95% confident that the proportion of all teenagers in the United States who use a video streaming service every day is between 55.85% and 62.15%.

(b) Since 0.5 is not contained in the 95% confidence interval (0.5585, 0.6215), the sample data provides convincing statistical evidence that the proportion of all teenagers in the United States who would use a streaming service every day is **NOT** 0.5

Problem 6.11.3 (Source: 2021 AP Statistics FRQ Problem 4) — The manager of a large company that sells pet supplies online wants to increase sales by encouraging repeat purchases. The manager believes that if past customers are offered \$10 off their next purchase, more than 40 percent of them will place an order. To investigate the belief, 90 customers who placed an order in the past year are selected at random. Each of the selected customers is sent an e-mail with a coupon for \$10 off the next purchase if the order is placed within 30 days. Of those who receive the coupon, 38 place an order.

(a) Is there convincing statistical evidence, at the significance level of $\alpha = 0.05$, that the manager's belief is correct? Complete the appropriate inference procedure to support your answer.

(b) Based on your conclusion from part (a), which of the two errors, Type I or Type II, could have been made? Interpret the consequence of the error in context.

Solution

(a) Let p represent the proportion of all customers who would place an order after receiving the coupon.
$H_0 : p = 0.40$ (The proportion of customers placing an order is 40%)
$H_a : p > 0.40$ (The proportion of customers placing an order is greater than 40%)

Plan: We will conduct a one-sample z-test for proportions.
✓Random - The sample of 90 customers was randomly selected from past customers.
✓Normal - The sampling distribution of $\hat{p}$ is approximately normal, since $n\hat{p} = 90(0.40) = 36 \geq 10, \quad n(1-\hat{p}) = 90(0.60) = 54 \geq 10$
✓Independent - We can assume that there are more than 900 customers in the population, ensuring independence.

Do:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $\hat{p} = \frac{38}{90} = 0.422$

$$z = \frac{0.422 - 0.40}{\sqrt{\frac{0.40(1-0.40)}{90}}} = \frac{0.022}{\sqrt{\frac{0.40 \times 0.60}{90}}} = \frac{0.022}{0.0516} = 0.426$$

Using our calculator, we find the p-value associated with $z = 0.426$ is approximately 0.335.

Conclude: Since our p-value of 0.335 is greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. There is not enough statistical evidence to support the manager's belief that more than 40% of past customers will place an order after receiving a coupon.

(b) Since we failed to reject the null hypothesis, it is possible that a Type II error could have been made. This means that the null hypothesis was not rejected when, in fact, more than 40% of past customers who received the coupon would place an order. The consequence of this error is that the manager might incorrectly conclude that offering the coupon is not effective in increasing sales, when in reality it is.

Problem 6.11.4 (Source: 2015 AP Statistics FRQ Problem 2) — To increase business, the owner of a restaurant is running a promotion in which a customer's bill can be randomly selected to receive a discount. When a customer's bill is printed, a program in the cash register randomly determines whether the customer will receive a discount on the bill. The program was written to generate a discount with a probability of 0.2, that is, giving 20 percent of the bills a discount in the long run. However, the owner is concerned that the program has a mistake that results in the program not generating the intended long-run proportion of 0.2.

The owner selected a random sample of bills and found that only 15 percent of them received discounts. A confidence interval for p, the proportion of bills that will receive a discount in the long run, is 0.15 ± 0.06. All conditions for inference were met.

(a) Consider the confidence interval 0.15 ± 0.06.

(i) Does the confidence interval provide convincing statistical evidence that the program is not working as intended? Justify your answer.

(ii) Does the confidence interval provide convincing statistical evidence that the program generates the discount with a probability of 0.2 ? Justify your answer.

A second random sample of bills was taken that was four times the size of the original sample. In the second sample 15 percent of the bills received the discount.

(b) Determine the value of the margin of error based on the second sample of bills that would be used to compute an interval for p with the same confidence level as that of the original interval.

(c) Based on the margin of error in part (b) that was obtained from the second sample, what do you conclude about whether the program is working as intended? Justify your answer.

Solution

(a) (i) The intended proportion of bills receiving a discount is $p = 0.2$. The confidence interval is $(0.09, 0.21)$. Since 0.2 is inside the confidence interval, the confidence interval does not provide convincing statistical evidence that the program is not working as intended.

(ii)No, the confidence interval does not provide convincing statistical evidence that the program generates a discount with a probability of 0.2. Although 0.2 is within the confidence interval, the interval is quite large, and any value within the interval is a plausable value of the true proportion.

(b) The margin of error for the original sample is 0.06. The second sample size is four times the size of the original sample. Since the margin of error decreases as the square root of the sample size increases, the margin of error for the second sample can be calculated as:

$$\text{New margin of error} = \frac{0.06}{\sqrt{4}} = \frac{0.06}{2} = 0.03$$

(c) The new confidence interval for the second sample is $(0.12, 0.18)$. Since the intended proportion of $p = 0.2$ is not within this new confidence interval, this provides convincing statistical evidence that the program is not generating discounts with a probability of 0.2.

Problem 6.11.5 (Source: 2015 AP Statistics FRQ Problem 4) — A researcher conducted a medical study to investigate whether taking a low-dose aspirin reduces the chance of developing colon cancer. As part of the study, 1,000 adult volunteers were randomly assigned to one of two groups. Half of the volunteers were assigned to the experimental group that took a low-dose aspirin each day, and the other half were assigned to the control group that took a placebo each day. At the end of six years, 15 of the people who took the low-dose aspirin had developed colon cancer and 26 of the people who took the placebo had developed colon cancer. At the significance level a 0.05, do the data provide convincing statistical evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers?

Solution State: Let p_1 represent the proportion of people who develop colon cancer in the low-dose aspirin group, and let p_2 represent the proportion of people who develop colon cancer in the placebo group.
$H_0 : p_1 - p_2 = 0$
$H_a : p_1 - p_2 < 0$

Plan: We will conduct a two-sample z-test for the difference in proportions.
✓Random - The 1,000 adult volunteers were randomly assigned to one of the two groups.
✓Normal - We check if the sampling distributions are approximately normal using the conditions: $n_1\hat{p}_1 = 500 \times \frac{15}{500} = 15 \geq 10, \quad n_1(1-\hat{p}_1) = 500 - 15 = 485 \geq 10$
$n_2\hat{p}_2 = 500 \times \frac{26}{500} = 26 \geq 10, \quad n_2(1-\hat{p}_2) = 500 - 26 = 474 \geq 10$
✓Independent - The populations are large enough, so we can assume the independence of observations.

Do:

$$\hat{p}_1 = \frac{15}{500} = 0.03, \quad \hat{p}_2 = \frac{26}{500} = 0.052$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{15 + 26}{500 + 500} = \frac{41}{1000} = 0.041$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.03 - 0.052 - 0}{\sqrt{0.041(1-0.041)\left(\frac{1}{500} + \frac{1}{500}\right)}} = \frac{-0.022}{\sqrt{0.041 \times 0.959 \times \left(\frac{1}{500} + \frac{1}{500}\right)}}$$

$$= \frac{-0.022}{\sqrt{0.0393 \times 0.004}} = \frac{-0.022}{0.0125} = -1.76$$

Using our calculator, we find the p-value associated with $z = -1.76$ is approximately 0.078.

Conclude: Since our p-value of 0.078 is greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. There is no statistically significant evidence at the 0.05 significance level to suggest that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers.

Problem 6.11.6 (Source: 2012 AP Statistics FRQ Problem 4) — A survey organization conducted telephone interviews in December 2008 in which 1,009 randomly selected adults in the United States responded to the following question.

At the present time, do you think television commercials are an effective way to promote a new product?

Of the 1,009 adults surveyed, 676 responded "yes." In December 2007, 622 of 1,020 randomly selected adults in the United States had responded "yes" to the same question. Do the data provide convincing evidence that the proportion of adults in the United States who would respond "yes" to the question changed from December 2007 to December 2008 ?

Solution

State: Let p_1 represent the proportion of adults who responded "yes" in December 2008, and let p_2 represent the proportion of adults who responded "yes" in December 2007.
$H_0 : p_1 - p_2 = 0$
$H_a : p_1 - p_2 \neq 0$

Plan: We will conduct a two-sample z-test for the difference in proportions.
✓Random - The samples from both years were randomly selected.
✓Normal - We check if the sampling distributions are approximately normal using the conditions: $n_1\hat{p}_1 = 1009 \times \frac{676}{1009} = 676 \geq 10, \quad n_1(1-\hat{p}_1) = 1009 - 676 = 333 \geq 10$ $n_2\hat{p}_2 = 1020 \times \frac{622}{1020} = 622 \geq 10, \quad n_2(1-\hat{p}_2) = 1020 - 622 = 398 \geq 10$ All conditions are satisfied.
✓Independent - The populations are large enough and the samples were drawn without replacement, so we can assume the independence of observations.

Do:

$$\hat{p}_1 = \frac{676}{1009} = 0.670, \quad \hat{p}_2 = \frac{622}{1020} = 0.610$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{676 + 622}{1009 + 1020} = \frac{1298}{2029} = 0.640$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.670 - 0.610 - 0}{\sqrt{0.640(1-0.640)\left(\frac{1}{1009} + \frac{1}{1020}\right)}} = \frac{0.060}{\sqrt{0.640 \times 0.360\left(\frac{1}{1009} + \frac{1}{1020}\right)}}$$

$$= \frac{0.060}{\sqrt{0.2304\left(\frac{1}{1009} + \frac{1}{1020}\right)}} = \frac{0.060}{\sqrt{0.2304 \times 0.00198}} = \frac{0.060}{0.0214} = 2.80$$

Using our calculator, we find the p-value associated with $z = 2.80$ is approximately 0.0051.

Conclude: Since our p-value of 0.0051 is less than the significance level $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence that the proportion of adults who would respond "yes" to the question changed from December 2007 to December 2008.

Problem 6.11.7 (Source: 2012 AP Statistics FRQ Problem 5) — A recent report stated that less than 35 percent of the adult residents in a certain city will be able to pass a physical fitness test. Consequently, the city's Recreation Department is trying to convince the City Council to fund more physical fitness programs. The council is facing budget constraints and is skeptical of the report. The council will fund more physical fitness programs only if the Recreation Department can provide convincing evidence that the report is true.

The Recreation Department plans to collect data from a sample of 185 adult residents in the city. A test of significance will be conducted at a significance level of a $= 0.05$ for the following hypotheses. $H_0 : p = 0.35$, $H_a : p < 0.35$ where p is the proportion of adult residents in the city who are able to pass the physical fitness test.

(a) Describe what a Type II error would be in the context of the study, and also describe a consequence of making this type of error.

(b) The Recreation Department recruits 185 adult residents who volunteer to take the physical fitness test. The test is passed by 77 of the 185 volunteers, resulting in a p-value of 0.97 for the hypotheses stated above. If it was reasonable to conduct a test of significance for the hypotheses stated above using the data collected from the 185 volunteers, what would the p-value of 0.97 lead you to conclude?

(c) Describe the primary flaw in the study described in part (b), and explain why it is a concern.

Solution

(a) A Type II error occurs when we fail to reject the null hypothesis when the alternative hypothesis is actually true. In this context, a Type II error would occur if the true proportion of adult residents who can pass the physical fitness test is actually less than 0.35, but we incorrectly conclude that it is not. The consequence of making a Type II error would be that the Recreation Department might not receive the necessary funding to improve physical fitness programs, even though a higher proportion of residents might indeed fail the test.

(b) The p-value of 0.97 is very high, indicating that the observed proportion of 0.416 passing the test is consistent with the null hypothesis that $p = 0.35$. This suggests that the sample data do not provide evidence that a lower proportion of residents can pass the test compared to the reported proportion of 0.35.

(c) The primary flaw in the study is that the sample was composed of volunteers who may not be representative of the general population. Volunteers might be more health-conscious or physically fit than the average adult resident thus the sample may over-represent those who are more likely to pass the test.

Problem 6.11.8 (Source: 2009 AP Statistics FRQ Problem 5) — For many years, the medically accepted practice of giving aid to a person experiencing a heart attack was to have the person who placed the emergency call administer chest compression (CC) plus standard mouth-to-mouth resuscitation (MMR) to the heart attack patient until the emergency response team arrived. However, some researchers believed that CC alone would be a more effective approach. In the 1990s a study was conducted in Seattle in which 518 cases were randomly assigned to treatments: 278 to CC plus standard MMR and 240 to CC alone. A total of 64 patients survived the heart attack: 29 in the group receiving CC plus standard MMR, and 35 in the group receiving CC alone. A test of significance was conducted on the following hypotheses.

H_0 : The survival rates for the treatments are equal.

H_a : The treatment that uses CC alone produces a higher survival rate.

This test resulted in a p-value of 0.0761.

(a) Interpret what this p-value measures in the context of this study.

(b) Based on this p-value and study design, what conclusion should be drawn in the context of this study? Use a significance level of $\alpha = 0.05$.

(c) Based on your conclusion in part (b), which type of error, Type I or Type II, could have been made? What is one potential consequence of this error?

Solution

(a) The p-value of 0.0761 measures the probability of observing a survival rate difference as extreme or more extreme than the one observed in this study, assuming that the true survival rates for both treatments are equal. In other words, if there is no real difference in survival rates between the two treatments, there is about a 7.61% chance of obtaining the observed difference in survival rates purely by random chance.

(b)Since the p-value of 0.0761 is greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. The study does not provide convincing statistical evidence at the 5% significance level to support the claim that the treatment using CC alone produces a higher survival rate than the treatment using CC plus MMR.

(c) Based on the conclusion in part (b), a Type II error could have been made, which occurs when we fail to reject the null hypothesis when the alternative hypothesis is actually true. In this context, a Type II error would mean that the treatment using CC alone actually does produce a higher survival rate, but the study failed to detect this difference. One potential consequence of this error is that medical professionals might continue using the less effective CC plus MMR treatment, potentially leading to worse outcomes for patients experiencing heart attacks.

7 Unit 7: Inference for Quantitative Data: Means

§7.1 Introducing Statistics: Should I Worry About Error?

In statistics, the concept of error is crucial, especially when dealing with quantitative data.

Definition 7.1.1 (What is Error?)

In statistics, error refers to the difference between the observed value and the true value of a parameter. It can arise from various sources such as sampling methods, measurement inaccuracies, or random fluctuations in the data.

For example, if we wanted to take a sample to estimate the average IQ of high school students in the U.S., we could take a sample of 50 high school students and calculate the mean IQ of our sample. However, this sample does perfectly represent our population of interest, and thus the sample mean will likely not equal the true population mean. The difference between our sample mean and population mean (whether the sample mean is way above or below the true mean) is called the **error**.

The purpose of our inference tests and confidence intervals is not to eliminate error (since that is impossible), but rather to learn how to control it and make reasonable claims about certain parameters.

Like the previous unit, this unit is heavily connected with the idea of sampling distributions. However, the normal distribution can only be used if the true population mean and standard deviation is known, which in most cases it is not. If we don't know these true parameters, we instead use a different distribution known as a t-distribution, which is much more accurate. The t-distribution is similar to the normal distribution except for the fact that has more area at its tails. While there is only one normal model, there are many variations of the t-distribution that depends on the number of degrees of freedom. The concept of degrees of freedom will be discussed in depth in the next section.

