



# Probability & Statistics Workbook Solutions

*krista king*  
MATH

## ONE-WAY TABLES

- 1. Identify the variables in the following data description and classify the variables as categorical or quantitative. If the variable is quantitative, list the units.

“The Indianapolis 500 is a car race that’s been taking place since 1911 and is often scheduled to take place over Memorial Day weekend. The race takes place at the Indianapolis Motor Speedway and a driver needs to complete 200 laps that cover a distance of 500 miles. Race results are reported by driver number, the driver’s name, the type of car the driver uses, and the time to the nearest ten-thousandth of a second. If a driver doesn’t finish the race, instead of the time to complete the race, their number of laps completed is recorded.”

*Solution:*

Remember that categorical variables can be represented as numbers. But they just don’t measure anything, and we can’t use them to perform a calculation. The driver’s number is a categorical variable because it’s not a measurement, but a way of keeping track of a person. The driver’s name and the type of car are also categorical.

The quantitative variables are measurements, like the time it takes a driver to finish the race, or the number of laps completed.



Categorical variables	Quantitative variables
Driver number	Time
Driver name	Number of laps
Type of car	

- 2. Casey is taking a survey of her senior class. She plans to ask the seniors this question:

“In general do you think things have gotten better or worse for our students over the course of the year?”

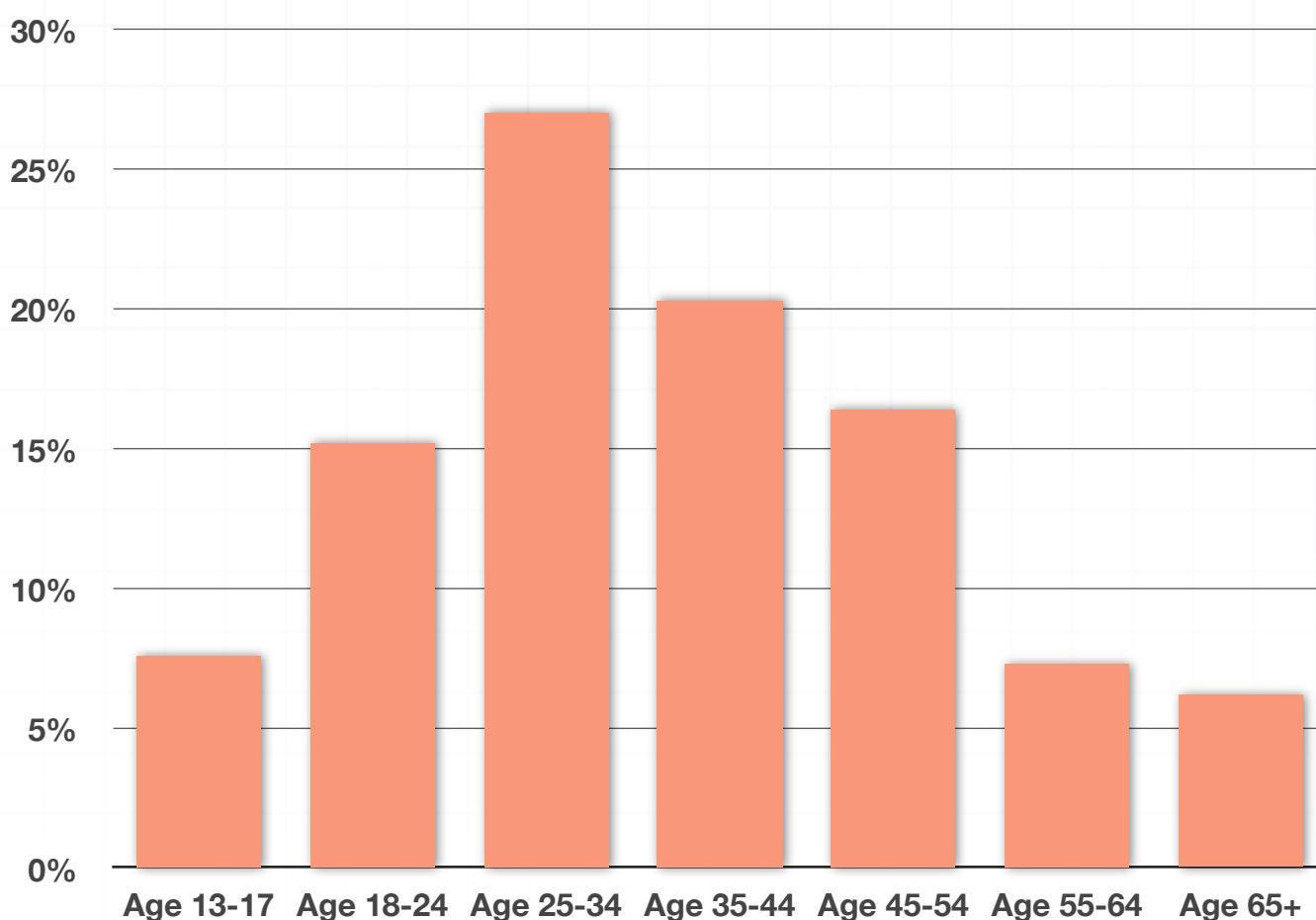
Her survey has a checklist with these responses: Better, Worse, Stayed the same, and Don’t know. Who are the individuals in the survey? What type of response variable is Casey looking for? Is it categorical or quantitative? What is the level of measurement of the data?

*Solution:*

Casey is surveying the senior class, so those students are the individuals of interest in the survey. Casey’s data is categorical because there’s not a unit of measurement included in the survey. Instead, she’s organizing each respondent’s answer into a category of Better, Worse, Stayed the same, or Don’t know. The data is measured at the ordinal scale because it’s categorical and can be ordered as four different categories.



3. The graph below shows the age breakdown of Apple iPad owners in the United States in February, 2011. Who are the individuals in the data? What is the variable? Is it categorical or quantitative?



Source: [www.statista.com](http://www.statista.com)

*Solution:*

The individuals in the data are the respondents, by age group. The percentage of each age group is a quantitative variable because it's a measurement.

4. The table below shows the number of rejected products by worker and shift. Can the data be used to build a one-way table? Why or why not?

Is the number of rejected products a discrete or continuous quantitative variable?

Worker ID	1st shift	2nd shift	3rd shift
1123	42	45	42
2256	45	74	32
6435	36	78	41

*Solution:*

This data can't be used to build a one-way table. In order to know the number of rejected products, we'd need to know two things: the individual worker ID, and the shift.

This means the data is now dependent on two independent things, not just one. In order to get to an answer to the question: "How many rejected products?," we'd need to ask something like "How many rejected products were there for worker 1123 during the first shift?" Since we need more than one reference point, this data can't be used to build a one-way table.

The number of rejected products is a discrete quantitative variable because it can only take on certain numeric values. We can't have, for example, 42.5 or 74.35 rejected products.

## ■ 5. Why is this table an example of a one-way data table?



Flavor	Scoops sold	Contains chocolate?	Smooth or chunky?
Vanilla	300	No	Smooth
Chocolate	450	Yes	Smooth
Cookies & Cream	275	Yes	Chunky
Mint Chocolate Chip	315	Yes	Chunky
Fudge Brownie	375	Yes	Chunky
Rocky Road	250	Yes	Chunky

*Solution:*

Even though this table has three different variables, if I'm given one individual and a category, I can answer a question about the data. For example I could ask: "How many scoops of vanilla were sold?" and I would know right away that the answer was 300.

If the data doesn't fit into a one-way table, I'd need to answer a question about both categories. For example, to answer a question like "How many scoops were sold?" we might need to ask something like "What flavor and which store?".

- 6. A botany student wants to test the claim of a diaper company that their product may be used in a compost pile. He creates 12 identical gardens and plants a random selection of 7 tomato plants in each one. He plans to have a fellow student use traditional compost on 6 of the garden plots and the compost from the diapers on the other 6. He does this so he



doesn't know which plot is which. He plans to check the tomato plants for disease every two days for a month, and record the number of tomato plants with disease after each check. Would this experiment result in a one-way data table? Why or why not?

*Solution:*

The experiment does not result in a one-way data table.

We can think of how the botany student would need to record his data to see whether or not this is an example of a one-way data table. He could create a table with the checks on the plants and the number of plots to record the number of tomato plants with disease.

	Check 1	Check 2	Check 3	Check 4	...	Check 15
Plot 1						
Plot 2						
Plot 3						
Plot 4						
...						
Plot 12						

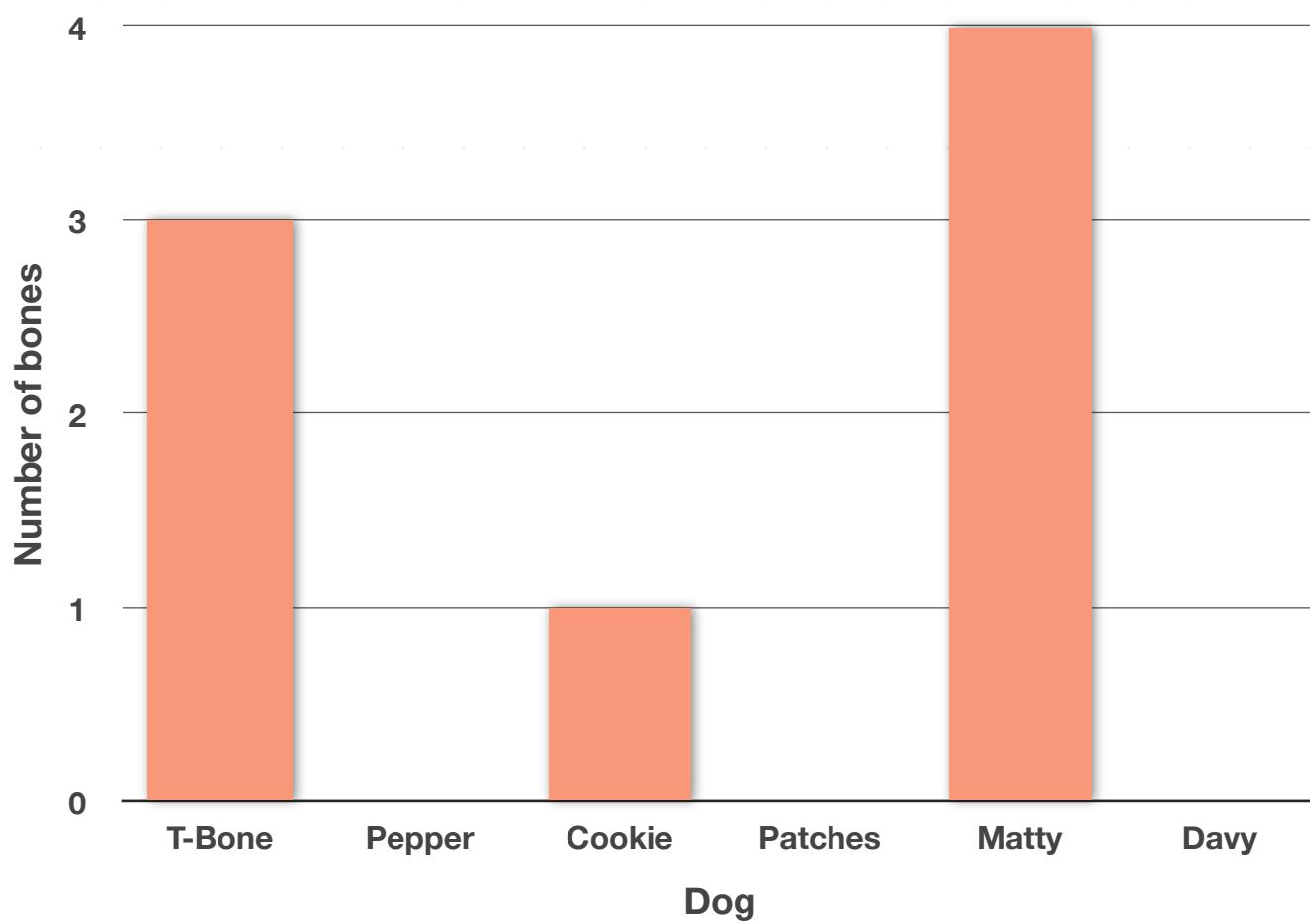
From the table, we can see that to answer a question like “How many tomato plants ended up with a disease?” we would need to know two things: which plot, and which check. This means the data is not an example of data that fits into a one-way table.



## BAR GRAPHS AND PIE CHARTS

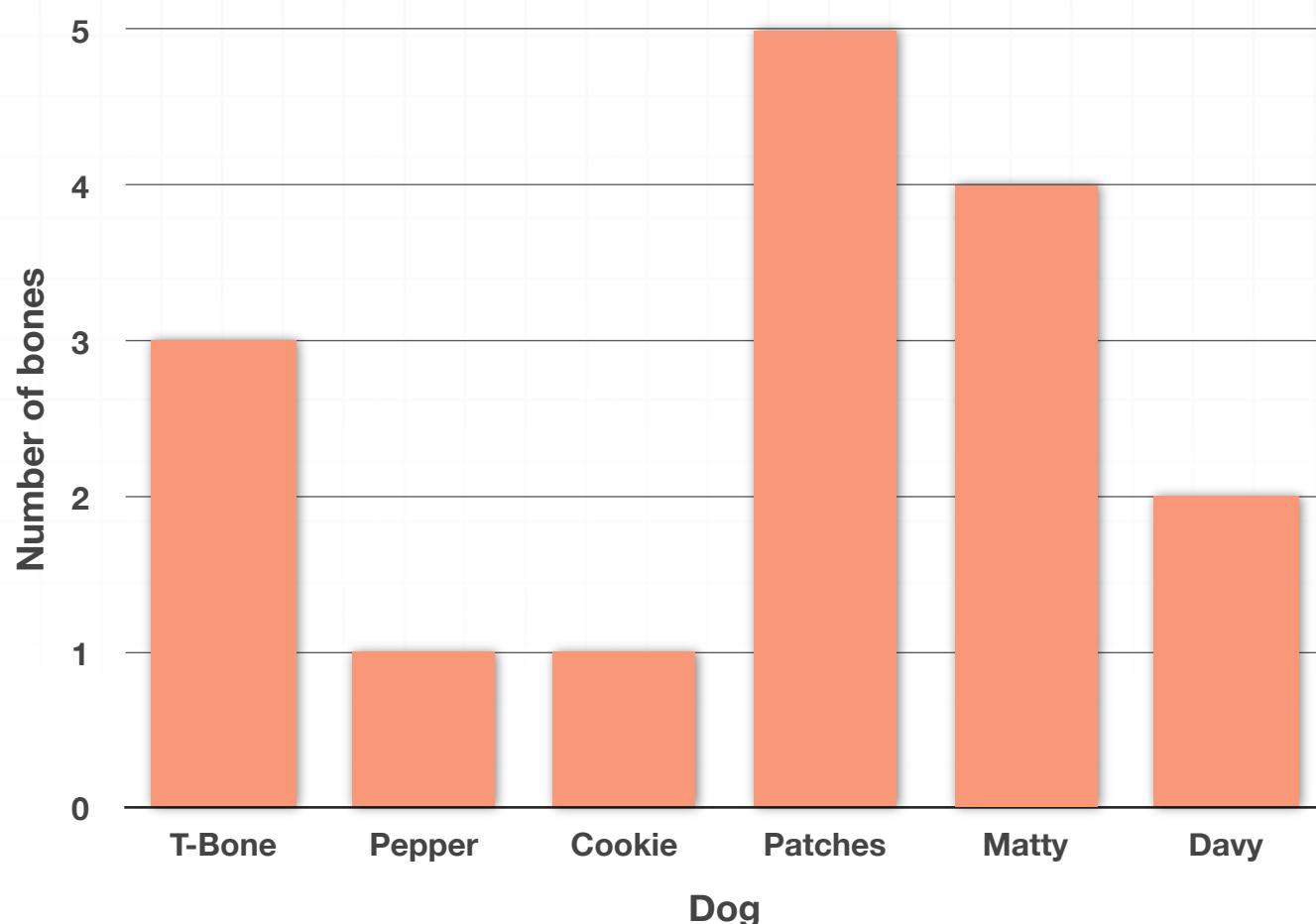
1. Both the bar graph and the table have missing information about the number of bones each dog consumed at doggie daycare. Use the graph and table together to fill in the missing pieces.

Dog	Number of bones
T-Bone	
Pepper	1
Cookie	
Patches	5
Matty	
Davy	2



*Solution:*

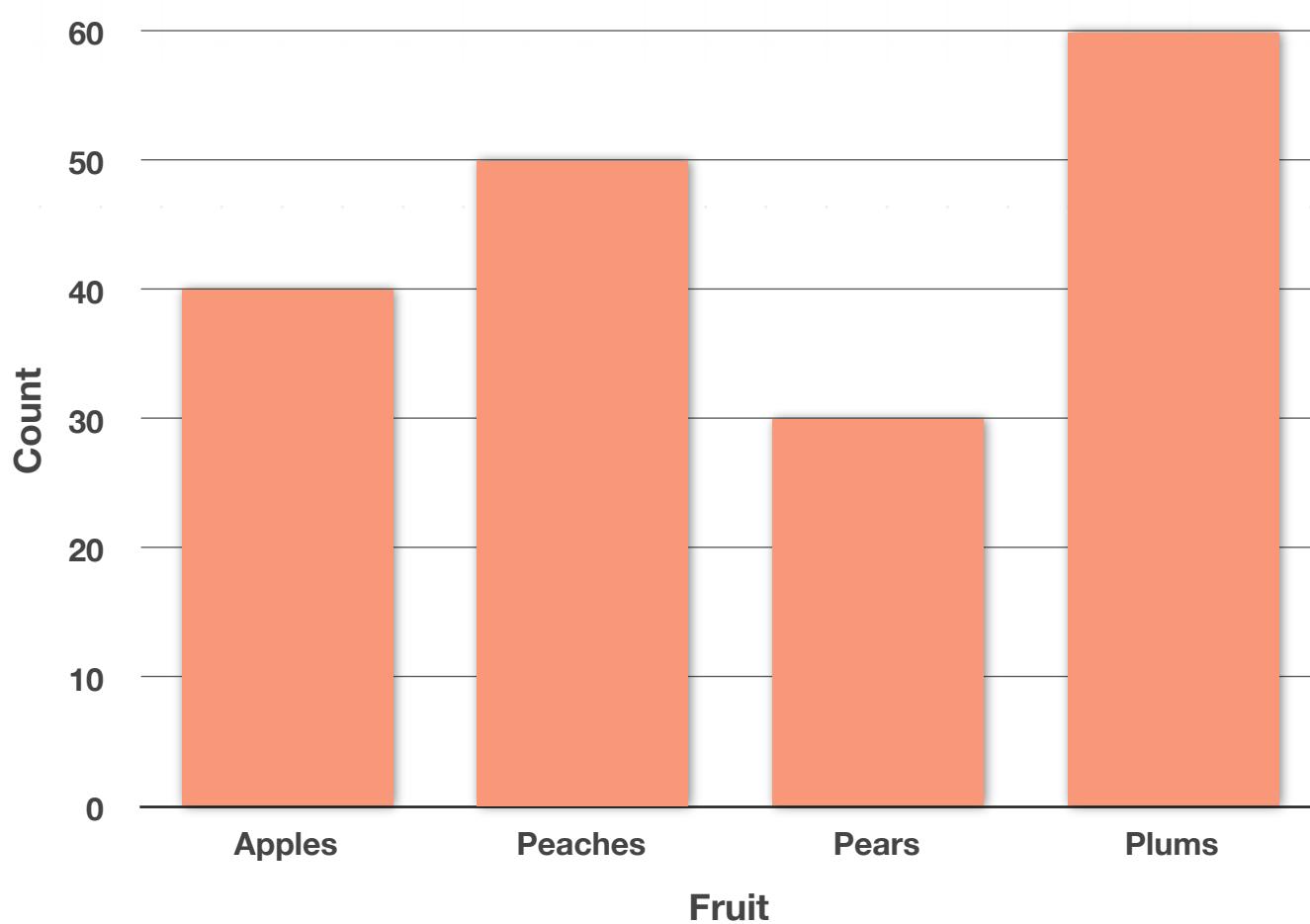
We can read from the table that Pepper ate 1 bone, Patches ate 5 bones, and Davy ate 2 bones. Therefore, the completed bar graph is



We can read from the bar graph that T-Bone ate 3 bones, Cookie ate 1 bone, and Matty ate 4 bones. Therefore, the completed table is

Dog	Number of bones
T-Bone	3
Pepper	1
Cookie	1
Patches	5
Matty	4
Davy	2

2. Eric's class went on a trip to an orchard. At the end of the trip they counted how many pieces of fruit came from each type of tree and graphed it in the bar graph shown below. Use the bar graph to create a pie chart of the data.



*Solution:*

To create a pie chart, we can divide the circle into fractional parts. We can see the students picked 40 apples, 50 peaches, 30 pears and 60 plums. That makes the total amount of fruit the class picked

$$40 + 50 + 30 + 60 = 180$$

The nice thing about this data is that it's all divisible by 10. We can therefore divide the pie chart into  $180 \div 10 = 18$  equal pieces, and then shade in the appropriate number of pieces for each fruit.

For apples:  $40 \div 10 = 4$

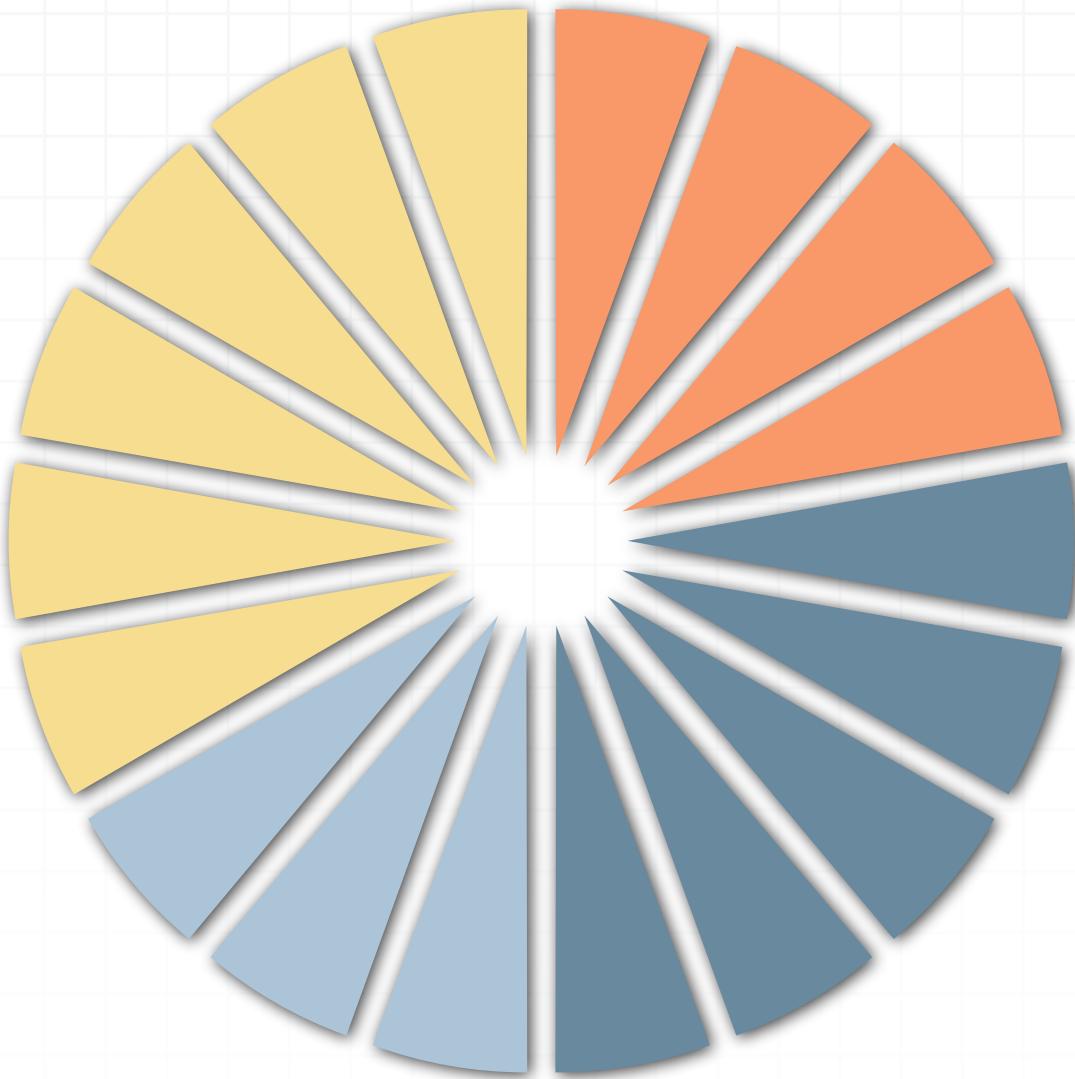
For peaches:  $50 \div 10 = 5$

For pears:  $30 \div 10 = 3$

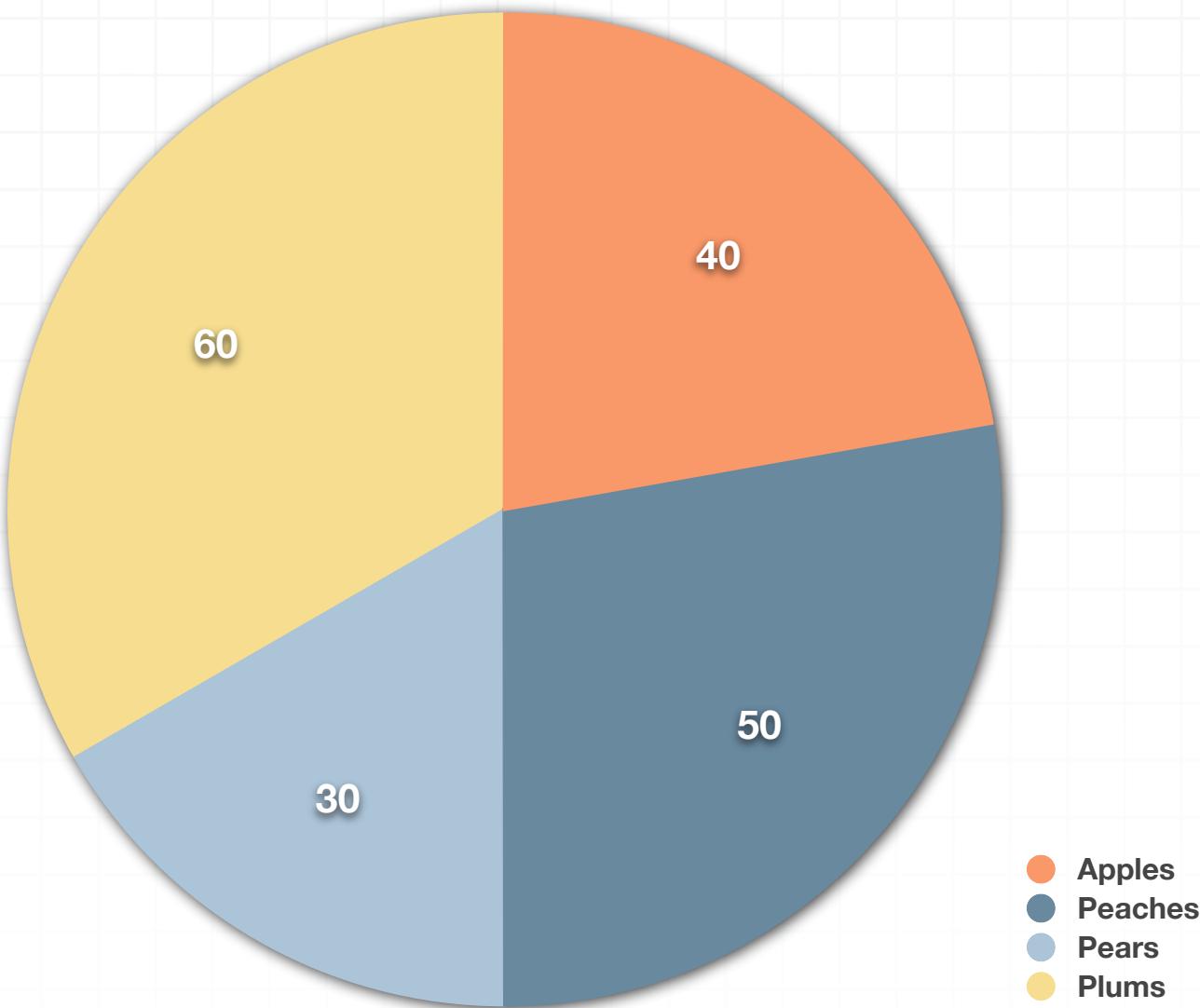
For plums:  $60 \div 10 = 6$

So if we use red for apples, dark blue for peaches, light blue for pears, and yellow for plums, we would shade 18 equal slices this way:

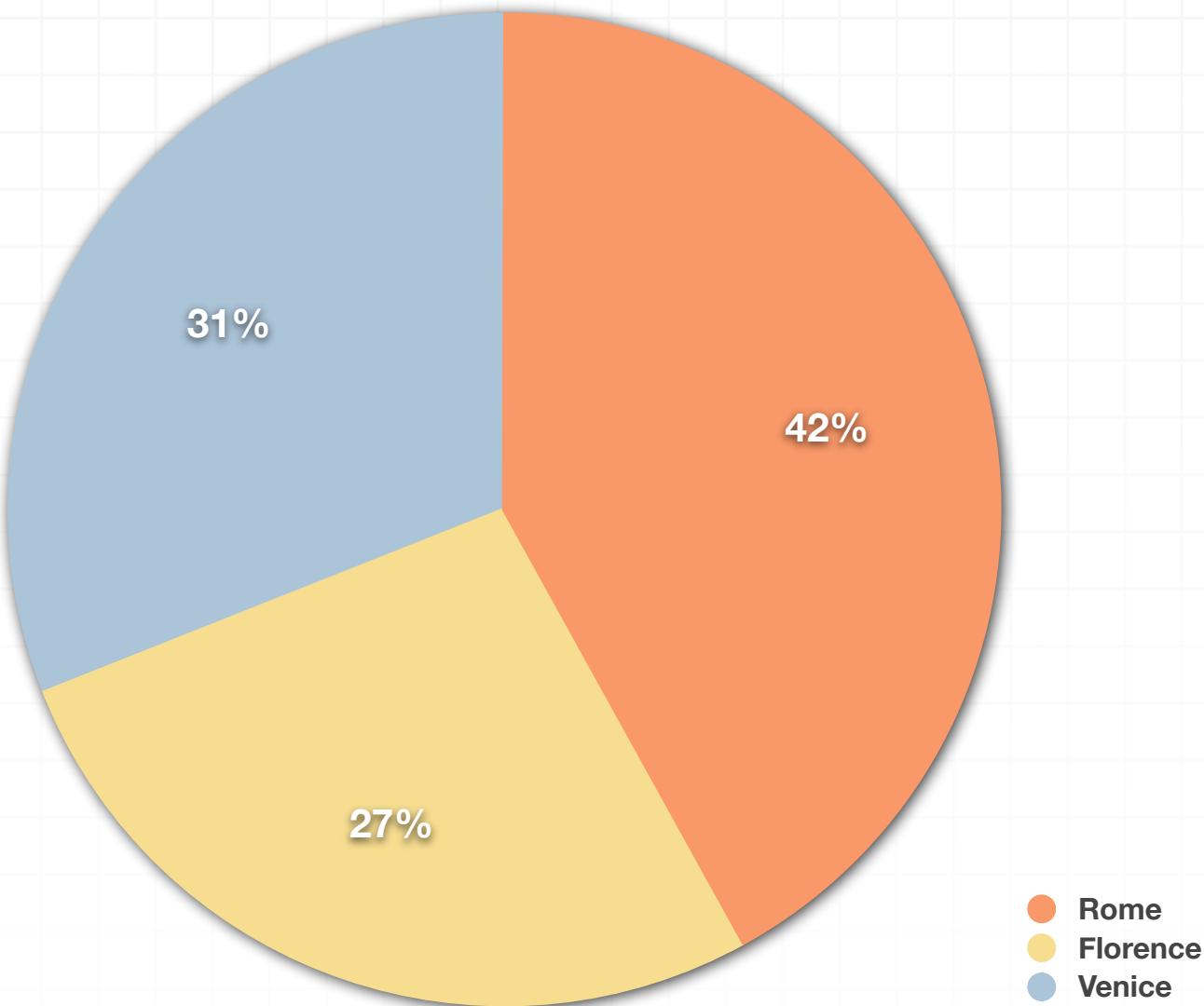




Then the finished pie chart is



3. A tourist company took a survey of 600 clients and asked them which Italian city they were most interested in visiting. How many clients said they wanted to visit Rome?

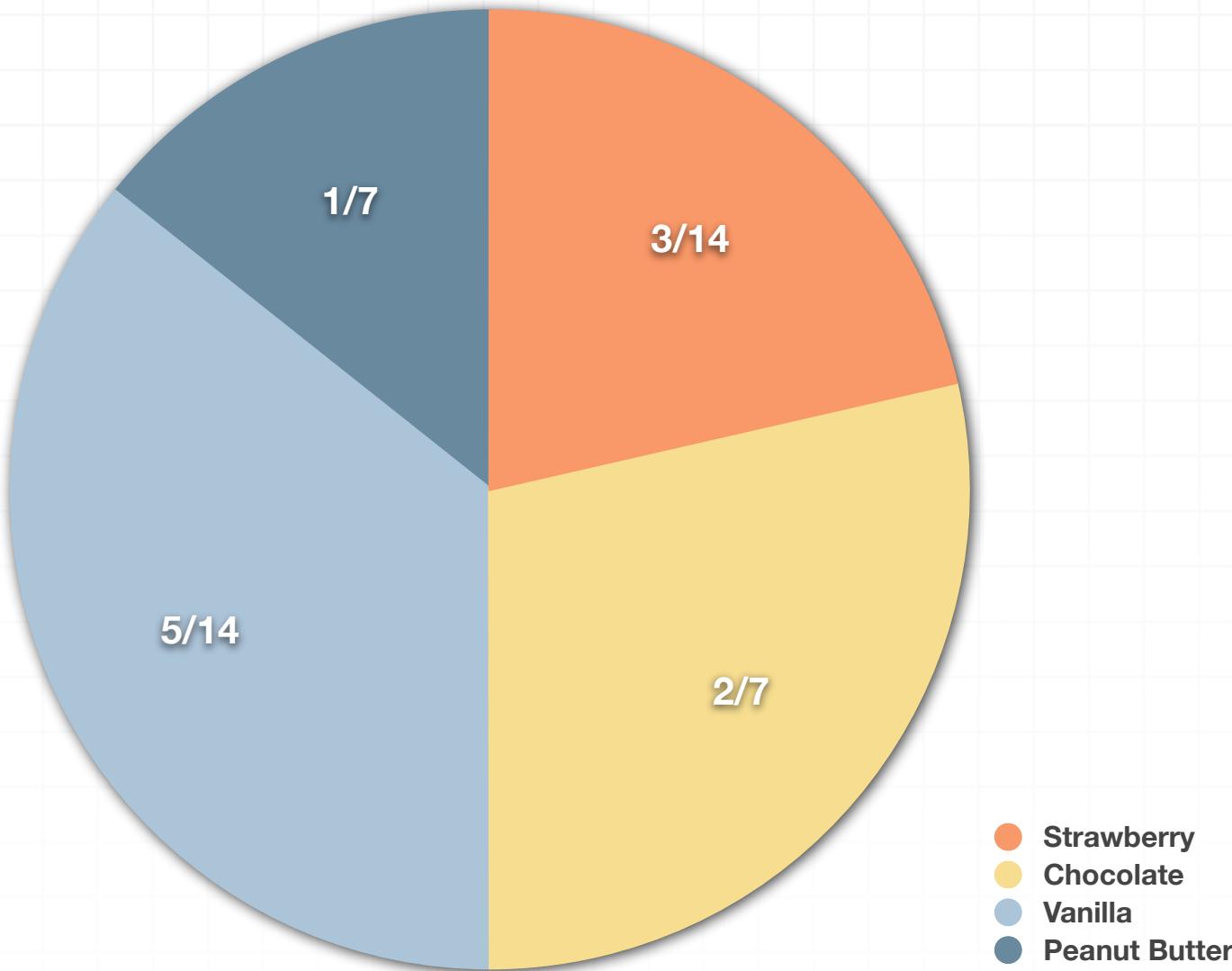


*Solution:*

Out of the 600 clients surveyed, 42 % of them said they wanted to visit Rome. 42 % of 600 is

$$600 \cdot 0.42 = 252 \text{ clients}$$

- 4. The pie chart shows how many ice cream cones of each flavor were sold. Assuming 280 total ice cream cones were sold in August, convert the pie chart to a bar graph.



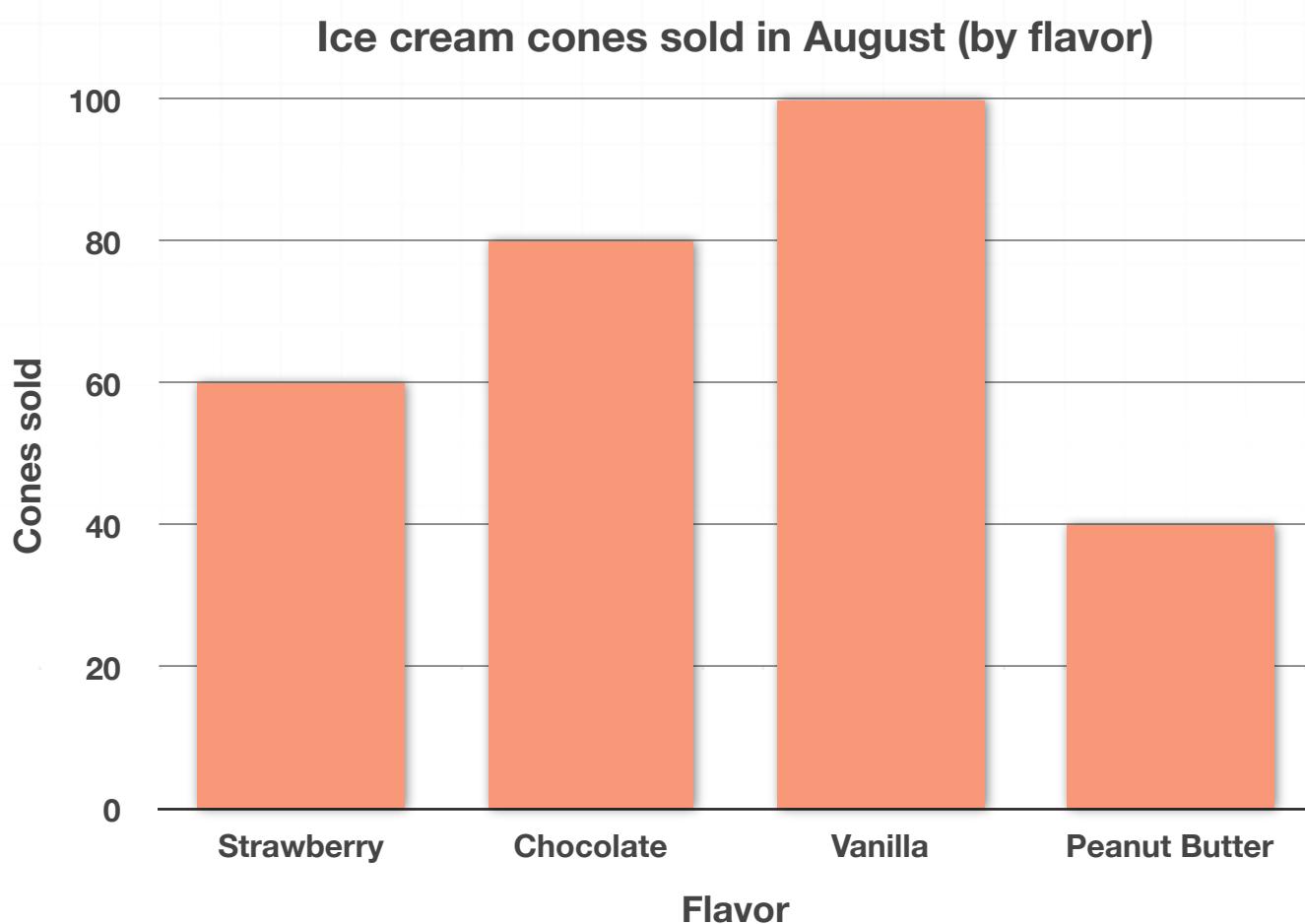
*Solution:*

We know that 280 ice cream cones were sold, and we have the amount of each flavor sold, as a fraction. For example, we know that  $\frac{3}{14}$  of the scoops of ice cream sold in August were strawberry.

Let's convert the information into a table first and also find the number of scoops sold of each type.

Flavor	Fraction	Cones sold
Strawberry	3/14	$(3/14)(280)=60$
Chocolate	2/7	$(2/7)(280)=80$
Vanilla	5/14	$(5/14)(280)=100$
Peanut Butter	1/7	$(1/7)(280)=40$

Now we can create a bar chart with the flavors on the horizontal axis and the number of cones sold on the vertical axis.



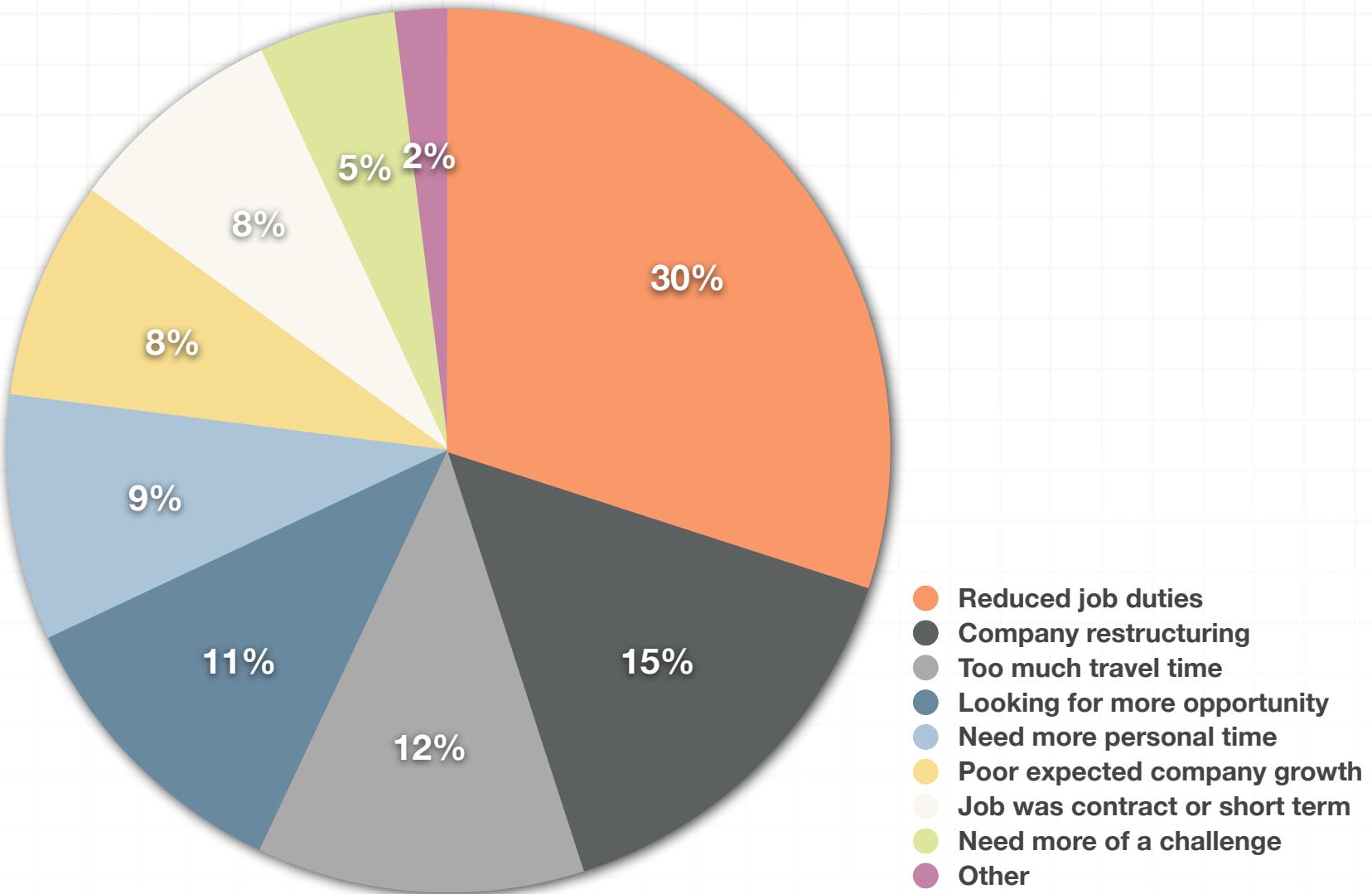
5. A company is analyzing the results from a recent survey about why people left their employment. The results are shown in the data table below. In general, is a bar graph or a pie chart a better choice to display the data? Why?

Reasons for leaving job	
Reduced job duties	30%
Company restructuring	15%
Too much travel time	12%
Looking for more opportunity	11%
Need more personal time	9%
Poor expected company growth	8%
Job was contract or short term	8%
Need more of a challenge	5%
Other	2%

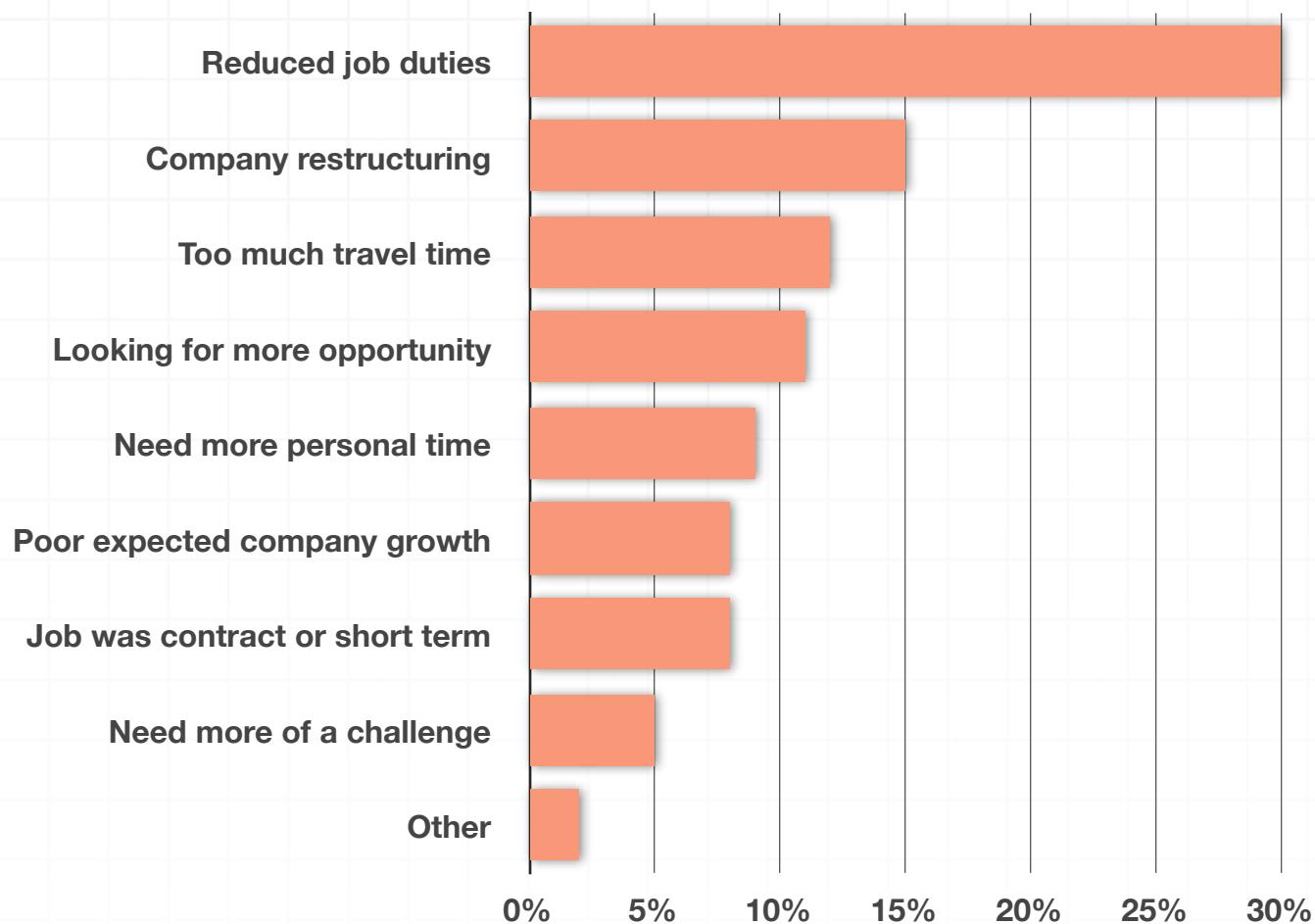
*Solution:*

A bar graph is a better choice to display the data because there are so many different categories. A pie chart can get cluttered when there are a lot of categories,

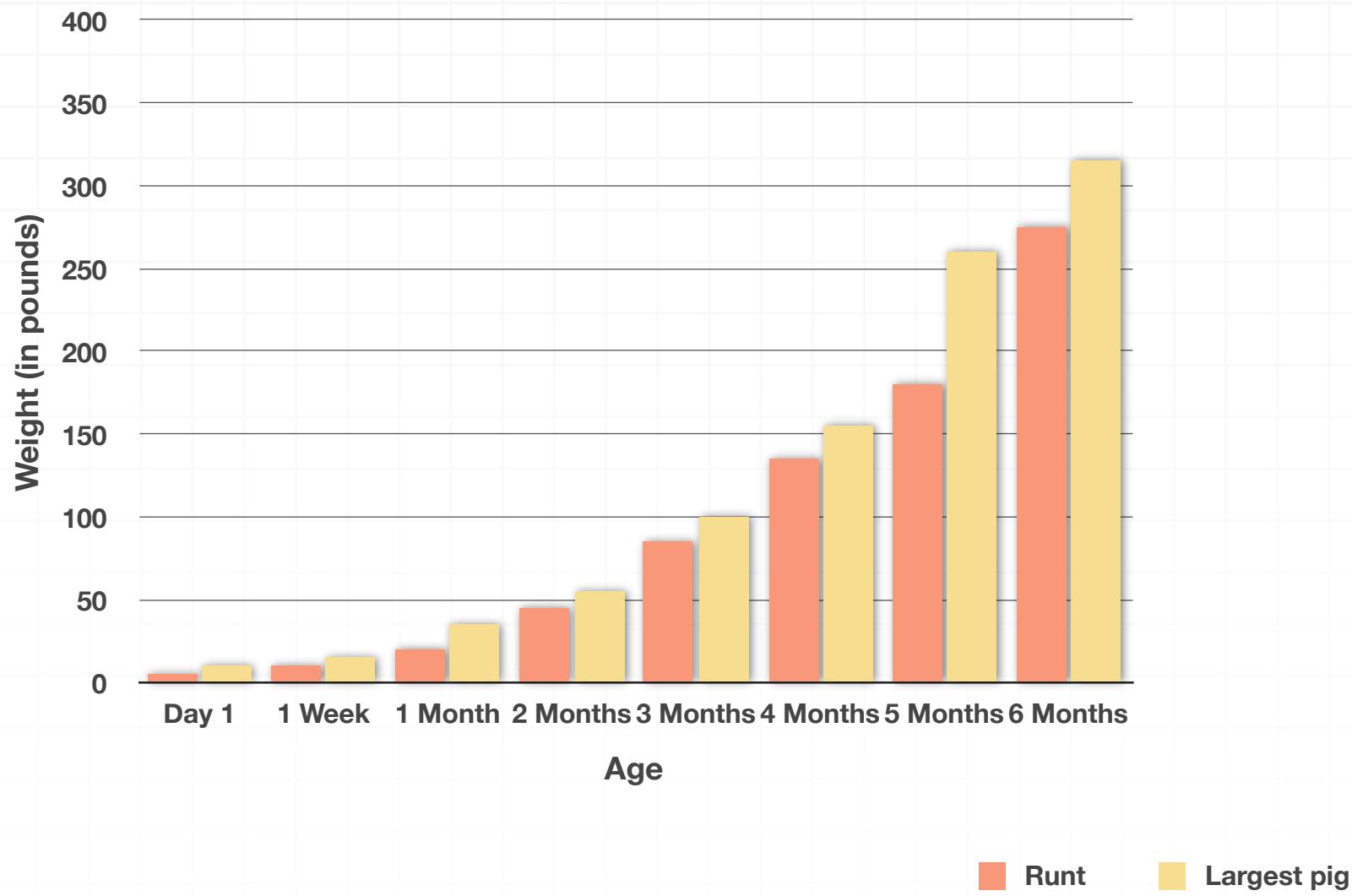




but a bar graph will remain fairly easy to read. Also notice that this is a good time to use a horizontal bar graph because the category titles are lengthy.



6. The comparison bar graph shows the growth of two pigs over their first 6 months of life. Which pig grew the most between 4 and 5 months?



*Solution:*

The largest pig in the litter grew from approximately 155 pounds to approximately 260 pounds, a change of about  $260 - 155 = 105$  pounds. The runt of the litter grew from approximately 135 pounds to approximately 175 pounds, a change of about  $175 - 135 = 40$  pounds. The largest pig in the litter grew much more than the runt.

## LINE GRAPHS AND OGIVES

- 1. Bethany started a sit-up program so that she can do 200 sit-ups in a day. At the end of week 6 she'll have completed 1,685 sit-ups. Create an ogive of the data.

Week	Number of sit-ups
Week 1	350
Week 2	455
Week 3	600
Week 4	540
Week 5	1,275
Week 6	1,685

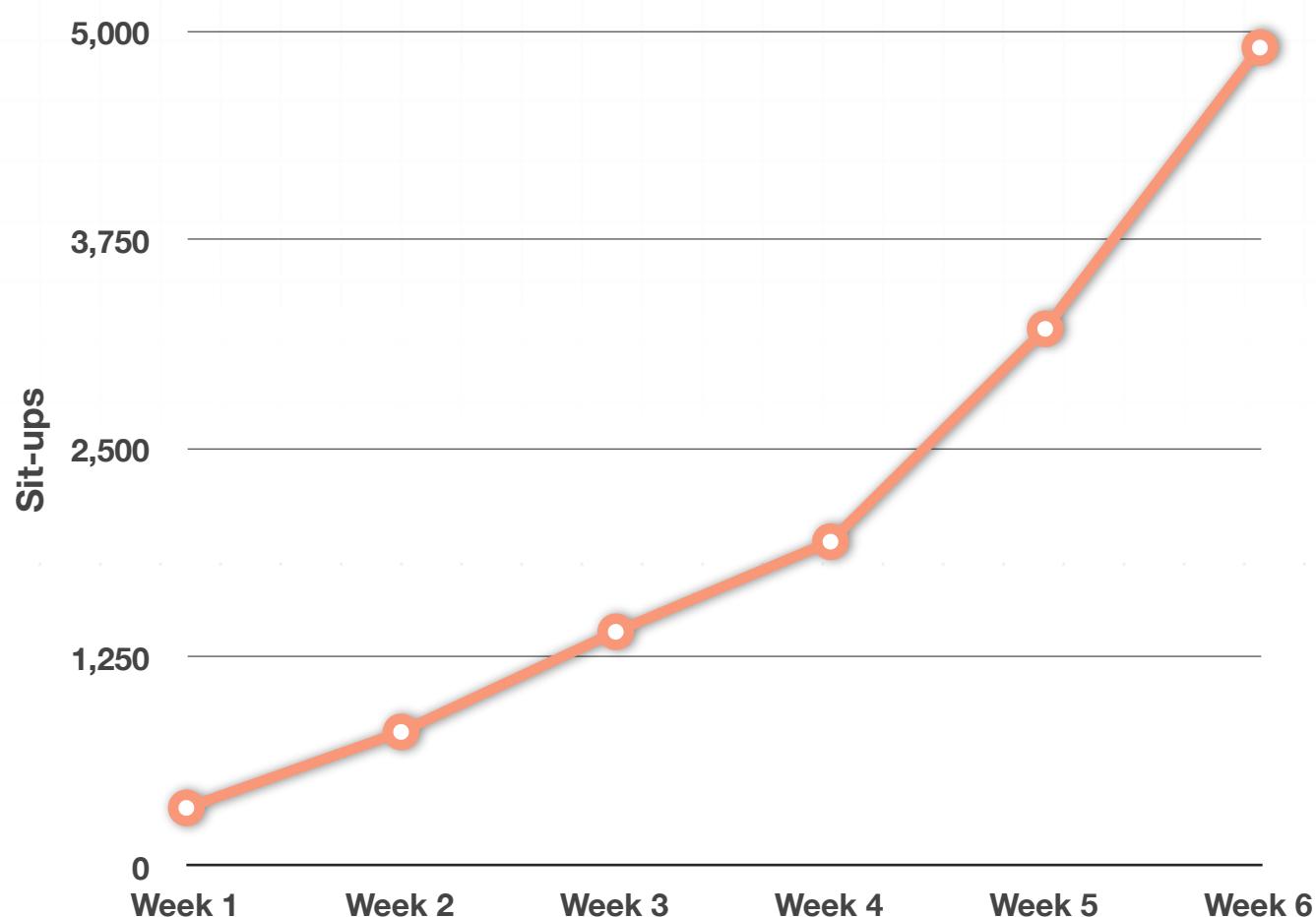
*Solution:*

We'll first add a total column to the table.



Week	Number of sit-ups	Total
Week 1	350	350
Week 2	455	805
Week 3	600	1,405
Week 4	540	1,945
Week 5	1,275	3,220
Week 6	1,685	4,905

Now, from the total column, we can create the ogive.

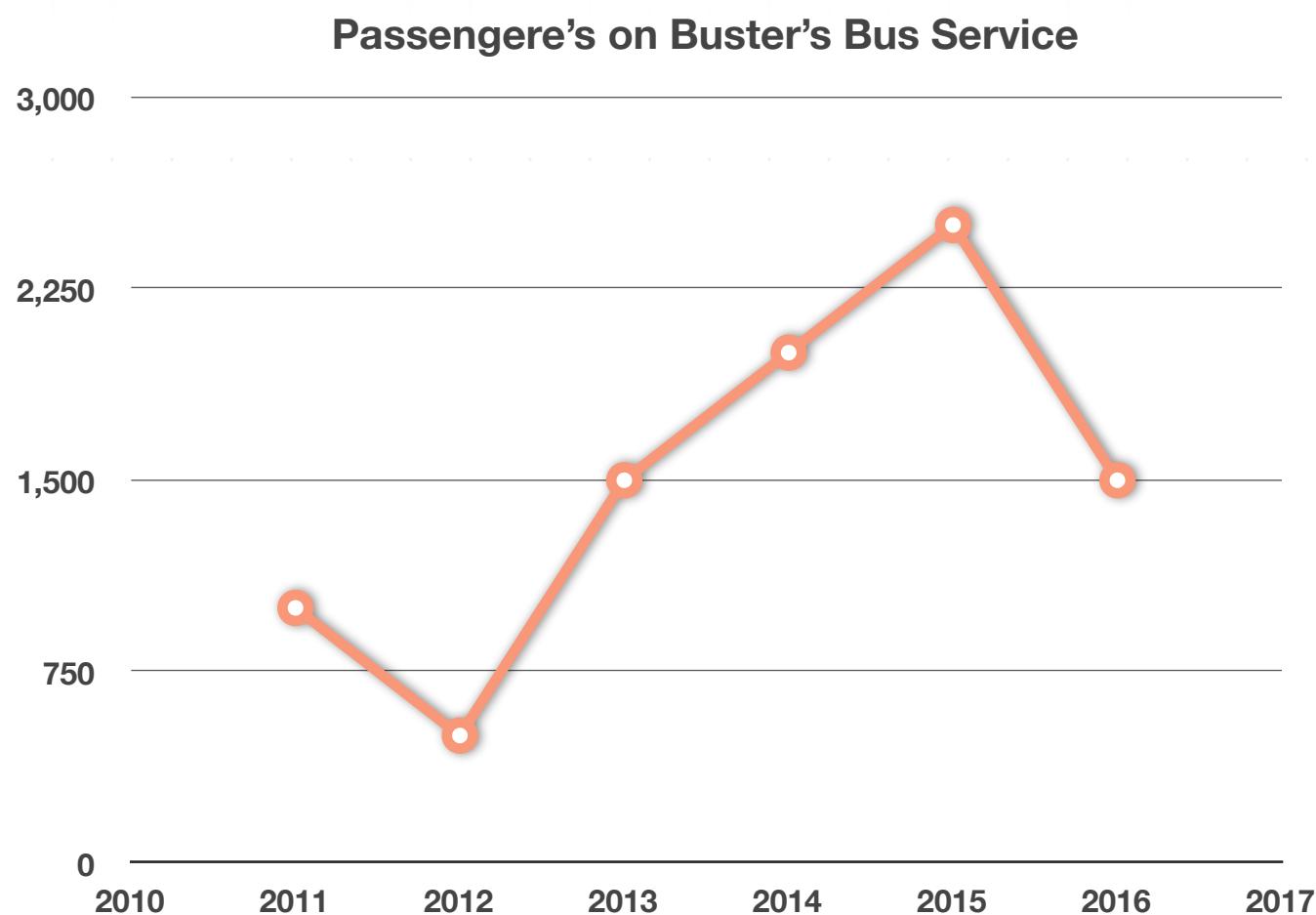


- 2. The table shows passengers by year for Buster's Bus Service. Create a line graph of the data in the table.

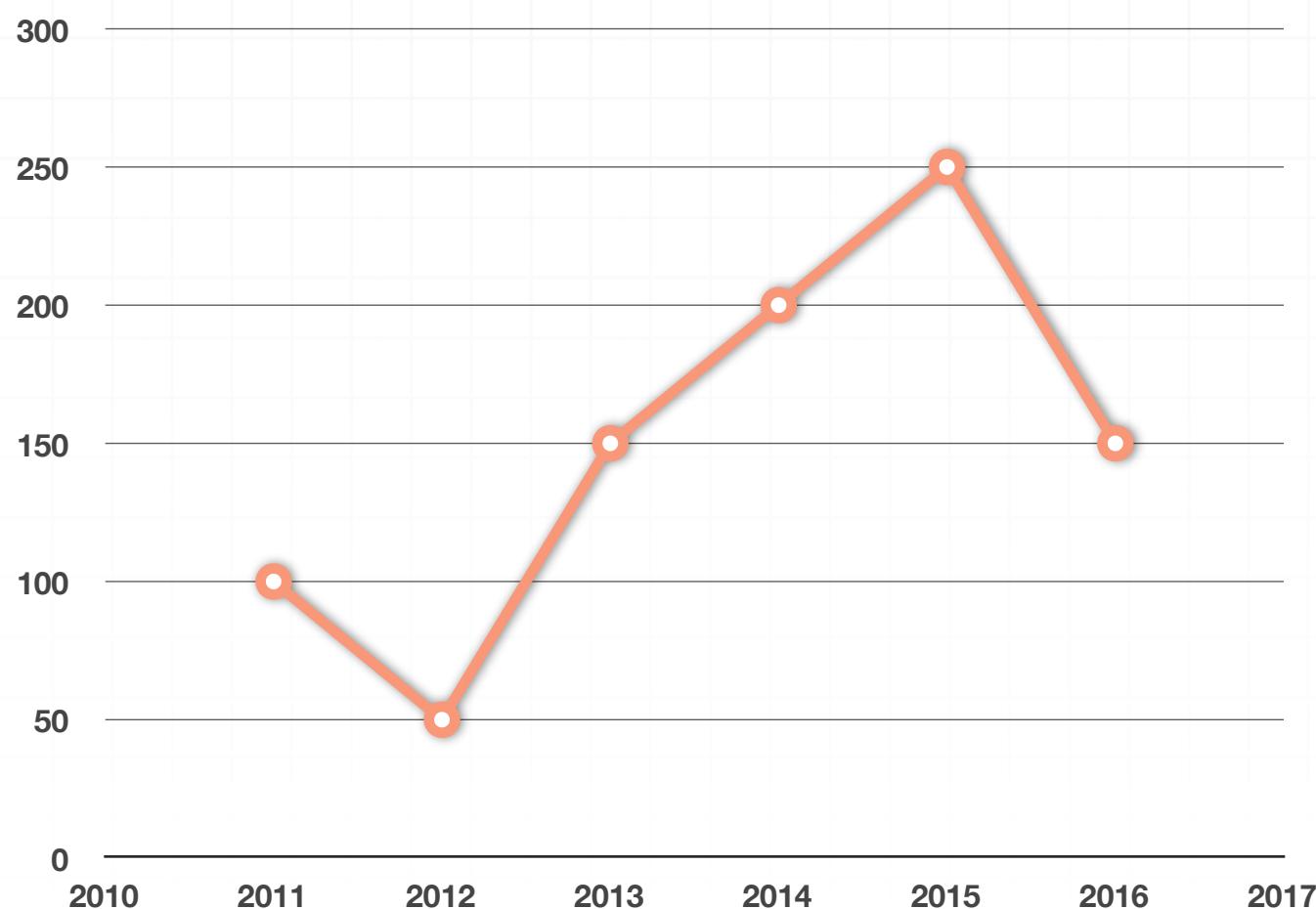
Year	Passengers
2011	1,000
2012	500
2013	1,500
2014	2,000
2015	2,500
2016	1,500

*Solution:*

We should start the line graph at the year 2010 and go by year along the horizontal axis to 2017, making sure we choose units on the vertical axis that make it easy to graph, as well as read.



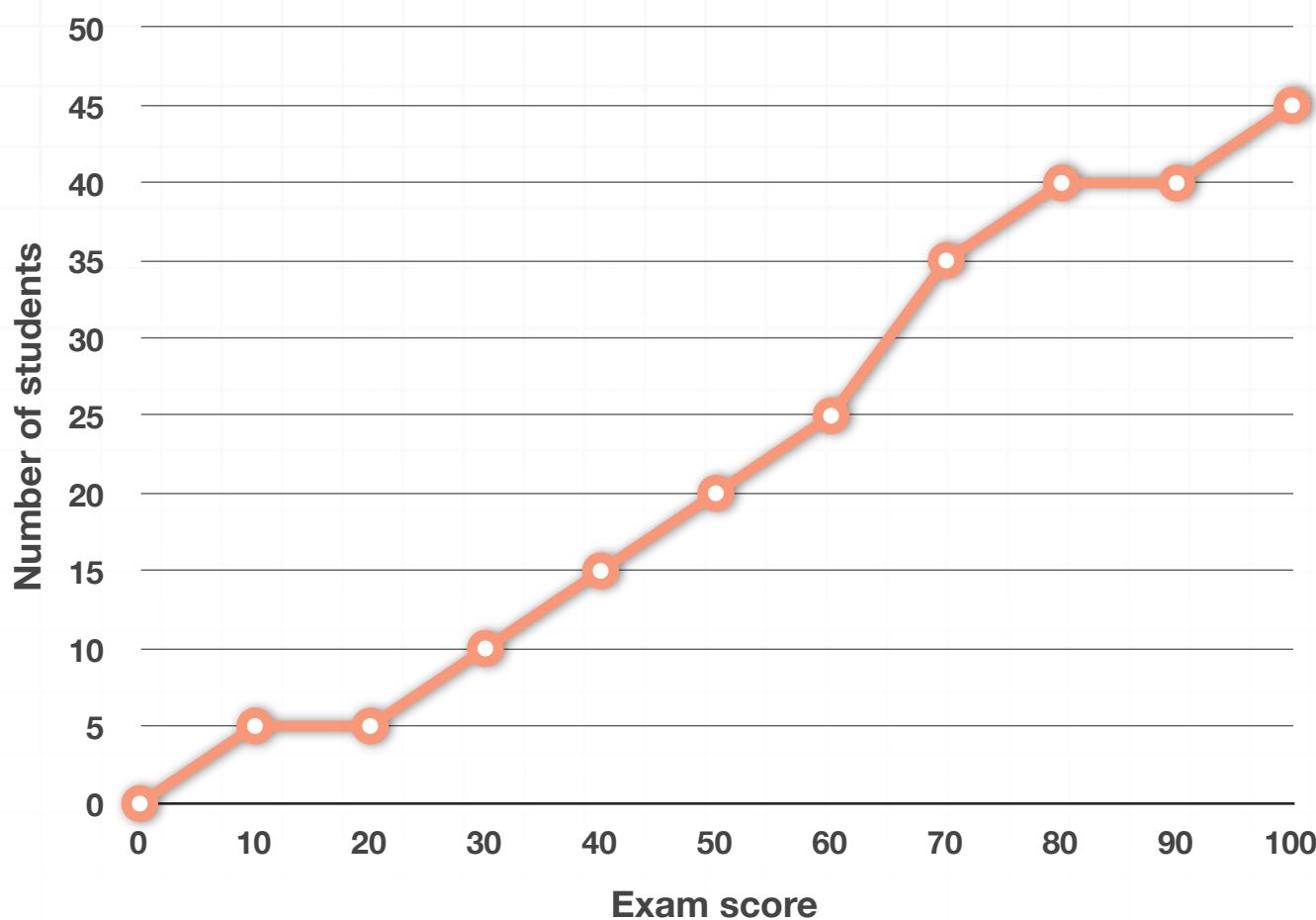
3. Between what two consecutive years was there the largest increase in car sales?



*Solution:*

The greatest increase in car sales was between 2012 and 2013. If we look at the line graph, we can see that the line increases at the sharpest rate between 2012 and 2013, so these are the years car sales increased the most.

4. Mrs. Moore gave her students a midterm exam, then she created this ogive of the 45 exam scores. How many students got a score between 70 % and 90 % ?