Note 7.1.2 (Conditions for inference for Quantitative Data: Means)

We end this section by introducing the conditions that must be met to do anything in this unit

- Random - Individual observations must be selected randomly from the population of interest
- Normal - The sampling distribution of sample means should be approximately normal. This condition is satisfied if one of the following apply:
  - Sample size is greater than 30
  - Population distribution is approximately normal (must be unimodal and symmetric)
  - Sample distribution is approximately normal (must be unimodal and symmetric)
- Independent - Individual observations are independent of eachother. This either means that sampling is done with replacement, or that the sample size is less than 10% the size of the population.

§7.2 Constructing a Confidence Interval for a Population Mean

Constructing a confidence interval for a population mean is very similar to constructing a confidence interval for a population proportion, with one key difference: **Instead of using a (z^*) value, we use a (t^*) value to determine our standard error (most of the time).** This is done to give a more accurate interval size when compared with the standard (z^*) statistic, which often underestimates the margin of error in population means. The only time we use a z-statistic when working with population means is if we know the **POPULATION** standard deviation. Otherwise, if we are only given the **SAMPLE** standard deviation, we must use a t-statistic since we are *estimating* our standard error rather then computing it exactly. With this new model, the value of the (t^*) statistic depends on the number of **degrees of freedom**.

Definition 7.2.1 (Degree of Freedom)

In statistics, **degrees of freedom** (often abbreviated as "**df**") refers to the number of independent values or quantities that can vary in a statistical calculation without violating any given constraints.

Although this might sound confusing, the number of degrees of freedom is just one less than the sample size. For example, if the sample size is 30, the number of degrees of freedom would be 29. Intuitively, this is because if you are given the mean of a set as well as the value $n-1$ data points, you can uniquely determine the last data point.

Theorem 7.2.2 (Confidence Interval Formula for a Population Mean)

$$\text{Confidence Interval} = \bar{x} \pm (t^*)(\frac{s}{\sqrt{n}}) \text{OR} \bar{x} \pm (z^*)(\frac{\sigma}{\sqrt{n}})$$

Where $\bar{x}$ is the sample mean, t^* is the corresponding critical value, s is the sample standard deviation, and n is the population size. Most of the time we will use a critical t value when dealing with means, categorical data, and slopes, since in those cases the population standard deviation is unknown, and we estimate it using sample standard deviation. However, in some problems, you may be given the population standard deviation, in which case you must use the critical z value.

Note 7.2.3 (Margin of Error)

The portion after the $\pm$ is often referred to as the margin of error.

Problem 7.2.4 (Source: Original) — A coffee shop wants to determine the average amount of coffee (in ounces) that its customers consume per visit. The shop manager randomly selects a sample of 30 customers and records the amount of coffee each customer drank during their visit. The sample data reveals an average consumption of 12.5 ounces with a standard deviation of 1.8 ounces. Construct and interpret a 95% confidence interval for the true mean amount of coffee consumed by all customers at this coffee shop per visit.

Solution

State: We will construct a 95% confidence interval for the true mean amount of coffee consumed by customers at this coffee shop per visit.

Plan: We will construct a 1-sample t-interval for means
✓ Random: The sample was randomly sampled from the population of interest.
✓ Normal: The sample size is at least 30.
✓ Independent: It is reasonable to assume that the coffee shop has at least 300 customers, satisfying the 10% rule.

Do:

$$\text{Confidence Interval}: \ \bar{x} \pm (t^*)(\frac{s}{\sqrt{n}}), \ df = n - 1 = 29$$

$$\text{Confidence Interval}: \ 12.5 \pm (2.05)(\frac{1.8}{\sqrt{30}})$$

$$\text{Confidence Interval}: \ (11.83, 13.17)$$

Conclude: We are 95% confident that the true mean amount of coffee consumed by customers at this coffee shop per visit lies within the interval $(11.83, 13.17)$

Problem 7.2.5 (Source: 2008 AP Statistics FRQ Problem 3, Form B) — A car manufacturer is interested in conducting a study to estimate the mean stopping distance for a new type of brakes when used in a car that is traveling at 60 miles per hour. These new brakes will be installed on cars of the same model and the stopping distance will be observed. The cost of each observation is \$100. A budget of \$12,000 is available to conduct the study and the goal is to carry it out in the most economical way possible. Preliminary studies indicate that $\alpha = 12$ feet for stopping distances.

(a) Are sufficient funds available to estimate the mean stopping distance to within 2 feet of the true mean stopping distance with 95% confidence? Explain your answer.

(b) A regulatory agency requires a 95% level of confidence for an estimate of mean stopping distance that is within 2 feet of the true mean stopping distance. The car manufacturer cannot exceed the budget of \$12,000 for the study. Discuss the consequences of these constraints.

(a) With a budget of \$12000 and a cost of \$100 per observation, we can only make a maximum of $\frac{12000}{100} = 120$ observations. Thus, $n \leq 120$. To estimate the mean stopping distance to within 2 feet of the true mean stopping distance with 95% confidence, the Margin of error will have to be equal to 2. Also, since we are given the **POPULATION STANDARD DEVIATION**, we use a critical z-value instead of a critical t-value.

$$\text{Margin of Error} = (z^*)(\frac{\sigma}{\sqrt{n}})$$

$$2 = (1.96)(\frac{12}{\sqrt{n}})$$

$$n = (\frac{(1.96)(12)}{2})^2 = 138.3$$

Since $138.3 > 120$, there are insufficient funds available to estimate the mean stopping distance to within 2 feet of the true mean stopping distance with 95% confidence.

(b) The consequence of these constraints is that there are not enough funds to conduct a estimate the mean stopping distance to within 2 feet of the true mean stopping distance with 95% confidence. To do that, we would need a sample size of 139, which would require $139 \times 100 = \$13900$, which is over the allocated budget.

§7.3 Justifying a Claim About a Population Mean Based on a Confidence Interval

After building the confidence interval, it is important to understand what it does and does not tell us. Recall from section 6.3 that when we say we are $C\%$ confident that the true population mean lies within our interval, it means that if we took many many samples from the same population with a confidence level of $C\%$, approximately $C\%$ of the intervals will contain the true mean.

Note 7.3.1 (Interpreting confidence intervals)

A few quick notes on confidence intervals:

- A confidence interval is **NOT** a probability statement. It does not mean that there is a $C\%$ chance that the interval contains the true mean since it either does or does not. Students often get tricked by these kinds of MC questions.
- If a confidence interval does not include a particular value, we can say that it is not likely that the particular value is the true population mean, no matter how close it is to the interval.

Problem 7.3.2 (Source: Original) — Maya is a production manager at a company that manufactures glass panels. Each panel is designed to have a thickness of 199 millimeters to meet industry standards. To ensure quality, Maya measures the thickness of 60 randomly selected points on a panel. The sample has a mean thickness of $\bar{x} = 198mm$ and a standard deviation of $s = 2.5mm$. Based on this data, a 95% confidence interval for the mean thickness is $(197.4, 198.6)$. Based on this interval, is it plausible that these panels meet the industry standard of 199 millimeters thickness?

Solution

Even though the interval $(197.4, 198.6)$ is very close to 199, 199 does not lie within it so thus it is not plausible that these panels meet the industry standard.

Problem 7.3.3 (Source: Original) — Liam read that the average employee at a tech company is 29 years old. Curious about the age distribution at his tech firm, Liam decided to estimate the mean age of employees at his company. He randomly sampled 35 employees and found that their mean age was $\bar{x} = 27.5$ years with a standard deviation of s=5.1 years. A 95% confidence interval for the mean age based on his sample was $(25.8, 29.2)$. Based on this interval, is it plausible that the mean age of employees at Liam's company matches the industry average of 29 years?

Solution

Yes, because the industry average of 29 years lies within the interval of $(25.8, 29, 2)$!

§7.4 Setting Up a Test for a Population Mean

The steps for setting up and carrying out a test for a population mean are the same as the steps for a population proportion.

Note 7.4.1 (Steps for Setting up a Test for a Population Mean)

Follow these steps:

1. Setting up the Hypotheses
   - Name the test being used
   - Build the null and alternative hypotheses for the parameter of interest

     Note 7.4.2

     Our null and alternative hypotheses should look something like this:
     $H_0 : \mu = 3$
     $H_a : \mu > 3$
     The null and alternative hypotheses should look **EXACTLY** the same, except for the comparison operator. In other words, the number and symbol in the hypotheses should be the same.
2. Checking the conditions for inference - Random, Normal, Independent

Before we go into an example, first we review the concept of a Matched Pairs experiment. Recall from Unit 2 that a matched pairs experiment is an experiment that compares the effect of two different treatments a population. In this design, individuals are paired based on their similar characteristics, and in each pair one person is given one of the treatments, and the other is given the alternative.

Later on in this unit, we will learn more about tests for the difference of two population means, which will look very similar to tests involving a matched pairs design. However, the matched pairs design still only contains **ONE** population, and thus should be treated as a test for a single population mean. For example, a matched pairs experiment might look something like the following: In a study to compare the effectiveness of two different studying techniques, students are paired based on their initial test scores. One student from each pair uses Technique A, while the other uses Technique B. After a period of time, their test scores are compared to assess which technique was more effective.

Pair	1	2	3	4	5	6	7	8	9	10
Study method A	78	97	85	62	94	77	39	50	94	88
Study method B	84	96	88	63	98	72	49	55	92	88
Difference (B-A)	6	−1	3	1	4	−5	10	5	−2	0

Conducting an inference test for a matched pairs test is the same as a test for a population mean, however, you are taking the mean of the **DIFFERENCES** between the different treatments. It's important to distinguish between a matched pairs design and a test for the difference between two population means, but just keep in mind that one only contains one population, while the other has two.

Problem 7.4.3 (Source: 2006 AP Statistics FRQ Problem 4, Form B) — The developers of a training program designed to improve manual dexterity claim that people who complete the 6-week training program will increase their manual dexterity. A random sample of 12 people enrolled in the training program was selected. A measure of each person's dexterity on a scale from 1 (lowest) to 9 (highest) was recorded just before the start of and just after the completion of the 6-week program. The data are shown in the table below.

Person	Before Program	After Program
A	6.7	7.8
B	5.4	5.9
C	7.0	7.6
D	6.6	6.6
E	6.9	7.6
F	7.2	7.7
G	5.5	6.0
H	7.1	7.0
I	7.9	7.8
J	5.9	6.4
K	8.4	8.7
L	6.5	6.5
Total	81.1	85.6

Can one conclude that the mean manual dexterity for people who have completed the 6-week training program has significantly increased? Support your conclusion with appropriate statistical evidence.

State: Let μ_d denote the mean of the differences of dexterity scores of individuals enrolled in the program (after program − before program).
$H_0 : \mu_d = 0$
$H_a : \mu_d > 0$

Plan: We will conduct a paired t-test for population means.
✓Random - The individuals were selected from a random sample
✓Normal - Since we only have a sample of 12 individuals, we must assume that the distribution of the differences between dexterity scores after and before the training program follows a somewhat normal distribution.
✓Independent - We must assume that there were at least 120 individuals in the training program, or that each individual was selected without replacement.

§7.5 Carrying Out a Test for a Population Mean

Now, all that's left is to find our test statistic and use that to get our associated p-value. Recall that the test statistic is essentially the number of standard error's away our observed mean is from the hypothesized mean. However, since we usually don't know the population standard deviation, most of the time we must estimate it using our sample standard deviation, which is why we use a t-statistic instead of z-statistic.

Theorem 7.5.1 (Test statistic for population means)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The numerator is essentially the distance between our observed mean and the hypothesized mean, while the denominator is the standard error.

Note 7.5.2 (Steps for carrying out a test for a population mean)

Here are the steps to carrying out a test for a population mean:

1. Obtain the corresponding test statistic by using the formula above
2. Use your calculator to obtain the associated p-value from the test statistic
3. Draw a conclusion based on the p-value. If it is lower than the significance level, you may reject the null hypothesis. Otherwise, you may not.

Now, we continue the problem from the previous section.

Problem 7.5.3 (Source: 2006 AP Statistics FRQ Problem 4, Form B) — The developers of a training program designed to improve manual dexterity claim that people who complete the 6-week training program will increase their manual dexterity. A random sample of 12 people enrolled in the training program was selected. A measure of each person's dexterity on a scale from 1 (lowest) to 9 (highest) was recorded just before the start of and just after the completion of the 6-week program. The data are shown in the table below.

Person	Before Program	After Program
A	6.7	7.8
B	5.4	5.9
C	7.0	7.6
D	6.6	6.6
E	6.9	7.6
F	7.2	7.7
G	5.5	6.0
H	7.1	7.0
I	7.9	7.8
J	5.9	6.4
K	8.4	8.7
L	6.5	6.5
Total	81.1	85.6

Can one conclude that the mean manual dexterity for people who have completed the 6-week training program has significantly increased? Support your conclusion with appropriate statistical evidence.

Person	Before Program	After Program	(After−Before)
A	6.7	7.8	1.1
B	5.4	5.9	0.5
C	7.0	7.6	0.6
D	6.6	6.6	0
E	6.9	7.6	0.7
F	7.2	7.7	0.5
G	5.5	6.0	0.5
H	7.1	7.0	−0.1
I	7.9	7.8	−0.1
J	5.9	6.4	0.5
K	8.4	8.7	0.3
L	6.5	6.5	0
Total	81.1	85.6	

Do: By using our calculator, we obtain $\bar{x_d} = 0.375,\ s_d = 0.367$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},\ df = 12 - 1 = 11$$

$$t = \frac{0.375 - 0}{\frac{0.367}{\sqrt{12}}} = 3.54$$

By using our calculator, we obtain our corresponding p-value of about 0.002

Conclude: Since our P-value of 0.002 is below the significance level of $\alpha = 0.05$, we reject the null hypothesis. There is significant evidence proving that the mean manual dexterity for people who have completed the 6-week training has significantly increased.

§7.6 Confidence Intervals for the Difference of Two Means

When comparing two independent populations, constructing a confidence interval for the difference of their means allows us to estimate the true difference between these populations. The formula for this interval is very similar to the confidence interval for a single mean but now takes into account both sample means and their standard errors. The most common scenario involves using the t distribution, as we usually do not know the population standard deviations. The confidence interval provides a range of plausible values for the difference between the population means, and the width of the interval reflects the uncertainty of this estimate. If the confidence interval includes zero, it suggests no significant difference between the population means. On the other hand, if the interval does not include zero, we can infer that a difference does exist.

Theorem 7.6.1 (Confidence Interval Formula for the Difference of Two Means)

$$\text{Confidence Interval: } (\bar{x_1} - \bar{x_2}) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where $\bar{x_1}$, $\bar{x_2}$ are the respective sample means, s_1, s_2 are the respective sample standard deviations, n_1, n_2 are the respective sample sizes, and t^* is the corresponding critical t-value.

Recall from unit 5 that to take the standard deviation of the sum of two random variables, you take the square root of the sum of their variances. However, since most of the time we do not know the true population standard deviation for both populations, we must estimate them with their sample standard deviations.

The degrees of freedom for the t-distribution with two samples are calculated using a very complicated formula that is outside the scope of the AP curriculum. Thus, on the actual AP exam, you may either use your ti-84 calculator to compute the number of degrees of freedom, **OR** use the method outlined below.

Note 7.6.2 (Degrees of Freedom)

The number of degrees of freedom for an inference test for the difference of two means is equal to the number of degrees of freedom of the smaller sample size. For example, if the first sample is size 30 and the second is size 40, the degrees of freedom is equal to $30 - 1 = 29$, since $30 < 40$. For the sake of simplicity, this book will be using this method in all the example problems

Problem 7.6.3 (Source: Original) — A researcher is studying the effectiveness of two different study techniques on high school student performances in North America. Two independent groups of students are randomly assigned to either Study Technique A or Study Technique B. After a month, the researcher records the scores on a standardized test for both groups. The results are as follows:

- Group A: $\bar{x_a} = 85$, $s_a = 8$, $n_a = 40$
- Group B: $\bar{x_b} = 80$, $s_b = 10$, $n_b = 30$

Construct and interpret a 95% confidence interval for the difference in mean test scores between students who used Study Technique A and those who used Study Technique B.

State: We will construct a 95% Confidence interval for the true mean difference of test scores between students who use Study Technique A and students who use Study Technique B.

Plan: We will conduct a 2-sample t-test for the difference of population means.
✓Random - Both samples were randomly chosen and assigned to the different study techniques

✓Normal - The sample sizes for both Study Techniques are both at least 30
✓Independent - We may assume that there are more than 750 high school students in North America

Do:

$$\text{Confidence Interval: } (\bar{x_1} - \bar{x_2}) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ df = 30 - 1 = 29$$

$$\text{Confidence Interval: } (40 - 30) \pm 1.699 \times \sqrt{\frac{8^2}{40} + \frac{10^2}{30}}$$

$$\text{Confidence Interval: } (6.226, 13.774)$$

Conclude: We are 95% confident that the true mean difference of test scores between students who use Study Technique A and students who use Study Technique B lies in the interval (6.226, 13.774).

§7.7 Justifying a Claim About the Difference of Two Means Based on a Confidence Interval

When we construct a confidence interval for the difference of two means, we can use it to justify claims about whether there is a significant difference between the two populations. In this section, we will focus on how to interpret a confidence interval in the context of comparing two population means and how this interpretation helps in justifying or rejecting claims.

Note 7.7.1 (Few notes on confidence intervals for the difference of two means)

A few key points to remember about confidence intervals for the difference of two means:

- If a confidence interval for the difference between two population means includes 0, it suggests that there is no statistically significant difference between the two population means. This means that the observed difference could be due to sampling variability.
- If the entire confidence interval lies above or below 0, this suggests that there is a statistically significant difference between the two population means. In this case, we have evidence to support a claim that the two means are different.
- The width of the confidence interval indicates the precision of the estimate of the difference between the two means. Narrower intervals provide more precise estimates, while wider intervals reflect greater uncertainty.

Problem 7.7.2 (Source: Original) — A researcher is studying the effectiveness of two different diet plans. Group 1 follows Diet A, and Group 2 follows Diet B. After 6 months, the researcher records the weight loss of 50 participants in each group. The sample data shows that the mean weight loss for Group 1 is $\bar{x}_1 = 15.2$ pounds with a standard deviation of $s_1 = 4.5$ pounds. For Group 2, the mean weight loss is $\bar{x}_2 = 13.4$ pounds with a standard deviation of $s_2 = 5.1$ pounds. The researcher constructs a 95% confidence interval for the difference in mean weight loss between the two groups as $(0.3, 3.3)$. Based on this confidence interval, can the researcher claim that there is a significant difference in the effectiveness of the two diet plans?

Solution

Yes, the researcher can claim that there is a significant difference in the effectiveness of the two diet plans. Since the entire confidence interval lies above 0 (the interval is $(0.3, 3.3)$), it suggests that Diet A leads to more weight loss on average than Diet B, and this difference is statistically significant.

Problem 7.7.3 (Source: Original) — An educational researcher is comparing the average test scores of students from two different teaching methods. Group 1 consists of 40 students taught using Method X, while Group 2 consists of 35 students taught using Method Y. After the final exam, the mean score for Group 1 is $\bar{x}_1 = 78$ with a standard deviation of $s_1 = 8$, and the mean score for Group 2 is $\bar{x}_2 = 76$ with a standard deviation of $s_2 = 9$. A 95% confidence interval for the difference in mean test scores is calculated to be $(-2.4, 6.4)$. Based on this interval, can the researcher claim that one teaching method is significantly better than the other?

Solution

No, the researcher cannot claim that one teaching method is significantly better than the other. Since the confidence interval $(-2.4, 6.4)$ includes 0, it suggests that the difference in average test scores between the two groups could be due to random variation, and there is no statistically significant difference between the two teaching methods.

§7.8 Setting Up a Test for the Difference of Two Population Means

The process of setting up and carrying out a hypothesis test for the difference of two population means is very similar to that for a single population mean. In this section, we will outline the key concepts and steps involved in conducting a test for the difference of two means.

Definition 7.8.1 (Null and Alternative Hypotheses for Two Means)

The null hypothesis for comparing two population means states that there is no difference between the means of the two populations. This can be written as:
$H_0 : \mu_1 - \mu_2 = 0$
The alternative hypothesis will depend on the context of the problem. It can take one of three forms:

- Right-tailed: $H_a : \mu_1 - \mu_2 > 0$ (if you believe that the first population mean is greater than the second)
- Left-tailed: $H_a : \mu_1 - \mu_2 < 0$ (if you believe that the first population mean is less than the second)
- Two-tailed: $H_a : \mu_1 - \mu_2 \neq 0$ (if you are testing whether the two means are simply different)

In the context of comparing two means, we often ask whether the observed difference between the sample means, $\bar{x}_1 - \bar{x}_2$, could have occurred by random chance if the true difference in population means is zero. If the observed difference is large enough, we may reject the null hypothesis in favor of the alternative hypothesis.

Note 7.8.2 (Setting up test for difference between population means)

Here are the steps for setting up a test for the difference between population means

1. State the Hypotheses: Clearly define the null and alternative hypotheses. For example, if we are comparing the average heights of men and women, our null hypothesis might be $H_0 : \mu_{\text{men}} - \mu_{\text{women}} = 0$, while our alternative hypothesis might be $H_a : \mu_{\text{men}} - \mu_{\text{women}} \neq 0$.
2. Name the inference test
3. Check Conditions for Inference:
   - Random: Ensure that both samples are randomly selected from their respective populations.
   - Normal: Each sample should come from a population that is approximately normal, or the sample sizes should be large enough ($n \geq 30$) to invoke the Central Limit Theorem.
   - Independent: The two samples should be independent of each other, and each sample size should be less than 10% of the respective population.

Problem 7.8.3 (Source: 2018 AP Statistics FRQ Problem 4) — The anterior crucial ligament (ACL) is one of the ligaments that help stabilize the knee. Surgery is often recommended if the ACL is completely torn, and recovery time from the surgery can be lengthy. A medical center developed a new surgical procedure designed to reduce the average recovery time from the surgery. To test the effectiveness of the new procedure, a study was conducted in which 210 patients needing surgery to repair a torn ACL were randomly assigned to receive either the standard procedure or the new procedure.

(b) Summary statistics on the recovery times from the surgery are shown in the table below. Do the data provide convincing statistical evidence that those who receive the new procedure will have less recovery time from the surgery, on average, than those who receive the standard procedure, for patients similar to those in the study?

Type of Procedure	Sample Size	Mean Recovery Time (days)	Standard Deviation Recovery Time (days)
Standard	110	217	34
New	100	186	29

State: Let μ_n, μ_s be the true mean recovery time among all patients receiving the new and standard treatments, respectively.
$H_0 : \mu_s = \mu_n$
$H_a : \mu_s > \mu_n$

Plan: We will conduct a two-sample t-test for a difference between means.
✓Random - Each individual was randomly selected and assigned to one of the two treatments
✓Normal - The sample size for both samples is greater than 30.
✓Independent - We can assume that there are at least 2100 patients needing surgery to repair a torn ACL.

§7.9 Carrying Out a Test for the Difference of Two Population Means

When Carrying out a test for the difference of two population means, we follow a similar structured process as we do for all other inference tests.

Note 7.9.1 (Steps for conducting the test)

Follow these steps for conducting the test:

1. Using the appropriate formulas, obtain the corresponding t-statistic
2. Use your calculator to obtain the p-value
3. Draw a conclusion based on the p-value. If the p-value is below the significance level, reject the null hypothesis. Otherwise, there is not statistically significant evidence to do so.

Theorem 7.9.2 (test statistic for the difference of two means)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

Note 7.9.3

In most cases on the AP exam, $\mu_1 - \mu_2 = 0$ since the null hypothesis virtually always states that there is no difference between the true means of the populations. Thus, you don't usually have to write the $\mu_1 - \mu_2$ on the AP exam.

This formula should seem familiar, as it follows the same structure as the formulas to calculate the test statistics for other inference tests. The numerator calculates the raw distance between the observed difference between the two means and the hypothesized difference between the two means, while the denominator calculates the standard error of the distribution of the difference between the two samples. As you can see, the test statistic is a numerical value that measures how far your sample data deviates from the null hypothesis in units of Standard Errors. It is a standardized value that allows you to compare your sample result with the expected result under the null hypothesis.

Now, we continue the problem from the previous section.

Problem 7.9.4 (Source: 2018 AP Statistics FRQ Problem 4) — The anterior crucial ligament (ACL) is one of the ligaments that help stabilize the knee. Surgery is often recommended if the ACL is completely torn, and recovery time from the surgery can be lengthy. A medical center developed a new surgical procedure designed to reduce the average recovery time from the surgery. To test the effectiveness of the new procedure, a study was conducted in which 210 patients needing surgery to repair a torn ACL were randomly assigned to receive either the standard procedure or the new procedure.

(b) Summary statistics on the recovery times from the surgery are shown in the table below. Do the data provide convincing statistical evidence that those who receive the new procedure will have less recovery time from the surgery, on average, than those who receive the standard procedure, for patients similar to those in the study?

Type of Procedure	Sample Size	Mean Recovery Time (days)	Standard Deviation Recovery Time (days)
Standard	110	217	34
New	100	186	29

Do:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \; df = n - 1 = 99$$

$$t = \frac{(217 - 186) - (0)}{\sqrt{\frac{34^2}{110} + \frac{29^2}{100}}} = 7.13$$

By using our calculator, we obtain a p-value of about $8.308 \times 10^{-}11$.

Conclude: Since our P-value of $8.308 \times 10^{-}11$ is below the significance level of $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence proving that for patients similar to those in the study, those who receive the new procedure will have less recovery time from the surgery, on average, than those who receive the standard procedure.

Problem 7.9.5 (Source: 2011 AP Statistics FRQ Problem 4) — High cholesterol levels in people can be reduced by exercise, diet, and medication. Twenty middle-aged males with cholesterol readings between 220 and 240 milligrams per deciliter (mg/dL) of blood were randomly selected from the population of such male patients at a large local hospital. Ten of the 20 males were randomly assigned to group A, advised on appropriate exercise and diet, and also received a placebo. The other 10 males were assigned to group B, received the same advice on appropriate exercise and diet, but received a drug intended to reduce cholesterol instead of a placebo. After three months, posttreatment cholesterol readings were taken for all 20 males and compared to pretreatment cholesterol readings. The tables below give the reduction in cholesterol level (pretreatment reading minus posttreatment reading) for each male in the study.

Group A (placebo):

Reduction (in mg/dL)	2	19	8	4	12	8	17	7	24	1

Mean Reduction: 10.20 Standard Deviation of Reduction: 7.66

Group B (Cholesterol drug):

Reduction (in mg/dL)	30	19	18	17	20	−4	23	10	9	22

Mean Reduction: 16.40 Standard Deviation of Reduction: 9.40

Do the data provide convincing evidence, at the = 0 01 . level, that the cholesterol drug is effective in producing a reduction in mean cholesterol level beyond that produced by exercise and diet?

Solution

State: Let μ_A, μ_B represent the Mean Reduction in cholesterol level for Group A and Group B respectively.
$H_0 : \mu_A - \mu_B = 0$
$H_a : \mu_A - \mu_B < 0$

Plan: We will conduct a two-sample t-test for the difference between two means.
✓Random - The 20 males were randomly selected from the population of interest, and randomly assigned to either Group A or Group B.
✓Normal - Looking at the table, there does not appear to be any major skewness in the data, so we may reasonably assume that the distribution is roughly normal.

✓Independent - We can assume that there are at least 200 middle-aged males with cholesterol readings between 220 and 240 milligrams per deciliter.

Do:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},\ df = 10 - 1 = 9$$

$$t = \frac{(16.4 - 10.2) - (0)}{\sqrt{\frac{7.66^2}{10} + \frac{9.4^2}{10}}} = \frac{6.2}{3.83} = 1.62$$

Using our calculator, we obtain the corresponding p-value of about 0.0698

Conclude: Since our P-value of 0.0698 is above the significance level of $\alpha = 0.01$, we fail to reject the null hypothesis. There is no statistically significant evidence proving that the cholesterol drug is effective in producing a reduction in mean cholesterol level beyond that produced by exercise and diet for middle-aged men similar to those in the study.

§7.10 Unit 7 Practice Problems

Problem 7.10.1 (Source: 2014 AP Statistics FRQ Problem 5) — A researcher conducted a study to investigate whether local car dealers tend to charge women more than men for the same car model. Using information from the county tax collector's records, the researcher randomly selected one man and one woman from among everyone who had purchased the same model of an identically equipped car from the same dealer. The process was repeated for a total of 8 randomly selected car models.

The purchase prices and the differences (woman – man) are shown in the table below. Summary statistics are also shown.

Car model	20100	17400	22300	32500	177100	21500	29600	46300
Car model	1	2	3	4	5	6	7	8
Car model	1	2	3	4	5	6	7	8
Car model	1	2	3	4	5	6	7	8

	Mean	Standard Deviation
Women	$\$25,926.25$	$\$9,846.61$
Men	$\$25,341.25$	$\$9,728.60$
Difference	585.00	530.71

Dotplots of the data and the differences are shown below. Do the data provide convincing evidence that, on average, women pay more than men in the county for the same car model?

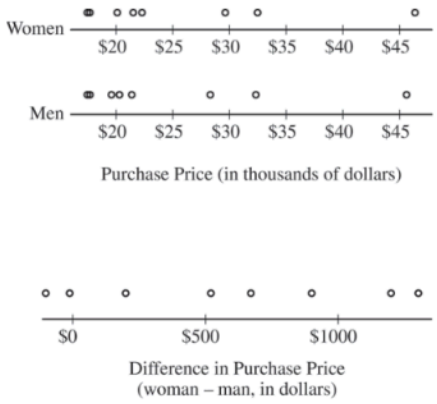


Figure 7.1: Caption

Solution

State: Let μ_d represent the mean difference (women - men) in prices for the same car model.

$H_0 : \mu_d = 0$

$H_a : \mu_d > 0$

Plan: We will conduct a paired t-test for the mean difference in prices paid by women and men for the same car model.

✓Random - The researcher randomly selected one man and one woman for each of the 8 car models.

✓Normal - The dotplot of the differences does not show strong skewness or outliers. With only 8 pairs, the sample size is small, but the data appear to be reasonably symmetric.
✓Independent - The differences between prices for women and men for each car model are assumed to be independent of each other.

Do:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},\ df = n - 1 = 7$$

$$t = \frac{585 - 0}{\frac{530.71}{\sqrt{8}}} = \frac{585}{187.69} = 3.12$$

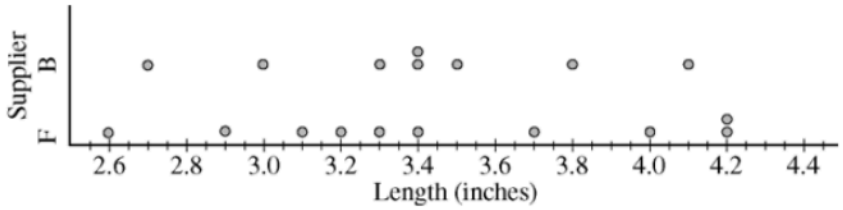
Using a calculator, we obtain a p-value of approximately 0.008.