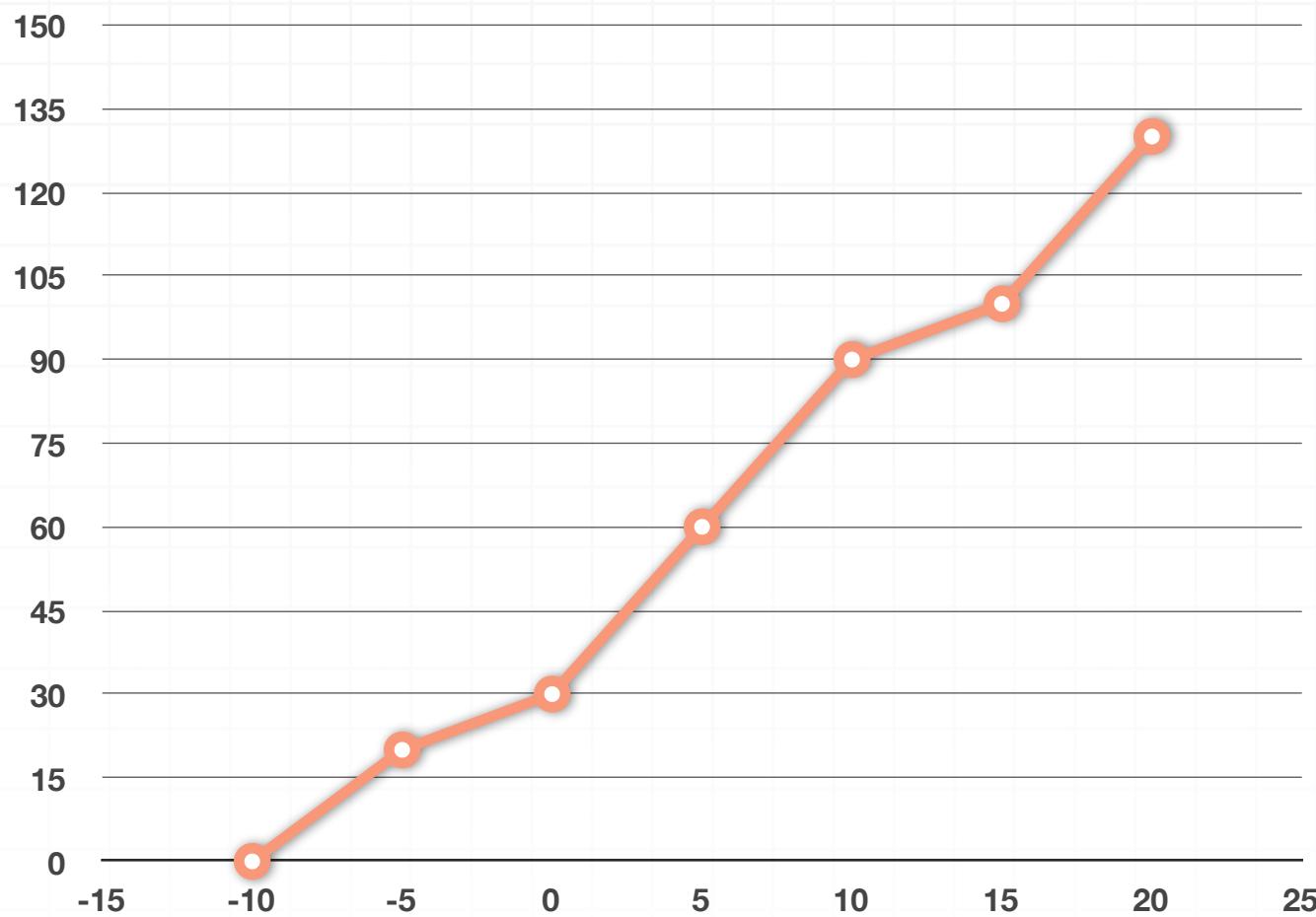
*Solution:*

We can tell from the ogive that 35 students scored lower than 70 % , and that 40 students scored lower than 90 %. Which means

$$40 - 35 = 5 \text{ students}$$

must have scored between 70 % and 90 % .

5. Draw the line graph that corresponds to the ogive below.



*Solution:*

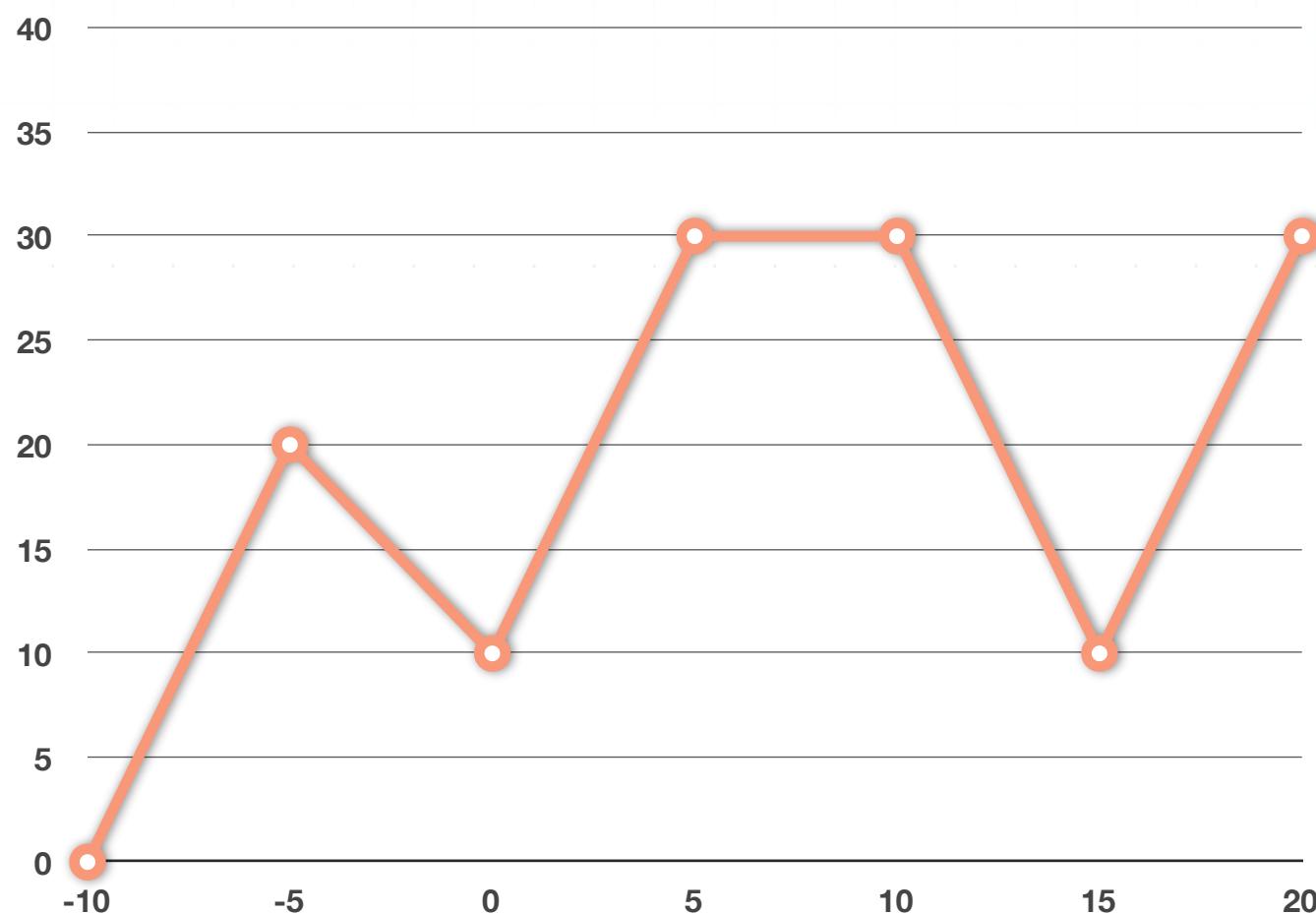
First we can create a table of the information from the ogive. The statistical name for the “total” is the “cumulative frequency.”

Horizontal value	Cumulative frequency
-10	0
-5	20
0	30
5	60
10	90
15	100
20	130

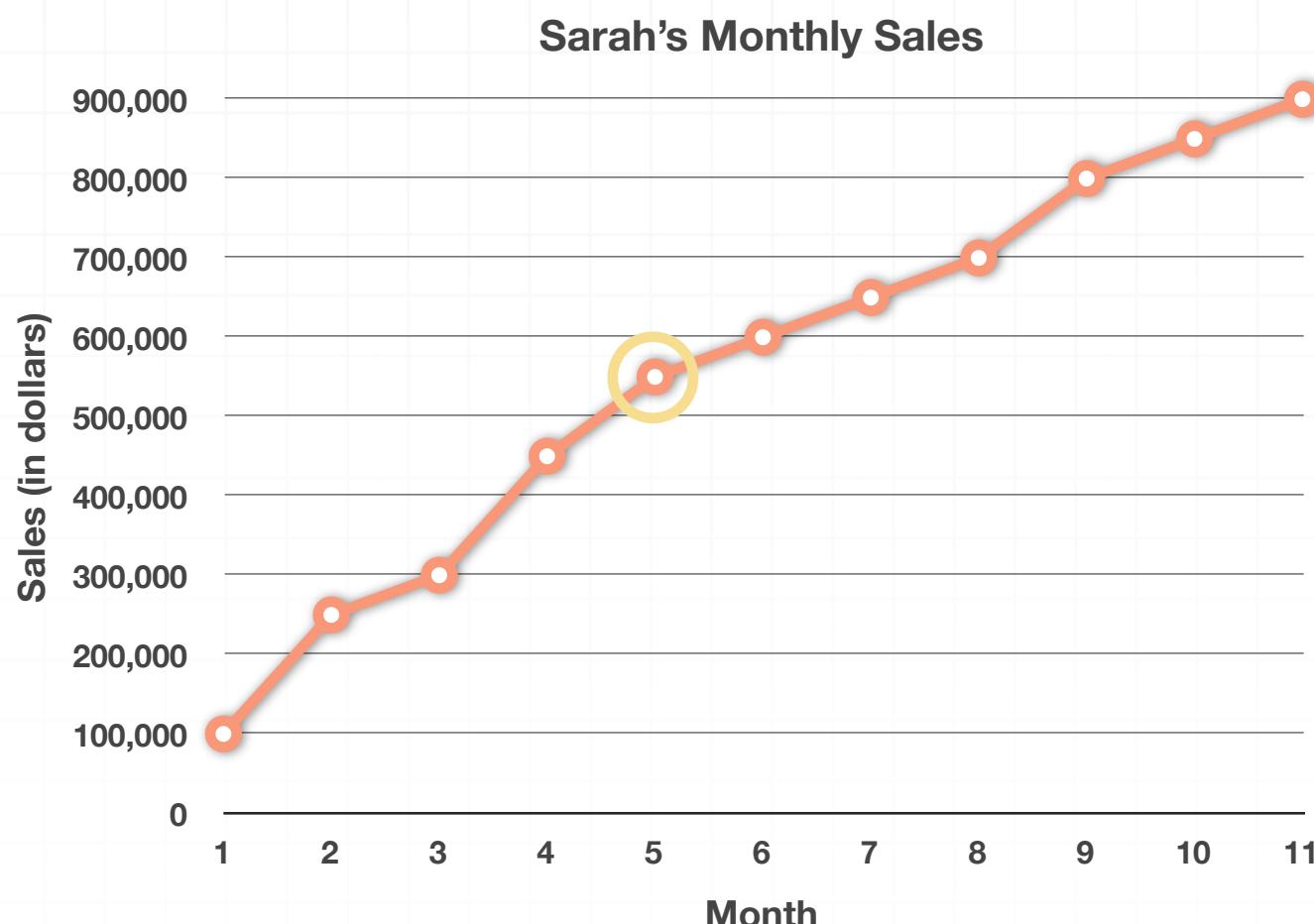
We can use the difference between the cumulative frequencies to find the frequency of each value and create a line graph.

Horizontal value	Cumulative frequency	Frequency
-10	0	0
-5	20	$20-0=20$
0	30	$30-20=10$
5	60	$60-30=30$
10	90	$90-60=30$
15	100	$100-90=10$
20	130	$130-100=30$

Now we'll create a line graph from the “frequency” column.



6. Sarah's monthly sales to date are shown in the ogive. What is the meaning of the circled point?



*Solution:*

By the time Sarah completed her fifth month at the company, she'd sold about \$550,000. This amounts to a little more than 60% of her total sales since she's worked at the company, because  $550/900 \approx 0.61$ .

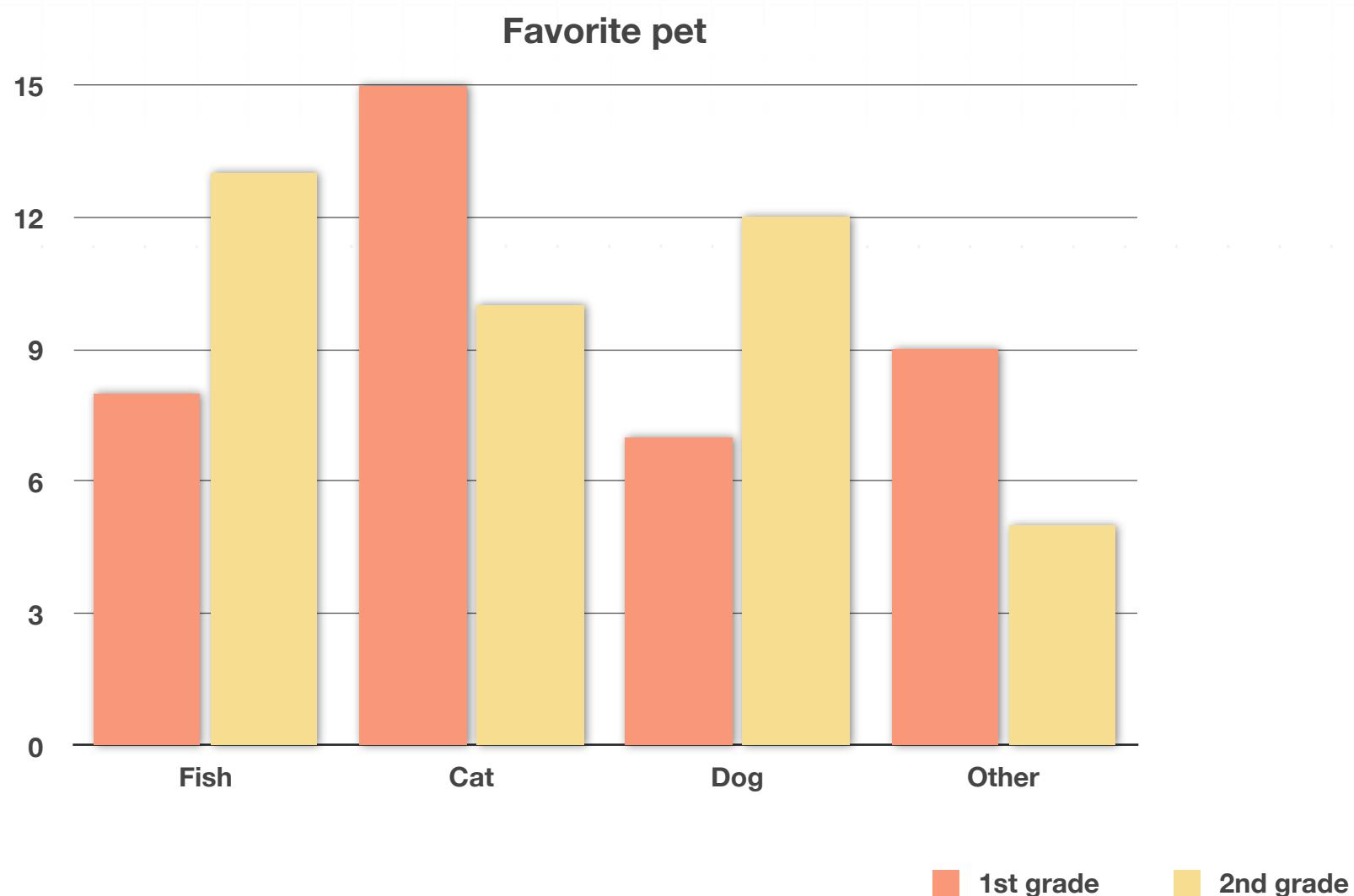
## TWO-WAY TABLES

1. Create a comparison bar graph for the two-way table.

Favorite pet	Fish	Cat	Dog	Other
1st grade	8	15	7	9
2nd grade	13	10	12	5

*Solution:*

We'll include a title, a key, start the vertical axis at 0, and plot the data.



2. A pizza parlor wants to know if the age range of their customers affects pizza preferences. The pizza parlor asks each customer two questions:

1. Which type of pizza is your favorite: pepperoni, cheese, supreme or veggie?
2. What is your age range: Under 18, or 18 and over?

The results of the survey are as follows:

Of the 50 customers who prefer pepperoni pizza, 25 are under 18.

Of the 20 customers who prefer cheese pizza, 18 are under 18.

Of the 30 customers who prefer supreme pizza, 24 are over 18.

Of the 25 customers who prefer veggie pizza, 19 are over 18.

Which type of table, one-way or two-way, can be created from the data that the pizza parlor is collecting? Create the best type of frequency table for the data.

*Solution:*

This is an example of data that can be organized into a two-way table because the data has two types of categories that can be organized together: age range and favorite pizza.



Since we're given the totals in each response, it can be easiest to start our table by filling in the totals of the people who like each type of pizza.

Favorite pizza	Pepperoni	Cheese	Supreme	Veggie	Total
Under 18					
18 and over					
Total	50	20	30	25	

Now we can use the rest of the information from each statement.

Favorite pizza	Pepperoni	Cheese	Supreme	Veggie	Total
Under 18	25	18			
18 and over			24	19	
Total	50	20	30	25	125

Now subtract each of these values from the totals to find the missing information.

Favorite pizza	Pepperoni	Cheese	Supreme	Veggie	Total
Under 18	25	18	30-24=6	25-19=6	
18 and over	50-25=25	20-18=2	24	19	
Total	50	20	30	25	125

Now find the totals to complete the table.

Favorite pizza	Pepperoni	Cheese	Supreme	Veggie	Total
Under 18	25	18	6	6	$25+18+6+6=55$
18 and over	25	2	24	19	$25+2+24+19=70$
Total	50	20	30	25	$55+70=125$

Therefore, the finished table is

Favorite pizza	Pepperoni	Cheese	Supreme	Veggie	Total
Under 18	25	18	6	6	55
18 and over	25	2	24	19	70
Total	50	20	30	25	125

- 3. An elementary school creates the following two-way table. What is the best name for the row variable and what is the best name for the column variable?

	Walk	School bus	Day care vehicle	Carpool
Pre-school	1	10	20	26
First	5	12	14	19
Second	10	22	5	15
Third	8	33	3	10

*Solution:*



Row or column variables are how we explain each section of the two-way table. The row variable describes the data in each row, and the column variable explains the data in each column. “Method of transportation” is one possible description for the column variable. “Student grade” is a possible description for the row variable.

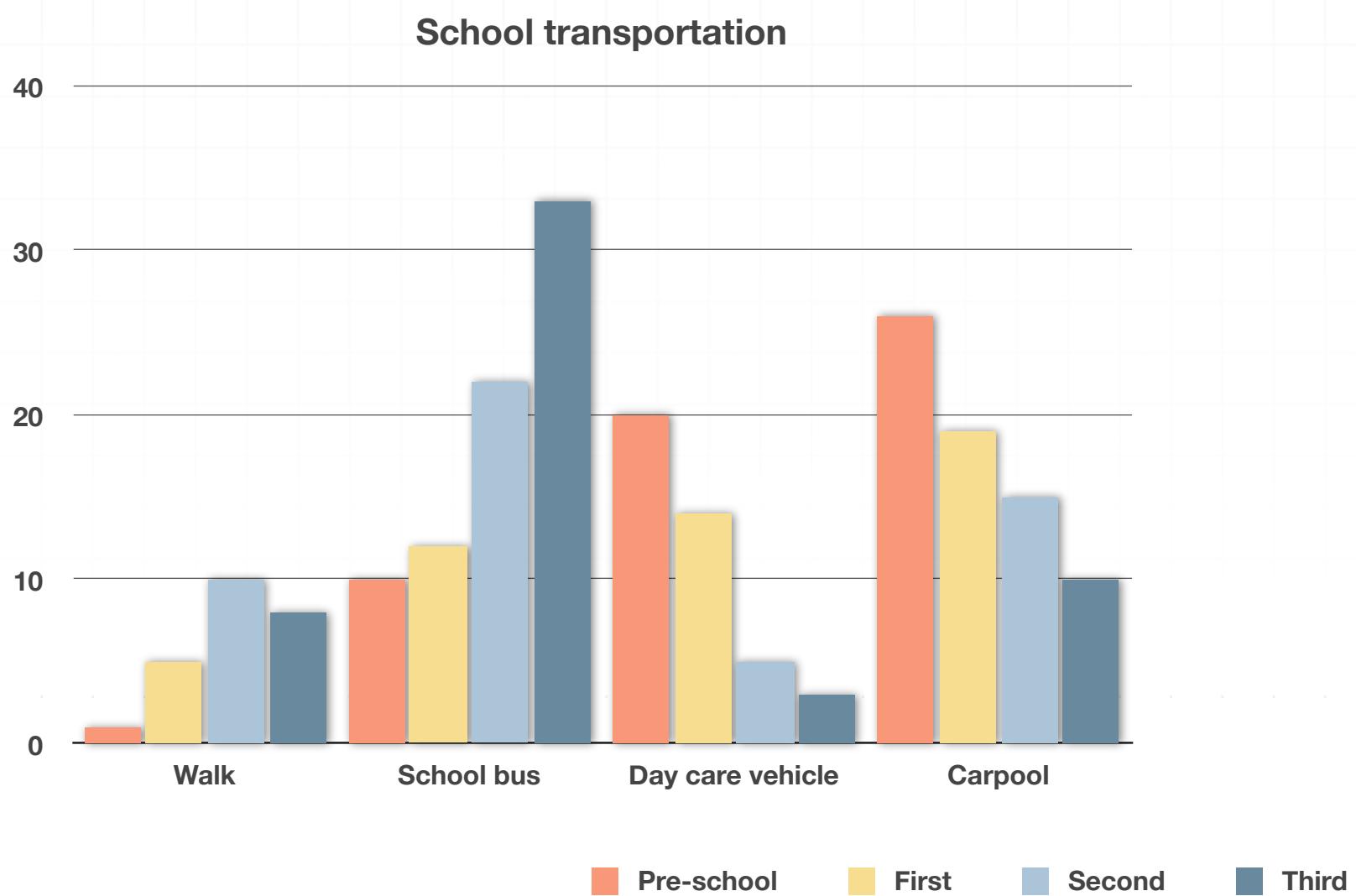
		Method of transportation			
		Walk	School bus	Day care vehicle	Carpool
Grade in school	Pre-school	1	10	20	26
	First	5	12	14	19
	Second	10	22	5	15
	Third	8	33	3	10

- 4. Decide whether a comparison bar graph or comparison line graph would be better at displaying the data in the two-way table, then create the graph.

		Method of transportation			
		Walk	School bus	Day care vehicle	Carpool
Grade in school	Pre-school	1	10	20	26
	First	5	12	14	19
	Second	10	22	5	15
	Third	8	33	3	10

**Solution:**

A comparison bar graph is the best choice for the data because a comparison line graph is used to show changes over time. We're comparing grades in school, so the comparison bar graph is the best choice. Remember, when we create a comparison bar graph, we need to include the title, key, and a reasonable scale on the vertical axis.



- 5. Eric creates a survey asking students who ate a snack in the morning between classes if they felt sleepy or not. Given his survey results below, create a two-way data table for Eric's survey.

<b>Snack</b>	Yes	Yes	No	No	No	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No
<b>Sleepy</b>	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes

*Solution:*

We could set up the table this way:

		Do you feel sleepy?		
		Yes	No	Total
Did you eat a snack?	Yes			
	No			
	Total			

There are 5 people who ate a snack and feel sleepy. There are 2 people who ate a snack but don't feel sleepy. There are 3 people who didn't eat a snack and feel sleepy. And there are 5 people who didn't eat a snack and don't feel sleepy.

		Do you feel sleepy?		
		Yes	No	Total
Did you eat a snack?	Yes	5	2	
	No	3	5	
	Total			

Now we just total everything up.

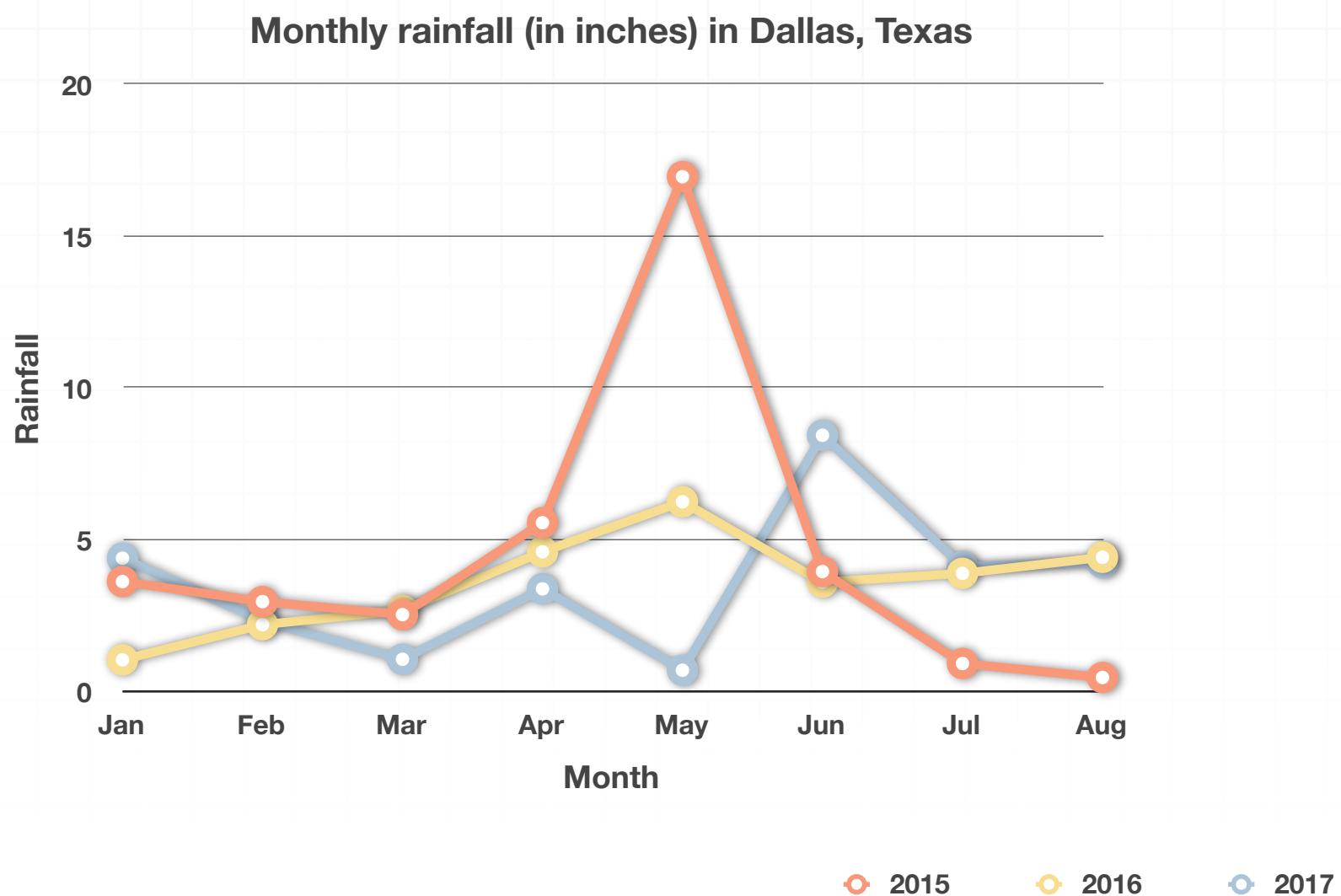
		Do you feel sleepy?		
		Yes	No	Total
Did you eat a snack?	Yes	5	2	7
	No	3	5	8
	Total	8	7	15

- 6. Is a comparison line graph an appropriate visual display for the data table, which shows monthly rainfall (in inches) for Dallas, Texas, January - August? Why or why not? If it's an appropriate display, create a comparison line graph. If it's not an appropriate display for the data, create a comparison bar graph.

	2015	2016	2017
January	3.62	1.04	4.39
February	2.96	2.20	2.33
March	2.53	2.67	1.06
April	5.56	4.60	3.38
May	16.96	6.25	0.70
June	3.95	3.60	8.44
July	0.92	3.89	4.12
August	0.46	4.42	4.24

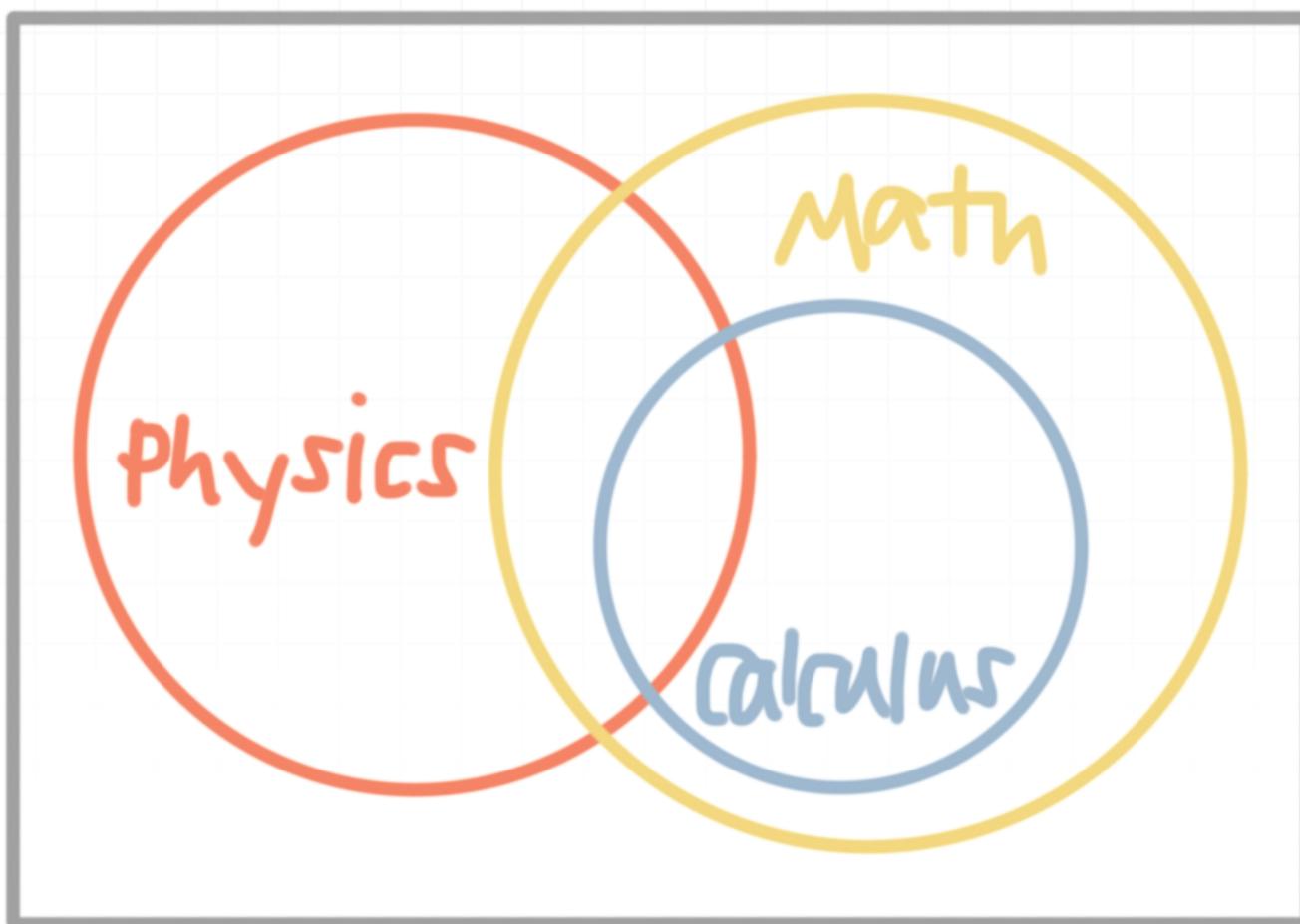
*Solution:*

Yes, a comparison line graph is an appropriate visual display for the data because it would be useful to track rainfall in Dallas over a given time period.



## VENN DIAGRAMS

- 1. What does the Venn diagram show about how Calculus is related to Physics and Mathematics?



*Solution:*

We can see from the Venn diagram that all of Calculus is a subset of Mathematics. Some Calculus is part of Physics, although there's also some Mathematics in Physics that does not include Calculus.

- 2. Draw the Venn diagram for the number of humans in a room and the number of frogs in a room, if the room has 12 frogs and 15 humans.

*Solution:*

Notice that it's possible for a Venn diagram to have no overlapping parts. In this case, the Frogs and Humans do not share any characteristics we're interested in, so the two circles do not overlap.



- 3. Students at Green Bow High School conducted a survey during lunch time to see what kind of music the students at the school liked. They recorded their results in a Venn diagram. How many students participated

in the survey? What percentage of the students who participated did not like Pop Music?



*Solution:*

Add up all of the data in the Venn diagram, and don't forget the 6 on the outside.

$$23 + 6 + 5 + 12 + 21 + 2 + 16 + 6 = 91$$

This is the number of students who participated in the survey. The students who did not like Pop music are those who only liked Rap (21), Country (16), Country and Rap (2), or something else (6). That adds to

$$21 + 2 + 16 + 6 = 45$$

The total number of students who participated in the survey was 91. Therefore, the percentage who didn't like Pop is then

$$\frac{45}{91} \approx 49\%$$

- 4. A survey team is collecting data on a type of minnow that lives where a river meets the sea. They place nets in the river, where the river and sea meet and where there is only sea. They count the minnows caught in each net. What percent of the minnows were living in the brackish water? Brackish water is water that's a combination of fresh and saltwater.



*Solution:*

The total minnows caught in the sample was  $22 + 51 + 29 = 102$ . We can read in the overlap that 51 of the minnows were caught in the brackish water since it's a combination of the saltwater and freshwater. Now we can calculate the percentage as  $51/102 = 0.50 = 50\%$ . This means 50% of the minnows were from brackish water.

■ 5. Fill in the Venn diagram using the following information.

18 people's favorite exercise was swimming.

13 people's favorite exercise was running.

10 people only liked weight lifting.

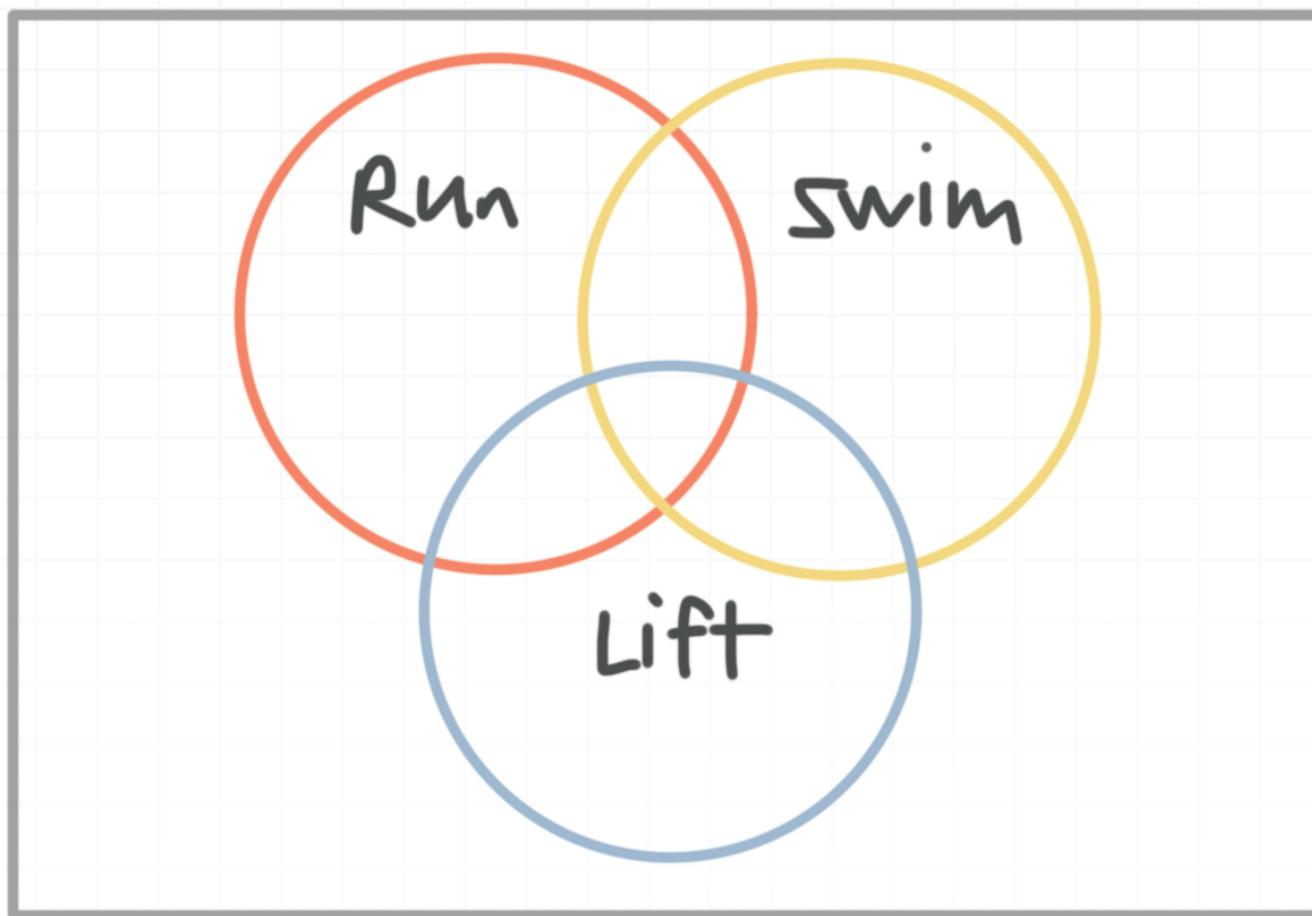
3 people liked swimming and weight lifting equally, but not running.

4 people liked running and weight lifting equally, but not swimming.

5 people liked running and swimming equally, but not weight lifting.

2 people liked all three equally.





*Solution:*

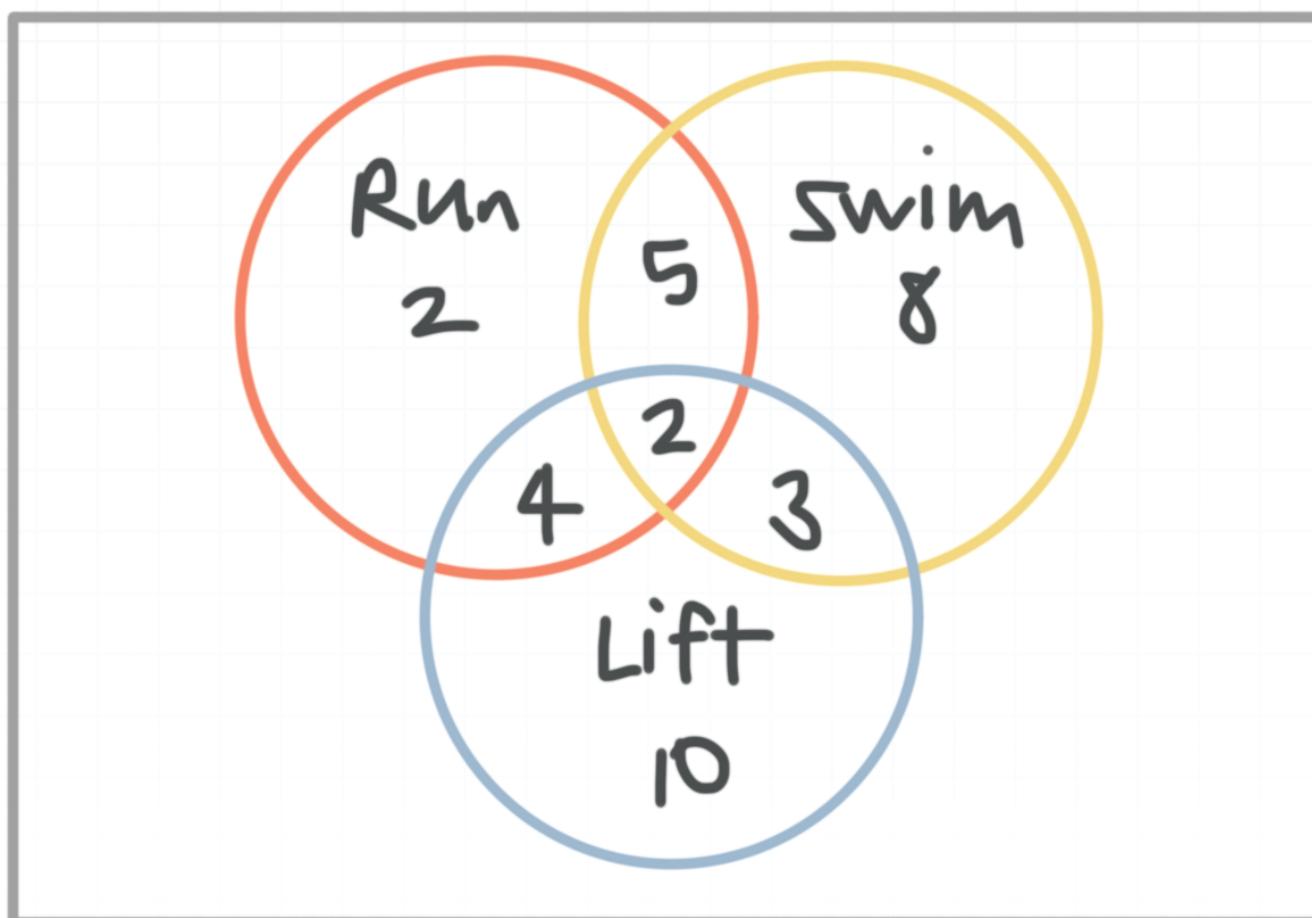
Since 10 people liked weight lifting only, we can put a 10 into the part of the weight lifting circle that doesn't intersect with the other two circles. We can put a 3 into the intersection between weight lifting and swimming, and we can put a 5 into the intersection between running and swimming. We can put a 2 into the intersection of all three circles.

The number of people who liked swimming only wouldn't be 18, since some people who liked swimming also liked either running, weight lifting, or both. Therefore, we need to put  $18 - 5 - 2 - 3 = 8$  into the part of the swimming circle that doesn't intersect with the other two circles.

The same thing is happening with people who liked running. Out of those whose favorite exercise was running, 5 people also liked swimming, 4

people also liked weight lifting, and 2 people liked all three exercises. Therefore,  $13 - 5 - 4 - 2 = 2$  people liked running only.

From the information we were given, the Venn diagram is

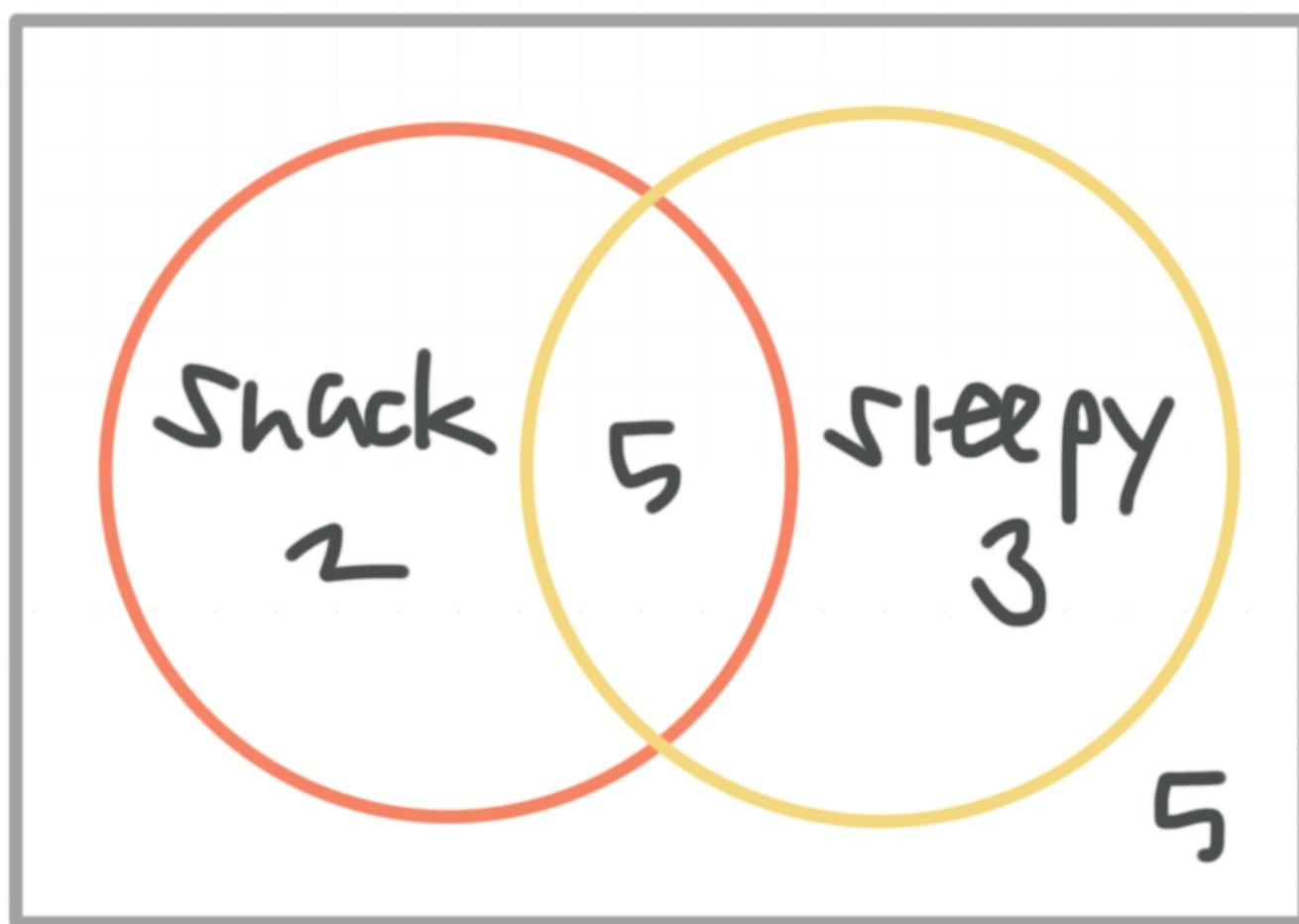


6. Eric creates a survey asking students who ate a snack in the morning between classes if they felt sleepy or not. He organizes his survey results into a two-way data table. Draw a Venn diagram for Eric's survey results.

		Do you feel sleepy?		
		Yes	No	Total
Did you eat a snack?	Yes	5	2	7
	No	3	5	8
	Total	8	7	15

*Solution:*

There are 5 students who ate a snack and feel sleepy, so we'll put a 5 in the middle. There are 3 students who didn't eat a snack but feel sleepy, so we'll put a 3 in the "feel sleepy" circle. There are 2 students who had a snack but don't feel sleepy, so we'll put a 2 in the "had a snack" circle. And there are 5 students who didn't have a snack and don't feel sleepy, so we'll put a 5 outside of both circles.



## FREQUENCY TABLES AND DOT PLOTS

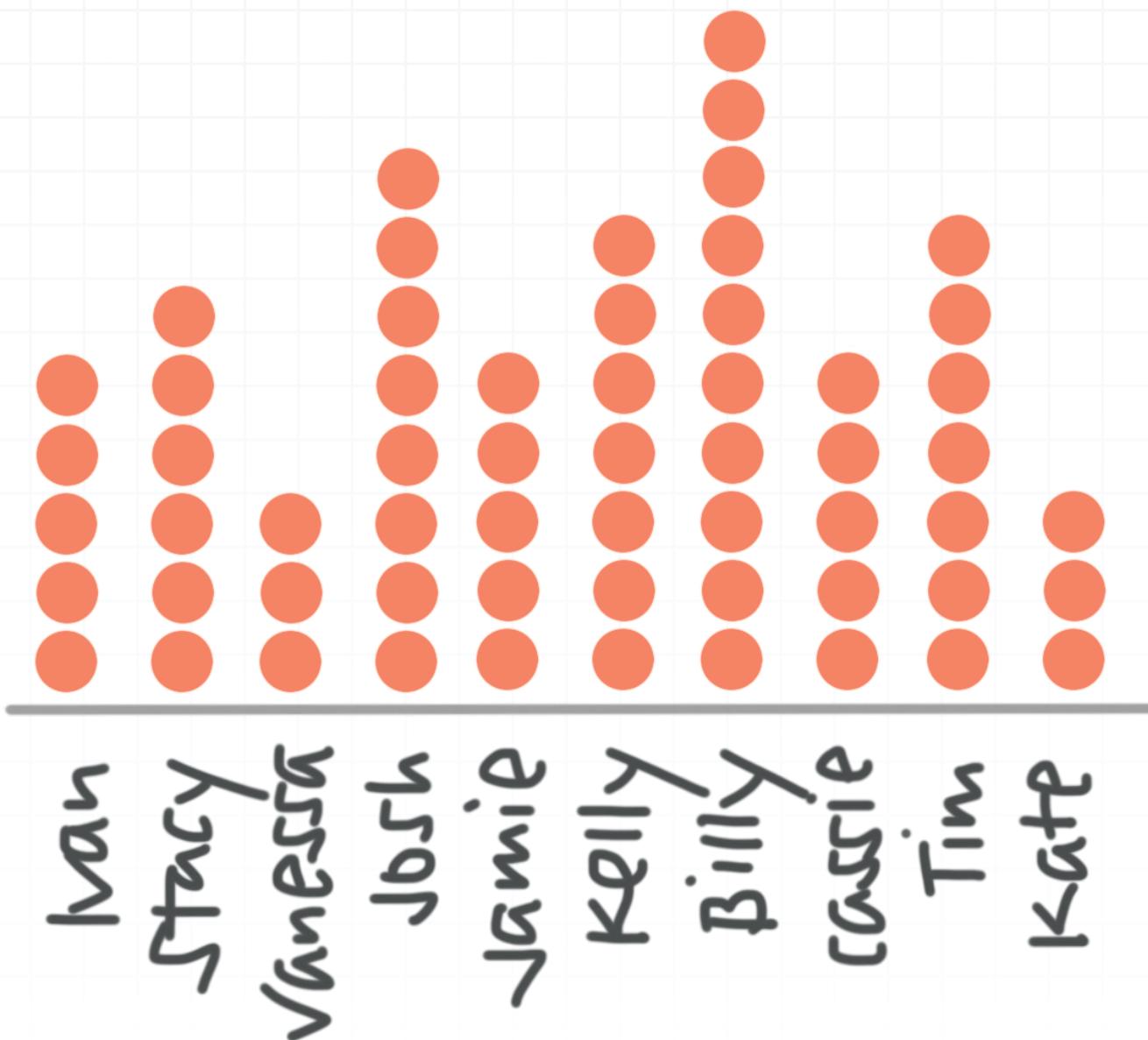
- 1. The frequency table shows the number of seed packets sold by each child during a pre-school fundraiser. Create a dot plot from the frequency table.

Name	packets sold
Ivan	5
Stacy	6
Vanessa	3
Josh	8
Jamie	5
Kelly	7
Billy	10
Cassie	5
Tim	7
Kate	3

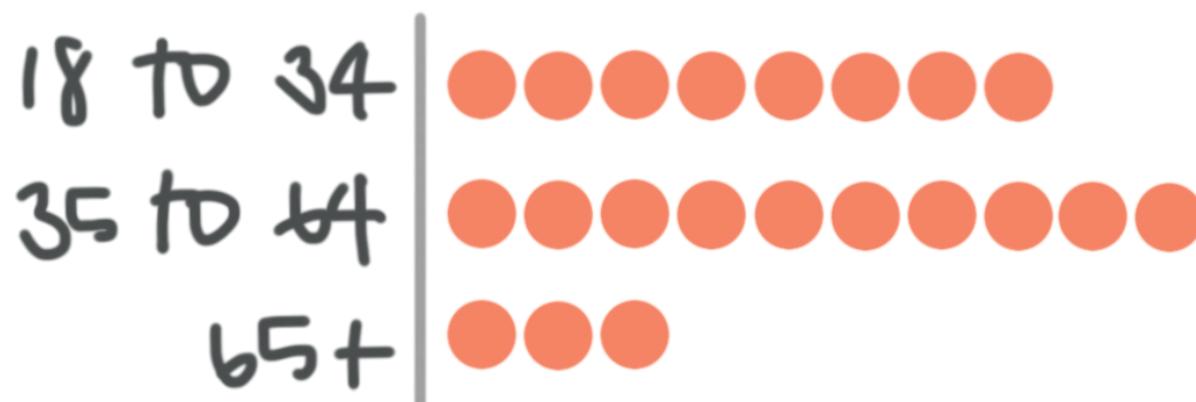
*Solution:*

Label the bottom of the dot plot with the names of the preschoolers. Put a dot to represent each packet sold. The dot plot would then look like this:





2. The dot plot shows the age of people who bought a bag of kale at a grocery store. Create a frequency table from the dot plot.



*Solution:*

Count the number of dots in each column. The number of dots are the frequency of each age group, so the frequency table would look like this:

Age of purchaser	Count
18 to 34	8
35 to 64	10
65+	3

- 3. The following data shows the number of red marbles drawn in a class lottery. Create a frequency table for the data.

0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 5, 5, 5, 7, 7

*Solution:*

Count the number of times the same amount of red marbles appeared and organize them into the table.



Red marbles	Frequency
0	3
1	5
2	5
5	3
7	2

4. The following data shows the favorite color of the students in Sebastian's kindergarten class. Create a frequency table for the data.

pink, pink, pink, pink, purple, purple, blue, blue, blue, blue, red, red, red, yellow, orange, orange, green, green, green, black

*Solution:*

To create the frequency table, we count how many of each color we have, then record the data in the table.



Color	Frequency
Pink	4
Purple	2
Blue	5
Red	3
Yellow	1
Orange	2
Green	3
Black	1

- 5. Kevin watches birds from his window and records what kind he sees. Create a dot plot from the data.

chickadee, redbird, redbird, redbird, chickadee, sparrow, sparrow, sparrow, sparrow, blue jay, crow, crow, redbird, chickadee, sparrow, sparrow, blue jay

*Solution:*

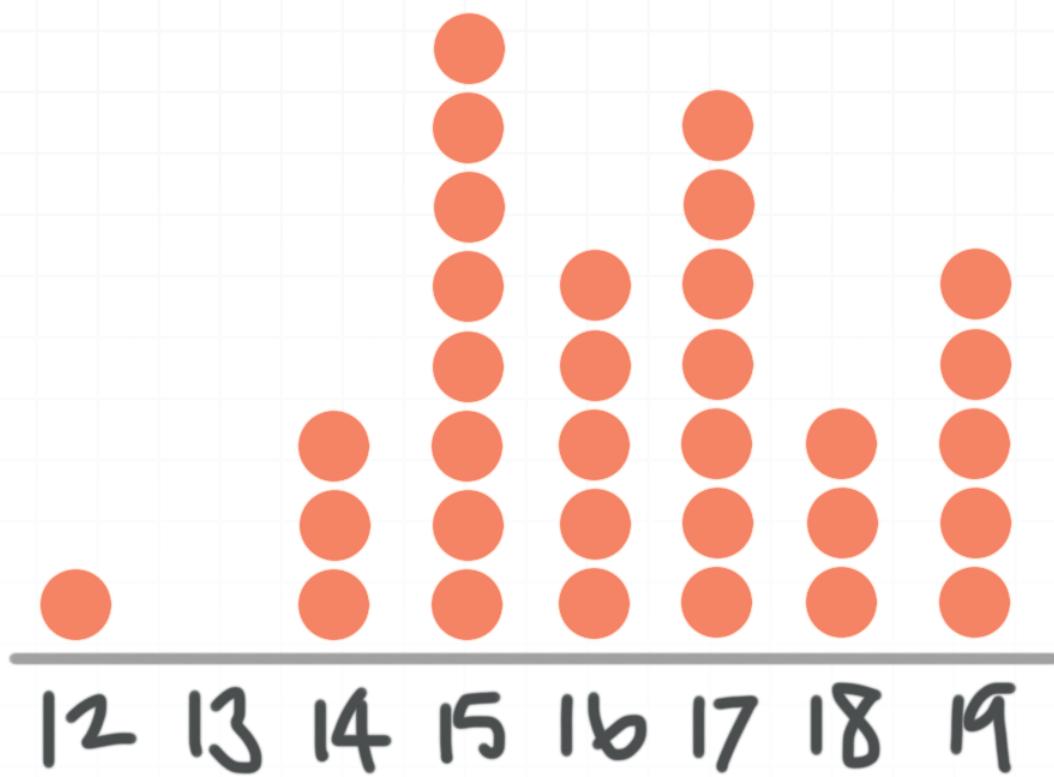
It can help to make a frequency table first. The birds are not grouped according to kind, so we'll make sure we count how many of each type we have.

Bird type	Frequency
Chickadee	3
Redbird	4
Sparrow	6
Blue jay	2
Crow	2

Now we create the dot plot with the same number of dots for each category as the table has frequencies.



6. The dot plot shows the ages of people in a lifeguard class at the local recreation center. How many people are enrolled in the class who are either 16, 17, or 18 years old?



*Solution:*

Use the dot plot to count how many lifeguards are 16, 17, and 18. There are 5 lifeguards who are 16, 7 lifeguards who are 17, and 3 lifeguards who are 18, which means there are 15 lifeguards in this age range.

$$5 + 7 + 3 = 15$$

## RELATIVE FREQUENCY TABLES

1. Blake is surveying students in his class (made up of juniors and seniors) about whether or not they play video games on a daily basis. What type of relative frequency table is shown? Finish filling in the table.

	Play at least one video game daily	Don't play any video games daily	Total
Junior	23%		75%
Senior		14%	
Total			100%

*Solution:*

This is a total-relative frequency table because there's a 100% in the grand total box. We can use the information in the table to fill out the rest of the information.

	Play at least one video game daily	Don't play any video games daily	Total
Junior	23%	52%	75%
Senior	11%	14%	25%
Total	34%	66%	100%

2. Create the row-relative frequency table for the frequency table below displaying 9th grade students who participate in an after school activity, and then answer the question: What percent of female 9th grade students do not participate in an after school activity?

	Participate	Don't participate
Male	62	40
Female	57	38

*Solution:*

40 % of female 9th grade students don't participate in an after school activity. To figure this out, we need to create a row-relative frequency. We'll start by finding row totals.

	Participate	Don't participate	Total
Male	62	40	$62+40=102$
Female	57	38	$57+38=95$

Now we can turn the table into a row-relative frequency table.

	Participate	Don't participate	Total
Male	$62/102=61\%$	$40/102=39\%$	$102/102=100\%$
Female	$57/95=60\%$	$38/95=40\%$	$95/95=100\%$

From the finalized table, we can see that 40 % of female 9th grade students don't participate in an after school activity.



	Participate	Don't participate	Total
Male	61%	39%	100%
Female	60%	40%	100%

- 3. Create the column-relative frequency table for this data table and then answer the question: What percentage of those who participate in an after school activity are male?

	Participate	Don't participate
Male	62	40
Female	57	38

*Solution:*

The first thing we need to do is to find the totals for each column.

	Participate	Don't participate
Male	62	40
Female	57	38
Total	$62+57=119$	$40+38=78$

Now use the column totals to calculate the column-relative frequencies for each column.

	Participate	Don't participate
Male	62/119=52%	40/78=51%
Female	57/119=48%	38/78=49%
Total	119/119=100%	78/78=100%

From the finalized table, we can see that 52% of 9th grade students who participate in an after school activity are male.

	Participate	Don't participate
Male	52%	51%
Female	48%	49%
Total	100%	100%

- 4. Create the total-relative frequency table for the data, and then answer this question: Carl is in charge of creating an activity for the students in his college dorm. If Carl wants the highest possible turnout, which activity should he choose? Why?

	Movie	Bowling	Pizza Party
Male	20	40	55
Female	35	50	62

*Solution:*

The largest percentage of students preferred a pizza party (45 %), so that is the event that Carl should choose. To figure this out, we need to create a total relative frequency table.

	Movie	Bowling	Pizza Party	Total
Male	20	40	55	$20+40+55=115$
Female	35	50	62	$35+50+62=147$
Total	$20+35=55$	$40+50=90$	$55+62=117$	$115+147=262$

So here is the finished frequency table.

	Movie	Bowling	Pizza Party	Total
Male	20	40	55	115
Female	35	50	62	147
Total	55	90	117	262

Now the total relative frequency table is:

	Movie	Bowling	Pizza Party	Total
Male	$20/262=8\%$	$40/262=15\%$	$55/262=21\%$	$115/262=44\%$
Female	$35/262=13\%$	$50/262=19\%$	$62/262=24\%$	$147/262=56\%$
Total	$55/262=21\%$	$90/262=34\%$	$117/262=45\%$	$262/262=100\%$

	Movie	Bowling	Pizza Party	Total
Male	8%	15%	21%	44%
Female	13%	19%	24%	56%
Total	21%	34%	45%	100%



It looks like the largest percentage of students preferred a pizza party (45 %), so that's the event that Carl should choose.

- 5. A city hall is looking into a dangerous intersection that has caused many bicycle accidents over the past month, due to rerouted traffic. They have counted the number of bicycle accidents and put them into a frequency table like the one below. Create the relative frequency table for the data and answer the following question: What day had the highest percentage of bicycle accidents?

Day of the week	Number of crashes
Sunday	13
Monday	10
Tuesday	8
Wednesday	6
Thursday	2
Friday	11
Saturday	14

*Solution:*

The highest percentage of bicycle accidents (22 %) happened on Saturday. We know this is true simply because the largest number of bicycle accidents happened on Saturday, but we could also create a relative frequency table, first by finding a total.



Day of the week	Number of crashes
Sunday	13
Monday	10
Tuesday	8
Wednesday	6
Thursday	2
Friday	11
Saturday	14
<b>Total</b>	<b>64</b>

Now we'll find the crashes on each day as a percentage of the total, and we can see that the highest percentage of accidents is still occurring on Saturday.

Day of the week	Number of crashes
Sunday	$13/64=20\%$
Monday	$10/64=16\%$
Tuesday	$8/64=13\%$
Wednesday	$6/64=9\%$
Thursday	$2/64=3\%$
Friday	$11/64=17\%$
Saturday	$14/64=22\%$
<b>Total</b>	<b><math>64/64=100\%</math></b>

6. Addie took a poll of the children in her neighborhood. She found that 15 of them watch 2 hours or more of cartoons per day. Out of the 15 that watch 2 hours or more, 10 watched the cartoons on a device other than the television. There were also 12 children who watched less than 2 hours of cartoons per day. For those 12 children, 2 of them watched cartoons on a device other than a television. Construct a two-way table to summarize the data and then construct a total-relative frequency table for the data.

*Solution:*

Given what we know, we can fill in the table with this information:

	< 2 hours	> 2 hours	Total
Watched on T.V.			
Watched on a different device	2	10	
Total	12	15	

And then we can fill in the rest of the table:

	< 2 hours	> 2 hours	Total
Watched on T.V.	10	5	15
Watched on a different device	2	10	12
Total	12	15	27

And then we can convert this to a total-relative frequency table by dividing by the total in the lower right.

	< 2 hours	> 2 hours	Total
Watched on T.V.	10/27=37%	5/27=19%	15/27=56%
Watched on a different device	2/27=7%	10/27=37%	12/27=44%
Total	12/27=44%	15/27=56%	27/27=100%

So the total-relative frequency table is:

	< 2 hours	> 2 hours	Total
Watched on T.V.	37%	19%	56%
Watched on a different device	7%	37%	44%
Total	44%	56%	100%

## JOINT DISTRIBUTIONS

- 1. To study the relationship between votes for a new park and people who have children, a community group surveyed voters. What percentage of those surveyed had children? Is this part of the joint, conditional, or marginal distribution?

	For	Against	No opinion
Children	125	50	30
No children	40	150	60

*Solution:*

About 45 % of those surveyed had children. This is part of the marginal distribution.

	For	Against	No opinion	Total
Children	125	50	30	205
No children	40	150	60	250
Total	165	200	90	455

Now we can calculate the percentage of the voters who had children:

$$\frac{205}{455} \approx 45\%$$

Since this calculation was done from the total column it is part of the marginal distribution.

- 2. To study the relationship between votes for a new park and people who have children, a community group surveyed voters. What percentage of those surveyed were for the park and had children? Is this part of the joint, conditional, or marginal distribution?

	For	Against	No opinion
Children	125	50	30
No children	40	150	60

*Solution:*

About 27% of those surveyed voted for the park and had children. This is part of the joint distribution.

	For	Against	No opinion	Total
Children	125	50	30	205
No children	40	150	60	250
Total	165	200	90	455

Now we can calculate the percentage of the voters who had children and voted for the park:

$$\frac{125}{455} \approx 27\%$$

Since this is dependent on the grand total, this is part of the joint distribution.

3. To study the relationship between votes for a new park and people who have children, a community group surveyed voters. What percentage of those with no children had no opinion? Is this part of the joint, conditional, or marginal distribution?

	For	Against	No opinion
Children	125	50	30
No children	40	150	60

*Solution:*

About 24% of those with no children had no opinion. This is part of a conditional distribution. Here we need to only look at the 250 people surveyed who had no children. Out of those we want to know who had “no opinion” on the park.

	For	Against	No opinion	Total
Children	125	50	30	205
No children	40	150	60	250
Total	165	200	90	455



Since we're interested in a subset of those surveyed, this is a conditional distribution. The percentage of those with no children who had no opinion is:

$$\frac{60}{250} \approx 24\%$$

- 4. Carl is in charge of creating an activity for the students in his college dorm, and he records their preferences by activity and gender. What percentage of the female students prefer pizza? To answer the question, should we use a marginal, joint, or conditional distribution?

	Movie	Bowling	Pizza Party
Male	20	40	55
Female	35	50	62

*Solution:*

To answer the question, we need to use a conditional distribution. We're interested in the percentage of female students who prefer pizza. So, we're interested only in the conditional distribution for the row of female students.

	Movie	Bowling	Pizza Party	Total
Female	35	50	62	147



The percentage of female students who preferred a pizza party was

$$\frac{62}{147} \approx 42\%$$

- 5. A pharmaceutical company is testing heart burn as a side effect of its new pain reliever. What conclusions can we draw from the marginal distributions of the study?

	Pain reliever	Placebo	Total
Minor heartburn	4	171	175
Major heartburn	102	25	127
No heartburn	10,568	10,478	21,046
Total	10,674	10,674	21,348

*Solution:*

Remember that for the marginal distributions, we're just looking at the total column and total row. 10,674 of the 21,348 people in the study took the pain reliever, and 10,674 took the placebo. This means

$$\frac{10,674}{21,348} \approx 50\%$$

took the pain reliever and about

$$\frac{10,674}{21,348} \approx 50\%$$



took the placebo. We also know 175 of those in the study experienced minor heartburn, or

$$\frac{175}{21,348} \approx 0.8\%$$

127 experienced major heartburn,

$$\frac{127}{21,348} \approx 0.6\%$$

and 21,046 or

$$\frac{21,046}{21,348} \approx 98.6\%$$

experienced no heartburn. Without calculating the conditional probabilities, we can't say much more about heartburn as a side effect of the pain reliever.

- 6. Consider the same data as the previous question. What do the conditional distributions (given the participant experienced minor heartburn, major heartburn, or no heartburn) tell us about the study?

	Pain reliever	Placebo	Total
Minor heartburn	4	171	175
Major heartburn	102	25	127
No heartburn	10,568	10,478	21,046
Total	10,674	10,674	21,348



*Solution:*

The conditional distributions described here are the row-relative frequencies. Of the 175 people in the study who had minor heartburn, 4 took the pain reliever and 171 took the placebo.

	Pain reliever	Placebo	Total
Minor heartburn	4/175=2.3%	171/175=97.7%	100%
	Pain reliever	Placebo	Total

More people who took the placebo suffered from minor heartburn than those that took the pain reliever. It's probably safe to say that taking the pain reliever doesn't cause minor heartburn.

Of the 127 people in the study who had major heartburn, 102 took the pain reliever and 25 took the placebo.

	Pain reliever	Placebo	Total
Major heartburn	102/127=80.3%	25/127=19.7%	100%
	Pain reliever	Placebo	Total

More people who took the pain reliever suffered from major heartburn than those that took the placebo.

	Pain reliever	Placebo	Total
No heartburn	10,568/21,046=50.2%	10,478/21,046=49.8%	100%
	Pain reliever	Placebo	Total

Those who took the pain reliever and placebo were symptom free at roughly the same rate. Due to the discrepancy between major and minor heartburn symptoms, it could be worthwhile to look at the study again to



see if there was a problem with the placebo used in the minor heartburn part of the study.



## HISTOGRAMS AND STEM-AND-LEAF PLOTS

- 1. A doctor recorded the weight of all the babies that visited her clinic last week. How many babies weighed no more than 24 pounds?

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

$$1 | 5 = 15$$

*Solution:*

“No more than” means we include all of the babies that weigh 24 pounds or less in our count. That means 6 babies weighed no more than 24 pounds.

- 2. The stem plot shows the number of clothing pieces on each rack at a clothing store. Create a histogram from the stem plot, and use buckets of size 10.