Conclude: Since the p-value of 0.008 is less than the significance level $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence that, on average, women pay more than men for the same car model in the county.

Problem 7.10.2 (Source: 2010 AP Statistics FRQ Problem 5) — A large pet store buys the identical species of adult tropical fish from two different suppliers—Buy-Rite Pets and Fish Friends. Several of the managers at the pet store suspect that the lengths of the fish from Fish Friends are consistently greater than the lengths of the fish from Buy-Rite Pets. Random samples of 8 adult fish of the species from Buy-Rite Pets and 10 adult fish of the same species from Fish Friends were selected and the lengths of the fish, in inches, were recorded, as shown in the table below.

	Length of Fish	Mean	S.D.
Bur-Rite Pets ($n_B = 8$)	3.4 2.7 3.3 4.1 3.5 3.4 3.0 3.8	3.40	0.434
Fish Friends ($n_F = 10$)	3.3 2.9 4.2 3.1 4.2 4.0 3.4 3.2 3.7 2.6	3.46	0.550

Do the data provide convincing evidence that the mean length of the adult fish of the species from Fish Friends is greater than the mean length of the adult fish of the same species from Buy-Rite Pets?



Solution

State: Let μ_F represent the mean length of fish from Fish Friends, and let μ_B represent the mean length of fish from Buy-Rite Pets. $H_0 : \mu_F = \mu_B$, $H_a : \mu_F > \mu_B$

Plan: We will conduct a two-sample t-test for the difference between means.
✓Random - The samples were randomly selected from the populations of interest.
✓Normal - The sample sizes are small ($n_B = 8$, $n_F = 10$), but the dotplots of the data suggest the fish lengths are reasonably symmetric without strong skewness or outliers.
✓Independent - The fish lengths from Buy-Rite Pets and Fish Friends are independent of each other.

Do:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \; df = n - 1 = 7$$

$$t = \frac{(3.46 - 3.40) - 0}{\sqrt{\frac{0.434^2}{8} + \frac{0.550^2}{10}}} = \frac{0.06}{0.23198} = 0.259$$

Using a calculator, we obtain a p-value of approximately 0.40.

Conclude: Since the p-value of 0.40 is greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. There is not enough evidence to conclude that the mean length of the fish from Fish Friends is greater than that of the fish from Buy-Rite Pets.

Problem 7.10.3 (2009 AP Statistics FRQ Problem 4) — One of the two fire stations in a certain town responds to calls in the northern half of the town, and the other fire station responds to calls in the southern half of the town. One of the town council members believes that the two fire stations have different mean response times. Response time is measured by the difference between the time an emergency call comes into the fire station and the time the first fire truck arrives at the scene of the fire. Data were collected to investigate whether the council member's belief is correct. A random sample of 50 calls selected from the northern fire station had a mean response time of 4.3 minutes with a standard deviation of 3.7 minutes. A random sample of 50 calls selected from the southern fire station had a mean response time of 5.3 minutes with a standard deviation of 3.2 minutes.

(a) Construct and interpret a 95 percent confidence interval for the difference in mean response times between the two fire stations.

(b) Does the confidence interval in part (a) support the council member's belief that the two fire stations have different mean response times? Explain.

Solution

(a) **State:** Let μ_N represent the mean response time for the northern fire station and μ_S represent the mean response time for the southern fire station. We will conduct a 95% confidence interval for the difference in mean response times between the two fire stations $(\mu_N - \mu_S)$.

Plan: We construct a two-sample t-interval for the difference in population means.
✓Random: The samples were randomly selected from both fire stations.
✓Normal: The sample sizes are both greater than 30.
✓Independent: The samples from the northern and southern fire stations are independent (neither affects the other). Furthermore, we can assume that each sample is at most 10% of their respective populations.

Do:

$$\text{Confidence interval: } (\bar{x_1} - \bar{x_2}) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ df = n - 1 = 49$$

$$\text{Confidence interval: } (4.3 - 5.3) \pm 2.010 \times \sqrt{\frac{3.7^2}{50} + \frac{3.2^2}{50}}$$

$$\text{Confidence interval: } -1 \pm 1.39$$

Conclude: We are 95% confident that the true difference in mean response times between the northern and southern fire stations is between -2.39 minutes and 0.39 minutes.

(b) The confidence interval $(-2.39, 0.39)$ includes 0, which means that we do not have convincing evidence that the mean response times are different between the two fire stations. Therefore, the confidence interval does not support the council member's belief that the two fire stations have different mean response times.

Problem 7.10.4 (Source: 2009 AP Statistics FRQ Problem 5, Form B) — A bottle-filling machine is set to dispense 12.1 fluid ounces into juice bottles. To ensure that the machine is filling accurately, every hour a worker randomly selects four bottles filled by the machine during the past hour and measures the contents. If there is convincing evidence that the mean amount of juice dispensed is different from 12.1 ounces or if there is convincing evidence that the standard deviation is greater than 0.05 ounce, the machine is shut down for recalibration. It can be assumed that the amount of juice that is dispensed into bottles is normally distributed.

During one hour, the mean number of fluid ounces of four randomly selected bottles was 12.05 and the standard deviation was 0.085 ounce.

(a) Perform a test of significance to determine whether the mean amount of juice dispensed is different from 12.1 fluid ounces. Assume the conditions for inference are met.

(b)To determine whether this sample of four bottles provides convincing evidence that the standard deviation of the amount of juice dispensed is greater than 0.05 ounce, a simulation study was performed. In the simulation study, 300 samples, each of size 4, were randomly generated from a normal population with a mean of 12.1 and a standard deviation of 0.05. The sample standard deviation was computed for each of the 300 samples. The dotplot below displays the values of the sample standard deviations. Use the results of this simulation study to explain why you think the sample provides or does not provide evidence that the standard deviation of the juice dispensed exceeds 0.05 fluid ounce.

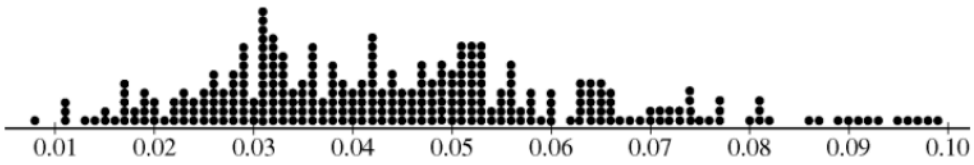


Figure 7.2: Caption

Solution

(a) **State:** We will conduct a significance test to determine whether the mean amount of juice dispensed is different from 12.1 fluid ounces.
$H_0 : \mu = 12.1$
$H_a : \mu \neq 12.1$

Plan: We will conduct a one-sample t-test for the population mean.
✓Random: The bottles were randomly selected.
✓Normal: We are told that the amount of juice dispensed follows a normal distribution.
✓Independent: We are told to assume that the individual samples are independent from each other.

Do:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}, \; df = n - 1 = 4 - 1 = 3$$

$$t = \frac{12.05 - 12.1}{\frac{0.085}{\sqrt{4}}} = \frac{-0.05}{\frac{0.085}{2}} = \frac{-0.05}{0.0425} = -1.18$$

Using our calculator, we obtain the associated p-value for the **TWO-SIDED** t-test, which is about 0.314.

Conclude: Since the p-value is 0.314, which is greater than $\alpha = 0.05$, we fail to reject the null hypothesis. There is no statistically significant evidence to conclude that the mean amount of juice dispensed is different from 12.1 ounces.

(b) Out of the 300 simulated sample standard deviations, only 12 had sample standard deviations greater than or equal to 0.85ounces. This gives us a p-value of $\frac{12}{300} = 0.04$, assuming that the null hypothesis is true (that the true population standard deviation is in fact 0.05). At the significance level of $\alpha = 0.05$, we would reject the null hypothesis. Thus, the sample from part (a) provides convincing statistical evidence that the standard deviation of the juice dispensed exceeds 0.05 fluid ounce.

Problem 7.10.5 (Source: Original) — A team of botanists conducted an experiment to determine whether the type of music played to plants affects their growth rate. They selected two groups of identical plants: Group A was exposed to classical music, and Group B was exposed to heavy metal music. Both groups were kept under identical conditions except for the type of music they heard for 8 hours per day. After a month, the increase in height (in centimeters) of each plant was recorded. The following summary statistics were calculated:

Group	Mean Growth (cm)	Standard Deviation (cm)	Sample Size
Group A	12.3	2.1	47
Group B	10.1	2.5	33

Researchers want to investigate whether the plants exposed to classical music grew more than those exposed to heavy metal music.

(a) State appropriate hypotheses to test whether plants exposed to classical music grew more than those exposed to heavy metal music.

(b) Construct a 95% confidence interval for the difference in mean growth between plants exposed to classical music and those exposed to heavy metal music. Interpret the interval in the context of the problem.

(c) Based on your confidence interval in part (b), does it appear that the type of music has a significant effect on plant growth? Justify your answer.

Solution

(a) Let μ_A be the mean growth (in cm) of plants exposed to classical music. Let μ_B be the mean growth (in cm) of plants exposed to heavy metal music.
$H_0 : \mu_A - \mu_B = 0$ (no difference in mean growth between the two groups)
$H_a : \mu_A - \mu_B > 0$ (plants exposed to classical music grew more on average than those exposed to heavy n

(b) **State:** We will conduct a 95% percent confidence interval for the difference in mean growth between plants exposed to classical music and those exposed to heavy metal music.

Plan: We will conduct a two-sample z-interval for the difference between sample means. First, we check the conditions for inference.
✓Random - Both samples were randomly sampled from the population of interest
✓Normal - Both sample sizes are greater than 30, so the normality condition is satisfied.
✓Independent - We may assume that the sample size is at most 10% of the population.

Do:

$$\text{Confidence Interval} : b \pm (t^*)(SE_b),\ df = 30 - 2 = 28$$

$$\text{Confidence Interval} : -0.85 \pm (1.701)(0.22)$$

$$\text{Confidence Interval} : (-1.224, -0.4758)$$

Conclude: We are 95% confident that the true difference in mean growth between plants exposed to classical music and those exposed to heavy metal music lies in the interval $(-1.224, -0.4758)$.

(c) Since the entire confidence interval is negative, this suggests that the mean growth of plants exposed to heavy metal music is greater than the mean growth of plants exposed to classical music. Therefore, the plants exposed to classical music did not grow significantly more than those exposed to heavy metal music.

Problem 7.10.6 (Source: 2018 AP Statistics FRQ Problem 6) — Systolic blood pressure is the amount of pressure that blood exerts on blood vessels while the heart is beating. The mean systolic blood pressure for people in the United States is reported to be 122 millimeters of mercury (mmHg) with a standard deviation of 15mmHg.

The wellness department of a large corporation is investigating whether the mean systolic blood pressure of its employees is greater than the reported national mean. A random sample of 100 employees will be selected, the systolic blood pressure of each employee in the sample will be measured, and the sample mean will be calculated.

Let μ represent the mean systolic blood pressure of all employees at the corporation. Consider the following hypotheses.

$$H_0 : \ \mu = 122, \ H_a : \ \mu > 122$$

(a) Describe a Type II error in the context of the hypothesis test.

(b) Assume that σ, the standard deviation of the systolic blood pressure of all employees as the corporation, is 15mmHg and a standard deviation of 1.5mmHg. What values of the sample mean $\bar{x}$ would represent sufficient evidence to reject the null hypothesis at the significance level of $\alpha = 0.05$

The actual mean systolic blood pressure of all employees at the corporation is 125mmHg, not the hypothesized value of 122mmHg, and the standard deviation is 15mmHg.

(c) Using the actual mean of 125mmHg and the results from part (b), determine the probability that the null hypothesis will be rejected.

(d) What statistical term is used for the probability found in part (c)?

(e) Suppose the size of the sample of employees to be selected is greater than 100. Would the probability of rejecting the null hypothesis be greater than, less than, or equal to the probability calculated in part (c)? Explain your reasoning.

Solution

(a) A Type II error occurs when we fail to reject the null hypothesis H_0 when in fact the alternative hypothesis H_a is true. In this context, a Type II error would mean concluding that the mean systolic blood pressure of the employees is not greater than 122 mmHg, when in reality, it is greater.

(b) We are performing a one-sample z-test for the mean with the following hypotheses:

$$H_0 : \mu = 122, \quad H_a : \mu > 122$$

We need to determine the values of the sample mean $\bar{x}$ that would provide sufficient evidence to reject H_0 at $\alpha = 0.05$. Given that the population standard deviation $\sigma = 15$

mmHg, and the sample size is $n = 100$, the standard deviation of the sample mean is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5\,\text{mmHg}$$

Using the standard normal distribution and a significance level of $\alpha = 0.05$ for a one-tailed test, the critical value for z is:

$$z_{0.05} = 1.645$$

The rejection region corresponds to values of $\bar{x}$ for which the z-test statistic is greater than 1.645. The z-test statistic is given by:

$$z = \frac{\bar{x} - 122}{1.5}$$

Setting $z = 1.645$ to find the critical value of $\bar{x}$:

$$1.645 = \frac{\bar{x} - 122}{1.5}$$

Solving for $\bar{x}$:

$$\bar{x} = 1.645 \times 1.5 + 122 = 2.4675 + 122 = 124.47$$

Thus, we will reject H_0 if $\bar{x} > 124.47$.

(c)

$$z = \frac{124.47 - 125}{1.5} = \frac{-0.53}{1.5} = -0.353$$

Using the standard normal distribution, the probability of obtaining a z-value greater than -0.353 is:

$$P(z > -0.353) = 0.638$$

Thus, the probability of rejecting H_0 when $\mu = 125$ mmHg is approximately 0.638.

(d) The probability calculated in part (c) is known as the **power** of the test.

(e) If the sample size were greater than 100, the probability of rejecting the null hypothesis would increase. This is because increasing the sample size decreases the standard error of the sample mean, making it easier to detect differences between the sample mean and the hypothesized mean. In other words, the test becomes more sensitive, and the power of the test increases as the sample size increases.

Problem 7.10.7 (Source: 2000 AP Statistics FRQ Problem 4) — Baby walkers are seats hanging from frames that allow babies to sit upright with their legs dangling and feet touching the floor. Walkers have wheels on their legs that allow the infant to propel the walker around the house long before he or she can walk or even crawl. Typically, babies use walkers between the ages of 4 months and 11 months.

Because most walkers have tray tables in front that block babies' views of their feet, child psychologists have begun to question whether walkers affect infants' cognitive development. One study compared mental skills of a random sample of those who used walkers with a random sample of those who never used walkers. Mental skill scores averaged 113 for 54 babies who used walkers (standard deviation of 12) and 123 for 55 babies who did not use walkers (standard deviation of 15).

(a) Is there evidence that the mean mental skill score of babies who use walkers is different from the mean mental skill score of babies who do not use walkers? Explain your answer.

(b) Suppose that a study using this design found a statistically significant result. Would it be reasonable to conclude that using a walker causes a change in mean mental skill score? Explain your answer.

Solution

(a) **State:** Let μ_1 represent the mean mental skill score of babies who use walkers, μ_2 represent the mean mental skill score of babies who do not use walkers.
$H_0 : \mu_1 = \mu_2$
$H_a : \mu_1 \neq \mu_2$.