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

$$1 | 0 = 10$$

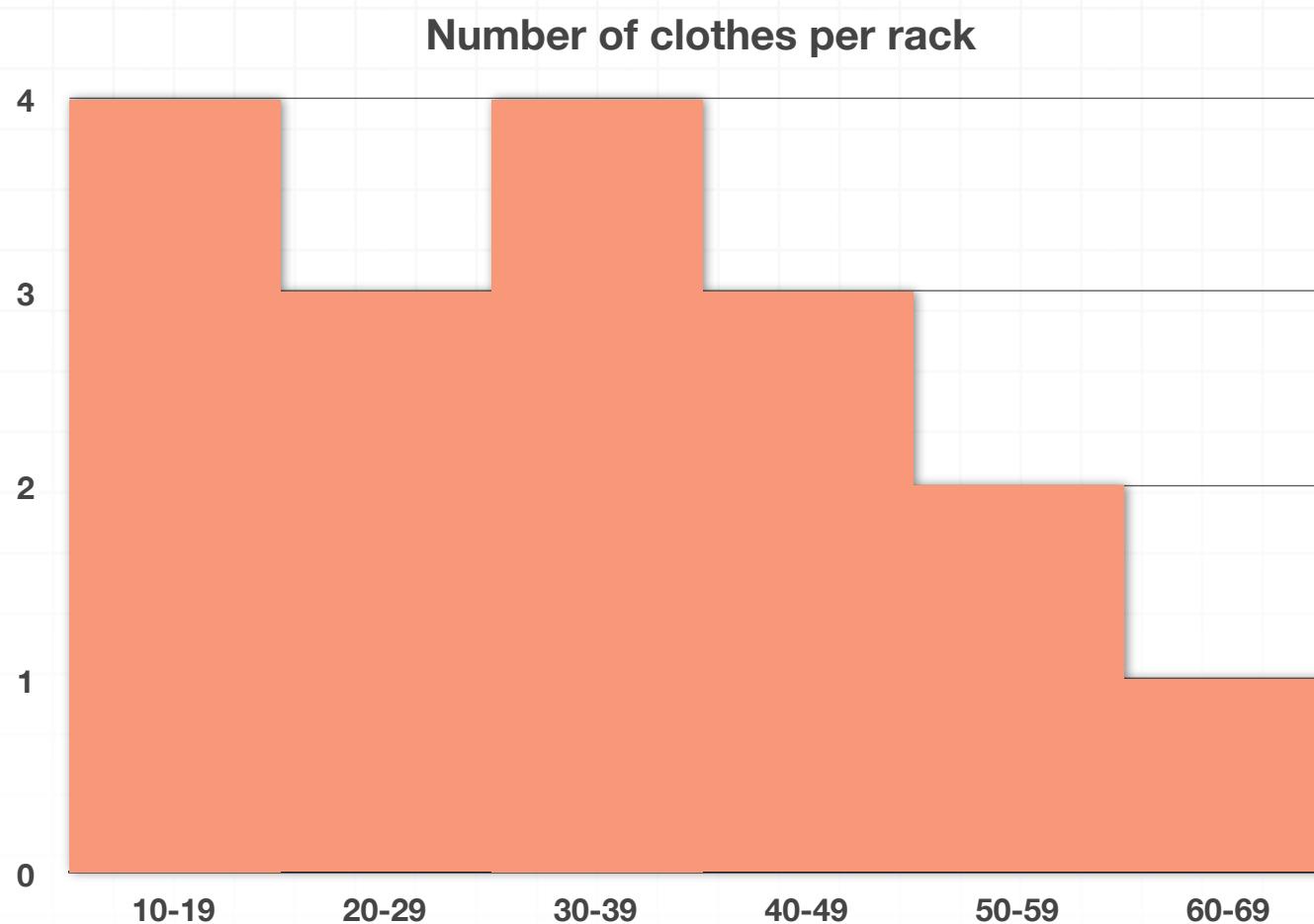
*Solution:*

The stem-and-leaf plot counts by 10s along the left-hand side. Then we can see how many data points fall into each bucket, because we're using buckets of size 10.

In other words, the stems become the buckets and the number of leaves become the frequencies graphed in the histogram.

Buckets	Number of leaves
10 - 19	0 1 2 8
20 - 29	8 8 8
30 - 39	2 6 8 9
40 - 49	4 4 4
50 - 59	2 6
60 - 69	0

Now we can turn this frequency table into a histogram.

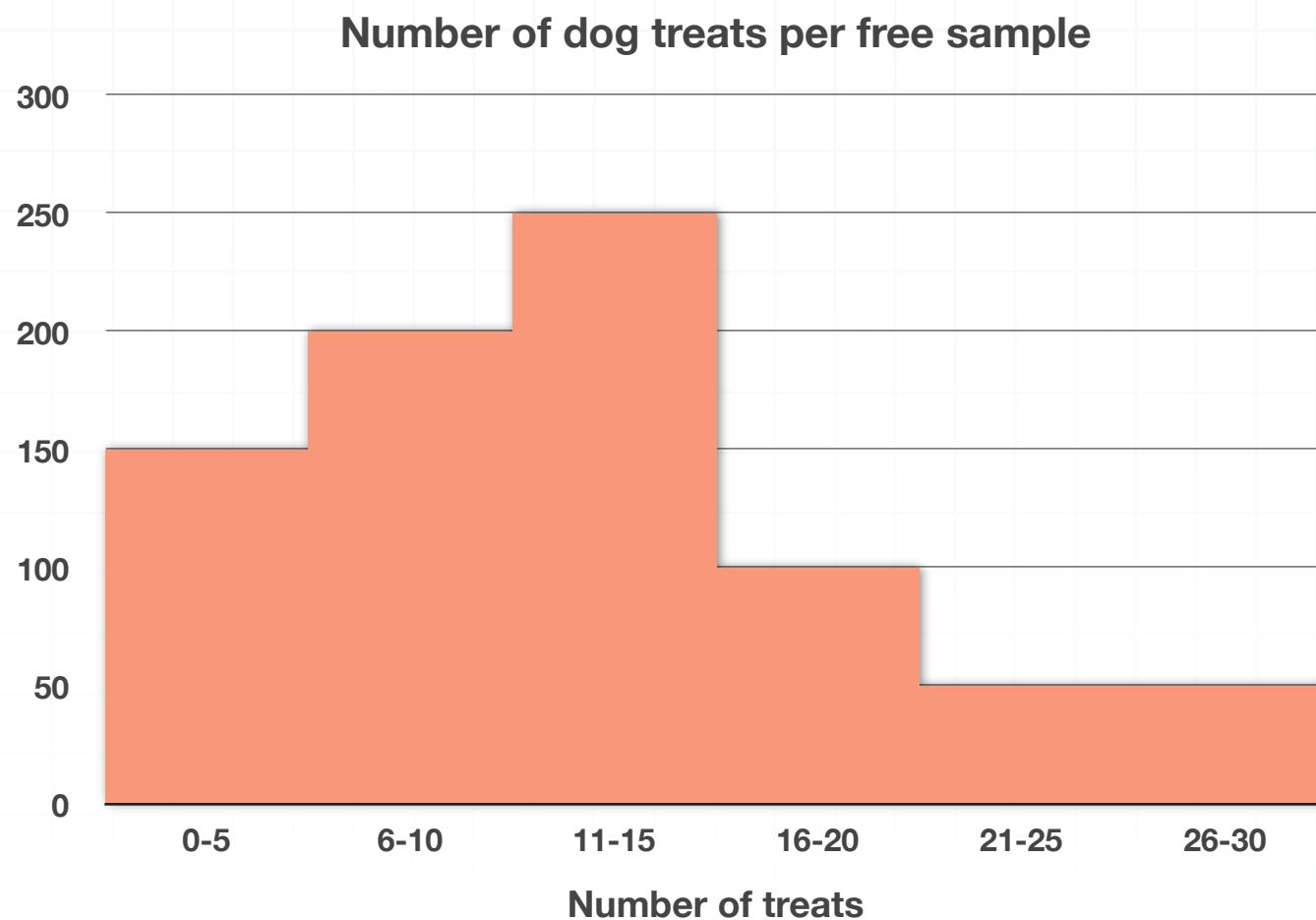


- 3. Is it possible to create a stem-and-leaf plot from a histogram? Why or why not?

*Solution:*

We can't make a stem-and-leaf plot from a histogram, because a stem-and-leaf records each data point, while a histogram records how many data points occur in a certain range. This means that a histogram doesn't contain specific enough information to create a stem-and-leaf plot.

4. A company mails out packets of dog treat samples based on a consumer's previous dog food purchases. How many times did the company mail a packet of 11 – 15 treats?

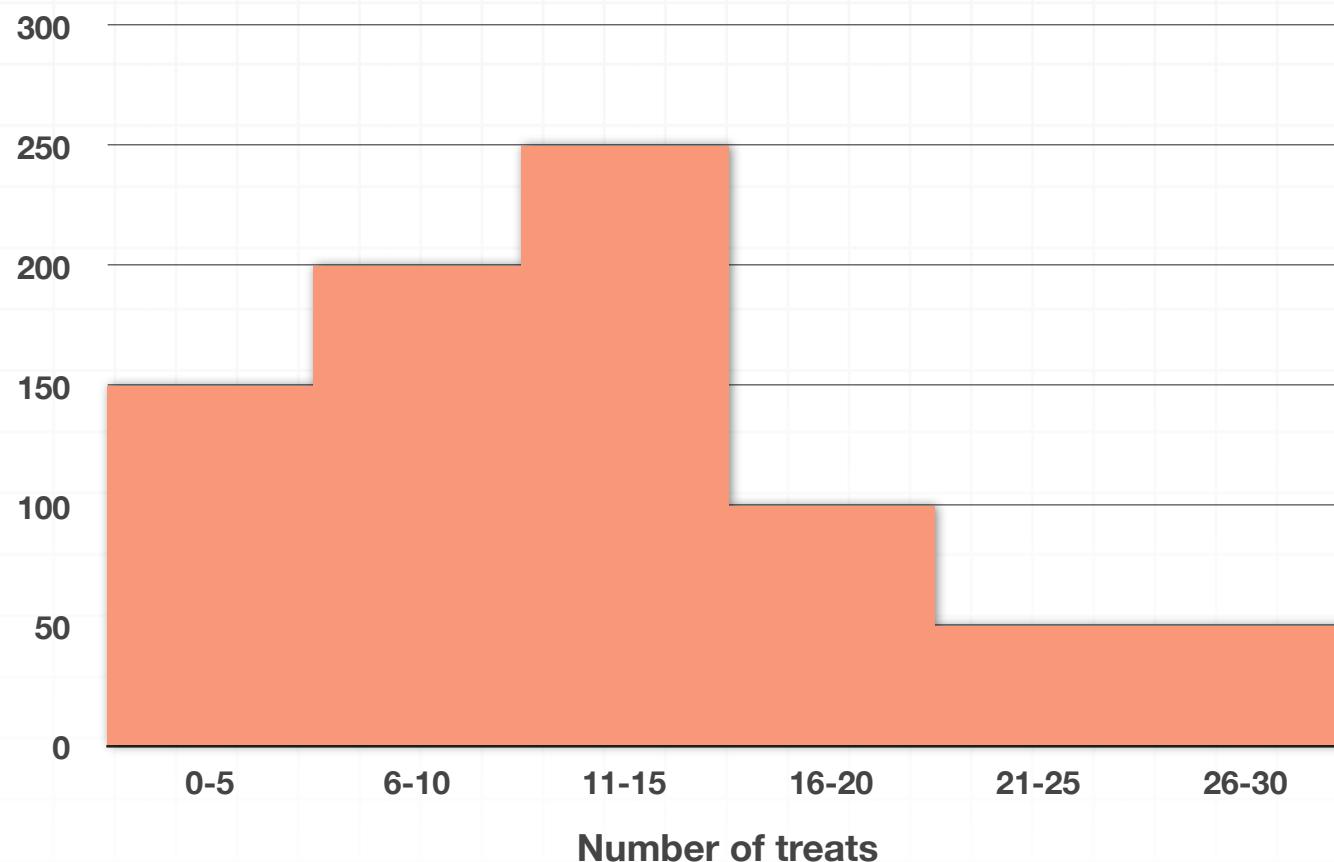


*Solution:*

The bar for the 11 – 15 interval has a frequency of 250. This means the company mailed out bags with 11 – 15 treats 250 times. So 250 samples contained 11 – 15 treats.

5. A company mails out packets of dog treat samples based on a consumer's previous dog food purchases. How many packets of dog treat samples did the company give out?

### Number of dog treats per free sample



*Solution:*

To find the total number of treats the company mailed out, add up the frequencies in the histogram.

$$150 + 200 + 250 + 100 + 50 + 50 = 800 \text{ samples}$$

The company mailed out 800 packets of treat samples.

■ 6. Create a stem-and-leaf chart from the list of student test scores.

60, 65, 80, 80, 81, 82, 88, 89, 90, 97, 98, 100, 100



*Solution:*

To make a stem-and-leaf plot, it's helpful to make sure all of our data is in order from smallest to largest. In this case, it makes sense to choose a stem of tens and leaves of ones.

60, 65, 80, 80, 81, 82, 88, 89, 90, 97, 98, 100, 100

Notice the data does not have any numbers in the 70s, so that stem is left blank. When we get to 100, the leaf needs to be 10 to represent ten 10s, or 100.

The stem-and-leaf plot for student test scores is

6	0 5
7	
8	0 0 1 2 8 9
9	0 7 8
10	0 0

$$6|0 = 60$$



## BUILDING HISTOGRAMS FROM DATA SETS

- 1. If the range of the data set is 36 and we want to divide it into 5 class intervals, which of the following would be the most appropriate class width?

*Solution:*

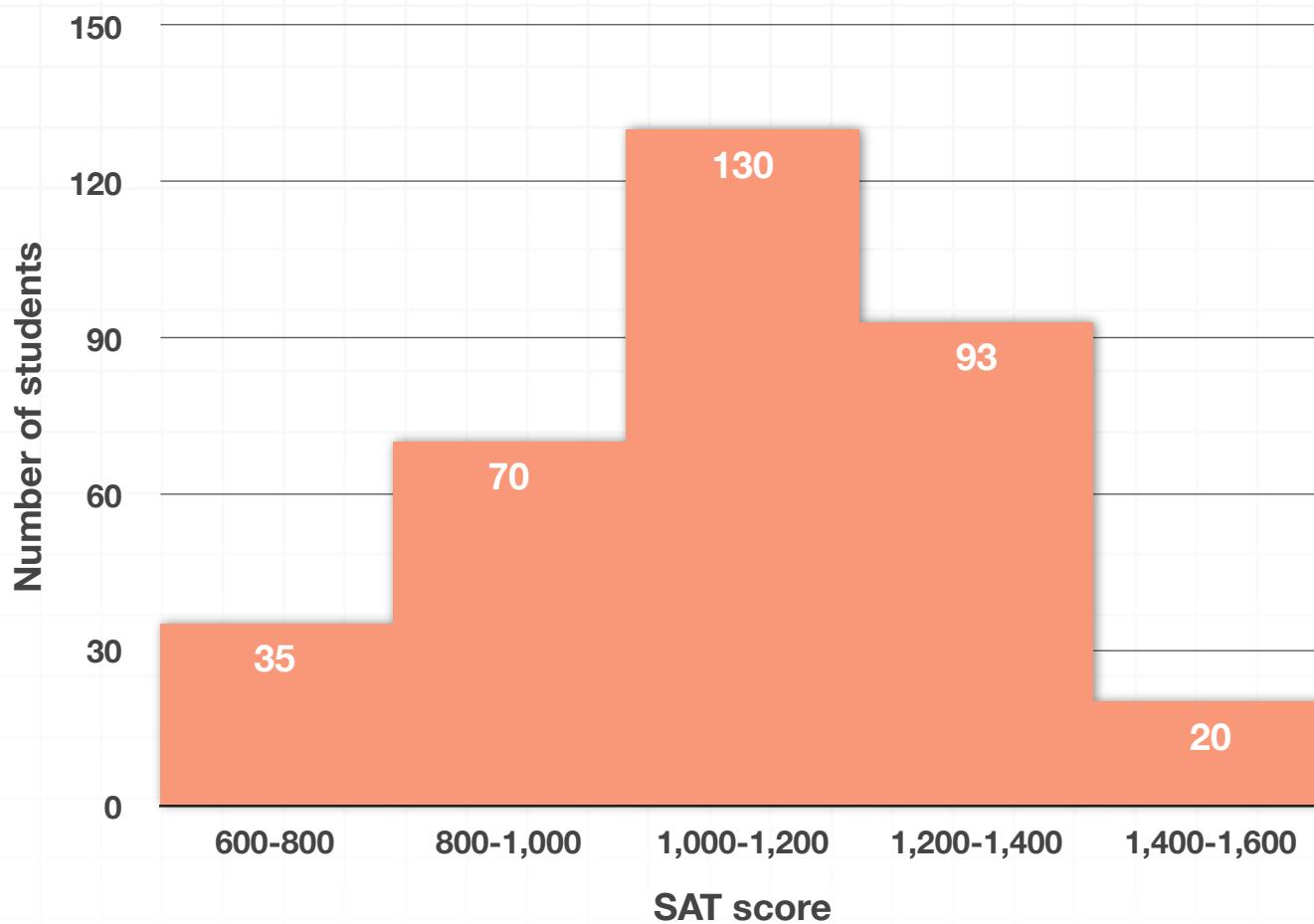
We can find the class width by dividing the range of the data set by the number of classes, or class intervals.

$$\frac{36}{5} = 7.2$$

As a rule of thumb, it's always better to round the decimal up to make sure we include the minimal and maximal values, so we get 8 as a class width.

- 2. Based on the histogram showing the distribution of the SAT scores for students at a local high school, what number of students scored between 1,000 and 1,400?





*Solution:*

The question asks us to find the number of students who scored between 1,000 and 1,400, which means we need to count the number of students in two class intervals, the 1,000 – 1,200 interval and the 1,200 – 1,400 interval.

$$130 + 93 = 223$$

- 3. If we set the first two classes for the data set below as 0 – 7 and 7 – 14, how many data points will fall into the first interval?

7, 3, 4, 12, 23, 34, 2, 13, 21, 8, 7

*Solution:*

The intervals  $0 - 7$  and  $7 - 14$  overlap. When we use overlapping classes, any data point that falls on that overlapping boundary gets included into the second (upper) class.

In other words, the lower class contains the values that are greater than or equal to the lower class limit and less than the upper class limit, and the value of the upper class limit, which is also the lower class limit of the next interval, will fall into the next class.

This means that the data points from 0 to 6 are included into the first class, and 7 will fall into the second class. Therefore, the first class  $0 - 7$  contains three data points, which are 2, 3, and 4. The second class  $7 - 14$  contains five data points, which are 7, 7, 8, 12, and 13.

- 4. Considering the table below, what is the midpoint of the class that includes the smallest number of the students?

Exam score	Number of students
90 - 100	25
80 - 90	54
70 - 80	64
60 - 70	8

*Solution:*

The class interval with the smallest number of students is 60 – 70. Then the class midpoint is given by

$$\frac{60 + 70}{2} = 65$$

- 5. A literature teacher asked 50 of his students how many hours they spent reading last week, then recorded his results in a table. What class width did he use?

Hours spent reading	Number of students
0 - 4	25
4 - 8	54
8 - 11	64

*Solution:*

The class width is given by the difference between either the lower or upper limits of consecutive classes. The lower limits of the first two classes are 0 and 4, so the class width is  $4 - 0 = 4$ .

- 6. A math teacher asks 25 of her students how many hours they spent on math homework last week. Given the responses below, build a histogram with 6 bins that displays the data.



4, 3, 5, 1, 0, 12, 11, 6, 4, 2, 13, 3, 7, 12, 9, 8, 10, 22, 13, 4, 5, 20, 1, 0, 7

*Solution:*

Put the data points in ascending order.

0, 0, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 7, 7, 8, 9, 10, 11, 12, 12, 13, 13, 20, 22

Then the range is

$$\text{Range} = 22 - 0 = 22$$

Divide the range by the number of classes to find the class width.

$$\frac{22}{6} \approx 3.67$$

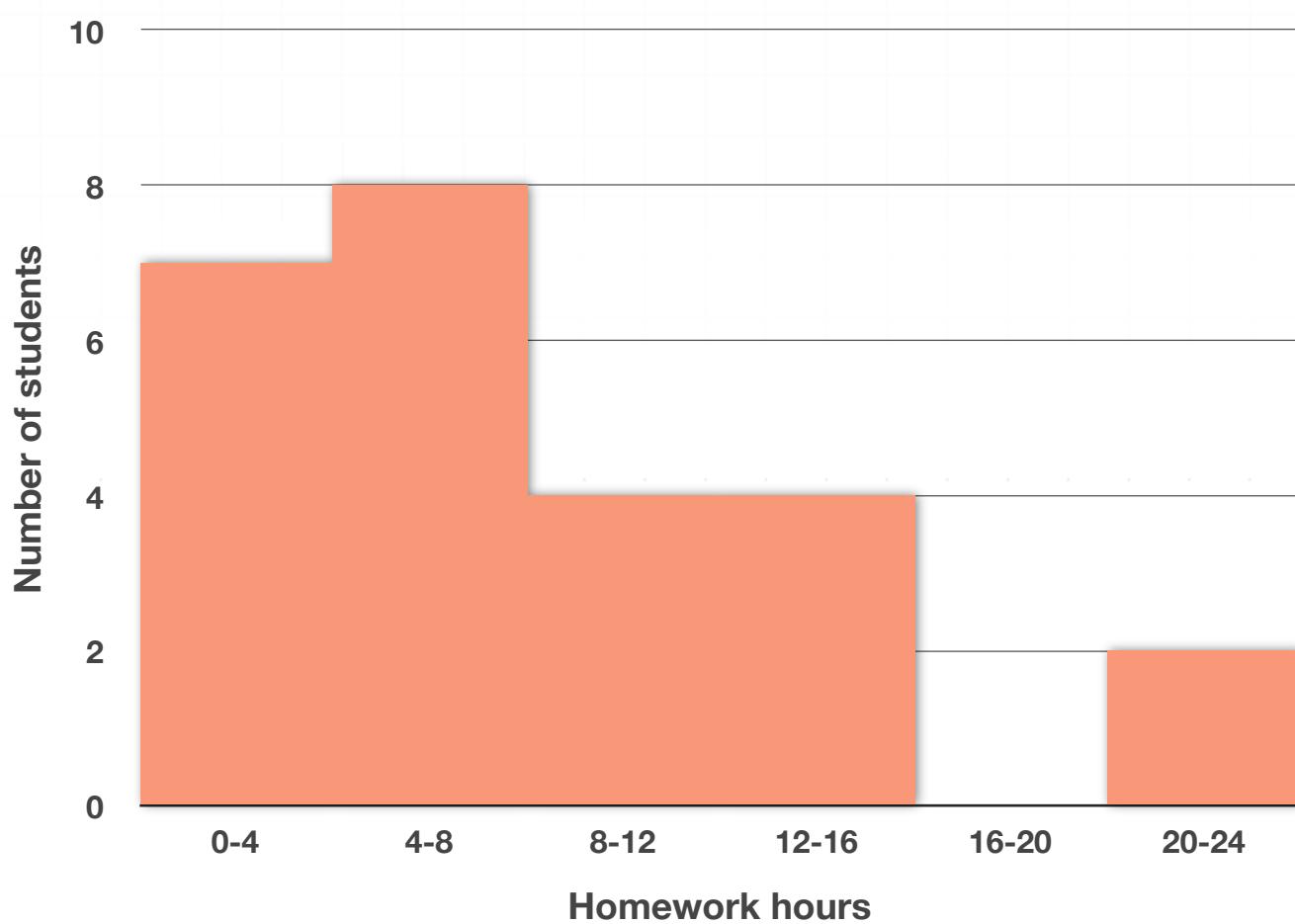
To make sure we include the smallest and the largest values in the data set, we'll round the result up and use a class width of 4.

Now we can set up the classes and count the number of data points that fall into each class.



Homework hours	Number of students
0 - 4	7
4 - 8	8
8 - 12	4
12 - 16	4
16 - 20	0
20 - 24	2

We have an empty interval, but we still include it in the histogram, so the result is



## MEASURES OF CENTRAL TENDENCY

### ■ 1. What is the mean of the data set?

105, 250, 358, 422

*Solution:*

To find the mean, add all the numbers in the data set, and then divide by the number of data points. This data set has 4 numbers so we get:

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

$$\mu = \frac{105 + 250 + 358 + 422}{4}$$

$$\mu = 283.75$$

The mean is 283.75.

### ■ 2. What is the median of the data set?

62, 64, 69, 70, 70, 71, 73, 74, 75, 77

*Solution:*



This data set has 10 values, which means to find the median we need to find the two middle numbers and take their mean.

~~62, 64, 69, 70, 70, 71, 73, 74, 75, 77~~

Now we need to find the mean of 70 and 71.

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

$$\mu = \frac{70 + 71}{2}$$

$$\mu = 70.5$$

The median of the data set is 70.5.

### ■ 3. What is the mode of the data set?

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

$$1 | 3 = 13$$

*Solution:*

The mode of a data set is the number that repeats the most often. In the stem plot, the mode is 35.

- 4. What number could we add to the data set that would give us a median of 15?

1, 2, 8, 13, 20, 30, 31

*Solution:*

The median of a data set is the middle number. In this data set, without changing anything, 13 is the middle number, so it's the median.

~~1, 2, 8, 13, 20, 30, 31~~

If we were to add one more number to the data set, then to find the median we would take the average of the two middle numbers. We want the median to be 15, which means we'll need to insert a number larger than 13.

~~1, 2, 8, 13, \_\_, 20, 30, 31~~

Let's call the missing number  $m$ . Then we can set up this equation:

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



$$15 = \frac{13 + m}{2}$$

$$15(2) = 13 + m$$

$$30 = 13 + m$$

$$17 = m$$

This means 17 is the number we can add to the data set to force the median to be 15.

- 5. A teacher lost Samantha's test after it was graded, but she knows the statistics for the rest of the class.

Class mean (including Samantha's test):  $\mu = 85$

Total number of students who took the test: 18

Class test scores for everyone but Samantha were:

75, 75, 75, 80, 80, 80, 80, 80, 82, 82, 82, 82, 95, 95, 95, 95, 98

What did Samantha score on her test?

*Solution:*

To find the mean, add the test scores, then divide by the number of test scores. We know the mean is 85, and that there were 18 students who took



the test. Let's call Samantha's missing test score  $s$ . Then we can set up this equation:

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

$$85 = \frac{s + 3(75) + 5(80) + 4(82) + 4(95) + 1(98)}{18}$$

When we solve for  $s$ , we get

$$85(18) = s + 3(75) + 5(80) + 4(82) + 4(95) + 1(98)$$

$$1,530 = s + 3(75) + 5(80) + 4(82) + 4(95) + 1(98)$$

$$1,530 = s + 225 + 400 + 328 + 380 + 98$$

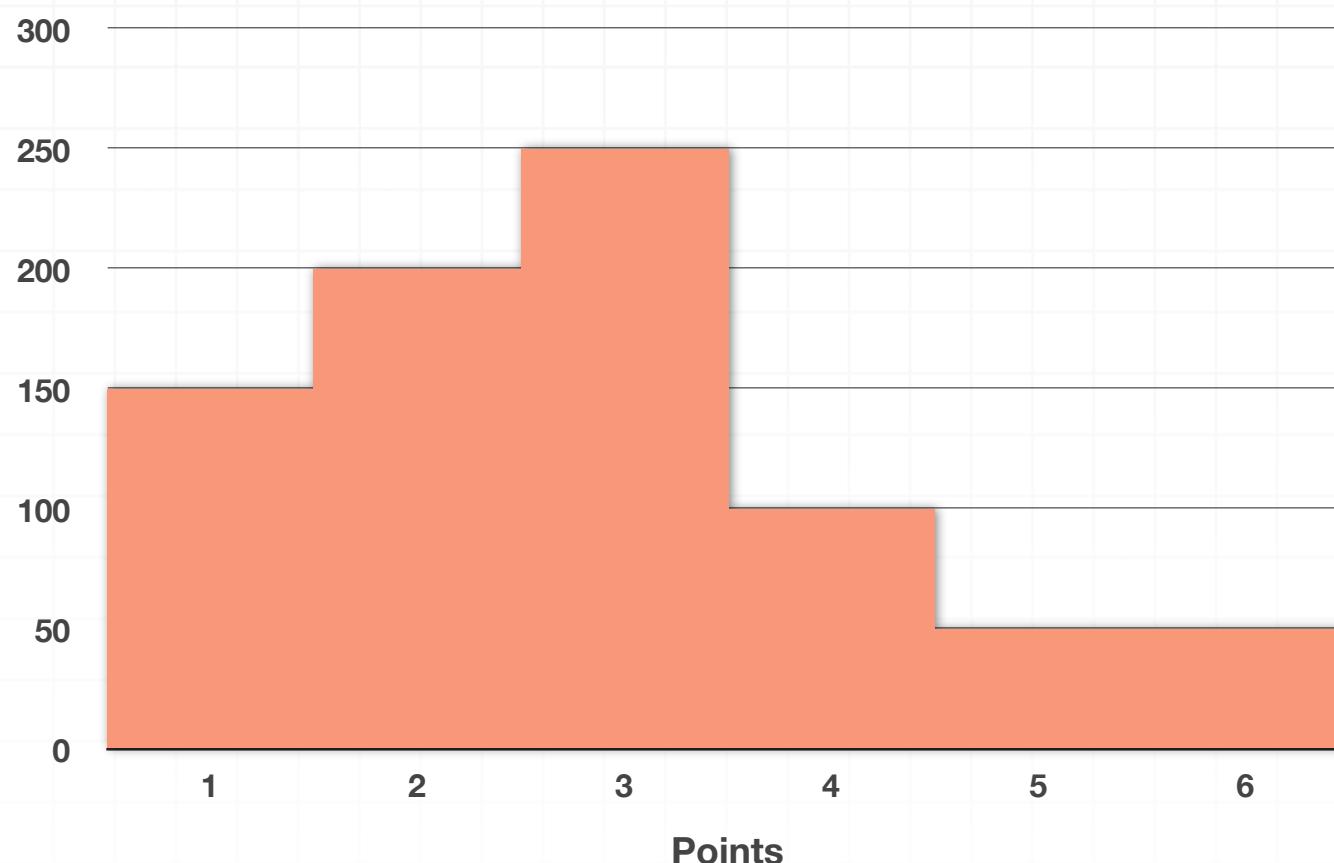
$$s = 1,530 - 225 - 400 - 328 - 380 - 98$$

$$s = 99$$

Samantha scored a 99 on her test.

## ■ 6. What is the mode of the data set?



**Points scored in a word game**

*Solution:*

This is a frequency histogram. More often than any other value, 3 points were scored in the word game. Therefore, the mode is 3.

## MEASURES OF SPREAD

- 1. Sarah is visiting dairy farms as part of a research project and counting the number of red cows at each farm she visits. Here is her data:

0, 1, 1, 1, 2, 5, 5, 7, 7, 18, 24, 24

Calculate the IQR and range of the data set.

*Solution:*

The range of the data is the largest number minus the smallest number.

The smallest number in the data set is 0, and the largest number in the data set is 24, so the range is  $24 - 0 = 24$ .

To find the IQR, we need to find the upper median (called the upper quartile) and the lower median (called the lower quartile). To do this, we divide the data into four equal parts.

This data set has 12 data items, so we can find the median by crossing out the first five numbers and the last five numbers, and then take the average of the middle two numbers.

~~0, 1, 1, 1, 2, 5, 5, 7, 7, 18, 24, 24~~

The median is

$$\frac{5 + 5}{2} = \frac{10}{2} = 5$$



The lower half of the data set is 0, 1, 1, 1, 2, 5, and its median is

$$\frac{1+1}{2} = \frac{2}{2} = 1$$

The upper half of the data set is 5, 7, 7, 18, 24, 24, and its median is

$$\frac{7+18}{2} = \frac{25}{2} = 12.5$$

Therefore, the IQR is  $12.5 - 1 = 11.5$ .

- 2. A dog boarding company kept track of the number of dogs staying overnight and the frequency. What is the range of the data?

Number of dogs	Frequency
20	2
25	3
32	1
38	1
39	2
40	3
43	2

*Solution:*

The largest number in the data set is 43, and the smallest number is 20, so the range is  $43 - 20 = 23$ .

3. Catherine counted the number of lizards she saw in her garden each week and recorded the data in a table. What is the interquartile range of the data?

Number of lizards	Frequency
2	5
5	2
8	1
12	2
13	2
15	3
21	1

*Solution:*

Let's create a list from the table. The data set is

2, 2, 2, 2, 2, 5, 5, 8, 12, 12, 13, 13, 15, 15, 15, 21

There are 16 items in the data set, so we can cross off the first seven and last seven, and then find the average of the middle two numbers to get the median.

~~2, 2, 2, 2, 2, 5, 5, 8, 12, 12, 13, 13, 15, 15, 15, 21~~

The median is

$$\frac{8 + 12}{2} = \frac{20}{2} = 10$$

The lower half of the data is 2, 2, 2, 2, 2, 5, 5, 8, so the median of the lower half is

$$\frac{2 + 2}{2} = \frac{4}{2} = 2$$

The upper half of the data is 12, 12, 13, 13, 15, 15, 15, 21, so the median of the upper half is

$$\frac{13 + 15}{2} = \frac{28}{2} = 14$$

Therefore, the interquartile range is  $14 - 2 = 12$ .

- 4. The median of the lower-half of a data set is 98. The interquartile range is 2. If the data set has 9 numbers, what can we say about the median of the entire data set?

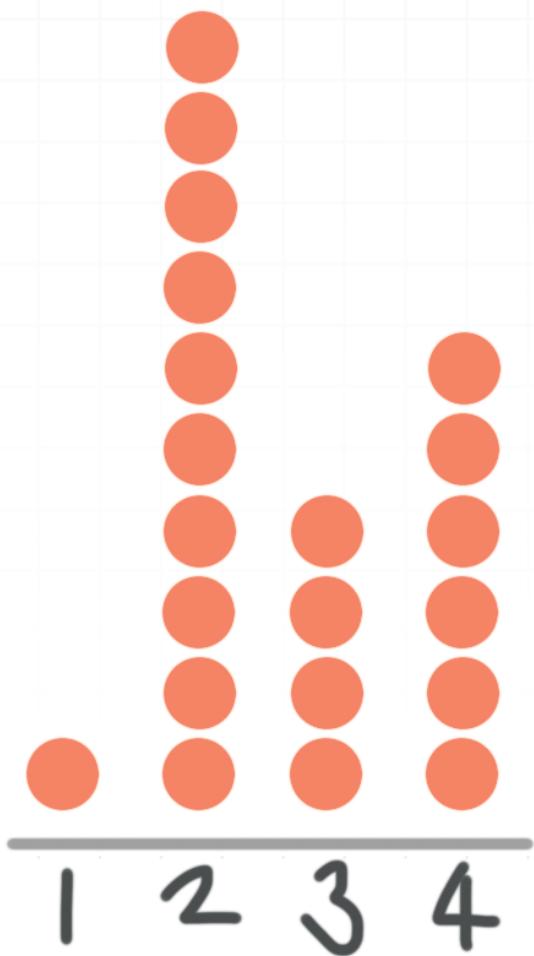
*Solution:*

Since the median of the lower half of the data is 98 and the interquartile range is 2, we can find the median of the upper half of the data as



$98 + 2 = 100$ . This means the median of the data set is any number between or including 98 and 100.

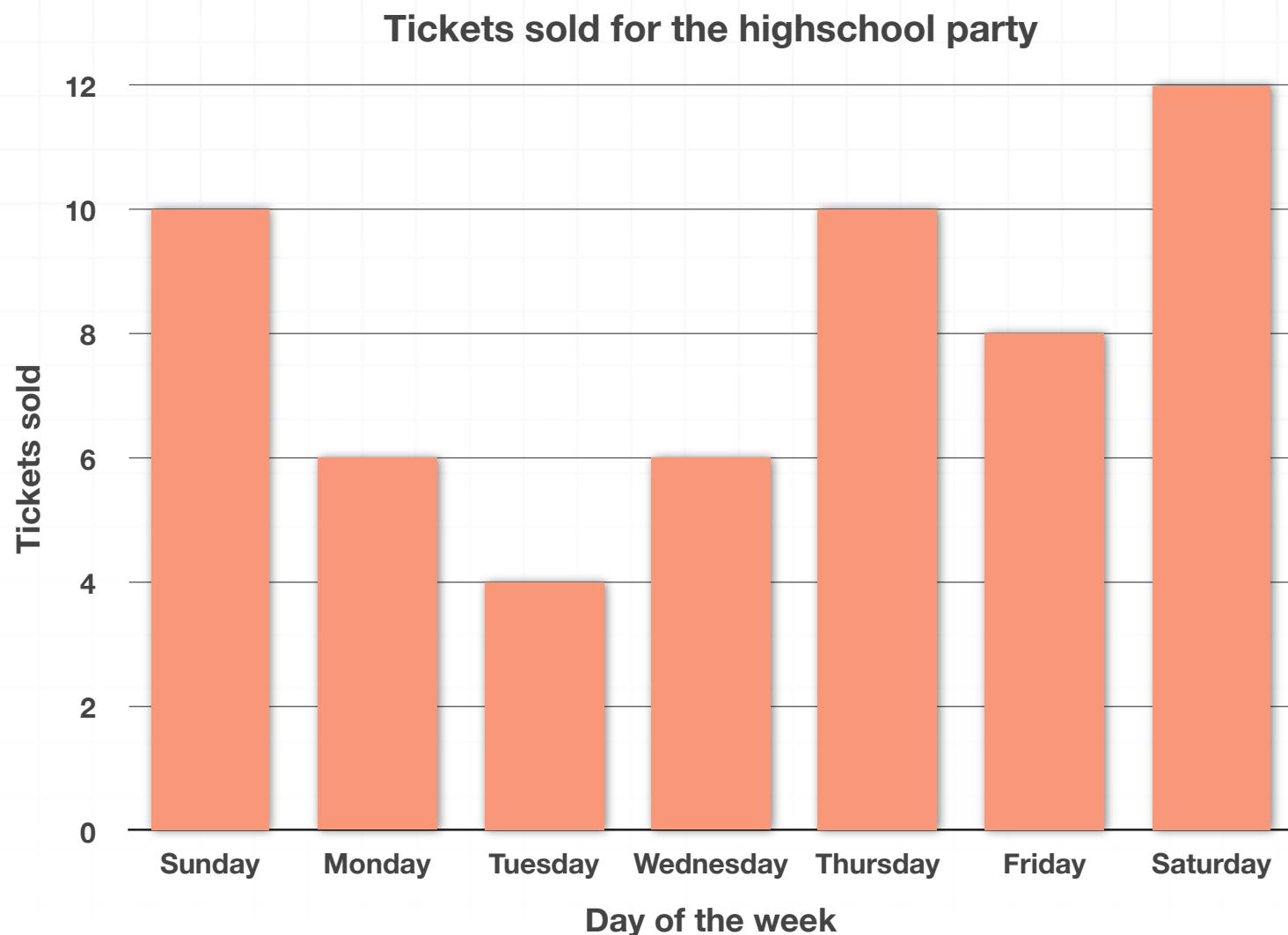
- 5. The dot plot shows the number of trips to the science museum for a class of 4th graders. What is the range of the data set?



*Solution:*

The range is the largest number in the data set minus the smallest number. The largest number is 4, and the smallest number is 1, so the range is  $4 - 1 = 3$ . The range is 3.

6. The bar graph shows the number of tickets sold for the high school party each day. What is the interquartile range of the data set?



*Solution:*

We could list the data from the bar graph as

10, 6, 4, 6, 10, 8, 12

Put the data in order so we can find the median.

4, 6, 6, 8, 10, 10, 12

The median is 8. The lower half of the data is 4, 6, 6, so the median of the lower half is 6. The upper half of the data is 10, 10, 12, so the median of the upper half is 10.

Therefore, the IQR of the data set is  $10 - 6 = 4$ .



## CHANGING THE DATA AND OUTLIERS

- 1. The students in an English class ended up with a mean score on their recent exam of 65 points. The range of exam scores was 25 points. If each score is increased by 10%, what are the new mean and range?

*Solution:*

Increasing the scores by 10% is the same as multiplying the data set by 1.10. This multiplication both increases the scores and spreads out the data. This means that both the mean and the range will be multiplied by 1.10.

The original mean is 65 and the new mean is  $65(1.10) = 71.5$ . The original range is 25, and the new range is  $25(1.10) = 27.5$ .

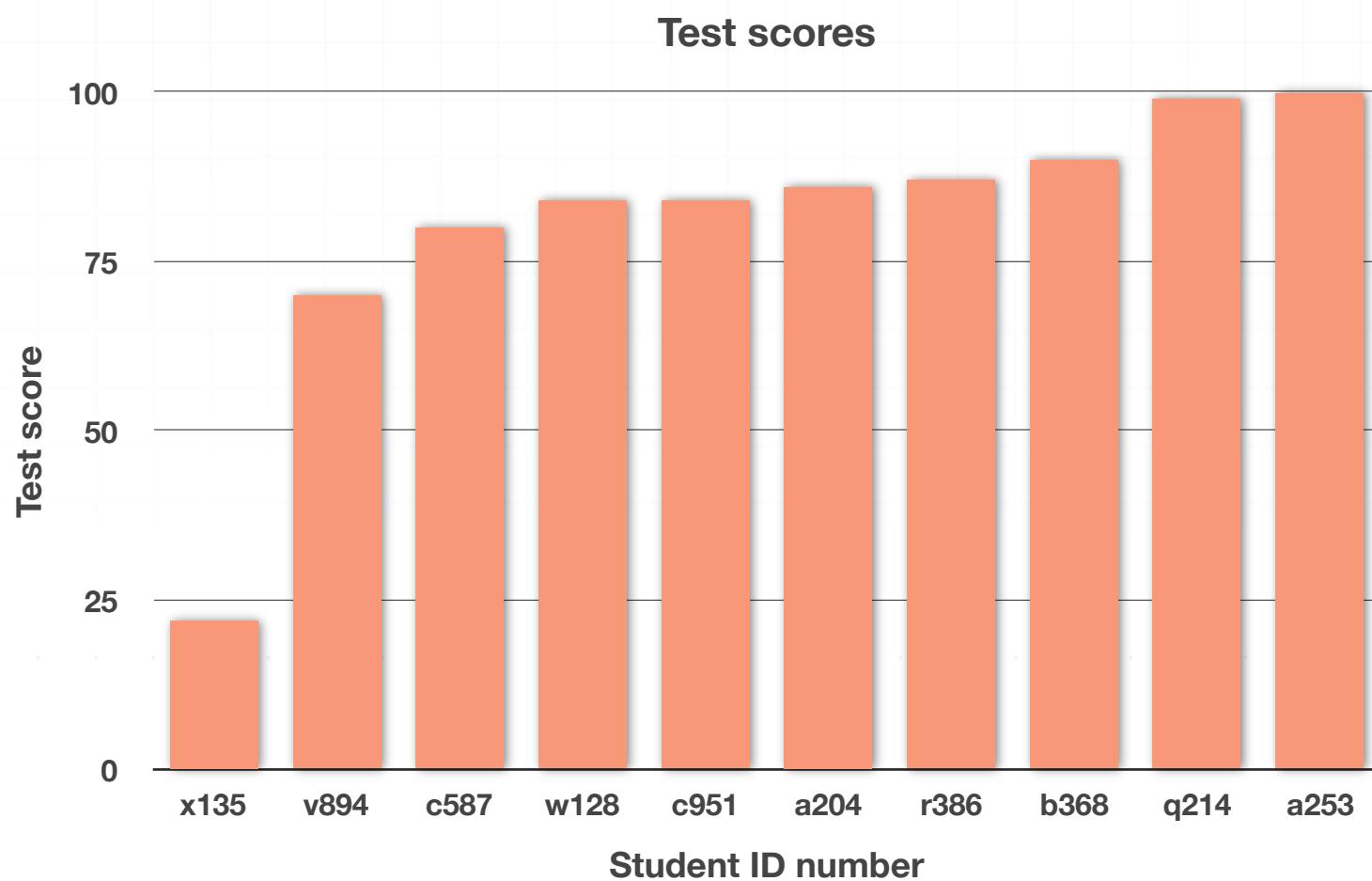
- 2. Spencer asked students at his high school what percentage of the school budget they thought was spent on extracurricular activities. The mean response was 8% and the median response was 5%. There was one outlier in the responses. What do the mean and median tell us about the outlier?

*Solution:*



The outlier was greater than the rest of the data because the mean is greater than the median. In other words, the outlier is pulling the mean toward the larger value. The median is more resistant to outliers, which is why it's much lower.

- 3. How does the mean compare to the median in the data from the bar graph?

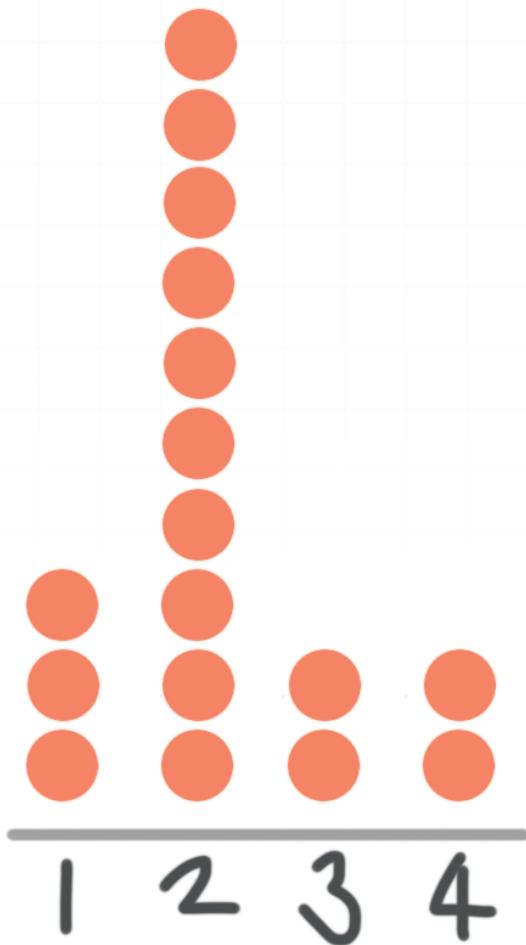


*Solution:*

In this bar graph, there's one score from student  $x135$  that's significantly lower than the rest. This means that the score is likely an outlier. This will

make the mean smaller than the median because it'll pull the mean score down.

- 4. The dot plot shows the number of trips to the science museum for a class of 4th graders. How does the mean compare to the median in the data set below, and what does it tell us about the potential outliers in the data set?



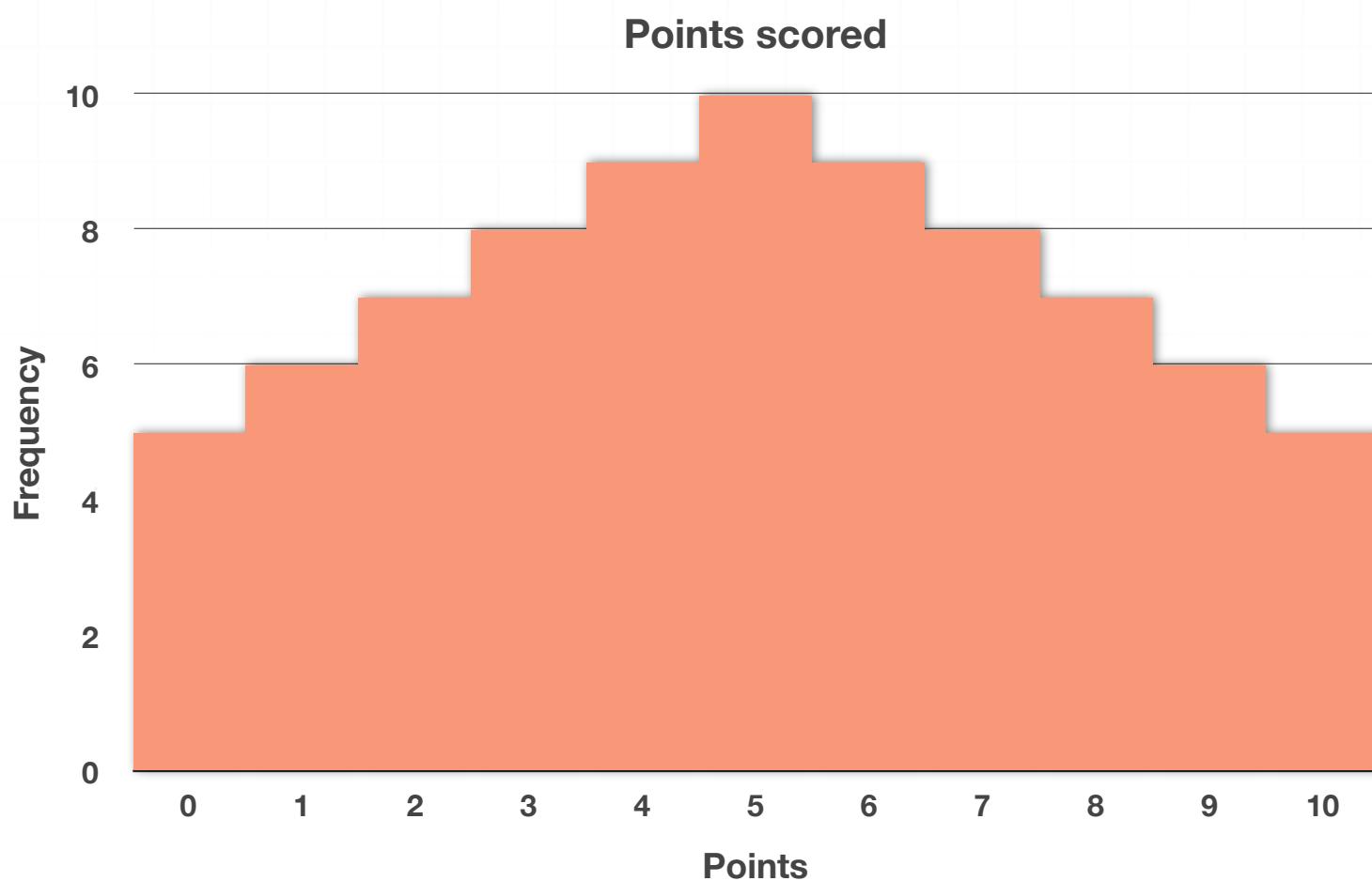
*Solution:*

We can calculate the mean and median from the dot plot. The median is 2 and the mean is

$$\mu = \frac{3(1) + 10(2) + 2(3) + 2(4)}{3 + 10 + 2 + 2} \approx 2.18$$

The mean and median are close together, so this data set doesn't have outliers. The mean is being pulled upward from the median a bit because there are more students who visited the museum more than 2 times than there are students who visited the museum less than 2 times.

- 5. What does the shape of this histogram tell us about the mean and median of the data?



*Solution:*

This data set is symmetric. The mean and median are equal to one another, and there are no outliers in the data.

- 6. An experiment is done in degrees Celsius. The original data had the following:

Mean:  $102^\circ$  Celsius

Median:  $101^\circ$  Celsius

Mode:  $99^\circ$  Celsius

Range:  $7^\circ$  Celsius

IQR:  $4^\circ$  Celsius

The formula to convert to degrees Fahrenheit is  $F = (9/5)C + 32$ . After the conversion to Fahrenheit, what are the new reported measures of the data set?

*Solution:*

Multiplying the data set by a constant value of  $9/5$  will multiply all of these measures of center and spread as well.

Mean:  $102^\circ(9/5) = 183.6^\circ$

Median:  $101^\circ(9/5) = 181.8^\circ$



Mode:  $99^{\circ}(9/5) = 178.2^{\circ}$

Range:  $7^{\circ}(9/5) = 12.6^{\circ}$

IQR:  $4^{\circ}(9/5) = 7.2^{\circ}$

Shifting the data set by adding 32, will add 32 to the new mean, median and mode. The range and IQR will stay the same.

Mean:  $183.6^{\circ} + 32^{\circ} = 215.6^{\circ}$  Fahrenheit

Median:  $181.8^{\circ} + 32^{\circ} = 213.8^{\circ}$  Fahrenheit

Mode:  $178.2^{\circ} + 32^{\circ} = 210.2^{\circ}$  Fahrenheit

Range:  $12.6^{\circ}$  Fahrenheit

IQR:  $7.2^{\circ}$  Fahrenheit



## BOX-AND-WHISKER PLOTS

- 1. What is the range and interquartile range of the data set?

Median: 617,594

Minimum: 216,290

Maximum: 845,300

First quartile: 324,528

Third quartile: 790,390

*Solution:*

The range is

$$845,300 - 216,290 = 629,010$$

The interquartile range is

$$790,390 - 324,528 = 465,862$$

- 2. These are average lifespans in years of various mammals:

35, 10, 40, 40, 20, 10, 15, 14, 18, 35

Find the five-number summary for the data.



*Solution:*

The five-number summary is the list of the minimum, first quartile, median, third quartile and maximum values. We need to divide the data set into four parts to find the five-number summary, which we can do by arranging the numbers from least to greatest.

10, 10, 14, 15, 18, 20, 35, 35, 40, 40

Now we can see that the minimum is 10, that the maximum is 40, and that the median is

$$\frac{18 + 20}{2} = 19$$

The lower half of the data set is 10, 10, 14, 15, 18, so the median of the lower half is 14. The upper half of the data set is 20, 35, 35, 40, 40, so the median of the upper half is 35. So we can summarize the five-number summary as

Min	Q1	Median	Q3	Max
10	14	19	35	40

■ 3. Create a box plot based on the following information about a data set.

Mode: 300

Minimum: 100

First Quartile: 300

Median: 2,000

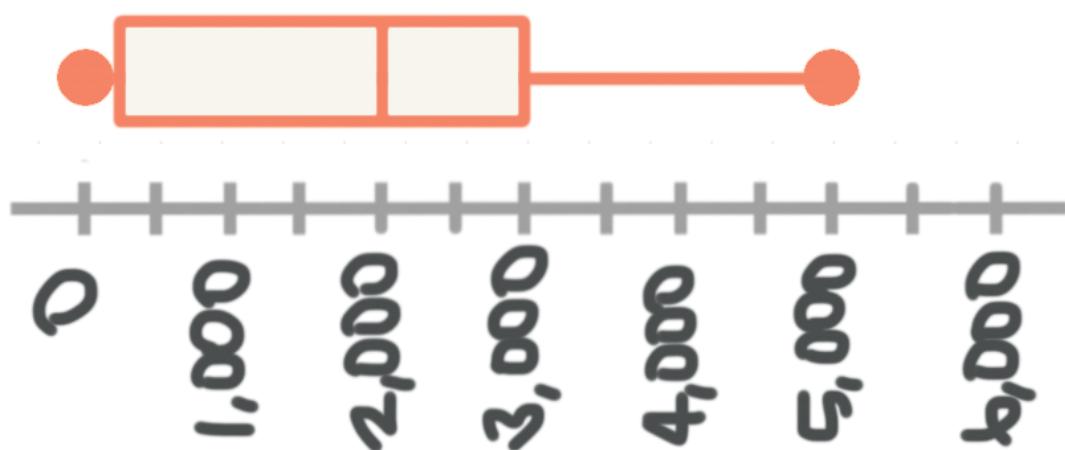
Mean: 1,887.5

Third Quartile: 3,050

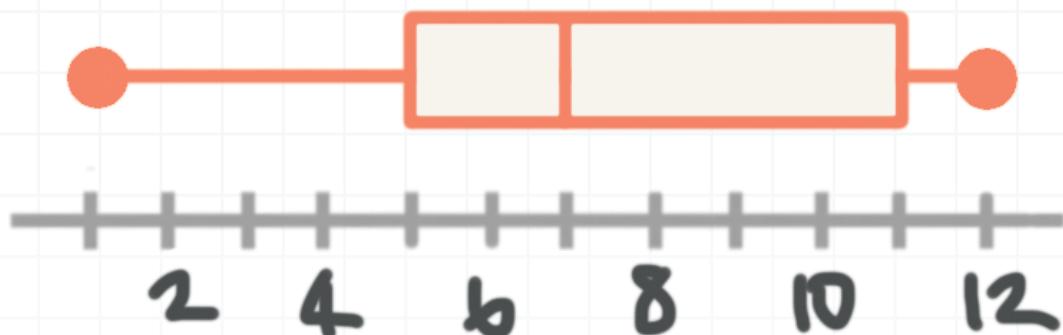
Maximum: 4,800

*Solution:*

To create a box plot, we just need to use the five-number summary. The five-number summary is the list of the minimum, first quartile, median, third quartile, and maximum values. If we take those values from the question, then we can create the box plot.



- 4. How does the amount of data between 1 and 5 compare to the amount of data between 11 and 12?

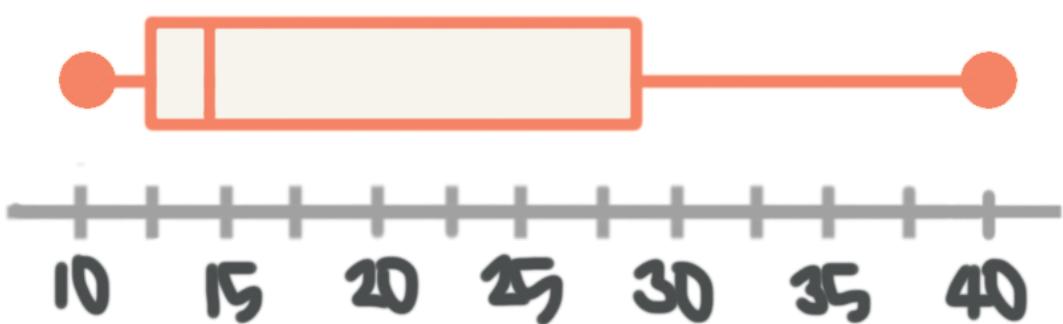


*Solution:*

The whisker from 1 to 5 is longer because the values in that quarter of the data set are more spread out than the values in the quarter of the data set between 11 and 12.

For example, if the data set had 24 values, 6 of the values would fall between 1 and 5, and 6 of the values would fall between 11 and 12.

■ 5. In which quarter of the data is the number 23 located?



*Solution:*

The number 23 lies between the median and the right edge of the box, which is the third quarter of the data set.

6. Create the box-and-whisker plot for the book ratings given in the stem and leaf plot.

Stem	Leaf
1	3 7 8
2	1 4 6
3	5 5
4	
5	2 6

Key: 1 | 3 = 13

*Solution:*

To create the box-and-whisker plot, we first need to create the five-number summary. The five-number summary is the list of the minimum, first quartile, median, third quartile, and maximum values. We need to divide the data set into four parts to find the five-number summary, so let's start by writing out the numbers in the data set.

13, 17, 18, 21, 24, 26, 35, 35, 52, 56

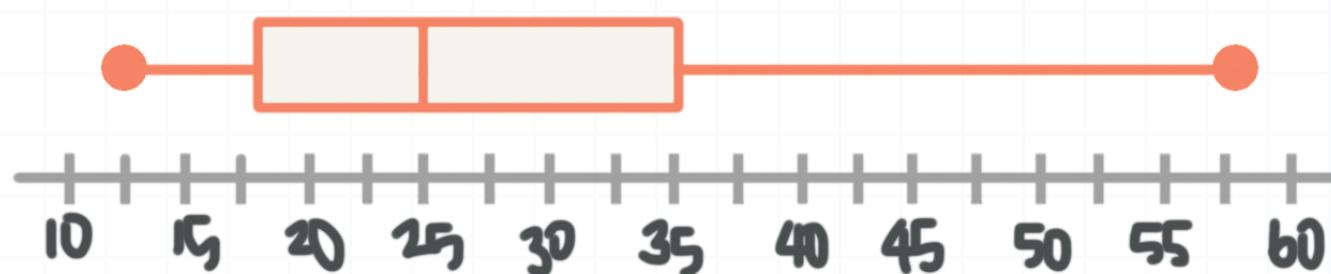
Now we can see that the minimum of the data set is 13, and the maximum of the data set is 56. The median is

$$\frac{24 + 26}{2} = 25$$



The lower half of the data is 13, 17, 18, 21, 24, so the median of the lower half is 18. The upper half of the data is 26, 35, 35, 52, 56, so the median of the upper half is 35.

Now that we have the five-number summary given by the minimum, first quartile, median, third quartile, and maximum, we can sketch the box plot.



## MEAN, VARIANCE, AND STANDARD DEVIATION

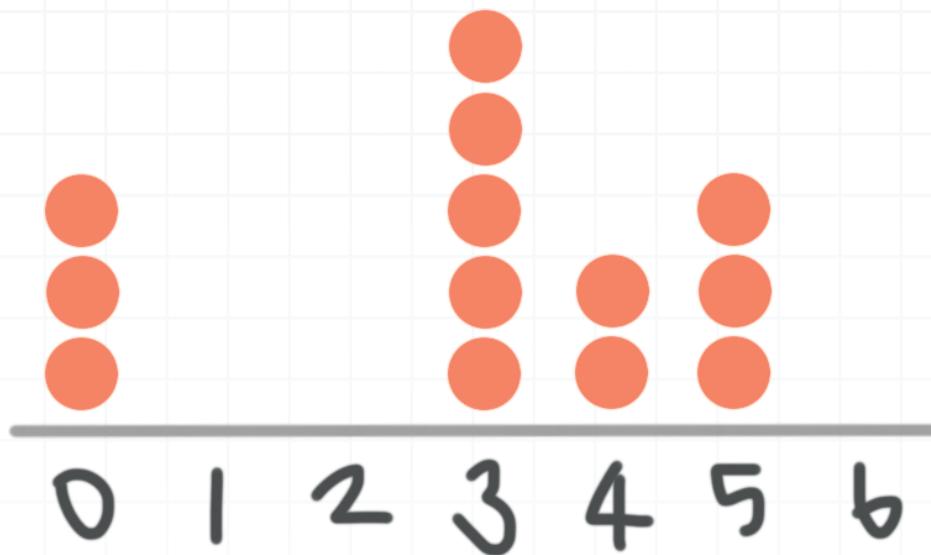
- 1. Mrs. Bayer's students take a test on Friday. She grades their tests over the weekend and notes that the average test score is 68 points with a population standard deviation of 5 points. She decided to add 10 points to all of the tests. What are the new mean and population standard deviation?

*Solution:*

The population standard deviation will remain the same, because adding the 10 points won't change the spread of the data. The population standard deviation of the old and new data will both be 5. Adding 10 points to all of the tests will increase the mean by 10 points. The old mean is 68 points, so the new mean is 78 points.

- 2. What is the sample variance of the data set, rounded to the nearest hundredth?





*Solution:*

The formula for the sample variance includes the sample mean, so we'll need to find that first. There are  $n = 13$  data points in the dot plot, so the sample mean is

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

$$\bar{x} = \frac{3(0) + 5(3) + 2(4) + 3(5)}{13}$$

$$\bar{x} = \frac{38}{13}$$

$$\bar{x} \approx 2.92$$

The sample variance is

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$$S^2 = \frac{3(0 - 2.92)^2 + 5(3 - 2.92)^2 + 2(4 - 2.92)^2 + 3(5 - 2.92)^2}{13 - 1}$$

$$S^2 = \frac{40.9232}{12}$$

$$S^2 \approx 3.41$$

- 3. Sometimes it can be helpful to calculate the standard deviation by using a table. Use the data to fill in the rest of the table and then use the table to calculate the sample standard deviation.

Data value	Data value - Mean	Squared difference
97		
110		
112		
121		
110		
98		
<b>Total</b>		

*Solution:*

We'll first calculate the mean of the data values given in the table.

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

$$\bar{x} = \frac{97 + 110 + 112 + 121 + 110 + 98}{6}$$

$$\bar{x} = \frac{648}{6}$$

$$\bar{x} = 108$$

Now we can fill out the table.

Data value	Data value - Mean	Squared difference
97	97-108=-11	(-11) <sup>2</sup> =121
110	110-108=2	2 <sup>2</sup> =4
112	112-108=4	4 <sup>2</sup> =16
121	121-108=13	13 <sup>2</sup> =169
110	110-108=2	2 <sup>2</sup> =4
98	98-108=-10	(-10) <sup>2</sup> =100
<b>Total</b>		<b>121+4+16+169+4+100=414</b>

The sum of the squared differences is 414. So sample variance is

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$$S^2 = \frac{414}{5}$$

$$S^2 = 82.8$$

So the sample standard deviation is



$$\sqrt{s^2} = \sqrt{82.8}$$

$$s \approx 9.099$$

- 4. The sum of the squared differences from the population mean for a data set is 212. If the data set has 25 items, what is the population standard deviation?

*Solution:*

The formula for population variance is

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

The numerator gives the sum of the squared differences, so we can plug in from the problem.

$$\sigma^2 = \frac{212}{25}$$

$$\sigma^2 = 8.48$$

The population standard deviation is therefore

$$\sqrt{\sigma^2} = \sqrt{8.48}$$

$$\sigma \approx 2.91$$



■ 5. For the data set 40, 44, 47, 55, 60, 60, 65, 80, find

$$\sum_{i=1}^n (x_i - \bar{x})$$

What does this say about why we square the  $(x_i - \bar{x})$  in the variance and standard deviation formulas?

*Solution:*

The value of

$$\sum_{i=1}^n (x_i - \bar{x})$$

will be 0 for any data set. The sum of the deviations from the mean will always be 0, because the negative and positive values will cancel each other out. This is one of the reasons that  $(x_i - \bar{x})$  is squared in the standard deviation formulas.

To prove that this value is 0 for this particular data set, we'll first find the mean.

$$\bar{x} = \frac{40 + 44 + 47 + 55 + 60 + 60 + 65 + 80}{8}$$

$$\bar{x} = 56.375$$

Then we can find the sum.



$$\sum_{i=1}^n (x_i - \bar{x}) = (40 - 56.375) + (44 - 56.375) + (47 - 56.375) + (55 - 56.375)$$

$$+(60 - 56.375) + (60 - 56.375) + (65 - 56.375) + (80 - 56.375)$$

$$\sum_{i=1}^n (x_i - \bar{x}) = -16.375 - 12.375 - 9.375 - 1.375 + 3.625 + 3.625 + 8.625 + 23.625$$

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

- 6. Give an example of a situation where \$5 could represent a large standard deviation and another where \$5 could represent a small standard deviation.

*Solution:*

The idea of how large or small the standard deviation of a data set is really depends on what it is we're measuring. If, for example, we were measuring the price of a soft drink at a state fair, and we found a standard deviation of \$5, that's a large standard deviation. It's large because soft drinks usually do not cost very much and this would tell us that we need to hunt around for the best price.

On the other hand, if we were purchasing a specific type of car and we found that the standard deviation was \$5 among the dealerships we were considering, that standard deviation would be very small. Small enough, in



fact, that it wouldn't matter much where we bought the car because the prices would all be pretty much the same.



## FREQUENCY HISTOGRAMS AND POLYGONS, AND DENSITY CURVES

- 1. A dog walking company keeps track of how many times each dog receives a walk. 40% of all the dogs walked by the company received between 25 and 40 walks, and no dogs received more than 40 walks. How many dogs received between 0 and 25 walks, if the company walks 400 dogs?

*Solution:*

Because no dogs received more than 40 walks, that means 100% of the dogs received between 0 and 40 walks. Since 40% of the dogs received between 25 and 40 walks, that must mean that  $100\% - 40\% = 60\%$  of the 400 dogs received between 0 and 25 walks. This means  $0.60(400) = 240$  dogs took between 0 and 25 walks.

- 2. The number of crayons in each student's pencil box is

4, 1, 5, 5, 9, 11, 15, 13, 15, 14, 16, 17, 20, 16, 16, 17

Complete the frequency and relative frequency tables for the data and use it to create a relative frequency histogram.



Crayons	Frequency	Relative Frequency
1-5		
6-10		
11-15		
16-20		
<b>Totals:</b>		<b>100%</b>

*Solution:*

First count the number of items in each frequency interval and add that to the table, as well as calculate the total number of crayons.

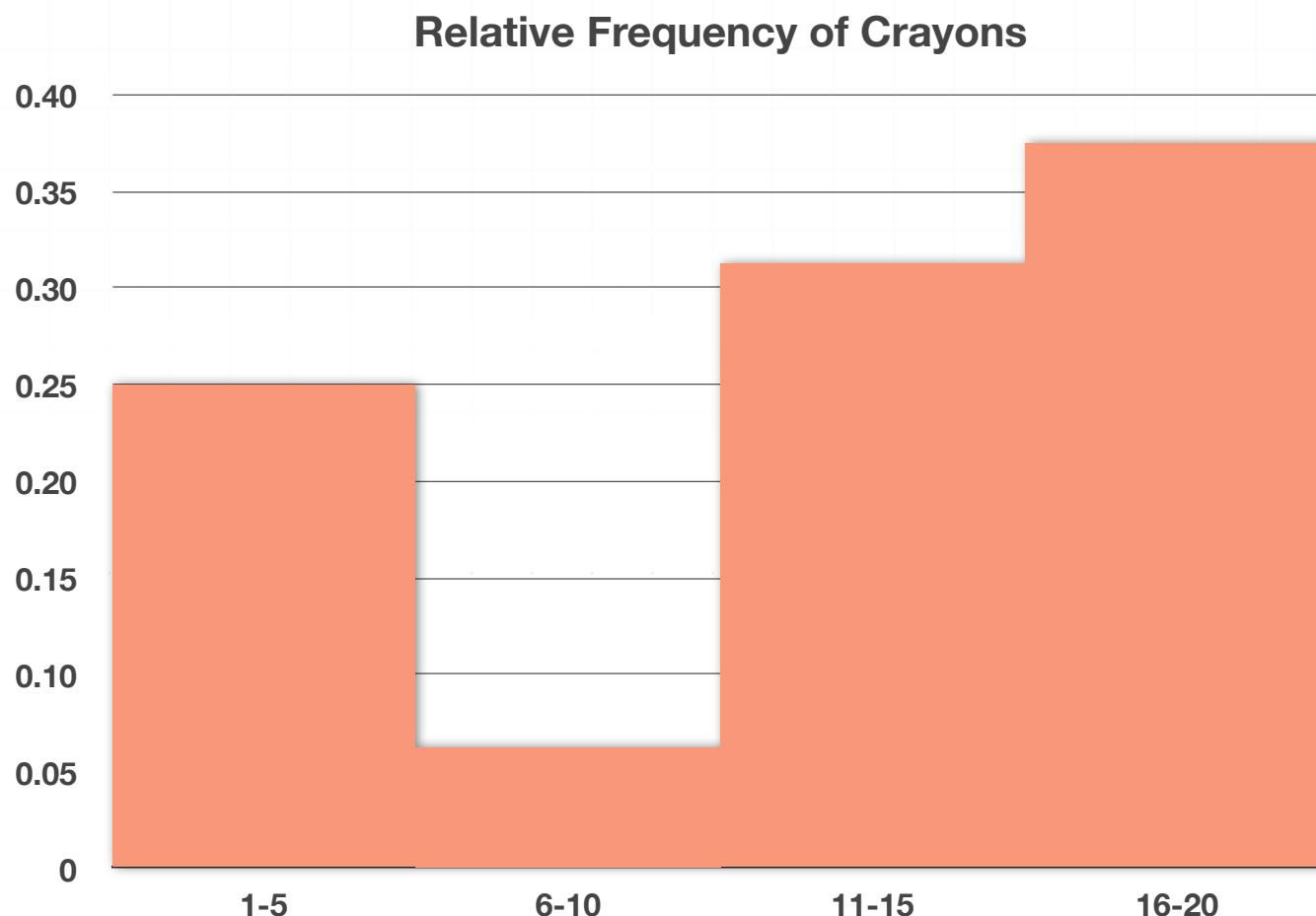
Crayons	Frequency	Relative Frequency
1-5	4	
6-10	1	
11-15	5	
16-20	6	
<b>Totals:</b>	<b>16</b>	<b>100%</b>

Next calculate the relative frequencies in the table by dividing the frequency by the total number of crayons.