Plan: We will perform a two-sample t-test for the difference in means. The following conditions must be checked for inference:
✓Random - Both samples of babies were random.
✓Normal - Both samples are larger than 30, so we satisfy the normality condition
✓Independent: We may assume that both samples are at most 10% of their respective populations.

Do:

$$t == \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},\ df = 54 - 1 = 53$$

$$t = \frac{(113 - 123)}{2.6} = \frac{-10}{2.6} = -3.85$$

Using our calculator, we obtain the p-value of 0.00032

Conclude: Since the p-value is 0.00032, which is extremely small, we reject the null hypothesis. There is statistically significant evidence to conclude that the mean mental skill score of babies who use walkers is different from the mean mental skill score of babies who do not use walkers.

(b) Even if a statistically significant result is found, it is not reasonable to conclude that using a walker causes a change in the mean mental skill score. The study is observational, meaning that other factors may be confounding the relationship between walker use and

mental skill scores. For example, there could be other variables such as socioeconomic status, parenting style, or general physical activity that affect both walker use and mental skill scores.

8 Unit 8: Inference for Categorical Data: Chi-Square

§8.1 Introducing Statistics: Are My Results Unexpected?

Back in Unit 6, we learned how to make inferences relating to a single population proportion, and the difference between two population proportions. But what if instead, we wanted to make inferences about the distribution of categories for categorical data?

Note 8.1.1 (Types of Chi-Square tests)

Luckily, we have three different types of tests that help us achieve this

1. **Chi-Square goodness of fit test** - This test shows how closely a sample distribution of categorical data resembles a hypothesized distribution to test whether it is valid.
2. **Chi-Squared test for Homogeneity** - This test shows how similar the distribution is for 1 variable across different populations.
3. **Chi-Squared test for Independence** - This test allows us to see the strength of the association between two variables within a population

Due to the randomness of our sampling, the distributions of our data will not look exactly as how we predicted it to. These tests aim to determine whether or not this is because of the variability with random sampling, or if the expected distribution is not plausible. Each of these tests will be explored in depth later in their respective sections. For now, just keep the following formula in the back of your mind, as it is the only formula used in this unit.

Theorem 8.1.2 (Chi-Square statistics)

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

Where χ^2 is the Chi-Square statistic, O is the observed count, and E is the expected count for a certain category.

We end this section by introducing the conditions that must be met to conduct any of the Chi-Square tests.

Note 8.1.3 (Conditions for Chi-Square tests)

Thankfully, there are only three:

- **Random** - Each individual in the sample must be selected at random
- **Large Counts** - The expected counts for each category must be at least 5
- **Independence** - Each individual in the sample must be selected independently from each other (either with replacement, or 10% rule)

§8.2 Setting Up a Chi-Square Goodness of Fit Test

The Chi-Square Goodness of Fit test determines if a **single** categorical variable follows a specific distribution across various categories. It allows us to compare observed data to expected proportions to see if they fit a hypothesized distribution.

Note 8.2.1 (Steps for setting up a Chi-Square Goodness of fit test)

Follow these steps to set up a Chi-Square Goodness of fit test:

1. Define the Hypotheses
   - **Null Hypothesis** (H_0): The observed distribution follows the expected proportions.
   - **Alternative Hypothesis** (H_a): The observed distribution does **not** follow the expected proportions.
2. Calculate the Expected Counts
   - For each category, calculate the expected count by multiplying the total sample size by the expected proportion for that category.

$$\text{Expected count} = (\text{Total Sample Size}) \times (\text{Expected Proportion})$$

3. Verify Conditions for the Test
   - **Random Sample**: Ensure that the sample is randomly selected.
   - **Large Counts**: All expected counts should be at least 5.
   - **Independent**: The population should be at least 10 times the sample size.

Problem 8.2.2 (Source: Original) — Let's imagine that you read a study that said the distribution of favorite colors among high school students should be: 25% choose Red, 40% choose Blue, 5% choose Green, and 30% choose Other. To test this, you ask 100 randomly selected high school students and gather the data in the table below.

	RED	BLUE	GREEN	OTHER	TOTAL
OBSERVED COUNT	31	28	6	35	100

Using this data, perform the appropriate inference test at the significance level $\alpha = 0.05$ to determine whether the observed distribution of favorite colors is consistent with the expected distribution from the study.

Solution

State: H_0: The distribution of favorite colors for High School Students is consistent with the expected distribution from the study.
H_a: The distribution of Favourite colors for High School Students is **NOT** consistent with the expected distribution from the study.

Plan: We conduct a Chi-Square Goodness of fit test. First, we check that the conditions of inference are satisfied.

	RED	BLUE	GREEN	OTHER	TOTAL
OBSERVED COUNT	31	28	6	35	100
EXPECTED COUNT	25	40	5	30	100

✓Random - The sample is a random sample from the population of interest.
✓Large Counts - The expected count for each category is at least 5.
✓Independent - We may assume that there are at least 1000 high school students.

§8.3 Carrying Out a Chi-Square Test for Goodness of Fit

In this section, we use our Chi-Square statistic to determine whether or not the hypothesized distribution is valid or not. Intuitively, the Chi-Square statistic measures how far each observed count is from the expected count, relative to the expected count. Thus, the higher the Chi-Square statistic, the less the observed distribution resembles the hypothesized distribution. We can then use this statistic to get an associated p value, and if this p value is below our pre-determined significance level, we can safely reject our null hypothesis.

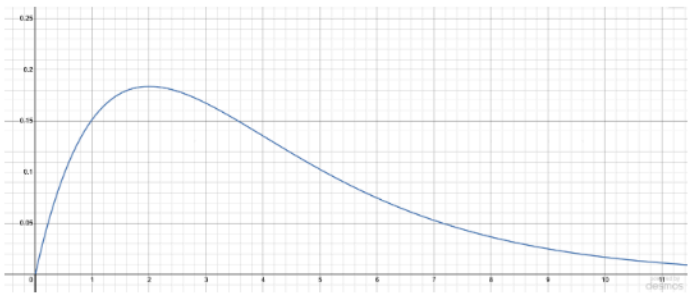


Figure 8.1: Chi-Square distribution for $df = 4$

Notice that unlike the **Z** and **T** distribution, the Chi-Square distribution is not symmetric,

but rather skewed to the right. This makes sense, since the Chi-Squared statistic is always positive (this should be obvious from the formula!). When conducting Chi-Square tests, if our Chi-Square statistic is to the ***right*** of our pre-determined significance level, that means that our observed distribution is too far from our hypothesized distribution. In this case we must reject the null hypothesis.

It is also important to note that the specific Chi-Square distribution varies based on the number of degrees of freedom (As degrees of freedom increases, the distribution becomes less skewed).

Theorem 8.3.1 (Degrees of Freedom for Goodness of Fit test)

$$\text{Degrees of freedom} = c - 1$$

Where c is the number of columns/categories.

Proof If we know the the counts for each category except for one, as well as the total sample size, we may easily find the count for the last category (one equation, one unknown).

Now, we are ready to continue the example problem from the last section.

Problem 8.3.2 (Source: Original) — Let's imagine that you read a study that said the distribution of favorite colors among high school students should be: 25% choose Red, 40% choose Blue, 5% choose Green, and 30% choose Other. To test this, you ask 100 randomly selected high school students and gather the data in the table below.

	RED	BLUE	GREEN	OTHER	TOTAL
OBSERVED COUNT	31	28	6	35	100

Using this data, perform the appropriate inference test at the significance level $\alpha = 0.05$ to determine whether the observed distribution of favorite colors is consistent with the expected distribution from the study.

Do:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E},\ df = c - 1 = 3$$

$$\chi^2 = 1.44 + 3.6 + 0.2 + 0.83$$

$$\chi^2 = 6.073$$

Using our calculator, we obtain a P-value of 0.1081.

Conclude: Since our P-value of 0.1081 is above the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. There is no significant evidence proving that the distribution of favorite colors among high school students does **NOT** match with our hypothesized distribution. This does not mean that our hypothesized distribution is necessarily true, we are just not able to prove that it isn't. Even though our observed counts did not match the hypothesized distribution perfectly, it wasn't that far off.

Problem 8.3.3 (Source: 2008 AP Statistics FRQ Problem 5) — A study was conducted to determine where moose are found in a region containing a large burned area. A map of the study area was partitioned into the following four habitat types.

1. Inside the burned area, not near the edge of the burned area,
2. Inside the burned area, near the edge,
3. Outside the burned area, near the edge, and
4. Outside the burned area, not near the edge.

The proportion of total acreage in each of the habitat types was determined for the study area. Using an aerial survey, moose locations were observed and classified into one of the four habitat types. The results are given in the table below.

Habitat Type	Proportion of Total Acreage	Number of Moose Observed
1	0.340	25
2	0.101	22
3	0.104	30
4	0.455	40
Total	1.000	117

(a) The researchers who are conducting the study expect the number of moose observed in a habitat type to be proportional to the amount of acreage of that type of habitat. Are the data consistent with this expectation? Conduct an appropriate statistical test to support your conclusion. Assume the conditions for inference are met.

(b) Relative to the proportion of total acreage, which habitat types did the moose seem to prefer? Explain.

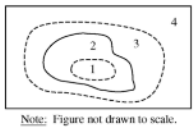


Figure 8.2: The four habitat types

Solution

(a) **State**:

H_0 : The number of moose observed in a habitat type is proportional to the amount of acreage of that type of habitat

H_a : The number of moose observed in a habitat type is **NOT** proportional to the amount of acreage of that type of habitat

Habitat Type	Expected Number of Moose	Number of Moose Observed
1	39.78	25
2	11.82	22
3	12.17	30
4	53.24	40
Total	117	117

Plan: We will conduct a Chi-Square goodness-of-fit test. The problem states that all conditions for inference are met (Random, Large Counts, Normal).

Do:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \ df = 4-1 = 3$$

$$\chi^2 = 5.491 + 8.768 + 26.122 + .293$$

$$\chi^2 = 43.67$$

Using our calculator, we get our corresponding p-value of about 1.756×10^{-9}

Conclude: Since our p-value of 1.756×10^{-9} is less than our significance value of $\alpha = 0.05$ we must reject our null hypothesis. Thus, the data is not consistent with the researcher's expectations. There is strong evidence that moose have a preference for habitat type.

(b) The moose appeared to prefer Habitat Types 2 and Habitat Type 3. A higher proportion of moose were observed in these two habitats relative to the proportion of total acreage. On the other hand, Habitat Types 1and Habitat Type 4 saw fewer moose than expected based on their acreage, indicating that these areas were less desirable for the moose. In summary, moose seemed to prefer areas near the edge, both inside and outside the burned region.

§8.4 Expected Counts in Two-Way Tables

A Two-way table is a table that contains **TWO** categorical variables instead of one. The next two Chi-Square tests will be involving Two-way tables. Recall from the Chi-Square formula that you need both the observed and expected counts in each category for the computation. You will always be given the observed counts, however you will often have to compute the expected counts on your own. To find the expected counts, we first need to have a good understanding of what the Chi-Square test for Homogenity and Independence are doing.

In a Chi-Square test for Homogenity, we are seeing if there is evidence of a significant difference in the distributions of a categorical variable across **MANY** different populations.

In a Chi-Square test for Independence (sometimes called Chi-Square test for Homogenity) we are seeing if there is evidence that **two categorical variables are associated** (that is, the distribution of one variable affects the distribution of another).

When calculating expected counts, we assume the null hypothesis is true, and work from there. For the Chi-Square test for Homogenity, this means assuming that there

is no difference in the distributions for the variable across different populations. For the Chi-Square test for Independence, this means assuming that there is no association between the two categorical variables. Luckily, the formula for expected count on both these tables are the same!

Theorem 8.4.1 (Expected counts of two-way tables)

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

Proof

We present the proof of the expected count formula of the Chi-Squared test for Homogenity and leave the proof for the other test to the reader. Choose an arbitrary row i and column j in the table, and label the row total and column total X_i and Y_j respectively. Let the total number of observations be T, and let our expected count of the desired square be E. Since we are assuming the null hypothesis is true, we assume that the distribution of our categorical variable is the same across all populations, Thus, we obtain the following equation:

$$\frac{Y_j}{T} = \frac{E}{X_i}$$

$$E = \frac{(X_i)(Y_i)}{T}$$

Problem 8.4.2 (Source: Original) — A high school principal wants to determine whether there is an association between students' grade level and their preferred mode of transportation to school (Bus, Car, Walk/Bike). The principal collects data from a random sample of 200 students. The results are summarized in the following two-way table:

	Bus	Car	Walk/Bike	Total
9th Grade	30	20	10	60
10th Grade	25	30	5	60
11th Grade	15	25	10	50
12th Grade	10	15	5	30
Total	80	90	30	200

(a) Calculate the expected count for 9th-grade students who prefer to take the bus.
(b) Calculate the expected count for 10th-grade students who prefer to Walk/Bike.
(c) Calculate the expected count for 12th-grade students who prefer to take the Car.

Each part of the question will require the formula: expected count $= \frac{(\text{row total})(\text{column total})}{\text{table total}}$

(a) expected count $= \frac{(80)(60)}{200} = 24$
(b) expected count $= \frac{(30)(60)}{200} = 9$
(c) expected count $= \frac{(90)(30)}{200} = 13.5$

§8.5 Setting Up a Chi-Square Test for Homogeneity or Independence

The first step to setting up a Chi-Square Test for Homogeneity of Independence is first identifying which test is being asked of by the question. Note that if the question contains **TWO** categorical variables, it will always be either for Homogeneity or Independence, and if it contains **ONE** then it is always for Goodness-of-fit. Differentiating tests for Homogeneity and Independence can sometimes be tricky.

Note 8.5.1 (Differences between tests for Homogeneity vs. Independence)

A Chi-Square test for Homogeneity always contains **MULTIPLE POPULATIONS AND MEASURES** 1 **CATEGORICAL VARIABLE**.
A Chi-Square test for Independence always contains 1 **POPULATION AND MEASURES** 2 **CATEGORICAL VARIABLES**.

For example, a Chi-Square test for Homogeneity could measure the difference in preferences of ice cream flavours across various cities, while a Chi-Square test for Independence could measure the relationship between a student's grade level and their preferred method of studying. Once the specific type of test is determined, the set up for the test follows the same steps as for the Goodness-of-fit test.

We now must determine what the null and alternative hypotheses should look like for both tests. For a test for Homogeneity, the null hypothesis is always that there is **NO** difference between the difference in distributions of the variable across different populations (I.e. nothing special happens). The alternative hypothesis states that there **IS** a difference. For a test for Independence, the null hypothesis states that there is **NO** association between the two categorical variables in the given population. The alternative hypothesis states that there **IS** a difference (I.e. the two variables depend on each other).

After stating the name of the inference test being used, we then must check that all the conditions for inference are met. This was covered back in unit 8.1, so we will not be going over these again. After that, we will use our knowledge from section 8.4 to find all the expected counts for each category, thus finishing up our setup for the test.

Note 8.5.2 (Setting up a Chi-Square test for homogeneity or independence)

Follow these steps:

1. Identify and state the name of the appropriate inference test
2. Find the expected counts for each category
3. Check that all the conditions for inference are met

§8.6 Carrying Out a Chi-Square Test for Homogeneity or Independence

The process of carrying out a Chi-Square Test for Homogeneity or Independence is identical to the process of carrying out a Goodness-of-fit test, with the only difference

being the number of degrees of freedom.