Crayons	Frequency	Relative Frequency
1-5	4	$4/16=25\%$
6-10	1	$1/16=6.25\%$
11-15	5	$5/16=31.25\%$
16-20	6	$6/16=37.5\%$
<b>Totals:</b>	<b>16</b>	<b>100%</b>

Use the intervals on the horizontal axis and the relative frequencies on the vertical axis to make the histogram.



- 3. The table shows the scores on the last history exam in Mr. Ru's class.

40	32	40	83
95	33	87	59
32	81	46	78
91	61	55	88
40	61	82	99
72	47	83	91
101	77	65	87

Complete the relative frequency table and create a frequency polygon for the data.

Score	Frequency	Relative Frequency
30-39		
40-49		
50-59		
60-69		
70-79		
80-89		
90-99		
100-109		
<b>Totals:</b>		

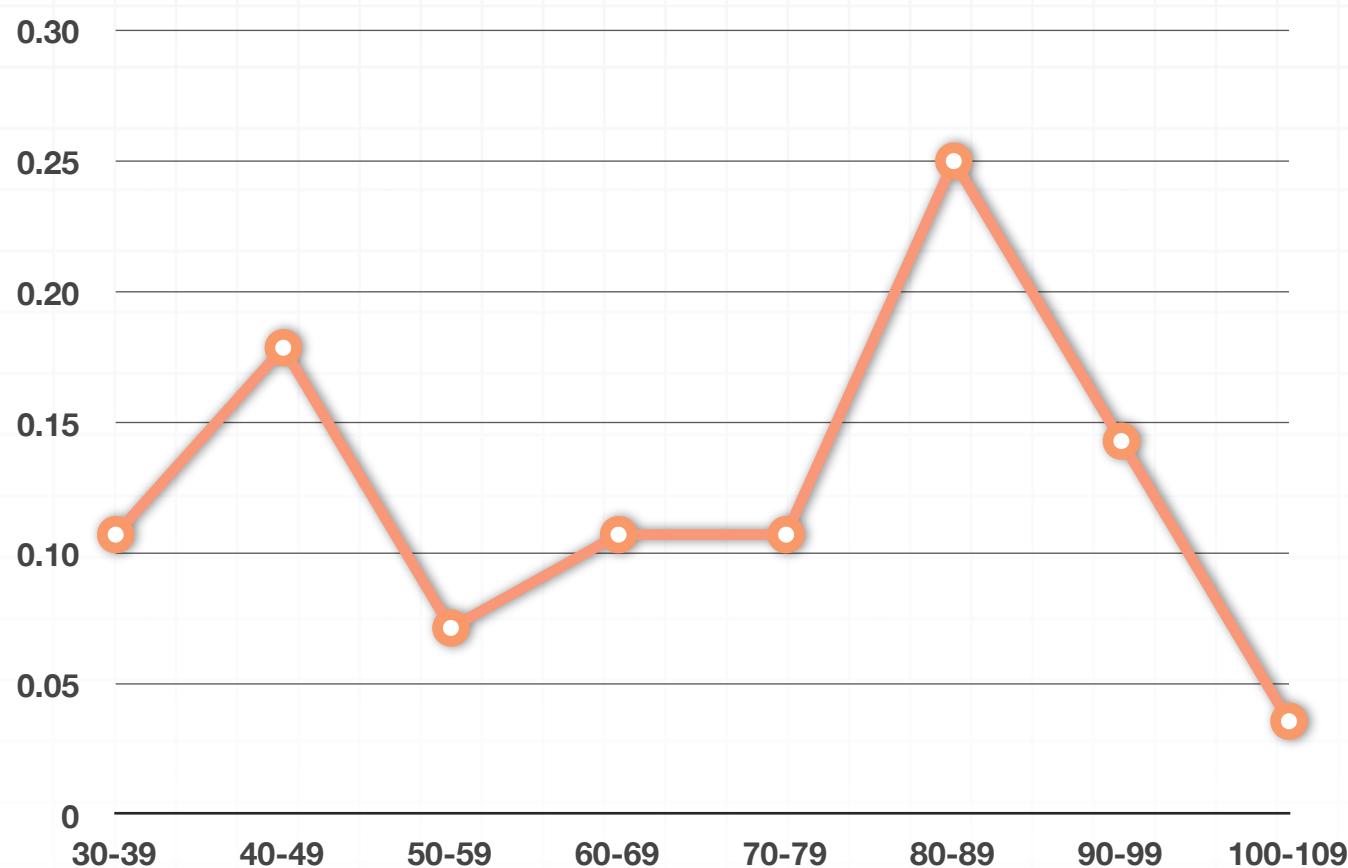
*Solution:*

The first step to completing the frequency table is to count the scores in each interval, then use those frequencies and the total number of test scores to calculate the relative frequencies.

Score	Frequency	Relative Frequency
30-39	3	$3/28=0.1071$
40-49	5	$5/28=0.1786$
50-59	2	$2/28=0.0714$
60-69	3	$3/28=0.1071$
70-79	3	$3/28=0.1071$
80-89	7	$7/28=0.2500$
90-99	4	$4/28=0.1429$
100-109	1	$1/28=0.0357$
<b>Totals:</b>	<b>28</b>	<b>100%</b>

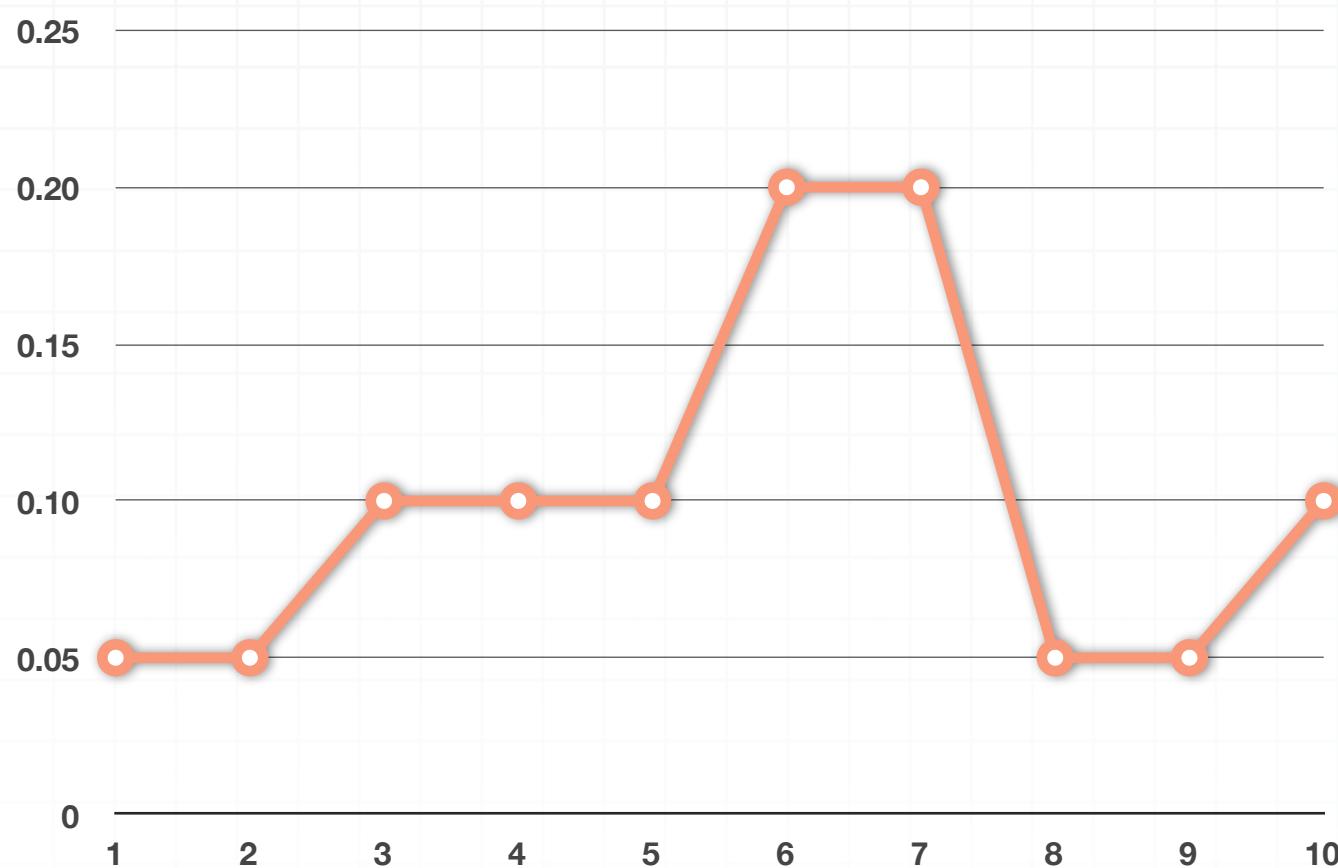
Use the intervals on the horizontal axis and the relative frequencies on the vertical axis to make the relative frequency polygon.



**Relative Frequency of Test Scores in Mr. Ru's Class**

- 4. Becky kept track of the number of ducks she saw at her neighborhood pond at 6 : 30 a.m. every morning for 365 days. On how many days did Becky see more than 5 ducks?

### Relative Frequency of Duck Sightings



*Solution:*

We want to know on how many days Becky saw 6, 7, 8, 9, and 10 ducks. We can organize our data into a table to read the values we want. Read the relative frequencies from the frequency polygon.

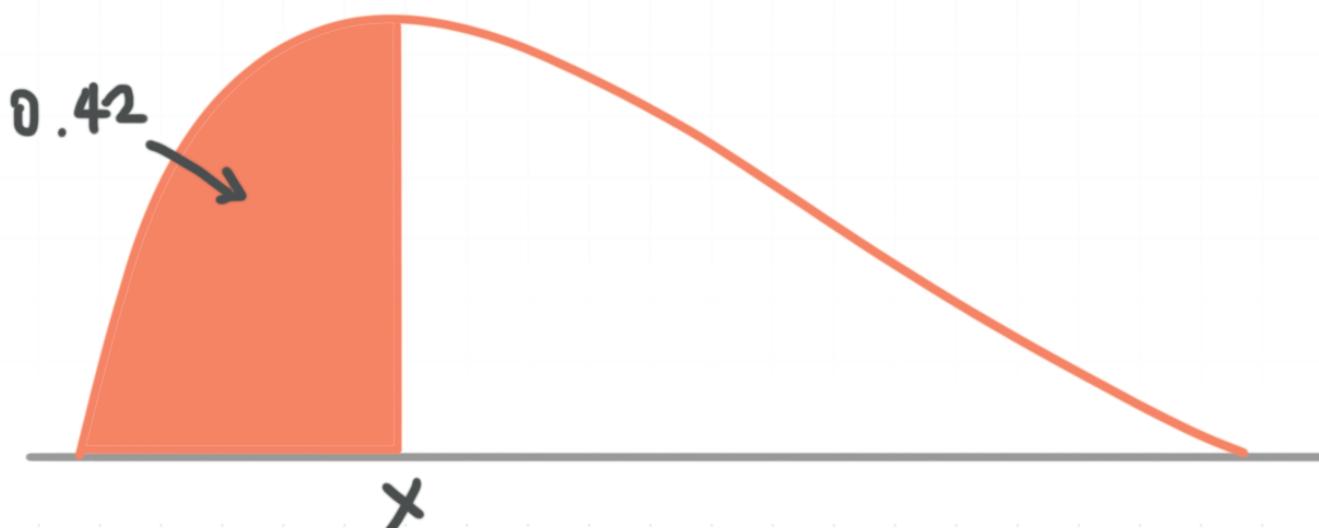
Ducks	Relative Frequency
6	0.20
7	0.20
8	0.05
9	0.05
10	0.10

Add the relative frequencies from 6 to 10. The cumulative relative frequency is

$$0.20 + 0.20 + 0.05 + 0.05 + 0.10 = 0.60 = 60\%$$

She took 365 days of data, which means she saw more than five ducks on  $0.60(365) = 219$  days.

■ 5. What percentage of the population is greater than  $x$  for the density curve?

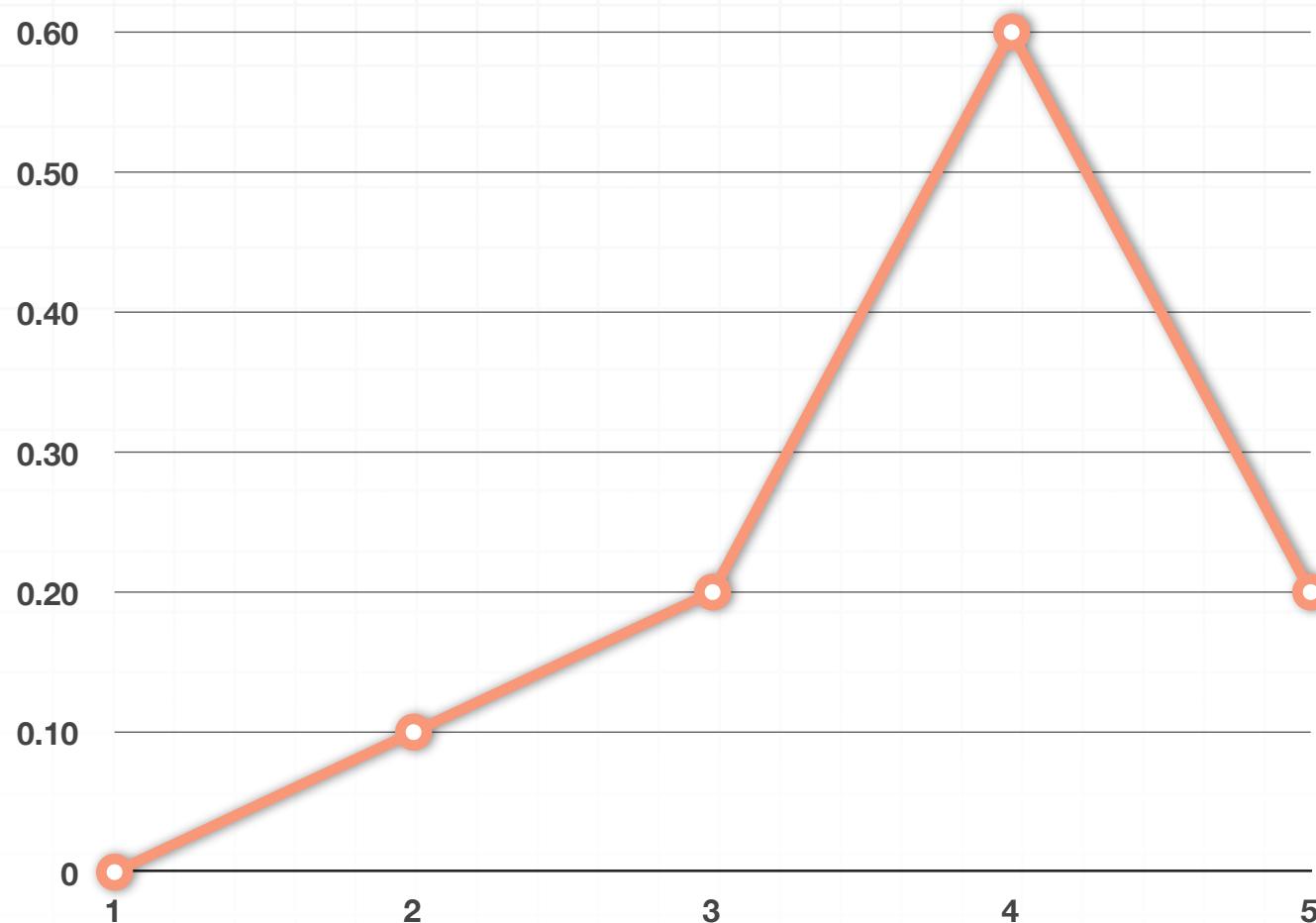


*Solution:*

Remember that the area under a density curve always adds to 1. Therefore everything greater than  $x$  must be

$$1 - 0.42 = 0.58 = 58\%$$

■ 6. What percentage of the area in the density curve is between 3 and 5?



*Solution:*

We know that for a density curve, the area under the curve adds to 1. We can use area formulas to find the density under certain parts of the curve.

The area under the curve between 1 and 3 is a triangle, so the area can be found as

$$A = \frac{1}{2}bh = \frac{1}{2}(2)(0.2) = 0.2$$

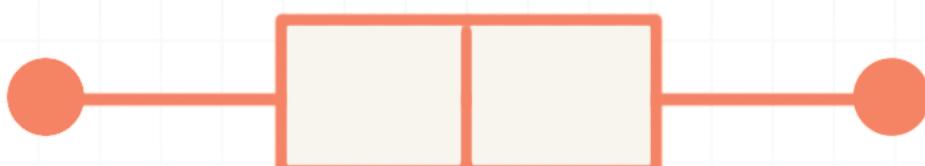
Which means the area under the rest of the polygon is the area between 3 and 5 and must be

$$1 - 0.2 = 0.8$$



## SYMMETRIC AND SKEWED DISTRIBUTIONS AND OUTLIERS

- 1. Which type of distribution is modeled in the box plot (symmetric, negatively skewed, or positively skewed)?



*Solution:*

This is an example of a symmetric distribution. The mean and the median are equal because the median of the data is in the middle of the box plot.

- 2. Which type of distribution is modeled in the box plot (symmetric, negatively skewed, or positively skewed)?



*Solution:*

This is an example of a positively skewed distribution. The median of the box plot is to the left of the middle of the box. This makes the mean greater than the median.

■ 3. The ages (in months) that babies spoke for the first time are

6, 8, 9, 10, 10, 11, 11, 12, 12, 13, 15, 15, 18, 19, 20, 20, 21

Are there outliers in the data set? If so, state what they are. What is the best measure of central tendency for the data? What is the best measure of spread?

*Solution:*

This data has no outliers, so the best measure of central tendency is the mean, and the best measure of spread is the standard deviation. To find if there are outliers in the data, use the 1.5-IQR rule.

Low outliers are given by  $Q_1 - 1.5(\text{IQR})$

High outliers are given by  $Q_3 + 1.5(\text{IQR})$

In the data set, the median is 12. And the first and third quartiles are

$$Q_1 = \frac{10 + 10}{2} = 10$$

$$Q_3 = \frac{18 + 19}{2} = 18.5$$

The interquartile range is

$$Q_3 - Q_1 = 18.5 - 10 = 8.5$$



Now we can calculate the boundary for outliers.

**Low outliers:**

$$Q_1 - 1.5(\text{IQR})$$

$$10 - 1.5(8.5)$$

$$-2.75$$

**High outliers:**

$$Q_3 + 1.5(\text{IQR})$$

$$18.5 + 1.5(8.5)$$

$$31.25$$

Since the data set has no values below  $-2.75$  or above  $31.25$ , there are no outliers in the data set.

■ 4. The number of text messages sent each day by Lucy's mom is

0, 18, 19, 20, 20, 20, 21, 23, 23, 23, 24, 24,

24, 24, 24, 25, 25, 25, 25, 25, 25, 30, 30, 31

Are there outliers in the data set? If so, state what they are. What is the best measure of central tendency for the data? What is the best measure of spread?



*Solution:*

This data has a low outlier of 0, so the best measure of central tendency is the median and the best measure of spread is the interquartile range. To see if there are outliers in the data use the 1.5-IQR rule.

Low outliers are given by  $Q_1 - 1.5(\text{IQR})$

High outliers are given by  $Q_3 + 1.5(\text{IQR})$

The median of the data set is 24. The first and third quartiles are

$$Q_1 = \frac{20 + 21}{2} = 20.5$$

$$Q_3 = \frac{25 + 25}{2} = 25$$

So the interquartile range is

$$Q_3 - Q_1 = 25 - 20.5 = 4.5$$

Now we can calculate where to look for outliers.

Low outliers:

$$Q_1 - 1.5(\text{IQR})$$

$$20.5 - 1.5(4.5)$$

$$13.75$$

High outliers:



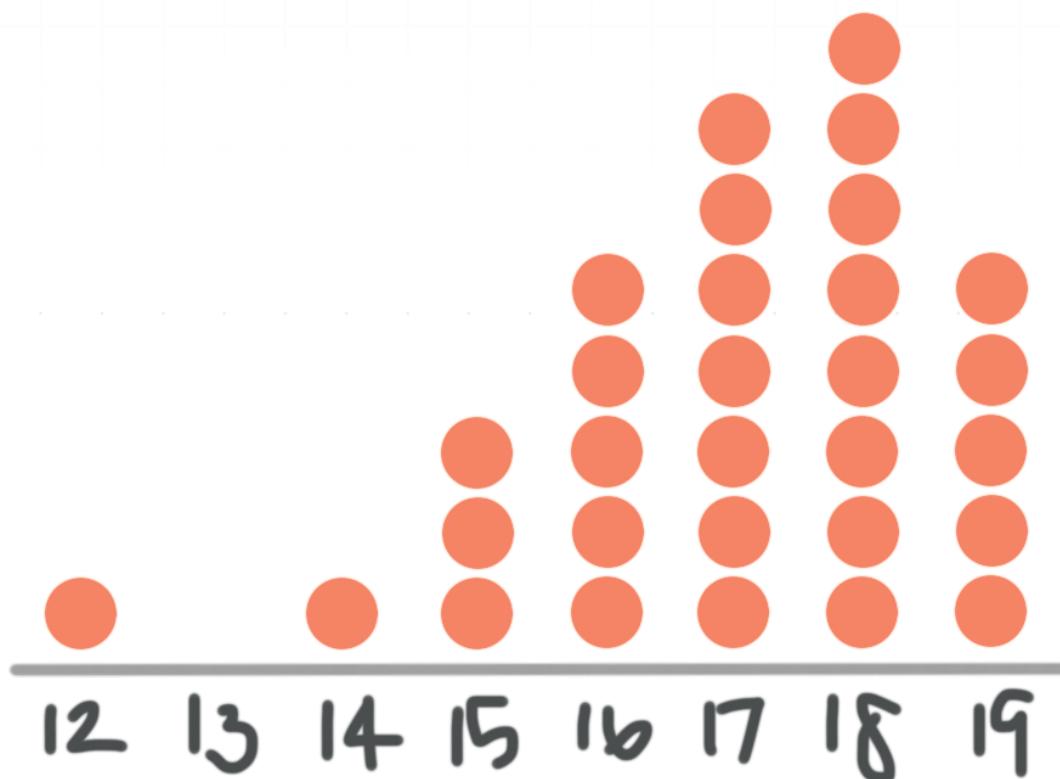
$$Q_3 + 1.5(\text{IQR})$$

$$25 + 1.5(4.5)$$

$$31.75$$

The data has a low outlier of 0 because it's less than 13.75. The data has no high outliers because no numbers in the set are greater than 31.75. Since the data has an outlier, the best measure of central tendency is the median and the best measure of spread is the interquartile range.

- 5. Describe the shape, center, and spread of the data. State if there are outliers and what they are if they exist.



*Solution:*

This data is negatively skewed, because it has a tail on the left-hand side with an outlier at 12. This means that the median will be the best measure of center and the interquartile range will be the best measure of spread. The median of the data is 17. The first and third quartile are  $Q_1 = 16$  and  $Q_3 = 18$ , so the interquartile range is  $Q_3 - Q_1 = 18 - 16 = 2$ .

This means that low outliers are any values less than  $Q_1 - 1.5(\text{IQR}) = 16 - 1.5(2) = 16 - 3 = 13$ , and high outliers are any values greater than  $Q_3 + 1.5(\text{IQR}) = 18 + 1.5(2) = 18 + 3 = 21$ . Based on the dot plot, 12 is a low outlier, and there are no high outliers.

- 6. Describe the shape, center and spread of the data. State if there are outliers and what they are if they exist.



*Solution:*

This is a symmetric distribution that is approximately normal. There are no outliers in the data set. The best measure of center will be the mean

(which is the same as the median) and the best measure of spread will be the standard deviation.

The mean of the data set is  $\mu = 25$  and the population standard deviation is 1.5811. To get the standard deviation, we would need to first calculate variance.

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{1(22 - 25)^2 + 2(23 - 25)^2 + 3(24 - 25)^2 + 4(25 - 25)^2 + 3(26 - 25)^2 + 2(27 - 25)^2 + 1(28 - 25)^2}{16}$$

$$\sigma^2 = \frac{1(-3)^2 + 2(-2)^2 + 3(-1)^2 + 4(0)^2 + 3(1)^2 + 2(2)^2 + 1(3)^2}{16}$$

$$\sigma^2 = \frac{1(9) + 2(4) + 3(1) + 4(0) + 3(1) + 2(4) + 1(9)}{16}$$

$$\sigma^2 = \frac{9 + 8 + 3 + 0 + 3 + 8 + 9}{16}$$

$$\sigma^2 = \frac{40}{16}$$

$$\sigma^2 = 2.5$$

Now take the square root of the population variance to find the population standard deviation.

$$\sqrt{\sigma^2} = \sqrt{2.5}$$

$$\sigma \approx 1.5811$$

## NORMAL DISTRIBUTIONS AND Z-SCORES

- 1. A population has a mean of 62 and a standard deviation of 5. What is the  $z$ -score for a value of 50?

*Solution:*

The formula for a  $z$ -score is:

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

We know the mean is  $\mu = 62$  and that the standard deviation is  $\sigma = 5$ . The value of interest is  $x = 50$ . So the  $z$ -score is

$$z = \frac{50 - 62}{5} = -\frac{12}{5} = -2.4$$

- 2. What percentile is a  $z$ -score of  $-1.68$ ?

*Solution:*

To find the percentile, we'll look up the  $z$ -score in the  $z$ -table. The amount in the table is 0.0465, which rounds to about 5%, so the  $z$ -score is associated with approximately the 5th percentile.



3. A population has a mean of 170 centimeters and a standard deviation of 8 centimeters. What percentage of the population has a value less than 154 centimeters?

*Solution:*

The mean is  $\mu = 170$  and that the standard deviation is  $\sigma = 8$ . The value we're interested in is  $x = 154$ . So the  $z$ -score is

$$z = \frac{154 - 170}{8} = -\frac{16}{8} = -2.00$$

If we look up this  $z$ -score in the  $z$ -table, we find 0.0228, so about 2.28 % of the population has a value less than 154 centimeters.

4. The mean diameter of a North American Native Pine tree is 18" with a standard deviation of 4". What is the approximate diameter for a tree in the 21st percentile for this distribution? Assume an approximately normal distribution.

*Solution:*

We know that the mean is  $\mu = 18$  and that the standard deviation is  $\sigma = 4$ . If we look up the 21st percentile, or 0.2100 in a  $z$ -table, we get a  $z$ -score of  $-0.81$ . Plugging all this into the  $z$ -score formula, we get



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

$$-0.81 = \frac{x - 18}{4}$$

$$-0.81(4) = x - 18$$

$$-3.24 = x - 18$$

$$14.76 = x$$

- 5. The mean diameter of a North American Native Pine tree is 18" with a standard deviation of 4". According to the Empirical Rule, 68 % of North American Native Pines have a diameter between which two values? Assume an approximately normal distribution.

*Solution:*

According to the Empirical Rule, 68 % of an approximately normal distribution is within one standard deviation of the mean. Since we know that  $\mu = 18$  and  $\sigma = 4$ , 68 % of these pines have a diameter on the interval

$$(18 - 4, 18 + 4)$$

$$(14, 22)$$



6. IQ scores are normally distributed with a mean of 100 and a standard deviation of 16. What percentage of the population has an IQ score between 120 and 140?

*Solution:*

First, we need to find the percentage of people who have an IQ of at most 120 and then the percentage of people with an IQ of at most 140, and then subtract those percentages. This means we find those  $z$ -scores and look up the percentages on the  $z$ -table.

Since we know that  $\mu = 100$  and  $\sigma = 16$ , the  $z$ -score for 120 is

$$z = \frac{120 - 100}{16} = \frac{20}{16} = 1.25$$

which gives .8944 in the  $z$ -table. The  $z$ -score for 140 is

$$z = \frac{140 - 100}{16} = \frac{40}{16} = 2.5$$

which gives .9938 in the  $z$ -table. Therefore, we can say that

$$.9938 - .8944 = .0994 = 9.94\%$$

percent of people have an IQ between 120 and 140.



## CHEBYSHEV'S THEOREM

- 1. If the Empirical Rule tells us that 95 % of the area under the normal distribution falls within two standard deviations of the mean, what will Chebyshev's Theorem say about the same number of standard deviations?

*Solution:*

Because Chebyshev's Theorem has to apply to distributions of all shapes, it's always more conservative than the Empirical Rule.

Therefore, we know that Chebyshev's Theorem will only be able to conclude that less than 95 % of the area under distribution will fall within two standard deviations of the mean.

In actuality, Chebyshev's Theorem says that at least 75 % of the data under the probability distribution will fall within two standard deviations of the mean.

- 2. A basket of strawberries has a mean weight of 2 ounces with a standard deviation of 0.35 ounces. What percentage of the strawberries in the basket have a weight between 1.5 and 2.5 ounces?



*Solution:*

Determine the distance from the mean of 1.5 and 2.5, in terms of standard deviations.

$$k = \frac{1.5 - 2}{0.35} = -\frac{0.5}{0.35} \approx -1.43$$

$$k = \frac{2.5 - 2}{0.35} = \frac{0.5}{0.35} \approx 1.43$$

With  $k = 1.43$ , Chebyshev's Theorem gives

$$1 - \frac{1}{k^2} = 1 - \frac{1}{1.43^2} \approx 1 - \frac{1}{2.04} \approx 0.51$$

Because we found 0.51, we know at least 51 % of the strawberries have a weight between 1.5 and 2.5 ounces.

- 3. A pod of 580 migrating whales travels a mean distance of 2,000 miles each year, with a standard deviation of 175 miles. How many whales in the pod travel between 1,600 and 2,400 miles?

*Solution:*

Determine the distance from the mean of 1,750 and 2,250, in terms of standard deviations.

$$k = \frac{1,600 - 2,000}{175} = -\frac{400}{175} \approx -2.29$$



$$k = \frac{2,400 - 2,000}{175} = \frac{400}{175} \approx 2.29$$

With  $k = 2.29$ , Chebyshev's Theorem gives

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.29^2} \approx 1 - \frac{1}{5.22} \approx 0.81$$

Because we found 0.81, we know at least 81% of the whales migrate between 1,600 and 2,400 miles. Then we can say that 81% of the 580-whale pod is approximately 469.8 whales.

Of course we can't take part of a whale, so in order not to overstate Chebyshev's Theorem, we round down to 469 whales.

- 4. A hockey team of 20 boys have a mean height of 73 inches, with a standard deviation of 1.8 inches. Find the height range for the central 90% of team members.

*Solution:*

Using Chebyshev's Theorem,

$$0.9 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.9$$

$$1 = 0.1k^2$$



$$k^2 = 10$$

$$k \approx 3.16$$

Approximately 3.16 standard deviations above the mean gives us a height of

$$73 + 3.16(1.8)$$

$$73 + 5.69$$

$$78.69$$

And 3.16 standard deviations below the mean gives us a height of

$$73 - 3.16(1.8)$$

$$73 - 5.69$$

$$67.31$$

So at least 90 % of the players have heights between 67.31 and 78.69 inches.

- 5. A university with 40,000 students accepts an average of 10,000 new students each year, with a standard deviation of 500 students. Find the values that make up the middle 75 % of the yearly acceptance range.

*Solution:*

Using Chebyshev's Theorem,



$$0.75 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.75$$

$$1 = 0.25k^2$$

$$k^2 = 4$$

$$k = 2$$

Two standard deviations above the mean gives us the upper end of students accepted.

$$10,000 + 2(500)$$

$$10,000 + 1,000$$

$$11,000$$

And two standard deviations below the mean gives us the lower end of students accepted.

$$10,000 - 2(500)$$

$$10,000 - 1,000$$

$$9,000$$

So at least 75 % of the time (75 % of years), the school accepts between 9,000 and 11,000 students.



6. A pack of 26 wolves have a mean weight of 100 pounds, with a standard deviation of 24 pounds. Find the weight range for the central 82% of the wolves.

*Solution:*

Using Chebyshev's Theorem,

$$0.82 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.82$$

$$1 = 0.18k^2$$

$$k^2 \approx 5.56$$

$$k \approx 2.36$$

Approximately 2.36 standard deviations above the mean gives us a weight of

$$100 + 2.36(24)$$

$$100 + 56.57$$

$$156.57$$

And 2.36 standard deviations below the mean gives us a weight of

$$100 - 2.36(24)$$

$$100 - 56.57$$

$$43.43$$

So at least 82 % of the wolves have weights between 43.43 and 156.57 pounds.



## COVARIANCE

- 1. A bakery records sales and number of customers for a sample of hours throughout the week. Calculate the covariance of customers and sales.

<b>Customers</b>	4	7	12	2	3	9	15
<b>Sales</b>	45.75	36.00	58.5	20.00	15.80	39.95	123.45

*Solution:*

We'll find the mean number of customers,

$$\bar{x} = \frac{4 + 7 + 12 + 2 + 3 + 9 + 15}{7} \approx 7.43$$

and the mean revenue in dollars.

$$\bar{y} = \frac{45.75 + 36 + 58.50 + 20 + 15.80 + 39.95 + 123.45}{7} \approx 48.49$$

Now we'll use the means to find the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (4 - 7.43)(45.75 - 48.49) + (7 - 7.43)(36 - 48.49)$$

$$+ (12 - 7.43)(58.50 - 48.49) + (2 - 7.43)(20 - 48.49)$$

$$+ (3 - 7.43)(15.80 - 48.49) + (9 - 7.43)(39.90 - 48.49)$$

$$+(15 - 7.43)(123.45 - 48.49)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) \approx 913.9929$$

$$s_{xy} \approx \frac{913.9929}{7 - 1}$$

$$s_{xy} \approx 152.33$$

- 2. The cost of the stock of two unrelated companies over five days is recorded in the table. Calculate the covariance of the stocks.

Company X	13	13.75	12.70	13.15	14.80
Company Y	21.05	21.55	20.95	21.75	21.50

*Solution:*

We'll find the mean cost of the stock for company  $X$ ,

$$\bar{X} = \frac{13 + 13.75 + 12.70 + 13.15 + 14.80}{5} = 13.48$$

and the mean cost of the stock for company  $Y$ .

$$\bar{Y} = \frac{21.05 + 21.55 + 20.95 + 21.75 + 21.50}{5} = 21.36$$

Now we'll use the means to find the sample covariance.



$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (13 - 13.48)(21.05 - 21.36)$$

$$+(13.75 - 13.48)(21.55 - 21.36) + (12.70 - 13.48)(20.95 - 21.36)$$

$$+(13.15 - 13.48)(21.75 - 21.36) + (14.80 - 13.48)(21.50 - 21.36)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 0.576$$

$$s_{XY} = \frac{0.576}{5 - 1}$$

$$s_{XY} = 0.144$$

- 3. The following table represents temperatures, in Celsius, during a sample of 5 days in two cities with a distance of 50 miles between them. Calculate the covariance.

City X	25	23	24.5	20	18
City Y	23	24	21	18	22

*Solution:*

We'll find the mean temperature in city X,

$$\bar{X} = \frac{25 + 23 + 24.5 + 20 + 18}{5} = 22.1$$

and the mean temperature in city  $Y$ .

$$\bar{Y} = \frac{23 + 24 + 21 + 18 + 22}{5} = 21.6$$

Now we'll use the means to find the sample covariance.

$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (25 - 22.1)(23 - 21.6) + (23 - 22.1)(24 - 21.6)$$

$$+ (24.5 - 22.1)(21 - 21.6) + (20 - 22.1)(18 - 21.6)$$

$$+ (18 - 22.1)(22 - 21.6)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 10.7$$

$$s_{XY} = \frac{10.7}{5 - 1}$$

$$s_{XY} = 2.675$$

- 4. David prepares for his annual math and physics exams and decides to take four practice tests for each subject. Calculate the covariance for his test scores for math and physics.

Math, X	85	89	89	93
Physics, Y	92	93	89	90

*Solution:*

We'll find the mean score of the math practice tests,

$$\bar{X} = \frac{85 + 89 + 89 + 93}{4} = 89$$

and the mean score of the physics practice tests.

$$\bar{Y} = \frac{92 + 93 + 89 + 90}{4} = 91$$

Now we'll use the means to find the sample covariance.

$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\begin{aligned} \sum (X_i - \bar{X})(Y_i - \bar{Y}) &= (85 - 89)(92 - 91) + (89 - 89)(93 - 91) \\ &\quad + (89 - 89)(89 - 91) + (93 - 89)(90 - 91) \end{aligned}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -4$$

$$s_{XY} = \frac{-4}{4 - 1}$$

$$s_{XY} \approx -1.33$$

- 5. Mark and John exercise daily and record their minutes of daily exercise over 10 days. Calculate the covariance.



<b>Mark</b>	53	57	63	55	45	50	65	60	59	70
<b>John</b>	65	55	60	53	30	45	25	65	57	50

*Solution:*

We'll find the mean amount of time, in minutes, that Mark spent exercising daily for the last 10 days,

$$\bar{M} = \frac{53 + 57 + 63 + 55 + 45 + 50 + 65 + 60 + 59 + 70}{10} = 57.7$$

and the mean amount of time, in minutes, that John spent exercising daily for the last 10 days.

$$\bar{J} = \frac{65 + 55 + 60 + 53 + 30 + 45 + 25 + 65 + 57 + 50}{10} = 50.5$$

Now we'll use the means to find the sample covariance.

$$s_{MJ} = \frac{\sum (M_i - \bar{M})(J_i - \bar{J})}{n - 1}$$

$$\begin{aligned} \sum (M_i - \bar{M})(J_i - \bar{J}) &= (53 - 57.7)(65 - 50.5) + (57 - 57.7)(55 - 50.5) \\ &\quad + (63 - 57.7)(60 - 50.5) + (55 - 57.7)(53 - 50.5) \\ &\quad + (45 - 57.7)(30 - 50.5) + (50 - 57.7)(45 - 50.5) \\ &\quad + (65 - 57.7)(25 - 50.5) + (60 - 57.7)(65 - 50.5) \\ &\quad + (59 - 57.7)(57 - 50.5) + (70 - 57.7)(50 - 50.5) \end{aligned}$$

$$+(70 - 57.7)(50 - 50.5)$$

$$\sum (M_i - \bar{M})(J_i - \bar{J}) = 118.35$$

$$s_{MJ} = \frac{118.35}{10 - 1}$$

$$s_{MJ} \approx 13.15$$

- 6. An annual return on investment of two stocks over the last 7 years is recorded in the table. Calculate the covariance.

<b>Stock X</b>	3.5	2.4	1.4	-0.5	0.7	1.1	0.5
<b>Stock Y</b>	2.4	1.7	2.1	1.8	2.1	-0.4	0.8

*Solution:*

We'll find the mean annual return of stock X,

$$\bar{X} = \frac{3.5 + 2.4 + 1.4 + (-0.5) + 0.7 + 1.1 + 0.5}{7} = 1.3$$

and the mean annual return of stock Y.

$$\bar{Y} = \frac{2.4 + 1.7 + 2.1 + 1.8 + 2.1 + (-0.4) + 0.8}{7} = 1.5$$

Now we'll use the means to find the sample covariance.



$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (3.5 - 1.3)(2.4 - 1.5) + (2.4 - 1.3)(1.7 - 1.5)$$

$$+(1.4 - 1.3)(2.1 - 1.5) + (-0.5 - 1.3)(1.8 - 1.5)$$

$$+(0.7 - 1.3)(2.1 - 1.5) + (1.1 - 1.3)(-0.4 - 1.5)$$

$$+(0.5 - 1.3)(0.8 - 1.5)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 2.3$$

$$s_{XY} = \frac{2.3}{7 - 1}$$

$$s_{XY} \approx 0.38$$



## CORRELATION COEFFICIENT

- 1. Calculate the correlation coefficient of the newborns' weight and body length, and then interpret the result.

Weight, kg	Body length, cm
3.55	51
4.01	54
3.05	50
5.35	60
4.22	52
6.12	61
7.45	63
5.95	59
6.35	68
6.98	74

*Solution:*

Start by calculating mean weight,

$$\bar{x} = \frac{3.55 + 4.01 + 3.05 + 5.35 + 4.22 + 6.12 + 7.45 + 5.95 + 6.35 + 6.98}{10}$$

$$\bar{x} = 5.303$$

and mean body length.



$$\bar{y} = \frac{51 + 54 + 50 + 60 + 52 + 61 + 63 + 59 + 68 + 74}{10}$$

$$\bar{y} = 59.2$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (3.55 - 5.303)(51 - 59.2)$$

$$+(4.01 - 5.303)(54 - 59.2) + (3.05 - 5.303)(50 - 59.2)$$

$$+(5.35 - 5.303)(60 - 59.2) + (4.22 - 5.303)(52 - 59.2)$$

$$+(6.12 - 5.303)(61 - 59.2) + (7.45 - 5.303)(63 - 59.2)$$

$$+(5.95 - 5.303)(59 - 59.2) + (6.35 - 5.303)(68 - 59.2)$$

$$+(6.98 - 5.303)(74 - 59.2)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 93.194$$

$$s_{xy} = \frac{93.194}{10 - 1}$$

$$s_{xy} \approx 10.35$$

Next we'll need the standard deviation for weight,

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (3.55 - 5.303)^2 + (4.01 - 5.303)^2 + (3.05 - 5.303)^2$$

$$+ (5.35 - 5.303)^2 + (4.22 - 5.303)^2 + (6.12 - 5.303)^2 + (7.45 - 5.303)^2$$

$$+ (5.95 - 5.303)^2 + (6.35 - 5.303)^2 + (6.98 - 5.303)^2$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 \approx 20.60$$

$$s_x \approx \sqrt{\frac{20.60}{10 - 1}}$$

$$s_x \approx 1.513$$

and the standard deviation for body length.

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (51 - 59.2)^2 + (54 - 59.2)^2 + (50 - 59.2)^2 + (60 - 59.2)^2$$

$$+ (52 - 59.2)^2 + (61 - 59.2)^2 + (63 - 59.2)^2 + (59 - 59.2)^2$$

$$+ (68 - 59.2)^2 + (74 - 59.2)^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 545.6$$

$$s_y = \sqrt{\frac{545.6}{10 - 1}}$$

$$s_y \approx 7.786$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{10.35}{1.513 \cdot 7.786}$$

$$r_{xy} \approx 0.8786$$

The value of the correlation coefficient indicates that there's a strong positive correlation between the weight and body length of newborns.

- 2. Oliver is wondering whether there's a correlation between the number of hours his classmates studied to prepare for the exam and their exam scores. He surveyed five classmates and recorded the data in a table. Calculate the correlation coefficient.

<b>Study hours</b>	6	2	11	7	5
<b>Exam score</b>	85	79	84	89	91

*Solution:*

Start by calculating mean number of study hours,



$$\bar{x} = \frac{6 + 2 + 11 + 7 + 5}{5}$$

$$\bar{x} = 6.2$$

and mean exam score.

$$\bar{y} = \frac{85 + 79 + 84 + 89 + 91}{5}$$

$$\bar{y} = 85.6$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (6 - 6.2)(85 - 85.6) + (2 - 6.2)(79 - 85.6)$$

$$+ (11 - 6.2)(84 - 85.6) + (7 - 6.2)(89 - 85.6) + (5 - 6.2)(91 - 85.6)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 16.4$$

$$s_{xy} = \frac{16.4}{5 - 1}$$

$$s_{xy} = 4.1$$

Next we'll need the standard deviation for the number of study hours,

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (6 - 6.2)^2 + (2 - 6.2)^2 + (11 - 6.2)^2 + (7 - 6.2)^2 + (5 - 6.2)^2$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 42.8$$

$$s_x = \sqrt{\frac{42.8}{5 - 1}}$$

$$s_x \approx 3.271$$

and the standard deviation for exam score.

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (85 - 85.6)^2 + (79 - 85.6)^2 + (84 - 85.6)^2$$

$$+(89 - 85.6)^2 + (91 - 85.6)^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 87.2$$

$$s_y = \sqrt{\frac{87.2}{5 - 1}}$$

$$s_y \approx 4.669$$

Now we can plug the covariance and standard deviations into the formula for correlation.



$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{4.1}{3.271 \cdot 4.669}$$

$$r_{xy} \approx 0.2685$$

There's a weak positive correlation between the number of study hours and exam score.

- 3. Calculate the value of the Pearson correlation coefficient for the age, in years, and blood glucose levels, in mg/dL, then interpret the result.

Age	28	35	58	42	21	63	46
Blood glucose	101	93	95	105	93	89	100

*Solution:*

Start by calculating mean age,

$$\bar{x} = \frac{28 + 35 + 58 + 42 + 21 + 63 + 46}{7}$$

$$\bar{x} \approx 41.86$$

and mean blood glucose.

$$\bar{y} = \frac{101 + 93 + 95 + 105 + 93 + 89 + 100}{7}$$

$$\bar{y} \approx 96.57$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (28 - 41.86)(101 - 96.57)$$

$$+(35 - 41.86)(93 - 96.57) + (58 - 41.86)(95 - 96.57)$$

$$+(42 - 41.86)(105 - 96.57) + (21 - 41.86)(93 - 96.57)$$

$$+(63 - 41.86)(89 - 96.57) + (46 - 41.86)(100 - 96.57)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = -56.7286$$

$$s_{xy} = \frac{-132.4286}{7 - 1}$$

$$s_{xy} \approx -22.071$$

Next we'll need the standard deviation for age,

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (28 - 41.86)^2 + (35 - 41.86)^2 + (58 - 41.86)^2$$

$$+(42 - 41.86)^2 + (21 - 41.86)^2 + (63 - 41.86)^2 + (46 - 41.86)^2$$



$$\sum_{i=1}^n (x_i - \bar{x})^2 = 1,398.8572$$

$$s_x = \sqrt{\frac{1,398.8572}{7 - 1}}$$

$$s_x \approx 15.269$$

and the standard deviation for blood glucose level.

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (101 - 96.57)^2 + (93 - 96.57)^2 + (95 - 96.57)^2$$

$$+ (105 - 96.57)^2 + (93 - 96.57)^2 + (89 - 96.57)^2 + (100 - 96.57)^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 187.7143$$

$$s_y = \sqrt{\frac{187.7143}{7 - 1}}$$

$$s_y \approx 5.593$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{-22.071}{15.269 \cdot 5.593}$$

$$r_{xy} \approx -0.2584$$

The value of the correlation coefficient indicates that there's a weak negative correlation between age and blood glucose levels.

- 4. Maria likes discovering interesting correlations. She decides to choose random six days and record the data for shark attacks and ice cream sales in her coastal city. How should she interpret the correlation coefficient.

<b>Shark attacks</b>	4	2	8	11	5	9
<b>Ice cream sales</b>	38	30	55	61	38	42

*Solution:*

Start by calculating mean shark attacks,

$$\bar{x} = \frac{4 + 2 + 8 + 11 + 5 + 9}{6}$$

$$\bar{x} = 6.5$$

and mean ice cream sales.

$$\bar{y} = \frac{38 + 30 + 55 + 61 + 38 + 42}{6}$$

$$\bar{y} = 44$$



Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (4 - 6.5)(38 - 44) + (2 - 6.5)(30 - 44) + (8 - 6.5)(55 - 44)$$

$$+ (11 - 6.5)(61 - 44) + (5 - 6.5)(38 - 44) + (9 - 6.5)(42 - 44)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 175$$

$$s_{xy} = \frac{175}{6 - 1}$$

$$s_{xy} = 35$$

Next we'll need the standard deviation for shark attacks,

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (4 - 6.5)^2 + (2 - 6.5)^2 + (8 - 6.5)^2 + (11 - 6.5)^2$$

$$+ (5 - 6.5)^2 + (9 - 6.5)^2$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 57.5$$

$$s_x = \sqrt{\frac{57.5}{6 - 1}}$$

$$s_x \approx 3.391$$

and the standard deviation for ice cream sales.

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= (38 - 44)^2 + (30 - 44)^2 + (55 - 44)^2 + (61 - 44)^2 \\ &\quad + (38 - 44)^2 + (42 - 44)^2 \end{aligned}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 682$$

$$s_y = \sqrt{\frac{682}{6 - 1}}$$

$$s_y \approx 11.679$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{35}{3.391 \cdot 11.679}$$

$$r_{xy} \approx 0.8838$$

There's strong positive correlation between the number of shark attacks and ice cream sales on a particular day.



However, it seems unlikely that an increase in ice cream sales is the cause of an increase in shark attacks. Instead, it seems more likely that a hot summer day causes more people to buy ice cream and more people to swim in the ocean, leading to more shark attacks.

■ 5. Calculate and interpret the correlation coefficient of the variables.

Hand length, cm	Height, cm
12	158
15	160
11	157
13	164
9	150
18	178
16	169
17	156

*Solution:*

Start by calculating mean hand length,

$$\bar{x} = \frac{12 + 15 + 11 + 13 + 9 + 18 + 16 + 17}{8}$$

$$\bar{x} = 13.875$$

and mean height.



$$\bar{y} = \frac{158 + 160 + 157 + 164 + 150 + 178 + 169 + 156}{8}$$

$$\bar{y} = 161.5$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\begin{aligned}\sum (x_i - \bar{x})(y_i - \bar{y}) &= (12 - 13.875)(158 - 161.5) + (15 - 13.875)(160 - 161.5) \\ &\quad + (11 - 13.875)(157 - 161.5) + (13 - 13.875)(164 - 161.5) \\ &\quad + (9 - 13.875)(150 - 161.5) + (18 - 13.875)(178 - 161.5) \\ &\quad + (16 - 13.875)(169 - 161.5) + (17 - 13.875)(156 - 161.5)\end{aligned}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 138.5$$

$$s_{xy} = \frac{138.5}{8 - 1}$$

$$s_{xy} \approx 19.786$$

Next we'll need the standard deviation for hand length,

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (12 - 13.875)^2 + (15 - 13.875)^2 + (11 - 13.875)^2$$



$$+(13 - 13.875)^2 + (9 - 13.875)^2 + (18 - 13.875)^2$$

$$+(16 - 13.875)^2 + (17 - 13.875)^2$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 68.875$$

$$s_x = \sqrt{\frac{68.875}{8 - 1}}$$

$$s_x \approx 3.137$$

and the standard deviation for height.

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (158 - 161.5)^2 + (160 - 161.5)^2 + (157 - 161.5)^2$$

$$+(164 - 161.5)^2 + (150 - 161.5)^2 + (178 - 161.5)^2$$

$$+(169 - 161.5)^2 + (156 - 161.5)^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 532$$

$$s_y = \sqrt{\frac{532}{8 - 1}}$$

$$s_y \approx 8.718$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{19.786}{3.137 \cdot 8.718}$$

$$r_{xy} \approx 0.7235$$

There's a strong positive correlation between hand length and height.

- 6. Calculate and interpret the value of the correlation coefficient for the correlation between systolic blood pressure, in mmHg, and weight, in lbs.

SBP, mmHg	Weight, lbs
138	167
125	153
145	149
156	165
132	170
148	175
160	180
135	140
150	190
155	155

*Solution:*

Start by calculating mean systolic blood pressure,

$$\bar{x} = \frac{138 + 125 + 145 + 156 + 132 + 148 + 160 + 135 + 150 + 155}{10}$$

$$\bar{x} = 144.4$$

and mean weight.

$$\bar{y} = \frac{167 + 153 + 149 + 165 + 170 + 175 + 180 + 140 + 190 + 155}{10}$$

$$\bar{y} = 164.4$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (138 - 144.4)(167 - 164.4) + (125 - 144.4)(153 - 164.4)$$

$$+ (145 - 144.4)(149 - 164.4) + (156 - 144.4)(165 - 164.4)$$

$$+ (132 - 144.4)(170 - 164.4) + (148 - 144.4)(175 - 164.4)$$

$$+ (160 - 144.4)(180 - 164.4) + (135 - 144.4)(140 - 164.4)$$

$$+ (150 - 144.4)(190 - 164.4) + (155 - 144.4)(155 - 164.4)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 687.4$$

$$s_{xy} = \frac{687.4}{10 - 1}$$

$$s_{xy} \approx 76.378$$

Next we'll need the standard deviation for systolic blood pressure,

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (138 - 144.4)^2 + (125 - 144.4)^2 + (145 - 144.4)^2$$

$$+ (156 - 144.4)^2 + (132 - 144.4)^2 + (148 - 144.4)^2 + (160 - 144.4)^2$$

$$+ (135 - 144.4)^2 + (150 - 144.4)^2 + (155 - 144.4)^2$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 1194.4$$

$$s_x = \sqrt{\frac{1194.4}{10 - 1}}$$

$$s_x \approx 11.52$$

and the standard deviation for weight.

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (167 - 164.4)^2 + (153 - 164.4)^2 + (149 - 164.4)^2$$

$$+ (165 - 164.4)^2 + (170 - 164.4)^2 + (175 - 164.4)^2 + (180 - 164.4)^2$$



$$+(140 - 164.4)^2 + (190 - 164.4)^2 + (155 - 164.4)^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 2,100.4$$

$$s_y = \sqrt{\frac{2,100.4}{10 - 1}}$$

$$s_y \approx 15.277$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{76.378}{11.52 \cdot 15.277}$$

$$r_{xy} \approx 0.4340$$

There's a moderate positive correlation between systolic blood pressure and weight.



## WEIGHTED MEANS AND GROUPED DATA

- 1. An investor purchases shares of a particular stock on the same date every month for 12 months. He records the price and number of shares each month. Calculate the mean share price.

Stock price	Shares
8	30
10	12
14	10
9	25
6	35
13	15
18	10
21	5
25	7
27	10
28	8
31	4

*Solution:*

We can calculate the weighted sample mean share price.

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

$$\sum_{i=1}^n w_i x_i = 30(8) + 12(10) + 10(14) + 25(9) + 35(6) + 15(13) + 10(18)$$

$$+ 5(21) + 7(25) + 10(27) + 8(28) + 4(31)$$

$$\sum_{i=1}^n w_i x_i = 2,208$$

$$\bar{x} = \frac{2,208}{30 + 12 + 10 + 25 + 35 + 15 + 10 + 5 + 7 + 10 + 8 + 4}$$

$$\bar{x} \approx 12.91$$

The mean share price is \$12.91.

- 2. A chemistry course teacher weights class discussions at 0.05, quizzes at 0.10, and group projects at 0.40. Given the grades for one student in the table below, calculate her final grade.

Assignment	Grade	Weight
Quiz 1	88	0.10
Discussion 1	92	0.05
Quiz 2	93	0.10
Discussion 2	90	0.05
Quiz 3	85	0.10
Discussion 3	94	0.05
Quiz 4	97	0.10
Discussion 4	80	0.05
Group Project	85	0.40

*Solution:*

This is the entire population of scores for the single student, so we'll calculate the weighted population mean.

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

$$\begin{aligned} \sum_{i=1}^N w_i x_i &= 0.10(88) + 0.05(92) + 0.10(93) + 0.05(90) + 0.10(85) \\ &\quad + 0.05(94) + 0.10(97) + 0.05(80) + 0.40(85) \end{aligned}$$

$$\sum_{i=1}^N w_i x_i = 88.1$$

$$\mu = \frac{88.1}{0.1 + 0.05 + 0.1 + 0.05 + 0.1 + 0.05 + 0.1 + 0.05 + 0.4}$$

$$\mu = 88.1$$

The student's final grade is 88.1.

- 3. Given the dataset {12, 15, 8, 21, 25, 14, 16, 18, 10}, divide it into four groups and find the sample variance and standard deviation.



*Solution:*

First, we need to put the data set into ascending order.

$$\{8, 10, 12, 14, 15, 16, 18, 21, 25\}$$

Now let's set up a table with four groups, find the midpoint of each group, and then the frequency of each group.

Group	Midpoint	Frequency
8 - 12	10	3
13 - 17	15	3
18 - 22	20	2
23 - 27	25	1

Then the estimate of the sample mean is

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

$$\bar{x} = \frac{(3)(10) + (3)(15) + (2)(20) + (1)(25)}{9}$$

$$\bar{x} = 15.55$$

We can use this mean to estimate the variance of the sample,

$$s^2 = \frac{\sum_{i=1}^n f_i (M_i - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{3(10 - 15.55)^2 + 3(15 - 15.55)^2 + 2(20 - 15.55)^2 + 1(25 - 15.55)^2}{9 - 1}$$

$$s^2 \approx 27.778$$

and then the standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{s^2}$$

$$s \approx \sqrt{27.778}$$

$$s \approx 5.270$$

- 4. A sample of book club members record the number of books they read last year. Calculate the mean number of books per member.

<b>Number of books</b>	0	1	2	3	4	5
<b>Number of people</b>	25	15	18	5	12	3

*Solution:*

We can calculate the weighted sample mean.

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

$$\bar{x} = \frac{25(0) + 15(1) + 18(2) + 5(3) + 12(4) + 3(5)}{25 + 15 + 18 + 5 + 12 + 3}$$

$$\bar{x} \approx 1.654$$

Each book club member read 1.654 books, on average, during the last year.

- 5. The frequency distribution represents the number of pizza orders a local pizza restaurant received each day over the last 20 days. Calculate the weighted sample mean, variance, and standard deviation.

Orders	Number of days
5 - 7	5
8 - 10	4
11 - 13	4
14 - 16	3
17 - 19	3
20 - 22	1

*Solution:*

First, we need to find the midpoint of each class.

Number of orders	Midpoint	Number of days
5 - 7	6	5
8 - 10	9	4
11 - 13	12	4
14 - 16	15	3
17 - 19	18	3
20 - 22	21	1

Then the estimate of the sample mean is

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

$$\bar{x} = \frac{5(6) + 4(9) + 4(12) + 3(15) + 3(18) + 1(21)}{5 + 4 + 4 + 3 + 3 + 1}$$

$$\bar{x} = 11.7$$

We can use this mean to estimate the variance of the sample,

$$s^2 = \frac{\sum_{i=1}^n f_i (M_i - \bar{x})^2}{n - 1}$$

$$\sum_{i=1}^n f_i (M_i - \bar{x})^2 = 5(6 - 11.7)^2 + 4(9 - 11.7)^2 + 4(12 - 11.7)^2$$

$$+ 3(15 - 11.7)^2 + 3(18 - 11.7)^2 + 1(21 - 11.7)^2$$

$$\sum_{i=1}^n f_i (M_i - \bar{x})^2 \approx 430.198$$

$$s^2 \approx \frac{430.198}{20 - 1}$$

$$s^2 \approx 22.642$$

and then the standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{s^2}$$

$$s \approx \sqrt{22.642}$$

$$s \approx 4.758$$



6. Use the sample data to find the mean, variance, and standard deviation of commute time.

Commute time	Number of people
1 - 5	1
6 - 10	4
11 - 15	6
16 - 20	3
21 - 25	10
26 - 30	13

*Solution:*

First, we need to find the midpoint of each group.

Commute time	Midpoint	Number of people
1 - 5	3	1
6 - 10	8	4
11 - 15	13	6
16 - 20	18	3
21 - 25	23	10
26 - 30	28	13

Then the estimate of the sample mean is



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

$$\bar{x} = \frac{1(3) + 4(8) + 6(13) + 3(18) + 10(23) + 13(28)}{1 + 4 + 6 + 3 + 10 + 13}$$

$$\bar{x} \approx 20.568$$

We can use this mean to estimate the variance of the sample,

$$s^2 = \frac{\sum_{i=1}^n f_i (M_i - \bar{x})^2}{n - 1}$$

$$\sum_{i=1}^n f_i (M_i - \bar{x})^2 = 1(3 - 20.578)^2 + 4(8 - 20.578)^2 + 6(13 - 20.578)^2$$

$$+ 3(18 - 20.578)^2 + 10(23 - 20.578)^2 + 13(28 - 20.578)^2$$

$$\sum_{i=1}^n f_i (M_i - \bar{x})^2 = 2,081.088$$

$$s^2 = \frac{2,081.088}{37 - 1}$$

$$s^2 \approx 57.808$$

and then the standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{s^2}$$

$$s \approx \sqrt{57.808}$$

$$s \approx 7.603$$

## SIMPLE PROBABILITY

- 1. A child drops a marble onto a board. Suppose that it is equally likely for it to fall anywhere on the board. What is the probability, to the nearest percent, that it lands on the red circle?



*Solution:*

We want to know the probability that the marble falls on the red area of the board. So we need to know

$$P(\text{red circle}) = \frac{\text{area of red circle}}{\text{area of full rectangle}}$$

This means we need to find the area of the circle,

$$A_{\text{circle}} = \pi r^2$$

$$A_{\text{circle}} = \pi(2.5)^2$$

$$A_{\text{circle}} \approx 19.63 \text{ cm}^2$$

and the rectangle.

$$A_{\text{rectangle}} = lw$$

$$A_{\text{rectangle}} = (15)(6)$$

$$A_{\text{rectangle}} = 90 \text{ cm}^2$$

So the probability that the marble lands on the red circle is

$$P(\text{red circle}) = \frac{19.63 \text{ cm}^2}{90 \text{ cm}^2} \approx 0.22$$

There's a 22% chance the marble lands on the blue circle.

- 2. A 12-sided number cube is rolled 60 times. Use the table to calculate  $P(\text{rolling an 11})$ . Is this theoretical or experimental probability? Why?

Number rolled	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	5	8	2	0	10	1	6	5	2	8	12	1

*Solution:*

This is an experimental probability because it's based on the results of actual trials. From the table, we can see that we rolled an 11 on the dice 12 times out of the 60 total rolls.



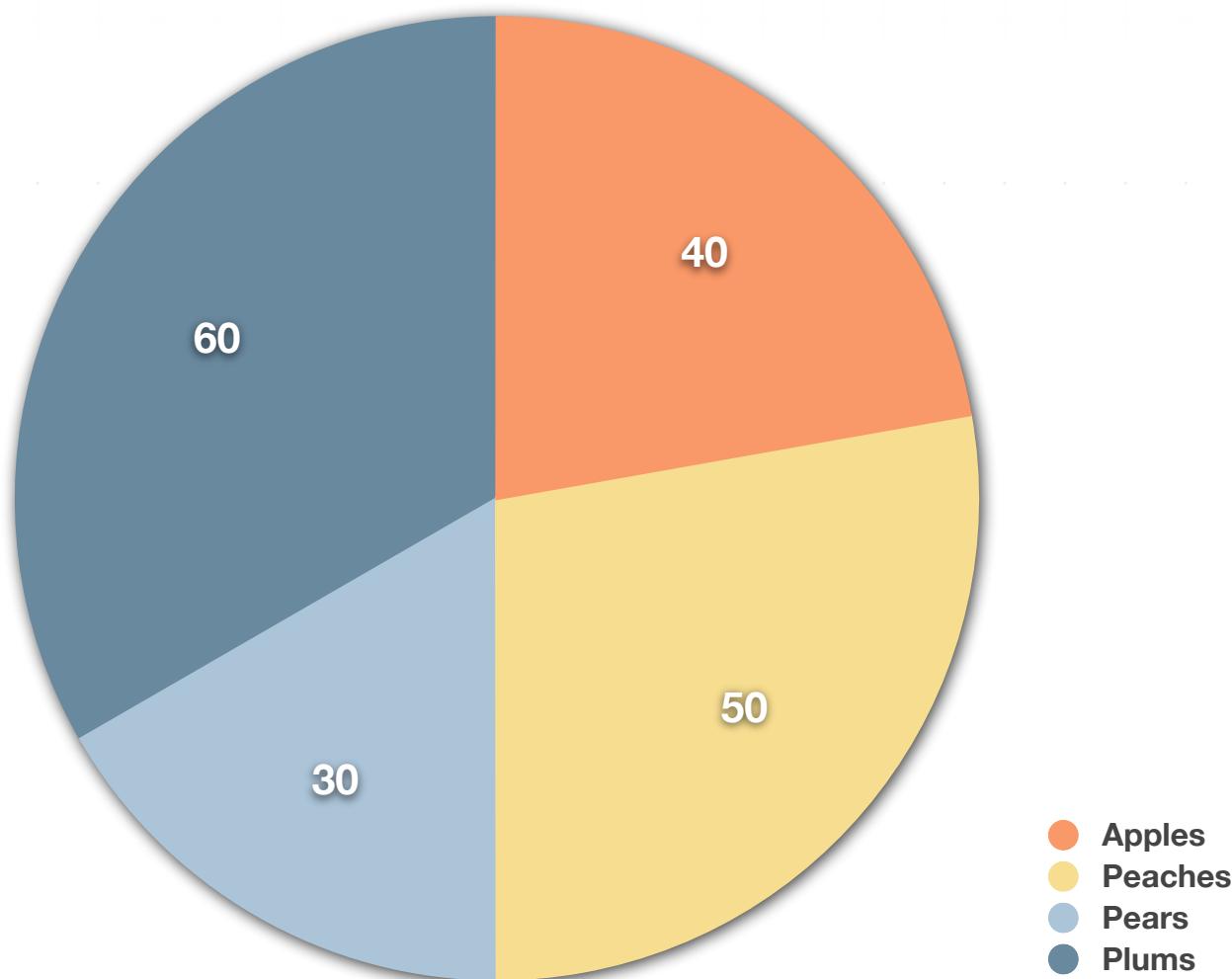
Number rolled	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	5	8	2	0	10	1	6	5	2	8	12	1

So  $P(\text{rolling an } 11)$  is

$$P(\text{rolling an } 11) = \frac{12}{60} = \frac{1}{5} = 0.2 = 20\%$$

3. Monica's class went on a trip to an orchard. At the end of the trip they put all of the fruit they picked into one big basket. The chance of picking any fruit from the basket is equally likely. Monica's teacher picks out a fruit for her to eat at random. What is the probability that it's a plum (Monica's favorite)? Is this an experimental or theoretical probability? Why?

Number of fruit picked from each tree



*Solution:*

This is a theoretical probability because it was calculated based on the knowledge of the sample space. Monica didn't perform repeated trials, so there was no experiment.

In this case, the outcomes that meet our criteria are the 60 plums. All possible outcomes can be found by adding all of the types of fruit together.

$$60 + 40 + 30 + 50 = 180$$

Therefore, the probability of getting a plum is

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

$$P(\text{plum}) = \frac{60}{180} = \frac{1}{3}$$

Monica has a  $1/3 \approx 33\%$  chance of getting a plum.

- 4. Jamal surveyed the people at his local park about their favorite hobby and recorded his results in a table. Based on the survey, what's the probability that someone who visits the park will choose Art as their favorite hobby? Is this a theoretical or experimental probability? Why?



Hobby	Count
Reading	14
Sports	28
Art	15
<b>Total</b>	<b>57</b>

*Solution:*

Jamal is not likely to have surveyed everyone who visits the park or everyone who will visit the park in the future. A survey is most often a sample of a larger population, so the results are an experimental probability.

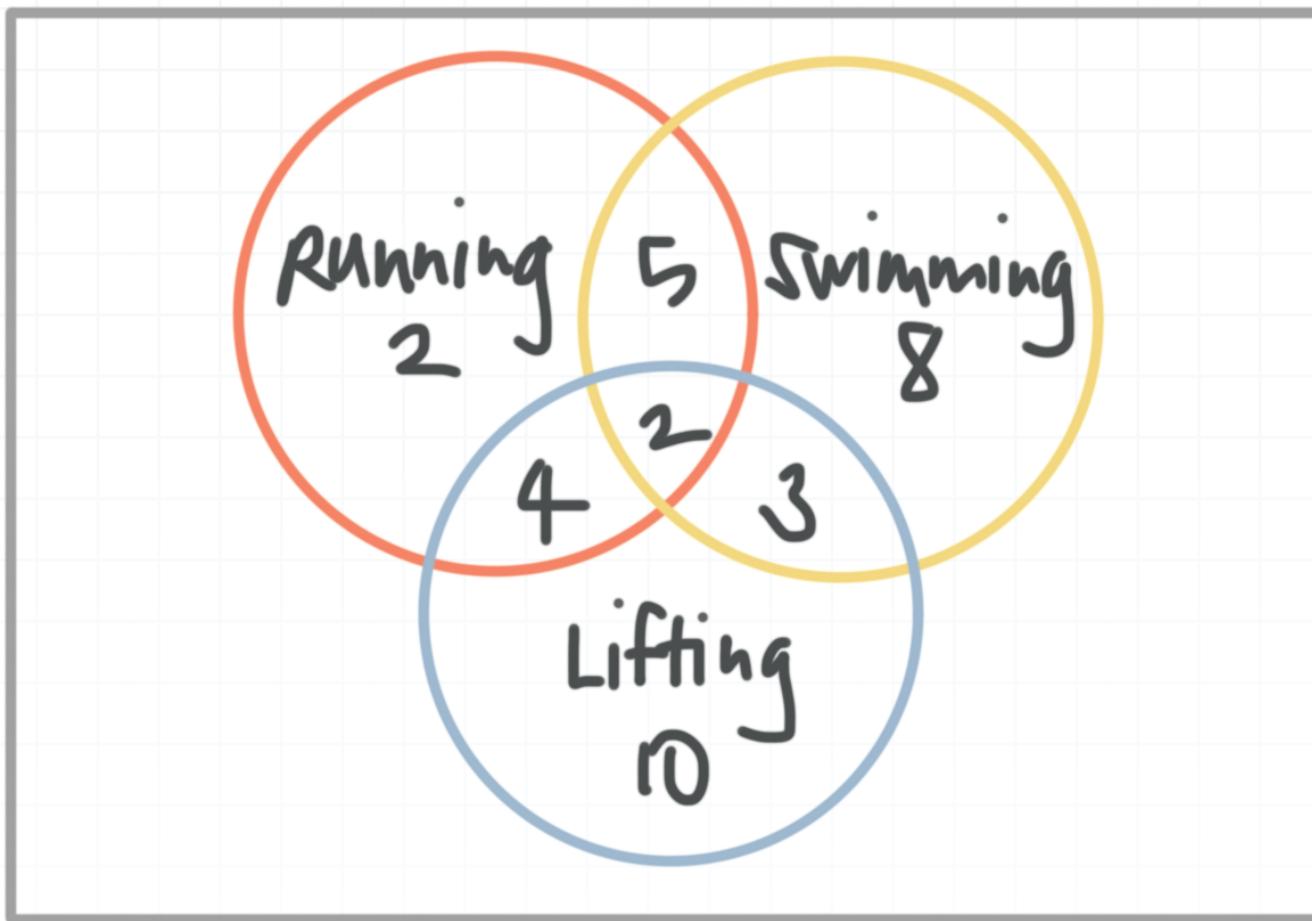
In this case, the outcomes that meet our criteria are the 15 people who selected Art as their favorite hobby. The total possible outcomes are the number of people surveyed, 57. Therefore, the probability that someone in Jamal's survey chooses Art is

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

$$P(\text{Art}) = \frac{15}{57} = \frac{5}{19}$$

- 5. What is the probability that someone's favorite exercise was weight lifting only?





*Solution:*

In this case, the outcomes that meet our criteria are the 10 people whose favorite exercise was weight lifting. The total of all possible outcomes are the total number of people included in the Venn diagram:

$$2 + 5 + 8 + 4 + 2 + 3 + 10 = 34$$

So the probability that someone in the survey chose weight lifting as their favorite exercise is

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

$$P(\text{weight lifting}) = \frac{10}{34} = \frac{5}{17}$$

6. What is the sample space for rolling two six-sided dice (the list of all possible outcomes)? What's the probability that the sum of the two dice is an odd number? Is this a theoretical or experimental probability? Why?

*Solution:*

We're asked to list the sample space for rolling two six-sided dice. This means we want to make a list of all the possible ways we could roll the dice (the total outcomes).

A nice way to make sure we include every combination is to make a table. We can represent one die by the top row and one die by the far-left column and then write down all of the combinations to find the sample space.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

The rolls that give an odd sum are

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

There are 36 total rolls in the sample space, and 18 that give an odd sum, so the probability of rolling an odd sum is

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

$$P(\text{odd sum}) = \frac{18}{36} = \frac{1}{2} = 0.5 = 50\%$$

This is an example of theoretical probability because we used the probability formula and did not perform an experiment.

## THE ADDITION RULE, AND UNION VS. INTERSECTION

- 1. Given the probabilities  $P(A) = 0.3$ ,  $P(B) = 0.6$  and  $P(A \cap B) = 0.05$ , what is  $P(A \cup B)$ ? Are  $A$  and  $B$  mutually exclusive events? Why or why not?

*Solution:*

Events  $A$  and  $B$  are not mutually exclusive events because sometimes they can happen at the same time. The problem even tells us that  $P(A \cap B) = 0.05$ , which means there's a 5% chance that both events happen at the same time. To find  $P(A \cup B)$ , we'll use

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

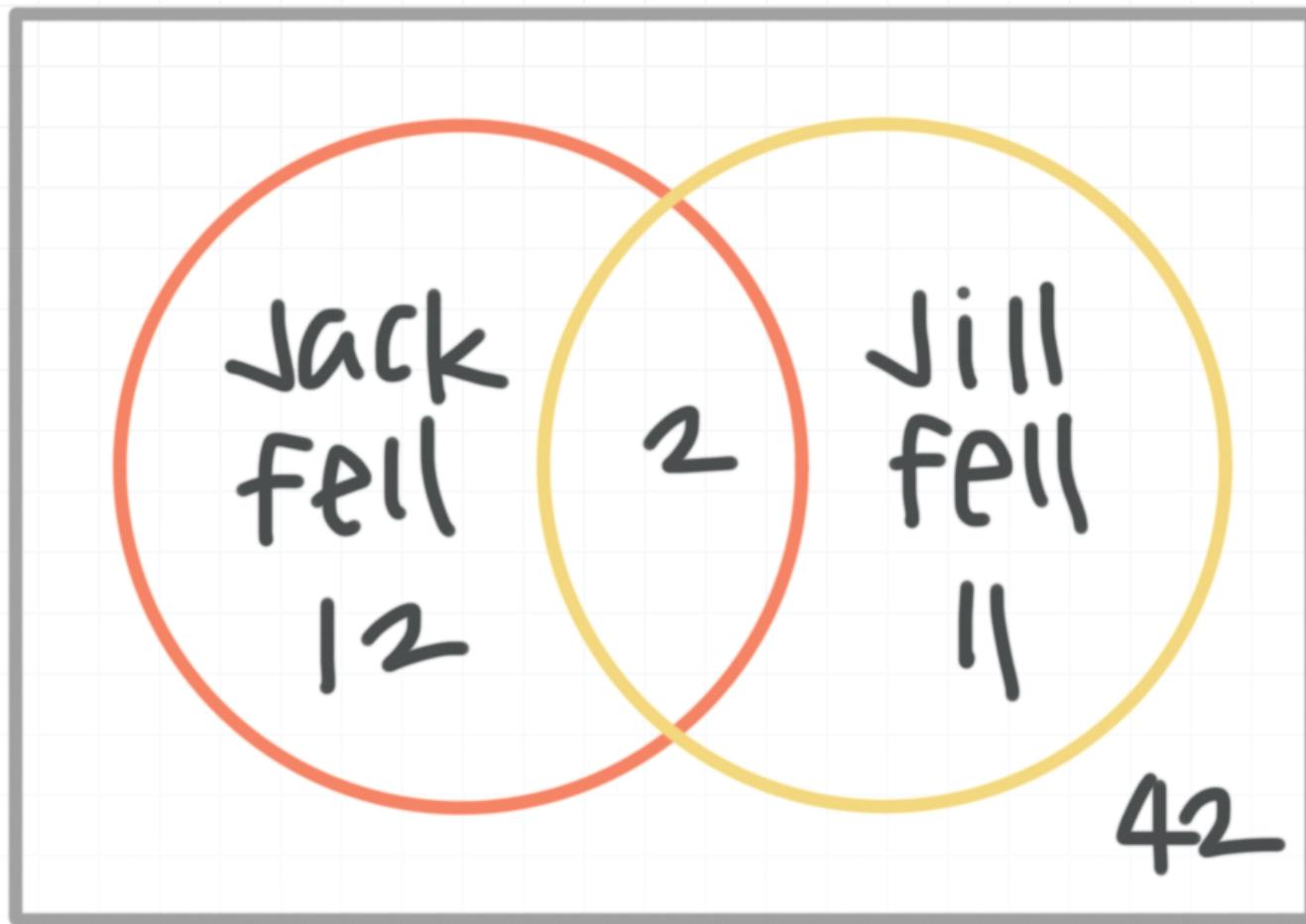
and plug in  $P(A) = 0.3$ ,  $P(B) = 0.6$ , and  $P(A \cap B) = 0.05$ .

$$P(A \cup B) = 0.3 + 0.6 - 0.05$$

$$P(A \cup B) = 0.85$$

- 2. Jack and Jill are taking multiple trips up a hill together. The Venn diagram shows the number of times Jack and Jill fell down on their various trips up the hill. What is the probability that Jack and Jill both fell down on any particular trip, and what is the probability that only Jack fell down or only Jill fell down on any particular trip?





*Solution:*

From the Venn diagram, we can add the numbers from each of the four sections to see that Jack and Jill made

$$12 + 2 + 11 + 42 = 67$$

trips up the hill together. From the 2 in the center of the Venn diagram where the circles overlap, we can tell that Jack and Jill both fell down on 2 of the trips up the hill. So the probability that Jack fell down and Jill fell down is

$$P(\text{Jack fell down} \cap \text{Jill fell down}) = \frac{2}{67}$$

From the Venn diagram, we know that they took 12 trips where only Jack fell down, and 11 trips where only Jill fell down. So the probability that either only Jack fell down or only Jill fell down is

$$P(\text{Jack fell down} \cup \text{Jill fell down}) = \frac{12}{67} + \frac{11}{67}$$

$$P(\text{Jack fell down} \cup \text{Jill fell down}) = \frac{23}{67}$$

- 3. When people buy a fish at a pet store the cashier can check off the color of the fish as mostly red, mostly orange or mostly yellow. Currently the probability of buying a red fish is 0.31, the probability of buying an orange fish is 0.23, and the probability of buying a mostly yellow fish is 0.13 (there are colors of fish other than red, orange, and yellow).

Are the events buying a mostly red fish and buying a mostly orange fish mutually exclusive? Find the probability that the purchase of a randomly selected fish is either mostly red or mostly orange.

*Solution:*

The events of buying a mostly red fish and buying a mostly orange fish are mutually exclusive because a single fish must be either mostly red or mostly orange. It can't be both, so there's no overlap in the two events.

The probability that the purchase of a randomly selected fish is either mostly red or mostly orange is

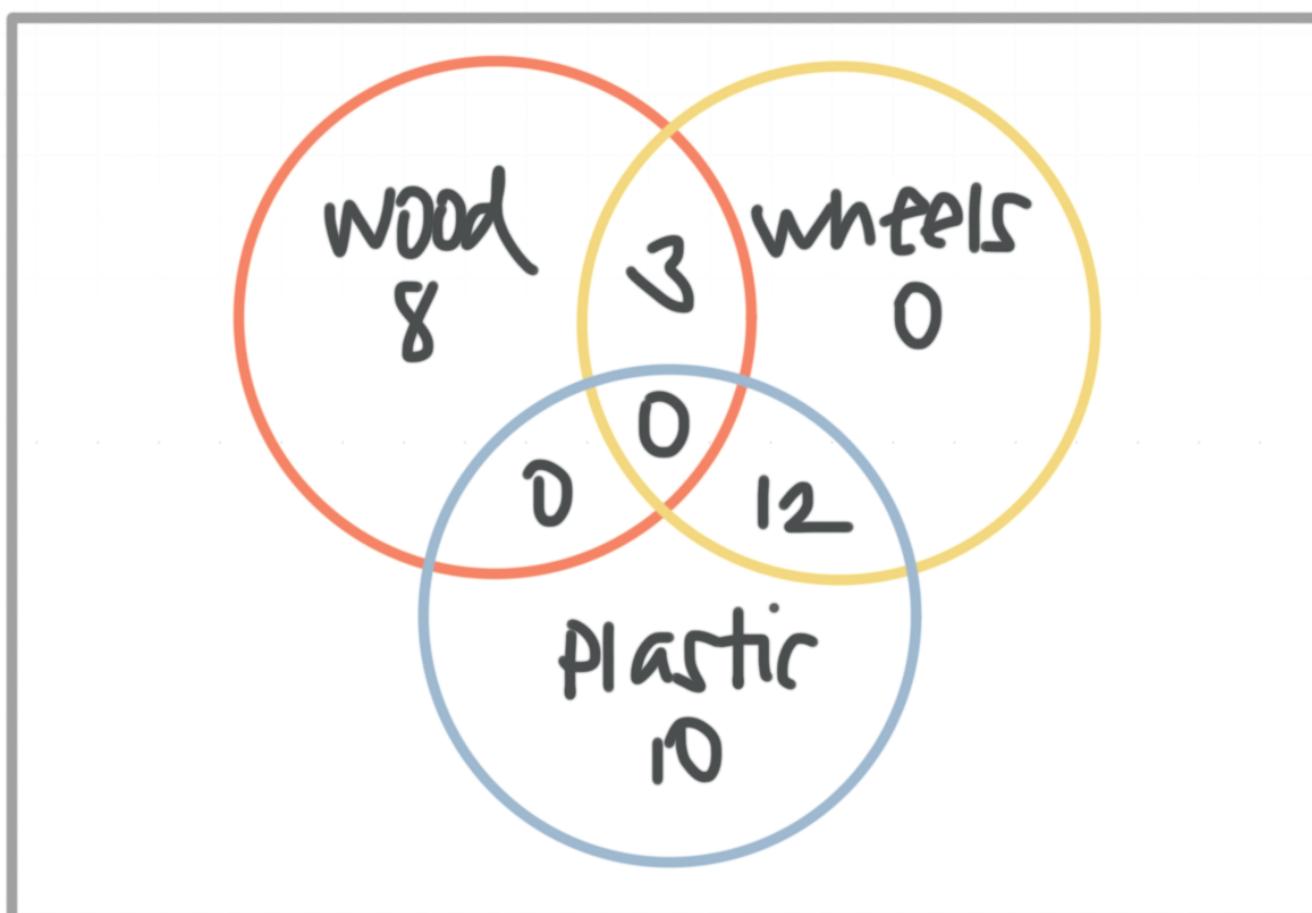


$$P(\text{mostly red} \cup \text{mostly orange}) = P(\text{mostly red}) + P(\text{mostly orange})$$

$$P(\text{mostly red} \cup \text{mostly orange}) = 0.31 + 0.23$$

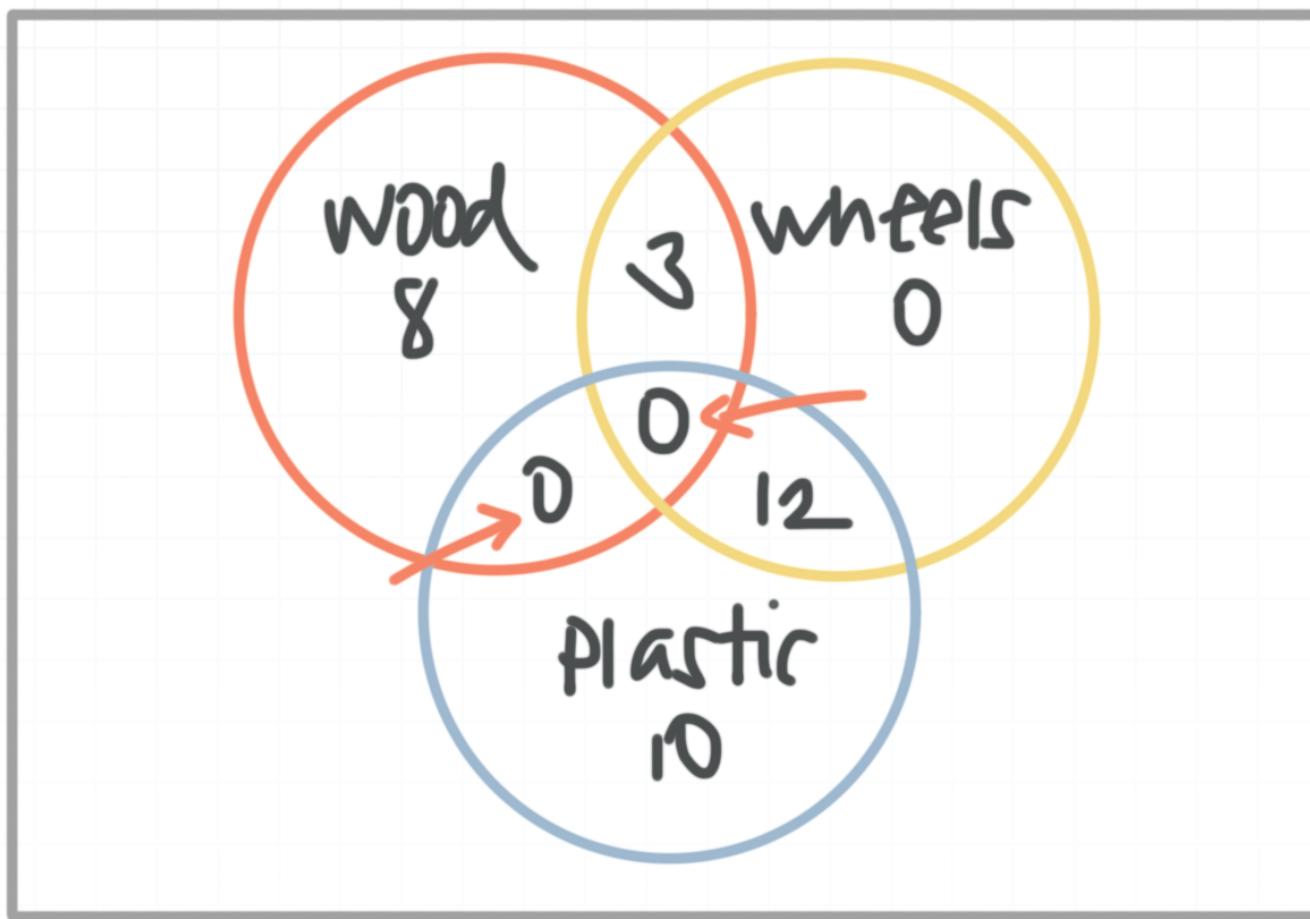
$$P(\text{mostly red} \cup \text{mostly orange}) = 0.54$$

- 4. The Venn diagram shows Mason's toy car collection. Are the events "plastic" and "wood" mutually exclusive? What is the probability that a vehicle is made from plastic or wood? Are the events "wood" and "wheels" mutually exclusive? What is the probability that a vehicle is made from wood and has wheels?



*Solution:*

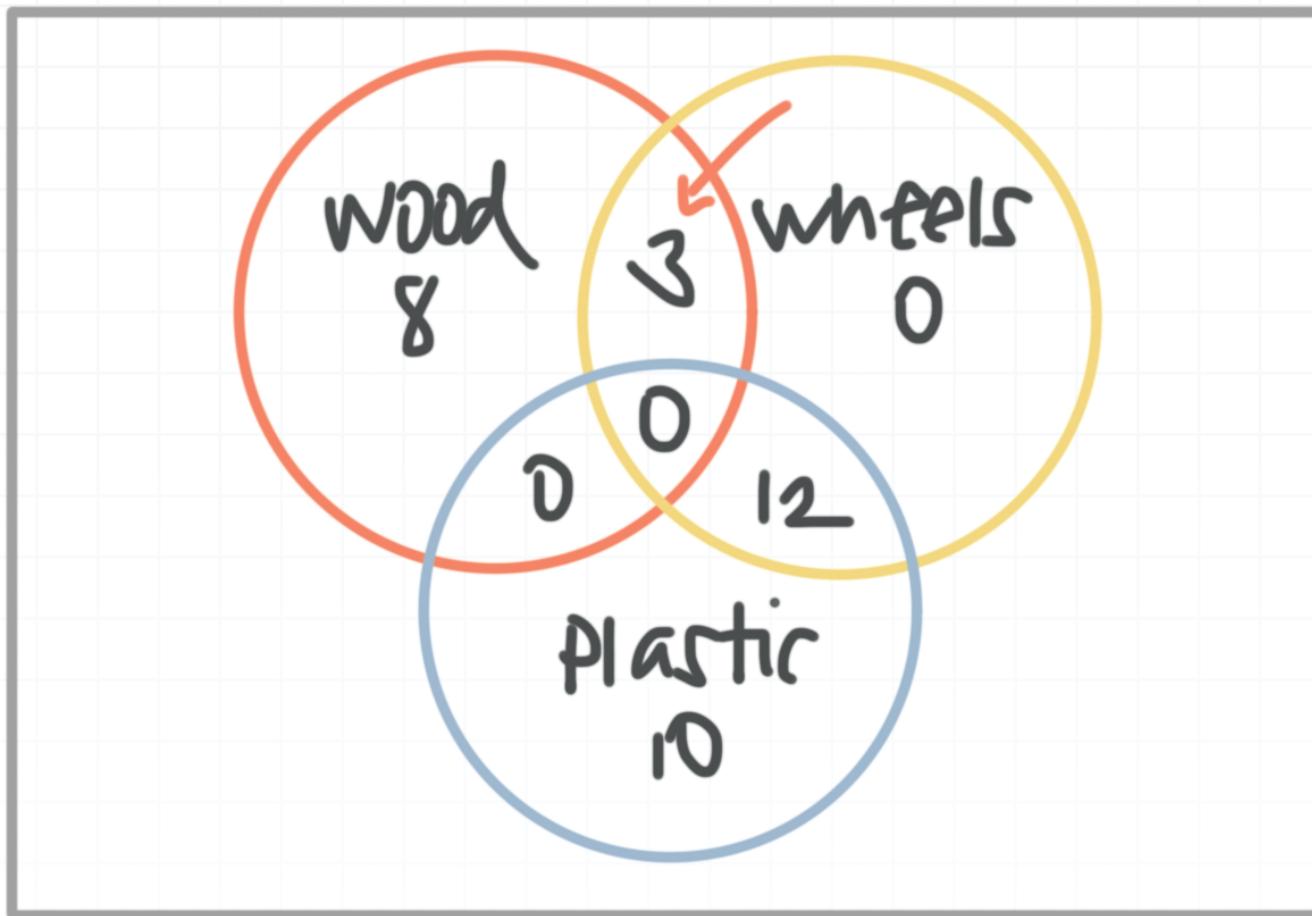
The events “plastic” and “wood” are mutually exclusive, because the intersection between them is 0.



The probability that a vehicle is made from plastic or wood is represented by  $P(\text{plastic} \cup \text{wood})$ . There are  $12 + 10 + 8 + 3 = 33$  total cars in the Venn diagram, and  $8 + 3 = 11$  of them are made with wood, while  $10 + 12 = 22$  of them with plastic. Which means that the probability that a car is made with wood or plastic is

$$P(\text{plastic} \cup \text{wood}) = \frac{11 + 22}{33} = \frac{33}{33} = 1 = 100\%$$

The events “wood” and “wheels” are not mutually exclusive because they have a non-zero number in their intersection.



The probability that a vehicle is made from wood and has wheels is represented by  $P(\text{wood} \cap \text{wheels})$ . Of all the vehicles in the Venn diagram, 3 are made from wood and have wheels, so

$$P(\text{wood} \cap \text{wheels}) = \frac{3}{33} = \frac{1}{11} \approx 9\%$$

- 5. Every student at a certain high school needs to choose exactly one fine arts elective. The frequency table shows the enrollment of electives for all students. Are the events “junior” and “architecture” mutually exclusive? What is the probability that a student is taking architecture and a junior? What is the probability that a student is a junior or is taking architecture?

		Extracurricular activities			
		Art	Architecture	Music	Total
Grade	Freshmen	40	25	55	120
	Sophomore	52	12	71	135
	Junior	56	45	54	155
	Senior	30	60	20	110
	Total	178	142	200	520

*Solution:*

The events “junior” and “architecture” are not mutually exclusive events because it’s possible for a student to be both a junior and enrolled in architecture.

The probability that a student is a junior and is taking architecture is given by

$$P(\text{junior} \cap \text{architecture}) = \frac{45}{520} = \frac{9}{104}$$



		Extracurricular activities			
		Art	Architecture	Music	Total
Grade	Freshmen	40	25	55	120
	Sophomore	52	12	71	135
	Junior	56	45	54	155
	Senior	30	60	20	110
	Total	178	142	200	520

The probability that a student is a junior or is taking architecture is given by

$$P(\text{junior} \cup \text{architecture}) = P(\text{junior}) + P(\text{architecture})$$

$$-P(\text{junior} \cap \text{architecture})$$

$$P(\text{junior} \cup \text{architecture}) = \frac{155}{520} + \frac{142}{520} - \frac{45}{520}$$

$$P(\text{junior} \cup \text{architecture}) = \frac{252}{520} = \frac{63}{130}$$

		Extracurricular activities			
		Art	Architecture	Music	Total
Grade	Freshmen	40	25	55	120
	Sophomore	52	12	71	135
	Junior	56	45	54	155
	Senior	30	60	20	110
	Total	178	142	200	520

These are not mutually exclusive events, which is why we need to subtract the overlap.

- 6. James tosses a coin and rolls a six-sided die. What is the sample space for this situation? What is the probability the coin lands on heads and the die lands on a 2 or a 3?

*Solution:*

We're asked to list the sample space for flipping a coin and rolling a six-sided die. This means we want to make a list of all the possible ways we could flip the coin and roll the die (the total outcomes). A nice way to make sure we include every combination is to make a table. We can represent one die in the top row and the coin in the far-left column. Then we can write down all of the combinations to find the sample space, in a similar way that we would make a multiplication table.

	1	2	3	4	5	6
Heads	Heads, 1	Heads, 2	Heads, 3	Heads, 4	Heads, 5	Heads, 6
Tails	Tails, 1	Tails, 2	Tails, 3	Tails, 4	Tails, 5	Tails, 6

Next, we're interested in the probability that the coin lands on heads and the die lands on a 2 or a 3. This means we need to find  $P(\text{heads} \cap 2 \text{ or } 3)$ . There are only two values from the sample space that give heads and a 2 or a 3.



	1	2	3	4	5	6
Heads	Heads, 1	Heads, 2	Heads, 3	Heads, 4	Heads, 5	Heads, 6
Tails	Tails, 1	Tails, 2	Tails, 3	Tails, 4	Tails, 5	Tails, 6

And there are 12 possible outcomes. So the probability is

$$P(\text{heads} \cap 2 \text{ or } 3) = \frac{2}{12} = \frac{1}{6}$$

## INDEPENDENT AND DEPENDENT EVENTS AND CONDITIONAL PROBABILITY

- 1. What is the probability of getting four heads in a row when we flip a fair coin four times?

*Solution:*

Each coin flip is an independent event. The probability of getting a head on each flip is  $1/2$  (there's one way to get a head out of two possible ways, heads or tails). Therefore,

$$P(HHHH) = \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

- 2. An old dog finds and eats 60% of food that's dropped on the floor. A toddler wanders through the house and drops 10 pieces of cereal. What's the probability the dog finds and eats all 10 pieces?

*Solution:*

The dog's success rate of finding dropped food is 60%. We can calculate the probability the dog finds all the pieces by saying that the dog finding the next piece of food is independent from finding the piece before. Then,



$$P(FFFFFFFFF) = (0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)$$

$$P(FFFFFFFFF) = (0.6)^{10}$$

$$P(FFFFFFFFF) \approx 0.006$$

There's a 0.6% chance the dog will find and eat all of the dropped cereal.

- 3. Amelia is choosing some pretty stones from the gift shop at the museum. The gift shop has a grab bag that contains 5 amethyst stones, 6 fluorite stones, 2 pink opals, and 7 yellow calcite stones. Amelia looks into the bag and takes out two stones, one at a time, at random. What is the probability that she gets an amethyst first and then a pink opal?

*Solution:*

There are a total of  $5 + 6 + 2 + 7 = 20$  stones. If Amelia pulls one stone from the grab bag, the probability of taking out an amethyst is

$$P(\text{amethyst}) = \frac{5}{20} = \frac{1}{4}$$

Once an amethyst is pulled out, there are only 19 stones left in the bag, 2 of which are pink opals, so the chance of pulling a pink opal is

$$P(\text{pink opal} | \text{amethyst}) = \frac{2}{19}$$



We can therefore say that the probability of pulling both stones in that specific order (these are dependent events) is

$$P(\text{amethyst then pink opal}) = \frac{1}{4} \cdot \frac{2}{19} = \frac{2}{76} = \frac{1}{38}$$

- 4. Emily counted the shape and type of blocks that her little sister owns and organized the information into a frequency table.

		Block Shape		
		Cube	Rectangular Prism	Total
Block Color	Red	5	9	14
	Blue	4	10	14
	Total	9	19	28

Are events  $A$  and  $B$  dependent or independent events? Use the formula to explain the answer.

Event  $A$  is that the block is a cube.

Event  $B$  is that block is red.

Let  $P(A)$  be the probability that a block drawn at random is a cube.

Let  $P(B)$  be the probability that a block drawn at random is red.

*Solution:*



The events are independent if we can show that  $P(A \text{ and } B) = P(A)P(B)$ .  $P(A)$  is the probability that a block drawn at random is a cube.  $P(A) = 9/28$ .

		Block Shape		Total
Block Color		Cube	Rectangular Prism	
	Red	5	9	14
	Blue	4	10	14
	Total	9	19	28

$P(B)$  is the probability that a block drawn at random is red.

$$P(B) = 14/28 = 1/2.$$

		Block Shape		Total
Block Color		Cube	Rectangular Prism	
	Red	5	9	14
	Blue	4	10	14
	Total	9	19	28

$P(A \text{ and } B)$  is the probability that the chosen block is both red and a cube.

$$P(A \text{ and } B) = 5/28.$$

		Block Shape		
		Cube	Rectangular Prism	Total
Block Color	Red	5	9	14
	Blue	4	10	14
	Total	9	19	28

Now we can check for independence by showing  $P(A \text{ and } B) = P(A)P(B)$ .

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

$$\frac{5}{28} = \frac{9}{28} \cdot \frac{1}{2}$$

$$\frac{5}{28} = \frac{9}{56}$$

$$\frac{10}{56} = \frac{9}{56}$$

Because the values are unequal,  $P(A)$  and  $P(B)$  are dependent events.

- 5. A bag has 4 cinnamon candies, 6 peppermint candies, and 12 cherry candies. Sasha draws 3 candies at random from the bag one at a time without replacement. Does the situation describe dependent or independent events? What is the probability of drawing a cinnamon first, then a cherry, and then a peppermint?

*Solution:*

These events are dependent events, because removing a candy from the bag changes what's inside and effects the probability of subsequent pulls.

We want to find the probability of drawing a cinnamon first, then a cherry, and then a peppermint last. There are  $4 + 6 + 12 = 22$  total candies in the bag. Let's look at the probability of getting a cinnamon first. Since there are 4 cinnamon candies, the probability of getting a cinnamon is

$$P(\text{cinnamon}) = \frac{4}{22} = \frac{2}{11}$$

Now there are 21 total candies remaining, 12 of which are cherry, so the probability of getting cherry next is

$$P(\text{cherry}) = \frac{12}{21} = \frac{4}{7}$$

Now there are 20 total candies remaining, 6 of which are peppermint, so the probability of getting peppermint next is

$$P(\text{peppermint}) = \frac{6}{20} = \frac{3}{10}$$

Therefore, the probability of drawing these three flavors in this particular order is

$$P(\text{Ci, Ch, Pe}) = \frac{2}{11} \cdot \frac{4}{7} \cdot \frac{3}{10}$$

$$P(\text{Ci, Ch, Pe}) = \frac{24}{770}$$



$$P(\text{Ci, Ch, Pe}) = \frac{12}{385}$$

- 6. Nyla has 12 stuffed animals, 7 of which are elephants (4 of the elephants play music and light up) and 5 of which are bears (2 of the bears play music and light up). Her mother randomly selects an animal to bring with them on vacation. Let  $A$  be the event that she selects an elephant and  $B$  be the event that she selects an animal that plays music and lights up.

Find  $P(A)$ ,  $P(B)$ ,  $P(A | B)$ , and  $P(B | A)$ . State if events  $A$  and  $B$  are dependent or independent events, then find  $P(A \text{ and } B)$ .

*Solution:*

There are  $7 + 5 = 12$  total stuffed animals.  $P(A)$  is the probability of selecting an elephant, and there are 7 elephants.

$$P(A) = \frac{7}{12}$$

$P(B)$  is the probability of selecting an animal that plays music and lights up. There are  $4 + 2 = 6$  animals that play music and light up.

$$P(B) = \frac{6}{12} = \frac{1}{2}$$

$P(A | B)$  is the probability of selecting an elephant, given that the animal plays music and lights up. There are 4 elephants that play music and light up out of  $4 + 2 = 6$  total animals that play music and light up.

$$P(A | B) = \frac{4}{6} = \frac{2}{3}$$

$P(B | A)$  is the probability of picking a toy that plays music and lights up given that the toy is an elephant. There are 4 elephants that play music and light up out of 7 total elephants.

$$P(B | A) = \frac{4}{7}$$

Because  $P(A) \neq P(A | B)$  and  $P(B) \neq P(B | A)$ ,  $A$  and  $B$  are dependent events.

$P(A \text{ and } B)$  is the probability of choosing an elephant that plays music and lights up. We know the events are dependent events, so

$$P(A \text{ and } B) = P(A) \cdot P(B | A)$$

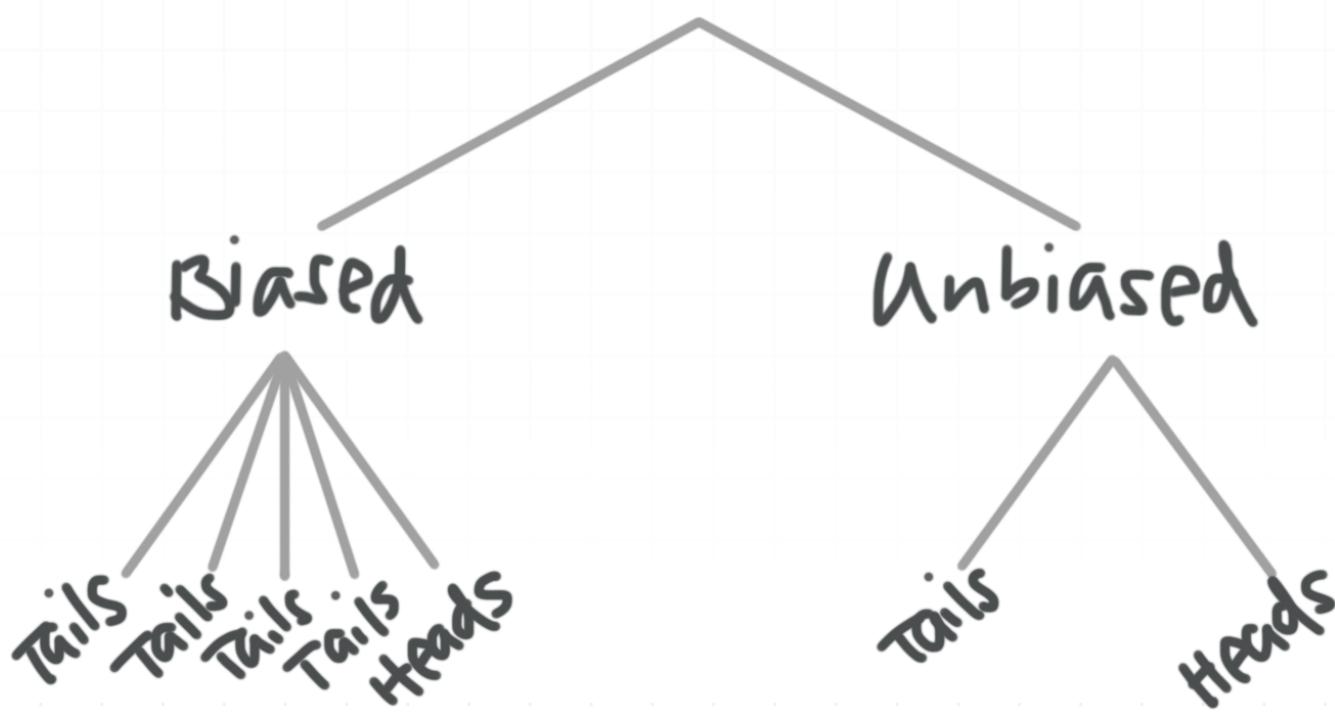
$$P(A \text{ and } B) = \frac{7}{12} \cdot \frac{4}{7}$$

$$P(A \text{ and } B) = \frac{28}{84} = \frac{4}{12}$$

$$P(A \text{ and } B) = \frac{1}{3}$$

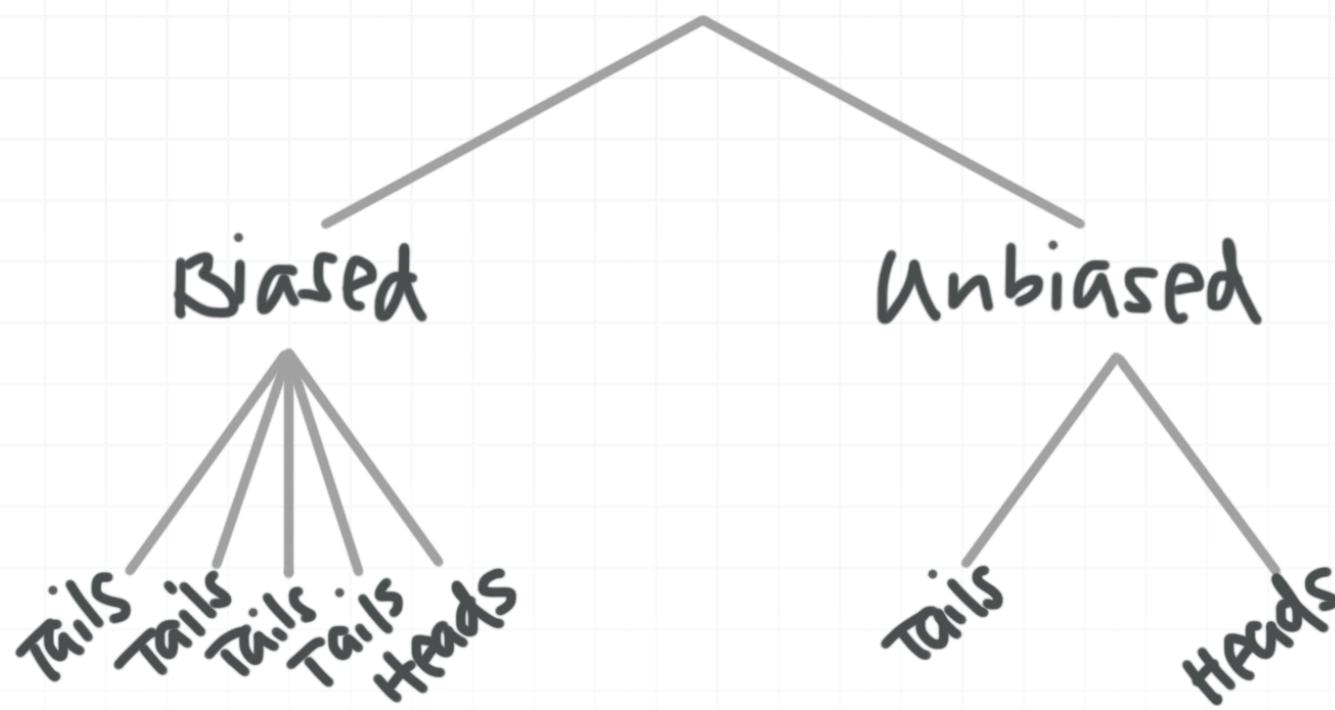
## BAYES' THEOREM

- 1. We have two coins. One is fair and the other one is weighted to land on tails  $\frac{4}{5}$  of the time. Without knowing which coin we're choosing, we pick one at random, toss the coin and get tails. What is the probability we flipped the biased coin? Complete the tree diagram to answer the question.

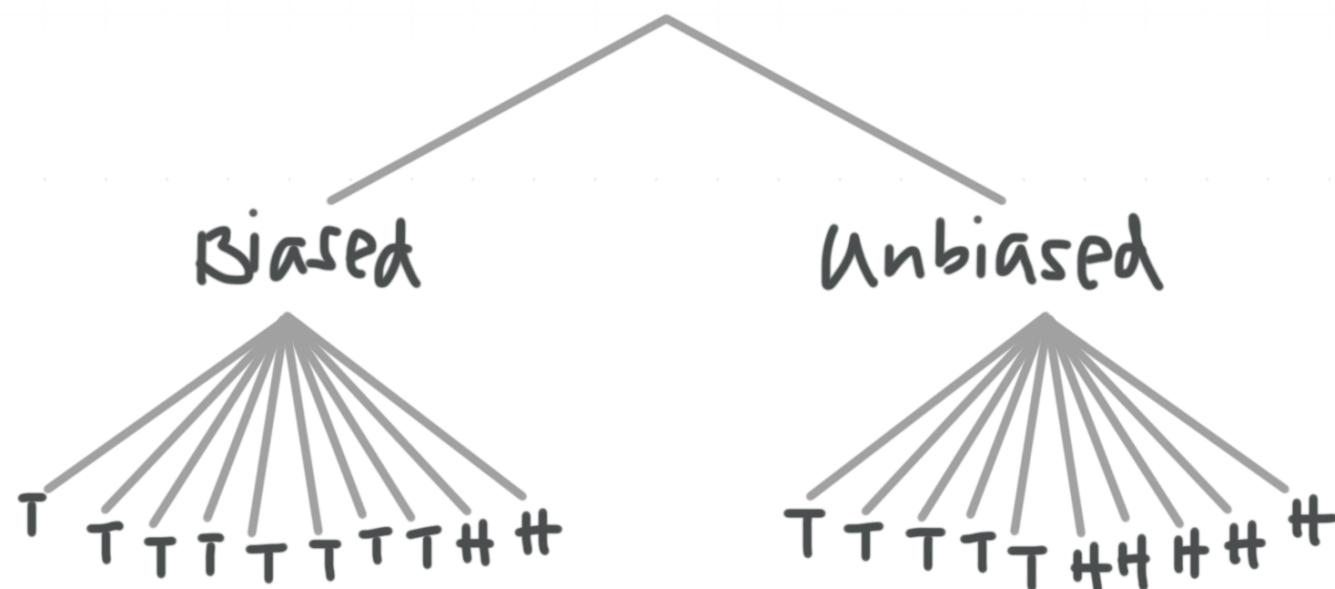


*Solution:*

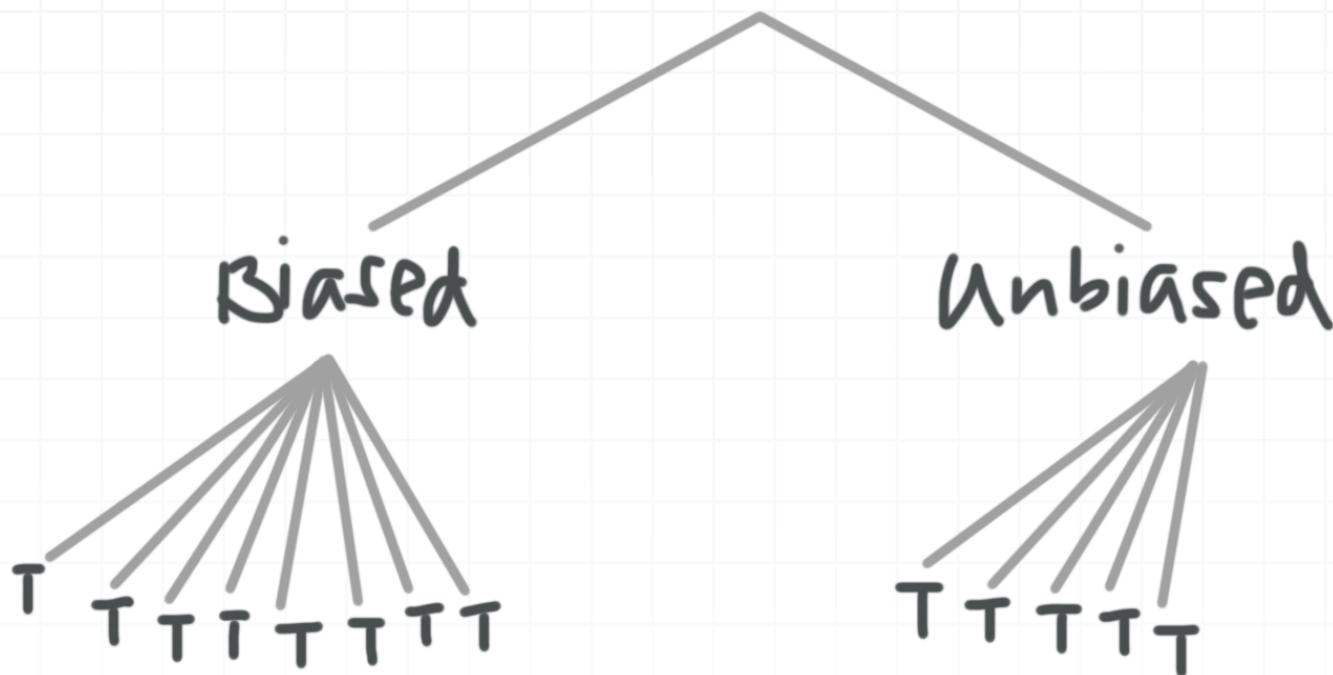
We're looking for the probability that the coin is biased given that we already flipped a tails, so we're looking for  $P(\text{biased} | \text{tails})$ .



The next step for the tree diagram is to make sure the branches are balanced. We use equivalent fractions to do this. For the biased side we know that we get tails 4 out of 5 times. This is the same as 8 out of 10 times. For the unbiased coin, we get tails 1 out of 2 times, which is the same as 5 out of 10 times.



We're only interested in tails, so now we need to trim the tree.



Now we're looking for the probability that we tossed the biased coin. 8 of the tails came from the biased coin and 5 did not.

$$P(\text{biased}) = \frac{8}{8+5} = \frac{8}{13}$$

The probability we tossed the biased coin, knowing that it landed on tails, is 8/13.

- 2. We have two dice. One is fair and the other is biased. The biased die is weighted to land on 6 every 1 out of 36 rolls. There's an equal probability for all of the other five faces on the biased die. Without knowing which one we're choosing, we pick one of the dice, roll it, and get a 6.

Calculate the following and use them to answer the question: What is the probability that we rolled the fair die?

$$P(6 | \text{fair})$$

$$P(\text{fair})$$

$P(6)$

*Solution:*

$P(6 \mid \text{fair})$  is the probability of rolling a 6, given that the die was fair. Since all outcomes are equally likely on the fair die, we have a 1 in 6 chance of rolling a 6.

$$P(6 \mid \text{fair}) = \frac{1}{6}$$

$P(\text{fair})$  is the probability of choosing the fair die. Each of the 2 dice has an equally likely chance of being chosen, so the probability of choosing the fair die is 1 in 2.

$$P(\text{fair}) = \frac{1}{2}$$

$P(6)$  is the probability of rolling a 6. This is the probability of choosing the biased die and rolling a 6 or the probability of choosing the fair die and rolling a 6. Let's find the probability that the die is fair and we roll a 6.

$$P(\text{fair and } 6) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

Now let's find the probability the die is biased and we roll a 6.

$$P(\text{biased and } 6) = \frac{1}{2} \cdot \frac{1}{36} = \frac{1}{72}$$

Therefore, the probability of rolling a 6 is



$$P(6) = \frac{1}{12} + \frac{1}{72}$$

$$P(6) = \frac{6}{72} + \frac{1}{72}$$

$$P(6) = \frac{7}{72}$$

Now we want to answer the question: “What is the probability that we rolled the fair die?” We’re looking for  $P(\text{fair} | 6)$ , and we have everything we need to use Bayes’ Theorem.

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

$$P(\text{fair} | 6) = \frac{P(6 | \text{fair}) \cdot P(\text{fair})}{P(6)}$$

$$P(\text{fair} | 6) = \frac{\frac{1}{6} \cdot \frac{1}{2}}{\frac{7}{72}}$$

$$P(\text{fair} | 6) = \frac{\frac{1}{12}}{\frac{7}{72}} = \frac{1}{12} \cdot \frac{72}{7} = \frac{72}{84} = \frac{6}{7}$$

The probability we rolled the fair die given that we rolled a 6 is  $6/7$ .

■ 3. Charlie knows that, at his school,

$$P(\text{senior}) = 0.40$$

$$P(\text{playing soccer}) = 0.15$$

$$P(\text{soccer and senior}) = 0.05$$

Solve for the probability  $P(\text{senior} \mid \text{soccer})$ , then state whether or not Bayes' Theorem can be used to solve the problem.

*Solution:*

Let's look to see if we can use Bayes' Theorem to find the probability. First let's take Bayes' Theorem and write it in terms of our problem. We want to solve for the probability  $P(\text{senior} \mid \text{soccer})$ , so

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

$$P(\text{senior} \mid \text{soccer}) = \frac{P(\text{soccer} \mid \text{senior}) \cdot P(\text{senior})}{P(\text{soccer})}$$

Remember that the multiplication rule says that  $P(B \text{ and } A) = P(B \mid A) \cdot P(A)$ . So we can also say that  $P(\text{soccer and senior}) = P(\text{soccer} \mid \text{senior}) \cdot P(\text{senior})$ . Then we can use Bayes' Theorem.

$$P(\text{senior} \mid \text{soccer}) = \frac{P(\text{soccer and senior})}{P(\text{soccer})}$$

Now we can use the information we've been given to solve the problem.

$$P(\text{soccer and senior}) = 0.05$$

$$P(\text{playing soccer}) = 0.15$$



$$P(\text{senior} \mid \text{soccer}) = \frac{0.05}{0.15} = \frac{1}{3} \approx 33\%$$

We could have also used a Venn diagram, instead of Bayes' Theorem, to solve this problem.

- 4. We have two coins. One is fair and the other is weighted to land on tails  $\frac{3}{4}$  of the time. Without knowing which coin we're choosing, we pick one at random, toss the coin, and get tails. What's the probability we flipped the biased coin?

*Solution:*

We're looking for the probability that the coin is biased, given that we already flipped a tails, so we're looking for  $P(\text{biased} \mid \text{tails})$ . We can solve this problem using Bayes' Theorem, or by creating a tree diagram. Let's use Bayes' Theorem.

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

That means that to use Bayes' Theorem, we have  $P(A) = P(\text{biased})$  and  $P(B) = P(\text{tails})$ . Then we need to find these values to plug into the formula:

$$P(\text{tails} \mid \text{biased})$$

$$P(\text{biased})$$

$$P(\text{tails})$$



We know from the problem that  $P(\text{tails} \mid \text{biased}) = 3/4$ . There are two coins, and it's equally likely that we choose either one, so  $P(\text{biased}) = 1/2$ . The probability of flipping a tails is the probability of flipping the biased coin and landing on tails or the probability of flipping the unbiased coin and landing on tails. Let's find the probability the coin is biased and it lands on tails.

$$P(\text{biased and tails}) = \frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$$

Now let's find the probability the coin is fair and lands on tails.

$$P(\text{fair and tails}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

So the probability of flipping a tails is

$$P(\text{tails}) = \frac{3}{8} + \frac{1}{4}$$

$$P(\text{tails}) = \frac{3}{8} + \frac{2}{8}$$

$$P(\text{tails}) = \frac{5}{8}$$

Putting these values into Bayes' Theorem, we get

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

$$P(A \mid B) = \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{5}{8}} = \frac{\frac{3}{8}}{\frac{5}{8}} = \frac{3}{8} \cdot \frac{8}{5} = \frac{3}{5}$$

The probability that we flipped the biased coin is  $3/5$ .

- 5. A company is giving a drug test to all of its employees. The test is 90 % accurate, given that a person is using drugs, and 85 % accurate, given that the person is not using drugs. It's also known that 10 % of the general population of employees uses drugs. What is the probability that an employee was actually using drugs, given that they tested positive?

Let  $P$  represent a positive test for an individual.

Let  $N$  represent a negative test for an individual.

Let  $D$  represent the event that an employee is a drug user.

*Solution:*

We're asked to determine the probability that an employee was using drugs, given that they tested positive, or  $P(D|P)$ . Let's use Bayes' Theorem.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(D|P) = \frac{P(P|D) \cdot P(D)}{P(P)}$$

$P(P|D)$  is the probability that an employee tests positive, given that they are a drug user. From the problem, we know that  $P(P|D) = 90\%$ .  $P(D)$  is the probability that an employee is a drug user, and from the problem, we



know that  $P(D) = 10\%$ .  $P(P)$  is the probability of testing positive, regardless of whether the result was accurate or inaccurate.

Let's find the probability the an employee tests positive, and the result is accurate, because they're a drug user. We know 10% of employees are drug users, and we know that 90% of drug users will test positive.

$$(0.10)(0.90) = 0.09$$

Now let's calculate the probability that an employee tested positive, but wasn't a drug user. The problem tells us that the test is 85% accurate for non drug users, which means that 15% of those who aren't using drugs will still test positive. Since 10% of the employees are drug users, 90% are not. So the probability of a false positive from a non drug user is

$$(0.90)(0.15) = 0.135$$

Now we can calculate  $P(D|P)$ .

$$P(D|P) = \frac{(0.90)(0.10)}{0.09 + 0.135}$$

$$P(D|P) = \frac{0.09}{0.225}$$

$$P(D|P) = 40\%$$

This means that, for an employee who tests positive, there's a 40% chance that employee is actually using drugs.



6. Two factories  $A$  and  $B$  produce heaters for car seats. A customer received a defective car seat heater and the manager at factory  $B$  would like to know if it came from her factory. Use the table below to determine the probability that the heater came from factory  $B$ .

Factory	% of production	Probability of defective heaters
A	0.55	0.020 $P(D A)$
B	0.45	0.014 $P(D B)$

*Solution:*

The manager wants to know the probability the heater came from her factory, given it was defective. So she's looking for  $P(B|D)$ . We can use Bayes' Theorem to find the probability. Substituting in with the given events, we get

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(B|D) = \frac{P(D|B) \cdot P(B)}{P(D)}$$

Let's find  $P(D|B)$ ,  $P(B)$ , and  $P(D)$ .  $P(D|B)$  is the probability the heater is defective, given it came from factory  $B$ . We have this probability in the table as  $P(D|B) = 0.014$ .  $P(B)$  is the probability the heater came from factory  $B$ . We also have this in the table as  $P(B) = 0.45$ . Next, we need  $P(D)$ , which is the probability the heater is defective. This is made of the probability

the heater comes from factory  $A$  and is defective and the probability it came from factory  $B$  and is defective. So we need to find  $P(A \cap D) + P(B \cap D)$ .

First let's find the probability that the heater comes from factory  $A$  and is defective.

Factory	% of production	Probability of defective heaters
A	0.55	0.020 $P(D A)$

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

$$P(A \cap D) = (0.55)(0.020)$$

$$P(A \cap D) = 0.011$$

Next let's find the probability the heater comes from factory  $B$  and is defective.

Factory	% of production	Probability of defective heaters
B	0.45	0.014 $P(D B)$

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

$$P(B \cap D) = (0.45)(0.014)$$

$$P(B \cap D) = 0.0063$$

Now we can find  $P(D)$ .

$$P(D) = P(A \cap D) + P(B \cap D)$$

$$P(D) = 0.011 + 0.0063$$

$$P(D) = 0.0173$$

Putting these values into Bayes' Theorem, we get

$$P(B|D) = \frac{P(D|B) \cdot P(B)}{P(D)}$$

$$P(B|D) = \frac{(0.014) \cdot (0.45)}{0.0173}$$

$$P(B|D) = \frac{0.0063}{0.0173}$$

$$P(B|D) \approx 36\%$$

There is about a 36% chance the defective heater came from factory  $B$ .



## DISCRETE PROBABILITY

- 1. Let  $X$  be a discrete random variable with the following probability distribution. Find  $P(X \geq 3)$ .

<b>X</b>	1	2	3	4	5
<b>P(X)</b>	0.35	0.25	0.20	0.15	?

*Solution:*

First, we need to find the  $P(X = 5)$ , which we'll do by subtracting all the other probabilities from 1.

$$P(X = 5) = 1 - 0.35 - 0.25 - 0.20 - 0.15$$

$$P(X = 5) = 1 - 0.95$$

$$P(X = 5) = 0.05$$

Then the probability that  $X \geq 3$  is

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(X \geq 3) = 0.20 + 0.15 + 0.05$$

$$P(X \geq 3) = 0.40$$

2. Let  $B$  be a discrete random variable with the following probability distribution. Find  $\mu_B$  and  $\sigma_B$ .

<b>B</b>	0	5	10	15
<b>P(B)</b>	1/5	1/5	2/5	1/5

*Solution:*

We'll weight each value of  $B$  by the probability that the value occurs,  $P(B)$ , in order to find the expected value  $\mu_B$ .

$$\mu_B = E(B) = \sum_{i=1}^4 B_i P(B_i) = 0 \left( \frac{1}{5} \right) + 5 \left( \frac{1}{5} \right) + 10 \left( \frac{2}{5} \right) + 15 \left( \frac{1}{5} \right)$$

$$\mu_B = 8$$

In order to find the standard deviation of  $B$ ,  $\sigma_B$ , we have to find variance first.

$$\sigma_B^2 = \sum_{i=1}^4 (B_i - \mu_B)^2 P(B_i)$$

$$\sigma_B^2 = (0 - 8)^2 \left( \frac{1}{5} \right) + (5 - 8)^2 \left( \frac{1}{5} \right) + (10 - 8)^2 \left( \frac{2}{5} \right) + (15 - 8)^2 \left( \frac{1}{5} \right)$$

$$\sigma_B^2 = 26$$

Then the standard deviation is

$$\sqrt{\sigma_B^2} = \sqrt{26}$$

$$\sigma_B \approx 5.099$$

- 3. The table shows the distribution of size of households in the U.S. for 2016. Suppose we select a household of size at least 2 at random. What is the probability that this household has a size of at least 4?

Size of household	1	2	3	4	5	6	7+
P(size)	0.281	0.340	?	0.129	0.060	0.023	0.013

*Solution:*

Find the probability that the household is a 3-person household.

$$P(X = 3) = 1 - 0.281 - 0.340 - 0.129 - 0.060 - 0.023 - 0.013$$

$$P(X = 3) = 0.154$$

The probability of “at least 4” is

$$P(\text{at least } 4) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)$$

$$P(\text{at least } 4) = 0.129 + 0.060 + 0.023 + 0.013$$

$$P(\text{at least } 4) = 0.225$$

and the probability of “at least 2” is



$$P = (\text{at least } 2) = 1 - P(X = 1)$$

$$P = (\text{at least } 2) = 1 - 0.281$$

$$P = (\text{at least } 2) = 0.719$$

Then the probability that the household size is at least 4, given that the household size is at least 2, is

$$P(\text{at least } 4, \text{ given at least } 2) = \frac{P(\text{size at least } 4)}{P(\text{size at least } 2)}$$

$$P(\text{at least } 4, \text{ given at least } 2) = \frac{0.225}{0.719} \approx 0.313$$

- 4. A standard deck of cards is shuffled, and two cards are selected without replacement. Let  $R$  be the number of red cards selected. Construct a probability distribution for  $R$ .

*Solution:*

If we draw two cards, we can find the probability that either both are red  $P(R = 2)$ , that one is red  $P(R = 1)$ , or that neither are red  $P(R = 0)$ .

$$P(R = 0) = \frac{26}{52} \left( \frac{25}{51} \right) = \frac{25}{102}$$

$$P(R = 2) = \frac{26}{52} \left( \frac{25}{51} \right) = \frac{25}{102}$$

Then the probability that one card is red is

$$P(R = 1) = 1 - P(R = 2) - P(R = 0)$$

$$P(R = 1) = 1 - \frac{25}{102} - \frac{25}{102}$$

$$P(R = 1) = \frac{102}{102} - \frac{25}{102} - \frac{25}{102}$$

$$P(R = 1) = \frac{52}{102} = \frac{26}{51}$$

Which means we can build a probability distribution for  $R$ .

R	0	1	2
P(R)	25/102	52/102	25/102

- 5. A local restaurant features a wheel we can spin before paying the bill. The wheel is split into 8 equal size pieces. One of the sections gives us a \$10 discount on the bill, two sections give a \$5 discount, three sections give a \$2 discount, and the rest of the sections give no discount. Find the expected value for the discount given by the wheel.

*Solution:*

Let  $X$  be the amount of the discount. Then the expected value, or mean of the discount is

$$E(X) = \sum X P(X) = 10 \left( \frac{1}{8} \right) + 5 \left( \frac{2}{8} \right) + 2 \left( \frac{3}{8} \right) + 0 \left( \frac{2}{8} \right)$$

$$E(X) = \$3.25$$

- 6. John stops at the local gas station and decides to buy lottery tickets. Each ticket has a 20% chance of being a winner. He will buy a lottery ticket and check to see if it's a winner. If it's a winner, he'll collect his money and be done. If it's not a winner, he'll buy another. He'll repeat this until he gets a winning ticket. But if he hasn't won by his fifth ticket, he won't buy any more tickets. Let  $L$  be the number of lottery tickets John will buy, then find  $E(L)$ .

*Solution:*

We could find the probability of winning on each of the first four tickets.

$$P(L = 1) = 0.2$$

$$P(L = 2) = (0.2)(0.8) = 0.16$$

$$P(L = 3) = (0.2)(0.8)(0.8) = 0.128$$

$$P(L = 4) = (0.2)(0.8)(0.8)(0.8) = 0.1024$$

If we continue this pattern, we might think that the probability of winning on the fifth ticket would be

$$P(L = 5) = (0.2)(0.8)(0.8)(0.8)(0.8) = 0.08192$$

But the question tells us that John will never buy more than five tickets. Because he's guaranteed to buy one, two, three, four, or five tickets, the probability that he's going to purchase one of those numbers of tickets must be 100 %. So the probability that he purchases five tickets is actually

$$P(L = 5) = 1 - P(L \leq 4)$$

$$P(L = 5) = 1 - (0.2 + 0.16 + 0.128 + 0.1024)$$

$$P(L = 5) = 1 - 0.5904$$

$$P(L = 5) = 0.4096$$

Then the expected value for the number of tickets he'll buy,  $L$ , is

$$E(L) = 1(0.2) + 2(0.16) + 3(0.128) + 4(0.1024) + 5(0.4096)$$

$$E(L) = 0.2 + 0.32 + 0.384 + 0.4096 + 2.048$$

$$E(L) = 3.3616$$



## TRANSFORMING RANDOM VARIABLES

### ■ 1. We use the formula

$$^{\circ}F = \frac{9}{5}^{\circ}C + 32$$

to convert from Celsius to Fahrenheit. August is the hottest month in Hawaii with a mean temperature of  $27^{\circ}C$ . What is the mean temperature in Hawaii in  $^{\circ}F$ .

*Solution:*

We'll plug  $27^{\circ}C$  into the conversion formula to get the corresponding value in Fahrenheit.

$$\mu_{^{\circ}F} = \frac{9}{5}\mu_{^{\circ}C} + 32 = \frac{9}{5}(27) + 32 = \frac{243}{5} + 32 = 80.6^{\circ}$$

### ■ 2. Let $Z$ be a random variable with $\sigma_Z^2 = 49$ . Let $W = (1/2)Z - 10$ . Find $\sigma_W$ .

*Solution:*

We've been given the variance of  $Z$ , so we need to use it first to find the standard deviation of  $Z$ .



$$\sqrt{\sigma_Z^2} = \sqrt{49}$$

$$\sigma_Z = 7$$

Standard deviation is effected by scaling, but not by shifting. So when we convert from  $Z$  to  $W$  using

$$W = \frac{1}{2}Z - 10$$

we need to multiply by  $1/2$ , but we don't need to shift by  $10$ . So we can say that the standard deviation  $\sigma_W$  is

$$\sigma_W = \frac{1}{2}\sigma_Z$$

$$\sigma_W = \frac{1}{2}(7)$$

$$\sigma_W = \frac{7}{2}$$

- 3. The students in each 8th period classroom were asked to donate money for a school fundraiser, and the class that raised the most money was awarded a pizza party. The school secretary recorded the amount raised by each class and made a five-number summary for the data.

Min	Q1	Median	Q3	Max
4.50	15.25	22.00	38.75	95.50

If a donor commits to matching equally the students' donations, create a new five-number summary of the total amount raised, including the donor's contribution.

*Solution:*

The donor is essentially scaling the data set, because she's doubling the students' donations. And we know that scaling a data set scales all the values in the five-number summary.

Therefore, after putting the donor's contribution together with the students' donations, we can give a the new five-number summary for the data set as

Min	Q1	Median	Q3	Max
4.5(2)	15.25(2)	22(2)	38.75(2)	95.5(2)

Min	Q1	Median	Q3	Max
9	30.5	44	77.5	191

- 4. The number of items sold at a concession stand is normally distributed with  $\mu = 323$  and  $\sigma = 30$ . The average price per item sold is \$1.25. Different student clubs volunteer to work the concession stand throughout the year and get to keep half of their sales to go towards their club's activities. What is the probability that a club will get to keep more than \$220 in sales?



*Solution:*

Let  $N$  be the number of items sold and  $S$  be the amount of money the club gets to keep. Then we could write an equation for the amount of money they keep as

$$S = \frac{1}{2}(1.25)(N)$$

We know the mean of  $N$  is  $\mu_N = 323$ , so we can use the conversion equation to find the mean of  $S$ .

$$\mu_S = \frac{1}{2}(1.25)(\mu_N) = \frac{1}{2}(1.25)(323) \approx 201.88$$

Using  $\sigma_N = 30$ , we can find the standard deviation of  $S$  in the same way.

$$\sigma_S = \frac{1}{2}(1.25)(\sigma_N) = \frac{1}{2}(1.25)(30) = 18.75$$

With a mean of  $\mu_S = 201.88$  and a standard deviation of  $\sigma_S = 18.75$ , we can find the probability that the club will take home more than \$220. We'll find the  $z$ -score associated with \$220.

$$z = \frac{220 - 201.88}{18.75} = 0.9664 \approx 0.97$$

Since we're looking at 0.97 standard deviations above the mean, we get 0.8340 from the  $z$ -table, which tells us that the probability that the club takes home more than \$220 is  $1 - 0.8340 = 0.166 \approx 17\%$ . They have an approximately 17% chance of taking home more than \$220.



- 5. The average length of a full-term new born baby is 20 inches with variance 0.81 inches. What are the mean and standard deviation of the length of a full-term new born, expressed in centimeters? Use 1 in = 2.54 cm.

*Solution:*

To convert between inches and centimeters, we'll say

$$\mu_{\text{length in cm}} = 2.54\mu_{\text{length in in}}$$

$$\mu_{\text{length in cm}} = 2.54(20)$$

$$\mu_{\text{length in cm}} = 50.8 \text{ cm}$$

We'll use the same conversion formula to convert the standard deviation.

$$\sigma_{\text{length in cm}} = 2.54\sigma_{\text{length in in}}$$

$$\sigma_{\text{length in cm}} = 2.54\sqrt{0.81}$$

$$\sigma_{\text{length in cm}} = 2.286 \text{ cm}$$

- 6. The weights of full-term new born babies are normally distributed with  $\mu = 120$  ounces and  $\sigma = 20$  ounces. Describe the shape, center, and spread



for the weights of full-term new born babies as measured in pounds. Use 1 pound = 16 ounces.

*Solution:*

We can use a conversion formula to convert the mean from pounds to ounces.

$$\mu_{\text{weight in pounds}} = \frac{1}{16} \mu_{\text{weight in ounces}}$$

$$\mu_{\text{weight in pounds}} = \frac{1}{16}(120)$$

$$\mu_{\text{weight in pounds}} = 7.5 \text{ pounds}$$

Now we'll convert the given standard deviation.

$$\sigma_{\text{weight in pounds}} = \frac{1}{16} \sigma_{\text{weight in ounces}}$$

$$\sigma_{\text{weight in pounds}} = \frac{1}{16}(20)$$

$$\sigma_{\text{weight in pounds}} = 1.25 \text{ pounds}$$

The distribution of weights of full-term new born babies remains normally distributed, even after converting from ounces to pounds. The mean is  $\mu = 7.5$  pounds and the standard deviation is  $\sigma = 1.25$  pounds.



## COMBINATIONS OF RANDOM VARIABLES

- 1.  $X$  and  $Y$  are independent random variables with  $E(X) = 48$ ,  $E(Y) = 54$ ,  $SD(X) = 3$  and  $SD(Y) = 5$ . Find  $E(X - Y)$  and  $SD(X - Y)$ .

*Solution:*

To find the expected value of the difference, we find the difference of the expected values.

$$E(X - Y) = E(X) - E(Y) = 48 - 54 = -6$$

To find the standard deviation of the difference, we have to square both standard deviations in order to get the variances. We get  $SD^2(X) = 3^2 = 9$  and  $SD^2(Y) = 5^2 = 25$ . Then we can find the standard deviation of the difference.

$$SD(X - Y) = \sqrt{SD^2(X) + SD^2(Y)} = \sqrt{3^2 + 5^2} = \sqrt{34} \approx 5.831$$

- 2.  $A$  and  $B$  are independent random variables with  $E(A) = 6.5$ ,  $E(B) = 4.4$ ,  $SD(A) = 1.6$ , and  $SD(B) = 2.1$ . Find  $E(4A + 2B)$  and  $SD(4A + 2B)$ .

*Solution:*



The expected value of the sum of variables is the sum of the expected values.

$$E(4A + 2B) = 4E(A) + 2E(B) = 4(6.5) + 2(4.4) = 34.8$$

Then we'll find the standard deviation of the combination. When we scale a variable by some constant  $k$ , its standard deviation gets scaled by  $k$  as well. So the standard deviations of  $4A$  and  $2B$  are

$$SD(A) = 1.6$$

$$SD(4A) = 4(1.6)$$

$$SD(4A) = 6.4$$

and

$$SD(B) = 2.1$$

$$SD(2B) = 2(2.1)$$

$$SD(2B) = 4.2$$

Then the variances of the variables  $4A$  and  $2B$  are

$$SD^2(4A) = 6.4^2$$

$$SD^2(4A) = 40.96$$

and

$$SD^2(2B) = 4.2^2$$

$$SD^2(2B) = 17.64$$

The variance of a combination is the sum of the variances, so

$$SD^2(4A) + SD^2(2B) = 40.96 + 17.64$$

$$SD^2(4A) + SD^2(2B) = 58.6$$

Then the standard deviation of the combination is

$$\sqrt{SD^2(4A) + SD^2(2B)} = \sqrt{58.6}$$

$$\sqrt{SD^2(4A) + SD^2(2B)} \approx 7.66$$

- 3. The time it takes students to complete multiple choice questions on an AP Statistics Exam has a mean of 55 seconds with a standard deviation of 12 seconds. If the exam consists of 40 multiple choice questions, find the mean total time to finish the exam. Then find the standard deviation in the total time. What assumption must be made?

*Solution:*

We have to assume that the questions are independent. Then we can say that the mean finishing time is

$$\mu_{Q_1} + \mu_{Q_2} + \mu_{Q_3} + \dots + \mu_{Q_{40}} = 40(55) = 2,200 \approx 36.67 \text{ minutes}$$

and that the variance of the finishing time is

$$\sigma_{Q_1}^2 + \sigma_{Q_2}^2 + \sigma_{Q_3}^2 + \dots + \sigma_{Q_{40}}^2 = 40(12^2) = 40(144) = 5,760 \text{ seconds}$$



such that the standard deviation of the finishing time is

$$\sigma = \sqrt{5,760} \approx 75.89 \approx 1.26 \text{ minutes}$$

- 4. Let  $M$  represent the height of a male over 21 years of age and let  $W$  represent the height of a female over 21 years of age. Let  $D$  represent the difference between their heights ( $D = M - W$ ). Let  $E(M) = 70$  inches,  $\sigma_M = 2.8$  inches,  $E(W) = 64.5$  inches and  $\sigma_W = 2.4$  inches.

What is the mean and standard deviation of the difference between the two heights?

*Solution:*

To find the mean of the difference, we'll find the difference of the means.

$$E(M - W) = E(M) - E(W) = 70 - 64.5 = 5.5 \text{ inches}$$

We'll find variance in order to get standard deviation. The variances are  $\sigma_M^2 = 2.8^2 = 7.84$  and  $\sigma_W^2 = 2.4^2 = 5.76$ . Therefore, the standard deviation of the difference is

$$\sigma(M - W) = \sqrt{\sigma_M^2 + \sigma_W^2} = \sqrt{7.84 + 5.76} = \sqrt{13.6} \approx 3.69 \text{ inches}$$

- 5. The Ironman is a challenge in which a competitor swims 2.4 miles, then bikes 112 miles, and finally runs 26.2 miles. Suppose the times for each



of the legs are normally distributed with the given means and standard deviations.