Theorem 8.6.1 (Degrees of freedom in a Chi-Square test for homogeneity or independence)

$$\text{degrees of freedom} = (\text{num. of rows} - 1)(\text{num. of columns} - 1)$$

Proof

If you do not understand the proof for the number of degrees of freedom of a Goodness-of-fit test, please revisit that. Now, imagine that we are given the observed counts for an $(r-1)(c-1)$ rectangle, where r and c are the number of rows and columns respectively.

	row 1	row 2	row 3	Total
column 1	30	20		60
column 2	25	30		60
column 3	15	25		50
column				30
Total	80	90	30	200

For any given row or column, we can find the missing entry since we know all but 1 value, as well as the total of the values in that row or column. Thus, there are $(r-1)(c-1)$ degrees of freedom.

Note 8.6.2 (Carrying out a Chi-Square test for homogeneity or independence)

Here are the steps for carrying out a Chi-Square test for homogeneity or independence:

1. Calculate the Chi-Square statistic and number of degrees of freedom for the two-way table
2. Use your calculator to obtain the associated p-value
3. Draw a conclusion for the inference test. Either reject or fail to reject the null hypothesis

Problem 8.6.3 (Source: 2013 AP Statistics FRQ Problem 4) — The Behavioral Risk Factor Surveillance System is an ongoing health survey system that tracks health conditions and risk behaviors in the United States. In one of their studies, a random sample of 8,866 adults answered the question "Do you consume five or more servings of fruits and vegetables per day?" The data are summarized by response and by age-group in the frequency table below.

Age-Group (years)	Yes	No	Total
$18-34$	231	741	972
$35-54$	669	2242	2911
55 or older	1291	3692	4983
Total	2191	6675	8866

Do the data provide convincing statistical evidence that there is an association between age-group and whether or not a person consumes five or more servings of fruits and vegetables per day for adults in the United States?

Solution

State:

H_0 : The consumption of fruits and vegetables is independent of one's age group within the population of adults within the United States

H_a : The consumption of fruits and vegetables is **NOT** independent of one's age group within the population of adults within the United States

Plan: Since the problem contains 1 population and measures 2 categorical variables, we will conduct a Chi-Square test for Independence.

Age-Group (years)	Yes	No	Total
$18-34$	231(240.2)	741(731.8)	972
$35-54$	669(719.4)	2242(2191.6)	2911
55 or older	1291 (1231.4)	3692(3751.6)	4983
Total	2191	6675	8866

We also check that the conditions for inference are met.

✓Random - The problem states that the data was collected through a random sample

✓Independence - We may assume that there are more than 88660 adults in the United States

✓Large Counts - The expected count for each categorical variable is greater than 5

Do:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E}, df = (2-1)(3-1) = 2$$

$$\chi^2 = \frac{(240.2-231)^2}{240.2} + \frac{(731.8-741)^2}{731.8} + \frac{(719.4-669)^2}{719.4} + \frac{(2191.6-2242)^2}{2191.6} + \frac{(1231.4-1291)^2}{1231.4} + \frac{(3751.6-3692)^2}{3741.6}$$

$$\chi^2 = 8.983$$

Using our calculator, we get our corresponding p-value of about 0.011

Conclude: Since our p-value of 0.011 is less than our significance value of $\alpha = 0.05$

we must reject our null hypothesis. Thus, we conclude that the consumption of fruits and vegetables is **NOT** independent of one's age group within the population of adults within the United States.

Problem 8.6.4 (Source: Original) — A research team wants to investigate whether the distribution of preferred study methods differs between two different universities. They surveyed students at each university using a random sample, asking them to choose their primary study method from the following options: Group Study, Solo Study, Online Resources, Tutoring.

Study Method	University A	University B	Total
Group Study	40	90	130
Solo Study	50	85	135
Online Resources	35	50	80
Tutoring	25	25	50
Total	150	250	400

Determine if there is evidence to suggest that the distribution of preferred study methods differs among the two universities. Assume a significance level of $\alpha = 0.05$

Solution

State:

H_0: The distribution of preferred studying methods is the same across both universities.
H_a: The distribution of preferred studying methods is **DIFFERENT** across both universities.

Plan:

Since the problem measures 1 categorical variable and 2 populations, we conduct a Chi-Square test for homogeneity.

Study Method	University A	University B	Total
Group Study	40(48.75)	90(81.25)	130
Solo Study	50(50.63)	85(84.38)	135
Online Resources	35(30)	50(50)	80
Tutoring	25(18.75)	25(31.25)	50
Total	150	250	400

✓Random - The problem states that the data was collected through a random sample
✓Independence - We may assume that there are more than 1500 students in University A and more than 2500 students in University B.
✓Large Counts - The expected count for each categorical variable is greater than 5

Do:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \; df = (2-1)(4-1) = 3$$

$$\chi^2 = \frac{(48.75-40)^2}{48.75} + \frac{(81.25-90)^2}{81.25} + \frac{(50.63-50)^2}{50.63} + \frac{(84.38-85)^2}{84.38} + \frac{(30-35)^2}{30} + \frac{(50-50)^2}{50} + \frac{(18.75-25)^2}{18.75}$$

$$\chi^2 = 6.35$$

Using our calculator, we obtain the corresponding p-value of 0.096

Conclude:
Since our p-value of 0.096 is less than our significance value of $\alpha = 0.05$ we must reject our null hypothesis. Thus, we conclude that the distribution of preferred studying methods is **DIFFERENT** across both universities.

§8.7 Unit 8 Practice Problems

Problem 8.7.1 (Source: 2004 AP Statistics FRQ Problem 5) — A rural country hospital offers several health services. The hospital administrators conducted a poll to determine whether the residents' satisfaction with the available services depends on their gender. A random sample of 1,000 adult county residents was selected. The gender of each respondent was recorded and each was asked whether he or she was satisfied with the services offered by the hospital. The resulting data are shown in the table below.

	Male	Female	Total
Satisfied	384	416	800
Not Satisfied	80	120	200
Total	464	536	1000

(a) Using a significance level of 0.05, conduct an appropriate test to determine if, for adult residents of this county, there is an association between gender and whether or not they were satisfied with services offered by the hospital.

(b) Is $\frac{800}{1000}$ a reasonable estimate for the proportion of all adult county residents who are satisfied with the services offered by this hospital? Explain why or why not.

Solution

(a) **State:**

H_0: There is no association between gender and the response to the statement.

H_a: There is an association between gender and the response to the statement.

Plan: We will conduct a Chi-Square Test for Independence. First, we check the conditions for inference.

Category	Male (Expected)	Female (Expected)
Satisfied	371.2	428.8
Not Satisfied	92.8	107.2

✓Random - The problem states that the data come from a random sample of 1,000 adult residents.

✓Large Counts - All expected counts are at least 5

✓Independence - We assume that there are more than 10,000 adult residents in the county, so the sample of 1,000 residents is at most than 10% of the population.

Do:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \; df = (r-1)(c-1) = (2-1)(2-1) = 1$$

$$\chi^2 = 0.441 + 0.382 + 1.766 + 1.528 = 4.117$$

Using our calculator, we obtain an associated p-value of about 0.0426

Conclude: Since our p-value of 0.0426 is less than our significance value of $\alpha = 0.05$ we must reject our null hypothesis. There is statistically significant evidence that, for adult residents in the country, there is an association between gender and whether or not they were satisfied with services offered by the hospital.

(b) This is a reasonable estimate of the satisfaction level for the entire county population, provided that the sample of 1,000 residents is representative of the population.

Since the sample was randomly selected, and the sample size is large, 0.80 should be a reasonable estimate for the proportion of all adult county residents who are satisfied with the hospital services.

Problem 8.7.2 (Source: 2003 AP Statistics FRQ Problem 5) — A random sample of 200 students was selected from a large college in the United States. Each selected student was asked to give his or her opinion about the following statement.

"The most important quality of a person who aspires to be the President of the United States is a knowledge of foreign affairs."

Each response was recorded in one of the five categories. The gender of each selected student was noted. The data are summarized in the table below. Is there sufficient evidence to indicate that the response is dependent on gender? Provide statistical evidence to support your conclusion.

	Response Category				
	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Male	10	15	15	25	25
Female	20	25	25	25	15

Solution

State:

H_0: There is no association between gender and satisfaction with the services offered by the hospital.

H_a: There is an association between gender and satisfaction with the services offered by the hospital.

Plan: We will conduct a chi-square test for association to determine if there is a significant relationship between gender and satisfaction with the services offered by the hospital. First, we check the conditions for inference.

Expected Counts:

13.5	18	18	22.5	18
16.5	22	22	27.5	22

✓Random - The sample is a simple random sample of 200 students.

✓Large Counts - All expected counts are at least 5

✓Independence - The sample size (200 students) is less than 10% of the total student population at the college.

Do:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \; df = (r-1)(c-1) = (2-1)(5-1) = 4$$

$$0.907 + 0.500 + 0.500 + 0.278 + 2.722 + 0.742 + 0.409 + 0.409 + 0.227 + +2.227$$

$$\chi^2 = 8.921$$

Using our calculator, we obtain the associated p-value of about 0.063

Conclude: Since our p-value of 0.063 is larger than our significance value of $\alpha = 0.05$ we fail to reject our null hypothesis. There is no statistically significant evidence proving that there is an association between gender and satisfaction with the services offered in the hospital, for people similar to those in the sample.

Problem 8.7.3 (Source: 1999 AP Statistics FRQ Problem 2) — The Colorado Rocky Mountain Rescue Service wishes to study the behavior of lost hikers. If more were known about the direction in which lost hikers tend to walk, then more effective search strategies could be devised. Two hundred hikers selected at random from those applying for hiking permits are asked whether they would head uphill, downhill, or remain in the same place if they became lost while hiking. Each hiker in the sample was also classified according to whether he or she was an experienced or novice hiker. The resulting data are summarized in the following table.

	Uphill	Downhill	Remain in Same Place
Novice	20	50	50
Experienced	10	30	40

Do these data provide convincing evidence of an association between the level of hiking expertise and the direction the hiker would head if lost?

Give appropriate statistical evidence to support your conclusion.

Solution

State:

H_0 : There is no association between the direction a hiker would head and hiking expertise.
H_a : There is an association between the direction a hiker would head and hiking expertise.

Plan: We will conduct a chi-square test for independence. First, we check the conditions for inference:

	Uphill	Downhill	Remain in Same Place
Novice	20 (18)	50 (48)	50 (54)
Experienced	10 (12)	30 (32)	40 (36)

✓Random - The sample of hikers were randomly selected.
✓Large Counts - All expected counts are larger than 5.
✓Independent - We may assume that there are more than 2000 hikers applying for hiking permits.

Do:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E},\ df = (r-1)(c-1) = (2-1)(5-1) = 4$$

$$\chi^2 = 0.222 + 0.083 + 0.296 + 0.333 + 0.125 + 0.444$$

$$\chi^2 = 1.503$$

Using our calculator, we obtain the associated p-value of about 0.472

Conclude: Since our p-value of 0.472 is relatively large, we fail to reject our null hypothesis. There is no statistically significant evidence proving that there is an association between the level of hiking expertise and the direction the hiker would head if lost.

Problem 8.7.4 (Source: 2017 AP Statistics FRQ Problem 5) — The table and the bar chart below summarize the age at diagnosis, in years, for a random sample of 207 men and women currently being treated for schizophrenia.

	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46	40	21	12	119
Men	53	23	9	3	88
Total	99	63	30	15	207

Do the data provide convincing statistical evidence of an association between age-group and gender in the diagnosis of schizophrenia?

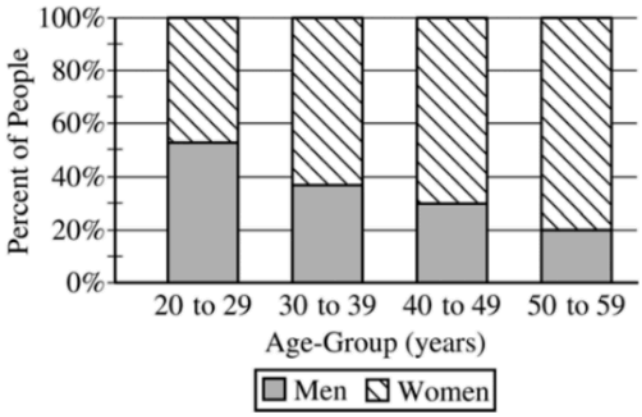


Figure 8.3: Caption

Solution

State:

H_0: There is no association between age group and gender in the diagnosis of schizophrenia.

H_a: There is an association between age group and gender in the diagnosis of schizophrenia.

Plan: We conduct a chi-square test for independence. First, we check the conditions for inference.

	20 *to* 29	30 *to* 39	40 *to* 49	50 *to* 59	Total
Women	46 (56.91)	40 (36.22)	21 (17.25)	12 (8.62)	119
Men	53 (42.09)	23 (26.78)	9 (12.75)	3 (6.38)	88
Total	99	63	30	15	207

✓Random - The data come from a random sample of 207 men and women.

✓Large Counts - All expected counts are at least 5.

✓Independent - We may assume that there are more than 1190 women and 880 men that are being treated for schizophrenia.

Do:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E},\ df = (r-1)(c-1) = (2-1)(4-1) = 3$$

$$\chi^2 = 2.10 + 2.85 + 0.39 + 0.53 + 0.82 + 1.10 + 1.32 + 1.79$$

$$\chi^2 = 10.90$$

Using our calculator, we obtain our associated p-value of about 0.0123.

Conclude: Since our p-value of 0.0123 is smaller than our significance level of $\alpha = 0.05$, we reject our null hypothesis. There is statistically significant evidence proving that there is an association between age-group and gender in the diagnosis of schizophrenia for patients similar to those in the study.

Problem 8.7.5 (Source: Original) — A school is trying to determine whether students' participation in extracurricular activities is associated with their year in school (freshman, sophomore, junior, or senior). The school conducted a survey of students, asking them to indicate their primary extracurricular activity (sports, music, or academic clubs). The results of the survey are shown in the table below.

Grade Level	Sports	Music	Academic Clubs	Total
9th Grade	12	9	6	27
10th Grade	10	6	5	21
11th Grade	8	9	7	24
12th Grade	9	6	7	22
Total	39	30	25	94

Is there evidence to suggest that the distribution of extracurricular activity participation is different across the four grade levels? Test the hypothesis using a significance level of $\alpha = 0.05$.

Solution

State:

H_0: The distribution of participation in extracurricular activities is the same for all four grade levels.

H_a: The distribution of participation in extracurricular activities is different across the four grade levels.

Plan: We will conduct a chi-square test for homogeneity. First, we check the conditions for inference.

Grade Level	Sports	Music	Academic Clubs	Total
9th Grade	12 (11.21)	9 (8.62)	6 (7.18)	27
10th Grade	10 (8.72)	6 (6.7)	5 (5.58)	21
11th Grade	8 (9.97)	9 (7.65)	7 (6.38)	24
12th Grade	9 (8.1)	6 (6.03)	7 (5.03)	22
Total	39	30	25	94

✓Random - The students were randomly selected for the survey.

✓Large Counts - All expected counts are at least 5.

✓Independent - We may assume that the number of students surveyed is at most 10% of the total population for each grade.