**Swim:**  $\mu_S = 76$  minutes and  $\sigma_S = 18$  minutes

**Bike:**  $\mu_B = 385$  minutes and  $\sigma_B = 32$  minutes

**Run:**  $\mu_R = 294$  minutes and  $\sigma_R = 25$  minutes

What percent of the competitors finish the Ironman in under 710 minutes?

*Solution:*

Let  $T$  be the total time to complete all three legs of the Ironman. Then the mean finishing time is

$$\mu_T = \mu_S + \mu_B + \mu_R = 76 + 385 + 294 = 755 \text{ minutes}$$

Assuming the legs are independent random variables, then we can find the sum of the variances to get the variance of the sum.

$$\sigma_T^2 = \sigma_S^2 + \sigma_B^2 + \sigma_R^2 = 18^2 + 32^2 + 25^2 = 1,973$$

Then the standard deviation of finishing time is

$$\sigma_T = \sqrt{1,973} \approx 44.42 \text{ minutes}$$

Since  $S$ ,  $B$ , and  $R$  are normally distributed,  $T$  will also be normally distributed. To find the probability that a finisher will finish in under 710 minutes, we'll find the  $z$ -score associated with 710 minutes.



$$z = \frac{710 - 755}{44.42} \approx -1.01$$

If we look up a  $z$ -score of  $z = -1.01$  in a  $z$ -table, we get 0.1562, which means there's an approximately 15.62% chance that a finisher finishes in under 710 minutes.

- 6. We buy a scratch-off lottery ticket for \$1 at the local gas station. If we get three hearts in a row on the scratch-off, the state will pay us \$500. Let  $X$  be the amount the state pays us and let  $X$  have the following probability distribution.

$X$	\$0	\$500
$P(X)$	0.999	0.001

Suppose we buy one of these scratch-off tickets every day for a week (7 days). Find the expected value and standard deviation of our total winnings.

*Solution:*

The expected value of our winnings on any one ticket is

$$E(X) = 0(0.999) + 500(0.001) = \$0.50$$

Find the standard deviation of the winnings by taking the sum of the variances, weighted by the associated probabilities.



$$SD(X) = \sqrt{(0 - 0.5)^2(0.999) + (500 - 0.5)^2(0.001)} = \sqrt{249.75} \approx 15.80$$

Let  $W$  be the amount the state pays us for 7 lottery tickets. The expected value of the total winnings for 7 lottery tickets is therefore  $E(W) = 7(0.5) = \$3.50$ . The standard deviation of the total winnings is

$$SD(W) = \sqrt{(15.80)^2 + (15.80)^2 + \dots + (15.80)^2}$$

$$SD(W) = \sqrt{7(15.80)^2}$$

$$SD(W) = \sqrt{1,747.48}$$

$$SD(W) \approx \$41.80$$

These are the mean and standard deviation of total winnings. They don't account for the price we paid for the scratch-off tickets. If we want to account for the cost of the tickets in order to calculate profit, instead of just winnings, then we'd use the probability distribution

$X$	-\$1	\$499
$P(X)$	0.999	0.001

in order to calculate the mean and standard deviation of our profit.



## PERMUTATIONS AND COMBINATIONS

### ■ 1. Calculate the binomial coefficient.

$$\binom{12}{7}$$

*Solution:*

Use the combination formula

$$\binom{n}{k} = {}^nC_k = \frac{n!}{k!(n-k)!}$$

Plug in  $n = 12$  and  $k = 7$ .

$$\binom{12}{7} = {}^{12}C_7 = \frac{12!}{7! \cdot 5!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

$$\binom{12}{7} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{5 \cdot 4 \cdot 3 \cdot 2}$$

$$\binom{12}{7} = 792$$

### ■ 2. Calculate ${}_{10}P_3$ .



*Solution:*

Use the permutation formula

$${}_n P_k = \frac{n!}{(n - k)!}$$

Plug in  $n = 10$  and  $k = 3$ .

$${}_{10} P_3 = \frac{10!}{(10 - 3)!} = \frac{10!}{7!}$$

$${}_{10} P_3 = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

$${}_{10} P_3 = 10 \cdot 9 \cdot 8$$

$${}_{10} P_3 = 720$$

■ 3. How much greater is  ${}_5 P_2$  than  ${}_5 C_2$ ?

*Solution:*

We'll calculate both values, then find the difference.

$${}_5 P_2 = \frac{5!}{(5 - 2)!} = \frac{5!}{3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 5 \cdot 4 = 20$$

$${}_5 C_2 = \frac{5!}{2!(5 - 2)!} = \frac{5!}{2! \cdot 3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} = \frac{5 \cdot 4}{2} = 10$$



The difference between  $_5P_2$  and  $_5C_2$  is

$$_5P_2 - _5C_2 = 20 - 10 = 10$$

- 4. The high school girls' basketball team has 8 players, 5 of whom are seniors. They need to figure out which senior will be captain and which senior will be co-captain. To make it fair, they choose two players out of a hat. The first drawn will be captain and the second will be co-captain. How many different captain/co-captain pairs are possible?

*Solution:*

Since the order matters, we have to calculate the permutations. There are 5 seniors we can choose from, and 2 spots to put them in.

$$_n P_k = \frac{n!}{(n-k)!} = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 5 \cdot 4 = 20$$

There are 20 possible captain/co-captain pairs.

- 5. How many different ways can the letters in the word "SUCCESS" be rearranged?

*Solution:*



Since the order matters, we have to calculate the permutations. There are 7 letters we can choose from, and 7 spots to put them in.

$${}_n P_k = \frac{n!}{(n - k)!} = \frac{7!}{(7 - 7)!} = \frac{7!}{0!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} = 5,040$$

But since the letter S repeats three times in the word and the letter C repeats twice, the actual number of unique rearrangements will be less than 5,040. We have to divide by 3! for the S and by 2! for the C.

$$\frac{5,040}{3! \cdot 2!} = \frac{5,040}{3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = \frac{5,040}{12} = 420$$

There are 420 possible arrangements of the letters.

- 6. Mrs. B's kindergarten class has 14 students and Mr. G's kindergarten class has 16 students. Three students will be selected at random from each of these classrooms to ride on a float in the school parade coming up next week. How many different groups of 6 can be chosen to ride the float?

*Solution:*

Since order doesn't matter, this is a combination question. We need to find the combination for Mrs. B's class, and then the combination for Mr. G's class. Then we'll multiply those together get the total number of combinations.

$${}_n C_k \cdot {}_n C_k = {}_{14} C_3 \cdot {}_{16} C_3 = 364 \cdot 560 = 203,840$$

## BINOMIAL RANDOM VARIABLES

- 1. We toss a fair coin 15 times and record the number of tails. Is this experiment modeled by a binomial random variable? If it isn't, explain why. If it is, determine its parameters  $n$  and  $p$  and express the binomial random variable as  $X \sim B(n, p)$ .

*Solution:*

Yes, this experiment results in a binomial random variable. Let  $X$  be the number of tails observed out of 15 tosses. We know that  $p = 0.5$  for each trial because there are only two possible outcomes, heads or tails. Therefore,  $X \sim B(15, 0.5)$ .

- 2. We randomly select students from our school until we find a student in the school band. Assume there are 900 students in the school and 80 participate in the school band. Is this experiment modeled by a binomial random variable? If it isn't, explain why. If it is, determine its parameters  $n$  and  $p$  and express the binomial random variable as  $X \sim B(n, p)$ .

*Solution:*

No, this experiment does not result in a binomial random variable. We do have a fixed probability of success,



$$p = \frac{80}{900} = \frac{4}{45} \approx 0.09$$

and the trials can be considered independent because we have a large population. But we're not using a fixed number of trials, because we're continuing to select students until we find one in the band, and we don't know how many trials that will take.

- 3. Let  $X \sim B(n, p)$  be a binomial random variable with  $n = 12$  and  $p = 0.08$ . Find  $P(X = 4)$ .

*Solution:*

We're being asked to find the probability that we get exactly 4 successes in 12 trials, if the probability of success is  $p = 0.08$ .

$$P(X = 4) = \binom{12}{4}(0.08)^4(1 - 0.08)^8$$

$$P(X = 4) = (495)(0.08)^4(1 - 0.08)^8$$

$$P(X = 4) \approx 0.0104$$

- 4. Let  $Y$  be the number of times we roll a 1 on a fair 6-sided die if we do 10 trials. Fill in the following probability distribution for  $Y$ , rounding each probability to 4 decimal places.



<b>Y</b>	0	1	2	3	4	5	6	7	8	9	10
<b>P(Y)</b>											

*Solution:*

With  $n = 10$ ,  $p = 1/6$ , and  $k = 0, 1, 2, 3, \dots, 10$ , find  $P(Y = k)$  for each value of  $k$  using

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

After rounding each value to 4 decimal places, the table is

<b>Y</b>	0	1	2	3	4	5	6	7	8	9	10
<b>P(Y)</b>	0.1615	0.3230	0.2907	0.1550	0.0543	0.0130	0.0022	0.0003	0.0000	0.0000	0.0000

■ 5. For each binomial random variable, determine whether the shape of the probability distribution will be skewed right, skewed left, or symmetrical.

1.  $X \sim B(n, p)$  with  $n = 10$  and  $p = 0.15$
2.  $Y \sim B(n, p)$  with  $n = 10$  and  $p = 0.75$
3.  $Z \sim B(n, p)$  with  $n = 10$  and  $p = 0.50$

*Solution:*

The probability distribution for  $X$  will be skewed right because the probability of success,  $p = 0.15$ , is less than 0.5.

The probability distribution for  $Y$  will be skewed left because the probability of success,  $p = 0.75$ , is greater than 0.5.

The probability distribution for  $Z$  will be symmetrical because the probability of success,  $p = 0.50$ , is exactly 0.5.

- 6. Suppose an environmental biologist is studying juvenile sunfish mortality. He finds that only 30 % of juvenile sunfish survive in a certain lake. Out of 8 randomly selected juvenile sunfish, what is the probability that exactly 3 will survive?

*Solution:*

We're finding the probability that we get exactly 3 successes in 8 trials.

$$P(X = 3) = \binom{8}{3}(0.3)^3(1 - 0.3)^5$$

$$P(X = 3) = (56)(0.3)^3(1 - 0.3)^5$$

$$P(X = 3) \approx 0.2541$$

## POISSON DISTRIBUTIONS

- 1. A student is able to solve 10 practice problems per hour, on average. Find the probability that she can solve 12 in the next hour.

*Solution:*

We know this is a Poisson experiment with the following values:

$\lambda = 10$ , the average number of practice problems solved in an hour

$x = 12$ , the number of homework problems she wants to complete in the next hour

The Poisson probability is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(12) = \frac{10^{12} e^{-10}}{12!}$$

$$P(12) \approx 0.0948$$

So the probability the student will solve 12 homework problems is approximately 0.0948 or 9.48 % .



- 2. A student is able to solve 6 practice problems per hour, on average. Find the probability that she can solve at least 4 in the next hour.

*Solution:*

The probability that the student solves at least 4 practice problems is the probability that she doesn't solve either 0, 1, 2, or 3 practice problems. So the probability we need to find is

$$P(X \geq 4) = 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3)$$

We know this is a Poisson experiment with  $\lambda = 6$ , the average number of practice problems solved in an hour, so the Poisson probability is

$$P(X \geq 4) = 1 - \frac{6^0 e^{-6}}{0!} - \frac{6^1 e^{-6}}{1!} - \frac{6^2 e^{-6}}{2!} - \frac{6^3 e^{-6}}{3!}$$

$$P(X \geq 4) = 1 - \frac{e^{-6}}{1} - \frac{6e^{-6}}{1} - \frac{36e^{-6}}{2} - \frac{216e^{-6}}{6}$$

$$P(X \geq 4) = 1 - \frac{1}{e^6} - \frac{6}{e^6} - \frac{18}{e^6} - \frac{36}{e^6}$$

$$P(X \geq 4) = 1 - \frac{1}{e^6}(1 + 6 + 18 + 36)$$

$$P(X \geq 4) = 1 - \frac{61}{e^6}$$

$$P(X \geq 4) \approx 0.8488$$

So the probability the student will solve at least 4 homework problems is approximately 0.8488 or 84.88 %.

- 3. A student is able to solve 5 practice problems per hour, on average. Find the probability that she solves at most 3 in the next hour.

*Solution:*

The probability that the student solves at most 3 practice problems is the probability that she solves either 0, 1, 2, or 3 practice problems. So the probability we need to find is

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

We know this is a Poisson experiment with  $\lambda = 5$ , the average number of practice problems solved in an hour, so the Poisson probability is

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$P(X \leq 3) = \frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!} + \frac{5^2 e^{-5}}{2!} + \frac{5^3 e^{-5}}{3!}$$

$$P(X \leq 3) = \frac{e^{-5}}{1} + \frac{5e^{-5}}{1} + \frac{25e^{-5}}{2} + \frac{125e^{-5}}{6}$$

$$P(X \leq 3) = \frac{1}{e^5} + \frac{5}{e^5} + \frac{25}{2e^5} + \frac{125}{6e^5}$$

$$P(X \leq 3) = \frac{1}{e^5} \left( 1 + 5 + \frac{25}{2} + \frac{125}{6} \right)$$



$$P(X \leq 3) = \frac{1}{e^5} \left( \frac{36}{6} + \frac{75}{6} + \frac{125}{6} \right)$$

$$P(X \leq 3) = \frac{118}{3e^5}$$

$$P(X \leq 3) \approx 0.2650$$

So the probability the student will solve at most 3 homework problems is approximately 0.2650 or 26.50 % .

- 4. A baker is able to bake 50 loaves of bread per day, on average. Find the probability that he can bake 60 on Friday.

*Solution:*

We know this is a Poisson experiment with the following values:

$\lambda = 50$ , the average number of loaves of bread baked per day

$x = 60$ , the number of loaves of bread he wants to bake on Friday

The Poisson probability is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(60) = \frac{50^{60} e^{-50}}{60!}$$



$$P(60) \approx 0.0201$$

So the probability the baker will bake 60 loaves of bread is approximately 0.0201 or 2.01 % .

- 5. A baker is able to bake 10 cakes per hour, on average. Find the probability that he can bake more than 5 in the next hour.

*Solution:*

The probability that the baker bakes more than 5 cakes is the probability that he doesn't bake either 0, 1, 2, 3, 4, or 5 cakes. So the probability we need to find is

$$P(X > 5) = 1 - P(X = 0) - P(X = 1) - P(X = 2)$$

$$-P(X = 3) - P(X = 4) - P(X = 5)$$

We know this is a Poisson experiment with  $\lambda = 10$ , the average number of cakes baked in an hour, so the Poisson probability is

$$P(X > 5) = 1 - \frac{10^0 e^{-10}}{0!} - \frac{10^1 e^{-10}}{1!} - \frac{10^2 e^{-10}}{2!}$$

$$-\frac{10^3 e^{-10}}{3!} - \frac{10^4 e^{-10}}{4!} - \frac{10^5 e^{-10}}{5!}$$

$$P(X > 5) = 1 - \frac{e^{-10}}{1} - \frac{10e^{-10}}{1} - \frac{100e^{-10}}{2}$$



$$\frac{1,000e^{-10}}{6} - \frac{10,000e^{-10}}{24} - \frac{100,000e^{-10}}{120}$$

$$P(X > 5) = 1 - \frac{1}{e^{10}} - \frac{10}{e^{10}} - \frac{50}{e^{10}} - \frac{500}{3e^{10}} - \frac{1,250}{3e^{10}} - \frac{2,500}{3e^{10}}$$

$$P(X > 5) = 1 - \frac{1}{e^{10}} \left( 1 + 10 + 50 + \frac{500}{3} + \frac{1,250}{3} + \frac{2,500}{3} \right)$$

$$P(X > 5) = 1 - \frac{4,433}{3e^{10}}$$

$$P(X > 5) \approx 0.9329$$

So the probability the baker will bake more than 5 cakes is approximately 0.9329 or 93.29 %.

- 6. A baker is able to frost 2 cakes per hour, on average. Find the probability that he frosts fewer than 5 cakes in the next hour.

*Solution:*

The probability that the baker frosts fewer than 5 cakes is the probability that he frosts either 0, 1, 2, 3, or 4 cakes. So the probability we need to find is

$$P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

We know this is a Poisson experiment with  $\lambda = 2$ , the average number of cakes frosted in an hour, so the Poisson probability is



$$P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$P(X < 5) = \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \frac{2^2 e^{-2}}{2!} + \frac{2^3 e^{-2}}{3!} + \frac{2^4 e^{-2}}{4!}$$

$$P(X < 5) = \frac{e^{-2}}{1} + \frac{2e^{-2}}{1} + \frac{4e^{-2}}{2} + \frac{8e^{-2}}{6} + \frac{16e^{-2}}{24}$$

$$P(X < 5) = \frac{1}{e^2} + \frac{2}{e^2} + \frac{2}{e^2} + \frac{4}{3e^2} + \frac{2}{3e^2}$$

$$P(X < 5) = \frac{1}{e^2} \left( 1 + 2 + 2 + \frac{4}{3} + \frac{2}{3} \right)$$

$$P(X < 5) = \frac{7}{e^2}$$

$$P(X < 5) \approx 0.9473$$

So the probability the baker will frost fewer than 5 cakes is approximately 0.9473 or 94.73 %.



## “AT LEAST” AND “AT MOST,” AND MEAN, VARIANCE, AND STANDARD DEVIATION

- 1. Assume  $X$  is a binomial random variable. Let  $X \sim B(n, p)$  with  $n = 15$  and  $p = 0.45$ . Find  $P(X > 7)$ .

*Solution:*

Since we’re running  $n = 15$  trials, and we want to find the probability that we get the first success *after* the 7th trial, we can express this as

$$P(X > 7) = P(X = 8) + P(X = 9) + \dots + P(X = 15)$$

which is the same as

$$1 - P(X \leq 7)$$

Find  $P(X \leq 7)$ .

$$\begin{aligned} P(X \leq 7) &= \binom{15}{0}(0.45)^0(1 - 0.45)^{15} + \binom{15}{1}(0.45)^1(1 - 0.45)^{14} \\ &\quad + \binom{15}{2}(0.45)^2(1 - 0.45)^{13} + \binom{15}{3}(0.45)^3(1 - 0.45)^{12} \\ &\quad + \binom{15}{4}(0.45)^4(1 - 0.45)^{11} + \binom{15}{5}(0.45)^5(1 - 0.45)^{10} \\ &\quad + \binom{15}{6}(0.45)^6(1 - 0.45)^9 + \binom{15}{7}(0.45)^7(1 - 0.45)^8 \end{aligned}$$



$$\begin{aligned}
 P(X \leq 7) &= 0.55^{15} + 15(0.45)(0.55^{14}) \\
 &\quad + 105(0.45^2)(0.55^{13}) + 455(0.45^3)(0.55^{12}) \\
 &\quad + 1,365(0.45^4)(0.55^{11}) + 3,003(0.45^5)(0.55^{10}) \\
 &\quad + 5,005(0.45^6)(0.55^9) + 6,435(0.45^7)(0.55^8)
 \end{aligned}$$

$$P(X \leq 7) \approx 0.6535$$

Then  $1 - P(X \leq 7)$  is

$$1 - P(X \leq 7) \approx 1 - 0.6535$$

$$1 - P(X \leq 7) \approx 0.3465$$

- 2. According to a 2017-2018 survey, 68 % of U.S. households own a pet. Suppose we select 12 households at random. What is the probability that fewer than 8 of them own a pet?

*Solution:*

Let  $X$  be the number of households that own a pet. Then we can express the variable as  $X \sim B(12, 0.68)$ . The probability that we'll have fewer than 8 successes is

$$P(X < 8) = P(X \leq 7) = P(X = 0) + P(X = 1) + \dots + P(X = 7)$$

Find  $P(X < 8)$ .

$$P(X < 8) = \binom{12}{0}(0.68)^0(1 - 0.68)^{12} + \binom{12}{1}(0.68)^1(1 - 0.68)^{11}$$

$$+ \binom{12}{2}(0.68)^2(1 - 0.68)^{10} + \binom{12}{3}(0.68)^3(1 - 0.68)^9$$

$$+ \binom{12}{4}(0.68)^4(1 - 0.68)^8 + \binom{12}{5}(0.68)^5(1 - 0.68)^7$$

$$+ \binom{12}{6}(0.68)^6(1 - 0.68)^6 + \binom{12}{7}(0.68)^7(1 - 0.68)^5$$

$$P(X < 8) = (0.32^{12}) + 12(0.68)(0.32^{11})$$

$$+ 66(0.68^2)(0.32^{10}) + 220(0.68^3)(0.32^9)$$

$$+ 495(0.68^4)(0.32^8) + 792(0.68^5)(0.32^7)$$

$$+ 924(0.68^6)(0.32^6) + 792(0.68^7)(0.32^5)$$

$$P(X < 8) \approx 0.3308$$

- 3. According to a 2017-2018 survey, 68 % of U.S. households own a pet. Suppose 200 households are selected at random. Find the expected value and standard deviation for the number of households that own a pet.

*Solution:*



Let  $X$  be the number of households that own a pet. Then we can express the variable as  $X \sim B(12,0.68)$ . The expected value is

$$\mu_X = E(X) = (200)(0.68) = 136 \text{ households}$$

The variance is

$$\sigma_X^2 = Var(X) = (200)(0.68)(1 - 0.68) = 43.52$$

which means the standard deviation is

$$\sigma_X = SD(X) = \sqrt{43.52} \approx 6.597 \text{ households}$$

- 4. 3% of runners in the Boston Marathon do not finish. Suppose we select a SRS of 140 Boston Marathon runners. How many do we expect to finish the race?

*Solution:*

Let  $X$  be the number of runners who finish the Boston Marathon. Then we can say  $X \sim B(140,0.97)$ . Then the expected value is

$$\mu_X = E(X) = (140)(0.97) = 135.8 \text{ runners}$$

- 5. We roll a fair die 6 times. What is the probability we'll observe an even number in at most 3 of the rolls?



*Solution:*

Let  $X$  be the number of even numbers observed. Then we can say  $X \sim B(6,0.5)$ .

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$P(X \leq 3) = \binom{6}{0}(0.5)^0(1 - 0.5)^6 + \binom{6}{1}(0.5)^1(1 - 0.5)^5$$

$$+ \binom{6}{2}(0.5)^2(1 - 0.5)^4 + \binom{6}{3}(0.5)^3(1 - 0.5)^3$$

$$P(X \leq 3) = (0.5^6) + 6(0.5)(0.5^5) + 15(0.5^2)(0.5^4) + 20(0.5^3)(0.5^3)$$

$$P(X \leq 3) = 0.5^6 + 6(0.5^6) + 15(0.5^6) + 20(0.5^6)$$

$$P(X \leq 3) = 42(0.5^6)$$

$$P(X \leq 3) \approx 0.6563$$

- 6. We roll two fair 6-sided die 10 times and observe the sum. What is the probability of rolling a sum of 7 on at least six of the rolls?

*Solution:*

Let  $X$  be the number of times we roll a sum of 7. Since there are 36 possible rolls when we roll two die, and 6 of them result in a sum of 7,

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

the probability is

$$P(\text{sum of } 7) = \frac{6}{36} = \frac{1}{6}$$

Therefore we can express  $X$  as

$$X \sim B\left(10, \frac{1}{6}\right)$$

So the probability of rolling a sum of 7 at least 6 times out of 10 rolls is

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$

or

$$1 - P(X \leq 5)$$

Find  $P(X \leq 5)$ .

$$P(X \leq 5) = \binom{10}{0} \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^{10} + \binom{10}{1} \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^9$$

$$+ \binom{10}{2} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^8 + \binom{10}{3} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^7$$

$$+ \binom{10}{4} \left(\frac{1}{6}\right)^4 \left(1 - \frac{1}{6}\right)^6 + \binom{10}{5} \left(\frac{1}{6}\right)^5 \left(1 - \frac{1}{6}\right)^5$$

$$P(X \leq 5) = \left(\frac{5}{6}\right)^{10} + 10 \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^9$$

$$+ 45 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8 + 120 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7$$

$$+ 210 \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 + 252 \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^5$$

$$P(X \leq 5) \approx 0.9976$$

Then  $1 - P(X \leq 5)$  is

$$1 - P(X \leq 5) \approx 1 - 0.9976$$

$$1 - P(X \leq 5) \approx 0.0024$$

or about a 0.24 % chance.

## BERNOULLI RANDOM VARIABLES

- 1. A game at the local county fair involves spinning a circular spinner that's divided into 8 congruent sections, only two of which are "winners." We buy 5 spins for \$3.00. If we land on "winner" on any of our 5 spins, we get to choose a stuffed animal. Is this an example of Bernoulli trials?

*Solution:*

The set of 5 spins can be considered Bernoulli trials because the spins are independent of one another, there are exactly two outcomes (land on a winning, or not), and the probability of success (landing on a winner) remains constant for each trial at  $p = 2/8 = 1/4 = 0.25 = 25\%$ .

- 2. A game at the local county fair involves spinning a circular spinner that's divided into 8 congruent sections, only two of which are "winners." We buy 5 spins for \$3.00. If we land on "winner" on any of our 5 spins, we get to choose a stuffed animal. Find the mean and standard deviation for each trial.

*Solution:*



We already know that the probability of winning on any single spin is  $p = 2/8 = 1/4 = 0.25 = 25\%$ , which means  $\mu = p = 0.25$ . The standard deviation will therefore be

$$\sigma = \sqrt{p(1-p)} = \sqrt{(0.25)(1-0.25)} = \sqrt{0.1875} \approx 0.4330$$

- 3. A game at the local county fair involves spinning a circular spinner that's divided into 8 congruent sections, only two of which are "winners." We buy 5 spins for \$3.00. If we land on "winner" on any of our 5 spins, we get to choose a stuffed animal. Find the mean and standard deviation for the number of winners expected in a set of 5 spins.

*Solution:*

We already know that the probability of winning on any single spin is  $p = 2/8 = 1/4 = 0.25 = 25\%$ , which means  $\mu = p = 0.25$ . Therefore, for 5 spins the mean will be  $\mu = np = 5(0.25) = 1.25$ . And the standard deviation for 5 spins will be

$$\sigma = \sqrt{np(1-p)} = \sqrt{(5)(0.25)(1-0.25)} = \sqrt{0.9375} \approx 0.9682$$

- 4. A game at the local county fair involves spinning a circular spinner that's divided into 8 congruent sections, only two of which are "winners." We buy 5 spins for \$3.00. If we land on "winner" on any of our 5 spins, we



get to choose a stuffed animal. Find the probability of observing no winners in a set of 5 spins.

*Solution:*

The probability of spinning a winner is

$$P(\text{winner}) = p = \frac{2}{8} = \frac{1}{4} = 0.25$$

$$P(\text{non-winner}) = 1 - p = 1 - 0.25 = 0.75$$

Therefore, the probability of no winners in 5 spins is

$$P(\text{no winners in 5 spins}) = (0.75)^5 = 0.2373$$

- 5. A game at the local county fair involves spinning a circular spinner that's divided into 8 congruent sections, only two of which are "winners." We buy 5 spins for \$3.00. If we land on "winner" on any of our 5 spins, we get to choose a stuffed animal. What is the probability of observing at least 1 winner in a set of 5 spins?

*Solution:*

If we observe at least one winner out of 5 spins, that means we're looking for the probability of getting 1, 2, 3, 4, or 5 winners. The only result we're excluding is the probability of getting 0 winners. Which means we could



flip this problem around and calculate the probability of at least 1 winner as

$$P(W \geq 1) = 1 - P(W = 0)$$

$$P(W \geq 1) = 1 - 0.2373$$

$$P(W \geq 1) = 1 - 0.7627$$

- 6. Our goal is to learn about the percentage of students with high ACT scores. We randomly select high school seniors and record their highest ACT score. Explain why these aren't Bernoulli trials. Then design a way to conduct the experiment differently so that they can be considered Bernoulli trials.

*Solution:*

These are not Bernoulli trials because actual ACT scores are recorded. This is a random variable, but the variable can take on many different values, not simply “success” or “failure.”

To change the experiment so that we're running Bernoulli trials, we could define a specific range of ACT scores as “failures” and another range as “successes.” For instance, we could define a success as a score of 28 or higher, and a failure as a score lower than 28 (27 or lower).



Then the probability of a senior having a score of 28 or higher will have some constant probability of success from trial to trial, so we now have an experiment in which we're using Bernoulli trials.



## GEOMETRIC RANDOM VARIABLES

- 1. We toss a coin until we get “tails.” Does this experiment represent a geometric random variable? If it doesn’t, explain why. If it does, determine its parameter  $p$  and express the variable as  $X \sim \text{Geom}(p)$ .

*Solution:*

Yes, this experiment results in a geometric random variable. Let  $X$  be the number of trials it takes to get our first “tails.” We know that  $p = 0.5$  for each trial because there are two equally likely outcomes when we flip a coin. So  $X \sim \text{Geom}(0.5)$ .

- 2. We randomly select students from our school until we find a student in the school band. Assume there are 900 students in the school and 80 participate in the school band. Does this experiment represent a geometric random variable? If it doesn’t, explain why. If it does, determine its parameter  $p$  and express the variable as  $X \sim \text{Geom}(p)$ .

*Solution:*

Yes, this experiment results in a geometric random variable. We do have a fixed probability of success,



$$p = \frac{80}{900} = \frac{4}{45} \approx 0.09$$

and the trials can be considered independent because we have a large population. We're selecting students until we find someone in the band. Therefore  $X \sim \text{Geom}(0.09)$ .

- 3. Let  $X \sim \text{Geom}(p)$  with  $p = 0.25$ . Find  $P(X = 5)$ .

*Solution:*

We're being asked to find the probability that we get our first success on the 5th trial, if the probability of success on any single trial is  $p = 0.25$ .

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

$$P(X = 5) = (0.25)(1 - 0.25)^{5-1}$$

$$P(X = 5) = (0.25)(0.75)^4$$

$$P(X = 5) \approx 0.0791$$

- 4. Suppose we roll a 6-sided fair die until we observe a 2. What is the probability that a 2 will be observed within the first 5 trials?

*Solution:*



The probability of success on any single trial is  $p = 1/6$ , which means the probability of failure is  $1 - p = 1 - (1/6) = 5/6$ . Therefore, the probability that we get a 2 within the first 5 trials is

$$P(X \leq 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(X \leq 5) = \frac{1}{6} \left(\frac{5}{6}\right)^{1-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{2-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{3-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{4-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{5-1}$$

$$P(X \leq 5) = \frac{1}{6} \left(\frac{5}{6}\right)^0 + \frac{1}{6} \left(\frac{5}{6}\right)^1 + \frac{1}{6} \left(\frac{5}{6}\right)^2 + \frac{1}{6} \left(\frac{5}{6}\right)^3 + \frac{1}{6} \left(\frac{5}{6}\right)^4$$

$$P(X \leq 5) \approx 0.5981$$

- 5. Suppose we roll a 6-sided fair die until we observe a 2. What is the probability that a 2 won't be observed until at least the 6th trial?

*Solution:*

The probability of success on any single trial is  $p = 1/6$ , which means the probability of failure is  $1 - p = 1 - (1/6) = 5/6$ . Therefore, the probability that we get a 2 within the first 5 trials is

$$P(X \leq 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(X \leq 5) = \frac{1}{6} \left(\frac{5}{6}\right)^{1-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{2-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{3-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{4-1} + \frac{1}{6} \left(\frac{5}{6}\right)^{5-1}$$

$$P(X \leq 5) = \frac{1}{6} \left(\frac{5}{6}\right)^0 + \frac{1}{6} \left(\frac{5}{6}\right)^1 + \frac{1}{6} \left(\frac{5}{6}\right)^2 + \frac{1}{6} \left(\frac{5}{6}\right)^3 + \frac{1}{6} \left(\frac{5}{6}\right)^4$$

$$P(X \leq 5) \approx 0.5981$$

Therefore, the probability that we don't observe a success until the 6th trial or later is

$$P(X \geq 6) \approx 1 - 0.5981$$

$$P(X \geq 6) \approx 0.4019$$

- 6. According to a 2017-2018 survey, 68 % of U.S. households own a pet. Suppose we start randomly surveying households and asking whether they are pet owners. How many do we expect we will need to survey to find our first household that owns a pet?

*Solution:*

Let  $X$  be the trial when we find our first pet owner. We know that  $X$  is a geometric random variable with  $X \sim \text{Geom}(0.68)$ . Then the expected value is

$$\mu_X = E(X) = \frac{1}{p} = \frac{1}{0.68} \approx 1.471$$

So we could say that we expect we'll need to survey somewhere between 1 and 2 households in order to find our first pet-owning household.



## TYPES OF STUDIES

- 1. The following table shows the age and shoe size of six children. Does the data have a positive correlation, negative correlation, or no correlation?

Age	Shoe size
3	7
3	6
5	9
6	12
6	11
7	13

*Solution:*

The data has a positive correlation because, as the age of the child increases, so does the size of shoe. Positive correlation occurs when two variables increase or decrease together, negative correlation occurs when one variable increases while the other decreases, and no correlation would have no discernible pattern.

- 2. A class conducts a survey and finds that 75 % of the school spends 2 or more hours on social media each day. Would the data fit into a one-way or two-way table? Is the study observational or experimental?

*Solution:*

The survey only shows data for one variable for a set of individuals, the amount of time spent on social media, so the data fits into a one-way table. The survey is an observational study because it records the results without manipulation.

- 3. The following table shows the number of classes from which students were absent and their final grade in the class. Does the data have a positive correlation, negative correlation, or no correlation?

Number of absences	0	0	1	2	3	3	3	5	5	6	7	10
Final grade	95%	97%	90%	86%	80%	74%	70%	65%	64%	58%	55%	45%

*Solution:*

The data has a negative correlation because, as the number of absences increases, the final grade in the class decreases. Positive correlation occurs when two variables increase or decrease together, negative correlation occurs when one variable increases while the other decreases, and no correlation would have no discernible pattern.



- 4. The table below shows the favorite winter activity of 50 adults. Is it a one-way data table? Why or why not?

	Skiing	Snowboarding	Ice Skating
Men	9	13	6
Women	8	7	7

*Solution:*

This is a two-way data table because we have the two categories of individuals: men and women, and the three categories of activities: skiing, snowboarding, and ice skating. We can use this data to examine the relationship between the two categorical variables.

- 5. Is the following experiment an example of a double-blind experiment? If not, what could be changed to make it a double-blind experiment?

“A soda company has developed a new flavor and wants to know how it compares in taste to competitor sodas. An employee of the soda company conducts a survey where participants are asked which soda tastes the best. The sodas are given to participants in unmarked plastic cups by the employee.”

*Solution:*

This experiment is an example of a blind experiment since the participants don't know which soda is being targeted. However, it's not a double-blind experiment since the employee of the soda company, who is also administering the survey, knows which soda is being targeted. To make it a double-blind experiment, the employee conducting the survey should have the sodas prepared by someone else so that neither the participants nor the employee administering the experiment know which soda is being targeted.

- 6. A new cancer drug is being used to treat cancer in children and adults. The hospital conducts a study to measure the effectiveness of the new drug. Cancer patients are placed into groups according to their age and each age range is split into two groups. One group is given traditional treatment of the cancer and the other group is given the new drug. Will the data fit into a one-way or two-way table? Is the study observational or experimental?

*Solution:*

The data fits into a two-way table because there's a control group and an experimental group, grouped according to age, and the data is about the effectiveness of the drug. It's an experimental study because the experimental group is being manipulated by receiving the new drug.



## SAMPLING AND BIAS

- 1. The zoo conducts a survey on why patrons enjoy coming to the zoo. They ask families with children about why they like to visit the zoo as they're leaving. Give a reason why the sampling method may be biased.

*Solution:*

The sampling method is selection biased since the zoo is only surveying families with children. An unbiased sampling method would include all zoo patrons. For example, the zoo could survey every 10th customer as they leave.

- 2. The owner of a restaurant gives a survey to each customer. Included in the survey is the question “Have you ever not tipped your waiter or waitress?” Give a reason why the sampling method may be biased.

*Solution:*

The sampling method is response biased because some people may not want to answer the question about tipping truthfully. This is also called “social desirability bias.” There might be less of a response bias if the wording were changed to, “Is there ever a circumstance where it’s acceptable to not tip your waiter or waitress?”



- 3. A health club wants to purchase a new machine and would like to know which machine members would most like to have. It creates a survey where members can rate the different machines that the health club is considering purchasing, and posts it at the reception desk for members to fill out if they choose to do so. Does the sample contain a bias? If so, what kind?

*Solution:*

The sampling method is biased because of voluntary response sampling. People who voluntarily participate in the survey may have different habits, opinions, or tendencies than people who choose not to participate.

- 4. A biologist wants to study a group of prairie dogs for parasites, but cannot examine the entire population. Which sampling method would be better in this case, a stratified random sample or a clustered random sample?

*Solution:*

A clustered random sample would be better. The biologist could divide the field into different sections and take a random sample from each section. This would give the biologist a representative sample of the entire



population. A stratified random sample would separate the prairie dogs by gender, age, or some other variable, and the results might vary based on those values.

- 5. A hospital is studying the health effects of obesity. They group patients into different groups according to a specific weight range and study a variety of biometrics. What type of sampling is this?

*Solution:*

The sampling method is a stratified random sample because people are the same weight range within each group. A simple random sample would study a group of people picked randomly with no regards to weight range. A clustered random sample might select a random sampling of people from each wing of the hospital.

- 6. A museum wants to find out the demographics of its patrons. They set up a survey and ask every 5th customer about their age, ethnicity, and gender. What type of sampling is this?

*Solution:*

This sampling method is systematic sampling. Patrons are randomly selected with no regard to groups or clusters.



## SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

- 1. The population of 32 year-old women in the United States have an average salary of \$42,000, but the distribution of their salaries is not normally distributed. A random sample of 24 women is taken. Does the sample meet the criteria to use the central limit theorem?

*Solution:*

Our sample space should be no more than 10% of our population, the sample should be selected randomly, and (if the population is not normal) the sample size must be greater than 30.

The sample is random, 24 is definitely less than 10% of all 32 year-old women in the United States, but 24 isn't greater than 30 and the population is not normal. So the sample does not meet the criteria to use the central limit theorem.

- 2. There are 130 dogs at a dog show who weigh an average of 11 pounds with a standard deviation of 3 pounds. A sample of 9 dogs is taken. What is the standard deviation of the sampling distribution of the sample mean?

*Solution:*



Find the standard deviation of the sampling distribution of the sample mean using  $\sigma = 3$  and  $n = 9$ , making sure to use the finite population correction factor in the standard deviation formula.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\bar{x}} = \frac{3}{\sqrt{9}} \sqrt{\frac{130-9}{130-1}}$$

$$\sigma_{\bar{x}} = \frac{3}{3} \sqrt{\frac{121}{129}}$$

$$\sigma_{\bar{x}} \approx 0.9685$$

- 3. A large university population has an average student age of 30 years old with a standard deviation of 5 years, and student age is normally distributed. A sample of 80 students is randomly taken. What is the probability that the mean of their ages will be less than 29?

*Solution:*

Our sample space should be no more than 10% of our population, the sample should be selected randomly, and (if the population is not normal) the sample size must be greater than 30.



The sample was collected randomly. It's safe to assume that 80 students is less than 10 % of the student population at a large university. The population is normal, so the sample size doesn't have to be greater than 30, but 80 is greater than 30 anyway. The sample space meets the conditions of normality.

Find the standard deviation of the sampling distribution of the sample mean.

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

$$\sigma_{\bar{x}} = \frac{5}{\sqrt{80}}$$

$$\sigma_{\bar{x}} \approx 0.559$$

We want to know the probability that the sample mean  $\bar{x}$  is less than 29. We need to express this in terms of standard deviations.

$$z = \frac{29 - 30}{0.559} = \frac{-1}{0.559} \approx -1.79$$

This means we want to know the probability of  $P(z < -1.79)$ . Using a  $z$ -table, a  $z$ -value of  $-1.79$  gives 0.0367, so  $P(z < -1.79) = 3.67\%$ . There's a 3.67 % chance that our sample mean will be less than 29.

- 4. A cereal company packages cereal in 12.5-ounce boxes with a standard deviation of 0.5 ounces. The amount of cereal put into each box is normally distributed. The company randomly selects 100 boxes to check



their weight. What is the probability that the mean weight will be greater than 12.6 ounces?

*Solution:*

Our sample space should be no more than 10% of our population, the sample should be selected randomly, and (if the population is not normal) the sample size must be greater than 30.

The sample was collected randomly. It's safe to assume that 100 boxes is less than 10% of the cereal boxes in the factory. The population is normal so the sample size doesn't have to be greater than 30, but 100 is greater than 30 anyway. The sample space meets the conditions of normality.

Find the standard deviation of the sampling distribution of the sample mean.

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

$$\sigma_{\bar{x}} = \frac{0.5}{\sqrt{100}}$$

$$\sigma_{\bar{x}} = 0.05$$

We want to know the probability that the sample mean  $\bar{x}$  is more than 12.6 ounces. We need to express this in terms of standard deviations.

$$z = \frac{12.6 - 12.5}{0.05} = \frac{0.1}{0.05} = 2$$



This means we want to know the probability of  $P(z > 2)$ .

Using the  $z$ -table, a  $z$ -value of 2 gives 0.9772, but we need to subtract this from 1 to find the probability that the sample mean is more than 12.6 ounces.

$$P(z > 2) = 1 - 0.9772$$

$$P(z > 2) = 0.0228$$

$$P(z > 2) = 2.28 \%$$

There's a 2.28% chance that our sample mean will be greater than 12.6 ounces.

- 5. A large hospital finds that the average body temperature of their patients is  $98.4^\circ$ , with a standard deviation of  $0.6^\circ$ , and we'll assume that body temperature is normally distributed. The hospital randomly selects 30 patients to check their temperature. What is the probability that the mean temperature of these patients  $\bar{x}$  is within  $0.2^\circ$  of the population mean?

*Solution:*

Our sample space should be no more than 10% of our population, the sample should be selected randomly, and (if the population is not normal) the sample size must be greater than 30.



The sample was collected randomly. It's safe to assume that 30 patients is less than 10% of the total patients in a large hospital. The population is normal so the sample size doesn't have to be greater than 30. The sample space meets the conditions of normality.

Find the standard deviation of the sampling distribution of the sample mean.

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

$$\sigma_{\bar{x}} = \frac{0.6}{\sqrt{30}}$$

$$\sigma_{\bar{x}} \approx 0.1095$$

We want to know the probability that the sample mean  $\bar{x}$  is within 0.2° of the population mean. We need to express 0.2° in terms of standard deviations.

$$\frac{0.2}{0.1095} \approx 1.83$$

This means we want to know the probability of  $P(-1.83 < z < 1.83)$ .

Using a  $z$ -table, a  $z$ -value of  $-1.83$  gives  $0.0336$  and a value of  $1.83$  gives  $0.9664$ . The probability under the normal curve between these  $z$ -scores is

$$P(-1.83 < z < 1.83) = 0.9664 - 0.0336$$

$$P(-1.83 < z < 1.83) = 0.9328$$

$$P(-1.83 < z < 1.83) = 93.28\%$$

There's a 93.28% chance that our sample mean will fall within  $0.2^\circ$  of the population mean of  $98.4^\circ$ .

- 6. A company produces volleyballs in a factory. Individual volleyballs are filled to an approximate pressure of 7.9 PSI (pounds per square inch), with a standard deviation of 0.2 PSI. Air pressure in the volleyballs is normally distributed. The company randomly selects 50 volleyballs to check their pressure. What is the probability that the mean amount of pressure in these balls  $\bar{x}$  is within 0.05 PSI of the population mean?

*Solution:*

Our sample space should be no more than 10% of our population, the sample should be selected randomly, and (if the population is not normal) the sample size must be greater than 30.

The sample was collected randomly. It's safe to assume that 50 volleyballs is less than 10% of all the volleyballs produced in the factory. The population is normal so the sample size doesn't have to be greater than 30, but 50 is greater than 30 anyway. The sample space meets the conditions of normality.

Find the standard deviation of the sampling distribution of the sample mean.



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

$$\sigma_{\bar{x}} = \frac{0.2}{\sqrt{50}}$$

$$\sigma_{\bar{x}} \approx 0.02828$$

We want to know the probability that the sample mean  $\bar{x}$  is within 0.05 PSI of the population mean. We need to express 0.05 in terms of standard deviations.

$$\frac{0.05}{0.02828} \approx 1.77$$

This means we want to know the probability of  $P(-1.77 < z < 1.77)$ .

Using a  $z$ -table, a  $z$ -value of  $-1.77$  gives 0.0384 and a  $z$ -value of  $1.77$  gives 0.9616. The probability under the normal curve between these  $z$ -scores is

$$P(-1.77 < z < 1.77) = 0.9616 - 0.0384$$

$$P(-1.77 < z < 1.77) = 0.9232$$

$$P(-1.77 < z < 1.77) = 92.32 \%$$

There's a 92.32 % chance that our sample mean will fall within 0.05 PSI of the population mean of 7.9 PSI.



## CONDITIONS FOR INFERENCE WITH THE SDSM

- 1. There are 1,000 students at our school, and we ask 150 of them to tell us their height as they exit school at the end of the day. Have we met the conditions for inference?

*Solution:*

Because we're asking students as they exit the school building at the end of the day, we're not taking a truly random sample, so we're violating the random condition.

We're also sampling with replacement, and whenever we sample without replacement, we have to keep the sample size below 10% of the population. Because 150 is greater than 10% of the population, we violate the independent condition as well.

- 2. We randomly sample 400 boxes (with replacement) in a very large, national shipping warehouse and record their weight in kilograms. Have we met the conditions for inference?

*Solution:*



We've taken a random sample with replacement, which means we've met the random and independent conditions. We don't know whether the population is normally distributed, but our sample size is large,  $n \geq 30$ , so we've met the normal condition as well, and we can move forward with using our data to answer probability questions.

3. A cookie company makes packages of cookies, where the weight of the packages is normally distributed with  $\mu = 500$  grams and  $\sigma = 4$  grams. If the cookie company's production manager randomly selects 100 packages of cookies, what is the probability that the sample mean is within 7.5 grams of the population mean?

*Solution:*

The production manager takes a random sample, and the sample size is 100 cookie packages, so we've met the random and normal conditions for inference. The question suggests that the manager samples without replacement, but we can safely assume that 100 packages is less than 10% of the total number of packages that the company produces, which means we've met the independence condition as well.

The mean of the sampling distribution of the sample mean will be  $\mu_{\bar{x}} = \mu = 500$ , and the standard error will be

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

$$\sigma_{\bar{x}} = \frac{4}{\sqrt{100}}$$

$$\sigma_{\bar{x}} = \frac{4}{10}$$

$$\sigma_{\bar{x}} = 0.4$$

We want to know the probability that the sample mean  $\bar{x}$  is within 7.5 grams of the population mean,  $\mu = 500$ . A 7.5 interval around 500 gives us the interval 492.5 to 507.5, so

$$P(492.5 < \bar{x} < 507.5) = P\left(\frac{492.5 - 500}{0.4} < z_{\bar{x}} < \frac{507.5 - 500}{0.4}\right)$$

$$P(492.5 < \bar{x} < 507.5) = P\left(-\frac{7.5}{0.4} < z_{\bar{x}} < \frac{7.5}{0.4}\right)$$

$$P(492.5 < \bar{x} < 507.5) = P(-18.75 < z_{\bar{x}} < 18.75)$$

The  $z$ -score  $z = 18.75$  is far higher than the largest value in the standard  $z$ -table, which means the probability under the normal curve between  $z = -18.75$  and  $z = 18.75$  is virtually 100 %.

Therefore, it's almost guaranteed that the sample mean will fall within 7.5 grams of the population mean  $\mu = 500$ .

- 4. A sushi chef builds a sushi roll approximately every 3 minutes, with a standard deviation of 15 seconds, every night while his restaurant is open between 5 : 00 p.m. and 10 : 00 p.m., Tuesday through Sunday. The time



spent to build sushi rolls is normally distributed. If the chef takes a random sample of 20 sushi rolls over the course of a week, what is the probability that the sample mean is within 5 seconds of the population mean?

*Solution:*

The chef takes a random sample, and the population is normally distributed, so we've met the random and normal conditions.

The restaurant is open 30 hours each week (5 hours per night for 6 nights per week), and the chef builds approximately  $60 \div 3 = 20$  sushi rolls per hour, so he builds about  $30 \cdot 20 = 600$  sushi rolls per week. The 20-roll sample is well below 10% of the population, so we've met the independent condition as well.

The mean of the sampling distribution of the sample mean will be  $\mu_{\bar{x}} = \mu = 3$  minutes (or 180 seconds). We'll calculate everything in seconds, and the standard error will be

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

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{20}}$$

$$\sigma_{\bar{x}} \approx 3.354$$



We want to know the probability that the sample mean  $\bar{x}$  is within 5 seconds of the population mean, 180. A 5-second interval around 180 gives us the interval 175 to 185, so

$$P(175 < \bar{x} < 185) = P\left(\frac{175 - 180}{3.354} < z_{\bar{x}} < \frac{185 - 180}{3.354}\right)$$

$$P(175 < \bar{x} < 185) = P\left(-\frac{5}{3.354} < z_{\bar{x}} < \frac{5}{3.354}\right)$$

$$P(175 < \bar{x} < 185) \approx P(-1.49 < z_{\bar{x}} < 1.49)$$

In the  $z$ -table, a  $z$ -value of 1.49 gives 0.9319,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441

and a  $z$ -value of  $-1.49$  gives 0.0681.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823

Which means the probability under the normal curve between these  $z$ -scores is

$$P(-1.49 < z_{\bar{x}} < 1.49) = 0.9319 - 0.0681$$

$$P(-1.49 < z_{\bar{x}} < 1.49) = 0.8638$$

$$P(-1.49 < z_{\bar{x}} < 1.49) = 86.4 \%$$

So there's an approximately 86.4 % chance that the mean  $\bar{x}$  of the 20-roll sample the chef takes will fall within 5 seconds of the population mean of  $\mu = 180$  seconds.

- 5. The time spent playing video games by competitive gamers is normally distributed with  $\mu = 40$  hours per week, and  $\sigma = 2.5$  hours. If we take a random sample with replacement of 75 players and record the number of hours they spend playing this week, what's the probability that our sample mean is within 30 minutes of the population mean?

*Solution:*

We take a random sample, and the population is normally distributed, so we've met the random and normal conditions. We sample with replacement, which means we've meet the independent condition as well.

The mean of the sampling distribution of the sample mean will be  $\mu_{\bar{x}} = \mu = 40$  hours, and the standard error will be

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

$$\sigma_{\bar{x}} = \frac{2.5}{\sqrt{75}}$$

$$\sigma_{\bar{x}} \approx 0.289$$

We want to know the probability that the sample mean  $\bar{x}$  is within 30 minutes (0.5 hours) of the population mean, 40. A 0.5-hour interval around 40 gives us the interval 39.5 to 40.5, so

$$P(39.5 < \bar{x} < 40.5) = P\left(\frac{39.5 - 40}{0.289} < z_{\bar{x}} < \frac{40.5 - 40}{0.289}\right)$$

$$P(39.5 < \bar{x} < 40.5) = P\left(-\frac{0.5}{0.289} < z_{\bar{x}} < \frac{0.5}{0.289}\right)$$

$$P(39.5 < \bar{x} < 40.5) \approx P(-1.73 < z_{\bar{x}} < 1.73)$$

In the  $z$ -table, a  $z$ -value of 1.73 gives 0.9582,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706

and a  $z$ -value of  $-1.73$  gives 0.0418.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455

Which means the probability under the normal curve between these  $z$ -scores is

$$P(-1.73 < z_{\bar{x}} < 1.73) = 0.9582 - 0.0418$$

$$P(-1.73 < z_{\bar{x}} < 1.73) = 0.9164$$

$$P(-1.73 < z_{\bar{x}} < 1.73) \approx 91.6\%$$

So there's an approximately 91.6% chance that the mean  $\bar{x}$  of the 75-player sample we take will fall within 30 minutes of the population mean of  $\mu = 40$  hours.

- 6. The time it takes for a roofing company to install a new roof on a single-story house normally distributed with  $\mu = 6$  hours and  $\sigma = 1$  hour. If the company's owner takes a random sample with replacement of 10 roofing jobs, what's the probability that his sample mean is within 45 minutes of the population mean?

*Solution:*

The company's owner takes a random sample, and the population is normally distributed, so he's met the random and normal conditions. Because he's sampling with replacement, he meets the independent condition as well.

The mean of the sampling distribution of the sample mean will be  $\mu_{\bar{x}} = \mu = 6$  hours, and the standard error will be

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

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{10}}$$



$$\sigma_{\bar{x}} \approx 0.316$$

The owner wants to know the probability that the sample mean  $\bar{x}$  is within 45 minutes (0.75 hours) of the population mean, 6. A 0.75-hour interval around 6 gives us the interval 5.25 to 6.75, so

$$P(5.25 < \bar{x} < 6.75) = P\left(\frac{5.25 - 6}{0.316} < z_{\bar{x}} < \frac{6.75 - 6}{0.316}\right)$$

$$P(5.25 < \bar{x} < 6.75) = P\left(-\frac{0.75}{0.316} < z_{\bar{x}} < \frac{0.75}{0.316}\right)$$

$$P(5.25 < \bar{x} < 6.75) \approx P(-2.37 < z_{\bar{x}} < 2.37)$$

In the  $z$ -table, a  $z$ -value of 2.37 gives 0.9911,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	<b>.9911</b>	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

and a  $z$ -value of  $-2.37$  gives 0.0089.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	<b>.0089</b>	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

Which means the probability under the normal curve between these  $z$ -scores is

$$P(-2.37 < z_{\bar{x}} < 2.37) = 0.9911 - 0.0089$$

$$P(-2.37 < z_{\bar{x}} < 2.37) = 0.9822$$

$$P(-2.37 < z_{\bar{x}} < 2.37) \approx 98.2\%$$

So there's an approximately 98.2% chance that the mean  $\bar{x}$  of the sample of 10 roofing jobs the owner takes will fall within 45 minutes of the population mean of  $\mu = 6$  hours.



## SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

- 1. The state representatives want to know how their constituents feel about the new tax to fund road improvements, so they send out a survey. Of the 5 million who reside in the state, 150,000 people respond. 40 % disapprove of the new tax and 60 % are in favor of the new tax because of the improvements they've seen to the roads. Does this sample meet the conditions for inference?

*Solution:*

Our sample space should be no more than 10 % of our population, the expected number of successes and failures should each be at least 10, and the sample should be selected randomly.

The sample space was random, but may have a bias since it was a voluntary sample. The sample space is no more than 10 % of the population:

$$\frac{150,000}{5,000,000} = 0.03 = 3 \% \leq 10 \%$$

And there are more than 10 expected successes and failures.

$$150,000(0.6) = 90,000 \geq 10$$

$$150,000(0.4) = 60,000 \geq 10$$

The sample space meets the conditions for inference. However, the voluntary bias should be noted and the direction of bias taken into account.

- 2. An ice cream shop states that only 5% of their 1,200 customers order a sugar cone. We want to verify this claim, so we randomly select 120 customers to see if they order a sugar cone. Does this sample meet the conditions for inference?

*Solution:*

Our sample space should be no more than 10% of our population, the expected number of successes and failures should each be at least 10, and the sample should be selected randomly.

The sample space was random, and was no more than 10% of the population:

$$\frac{120}{1,200} = 0.1 = 10\% \leq 10\%$$

But there are not at least 10 expected successes and failures.

$$120(0.05) = 6 \not\geq 10$$

$$120(0.95) = 114 \geq 10$$



The sample space doesn't meet the conditions for inference because the success of a customer ordering a sugar cone is 6, which is less than 10.

- 3. The zoo conducts a study about the demographics of its patrons, and wants to learn about how many groups that visit the zoo bring children under age 12. Every 10th customer or group is recorded as a “family,” and classified as either “including children under 12” or “not including children under 12.” The zoo collected data on 65 families, and 45 of them are classified as “not including children under 12.” That day, 650 families came to the zoo. What is the standard error of the sampling distribution of the sample proportion?

*Solution:*

Our sample space should be no more than 10% of our population, the expected number of successes and failures should each be at least 10, and the sample should be selected randomly.

The sample space was random, and was no more than 10% of the population:

$$\frac{65}{650} = 0.1 = 10\% \leq 10\%$$

The “success” rate was  $45/65 = \approx 0.69$ , which means the failure rate was  $1 - 0.69 \approx 0.31$ . Which means there are more than 10 expected successes and failures.



$$65(0.69) = 45 \geq 10$$

$$65(0.31) = 20 \geq 10$$

We've met the conditions for inference, so we'll identify the sample size  $n = 65$  and the population proportion as

$$\hat{p} = \frac{45}{65} = 0.69$$

Now we can calculate the standard error of the proportion, remembering to add in the finite population correction factor, since we have a finite population of 650 families.

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \sqrt{\frac{N - n}{N - 1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.69(1 - 0.69)}{65}} \sqrt{\frac{650 - 65}{650 - 1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.69(0.31)}{65}} \sqrt{\frac{585}{649}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2139 \cdot 585}{65 \cdot 649}}$$

$$\sigma_{\hat{p}} \approx 0.054463$$

- 4. A pizza shop finds that 80 % of the 75 randomly selected pizzas ordered during the week have pepperoni. What is the standard error of



the proportion if the pizza shop has a total of 1,000 pizzas ordered during the week?

*Solution:*

Our sample space should be no more than 10 % of our population, the expected number of successes and failures should each be at least 10, and the sample should be selected randomly.

The sample space was random, and was no more than 10 % of the population:

$$\frac{75}{1,000} = 0.075 = 7.5 \% \leq 10 \%$$

And there are more than 10 expected successes and failures.

$$75(0.8) = 60 \geq 10$$

$$75(0.2) = 15 \geq 10$$

We've met the conditions for inference, so we'll identify the sample size  $n = 75$  and the population proportion as  $p = 0.8$ . Now we can calculate standard error of the proportion, remembering to add in the finite population correction factor, since we have a finite population of 1,000 families.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.8(1 - 0.8)}{75}} \sqrt{\frac{1,000 - 75}{1,000 - 1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.8(0.2)}{75}} \sqrt{\frac{925}{999}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.16 \cdot 925}{75 \cdot 999}}$$

$$\sigma_{\hat{p}} \approx 0.044444$$

- 5. A hospital conducts a survey and finds that 10 patients of 30 who are randomly selected on a given day have high blood pressure. There were 325 patients in the hospital that day. What is the standard error of the proportion?

*Solution:*

Our sample space should be no more than 10% of our population, the expected number of successes and failures should each be at least 10, and the sample should be selected randomly.

The sample space was random, and was no more than 10% of the population:

$$\frac{30}{325} \approx 0.0923 \approx 9.23\% \leq 10\%$$

And there are more than 10 expected successes and failures.

$$30(0.33) \approx 10 \geq 10$$

$$30(0.67) \approx 20 \geq 10$$

We've met the conditions for inference, so we'll identify the sample size  $n = 30$  and the population proportion as

$$p = \frac{10}{30} \approx 0.33$$

Now we can calculate standard error of the proportion, remembering to add in the finite population correction factor, since we have a finite population of 325 patients.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.33(1-0.33)}{30}} \sqrt{\frac{325-30}{325-1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.33(0.67)}{30}} \sqrt{\frac{295}{324}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2211 \cdot 295}{30 \cdot 324}}$$

$$\sigma_{\hat{p}} \approx 0.081917$$



6. A study claims that first-born children are more likely to become leaders. The study finds that 72% of 2,000 first-born children are currently in or have held leadership roles in their careers. Another group of scientists wants to verify the claim, but can't survey all 2,000 people, so they randomly sample 175 of the participants. What is the probability that their results are within 2% of the first study's claim?

*Solution:*

Our sample space should be no more than 10% of our population, the expected number of successes and failures should each be at least 10, and the sample should be selected randomly.

The sample space was random, and was no more than 10% of the population:

$$\frac{175}{2,000} = 0.0875 = 8.75\% \leq 10\%$$

And there are more than 10 expected successes and failures.

$$175(0.72) = 126 \geq 10$$

$$175(0.28) = 49 \geq 10$$

We've met the conditions for inference. The original study found the population proportion to be  $p = 72\%$ . So the standard error of the proportion will be



$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.72(0.28)}{175}} \sqrt{\frac{2,000 - 175}{2,000 - 1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2016}{175}} \sqrt{\frac{1,825}{1,999}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2016 \cdot 1,825}{175 \cdot 1,999}}$$

$$\sigma_{\hat{p}} \approx 0.03243$$

We need to find the probability that our results are within 2% of the population proportion  $p = 72\%$ . This means, how likely is it that the mean of the sampling distribution of the sample proportion falls between 70% and 74%? We need to express 2% in terms of standard deviations:

$$\frac{0.02}{0.03243} \approx 0.616713$$

This means we want to know the probability of  $P(-0.62 < z < 0.62)$ . Using a  $z$ -table,  $-0.62$  gives us 0.2676 and  $0.62$  gives us 0.7324, so the probability is

$$P(-0.62 < z < 0.62) = 0.7324 - 0.2676$$

$$P(-0.62 < z < 0.62) = 0.4648$$

There's a 46.48% chance that our sample proportion will fall within 2% of the first study's claim.



## CONDITIONS FOR INFERENCE WITH THE SDSP

- 1. A gym owner takes a random sample of 10 local fitness instructors and asks them whether or not they train clients at multiple gyms. He finds that  $\hat{p} = 70\%$  of them report training clients at multiple gyms. Can he proceed with a hypothesis test?

*Solution:*

With a sample of only 10 instructors, the gym owner finds  $n\hat{p} = 10(0.7) = 7$  “successes,” and  $n(1 - \hat{p}) = 10(0.3) = 3$  “failures” in his sample. He needs at least 5 successes and at least 5 failures in order to guarantee normality. Therefore, he fails to meet the normal condition, and can’t proceed with a hypothesis test.

- 2. A professional basketball player makes 87.5% of his free throws. If he takes a random sample (without replacement) of 100 of his own free throws, can he move forward with a hypothesis test?

*Solution:*

We’re told that the player takes a random sample, so we assume he’s met the random condition. With the given population proportion of  $p = 0.875$ , we can expect  $np = 100(0.875) = 87.5$  “successes” and



$n(1 - p) = 100(0.125) = 12.5$  “failures.” Both of these values are greater than 5, so the player has met the normal condition.

Finally, even though the player samples without replacement, we can assume that the population of all his free throws is significantly larger than 10 times his sample size, and he therefore meets the independent condition.

Because he’s met all three conditions for inference, he can proceed with a hypothesis test.

- 3. If the basketball player from the previous question finds a sample proportion  $\hat{p} = 0.85$  in his sample of 100 free throws, calculate his test statistic. Remember that  $p = 0.875$ .

*Solution:*

When he runs a hypothesis test, the player will calculate his test statistic as

$$z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z_{\hat{p}} = \frac{0.85 - 0.875}{\sqrt{\frac{0.875(1 - 0.875)}{100}}}$$



$$z_{\hat{p}} = \frac{-0.025}{\sqrt{\frac{0.875(0.125)}{100}}}$$

$$z_{\hat{p}} = -0.025 \sqrt{\frac{100}{0.875(0.125)}}$$

$$z_{\hat{p}} = -0.025 \left( \frac{10}{\sqrt{0.109375}} \right)$$

$$z_{\hat{p}} \approx -0.76$$

- 4. A grocery chain claims that 75% of their customers say that they are “satisfied” with their local store. We want to verify this claim, so we take a random sample of 45 of their customers and ask them whether or not they are “satisfied.” How likely is it that our results are within 2% of the chain’s claim?

*Solution:*

The sample size is  $n = 45$  and  $p = 0.75$ , which means  $1 - p = 0.25$ , so

$$np = 45(0.75) = 33.75 \geq 5$$

$$n(1 - p) = 45(1 - 0.75) = 45(0.25) = 11.25 \geq 5$$

and we’ve verified normality. We were told in the question that our sample was random, and our sample  $n = 45$  is certainly less than 10% of the total



population of the grocery chain's customers, so we're not violating the 10% rule. With the conditions for inference satisfied, we can calculate the mean and standard deviation of the sampling distribution of the sample proportion. The mean is

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

$$\mu_{\hat{p}} = 0.75$$

The standard deviation of the sampling distribution (the standard error) is

$$\sigma_{\hat{p}} = \sqrt{\frac{0.75(1 - 0.75)}{45}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.75(0.25)}{45}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.1875}{45}}$$

$$\sigma_{\hat{p}} \approx 0.0645$$

The question asks us for the probability that our sample proportion is within 2% of the population proportion  $p = 75\%$ . In other words, how likely is it that the sample proportion falls between 73% and 77%?

$$P(0.73 < \hat{p} < 0.77) \approx P\left(\frac{0.73 - 0.75}{0.0645} < z < \frac{0.77 - 0.75}{0.0645}\right)$$

$$P(0.73 < \hat{p} < 0.77) \approx P\left(-\frac{0.02}{0.0645} < z < \frac{0.02}{0.0645}\right)$$

$$P(0.73 < \hat{p} < 0.77) \approx P(-0.31 < z < 0.31)$$

In a  $z$ -table, a  $z$ -value of 0.31 gives 0.6217,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879

and a value of  $-0.31$  gives 0.3783.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859

Which means the probability under the normal curve of the sampling distribution of the sample proportion between these  $z$ -scores is

$$P(-0.31 < z < 0.31) \approx 0.6217 - 0.3783$$

$$P(-0.31 < z < 0.31) \approx 0.2434$$

$$P(-0.31 < z < 0.31) \approx 24.34\%$$

Which means there's an approximately 24% chance that our sample proportion will fall within 2% of the grocery chain's claim.

- 5. A professional pickleball player claims that he wins 60% of the points he plays in championship matches. We want to verify this claim, so we take a random sample of 25 of his points in championship matches and



record whether or not he wins each point. How likely is it that our results are within 5% of the player's claim?

*Solution:*

The sample size is  $n = 25$  and  $p = 0.6$ , which means  $1 - p = 0.4$ , so

$$np = 25(0.6) = 15 \geq 5$$

$$n(1 - p) = 25(1 - 0.6) = 25(0.4) = 10 \geq 5$$

and we've verified normality. We were told in the question that our sample was random, and our sample  $n = 25$  is certainly less than 10% of the total population of points played in championship matches by the professional player, so we're not violating the 10% rule. With the conditions for inference satisfied, we can calculate the mean and standard deviation of the sampling distribution of the sample proportion. The mean is

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

$$\mu_{\hat{p}} = 0.6$$

The standard deviation of the sampling distribution (the standard error) is

$$\sigma_{\hat{p}} = \sqrt{\frac{0.6(1 - 0.6)}{25}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.6(0.4)}{25}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.24}{25}}$$

$$\sigma_{\hat{p}} \approx 0.0980$$

The question asks us for the probability that our sample proportion is within 5% of the population proportion  $p = 60\%$ . In other words, how likely is it that the sample proportion falls between 55% and 65%?

$$P(0.55 < \hat{p} < 0.65) \approx P\left(\frac{0.55 - 0.6}{0.0980} < z < \frac{0.65 - 0.6}{0.0980}\right)$$

$$P(0.55 < \hat{p} < 0.65) \approx P\left(-\frac{0.05}{0.0980} < z < \frac{0.05}{0.0980}\right)$$

$$P(0.55 < \hat{p} < 0.65) \approx P(-0.5102 < z < 0.5102)$$

$$P(0.55 < \hat{p} < 0.65) \approx P(-0.51 < z < 0.51)$$

In a  $z$ -table, a  $z$ -value of 0.51 gives 0.6950,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	<b>.6950</b>	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

and a value of  $-0.51$  gives 0.3050.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	<b>.3050</b>	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121

Which means the probability under the normal curve of the sampling distribution of the sample proportion between these  $z$ -scores is

$$P(-0.51 < z < 0.51) \approx 0.6950 - 0.3050$$

$$P(-0.51 < z < 0.51) \approx 0.39$$

$$P(-0.51 < z < 0.51) \approx 39\%$$

Which means there's an approximately 39% chance that our sample proportion will fall within 5% of the pickleball player's claim.

- 6. A company reports that the proportion of its invoices that get paid on time is  $p = 35\%$ . A clerk on the Accounts Receivable team wants to verify this claim, so she takes a random sample of 80 invoices and records whether or not they were paid on time. How likely is it that her sample proportion will fall within 10% of the company's claim?

*Solution:*

The sample size is  $n = 80$  and  $p = 0.35$ , which means  $1 - p = 0.65$ , so

$$np = 80(0.35) = 28 \geq 5$$

$$n(1 - p) = 80(1 - 0.35) = 80(0.65) = 52 \geq 5$$

and we've verified normality. We were told in the question that the clerk's sample was random, and her sample  $n = 80$  is less than 10% of the total population of invoices issued by the company, so she's not violating the



10% rule. With the conditions for inference satisfied, she can calculate the mean and standard deviation of the sampling distribution of the sample proportion. The mean is

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

$$\mu_{\hat{p}} = 0.35$$

The standard deviation of the sampling distribution (the standard error) is

$$\sigma_{\hat{p}} = \sqrt{\frac{0.35(1 - 0.35)}{80}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.35(0.65)}{80}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2275}{80}}$$

$$\sigma_{\hat{p}} = 0.053$$

The question asks us for the probability that the clerk's sample proportion is within 10% of the population proportion  $p = 35\%$ . In other words, how likely is it that the sample proportion falls between 25% and 45%?

$$P(0.25 < \hat{p} < 0.45) = P\left(\frac{0.25 - 0.35}{0.053} < z < \frac{0.45 - 0.35}{0.053}\right)$$

$$P(0.25 < \hat{p} < 0.45) = P\left(-\frac{0.1}{0.053} < z < \frac{0.1}{0.053}\right)$$

$$P(0.25 < \hat{p} < 0.45) \approx P(-1.89 < z < 1.89)$$



In a  $z$ -table, a  $z$ -value of 1.89 gives 0.9706,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

and a value of  $-1.89$  gives 0.0294.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

Which means the probability under the normal curve of the sampling distribution of the sample proportion between these  $z$ -scores is

$$P(-1.89 < z < 1.89) \approx 0.9706 - 0.0294$$

$$P(-1.89 < z < 1.89) \approx 0.9412$$

$$P(-1.89 < z < 1.89) \approx 94.12\%$$

Which means there's an approximately 94 % chance that the clerk's sample proportion will fall within 10% of the company's claim.

## THE STUDENT'S T-DISTRIBUTION

- 1. We take a random sample of size  $n = 25$ , and we want to be 99 % confident about our results. What  $t$ -score will we find?

*Solution:*

If we look up  $df = n - 1 = 25 - 1 = 24$ , along with a 99 % confidence level in the  $t$ -table, we find a  $t$ -score of  $t = 2.797$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

- 2. We take a random sample of size  $n = 18$ , and we want to be 90 % confident about our results. What  $t$ -score will we find?

*Solution:*

If we look up  $df = n - 1 = 18 - 1 = 17$ , along with a 90 % confidence level in the  $t$ -table, we find a  $t$ -score of  $t = 1.740$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

- 3. We take a random sample of size  $n = 8$ , and we want to be 95 % confident about our results. What  $t$ -score will we find?

*Solution:*

If we look up  $df = n - 1 = 8 - 1 = 7$ , along with a 95 % confidence level in the  $t$ -table, we find a  $t$ -score of  $t = 2.365$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

- 4. We take a random sample of size  $n = 14$ , and our upper-tail probability will be 0.05. What  $t$ -score will we find?



*Solution:*

If we look up  $df = n - 1 = 14 - 1 = 13$ , along with an upper-tail probability of 0.05 in the  $t$ -table, we find a  $t$ -score of  $t = 1.771$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

■ 5. We take a random sample of size  $n = 21$ , and our upper-tail probability will be 0.001. What  $t$ -score will we find?

*Solution:*

If we look up  $df = n - 1 = 21 - 1 = 20$ , along with an upper-tail probability of 0.001 in the  $t$ -table, we find a  $t$ -score of  $t = 3.552$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									



6. We take a random sample of size  $n = 3$ , and our upper-tail probability will be 0.025. What  $t$ -score will we find?

*Solution:*

If we look up  $df = n - 1 = 3 - 1 = 2$ , along with an upper-tail probability of 0.025 in the  $t$ -table, we find a  $t$ -score of  $t = 4.303$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	0.987	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

## CONFIDENCE INTERVAL FOR THE MEAN

- 1. We want to determine the mean of calories served in a restaurant meal in America. The government has already done a study to find this mean, and they found  $\sigma = 350.2$ . We randomly sample 31 meals and find  $\bar{x} = 1,500$ . Construct and interpret a 95 % confidence interval for the mean number of calories in a restaurant meal.

*Solution:*

We have population standard deviation, so with sample mean  $\bar{x} = 1,500$ , standard error  $\sigma = 350.2$ , and critical values of 1.96 associated with 95 % confidence, the confidence interval is given by

$$(a, b) = \bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

$$(a, b) = 1,500 \pm 1.96 \cdot \frac{350.2}{\sqrt{31}}$$

$$(a, b) = 1,500 \pm 123.28$$

$$(a, b) \approx (1,376.72, 1,623.28)$$

Based on the sample, we're 95 % confident that the average number of calories in a restaurant meal was between 1,376.72 and 1,623.28 calories.



2. A bus travels between Kansas City and Denver. We take a sample of 30 trips and find a mean travel time of  $\bar{x} = 12$  hours with standard deviation  $s = 0.25$  hours. Construct and interpret a 95 % confidence interval for the mean bus trip time in hours from Kansas City to Denver.

*Solution:*

We don't have population standard deviation, so we'll have to use the standard deviation from the sample instead. So the confidence interval is given by

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

From the  $t$ -table, we find

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Then the confidence interval is

$$(a, b) = 12 \pm 2.045 \cdot \frac{0.25}{\sqrt{30}}$$

$$(a, b) \approx 12 \pm 0.0933$$

$$(a, b) \approx (11.91, 12.09)$$

Based on the sample, we're 95 % confident that the average bus trip from Kansas City to Denver takes between 11.91 and 12.09 hours.

3. A student wanted to know how many chocolates were in the small bags of chocolate candies her school was selling for a fundraiser. She took a simple random sample of 20 small bags of chocolate candy. From the sample, she found an average of 17 pieces of candy per bag with a standard deviation of 2.03.

A box-plot of the data from the sample showed the distribution to be approximately normal. Compute and interpret a 95 % confidence interval for the mean number of chocolate candies per bag.

*Solution:*

We're told in the problem that the distribution is approximately normal and that it's from a simple random sample. We have a small sample size of 20 bags of candy and an unknown population standard deviation. This means we need to use a  $t$  test-statistic in the confidence interval.

$t^*$  is the test statistic, so we'll look this up in the  $t$ -table. We need to use the confidence level and the degrees of freedom. The confidence level is 95 %, and the degrees of freedom is  $n - 1 = 20 - 1 = 19$ . The value we get from the  $t$ -table is 2.093.



df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

With a sample mean of  $\bar{x} = 17$ , a standard error of  $s = 2.030$ , a sample size of  $n = 20$ , and a critical value from the  $t$ -table of 2.093, the confidence interval is

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

$$(a, b) = 17 \pm 2.093 \cdot \frac{2.03}{\sqrt{20}}$$

$$(a, b) \approx 17 \pm 0.9501$$

$$(a, b) \approx (16.0499, 17.9501)$$

Based on the sample, we're 95 % confident that the average number of chocolates per bag is between 16.0499 and 17.9501 pieces.

- 4. Consider the formula for a confidence interval for a population mean with an unknown sample standard deviation. How does doubling the sample size affect the confidence interval?

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

*Solution:*

Doubling the sample size makes the confidence interval narrower, which means we would get a better estimate of the population mean.

The confidence interval has the formula:

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

If we double the sample size, we multiply  $n$  by 2.

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{2n}}$$

We can choose some numbers for our confidence interval just to look at what's happening. Let's randomly choose some numbers for the sample mean, sample standard deviation and sample size.

$$\bar{x} = 17$$

$$s = 2.030$$

$$n = 11$$

Let's choose a confidence interval of 95 %. Then we can choose the test-statistic based on the sample size. Here we choose the test statistic for  $n = 11$  as  $t^* = 2.228$  and the test statistic for  $2n = 2(11) = 22$  as  $t^* = 2.080$ .

Let's set up the confidence interval with the first sample size.



$$(a, b) = 17 \pm 2.228 \cdot \frac{2.030}{\sqrt{11}}$$

$$(a, b) = 17 \pm 1.3637$$

Now let's look at what happens when the sample size is doubled.

$$(a, b) = 17 \pm 0.9002$$

Here we can see that we're adding and subtracting a smaller amount when the sample size is doubled. This would make the confidence interval narrower, which means we would get a better estimate of the population mean.

- 5. A magazine took a random sample of 30 people and reported the average spending on an Easter basket this year to be \$44.78 per basket with a sample standard deviation of \$18.10. Construct and interpret a 98 % confidence interval for the data.

*Solution:*

We're told in the problem that the data is from a simple random sample. We have a large sample size of 30 people and an unknown population standard deviation. Because population standard deviation is unknown, we'll use a *t*-test.



Let's set up the values we need for the calculation. The sample mean is  $\bar{x} = \$44.78$ , and the sample standard deviation is  $s = \$18.10$ . We also know the sample size is  $n = 30$ .

To find the  $t$ -value associated with a 98 % confidence interval, we realize that  $\alpha/2 = 2\% / 2 = 1\%$ . So we'll look up the intersection of 0.01 and  $df = 29$  in the body of the  $t$ -table. The  $t$ -value is  $t = 2.462$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	<b>2.462</b>	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

So the confidence interval will be

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

$$(a, b) = 44.78 \pm 2.462 \cdot \frac{18.10}{\sqrt{30}}$$

$$(a, b) \approx 44.78 \pm 8.1359$$

$$(a, b) \approx (36.64, 52.92)$$

Based on the sample, we're 98 % confident that the average amount spent on Easter baskets was between \$36.64 and \$52.92.



6. A confidence interval for a study is (11.5,18.5). What was the value of the sample mean?

*Solution:*

The sample mean is always in the middle of the confidence interval. If we find the middle of (11.5,18.5), then we know the sample mean.

$$\bar{x} = \frac{11.5 + 18.5}{2} = \frac{30}{2} = 15$$



## CONFIDENCE INTERVAL FOR THE PROPORTION

- 1. According to a recent poll, 47 % of the 648 Americans surveyed make weekend plans based on the weather. Construct and interpret a 99 % confidence interval for the percentage of Americans who make weekend plans based on the weather.

*Solution:*

The sample proportion is  $\hat{p} = 0.47$  and the confidence level is 99 %. The test statistic for this confidence level is  $z^* = 2.58$  and the sample size is  $n = 648$ . So the confidence interval is

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$(a, b) = 0.47 \pm 2.58 \sqrt{\frac{0.47(1 - 0.47)}{648}}$$

$$(a, b) \approx 0.47 \pm 0.0506$$

$$(a, b) \approx (0.42, 0.52)$$

This means that we're 99 % confident that the percentage of Americans who make weekend plans based on weather is between 42 % and 52 % .



- 2. We want to determine the proportion of teenagers who own their own cell phone. We take a random sample of 100 teenagers and find that 86 of them own a cell phone. At 90 % confidence, build a confidence interval for the population proportion.

*Solution:*

The sample proportion is  $\hat{p} = 86/100 = 0.86$  and the confidence level is 90 %. The test statistic for this confidence level is  $z^* = 1.65$  and the sample size is  $n = 100$ . So the confidence interval is

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$(a, b) = 0.86 \pm 1.65 \sqrt{\frac{0.86(1 - 0.86)}{100}}$$

$$(a, b) \approx 0.86 \pm 0.0573$$

$$(a, b) \approx (0.80, 0.92)$$

This means that we're 90 % confident that the proportion of teenagers who own their own cell phone is between 80 % and 92 %.

- 3. A biologist is trying to determine the proportion of plants in a jungle that are ferns. She takes a random sample of 82 plants and finds that 31 of them can be classified as ferns. At 95 % confidence, what is the confidence interval for the population proportion?



*Solution:*

The sample proportion is  $\hat{p} = 31/82 \approx 0.3780$  and the confidence level is 95 %. The test statistic for this confidence level is  $z^* = 1.96$  and the sample size is  $n = 82$ . So the confidence interval is

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$(a, b) = 0.3780 \pm 1.96 \sqrt{\frac{0.3780(1 - 0.3780)}{82}}$$

$$(a, b) \approx 0.3780 \pm 0.1050$$

$$(a, b) \approx (0.27, 0.48)$$

This means that the biologist can be 95 % confident that the proportion of plants in the jungle that are ferns is between 27 % and 48 % .

- 4. A statistics teacher at a university conducted a study and found that 80 % of university students are interested in taking a statistics class. We want to see if this proportion holds at your own university. Find the minimum sample size we can use to keep a margin of error of 0.02 at a 99 % confidence level.

*Solution:*



The given proportion is  $\hat{p} = 80\% = 0.8$ . The confidence level is 99% and the test statistic for this confidence level is  $z^* = 2.58$ . The margin of error is  $ME = 0.02$ . Plug these values into the formula for margin of error from the confidence interval for a population proportion.

$$ME = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.02 = 2.58 \sqrt{\frac{0.8(1 - 0.8)}{n}}$$

$$0.02 = 2.58 \sqrt{\frac{0.16}{n}}$$

Solve for  $n$ .

$$0.02 = 2.58 \frac{\sqrt{0.16}}{\sqrt{n}}$$

$$0.02\sqrt{n} = 2.58\sqrt{0.16}$$

$$\sqrt{n} = \frac{2.58\sqrt{0.16}}{0.02}$$

$$n = 0.16 \left( \frac{2.58}{0.02} \right)^2$$

$$n = 2,662.56$$

Since we need more than 2,662 university students for the sample, we have to round up to 2,663 students, so we can say  $n = 2,663$ .



5. Sarah is conducting a class survey to determine if the percentage of juniors in favor of having the next dance at a local bowling alley is 65 %. How many juniors should she survey in order to be 90 % confident with a margin of error of 0.08?

*Solution:*

The pre-determined success rate is  $\hat{p} = 65\% = 0.65$ . The confidence level is 90 % and the test statistic for this confidence level is  $z^* = 1.65$ . The margin of error is  $ME = 0.08$ . Plug these values into the formula for margin of error from the confidence interval for a population proportion.

$$ME = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.08 = 1.65 \sqrt{\frac{0.65(1 - 0.65)}{n}}$$

$$0.08 = 1.65 \sqrt{\frac{0.2275}{n}}$$

Solve for  $n$ .

$$0.08 = 1.65 \frac{\sqrt{0.2275}}{\sqrt{n}}$$

$$0.08\sqrt{n} = 1.65\sqrt{0.2275}$$

$$\sqrt{n} = \frac{1.65\sqrt{0.2275}}{0.08}$$

$$n = 0.2275 \left( \frac{1.65}{0.08} \right)^2$$

$$n \approx 96.78$$

Since we need more than 96 juniors for the sample, we have to round up to 97 juniors, so we can say  $n = 97$ .

- 6. A study suggests that 10% of practicing physicians are cognitively impaired. What random sample of practicing physicians is needed to confirm this finding at a confidence level of 95% with a margin of error of 0.05?

*Solution:*

The sample proportion is given as 0.10. The confidence level is 95% and the test statistic for this confidence level is  $z^* = 1.96$ . The margin of error is  $ME = 0.05$ . Plug these values into the formula for margin of error from the confidence interval for a population proportion.

$$ME = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.05 = 1.96 \sqrt{\frac{0.10(1 - 0.10)}{n}}$$



$$0.05 = 1.96 \sqrt{\frac{0.09}{n}}$$

Solve for  $n$ .

$$0.05 = 1.96 \frac{\sqrt{0.09}}{\sqrt{n}}$$

$$0.05\sqrt{n} = 1.96\sqrt{0.09}$$

$$\sqrt{n} = \frac{1.96\sqrt{0.09}}{0.05}$$

$$n = 0.09 \left( \frac{1.96}{0.05} \right)^2$$

$$n \approx 138.30$$

Since we need more than 138 physicians for the sample, we have to round up to 139 physicians, so we can say  $n = 139$ .



## INFERENTIAL STATISTICS AND HYPOTHESES

- 1. A current pain reliever has an 85 % success rate of treating pain. A company develops a new pain reliever and wants to show that its success rate of treating pain is better than the current option. Decide if the hypothesis statement would require a population proportion or a population mean, then set up the statistical hypothesis statements for the situation.

*Solution:*

We're interested in finding out if the new pain reliever has a better success rate than the current one. Since we're given a percentage of success, we'll be using a population proportion  $p$ , instead of a population mean  $\mu$ . And since we're looking at how much better the pain reliever will perform, we use the  $>$  symbol in our alternative hypothesis, which means the null hypothesis has to have the  $\leq$  symbol.

$$H_0 : p \leq 0.85$$

$$H_a : p > 0.85$$

- 2. A research study on people who quit smoking wants to show that the average number of attempts to quit before a smoker is successful is less than 3.5 attempts. How should they set up their hypothesis statements?



*Solution:*

We're interested in finding out if the mean number of attempts is less than 3.5, so we'll be using a population mean  $\mu$ . And since we're looking at whether the mean is less than 3.5, we use the  $<$  symbol in our alternative hypothesis, which means the null hypothesis has to have the  $\geq$  symbol.

$$H_0 : \mu \geq 3.5$$

$$H_a : \mu < 3.5$$

- 3. A factory creates a small metal cylindrical part that later becomes part of a car engine. Because of variations in the process of manufacturing, the diameters are not always identical. The machine was calibrated to create cylinders with an average diameter of  $1/16$  of an inch. During a periodic inspection, it became clear that further investigation was needed to determine whether or not the machine responsible for making the part needed recalibration. Write statistical hypothesis statements.

*Solution:*

The factory wants the mean diameter of the parts it produces to match the diameter that they need,  $1/16$  of an inch. That means this is an example of a statistical hypothesis statement that uses the population mean.



Both parts that are too small or too large could create problems, which means the alternative hypothesis needs to have a  $\neq$  sign. Which means the null hypothesis will include an = sign.

$$H_0 : \mu = \frac{1}{16}$$

$$H_a : \mu \neq \frac{1}{16}$$

- 4. A marketing study for a clothing company concluded that the mean percentage increase in sales could potentially be over 17% for creating a clothing line that focused on lime green and polka dots. Which hypothesis statements do they need to write in order to test their theory?

*Solution:*

The claim of the marketing study is that creating the clothing line that focuses on lime green and polka dots will increase sales by over 17%. Which means the alternative hypothesis would need to include the  $>$  sign, and therefore that the null hypothesis has to include a  $\leq$  sign.

$$H_0 : \mu \leq 0.17$$

$$H_a : \mu > 0.17$$



- 5. A food company wants to ensure that less than 0.0001 % of its product is contaminated. Which hypothesis statements will it write if it wants to test for this?

*Solution:*

The food company wants the proportion of contaminated product to be less than 0.0001 %, so they'll be using a population proportion  $p$ . And since they're looking at whether the proportion is less than 0.0001 %, they'll use the  $<$  symbol in the alternative hypothesis, which means the null hypothesis has to have the  $\geq$  symbol.

$$H_0 : \mu \geq 0.0001 \%$$

$$H_a : \mu < 0.0001 \%$$

- 6. A new medication is being developed to prevent heart worms in dogs, and the developer wants it to work better than the current medication. The current medication prevents heart worms at a rate of 75 %. What hypothesis statements should they write if they want to test whether or not the new medication works better than the existing one?

*Solution:*

The developer wants the proportion of dogs in which heart worm is prevented by their medication to be greater than 0.75, so they'll be using a



population proportion  $p$ . And since we're looking at whether the proportion is greater than 0.75, we use the  $>$  symbol in our alternative hypothesis, which means the null hypothesis has to have the  $\leq$  symbol.

$$H_0 : \mu \leq 0.75$$

$$H_a : \mu > 0.75$$



## SIGNIFICANCE LEVEL AND TYPE I AND II ERRORS

- 1. We're running a statistical test on a new pharmaceutical drug. The stakes are high, because the side effects of the drug could potentially be serious, or even fatal. If we want to reduce the Type I and Type II error rates as low as possible to avoid rejecting the null when it's true or accepting the null when it's false, what should we do when we take the sample?

*Solution:*

The only way to reduce both the Type I error rate and Type II error rate simultaneously is to increase the sample size. Therefore, if it's important that we reduce error rate as low as possible, we should take the largest possible sample.

- 2. If the probability of making a Type II error in a statistical test is 5 % , what is the power of the test?

*Solution:*

The power of a statistical test is the probability that we'll reject the null hypothesis when it's false (make that particular correct choice).



	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	<b>CORRECT</b> Power
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

Power is always equivalent to  $1 - \beta$ , and  $\beta$  is another name for Type II error rate. So

$$\text{Power} = 1 - \beta$$

$$\text{Power} = 1 - \text{Type II error rate}$$

$$\text{Power} = 1 - 0.05$$

$$\text{Power} = 0.95$$

The power of the statistical test, given that the probability of making a Type II error is 5 % , is Power = 95 % .

- 3. On average, professional golfers make 75 % of putts within 5 feet. One golfer believes he does better than this, and wants to use a statistical test to see whether or not he's correct. Unbeknownst to him, in actuality this golfer makes 7 out of 10 of these kinds of putts. When he takes a sample of his putts, he finds  $\hat{p} = 0.92$ . What kind of error might he be in danger of making?

*Solution:*



The golfer's null and alternative hypotheses are

$$H_0 : p \leq 0.75$$

$$H_a : p > 0.75$$

In reality, his null hypothesis is true, but based on the sample proportion  $\hat{p} = 0.92$ , he may be in danger of rejecting the null when he shouldn't.

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

Which means the golfer is in danger of making a Type I error.

- 4. The average age of a guest at an amusement park is 15 years old. One amusement park believes the average age of their guests is younger than this, and wants to use a statistical test to see whether or not they're correct. Unbeknownst to them, in actuality the average guest age at this particular amusement park is 12 years old. When they take a sample of his guests, they find  $\bar{x} = 16$  years. What kind of error might they be in danger of making?

*Solution:*

The park's null and alternative hypotheses are

$$H_0 : \mu \geq 15$$

$$H_a : \mu < 15$$

In reality, their null hypothesis is false, but based on the sample mean  $\bar{x} = 16$ , they may be in danger of accepting the null when they shouldn't.

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

Which means the amusement park is in danger of making a Type II error.

- 5. Of all political donations, 70% come from corporations and lobbies, not from individual citizens. One politician believes he receives less than 70% of his own donations from corporations and lobbies, and wants to use a statistical test to see whether or not he's correct. Unbeknownst to him, in actuality the proportion of his donations that come from corporations and lobbies is 65%. When he takes a sample of his donations that come from corporations and lobbies, he finds  $\hat{p} = 0.72$ . What kind of error might he be in danger of making?

*Solution:*

The politician's null and alternative hypotheses are



$$H_0 : p \geq 0.7$$

$$H_a : p < 0.7$$

In reality, his null hypothesis is false, but based on the sample proportion  $\hat{p} = 0.72$ , he may be in danger of accepting the null when he shouldn't.

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

Which means the politician is in danger of making a Type II error.

- 6. A coffee shop owner believes that he sells 500 cups of coffee each day, on average, and he wants to test this assumption. The truth is, he actually sells fewer than 500 cups each day. He takes a random sample of 10 days and records the number of cups he sells each of those days. What kind of error is the coffee shop owner in danger of making?

Day	1	2	3	4	5	6	7	8	9	10
Cups sold	488	502	496	506	492	489	510	511	506	500

*Solution:*

The owner's null and alternative hypotheses are



$$H_0 : \mu = 500$$

$$H_a : \mu \neq 500$$

If we look at the data, we can see that the sample mean is  $\bar{x} = 500$ . In reality, his null hypothesis is false, but based on the sample mean  $\bar{x} = 500$ , he may be in danger of accepting the null when he shouldn't.

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

Which means the coffee shop owner is in danger of making a Type II error.



## TEST STATISTICS FOR ONE- AND TWO-TAILED TESTS

- 1. A local high school states that its students perform much better than average on a state exam. The average score for all high school students in the state is 106 points. A sample of 256 students at this particular school had an average test score of 129 points with a sample standard deviation of 26.8. Choose and calculate the appropriate test statistic.

*Solution:*

The sample is comparing average scores, which means the population parameter is a population mean (not a proportion) with an unknown standard deviation (since we have the sample standard deviation and not the population standard deviation).

The sample size is large enough at 256 high schoolers that we can assume the distribution is approximately normal. In this case, we use a *t*-test statistic.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{129 - 106}{\frac{26.8}{\sqrt{256}}} \approx 13.73$$

- 2. A dietitian is looking into the claim at a local restaurant that the number of calories in its portion sizes is lower than the national average. The national average is 1,500 calories per meal. She samples 35 meals at



the restaurant and finds they contain an average of 1,250 calories per meal with a sample standard deviation of 350.2. Choose and calculate the appropriate test statistic.

*Solution:*

The sample is comparing average number of calories, which means the population parameter is a population mean (not a proportion) with an unknown standard deviation (since we have the sample standard deviation and not the population standard deviation).

The sample size is large enough at 35 meals that we can assume the distribution is approximately normal. In this case, we use a *t*-test statistic.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1,250 - 1,500}{\frac{350.2}{\sqrt{35}}} \approx -4.22$$

- 3. In a recent survey, 567 out of a 768 randomly selected dog owners said they used a kennel that was run by their veterinary office to board their dogs while they were away on vacation. The study would like to make a conclusion that the majority (more than 50 % ) of dog owners use a kennel run by their veterinary office when the owners go on vacation. Choose and calculate the appropriate test statistic.

*Solution:*



The sample size is large enough at 768 randomly selected individuals that we can state the distribution is approximately normal. We can show this by using the checks for the population proportion,  $np \geq 10$  and  $n(1 - p) \geq 10$ .

The sample size is  $n = 768$  and the population proportion is

$$\hat{p} = \frac{567}{768} \approx 0.738$$

Therefore,

$$n\hat{p} = (768)(0.738) \approx 567 \geq 10$$

$$n(1 - \hat{p}) = (768)(1 - 0.738) \approx 201 \geq 10$$

So we can say that the test statistic will be the  $z$ -test statistic for a population proportion.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.738 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{768}}} \approx 13.21$$

■ 4. We want to open a day care center, so we take a random sample of 500 households in our town with children under preschool age, and find that 243 of them were using a family member to care for those children. We want to determine if, at a statistically significant level, fewer than half of households in our town are using a family member to care for the kids.

1. Set up the hypothesis statements.
2. Check that the conditions for normality are met.



3. State the type of test: upper-tailed, lower-tailed, or two-tailed.
4. Calculate the test statistic using the appropriate formula.

*Solution:*

The hypothesis statements would be

$$H_0 : p \geq 0.5$$

$$H_a : p < 0.5$$

We need to see if we have an approximately normal distribution by using the checks for a population proportion. The sample size is from a simple random sample of  $n = 500$  households. The proportion is the 243 out of the 500 households, so  $\hat{p} = 243/500 = 0.486$ .

$$n\hat{p} = (500)(0.486) = 243 \geq 10$$

$$n(1 - \hat{p}) = (500)(1 - 0.486) = 257 \geq 10$$

Because both values are greater than 10, the distribution is approximately normal. This is a lower-tailed test because the alternative hypothesis uses the  $<$  sign.

This is a population proportion, so we'll calculate a  $z$ -test statistic for a population proportion.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.486 - 0.50}{\sqrt{\frac{0.5(1 - 0.5)}{500}}} \approx -0.6261$$



5. The highest allowable amount of bromate in drinking water is  $0.0100 \text{ mg/L}^2$ . A survey of a city's water quality took 50 water samples in random locations around the city and found an average of  $0.0102 \text{ mg/L}^2$  of bromate with a sample standard deviation of  $0.0025 \text{ mg/L}$ . The survey committee is interested in testing if the amount of bromate found in the water samples is higher than the allowable amount at a statistically significant level.

1. Set up the hypothesis statements.
2. Check that the conditions for normality are met.
3. State the type of test: upper-tailed, lower-tailed, or two-tailed.
4. Calculate the test statistic using the appropriate formula.

*Solution:*

The hypothesis statements would be

$$H_0 : \mu \leq 0.0100$$

$$H_a : \mu > 0.0100$$

The sample size is a simple random sample of 50 samples, so the distribution is approximately normal. This is an upper-tailed test because the alternative hypothesis uses the greater than sign.

This is a population mean with an unknown population standard deviation, so we'll calculate a  $t$ -test statistic with the population mean formula.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.0102 - 0.0100}{\frac{0.0025}{\sqrt{50}}} \approx 0.5657$$

■ 6. A farmer reads a study that states: The average weight of a day-old chick upon hatching is  $\mu_0 = 38.60$  grams with a population standard deviation of  $\sigma = 5.7$  grams. The farmer wants to see if her day-old chicks have the same average. She takes a simple random sample of 60 of her day-old chicks and finds their average weight is  $\bar{x} = 39.1$  grams.

1. Set up the hypothesis statements.
2. Check that the conditions for normality are met.
3. State the type of test: upper-tailed, lower-tailed, or two-tailed.
4. Calculate the test statistic using the appropriate formula.

*Solution:*

The hypothesis statements would be

$$H_0 : \mu = 38.60$$

$$H_a : \mu \neq 38.60$$



The sample size is a simple random sample of 60 of her day-old chicks so we can say the distribution is approximately normal. This is a two-tailed test because the alternative hypothesis uses the  $\neq$  sign.

This is a population mean with a known population standard deviation, so we'll calculate a  $z$ -test statistic with the population mean formula.

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{39.10 - 38.60}{\frac{5.7}{\sqrt{60}}} \approx 0.6795$$



## THE P-VALUE AND REJECTING THE NULL

- 1. A medical trial is conducted to test whether or not a new medicine reduces total cholesterol, when the national average is 230 mg/dL with a standard deviation of 16 mg/dL. The trial takes a simple random sample of 223 adults who take the new medicine, and finds  $\bar{x} = 227$  mg/dL. What can the trial conclude at a significance level of  $\alpha = 0.01$ ?

*Solution:*

The hypothesis statements will be

$$H_0 : \mu \geq 230$$

$$H_a : \mu < 230$$

The test statistic will be

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{227 - 230}{\frac{16}{\sqrt{223}}} = -\frac{3\sqrt{223}}{16} \approx -2.80$$

From the  $z$ -table we get 0.0026.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	<b>.0026</b>	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Because this is a lower-tail test, the  $p$ -value is just this value we found,  $p = 0.0026$ . Comparing this to  $\alpha = 0.01$ , we can see that  $p \leq \alpha$ , which means we'll reject the null hypothesis.

Therefore, at a significance level of  $\alpha = 0.01$ , the trial can conclude that the new medicine reduces cholesterol. Because the  $p$ -value we found is even more significant, the trial could go even further, stating that the new medicine reduces cholesterol at a significance level of  $p = 0.0026$ .

- 2. The national average length of pregnancy is 283.6 days with a population standard deviation of 10.5 days. A hospital wants to know if the average length of a pregnancy at their hospital deviates from the national average. They use a sample of 9,411 births at the hospital to calculate a test statistic of  $z = -1.60$ . Set up the hypothesis statements and find the  $p$ -value.

*Solution:*

The hospital wants to know if mean length of pregnancy at their hospital is different than the national average in a significant way.

$$H_0 : \mu = 283.6$$

$$H_a : \mu \neq 283.6$$



Because the alternative hypothesis uses a  $\neq$  sign, this is a two-tailed test. We were told in the problem that the test statistic is  $z = -1.60$ , so we'll look that up in the  $z$ -table.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	<b>.0548</b>	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

For the lower tail,  $z = -1.60$  gives an area of 0.0548. Now to calculate our  $p$ -value, we multiply this by 2.

$$p = 2(0.0548)$$

$$p = 0.1096$$

- 3. The highest allowable amount of bromate in drinking water is 0.0100 (mg/L)<sup>2</sup>. A survey of a city's water quality took 31 water samples in random locations around the city and used the data to calculate a test statistic of  $t = 2.04$ . The city wants to know if the amount of bromate in their drinking water is too high. Set up the hypothesis statements and determine the type of test, then find the  $p$ -value.

*Solution:*

The city wants to know if the amount of bromate in their drinking water is higher than the allowable amount in a significant way.



$$H_0 : \mu \leq 0.0100$$

$$H_a : \mu > 0.0100$$

Because the alternative hypothesis uses a  $>$  sign, this is an upper-tailed test. We were told in the problem that the test statistic is  $t = 2.04$ , so we'll look that up in the  $t$ -table, but we'll also need to know the degrees of freedom. We know the study included 31 samples, so degrees of freedom is  $n - 1 = 31 - 1 = 30$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Because the test statistic and degrees of freedom gives a  $p$ -value just under  $p = 0.025$ , we'll round the  $p$ -value to  $p = 0.025$ .

- 4. A paint company produces glow in the dark paint with an advertised glow time of 15 min. A painter is interested in finding out if the product behaves worse than advertised. She sets up her hypothesis statements as  $H_0 : \mu \geq 15$  and  $H_a : \mu < 15$ , then calculates a test statistic of  $z = -2.30$ . What would be the conclusions of her hypothesis test at significance levels of  $\alpha = 0.05$ ,  $\alpha = 0.01$ , and  $\alpha = 0.001$ ?



*Solution:*

We need to look up  $z = -2.30$  in the  $z$ -table.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	<b>.0107</b>	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

For a lower-tail test, the  $p$ -value is given by this value we found in the  $z$ -table, so  $p = 0.0107$ .

We know that

If  $p \leq \alpha$ , reject the null hypothesis

If  $p > \alpha$ , do not reject the null hypothesis

Therefore,

- For  $p = 0.0107$  and  $\alpha = 0.05$ ,  $p \leq \alpha$ , so she'd reject the null
- For  $p = 0.0107$  and  $\alpha = 0.01$ ,  $p > \alpha$ , so she'd fail to reject the null
- For  $p = 0.0107$  and  $\alpha = 0.001$ ,  $p > \alpha$ , so she'd fail to reject the null

- 5. An article reports that the average wasted time by an employee is 125 minutes every day. A manager takes a small random sample of 16 employees and monitors their wasted time, calculating that average wasted time for her employees is 118 minutes with a standard deviation of 28.7 minutes. She wants to know if 118 minutes is below average at a

significance level of  $\alpha = 0.05$ . She assumes the population is normally distributed.

1. State the population parameter and whether a  $t$ -test or  $z$ -test should be used.
2. Check that the conditions for performing the statistical test are met.
3. Set up the hypothesis statements.
4. State the type of test: upper-tailed, lower-tailed, or two-tailed.
5. Calculate the test statistic using the appropriate formula.
6. Calculate the  $p$ -value.
7. Compare the  $p$ -value to the significance level and draw a conclusion.

*Solution:*

This is a population mean with an unknown population standard deviation because the manager is going to do her analysis based on the sample standard deviation. She also has a small sample size of 16 employees. This means we should use the  $t$ -test statistic because we have a small sample size and also an unknown population standard deviation.



The conditions for performing a  $t$ -test with a population mean are an approximately normal distribution and a simple random sample, and we've been told in the problem that both of those conditions are met.

The manager wants to know if 118 minutes is below average. We're comparing 118 minutes to the stated average of 125 minutes. Since she wants to know if her measurement is below average, we should use the less than symbol in our alternative hypothesis.

$$H_0 : \mu \geq 125$$

$$H_a : \mu < 125$$

The test statistic will be

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{118 - 125}{\frac{28.7}{\sqrt{16}}} \approx -0.9756$$

The next step is to find the  $p$ -value by looking up the test statistic in the  $t$ -table. To look up a  $t$ -value, we'll also need to know the degrees of freedom from the problem. We know the study included 16 samples, so the degrees of freedom are  $16 - 1 = 15$ .

We calculated the test statistic as  $t \approx -0.9756$ . We're looking for the area in the lower tail, but the table will give us the area in the upper tail when  $t = 0.9756$ . Remember these values are equal because the  $t$ -curve is symmetric. Now we look up where our test statistic and degrees of freedom intersect. The value we read from the  $t$ -table is somewhere between  $p = 0.20$  and  $p = 0.15$ .



df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Regardless of the exact value of  $p$  between  $p = 0.20$  and  $p = 0.15$ , at a significance level of  $\alpha = 0.05$ , we can say  $p > \alpha$ , so the manager will fail to reject the null hypothesis, and conclude that there's not enough evidence to conclude that her employees waste less time than the average rate of 125 minutes per day at the significance level of  $\alpha = 0.05$ .

- 6. We want to test if college students take fewer than than 5 years to graduate, on average, so we take a simple random sample of 30 students and record their years to graduate. For the sample,  $\bar{x} = 4.9$  and  $s = 0.5$ . What can we conclude at 90 % confidence?

*Solution:*

The hypothesis statements will be

$$H_0 : \mu \geq 5$$

$$H_a : \mu < 5$$

Find the test statistic.



$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{4.9 - 5}{\frac{0.5}{\sqrt{30}}} = \frac{-0.1}{\frac{0.5}{\sqrt{30}}} \approx -1.0954$$

The next step is to look up the test statistic in the  $t$ -table. To look up a  $t$ -value, we'll also need to know the degrees of freedom from the problem. We know the study included 30 students, so the degrees of freedom are  $30 - 1 = 29$ .

We calculated the test statistic as  $t \approx -1.0954$ . We're looking for the area in the lower tail, but the table will give us the area in the upper tail when  $t = 1.0954$ . Remember these values are equal because the  $t$ -curve is symmetric. Now we look up where our test statistic and degrees of freedom intersect. The value we read from the  $t$ -table is somewhere between  $p = 0.15$  and  $p = 0.10$ .

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Regardless of the exact value of  $p$  between  $p = 0.15$  and  $p = 0.10$ , at a significance level of  $\alpha = 0.10$ , we can say  $p > \alpha$ , so we'll fail to reject the null hypothesis, because there's not enough evidence that college students take less than 5 years to graduate at a significance level of  $\alpha = 0.1$ .

## HYPOTHESIS TESTING FOR THE POPULATION PROPORTION

■ 1. A large electric company claims that at least 80 % of the company's 1,000,000 customers are very satisfied. Using a simple random sample, 100 customers were surveyed and 73 % of the participants were very satisfied. Based on these results, should we use a one- or two- tailed test, and should we accept or reject the company's hypothesis? Assume a significance level of 0.05.

*Solution:*

The first step is to state the null and alternative hypotheses for the survey.

$$H_0 : p \geq 0.80$$

$$H_a : p < 0.80$$

These hypotheses require a one-tailed test, specifically a lower-tail test. The null hypothesis will be rejected only if the sample proportion is significantly less than 80 % .

We calculate standard error based on the sample,

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.8(1 - 0.8)}{100}} = 0.04$$

and then compute the  $z$ -score test statistic.



$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.73 - 0.80}{0.04} = -1.75$$

The  $z$ -table gives 0.0401 for a  $z$ -score of  $z = -1.75$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455

Since we have a one-tailed test, the  $p$ -value is  $p = 0.0401$ , and we were told in the problem that  $\alpha = 0.05$ . Because the  $p$ -value is less than the  $\alpha$ -level,  $p < \alpha$ , we'll reject the null hypothesis.

- 2. A university is conducting a statistical test to determine whether the percentage of its students who live on its campus is above the national average of 64 %. They've calculated the test statistic to be  $z = 1.40$ . Set up hypothesis statements and find the  $p$ -value.

*Solution:*

The university wants to know if the proportion of students who live on campus is above the national average in a statistically significant way.

$$H_0 : p \leq 64 \%$$

$$H_a : p > 64 \%$$

Because the alternative hypothesis uses a  $>$  sign, this is a one-tail, upper-tailed test. We were told that the test statistic is  $z = 1.40$ , so we'll look that up in the  $z$ -table.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	<b>.9192</b>	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441

This is an upper-tail test, which means the  $p$ -value is the area outside of 0.9192, or

$$p = 1 - 0.9192$$

$$p = 0.0808$$

- 3. A report claims that 60 % of American families take fewer than 6 months to purchase a home, from the time they start looking to the time they make their first offer. A realtor wants to know if her clients purchase at the same rate, so she takes a simple random sample of 50 of her clients and finds  $\hat{p} = 0.64$  and  $\sigma_{\hat{p}} = \sqrt{0.0048}$  from the sample. What can she conclude with 90 % confidence?

*Solution:*

The first step is to state the null and alternative hypotheses for the survey.

$$H_0 : p = 0.60$$

$$H_a : p \neq 0.60$$

These hypotheses require a two-tailed test. The null hypothesis will be rejected only if the sample proportion is significantly different than 60%.

The  $z$ -score will be

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.64 - 0.60}{\sqrt{0.0048}} = \frac{0.04}{\sqrt{0.0048}} \approx 0.58$$

The  $z$ -table gives 0.7190 for a  $z$ -score of  $z = 0.58$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

The  $p$ -value is the area outside of 0.7190, or

$$p = 1 - 0.7190$$

$$p = 0.2810$$

Since we have a two-tailed test, the  $p$ -value is double this, or

$$p = 2(0.2810)$$

$$p = 0.5620$$

We were told in the problem that  $\alpha = 0.10$ , so  $p \geq \alpha$ , which means the realtor will fail to reject the null hypothesis. So she can't say that her clients purchase at a different rate than the report claims.



4. A gambler wins 48 % of the hands he plays, but he feels like he's on a losing streak recently, winning fewer hands than normal. He takes a random sample of 40 of his recent hands, and finds the proportion of winning hands in the sample to be  $\hat{p} = 0.45$  with  $\sigma_{\hat{p}} = \sqrt{0.00624}$ . What can he conclude with 90 % confidence?

*Solution:*

The first step is to state the null and alternative hypotheses for the survey.

$$H_0 : p \geq 0.48$$

$$H_a : p < 0.48$$

These hypotheses require a one-tailed test, specifically a lower-tail test. The null hypothesis will be rejected only if the sample proportion is significantly lower than 48 %.

The  $z$ -score test statistic will be

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.45 - 0.48}{\sqrt{0.00624}} = \frac{-0.03}{\sqrt{0.00624}} \approx -0.38$$

The  $z$ -table gives 0.3520 for a  $z$ -score of  $z = -0.38$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483

This is a lower-tail test, which means this is also the  $p$ -value,  $p = 0.3520$ .

We were told in the problem that  $\alpha = 0.10$ . Because the  $p$ -value is greater than the  $\alpha$ -level,  $p < \alpha$ , the gambler will fail to reject the null hypothesis, and conclude that he hasn't been on a losing streak at a statistically significant level.

■ 5. A study claims that the proportion of new homeowners who purchase an internet subscription plan is 0.92. We take a random sample of 140 new homeowners to test this claim, and find  $\hat{p} = 0.9$  with  $\sigma_{\hat{p}} \approx 0.0229$ . What can we conclude at a significance level of  $\alpha = 0.05$ ?

*Solution:*

The first step is to state the null and alternative hypotheses for the survey.

$$H_0 : p = 0.92$$

$$H_a : p \neq 0.92$$

These hypotheses require a two-tailed test. The null hypothesis will be rejected only if the sample proportion is significantly different than 92%.

The  $z$ -score test statistic will be

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.90 - 0.92}{0.0229} = -\frac{0.02}{0.0229} \approx -0.87$$



The  $z$ -table gives 0.1922 for a  $z$ -score of  $z \approx -0.87$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148

Since this is a two-tailed test, we double this to find the  $p$ -value, and we get

$$p = 2(0.1922)$$

$$p = 0.3844$$

We were told in the problem that  $\alpha = 0.05$ . Because the  $p$ -value is greater than the  $\alpha$ -level,  $p > \alpha$ , we'll fail to reject the null hypothesis, which means that we can't conclude that the number of homeowners who purchase an internet subscription plan is different than 92 %.

- 6. A recent study reported that the 15.3 % of patients who are admitted to the hospital with a heart attack die within 30 days of admission. The same study reported that 16.7 % of the 3,153 patients who went to the hospital with a heart attack died within 30 days of admission when the lead cardiologist was away.

Is there enough evidence to conclude that the percentage of patients who die when the lead cardiologist is away is any different than when they're present? Make conclusions at significance levels of  $\alpha = 0.05$  and  $\alpha = 0.01$ .



1. State the population parameter and whether a  $t$ -test or  $z$ -test should be used.
2. Check that the conditions for performing the statistical test are met.
3. Set up the hypothesis statements.
4. State the type of test: upper-tailed, lower-tailed, or two-tailed.
5. Calculate the test statistic using the appropriate formula.
6. Calculate the  $p$ -value.
7. Compare the  $p$ -value to the significance level and draw a conclusion.

*Solution:*

This is a population proportion because the data is looking at the proportion of heart attack patients admitted to the hospital who die within 30 days of admittance.

The sample size is large at 3,153 with a population proportion of 16.7 %, but to continue with the test we need to assume that the sample was a simple random sample (since it's not stated in the problem).

This sample size is large enough to meet the conditions:

$$np = (3,153)(0.167) \approx 527 \geq 10$$



$$n(1 - p) = (3,153)(1 - 0.167) \approx 2,626 \geq 10$$

When these two conditions are met, then the distribution is approximately normal. Then we can continue with the hypothesis test.

According to the problem, we want to know if the percentage of patients who went to the hospital with a heart attack and died within 30 days of admission when the leading cardiologist was away differs from when they were not away. This means we need to use the  $\neq$  symbol in our hypothesis statement.

$$H_0 : p = 0.153$$

$$H_a : p \neq 0.153$$

Since we're dealing with a population proportion, the  $z$ -test statistic will be

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.167 - 0.153}{\sqrt{\frac{0.153(1 - 0.153)}{3,153}}} \approx 2.1837$$

The next step is to find the  $p$ -value by looking up the test statistic in the  $z$ -table. Since this is a two-tailed test, we'll need to double the area we find in either the upper or lower tail. From the  $z$ -table, we find a value of 0.9854.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890



But this is the area below the upper tail. Before we can do anything else we need to find the area in the upper tail. The total area under the curve is 1, so we'll subtract this value from 1.

$$1 - 0.9854 = 0.0146$$

Now to calculate the  $p$ -value, we multiply the upper tail by 2.

$$p = 2(0.0146)$$

$$p = 0.0292$$

We know that

If  $p \leq \alpha$ , reject the null hypothesis

If  $p > \alpha$ , do not reject the null hypothesis

Therefore,

- For  $p = 0.0292$  and  $\alpha = 0.05$ ,  $p \leq \alpha$ , so we'd reject the null
- For  $p = 0.0292$  and  $\alpha = 0.01$ ,  $p > \alpha$ , so we'd fail to reject the null

Which means there's enough evidence to conclude that the percentage of patients who went to the hospital with a heart attack and died within 30 days of admission when the leading cardiologist was away is different than when the leading cardiologist is present, at a statistically significant level of  $\alpha = 0.05$ , but not at  $\alpha = 0.01$ .



## CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS

- 1. A researcher wants to compare the effectiveness of new blood pressure medication for males and females. He takes a simple random sample of 25 males and 25 females and finds an average drop in blood pressure of 4.5 with a standard deviation of 0.35 for males, and an average drop in blood pressure of 4.85 with a standard deviation of 0.22 for females. Can he use pooled standard deviation to find the confidence interval?

*Solution:*

The sample variances are  $s_1^2 = 0.35^2 = 0.1225$  and  $s_2^2 = 0.22^2 = 0.0484$ . The variance  $s_1^2$  is more than twice the other variance, so the researcher will assume unequal population variances.

- 2. A grocery store wants to know whether families of 3 spend more on groceries than families of 2. They randomly survey ten 2-person families and find a mean weekly grocery spend of \$252 with a standard deviation of \$26, then randomly survey ten 3-person families and find a mean weekly grocery spend of \$258 with a standard deviation of \$22. Calculate the number of degrees of freedom.

*Solution:*



The sample variances are  $s_1^2 = 26^2 = 676$  and  $s_2^2 = 22^2 = 484$ , and neither sample variance is more than twice the other, so we can assume equal population variances.

Which means that the degrees of freedom will be given by

$$\text{df} = n_1 + n_2 - 2$$

$$\text{df} = 10 + 10 - 2$$

$$\text{df} = 18$$

- 3. For the last question, calculate a 95 % confidence interval around the difference in mean weekly grocery spending for 2-and 3-person families.

*Solution:*

Because we already determined in the previous solution that we're working with equal population variances, we'll calculate pooled standard deviation.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(10 - 1)26^2 + (10 - 1)22^2}{10 + 10 - 2}}$$

$$s_p = \sqrt{\frac{9(26^2) + 9(22^2)}{18}}$$

$$s_p = \sqrt{\frac{26^2 + 22^2}{2}}$$

$$s_p \approx 24.083$$

At 95 % confidence and df = 18, the *t*-table gives 2.101.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Then the confidence interval is

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(a, b) \approx (252 - 258) \pm 2.101 \times 24.083 \sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$(a, b) \approx -6 \pm 22.63$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (-6 - 22.63, -6 + 22.63)$$

$$(a, b) \approx (-28.63, 16.63)$$

We can be 95% confident that the true difference of mean weekly grocery spend for 2- and 3-person families will fall between  $-\$28.63$  and  $\$16.63$ . But because the confidence interval contains 0, it means there's likely no difference between the population means.

- 4. A researcher is interested in whether a new fitness program lowers systolic blood pressure. He enrolls 50 participants into the study and randomly splits them into two groups of 25 each. The first group kept their same physical activity habits, while the second group followed the new fitness program. After a month, the mean systolic blood pressure in the group of exercisers was 123 with standard deviation of 4, and the mean systolic pressure in the group of non-exercisers was 131 with a standard deviation of 5.5. Calculate the margin of error at 99% confidence.

*Solution:*

The sample variances are  $s_1^2 = 4^2 = 16$  and  $s_2^2 = 5.5^2 = 30.25$ , and neither sample variance is more than twice the other, so we'll assume equal population variances, which means the degrees of freedom will be

$$\text{df} = n_1 + n_2 - 2$$

$$\text{df} = 25 + 25 - 2$$

$$\text{df} = 48$$



The pooled standard deviation will be

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(25 - 1)(16) + (25 - 1)(30.25)}{25 + 25 - 2}}$$

$$s_p = \sqrt{\frac{384 + 726}{48}}$$

$$s_p = \sqrt{23.125}$$

At  $df = 48$  and 99 % confidence, the  $t$ -table gives  $t = 2.682$ . Now we can calculate the margin of error as

$$ME = t_{\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$ME = 2.682 \times \sqrt{23.125} \sqrt{\frac{1}{25} + \frac{1}{25}}$$

$$ME \approx 2.682\sqrt{1.85}$$

$$ME \approx 3.6479$$

- 5. Given population standard deviations  $\sigma_1 = 2.25$  and  $\sigma_2 = 2.02$ , with sample means  $\bar{x}_1 = 14.5$  and  $\bar{x}_2 = 13.6$  and sample sizes  $n_1 = 250$  and  $n_2 = 250$ , calculate a 90 % confidence interval around the difference of means.



*Solution:*

A 90% confidence level is associated with  $z$ -scores of  $z = \pm 1.65$ , so the confidence interval will be

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(a, b) = (14.5 - 13.6) \pm 1.65 \sqrt{\frac{2.25^2}{250} + \frac{2.02^2}{250}}$$

$$(a, b) \approx 0.9 \pm 0.316$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (0.9 - 0.316, 0.9 + 0.316)$$

$$(a, b) \approx (0.584, 1.216)$$

6. Owners of a large shopping center want to determine whether or not there's a difference in the amount of time that men and women spend per visit to the shopping center. Previous studies showed a standard deviation of 0.4 hours for men and 0.2 hours for women. The owners sample 500 men and 500 women and find that the mean time spent per visit was 1.6 hours for men and 2.5 hours for women. Find a 98% confidence interval around the difference of means.



*Solution:*

The  $z$ -values associated with a 98 % confidence level are  $z \pm 2.33$ , so because the population standard deviations are known, the confidence interval will be given by

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(a, b) = (1.6 - 2.5) \pm 2.33 \sqrt{\frac{0.4^2}{500} + \frac{0.2^2}{500}}$$

$$(a, b) = -0.9 \pm 0.0466$$

Therefore, we can say that the confidence interval is

$$(a, b) = (-0.9 - 0.0466, -0.9 + 0.0466)$$

$$(a, b) = (-0.9466, -0.8534)$$

We can be 98 % confidence that the true difference between mean time spent in the shopping center per visit by men and women will fall between  $-0.95$  and  $-0.85$  hours. Therefore, we've provided support for the hypothesis that men spend less time in the shopping center per visit than women.



## HYPOTHESIS TESTING FOR THE DIFFERENCE OF MEANS

- 1. An ice cream shop owner believes his average daily revenue is higher in August than it is in September. He calculated average daily revenue of \$496 in August and \$456 in September, with standard deviations of \$14 and \$21.5, respectively. What can he conclude at a 0.05 significance level using a  $p$ -value approach.

*Solution:*

The sample variances are  $s_1^2 = 14^2 = 196$  and  $s_2^2 = 21.5^2 = 462.25$ , so the second sample variance is more than twice the first. Therefore, because the sample variances are unequal, we can assume unequal population variances. Additionally, we have large samples  $n_A = 31$  and  $n_S = 30$ , since there are 31 days in August and 30 days in September, so we'll use a  $z$ -test.

We'll run an upper-tailed test because the shop owner believes the average daily revenue in August was higher than in September.

$H_0 : \mu_A - \mu_S \leq 0$ ; the average daily revenue in August is not higher than in September.

$H_a : \mu_A - \mu_S > 0$ ; the average daily revenue in August is higher than in September.

Then the test statistic is



$$z = \frac{\bar{x}_A - \bar{x}_S}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_S^2}{n_S}}}$$

$$z = \frac{496 - 456}{\sqrt{\frac{14^2}{31} + \frac{21.5^2}{30}}}$$

$$z = \frac{40}{\sqrt{\frac{196}{31} + \frac{462.25}{30}}}$$

$$z \approx 8.581$$

The test statistic is much larger than the largest  $z$ -value in the  $z$ -table, so we could say that the probability of finding  $z \approx 8.581$  is almost 0. Therefore,  $p \leq \alpha$  and we can reject the null hypothesis, concluding that the average daily revenue in August was higher than in September.

- 2. A fitness coach wants to determine whether his new weight loss program is more effective than his old program. He randomly samples 50 of his clients following each program, and finds a mean weight loss of 5.5 pounds with a standard deviation of 1.05 pounds for those following the old program, and a mean weight loss of 6.12 pounds with a standard deviation of 0.95 pounds for those following the new program. Using a critical value approach, what can the coach conclude at a 0.01 level of significance?



*Solution:*

The sample variances are  $s_1^2 = 1.05^2 = 1.1025$  and  $s_2^2 = 0.95^2 = 0.9025$ . Neither sample variance is more than twice the other, which means we can assume equal sample variances, and therefore equal population variances.

The null and alternative hypotheses for the upper-tailed test are

$$H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

Since both samples are larger than 30 and the population variances are equal, we can calculate pooled standard deviation.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(50 - 1)1.05^2 + (50 - 1)0.95^2}{50 + 50 - 2}}$$

$$s_p = \sqrt{\frac{49(1.05^2) + 49(0.95^2)}{98}}$$

$$s_p = \sqrt{\frac{1.05^2 + 0.95^2}{2}}$$

$$s_p \approx 1.00125$$

Then the  $z$ -statistic is



$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$z = \frac{6.12 - 5.5}{1.00125 \sqrt{\frac{1}{50} + \frac{1}{50}}}$$

$$z = \frac{0.62}{1.00125 \sqrt{\frac{1}{25}}}$$

$$z \approx 3.10$$

For an upper-tailed test and  $\alpha = 0.01$ , the critical  $z$ -value is  $z = 2.33$ . Since  $3.10 > 2.33$ , the coach can reject the null hypothesis and conclude that the new weight loss program is more effective than the old program.

- 3. Test the claim that, in 2006, the mean weight of men in the US was not significantly different from the mean weight of women. Previous research showed population standard deviations were 10.25 pounds for men and 8.58 pounds for women. A random sample of 1,500 men has a mean weight of 193.5 pounds and a random sample of 1,500 women has a mean weight of 185.3 pounds. Assuming the population variances are unequal, use a  $p$ -value approach to formulate a decision at the 0.05 significance level.

*Solution:*



Given the sample mean  $\mu_m = 193.5$  and population standard deviation  $\sigma_m = 10.25$  for men, and the sample mean  $\mu_w = 185.3$  and population standard deviation  $\sigma_w = 8.58$  for women, the null and alternative hypotheses for the two-tailed test will be

$$H_0 : \mu_m - \mu_w = 0$$

$$H_a : \mu_m - \mu_w \neq 0$$

With unequal population variances and large samples, the test statistic will be

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{193.5 - 185.3}{\sqrt{\frac{10.25^2}{1,500} + \frac{8.58^2}{1,500}}}$$

$$z = \frac{8.2}{\sqrt{\frac{105.0625 + 73.6164}{1,500}}}$$

$$z = 8.2 \sqrt{\frac{1,500}{105.0625 + 73.6164}}$$

$$z \approx 23.76$$

The test statistic is much larger than the largest  $z$ -value in the  $z$ -table, so we could say that the probability of finding  $z \approx 23.76$  is almost 0. Therefore,



$p \leq \alpha$  and we can reject the null hypothesis, concluding that the mean weight of men and women was significantly different.

- 4. A research team wants to determine whether men and women drink a different amount of water each day. They randomly sample 25 men and 25 women and find that the men consumed 1.48 liters of water with a standard deviation of 0.13 liters, and that the women consumed 1.62 liters of water with a standard deviation of 0.20 liters. Using a critical value approach, what can the research team conclude at a 0.10 level of significance?

*Solution:*

The null and alternative hypotheses for the two-tailed test will be

$$H_0 : \mu_m - \mu_w = 0$$

$$H_a : \mu_m - \mu_w \neq 0$$

With small samples and unequal population variances ( $s_2^2 = 0.2^2 = 0.04$  is more than twice  $s_1^2 = 0.13^2 = 0.0169$ ), the  $t$ -statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



$$t = \frac{1.48 - 1.62}{\sqrt{\frac{0.0169}{25} + \frac{0.04}{25}}}$$

$$t = \frac{-0.14}{\sqrt{\frac{0.0569}{25}}}$$

$$t \approx -2.9346$$

The number of degrees of freedom will be

$$df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left( \frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left( \frac{s_2^2}{n_2} \right)^2}$$

$$df = \frac{\left( \frac{0.0169}{25} + \frac{0.04}{25} \right)^2}{\frac{1}{25 - 1} \left( \frac{0.0169}{25} \right)^2 + \frac{1}{25 - 1} \left( \frac{0.04}{25} \right)^2}$$

$$df \approx 27.9966$$

With  $df = 27$  and  $\alpha = 0.10$ , we find critical  $t$ -values of  $t \pm 1.703$ . Since  $-2.9346 < -1.703$ , the research team can reject the null hypothesis and conclude that there's a difference in the mean amount of water that men and women drink each day.

- 5. Given  $\bar{x}_1 = 23.55$  and  $\bar{x}_2 = 20.12$  with  $s_1 = 2.3$ ,  $s_2 = 2.9$ ,  $n_1 = 10$ , and  $n_2 = 15$ , determine whether the two population means differ significantly. Using a critical value approach, and assuming population standard deviations are unequal, what can we conclude at a 0.01 level of significance?

*Solution:*

We want to determine whether there's a difference in population means, so we need to use a two-tailed test, and our hypothesis statements will be

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

With small samples and unequal population variances, we should calculate a *t*-statistic.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{23.55 - 20.12}{\sqrt{\frac{2.3^2}{10} + \frac{2.9^2}{15}}}$$

$$t = \frac{3.43}{\sqrt{\frac{5.29}{10} + \frac{8.41}{15}}}$$

$$t \approx 3.286$$

Calculate the number of degrees of freedom.

$$df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left( \frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left( \frac{s_2^2}{n_2} \right)^2}$$

$$df = \frac{\left( \frac{2.3^2}{10} + \frac{2.9^2}{15} \right)^2}{\frac{2.3^4}{10^2(10 - 1)} + \frac{2.9^4}{15^2(15 - 1)}}$$

$$df \approx 22.17$$

Rounding down to the nearest whole number gives  $df = 22$ . From the  $t$ -table, we find that the critical  $t$ -value for the two-tailed test with  $\alpha = 0.01$  and  $df = 22$  is  $t = 2.819$ .

Because  $3.286 > 2.819$ , we can reject the null hypothesis and conclude that there's a significant difference in population means.

- 6. John claims that the temperature in July is higher than the temperature in August. He recorded the temperature daily at 12 : 00 p.m. throughout July and August. He found a mean temperature of  $28.4^\circ \text{C}$  with a standard deviation of  $2.1^\circ \text{C}$  in July, and a mean temperature of  $27.3^\circ \text{C}$  with a standard deviation of  $1.7^\circ \text{C}$  in August. Using a critical value approach and assuming the population variances are unequal, what can John conclude at a 0.05 level of significance?



*Solution:*

Using  $\mu_1$ ,  $\bar{x}_1$ , and  $s_1$  for July and  $\mu_2$ ,  $\bar{x}_2$ , and  $s_2$  for August, the hypothesis statements for the John's upper-tailed test will be

$$H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

With large samples and unequal population variances, John's test statistic will be

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$z = \frac{28.4 - 27.3}{\sqrt{\frac{2.1^2}{31} + \frac{1.7^2}{31}}}$$

$$z = 1.1 \sqrt{\frac{31}{4.41 + 2.89}}$$

$$z \approx 2.27$$

The critical  $z$ -value for  $\alpha = 0.05$  with an upper-tailed test is  $z = 1.65$ . Since  $2.27 > 1.65$ , John can reject the null hypothesis and conclude that the mean temperature in July is higher than in August.



## MATCHED-PAIR HYPOTHESIS TESTING

- 1. A golf club manufacturer claims that their new driver delivers 15 yards of extra driving distance. They record the before and after driving distances of 10 top professional players.

Player	1	2	3	4	5	6	7	8	9	10
Before $x_1$	303	308	295	305	301	312	287	294	300	301
After $x_2$	307	320	297	315	305	316	299	302	307	315
Difference, $d$	4	12	2	10	4	4	12	8	7	14
$d^2$	16	144	4	100	16	16	144	64	49	196

Can the manufacturer conclude at a 5% significance level that their driver delivers 15 yards of extra driving distance?

*Solution:*

The manufacturer will define the “before” responses as Population 1, and the “after” responses as Population 2, and their null and alternative hypotheses will be

$$H_0 : \mu_2 - \mu_1 \leq 15$$

$$H_a : \mu_2 - \mu_1 > 15$$

where  $\mu_1$  is the mean driving distance with the players’ current drivers, and  $\mu_2$  is the mean driving distance with the manufacturer’s new driver. And

because  $\mu_2 - \mu_1$  is the difference in distance, the hypothesis statements could also be written as

$$H_0 : \mu_d \leq 15$$

$$H_a : \mu_d > 15$$

where  $\mu_d$  is the mean difference between the two populations.

To find the mean difference, we'll sum the differences and divide by the number of matched-pairs in our sample,  $n = 10$ .

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{4 + 12 + 2 + 10 + 4 + 4 + 12 + 8 + 7 + 14}{10} = \frac{77}{10} = 7.7$$

So the sample mean tells us that mean distance gained is 7.7 yards. Then the sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

To calculate this, we'll first find

$$\sum_{i=1}^n (d_i - \bar{d})^2$$

$$(4 - 7.7)^2 + (12 - 7.7)^2 + (2 - 7.7)^2 + (10 - 7.7)^2 + (4 - 7.7)^2$$

$$+(4 - 7.7)^2 + (12 - 7.7)^2 + (8 - 7.7)^2 + (7 - 7.7)^2 + (14 - 7.7)^2$$

$$(-3.7)^2 + 4.3^2 + (-5.7)^2 + 2.3^2 + (-3.7)^2 + (-3.7)^2 + 4.3^2 + 0.3^2 + (-0.7)^2 + 6.3^2$$

$$13.69 + 18.49 + 32.49 + 5.29 + 13.69 + 13.69 + 18.49 + 0.09 + 0.49 + 39.69$$



156.1

Then the sample standard deviation is

$$s_d = \sqrt{\frac{156.1}{9}}$$

$$s_d \approx \sqrt{17.34}$$

$$s_d \approx 4.165$$

Because the population standard deviations are unknown, and/or because both sample sizes are small,  $n_1, n_2 < 30$ , the test statistic will be

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$$t \approx \frac{7.7 - 15}{\frac{4.165}{\sqrt{10}}}$$

$$t \approx -7.3 \cdot \frac{\sqrt{10}}{4.165}$$

$$t \approx -5.543$$

and the degrees of freedom are

$$\text{df} = n - 1 = 10 - 1 = 9$$

At a significance level of 5% (a confidence level of 95% for an upper-tailed test), and  $\text{df} = 9$ , the  $t$ -table gives 1.833.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

The manufacturer's *t*-test statistic  $t \approx -5.543$  doesn't meet the threshold  $t = 1.833$ , so the critical value approach tells them that they can't reject the null hypothesis, and therefore can't conclude that their new driver adds 15 yards of extra distance for the professional players.

- 2. A car company believes that the changes they've made to their hybrid engine will increase miles per gallon by 4. They send out one car with the old engine and one car with the new engine to drive the same route, and record the miles per gallon of each pair of cars.

Route	1	2	3	4	5	6	7	8	9	10
Old engine	39	39	38	42	44	43	42	47	47	47
New engine	50	49	45	46	46	41	42	43	43	49
Difference, d	11	10	7	4	2	-2	0	-4	-4	2
d <sup>2</sup>	121	100	49	16	4	4	0	16	16	4

Can the car company conclude at a 1% significance level that the changes they've made to the hybrid engine deliver 4 extra miles per gallon?

**Solution:**

The car company will define the values for the old engine as Population 1, and the values for the new engine as Population 2, and their null and alternative hypotheses will be

$$H_0 : \mu_2 - \mu_1 \leq 4$$

$$H_a : \mu_2 - \mu_1 > 4$$

where  $\mu_1$  is the miles per gallon obtained by the old engine, and  $\mu_2$  is the miles per gallon obtained by the new engine. And because  $\mu_2 - \mu_1$  is the difference in miles per gallon, the hypothesis statements could also be written as

$$H_0 : \mu_d \leq 4$$

$$H_a : \mu_d > 4$$

where  $\mu_d$  is the mean difference between the two populations.

To find the mean difference, we'll sum the differences and divide by the number of matched-pairs in our sample,  $n = 10$ .

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{11 + 10 + 7 + 4 + 2 + (-2) + 0 + (-4) + (-4) + 2}{10} = \frac{26}{10} = 2.6$$

So the sample mean tells us that mean difference is 2.6 miles per gallon. Then the sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$



To calculate this, we'll first find

$$\sum_{i=1}^n (d_i - \bar{d})^2$$

$$(11 - 2.6)^2 + (10 - 2.6)^2 + (7 - 2.6)^2 + (4 - 2.6)^2 + (2 - 2.6)^2$$

$$+(-2 - 2.6)^2 + (0 - 2.6)^2 + (-4 - 2.6)^2 + (-4 - 2.6)^2 + (2 - 2.6)^2$$

$$8.4^2 + 7.4^2 + 4.4^2 + 1.4^2 + (-0.6)^2$$

$$+(-4.6)^2 + (-2.6)^2 + (-6.6)^2 + (-6.6)^2 + (-0.6)^2$$

$$70.56 + 54.76 + 19.36 + 1.96 + 0.36 + 21.16 + 6.67 + 43.56 + 43.56 + 0.36$$

$$262.31$$

Then the sample standard deviation is

$$s_d = \sqrt{\frac{262.31}{9}}$$

$$s_d \approx \sqrt{29.15}$$

$$s_d \approx 5.399$$

Because the population standard deviations are unknown, and/or because both sample sizes are small,  $n_1, n_2 < 30$ , the test statistic will be

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$$t \approx \frac{2.6 - 4}{\sqrt{\frac{5.399}{10}}}$$

$$t \approx -1.4 \cdot \frac{\sqrt{10}}{5.399}$$

$$t \approx -0.82$$

and the degrees of freedom are

$$df = n - 1 = 10 - 1 = 9$$

At a significance level of 1% (a confidence level of 99%) for an upper-tailed test, and  $df = 9$ , the  $t$ -table gives 2.821.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

The car company's  $t$ -test statistic  $t \approx -0.82$  doesn't meet the threshold  $t = 2.821$ , so the critical value approach tells them that they can't reject the null hypothesis, and therefore can't conclude that their new engine adds 4 miles per gallon.

- 3. We want to test the claim that listening to classical music while studying makes students complete their homework faster. We ask 10

students to study in silence for the first semester, and study with classical music for the second semester, then we record the mean number of hours spent on homework per week in each semester.

Student	1	2	3	4	5	6	7	8	9	10
In silence	14	13	16	21	15	19	11	20	19	16
With music	12	13	15	22	16	19	8	17	18	17
Difference, $d$	-2	0	-1	1	1	0	-3	-3	-1	1
$d^2$	4	0	1	1	1	0	9	9	1	1

Can we conclude at a 10% significance level that studying with classical music reduces the number of hours spent per week on homework?

*Solution:*

We'll define the hours spent studying in silence as Population 1, and the hours spent studying with classical music as Population 2, and our null and alternative hypotheses will be

$$H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

where  $\mu_1$  is the mean number hours spent studying in silence, and  $\mu_2$  is the mean number of hours spent studying with classical music. And because  $\mu_1 - \mu_2$  is the difference in study time, the hypothesis statements could also be written as

$$H_0 : \mu_d \leq 0$$

$$H_a : \mu_d > 0$$

where  $\mu_d$  is the mean difference between the two populations.

To find the mean difference, we'll sum the differences and divide by the number of matched-pairs in our sample,  $n = 10$ .

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{-2 + 0 + (-1) + 1 + 1 + 0 + (-3) + (-3) + (-1) + 1}{10} = -\frac{7}{10} = -0.7$$

So the sample mean tells us that mean difference is  $-0.7$  studying hours. Then the sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

To calculate this, we'll first find

$$\sum_{i=1}^n (d_i - \bar{d})^2$$

$$(-2 - (-0.7))^2 + (0 - (-0.7))^2 + (-1 - (-0.7))^2 + (1 - (-0.7))^2$$

$$+(1 - (-0.7))^2 + (0 - (-0.7))^2 + (-3 - (-0.7))^2 + (-3 - (-0.7))^2$$

$$+(-1 - (-0.7))^2 + (1 - (-0.7))^2$$

$$(-2 + 0.7)^2 + (0 + 0.7)^2 + (-1 + 0.7)^2 + (1 + 0.7)^2 + (1 + 0.7)^2$$

$$+(0 + 0.7)^2 + (-3 + 0.7)^2 + (-3 + 0.7)^2 + (-1 + 0.7)^2 + (1 + 0.7)^2$$



$$(-1.3)^2 + 0.7^2 + (-0.3)^2 + 1.7^2 + 1.7^2 + 0.7^2 + (-2.3)^2 + (-2.3)^2 + (-0.3)^2 + 1.7^2$$

$$1.69 + 0.49 + 0.09 + 2.89 + 2.89 + 0.49 + 5.29 + 5.29 + 0.09 + 2.89$$

22.1

Then the sample standard deviation is

$$s_d = \sqrt{\frac{22.1}{9}}$$

$$s_d \approx \sqrt{2.46}$$

$$s_d \approx 1.567$$

Because the population standard deviations are unknown, and/or because both sample sizes are small,  $n_1, n_2 < 30$ , the test statistic will be

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$$t = \frac{-0.7 - 0}{\frac{1.567}{\sqrt{10}}}$$

$$t = -0.7 \cdot \frac{\sqrt{10}}{1.567}$$

$$t \approx -1.413$$

and the degrees of freedom are

$$\text{df} = n - 1 = 10 - 1 = 9$$



At a significance level of 10% (a confidence level of 90%) for an upper-tailed test, and  $df = 9$ , the  $t$ -table gives 1.383.

df	Upper-tail probability $p$									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Our  $t$ -test statistic  $t \approx -1.413$  meets the threshold  $t = 1.383$ , so the critical value approach tells us that we can reject the null hypothesis, and therefore conclude that studying with classical music reduces time spent on homework.

- 4. A clothing store wants to test the claim that customers who join their VIP program return less merchandise. They track the mean monthly merchandise returns of 10 customers for one year before and after joining the VIP program, then record the mean returns per month.