Do:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \; df = (r-1)(c-1) = (4-1)(3-1) = 6$$

$$\chi^2 = 0.057+0.017+0.194+0.183+0.073+0.062+0.389+0.222+0.061+0.002+0.146+0.249$$

$$\chi^2 = 1.655$$

Using our calculator, we obtain the corresponding p-value of about 0.949

Conclude: Since our p-value of 0.949 is larger than our significance level of $\alpha = 0.05$, we fail to reject our null hypothesis. There is no statistically significant evidence proving that there is a significant difference in the distributions of extracurricular activities across the four grade levels.

Problem 8.7.6 (Source: 2016 AP Statistics FRQ Problem 2) — Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in an experiment. One ad was for a new sugary snack called Choco-Zuties, and the other ad was for a new healthy snack called Apple-Zuties.

For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

- The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
- The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
- The advertising for group C included neither the Choco-Zuties ad nor the Apple-Zuties ad.

After the program, the children were offered a choice between the two snacks. The table below summarizes their choices

(a) Do the data provide convincing statistical evidence that there is an association between the type of ad and children's choice of snack among all children similar to those who participated in the experiment?

(b) Write a few sentences describing the effect of each ad on children's choice of snack.

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

Figure 8.4: Caption

Solution

(a)**State:**

H_0: There is no association between the type of ad and the children's snack choice.

H_a: There is an association between the type of ad and the children's snack choice.

Plan: We will conduct a Chi-Square test for Independence. First, we check the conditions for inference.

Expected Counts:

18.67	6.33
18.67	6.33
18.67	6.33

✓Random - The children were randomly selected and assigned to one of the three groups
✓Large Counts - The expected counts are at least 5.
✓Independence - We may assume that the sample size it at most 10% of the population.

Do:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E},\ df = (r-1)(c-1) = (3-1)(2-1) = 2$$

$$\chi^2 = 0.29 + 0.86 + 1.72 + 5.06 + 0.60 + 1.75 = 10.28$$

Using our calculator, we obtain our associated p-value of about 0.006

Conclude: Since our p-value of 0.006 is less than our significance level of $\alpha = 0.05$, we reject our null hypothesis. There is statistically significant evidence proving that there is an association between the type of ad and children's choice of snack among all children similar to those who participated in the experiment.

(b) Group A (Choco-Zuties Ad): 84% of children chose Choco-Zuties, similar to the 88% in the group with no ad. This suggests the Choco-Zuties ad had little effect on snack choice.

Group B (Apple-Zuties Ad): 52% of children chose Choco-Zuties, while 48% chose Apple-Zuties, indicating the Apple-Zuties ad significantly increased preference for the healthier snack.

Group C (No Ad): Without any ad, 88% of children chose Choco-Zuties, showing a strong baseline preference for the sugary snack.

9 Unit 9: Inference for Quantitative Data: Slopes

§9.1 Introducing Statistics: Do Those Points Align?

Recall from Unit 2 that the least squares regression line for a data set comparing two quantitative variables can be expressed in the form $\hat{y} = a + bx$ where a is the y-intercept, and b is the slope (note that the hat above y indicates that it is the predicted value, not the actual value).

We can take a least square regression line for any sample of data, but when we do, there are many things to be skeptical about. For example, how do we know if the values of a and b accurately represent the data set? Or worse, what if our sample data set shows a linear association when there isn't one in the true population? This unit aims to solve these questions.

This unit aims to find a way to estimate the true population regression line (that is, if the relationship is even linear) to understand the relationship between the variables we care about. We will only concern ourselves with b (the slope) in this unit. However, before we can make an inference about the true population regression line, there a few conditions that must be met

Note 9.1.1 (Conditions for Inference for Quantitative Data: Slopes)

An easy way to remember the conditions is through the acronym **LINER**

- **L**inear - There must be a linear relationship between the two quantitative variables
- **I**ndependent - Individual observations must be independent of each other (either sampling with replacement, or the 10% rule.
- **N**ormal - For any fixed value of x, the response y varies according to a normal distribution
- **E**qual Standard Deviation: The standard deviation for all y is the same regardless of the value of x
- **R**andom - Each data point is chosen randomly from the population of interest

AP questions will only ever ask you to assume the conditions are met, or to list them out since the Normal and Equal Standard Deviation conditions are beyond the AP curriculum. Thus, we will not be going into much detail on these conditions.

Just like how we have sampling distributions for sample means and proportions, we also have them for sample slopes as well. Also, just like in other sampling distributions, the mean of the sampling distribution of sample slopes is equal to the true mean of the population slope. However, the standard deviation of the sampling distribution for sample slopes is much more complicated (and outside the scope of the AP curriculum),

which is why it will ALWAYS be given to you on the AP exam.

For the sake of the example, let's imagine we are conducting an inference test for the relationship between caffeine consumed (mg) and hours slept among high school students. If this were an AP question, we very well might see something like the following:

Predictor	Coef	SE Coef	T	P
Constant	480	20		
Caffeine Consumed (mg)	-2.3	0.8	-2.875	0.006

$$S = 25.4,\ R^2 = 21\%$$

This is called a Regression output, and it appears often with these types of questions. In our linear regression formula $\hat{y} = a + bx$, the Constant Coefficient is our a value, or in other words our y-intercept. The coefficient of Caffeine Consumed is our b value, or in other words our slope. The SE Coef column is very important, as it gives us the Standard Error of its respective coefficient. For example, the standard error of the slope would thus be 0.8. The S value represents the standard deviation of the residuals, which essentially calculates how much our observed data points are spread around the regression line. You may recall from Unit 2 that the R^2 value is an indicator of how much of the variance in the y variable can be explained by the variance of the x variable, and thus it is another tool to describe how well the line of least squares regression describes the sample data (if $R^2 = 100\%$. For now, we won't worry about the T and P columns

§9.2 Confidence Intervals for the Slope of a Regression Model

As mentioned in the previous section, the mean of the sampling distribution of sample means is always equal to the true mean, just like in other distributions. The true standard deviation is much more complicated and not required for the AP statistics exam, but we present it anyway for the sake of improving our understanding.

Theorem 9.2.1 (Standard deviation of sampling distribution for slopes)

$$\sigma_b = \frac{\sigma}{(\sigma_x)(\sqrt{n})}$$

where σ is the standard deviation of the true regression line and σ_x is the standard deviation of the x-values for all the data points in the population.

However, since we will never know σ and σ_x), we must replace them with s and s_x respectively. Once again, since we are taking an estimate for our standard deviations, we call this our standard error.

Theorem 9.2.2 (Standard error of sampling distribution for slopes)

$$SE_b = \frac{s}{(s_x)(\sqrt{n-1})}$$

Note 9.2.3

You don't need to know this formula either, as the Standard Error of the slope will ALWAYS be given to you on AP exams when required.

Since we are using the Standard Error of the slope, our sampling distribution follows a t-model, which we should already be familiar with. Since we are using a t-model, we of course have to figure out the number of degrees of freedom.

Theorem 9.2.4 (Degrees of freedom for slopes)

$$\text{degrees of freedom} = n - 2$$

Where n is the number of data points.

Lastly, we introduce the formula for constructing a confidence interval for slope. It is not that different from the confidence intervals we have seen earlier.

Theorem 9.2.5 (Confidence Interval Formula for a true population regression line)

$$\text{Confidence Interval} : b \pm (t*)(SE_b)$$

Where b is the sample slope, (t^*) is the corresponding critical value and SE_b is the standard error for the sample slope.

Now that we have all the information we need to construct a confidence interval for slopes, let's look at an example.

Problem 9.2.6 (Source: Original) — A health researcher is studying the relationship between the number of hours spent on physical exercise per week and the amount of weight loss (in pounds) among adults participating in a fitness program. The researcher collects data from a random sample of 30 participants and performs a linear regression analysis. The regression output is summarized below:

Predictor	Coef	SE Coef	T	P
Constant	12.5	1.8		
Caffeine Consumed (mg)	−0.85	0.22		

$$S = 5.4, R^2 = 29\%$$

Construct and interpret a 95% confidence interval for the true slope of the regression line relating hours of exercise per week to weight loss. Assume that all the conditions for inference are met.

State: We will construct a 95% confidence interval for the true slope of the regression line relating hours of exercise per week to weight loss.

Plan: We conduct a 1-sample t-interval for slope. We were told to assume that all conditions for inference were met (Linear, Independent, Normal residuals, Equal variance, Random).

Do:

$$\text{Confidence Interval} : b \pm (t^*)(SE_b),\ df = 30 - 2 = 28$$

$$\text{Confidence Interval} : -0.85 \pm (1.701)(0.22)$$

$$\text{Confidence Interval} : (-1.224, -0.4758)$$

Conclude: We are 95% confident that the true slope of the regression line relating hours of exercise per week to weight loss lies in the interval $(-1.224, -0.4758)$

§9.3 Justifying a Claim About the Slope of a Regression Model Based on a Confidence Interval

This unit has little to learn, since it is identical to justifying a population mean and proportion claim. Just remember to try to always be as SPECIFIC as possible when making justifying a claim for the slope of a regression Model for slopes.

Note 9.3.1 (Making a claim about the slope of a regression model based on a confidence interval)

In this section, we present a few simple steps for making sure to always

1. Reference the sample from which you are obtaining the data from
2. State your confidence level
3. Properly interpret the slope in the context of the question (do not just say that the slope is between x and y

Example 1:
let's say we wanted to study the relationship between daily sugar (g) and body fat percentage. After randomly sampling 60 individuals, the regression output estimates the slope as 0.8 with a Standard Error of 0.3. A 95% confidence interval yields the interval $(0.2, 1.4)$. Draw a conclusion based on this information.

Bad example: We are 95% confident that the slope is somewhere in the interval $(0.2, 1.4)$

Good example: Based on the 60 randomly selected individuals in the sample, we are 95% confident that the slope of the true population regression line predicts that for every gram of additional average daily sugar intake, your body fat percentage will increase anywhere from 0.2 to 1.4 percent.

Example 2:

A sleep researcher is studying the relationship between the number of hours spent on social media each day and the total sleep duration (in hours) per night among high school students. After randomly sampling 75 students, the regression output yields a slope of -0.4 with a standard error of 0.15. Draw a conclusion from the given data.

Bad example: There is a 95% probability that the true slope falls between -0.7 and -0.1

Good example: Based on the 75 randomly selected students in the sample, we are 95% confident that the slope of the true population regression line predicts that for every additional hour spent on social media per day, the average sleep duration will decrease by anywhere from 0.1 to 0.7 hours. This suggests a significant negative relationship between social media use and sleep duration.

§9.4 Setting Up a Test for the Slope of a Regression Model

By now you should already be very familiar with the steps for setting up any inference test.

Note 9.4.1 (Setting up a test for the Slope of a Regression Model)
Steps:

1. State the test name and Null/Alternative hypotheses
2. Check that all conditions for inference are met (most of the time the question will ask you to assume they are)

In the null hypothesis, we always assume that there is nothing special going on, or in other words that the slope of the true regression line is equal to that of the sample regression line. The alternative hypothesis can either hypothesize that the slope of the true regression line is either greater than, less than, or simply not equal to the sample slope.

Problem 9.4.2 (Source: Original) — A researcher is studying the relationship between the number of hours spent studying per week (denoted by x) and the scores on a standardized test (denoted by y) of high school students. The researcher collects data from 30 high school students and fits a simple linear regression model, resulting in the following output:

$$\hat{y} = 50 + 2.2, SE_b = 0.7$$

The researcher believes that the slope of the true regression line should be $b = 1$. Assuming all the conditions for inference are met, conduct an inference test with the significance level of $\alpha = 0,05$ to determine whether the researcher's claim is valid.

In this section, we set up the test and finish it in the next section

State: We will be constructing a t-test for the slope between the number of hours spent studying per week and the scores of high school students on a standardized test.
H_0: $b = 1$
H_a: $b \neq 1$

Plan: We were told to assume that the conditions for inference were met. This means that...
✓Linear - We may assume that the true population scatterplot follows a somewhat linear relationship
✓Independent - There are at least 300 students in the population (since 10% of 300 is 30, or that the students were chosen with replacement)
✓Normal - The RESIDUALS must follow an approximately normal distribution (recall that $residual = y - \hat{y}$)
✓Equal Standard Deviation - The variance (also the standard deviation) of all the RESIDUALS is approximately the same for ALL values of x.
✓Random - Each individual was randomly selected from the population of interest (high school students)

§9.5 Carrying Out a Test for the Slope of a Regression Model

Out of all the inference tests on the AP exam, the inference test for a linear regression should be the easiest since there is the least amount of calculation involved in finding the p-value. In other inference tests, we often had to calculate the standard effect of our desired statistic, however, for this unit, it will **ALWAYS** be given! Once we have completed the setup for our inference test, we now must find our t-score for the sample.

Theorem 9.5.1 (T-score for an inference on slope)
$t = \frac{b - b_0}{SE_b}$
where b is the sample slope, b_0 is the hypothesized slope, and SE_b is the standard error of the sampling distribution of sample slopes. This formula should not feel new, as we are simply finding out how many standard errors our sample slope is away from our hypothesized slope.

Note 9.5.2
Sometimes the question will ask you to determine whether or not there is any linear relationship between two quantitative variables, and in that case, the null hypothesis is simply $b_0 = 0$. This is because if we assume that there is no linear relationship, the data points should be scattered randomly and have no apparent relationship.

Once we have our t-score, all that's left to do is, with the help of our calculator, use the *tcdf* function to get our associated p-value. Now, we will continue with our example from the previous section.

Do:
$t = \frac{b - b_0}{SE_b}$, $df = 30 - 2 = 28$
$t = \frac{2.2 - 1}{0.7}$
$t = 1.714$
Plugging this into our calculator, we obtain the ONE sided p-value of 0.04879. Since our test is a two-tailed test, we multiply our one-sided p-value by 2.
$2 \times 0.04879 = 0.09758$ is higher than our significance level of $\alpha = 0.05$

Conclude:
Since our p-value of 0.09758 is higher than our significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. There is no statistically significant evidence that the slope of the true population regression line between the number of hours spent studying per week and the scores of high school students on a standardized test is different from $b = 1$.

Note 9.5.3 (Important note on inferences for regression lines)

Looking back at our regression output, we still haven't talked about the T and P columns. Below is the regression output table of the example problem that we just solved, which was not given before.

Predictor	Coef	SE Coef	T	P
Constant				
hours spent studying per week	2.2	0.7	1.71	0.049

The T-column actually gives the corresponding t-statistic of the variable assuming the null hypothesis is true, and the P column gives the corresponding ONE sided p-value for that variable (notice how the numbers in the T and P column are the same as the ones we already calculated). In other words, if you are ever given the full regression output table on an AP statistics problem, you won't have to do any calculations when conducting an inference test for the slope! Just be sure to remember that the P column only gives the ONE sided P-value, so if you are conducting a two-sided t-test as we did in the example, you would have to multiply the P-value by two.

Problem 9.5.4 (Source: 2011 AP Statistics FRQ Problem 5) — Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on twenty-five randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 miles per hour to 40 miles per hour.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
predicted electricity production	0.240	0.019	12.63	0.000

(a) Use the computer output above to determine the equation of the least squares regression line. Identify all variables used in the equation.

(b) How much more electricity would the windmill be expected to produce on a day when the wind velocity is 25 mph than on a day when the wind velocity is 15 mph? Show how you arrived at your answer.

(c) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity?

(d) Is there statistically convincing evidence that electricity production by the windmill is related to wind velocity? Explain.

(a)

$$\text{predicted electricity production} = 0.137 + 0.240 \times (\text{wind velocity})$$

(b) Note that the slope of $b = 0.240$ indicates that for each increase of 1mph in the wind velocity, the expected electricity production will increase by 0.240amperes. The difference in mph we are asked to compute is:
$25 - 15 = 10$
$10 \times 0.240 = 2.4$
Thus, the windmill would be expected to produce 2.4 more amperes

(c) Recall that the definition of our R^2 value is the proportion of the variation of the y variable that may be explained by our variation in the x variable. Thus, using our regression output table, we deduce that 0.873% of the variation in electricity production is explained by its linear relationship with wind velocity.

(d) By looking at the given regression output table, we obtain our p-value rounded to three decimal places which is 0.000. We can safely reject our null hypothesis since this p-value is much less than 0.01. Thus, there is statistically significant evidence that electricity production by windmill is related to wind velocity.

§9.6 Unit 9 Practice Problems

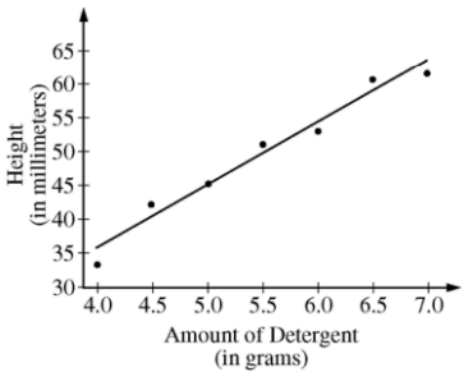
Problem 9.6.1 (Source: 2006 AP Statistics FRQ Problem 2) — A manufacturer of dish detergent believes the height of soapsuds in the dishpan depends on the amount of detergent used. A study of the suds' heights for a new dish detergent was conducted. Seven pans of water were prepared. All pans were of the same size and type and contained the same amount of water. The temperature of the water was the same for each pan. An amount of dish detergent was assigned at random to each pan, and that amount of detergent was added to the pan. Then the water in the dishpan was agitated for a set amount of time, and the height of the resulting suds was measured.

A plot of the data and the computer output from fitting a least squares regression line to the data are shown below.

(a) Write the equation of the fitted regression line. Define any variables used in this equation.

(b) Note that $s = 1.99821$ in the computer output. Interpret this value in the context of this study.

(c) Identify and interpret the standard error of the slope.



Predictor	Coef	SE Coef	T	P
Constant	−2.679	4.222	−0.63	0.554
Amount	9.5000	0.7553	12.58	0.000

S = 1.99821 R–Sq = 96.9% R–Sq(adj) = 96.3%

Figure 9.1: Caption

(a) $\hat{y} = -2.679 + 9.5x$ Where $\hat{y}$ is the predicted mean height of the soapsuds and x represents the amount of detergent added to the pan.

(b) The value $S = 1.99821$mm represents the standard deviation of the residuals. This statistic helps measure the variability of our data with respect to our regression line.

(c) The standard error of the slope is 0.7553. In other words, the standard deviation of the sampling distribution of the slope is 0.7553. This means that if we were to take repeated samples, the estimated slopes would typically vary by approximately 0.7553

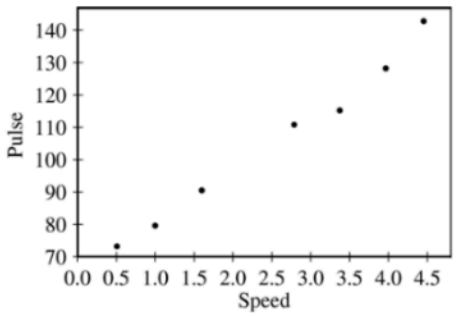
from the true population slope. A smaller standard error indicates that our estimate of the slope is more precise, while a larger standard error suggests more variability in the slope estimates across different samples.

Problem 9.6.2 (Source: 2005 AP Statistics FRQ Problem 5, Form B) — John believes that as he increases his walking speed, his pulse rate will increase. He wants to model this relationship. John records his pulse rate, in beats per minute (bpm), while walking at each of seven different speeds, in miles per hour (mph). A scatterplot and regression output are shown below.

(a) Using the regression output, write the equation of the fitted regression line.

(b) Do your estimates of the slope and intercept parameters have meaningful interpretations in the context of this question? If so, provide interpretations in this context. If not, explain why not.

(c) John wants to provide a 98 percent confidence interval for the slope parameter in his final report. Compute the margin of error that John should use. Assume that conditions for inference are satisfied.



Regression Analysis: Pulse Versus Speed

Predictor	Coef	SE Coef	T	P
Constant	63.457	2.387	26.58	0.000
Speed	16.2809	0.8192	19.88	0.000

S = 3.087 R-Sq = 98.7% R-Sq (adj) = 98.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3763.2	3763.2	396.13	0.000
Residual	5	47.6	9.5		
Total	6	3810.9			

Figure 9.2: Caption

Solution

(a) $\hat{y} = 63.457 + 16.2809x$ where $\hat{y}$ is the predicted pulse in bpm, and x is the walking speed in mph.

(b) The slope provides an estimate for the mean increase in predicted pulse (bpm) as his speed is increased by one mile per hour. The y-intercept represents his pulse rate if he is not moving (bpm).

(c)

$$\text{Margin of Error} = (t)(SE_b)$$

$$\text{Margin of Error} = (3.365)(0.8192) = 2.7566\text{bpm}$$

$$\text{Margin of Error} = 2.7566\text{bpm}$$

Problem 9.6.3 (Source: 2008 AP Statistics FRQ Problem 6) — Administrators in a large school district wanted to determine whether students who attended a new magnet school for one year achieved greater improvement in science test performance than students who did not attend the magnet school. Knowing that more parents would want to enroll their children in the magnet school than there was space available for those children, the district administrators decided to conduct a lottery of all families who expressed interest in participating. In their data analysis, the administrators would then compare the change in test scores of those children who were selected to attend the magnet school with the change in test scores of those who applied to attend the magnet school but who were not selected. The tables below show the scores on the same science pretest and the same science posttest for 20 students. Of the 20 students, 8 were randomly selected from the magnet school and 12 were randomly selected from those who applied to attend the magnet school but who were not selected and then attended their original school.

Pretest Score	Posttest Score	Posttest−Pretest
80	97	17
78	98	20
86	84	−2
78	79	1
64	89	25
71	77	6
71	83	12
73	88	15
$\bar{x} = 75.125$	$\bar{x} = 86.875$	$\bar{x} = 11.750$
$s = 6.770$	$s = 7.699$	$s = 9.407$

Pretest Score	Posttest Score	Posttest−Pretest
83	80	17
80	89	20
63	65	−2
79	78	1
83	93	25
77	79	6
66	70	12
80	84	15
73	80	15
90	90	15
77	78	15
90	91	15
$\bar{x} = 78.417$	$\bar{x} = 81.417$	$\bar{x} = 3.000$
$s = 8.207$	$s = 8.512$	$s = 3.977$

Regression Analysis: Post_Magnet versus Pre_Magnet

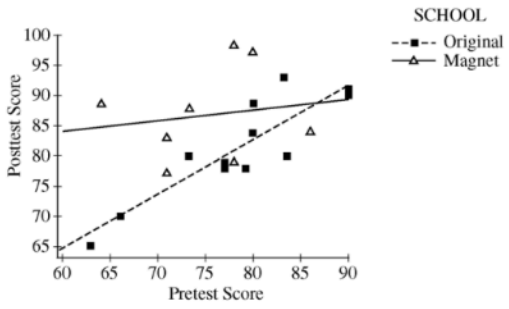
Predictor	Coef	SE Coef	T	P
Constant	73.27	34.55	2.12	0.078
Pre_Magnet	0.1811	0.4583	0.40	0.706

S = 8.20920 R-Sq = 2.5% R-Sq(adj) = 0.0%

Regression Analysis: Post_Original versus Pre_Original

Predictor	Coef	SE Coef	T	P
Constant	9.24	11.91	0.78	0.456
Pre_Original	0.9204	0.1512	6.09	0.000

S = 4.11463 R-Sq = 78.8% R-Sq(adj) = 76.6%



Problem 9.6.4 (Source: 2008 AP Statistics FRQ Problem 6 - Continued) — (a) Perform a test to determine whether students who attend the magnet school demonstrate a significantly higher mean difference in test scores (Posttest − Pretest) than students who applied to attend the magnet school but who were not selected and then attended their original school.

Administrators were also interested in using pretest scores on this test as a predictor of posttest scores on the test. The following computer output contains the results from separate regression analyses on the magnet school scores and on the original school scores. The accompanying graph displays the data and separate regression lines for the magnet and original schools.

(b) (i) State the equation of the regression line for the magnet school and interpret its slope in the context of the question.

(ii) State the equation of the regression line for the original school and interpret its slope in the context of the question.

(c) To determine whether there is a significant correlation between pretest score and posttest score, a test of the following hypotheses will be performed.

H_0: There is no correlation between pretest score and posttest score (true slope$= 0$)
versus
H_a: There is a correlation between pretest score and posttest score (true slope$\neq 0$)

(i) Using the regression output, state the p-value and conclusion for this test at the magnet school. Assume the conditions for inference have been met.

(ii) Using the regression output, state the p-value and conclusion for this test at the original school. Assume the conditions for inference have been met.

(d) What additional information do the regression analyses give you about student performance on the science test at the two schools beyond the comparison of mean differences in part (a)?

Solution

(a) **State:** Let μ_1 represent the true mean difference in test scores (posttest−pretest) of

students attending the magnet school. Let μ_2 represent the true mean difference in test scores (posttest−pretest) of students who were rejected from the magnet school.
H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 > 0$

Plan: We will conduct a two-sample t-test for the difference between population means. First we check the conditions of inference.
✓Random - Both samples were selected at random
✓Normal - Based on the fact that there are no obvious outliers, we may assume that distribution for the differences in test scores for both the magnet school and the original school are roughly normal.
✓Independent - We may assume that there are at least 80 students from the magnet school, and at least 120 students who applied to the magnet school but were rejected.

Do:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},\ df = 8 - 2 = 6$$

$$t = \frac{(11.750 - 3.000) - (0)}{\sqrt{\frac{9.407^2}{8} + \frac{3.977^2}{12}}}$$

$$t = \frac{(11.750 - 3.000) - (0)}{\sqrt{\frac{9.407^2}{8} + \frac{3.977^2}{12}}}$$

$$t = 2.487$$

Using our calculator, we obtain our associated p-value of about 0.0177

Conclude: Since our p-value of 0.0177 is less than our significance level of $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence that students who attend the magnet school demonstrate a significantly higher mean difference in test scores (posttest-pretest) than students who applied to attend the magnet school but who were not selected and then attended their original school.

(b) Let y and x be the posttest and pretest score respectively.
(i) $\hat{y} = 73.27 + 0.1811x$. The slope of 0.1811 indicates that for every 1-point increase in the pretest score, the predicted posttest score increases by 0.1811 for students attending the magnet school.

(ii) $\hat{y} = 9.24 + 0.0.9204x$. The slope of 0.9204 indicates that for every 1-point increase in the pretest score, the predicted posttest score increases by 0.1811 for students who applied to the magnet school, but got rejected.

(c)
(i) At the magnet school, the p-value for the hypothesis test is 0.706. Since this significance level is relatively high, we fail to reject the null hypothesis. There is no statistically significant evidence supporting the claim that there is a correlation between the pretest score and the posttest score.

(ii) At the magnet school, the p-value for the hypothesis test is 0.000. Since this

significance level is very low, we reject the null hypothesis. There is statistically significant evidence supporting the claim that there is a correlation between the pretest score and the posttest score.

(d) In contrast to the two-sample analysis in part (a), the regression analysis allows us to examine the relationship between pretest and posttest scores for each school. The results from the regression output and the graph suggest that students with lower pretest scores tend to benefit more from attending the magnet school than students at the original school. Additionally, at the magnet school, the impact is stronger for students with low pretest scores compared to those with higher pretest scores. This indicates that students at the magnet school generally perform well on the posttest, regardless of their initial scores. On the other hand, at the original school, only the students who performed well on the pretest tended to achieve high posttest scores. This result highlights how magnet schools may provide an equalizing effect, helping lower-performing students achieve similar posttest outcomes to their higher-performing peers, something that is less evident at the original school.

Problem 9.6.5 (Source: Original) — A researcher is studying the relationship between the number of hours a student studies per week and their score on a standardized test. A random sample of 15 students was selected, and the data were collected. The linear regression equation based on the sample data is:

$$\hat{y} = 50 + 3.2x$$

Where:

- $\hat{y}$ is the predicted test score.
- x is the number of hours spent studying per week.

The standard error of the slope is $SE_{\hat{\beta}} = 0.7$. The researcher wants to test whether there is evidence that the slope of the population regression line is different from 0. Use a significance level of $\alpha = 0.05$. Assume the conditions for inference are met.

Solution

State:

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

Plan: We will conduct a one-sample t-test for slope. We were told to assume that the conditions of inference are met (Linear, Independent, Normal, Equal variance, Random).

Do:

$$t = \frac{b - b_0}{SE_b}, \; df = n - 2 = 13$$

$$t = \frac{3.2 - 0}{0.7}$$

Using our calculator, we obtain our associated p-value of about 0.0005256

Conclude: Since our p-value of 0.0005256 is less than our significance level of $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence that the slope of the true population regression line of the relationship between the number of hours a student studies per week and their score on a standardized test is greater than zero.

Problem 9.6.6 (Source: Original) — A car manufacturer wants to study the relationship between the weight of a car (in thousands of pounds) and its fuel efficiency (measured in miles per gallon, mpg). A random sample of 15 cars was selected, and a linear regression analysis was performed to predict fuel efficiency from the weight of the car. The regression equation is given by:

$$\hat{y} = 40 - 2.5x$$

where $\hat{y}$ is the predicted fuel efficiency in mpg and x is the weight of the car in thousands of pounds. The manufacturer also obtained the following output:

Coefficient	Estimate	Standard Error	t	p
Intercept	40	3.2	–	–
Slope	-2.5	0.75	–	–

The manufacturer is interested in the following:

(a)Interpret the slope and intercept of the regression equation in the context of the problem.

(b) Test whether there is a significant linear relationship between the weight of the car and its fuel efficiency. Use a significance level of $\alpha = 0.05$.

Solution

(a) The intercept of 40 means that when the car's weight is 0 (which is not realistic for an actual car), the predicted fuel efficiency would be 40 miles per gallon. While this value doesn't have practical meaning in this context, it helps in formulating the regression equation.

The slope of -4.5 indicates that for every additional $1,000$ pounds of car weight, the fuel efficiency decreases by 4.5 miles per gallon, on average. This negative slope confirms that heavier cars tend to have lower fuel efficiency.

(b) **State:**
$H_0 : \beta_1 = 0$ (no linear relationship between car weight and fuel efficiency)
$H_a : \beta_1 \neq 0$ (there is a linear relationship between car weight and fuel efficiency)

Plan: We will conduct a $1 - sample$ z-test for slope. We were told to assume that the conditions for inference were met (Linear, Independent, Normal, Equal Variance, Random).

Do:

$$t = \frac{b - b_0}{SE_b}, \; df = 15 - 2 = 13$$

$$t = \frac{-2.5 - 0}{0.75}$$

$$t = -3.33$$

Using our calculator, we obtain the associated p-value of about 0.0054

Conclude: Since our p-value of 0.0054 is less than our significance level of $\alpha = 0.05$, we reject the null hypothesis. There is statistically significant evidence of a linear relationship between the weight of a car and its fuel efficiency.