Customer	1	2	3	4	5	6	7	8	9	10
Before VIP	12	55	48	23	97	103	33	44	17	29
After VIP	15	44	35	20	100	97	30	41	24	40
Difference, d	3	-11	-13	-3	3	-6	-3	-3	7	11
$d^2$	9	121	169	9	9	36	9	9	49	121

Can they conclude at a 5 % significance level that joining the VIP program reduces the amount of merchandise returns?

*Solution:*

The clothing store will define the values for returns before enrolling in the VIP program as Population 1, and the values for returns after enrolling in the VIP program as Population 2, and their null and alternative hypotheses will be

$$H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

where  $\mu_1$  is the mean monthly merchandise returns before the VIP program, and  $\mu_2$  is the mean monthly merchandise returns after the VIP program. And because  $\mu_1 - \mu_2$  is the difference in monthly merchandise returns, the hypothesis statements could also be written as

$$H_0 : \mu_d \leq 0$$

$$H_a : \mu_d > 0$$

where  $\mu_d$  is the mean difference between the two populations.

To find the mean difference, we'll sum the differences and divide by the number of matched-pairs in our sample,  $n = 10$ .

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{3 + (-11) + (-13) + (-3) + 3 + (-6) + (-3) + (-3) + 7 + 11}{10} = -\frac{15}{10} = -1.5$$



So the sample mean tells us that mean difference is  $-1.5$  in merchandise returns. Then the sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

To calculate this, we'll first find

$$\sum_{i=1}^n (d_i - \bar{d})^2$$

$$(3 - (-1.5))^2 + (-11 - (-1.5))^2 + (-13 - (-1.5))^2 + (-3 - (-1.5))^2 \\ + (3 - (-1.5))^2 + (-6 - (-1.5))^2 + (-3 - (-1.5))^2 + (-3 - (-1.5))^2 \\ + (7 - (-1.5))^2 + (11 - (-1.5))^2 \\ (3 + 1.5)^2 + (-11 + 1.5)^2 + (-13 + 1.5)^2 + (-3 + 1.5)^2 + (3 + 1.5)^2 \\ + (-6 + 1.5)^2 + (-3 + 1.5)^2 + (-3 + 1.5)^2 + (7 + 1.5)^2 + (11 + 1.5)^2 \\ 4.5^2 + (-9.5)^2 + (-11.5)^2 + (-1.5)^2 + 4.5^2 \\ + (-4.5)^2 + (-1.5)^2 + (-1.5)^2 + 8.5^2 + 12.5^2$$

$$20.25 + 90.25 + 132.25 + 2.25 + 20.25 + 20.25 + 2.25 + 2.25 + 72.25 + 156.25$$

$$518.5$$

Then the sample standard deviation is

$$s_d = \sqrt{\frac{518.5}{9}}$$

$$s_d \approx \sqrt{57.61}$$

$$s_d \approx 7.59$$

Because the population standard deviations are unknown, and/or because both sample sizes are small,  $n_1, n_2 < 30$ , the test statistic will be

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$$t \approx \frac{-1.5 - 0}{\frac{7.59}{\sqrt{10}}}$$

$$t \approx -1.5 \cdot \frac{\sqrt{10}}{7.59}$$

$$t \approx -0.625$$

and the degrees of freedom are

$$df = n - 1 = 10 - 1 = 9$$

At a significance level of 5% (a confidence level of 95%) for an upper-tailed test, and  $df = 9$ , the  $t$ -table gives 1.833.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									



The clothing store's  $t$ -test statistic  $t \approx -0.625$  doesn't meet the threshold  $t = 1.833$ , so the critical value approach tells them that they can't reject the null hypothesis, and therefore can't conclude that their VIP program causes customers to return less merchandise.

- 5. If the mean difference is  $\bar{d} = 10$  on a sample of  $n = 25$  with sample standard deviation  $s_d = 2.5$ , calculate the 95 % confidence interval around  $\bar{d}$ .

*Solution:*

For 95 % confidence with  $df = n - 1 = 25 - 1 = 24$ , and because we have a small sample  $n < 30$ , the confidence interval will be

$$(a, b) = \bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

$$(a, b) = 10 \pm 2.064 \left( \frac{2.5}{\sqrt{25}} \right)$$

$$(a, b) = 10 \pm 2.064(0.5)$$

$$(a, b) = 10 \pm 1.032$$

Then the confidence interval will be

$$(a, b) = (10 - 1.032, 10 + 1.032)$$

$$(a, b) = (8.968, 11.032)$$

So we're 95 % confident that the mean difference falls between 8.968 and 11.032.

- 6. If the mean difference is  $\bar{d} = 24$  on a sample of  $n = 49$  with population standard deviation  $\sigma_d = 3.2$ , calculate the 99 % confidence interval around  $\bar{d}$ .

*Solution:*

For 99 % confidence with critical values  $z = \pm 2.58$  and a large sample  $n \geq 30$ , the confidence interval will be

$$(a, b) = \bar{d} \pm z_{\alpha/2} \frac{\sigma_d}{\sqrt{n}}$$

$$(a, b) = 24 \pm 2.58 \left( \frac{3.2}{\sqrt{49}} \right)$$

$$(a, b) \approx 24 \pm 2.58(0.457)$$

$$(a, b) \approx 24 \pm 1.179$$

Then the confidence interval will be

$$(a, b) \approx (24 - 1.179, 24 + 1.179)$$

$$(a, b) \approx (22.821, 25.179)$$

So we're 99 % confident that the mean difference falls between 22.821 and 25.179.



## CONFIDENCE INTERVAL FOR THE DIFFERENCE OF PROPORTIONS

- 1. Given  $x_1 = 54$  successes in the first sample  $n_1 = 150$ , and  $x_2 = 47$  successes in the second sample  $n_2 = 160$ , calculate a 95 % confidence interval.

*Solution:*

The sample proportions are

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{54}{150} = 0.36$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{47}{160} \approx 0.294$$

At 95 % confidence, the critical  $z$ -values are  $z = \pm 1.96$ , so the confidence interval will be

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) \approx (0.36 - 0.294) \pm 1.96 \sqrt{\frac{0.36(1 - 0.36)}{150} + \frac{0.294(1 - 0.294)}{160}}$$

$$(a, b) \approx 0.066 \pm 1.96 \sqrt{\frac{0.36(0.64)}{150} + \frac{0.294(0.706)}{160}}$$

$$(a, b) \approx 0.066 \pm 0.104$$

Then the 95 % confidence interval is

$$(a, b) \approx (0.066 - 0.104, 0.066 + 0.104)$$

$$(a, b) \approx (-0.038, 0.170)$$

Because the confidence interval includes 0, we can't conclude that there's a difference between the population proportions.

- 2. A light bulb manufacturer wants to know whether their own bulbs last longer than a competitor's bulb. They randomly sampled 150 people who bought their bulb, and 72 of them reported that it lasted longer than 250 days. They randomly sampled 150 people who bought the competitor's bulb, and 69 of them reported that it lasted for more than 250 days. Find a 90 % confidence interval around the difference of proportions.

*Solution:*

The sample proportions are

$$\hat{p}_1 = \frac{72}{150} = 0.48$$

$$\hat{p}_2 = \frac{69}{150} = 0.46$$

At 90 % confidence, the critical  $z$ -values are  $z = \pm 1.65$ , so the confidence interval will be



$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.48 - 0.46) \pm 1.65 \sqrt{\frac{0.48(1 - 0.48)}{150} + \frac{0.46(1 - 0.46)}{150}}$$

$$(a, b) = 0.02 \pm 1.65 \sqrt{\frac{0.48(0.52)}{150} + \frac{0.46(0.54)}{150}}$$

$$(a, b) \approx 0.02 \pm 0.095$$

Then the 90% confidence interval is

$$(a, b) \approx (0.02 - 0.095, 0.02 + 0.095)$$

$$(a, b) \approx (-0.075, 0.115)$$

Because the confidence interval includes 0, we can't conclude that there's a difference between the proportion of bulbs from each company that last longer than 250 days.

- 3. A research team wants to know whether Vitamin C shortens recovery time from the common cold. They chose 100 patients with the common cold and randomly assigned 50 of them to the Vitamin C treatment group and 50 of them to the placebo group. In the Vitamin C group, 38 patients recovered in less than 7 days, while 24 patients in the placebo group recovered in less than 7 days. Find a 99% confidence interval around the difference in population proportions.



*Solution:*

The sample proportions are

$$\hat{p}_1 = \frac{38}{50} = 0.76$$

$$\hat{p}_2 = \frac{24}{50} = 0.48$$

At 99 % confidence, the critical  $z$ -values are  $z = \pm 2.58$ , so the confidence interval will be

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.76 - 0.48) \pm 2.58 \sqrt{\frac{0.76(1 - 0.76)}{50} + \frac{0.48(1 - 0.48)}{50}}$$

$$(a, b) = 0.28 \pm 2.58 \sqrt{\frac{0.76(0.24)}{50} + \frac{0.48(0.52)}{50}}$$

$$(a, b) \approx 0.28 \pm 0.24$$

Then the 99 % confidence interval is

$$(a, b) \approx (0.28 - 0.24, 0.28 + 0.24)$$

$$(a, b) \approx (0.04, 0.52)$$

We can be 99 % confident that the true difference between population proportions is between 0.04 and 0.52. Which means we can be 99 %



confident that the Vitamin C treatment shortens recovery time from the common cold.

- 4. A researcher randomly chose 900 smokers, 450 men and 450 women. He found that 357 of the male smokers have been diagnosed with coronary artery disease, while 295 of the female smokers have been diagnosed with coronary artery disease. Construct a 95 % confidence interval to estimate the difference between the proportions of male and female smokers who have been diagnosed with coronary artery disease.

*Solution:*

The sample proportions are

$$\hat{p}_1 = \frac{357}{450} \approx 0.793$$

$$\hat{p}_2 = \frac{295}{450} \approx 0.656$$

At 95 % confidence, the critical  $z$ -values are  $z = \pm 1.96$ , so the confidence interval will be

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.793 - 0.656) \pm 1.96 \sqrt{\frac{0.793(1 - 0.793)}{450} + \frac{0.656(1 - 0.656)}{450}}$$



$$(a, b) = 0.137 \pm 1.96 \sqrt{\frac{0.793(0.207)}{450} + \frac{0.656(0.344)}{450}}$$

$$(a, b) \approx 0.137 \pm 0.058$$

Then the 95 % confidence interval is

$$(a, b) \approx (0.137 - 0.058, 0.137 + 0.058)$$

$$(a, b) \approx (0.079, 0.195)$$

We can be 95 % confident that the true difference between the proportion of male smokers with coronary artery disease and the proportion of female smokers with coronary artery disease is between 0.079 and 0.195.

- 5. In a simple random sample of 1,000 people aged 20 – 24, 7 % said they ran at least one marathon in the last year. In a simple random sample of 1,200 people aged 25 – 29, 12 % said they ran at least one marathon in the last year. Find a 99 % confidence interval around the difference of population proportions.

*Solution:*

With  $\hat{p}_1 = 0.07$  for  $n_1 = 1,000$  and  $\hat{p}_2 = 0.12$  for  $n_2 = 1,200$ , and critical values of  $z = \pm 2.58$  for a 99 % confidence level, the confidence interval will be

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.07 - 0.12) \pm 2.58 \sqrt{\frac{0.07(1 - 0.07)}{1,000} + \frac{0.12(1 - 0.12)}{1,200}}$$

$$(a, b) = -0.05 \pm 2.58 \sqrt{\frac{0.07(0.93)}{1,000} + \frac{0.12(0.88)}{1,200}}$$

$$(a, b) \approx -0.05 \pm 0.032$$

Then the 99 % confidence interval is

$$(a, b) \approx (-0.05 - 0.032, -0.05 + 0.032)$$

$$(a, b) \approx (-0.082, -0.018)$$

We can be 99 % confident that the true proportion is between -0.082 and -0.018. Which means it's likely that more people aged 25 – 29 ran at least one marathon in the last year, compared to people aged 20 – 24.

- 6. In a simple random sample of 280 Masters students from one university, 24 said they planned to pursue a PhD. In a simple random sample of 350 Masters students at a second university, 34 said they planned to pursue a PhD. Build a 98 % confidence interval around the difference of proportions.

*Solution:*

The sample proportions are



$$\hat{p}_1 = \frac{24}{280} \approx 0.086$$

$$\hat{p}_2 = \frac{34}{350} \approx 0.097$$

At 98 % confidence, the critical  $z$ -values are  $z = \pm 2.33$ , so the confidence interval will be

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.086 - 0.097) \pm 2.33 \sqrt{\frac{0.086(1 - 0.086)}{280} + \frac{0.097(1 - 0.097)}{350}}$$

$$(a, b) = -0.011 \pm 2.33 \sqrt{\frac{0.086(0.914)}{280} + \frac{0.097(0.903)}{350}}$$

$$(a, b) \approx -0.011 \pm 0.054$$

Then the 98 % confidence interval is

$$(a, b) \approx (-0.011 - 0.054, -0.011 + 0.054)$$

$$(a, b) \approx (-0.065, 0.043)$$

Because the confidence interval includes 0, we can't conclude that there's a difference between the proportion of Masters students at each university who want to pursue a PhD.

## HYPOTHESIS TESTING FOR THE DIFFERENCE OF PROPORTIONS

- 1. We defined the hypothesis statements below, and then found sample proportions of  $\hat{p}_1 = 0.456$  for  $n_1 = 278$  and  $\hat{p}_2 = 0.384$  for  $n_2 = 310$ . Using a critical value approach, can we reject the null hypothesis at a confidence level of 95 % ?

$$H_0 : p_1 - p_2 \leq 0$$

$$H_a : p_1 - p_2 > 0$$

*Solution:*

The pooled proportion is

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$\hat{p} = \frac{0.456(278) + 0.384(310)}{278 + 310}$$

$$\hat{p} \approx 0.418$$

Then the  $z$ -statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



$$z \approx \frac{0.456 - 0.384}{\sqrt{0.418(1 - 0.418)\left(\frac{1}{278} + \frac{1}{310}\right)}}$$

$$z \approx \frac{0.072}{\sqrt{0.418(0.582)\left(\frac{1}{278} + \frac{1}{310}\right)}}$$

$$z \approx 1.767$$

For an upper-tailed test at a confidence level of 95 % , the critical value is  $z = 1.65$ . Since  $1.767 > 1.65$ , we can reject the null hypothesis and conclude that  $p_1 > p_2$  at a 95 % confidence level.

- 2. Given the hypothesis statements below,  $x_1 = 234$  with  $n_1 = 1,150$  and  $x_2 = 327$  with  $n_2 = 1,320$ , calculate the test statistic.

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 \neq 0$$

*Solution:*

First calculate the sample proportions.

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{234}{1,150} \approx 0.203$$



$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{327}{1,320} \approx 0.248$$

The pooled proportion is

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$\hat{p} = \frac{0.203(1,150) + 0.248(1,320)}{1,150 + 1,320}$$

$$\hat{p} \approx 0.227$$

Then the  $z$ -statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z \approx \frac{0.203 - 0.248}{\sqrt{0.227(1 - 0.227)\left(\frac{1}{1,150} + \frac{1}{1,320}\right)}}$$

$$z \approx \frac{-0.045}{\sqrt{0.227(0.773)\left(\frac{1}{1,150} + \frac{1}{1,320}\right)}}$$

$$z \approx -2.663$$



3. A cinema owner wants to know whether there's a difference in the number of boys and girls who watched a new movie last week. She randomly sampled 76 boys and 75 girls and found that 45 boys and 58 girls watched the movie. What can she conclude about the difference of proportions at a 99 % confidence level?

*Solution:*

The owner is running a two-tailed test, so her hypothesis statements will be

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 \neq 0$$

The sample proportions are

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{45}{76} \approx 0.592$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{58}{75} \approx 0.773$$

Then the pooled proportion is

$$\hat{p} = \frac{45 + 58}{76 + 75} = \frac{103}{151} \approx 0.682$$

$$\hat{p} \approx 0.682$$

and the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.592 - 0.773}{\sqrt{0.682(1 - 0.682)\left(\frac{1}{76} + \frac{1}{75}\right)}}$$

$$z = \frac{-0.181}{\sqrt{0.682(0.318)\left(\frac{1}{76} + \frac{1}{75}\right)}}$$

$$z \approx -2.39$$

For a two-tailed test at 99 % confidence, the critical value will be  $z = -2.58$ . Because  $-2.58 < -2.39$ , the cinema owner fails to reject the null hypothesis, and can't conclude that more boys than girls watched the new movie last week.

- 4. A store owner believes that women spend at least 22 % more in his store than men. He randomly chooses 64 visitors, 32 men and 32 women, and finds that 14 men spent more than \$100, while 23 women spent more than \$100. Using a  $p$ -value approach, what can he conclude at a 90 % confidence level?

*Solution:*

Assuming  $p_1$  is the population proportion of women who spend more than \$100 and  $p_2$  is the population proportion of men who spend more than \$100, the store owner's hypothesis statements for the upper-tailed test will be

$$H_0 : p_1 - p_2 \leq 0.22$$

$$H_a : p_1 - p_2 > 0.22$$

The sample proportions are

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{23}{32} \approx 0.7188$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{14}{32} = 0.4375$$

The pooled proportion is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{p} = \frac{23 + 14}{32 + 32}$$

$$\hat{p} \approx 0.578$$

Then the test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



$$z \approx \frac{(0.7188 - 0.4375) - 0.22}{\sqrt{0.578(1 - 0.578)\left(\frac{1}{32} + \frac{1}{32}\right)}}$$

$$z \approx \frac{0.0613}{\sqrt{0.578(0.422)\left(\frac{1}{16}\right)}}$$

$$z \approx 0.50$$

This  $z$ -value gives

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	<b>.6915</b>	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

we can see that the area to the left of  $z = 0.50$  is 0.6915. Because this is an upper-tailed test, we're interested in the area to the left of  $z = 0.50$ , so  $1 - 0.6915 = 0.3085$ . Therefore,  $p = 0.3085$ . Since  $0.3085 > 0.1$ , the store owner fails to reject the null hypothesis. There's not enough evidence to conclude that women spend 22% more than men.

- 5. In a random sample of 60 people under the age of 30, 14% said they're planning to go hiking next month. In a random sample of 75 people older than 50, 23% said they're planning to go hiking next month. Using a critical value approach at a 95% confidence level, is there enough evidence to conclude that a higher proportion of people over age 50 plan to go hiking next month than the proportion of people under 30 who plan to go hiking?



*Solution:*

If  $p_1$  is the proportion of people under 30 who plan to hike, and  $p_2$  is the proportion of people over 50 who plan to hike, then the hypothesis statements are

$$H_0 : p_1 - p_2 \geq 0$$

$$H_a : p_1 - p_2 < 0$$

The sample proportions are

$$\hat{p}_1 = 0.14$$

$$\hat{p}_2 = 0.23$$

The pooled proportion is

$$\hat{p} = \frac{0.14(60) + 0.23(75)}{60 + 75}$$

$$\hat{p} = 0.19$$

Then the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



$$z = \frac{0.14 - 0.23}{\sqrt{0.19(1 - 0.19)\left(\frac{1}{60} + \frac{1}{75}\right)}}$$

$$z = \frac{-0.09}{\sqrt{0.19(0.81)\left(\frac{1}{60} + \frac{1}{75}\right)}}$$

$$z = -1.325$$

For a 95 % confidence interval and a lower-tailed test, the critical  $z$ -value is  $z = -1.65$ . Since  $-1.325 > -1.65$ , we fail to reject the null hypothesis. There's not enough evidence to conclude that the proportion of people over 50 who plan to hike is higher than the proportion of people under 30 who plan to hike.

- 6. John and Steven are two fitness trainers who want to compare their client satisfaction rate. John chose a random sample of 85 clients and Steven chose a random sample of 72 clients. John found that 89 % of his clients were satisfied and Steve found that 91 % of his clients were satisfied. Using a critical value approach at a 95 % confidence level, is there a significant difference between proportions?

*Solution:*

John and Steven are running a two-tailed test, so their hypothesis statements are



$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 \neq 0$$

The sample proportions are

$$\hat{p}_1 = 0.89$$

$$\hat{p}_2 = 0.91$$

The pooled proportion is

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$\hat{p} = \frac{0.89(85) + 0.91(72)}{85 + 72}$$

$$\hat{p} \approx 0.899$$

Then the test statistic is

$$z = \frac{0.89 - 0.91}{\sqrt{0.899(1 - 0.899)\left(\frac{1}{85} + \frac{1}{72}\right)}}$$

$$z = \frac{-0.02}{\sqrt{0.899(0.101)\left(\frac{1}{85} + \frac{1}{72}\right)}}$$

$$z \approx -0.414$$



For a 95 % confidence level and a two-tailed test, the critical  $z$ -values are  $z = \pm 1.96$ . Since  $-0.414$  falls between  $-1.96$  and  $1.96$ , we fail to reject the null hypothesis. There's not enough evidence to conclude that there's a significant difference between John and Steven's client satisfaction rate.



## SCATTERPLOTS AND REGRESSION

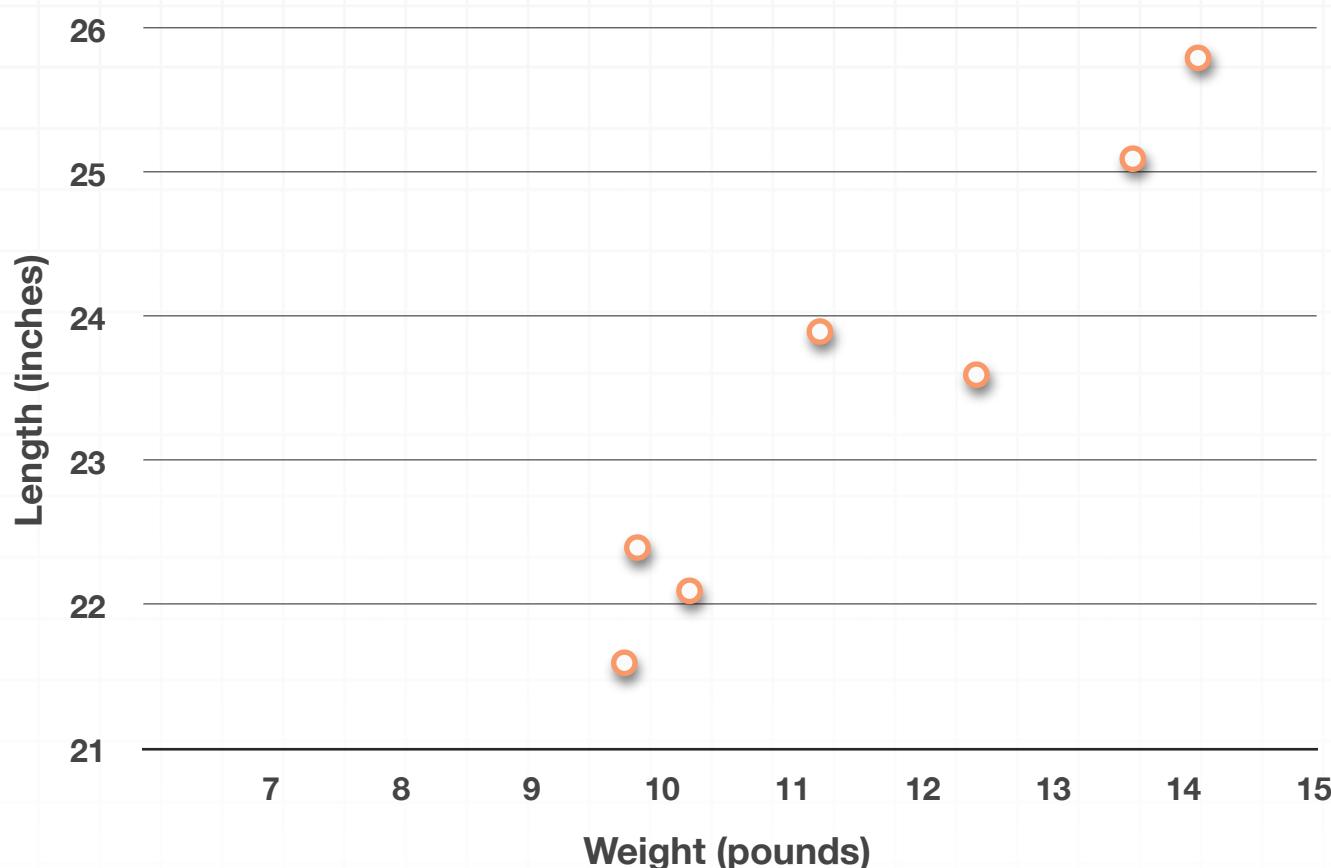
- 1. The table gives weight in pounds and length in inches for 3-month-old baby girls. Graph the points from the table in a scatterplot and describe the trend.

Weight (lbs)	Length (in)
9.7	21.6
10.2	22.1
12.4	23.6
13.6	25.1
9.8	22.4
11.2	23.9
14.1	25.8

*Solution:*

Sketch the scatterplot.

### 3-month-old baby girls



The points rise from left to right and are fairly linear. We can say that there is a strong positive linear correlation between the points. There do not appear to be any outliers in the data.

- 2. The following values have been computed for a data set of 14 points. Calculate the line of best fit.

$$\sum x = 86$$

$$\sum y = 89.7$$

$$\sum xy = 680.46$$

$$\sum x^2 = 654.56$$

*Solution:*

We're told that there are 14 items in the data set, so  $n = 14$ .

To find the line of best fit, we need its slope and  $y$ -intercept. The slope is given by

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{14(680.46) - (86)(89.7)}{14(654.56) - (86)^2}$$

$$b = \frac{9,526.44 - 7,714.2}{9,163.84 - 7,396}$$

$$b = \frac{1,812.24}{1,767.84}$$

$$b \approx 1.0251$$

The  $y$ -intercept is given by

$$a = \frac{\sum y - b \sum x}{n}$$

$$a = \frac{89.7 - 1.0251(86)}{14}$$

$$a = \frac{89.7 - 88.1599}{14}$$



$$a = \frac{1.5401}{14}$$

$$a \approx 0.1100$$

So the line of best fit is

$$\hat{y} = bx + a$$

$$\hat{y} = 1.0251x + 0.1100$$

- 3. For the data set given in the table, calculate each of the following values:

$$n, \sum x, \sum y, \sum xy, \sum x^2, \left( \sum x \right)^2$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	73	73	75	75	77	79	79	81	81	81	77	75

*Solution:*

Expand the original table to calculate the values.



Month, x	Temperature, y	xy	$x^2$
1	73	$1(73)=73$	$1^2=1$
2	73	$2(73)=146$	$2^2=4$
3	75	$3(75)=225$	$3^2=9$
4	75	$4(75)=300$	$4^2=16$
5	77	$5(77)=385$	$5^2=25$
6	79	$6(79)=474$	$6^2=36$
7	79	$7(79)=553$	$7^2=49$
8	81	$8(81)=648$	$8^2=64$
9	81	$9(81)=729$	$9^2=81$
10	81	$10(81)=810$	$10^2=100$
11	77	$11(77)=847$	$11^2=121$
12	75	$12(75)=900$	$12^2=144$

Summing the first column gives

$$\sum x = 78$$

Summing the second column gives

$$\sum y = 926$$

Summing the third column gives

$$\sum xy = 6,090$$

Summing the fourth column gives

$$\sum x^2 = 650$$

Squaring the sum from the first column gives

$$\left(\sum x\right)^2 = 78^2 = 6,084$$



4. Use the Average Global Sea Surface Temperatures data shown in the table to create a line of best fit for the data. Consider 1910 as year 10. Use the equation to predict the average global sea surface temperature in the year 2050.

Year	Temperature, F
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

*Solution:*

Start by expanding the table.

Year	Temperature, F	xy	$x^2$
10	-1.11277	-11.1277	100
20	-0.71965	-14.393	400
30	-0.58358	-17.5074	900
40	-0.17977	-7.1908	1,600
50	-0.55318	-27.659	2,500
60	-0.30358	-18.2148	3,600
70	-0.30863	-21.6041	4,900
80	0.077197	6.17576	6,400
90	0.274842	24.73578	8,100
100	0.232502	23.2502	10,000
110	0.612718	67.39898	12,100

Summing the first column gives

$$\sum x = 660$$

Summing the second column gives

$$\sum y = -2.5639$$

Summing the third column gives

$$\sum xy = 3.86392$$

Summing the fourth column gives

$$\sum x^2 = 50,600$$

Squaring the sum from the first column gives

$$\left( \sum x \right)^2 = 660^2 = 435,600$$

To find the regression line for the data, we need the slope and  $y$ -intercept of the line. The slope is

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{11(3.86392) - (660)(-2.5639)}{11(50,600) - 435,600}$$

$$b = \frac{42.50312 + 1,692.174}{556,600 - 435,600}$$

$$b = \frac{1,734.67712}{121,600}$$

$$b \approx 0.0143$$

The  $y$ -intercept is

$$a = \frac{\sum y - b \sum x}{n}$$

$$a = \frac{-2.5639 - 0.0143(660)}{11}$$

$$a = \frac{-2.5639 - 9.415188}{11}$$

$$a = \frac{-11.979088}{11}$$

$$a \approx -1.0890$$

Then the equation of the trend line is

$$\hat{y} = bx + a$$

$$\hat{y} = 0.0143x - 1.0890$$

To predict average global sea surface temperature in 2050, we'll need to plug 150 into this equation.

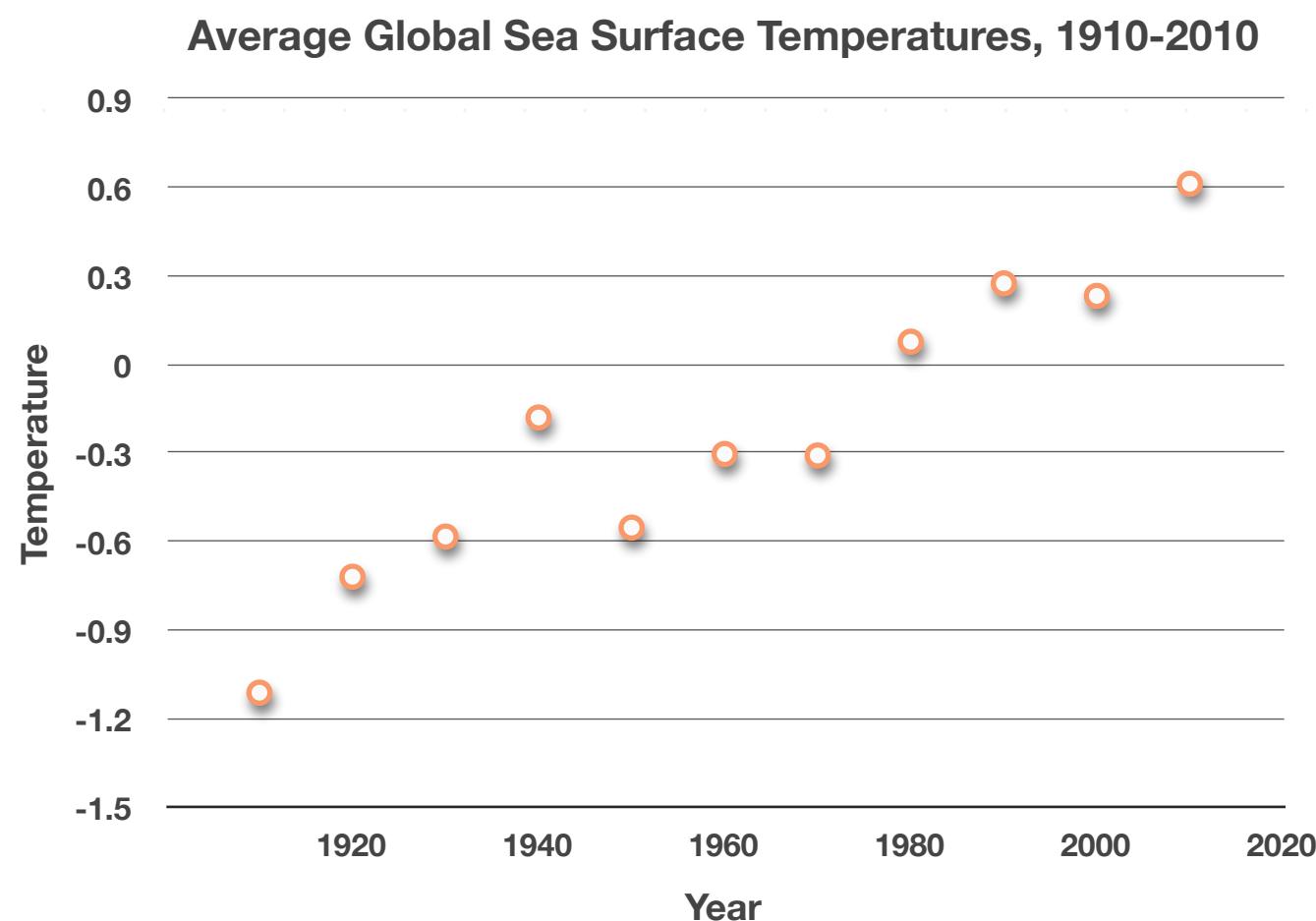
$$\hat{y} = 0.0143(150) - 1.0890$$

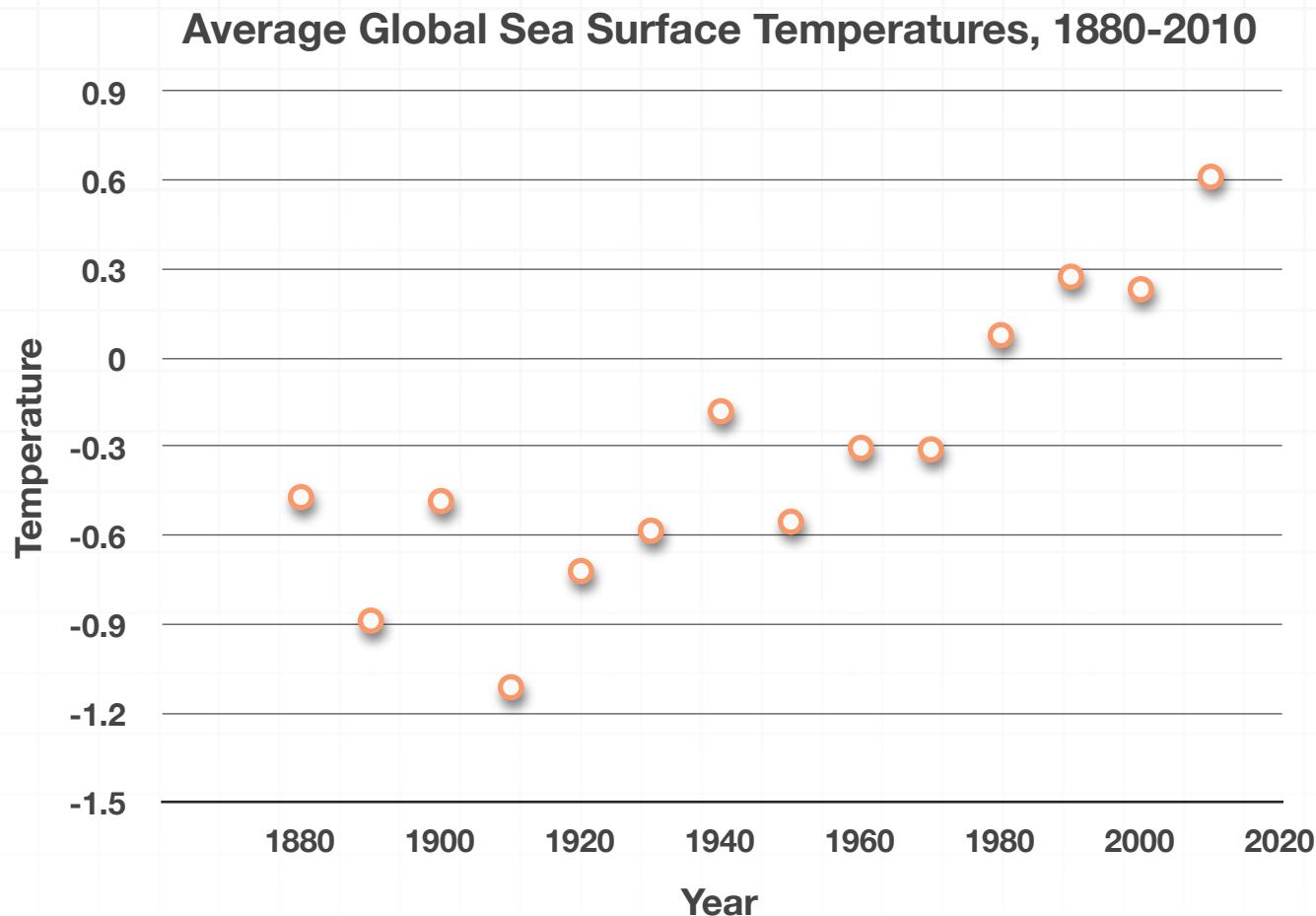
$$\hat{y} = 2.145 - 1.0890$$

$$\hat{y} = 1.056$$

So the predicted sea surface temperature in 2050 is about  $1.056^{\circ}$  F.

- 5. Compare the scatterplots. The second graph includes extra data starting in 1880. How does this compare to the plot that only shows 1910 to 2010? Explain trends in the data, and how the regression line changes by adding in these extra points. Which trend line would be best for predicting the temperature in 2050?





*Solution:*

Adding in these extra three points make the graph from 1880 to 2010 appear more scattered and not as linear as the graph that only includes the points from 1910 to 2010.

Both data sets have a positive correlation because the general trend of the scatterplot is to increase as we move from left to right, but we might consider graphs that are exponential in shape instead of linear. If we use a line of best fit for the data from 1880 to 2010, it might not be as accurate as a line predicting only the points from 1910 to 2010.

In other words, the best fit line for 1880 to 2010 would have a weaker correlation than the line for 1910 to 2010, because the additional points to the left of the graph are more spread out.

But even though cutting off the points makes the line of best fit have a stronger correlation, it would be good to include them in the data so that our line of best fit is not misleading.

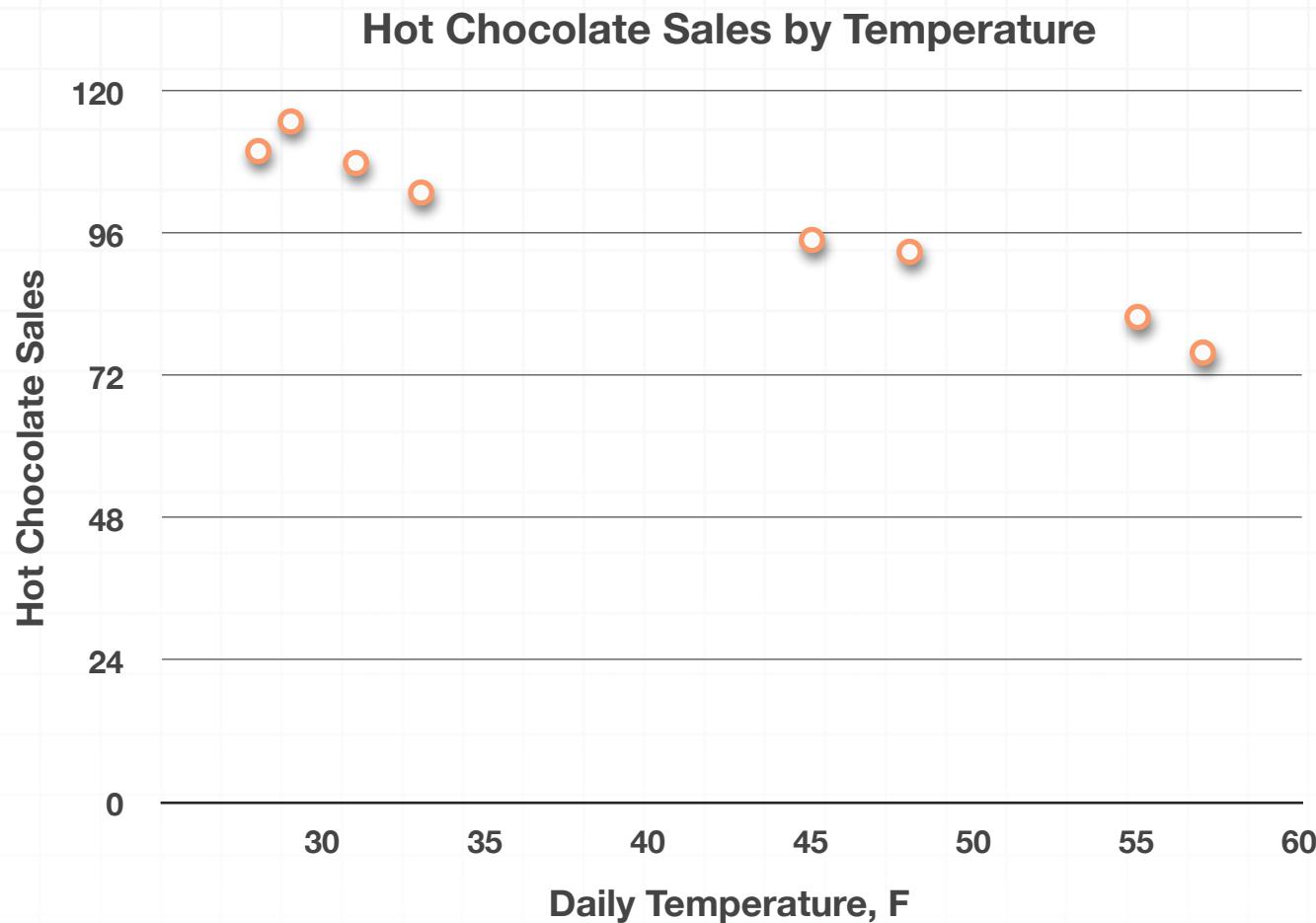
- 6. A small coffee shop wants to know how hot chocolate sales are affected by daily temperature. Find the rate of change of hot chocolate sales, with respect to temperature.

Daily Temperature, F	Hot Chocolate Sales
28	110
29	115
31	108
33	103
45	95
48	93
55	82
57	76

*Solution:*

Create a scatterplot.





From the plot we can see there's a relatively strong, negative linear relationship with no outliers. The rate of change is the slope, so we need to look at the slope of the line of best fit for the data set. The formula for the slope of the best-fit line is

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

Extend the table to find the values we need for the formula.

	Daily Temperature, F	Hot Chocolate Sales	xy	x <sup>2</sup>
	28	110	3,080	784
	29	115	3,335	841
	31	108	3,348	961
	33	103	3,399	1,089
	45	95	4,275	2,025
	48	93	4,464	2,304
	55	82	4,510	3,025
	57	76	4,332	3,249
<b>Sum:</b>	<b>326</b>	<b>782</b>	<b>30,743</b>	<b>14,278</b>

Plug these values into the slope formula.

$$b = \frac{8(30,743) - (326)(782)}{8(14,278) - (326)^2}$$

$$b = \frac{245,944 - 254,932}{114,224 - 106,276}$$

$$b = \frac{-8,988}{7,946}$$

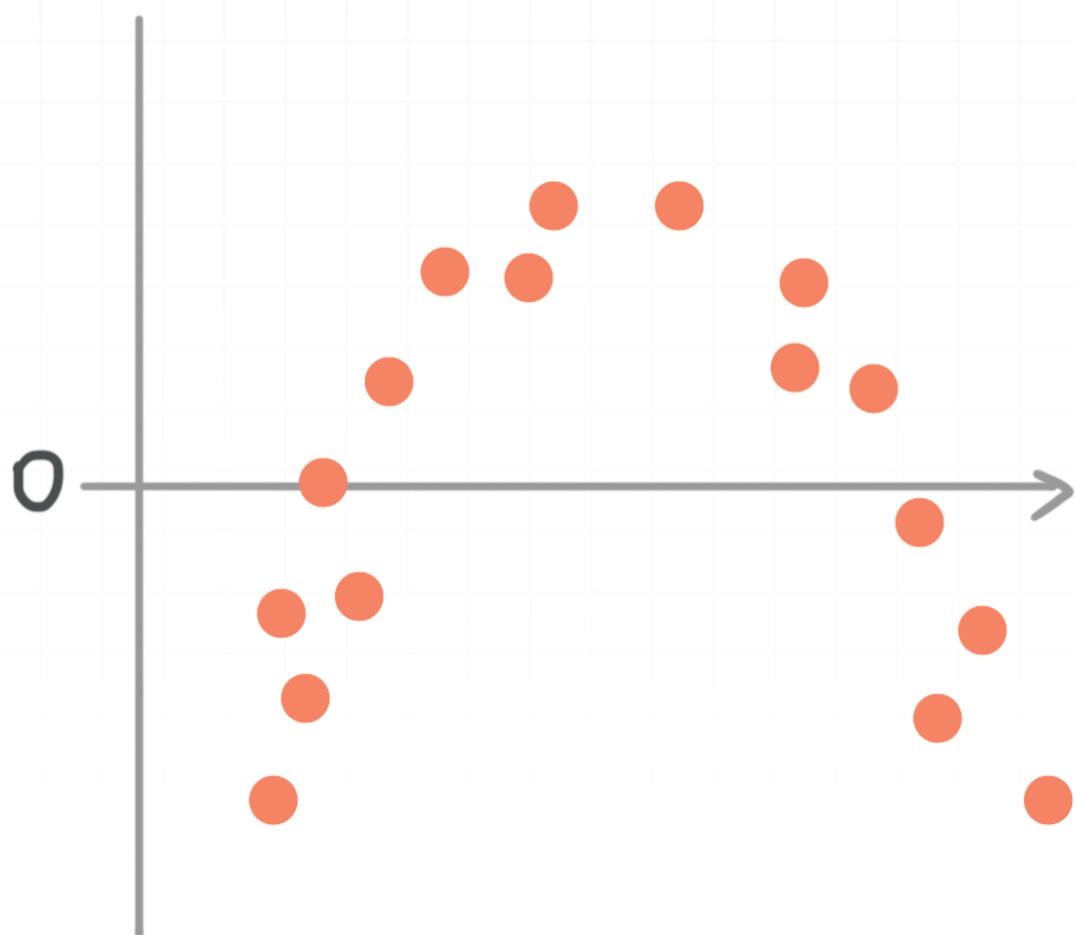
$$b \approx -1.1311$$

The units of the slope are “hot chocolate sales per degree Fahrenheit.” So the shop can expect hot chocolate sales to decrease by 1.1311 cups for every one degree increase in temperature.



## CORRELATION COEFFICIENT AND THE RESIDUAL

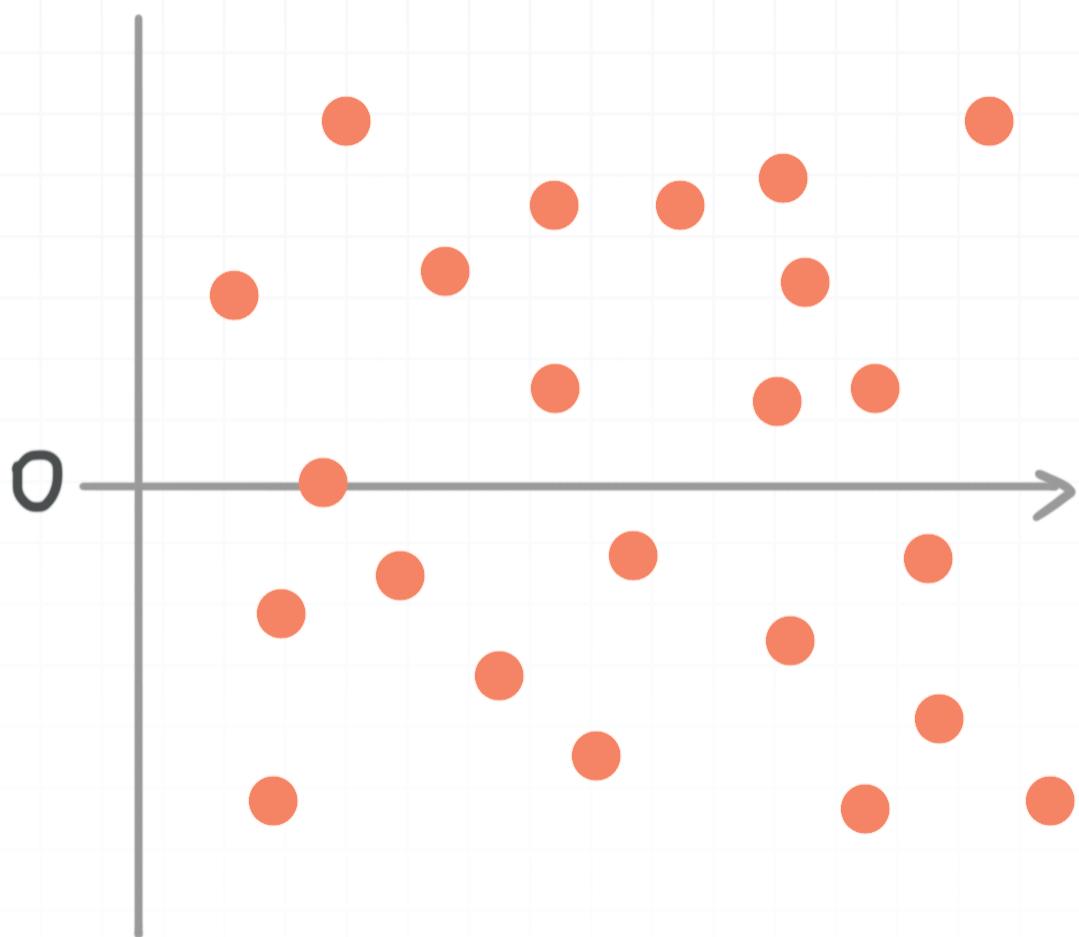
- 1. What does the shape of this residual plot tell us about the line of best fit that was created for the data?



*Solution:*

The shape of this graph tells us that the linear model is probably not the best choice for our data set, and that we should consider another type of regression curve, probably one that's quadratic.

- 2. What does the shape of this residual plot tell us about the line of best fit that was created for the data?



*Solution:*

The points in this residual plot are evenly spaced around the line  $y = 0$ . It has about the same number of points on the left and right, and about the same number of points above and below 0. It doesn't appear to have outliers or interesting features. So the line of best fit for the data is probably a good one and can be useful for making predictions.

- 3. Calculate and interpret the correlation coefficient for the data set.

x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

*Solution:*

To find the correlation coefficient, we use the formula

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

We need to start by finding the means and standard deviations for both  $x$  and  $y$ . The means are

$$\bar{x} = \frac{54 + 57 + 62 + 77 + 81 + 93}{6}$$

$$\bar{x} \approx 70.6667$$

and

$$\bar{y} = \frac{0.162 + 0.127 + 0.864 + 0.895 + 0.943 + 1.206}{6}$$

$$\bar{y} \approx 0.6995$$

and the standard deviations are

$$s_x = \sqrt{\frac{\sum_{i=1}^6 (x_i - \bar{x})^2}{n - 1}}$$

$$s_x \approx \sqrt{\frac{277.7790 + 186.7790 + 75.1117 + 40.1107 + 106.7770 + 498.7760}{5}}$$

$$s_x \approx 15.3970$$

and

$$s_y = \sqrt{\frac{\sum_{i=1}^6 (y_i - \bar{y})^2}{n - 1}}$$

$$s_y \approx \sqrt{\frac{0.2889 + 0.3278 + 0.0271 + 0.0382 + 0.0593 + 0.2565}{5}}$$

$$s_y \approx 0.4467$$

Plug these values into the correlation coefficient formula.

$$\begin{aligned} r = & \frac{1}{6 - 1} \left[ \left( \frac{54 - 70.6667}{15.3970} \right) \left( \frac{0.162 - 0.6995}{0.4467} \right) + \left( \frac{57 - 70.6667}{15.3970} \right) \left( \frac{0.127 - 0.6995}{0.4467} \right) \right. \\ & + \left( \frac{62 - 70.6667}{15.3970} \right) \left( \frac{0.864 - 0.6995}{0.4467} \right) + \left( \frac{77 - 70.6667}{15.3970} \right) \left( \frac{0.895 - 0.6995}{0.4467} \right) \\ & \left. + \left( \frac{81 - 70.6667}{15.3970} \right) \left( \frac{0.943 - 0.6995}{0.4467} \right) + \left( \frac{93 - 70.6667}{15.3970} \right) \left( \frac{1.206 - 0.6995}{0.4467} \right) \right] \end{aligned}$$



$$r = \frac{1}{5} \left[ (-1.0825)(-1.2033) + (-0.8876)(-1.2816) + (-0.5629)(0.3683) \right]$$

$$+ (0.4113)(0.4189) + (0.6711)(0.5217) + (1.4505)(1.1339) \right]$$

$$r = \frac{1}{5}(1.3026 + 1.1375 - 0.2073 + 0.1723 + 0.3501 + 1.6447)$$

$$r = \frac{1}{5}(4.3999)$$

$$r \approx 0.88$$

The positive correlation coefficient tells us that the regression line has a positive slope. The fact that the positive value is closer to 1 than it is to 0 tells us the data is strongly correlated, or that it most likely has a strong linear relationship. If we looked at a scatterplot of the data and sketched in the regression line, we'd see that this was true.

- 4. Calculate the residuals, draw the residual plot, and interpret the results. Compare the results to the  $r$ -value in the previous problem. The equation of the line of best fit for the data is

$$\hat{y} = 0.0257x - 1.1142$$



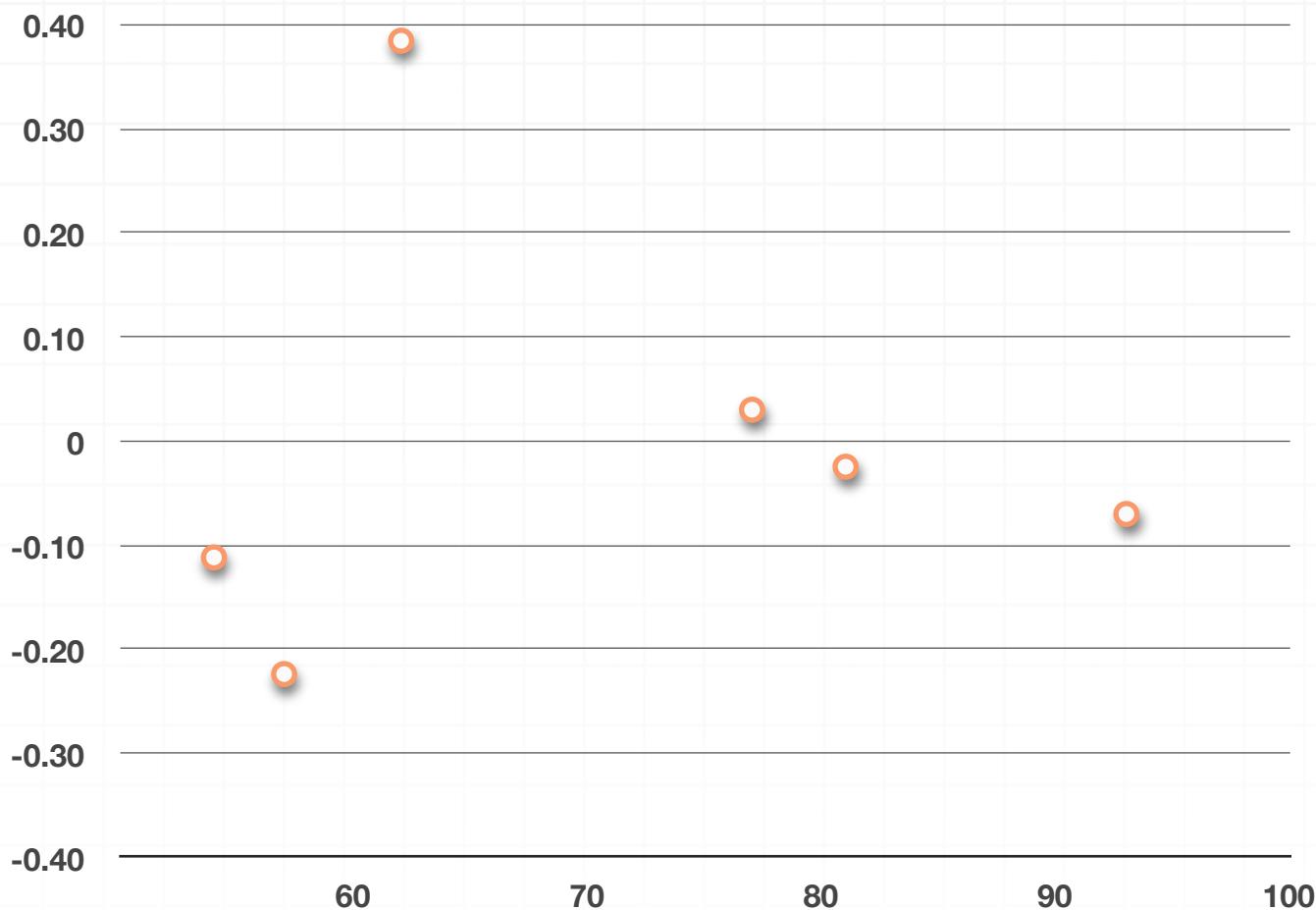
x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

*Solution:*

Create a table to find the residual of each value.

x	Actual y	Predicted y	Residual
54	0.162	0.2736	-0.1116
57	0.127	0.3507	-0.2237
62	0.864	0.4792	0.3848
77	0.895	0.8647	0.0303
81	0.943	0.9675	-0.0245
93	1.206	1.2759	-0.0699

A plot of the residuals is



From the residual plot, it looks like the data had an outlier at  $x = 62$ . We already have a somewhat strong positive linear correlation from the correlation coefficient from the previous problem of  $r \approx 0.88$ , so it's likely the relationship would be even stronger without the outlier.

- 5. The table shows average global sea surface temperature by year. Calculate and interpret the correlation coefficient for the data set. Leave the years as they are.

Year	Temperature, F
1880	-0.47001
1890	-0.88758
1900	-0.48331
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

*Solution:*

Since this is a larger set of data, it can be nice to use a program like Excel to expand the table.

	Year	Temperature, F	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	1880	-0.47001	4,225	0.02414
	1890	-0.88758	3,025	0.32827
	1900	-0.48331	2,025	0.02845
	1910	-1.11277	1,225	0.63703
	1920	-0.71965	625	0.16404
	1930	-0.58358	225	0.07233
	1940	-0.17977	25	0.01819
	1950	-0.55318	25	0.05691
	1960	-0.30358	225	0.00012
	1970	-0.30863	625	0.00004
	1980	0.077197	1,225	0.15353
	1990	0.274842	2,025	0.34748
	2000	0.232502	3,025	0.29935
	2010	0.612718	4,225	0.85997
<b>Sum:</b>	<b>27,230</b>	<b>-4.40480</b>	<b>22,750</b>	<b>2.98985</b>
<b>Mean:</b>	<b>1,945</b>	<b>-0.31463</b>		

The standard deviation for  $x$  and  $y$  are

$$s_x = \sqrt{\frac{\sum_{i=1}^{14} (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{22,750}{13}} \approx 41.8330$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{14} (y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{2.98986}{13}} \approx 0.4796$$

Now that we have the means and standard deviations, we can find the correlation coefficient. If we expand the table, then we'll be able to pull just the one sum out of the table to plug into the correlation coefficient formula.

	Year	Temp, F	$(x_i - \bar{x})$	$(x_i - \bar{x})/s_x$	$(y_i - \bar{y})$	$(y_i - \bar{y})/s_y$	$((x_i - \bar{x})/s_x)((y_i - \bar{y})/s_y)$
	1880	-0.47001	-65	-1.55380	-0.15538	-0.32398	0.50340
	1890	-0.88758	-55	-1.31475	-0.57925	-1.20778	1.58793
	1900	-0.48331	-45	-1.07571	-0.16868	-0.35171	0.37834
	1910	-1.11277	-35	-0.83666	-0.79814	-1.66418	1.39235
	1920	-0.71965	-25	-0.59761	-0.40502	-0.84450	0.50468
	1930	-0.58358	-15	-0.35857	-0.26895	-0.56078	0.20108
	1940	-0.17977	-5	-0.11952	0.13486	0.28119	-0.03361
	1950	-0.55318	5	0.11952	-0.23855	-0.49739	-0.05945
	1960	-0.30358	15	0.35857	0.01105	0.02304	0.00826
	1970	-0.30863	25	0.59761	0.00600	0.01251	0.00748
	1980	0.077197	35	0.83666	0.39183	0.81699	0.68354
	1990	0.274842	45	1.07571	0.58947	1.22909	1.32214
	2000	0.232502	55	1.31475	0.54713	1.14080	1.49987
	2010	0.612718	65	1.55380	0.92735	1.93359	3.00441
<b>Sum:</b>							<b>11.00042</b>

The correlation coefficient is then

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

$$r = \frac{1}{14 - 1}(10.98314)$$

$$r = \frac{1}{13}(10.98314)$$

$$r \approx 0.8449$$

There's a strong positive linear relationship between the year and the temperature of the ocean's surface.

- 6. Calculate the residuals and create the residual plot for the data in the table. Compare this with the  $r$ -value we calculated in the last question and interpret the results. Use the equation for the regression line

$$\hat{y} = 0.0143x - 28.332.$$



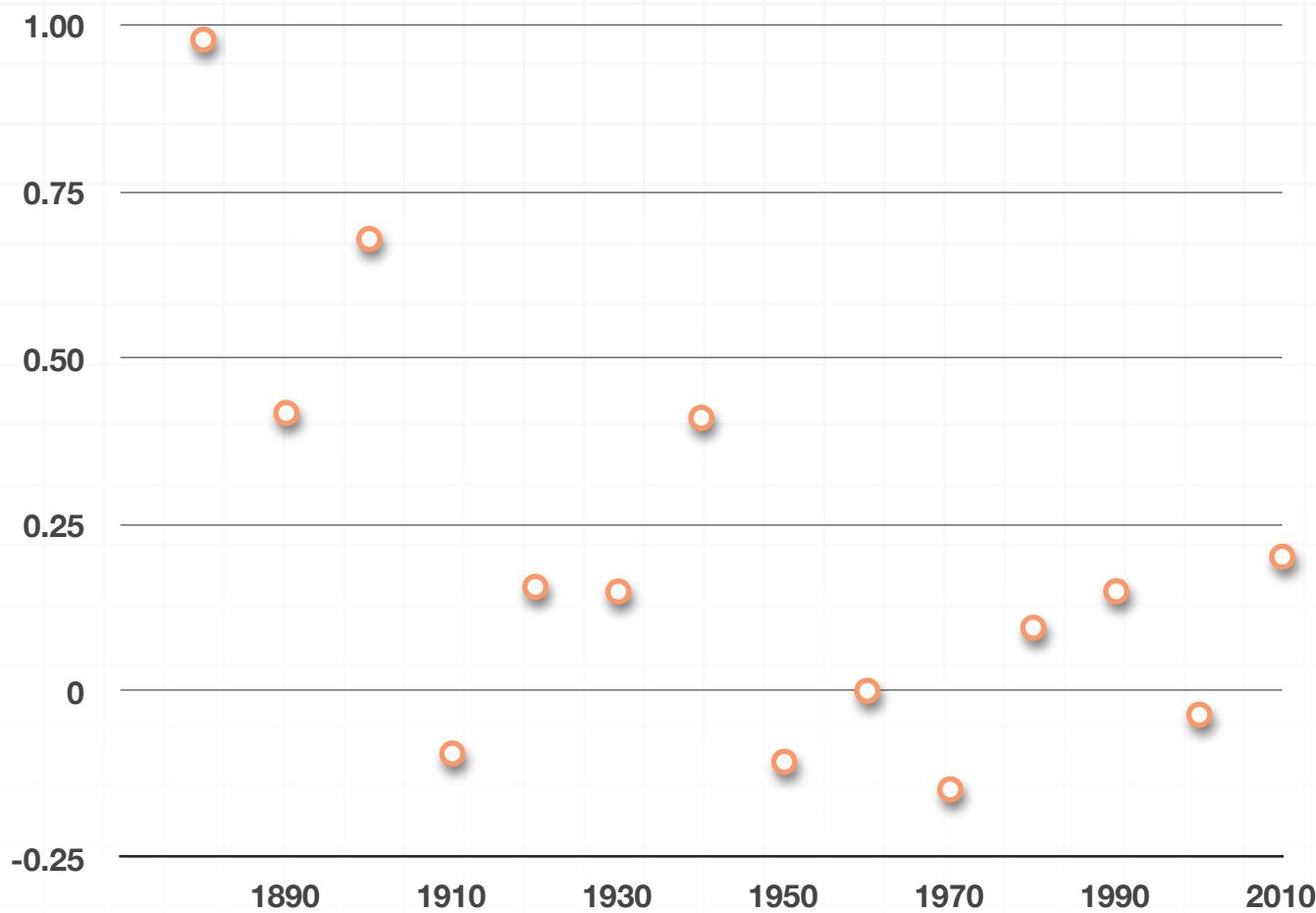
Year	Temperature, F
1880	-0.47001
1890	-0.88758
1900	-0.48331
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

*Solution:*

Use the equation of the regression line to find predicted values of temperature, and add those values to the table.

Year	Actual y	Predicted y	Residual
1880	-0.47001	-1.448	0.97799
1890	-0.88758	-1.305	0.41742
1900	-0.48331	-1.162	0.67869
1910	-1.11277	-1.019	-0.09337
1920	-0.71965	-0.876	0.15635
1930	-0.58358	-0.733	0.14942
1940	-0.17977	-0.59	0.41023
1950	-0.55318	-0.447	-0.10618
1960	-0.30358	-0.304	0.00042
1970	-0.30863	-0.161	-0.14763
1980	0.077197	-0.018	0.15917
1990	0.274842	0.125	0.149842
2000	0.232502	0.268	-0.035498
2010	0.612718	0.411	0.201718

Make a plot of the residuals.



The residual plot is a good example of why finding the correlation coefficient is not enough.

This plot makes it look like using an exponential regression would be a better fit for the data. The residuals are not above and below the line  $y = 0$  in a random pattern. Which means that even though the correlation coefficient of  $r \approx 0.8449$  says there is a strong positive linear relationship between year and temperature, it probably won't do as good of a job as we think at making predictions for the future because another type of model would be better.

## COEFFICIENT OF DETERMINATION AND RMSE

- 1. Linda read an article about the predictions of high school students and their GPA. The article studied three factors, the number of volunteer organizations each student participated in, the number of hours spent on homework, and the student's individual scores on standardized tests.

The article concluded that the number of hours spent on homework are the best predictor of GPA, because they found 24 % of the variance in GPA to be from hours spent on homework, 15 % from the number of volunteer organizations, and 11.5 % from individual scores on standardized tests.

What is the coefficient of determination for the line-of-best-fit that has  $y$ -values of high school GPA and  $x$ -values of hours spent on homework? Is the line of best fit a good predictor of the data? Why or why not?

*Solution:*

Remember the percent of the variation in  $y$  that can be explained by the  $x$ -values is the coefficient of determination or the  $r^2$  value.

In this context, the percent of the variance in GPA due to hours spent on homework is 24 %. So, we're talking about a least squares line where  $r^2 = 0.24$ . This is a very weak positive relationship, so the line of best fit is probably not a good predictor of the connection between hours spent on homework and GPA.



2. For the data in the table, calculate the sum of the squared residuals based on the mean of the  $y$ -values.

x	y
1	3.1
2	3.4
3	3.7
4	3.9
5	4.1

*Solution:*

First calculate  $\bar{y}$ .

$$\bar{y} = \frac{3.1 + 3.4 + 3.7 + 3.9 + 4.1}{5}$$

$$\bar{y} = \frac{18.2}{5}$$

$$\bar{y} = 3.64$$

The formula for a residual is

$$\text{residual} = \text{actual} - \text{predicted}$$

In this case, the predicted value is the mean of the  $y$ -values,  $\bar{y} = 3.64$ . Let's expand the table and calculate the residuals.

x	y	e
1	3.1	-0.54
2	3.4	-0.24
3	3.7	0.06
4	3.9	0.26
5	4.1	0.46

Now we just need to find the squares of these residuals and add them together.

	x	y	e	$e^2$
	1	3.1	-0.54	0.2916
	2	3.4	-0.24	0.0576
	3	3.7	0.06	0.0036
	4	3.9	0.26	0.0676
	5	4.1	0.46	0.2116
<b>Sum:</b>				<b>0.632</b>

So the sum of the squared residuals is about 0.632.

- 3. Use the same data as the previous question to calculate the sum of the squared residuals based on the least squares regression line,  
 $\hat{y} = 0.25x + 2.89$ .

*Solution:*

The formula for a residual is

$$\text{residual} = \text{actual} - \text{predicted}$$

In this case the predicted value is based on the regression line,  
 $\hat{y} = 0.25x + 2.89$ . Let's expand the table and calculate the residuals.

x	Actual y	Predicted y	e
1	3.1	3.14	-0.04
2	3.4	3.39	0.01
3	3.7	3.64	0.06
4	3.9	3.89	0.01
5	4.1	4.14	-0.04

Now we just need to find the squares of these residuals and add them together.

	x	Actual y	Predicted y	e	$e^2$
	1	3.1	3.14	-0.04	0.0016
	2	3.4	3.39	0.01	0.0001
	3	3.7	3.64	0.06	0.0036
	4	3.9	3.89	0.01	0.0001
	5	4.1	4.14	-0.04	0.0016
<b>Sum:</b>					<b>0.007</b>

So the sum of the squared residuals is 0.007.



- 4. Based on the previous two questions, in which we found the sum of the squared residuals based on the mean of the  $y$ -values and then the line of best fit, what percentage of error did we eliminate by using the least squares line? What is the term for this error?

*Solution:*

The sum of the squared residuals for the mean of the  $y$ -values was

$$\sum \text{residuals}^2 = 0.632$$

The sum of the squared residuals for the line of best fit was

$$\sum \text{residuals}^2 = 0.007$$

This means using the line of best fit reduces the error by

$$0.632 - 0.007$$

$$0.625$$

This is

$$\frac{0.625}{0.632} = 0.9889 = 98.89\%$$

of a reduction in error by using the least squares regression line. This is another way to calculate the coefficient of determination, so  $r^2 = 0.9889$ .



**5. What is the RMSE of the data set and what does it mean?**

x	y
1	3.1
2	3.4
3	3.7
4	3.9
5	4.1

*Solution:*

To find RMSE, we'll use the formula

$$\text{RMSE} = \sqrt{\frac{\sum \text{residuals}^2}{n}}$$

We already calculated the residual sum of squares.



	x	Actual y	Predicted y	e	e <sup>2</sup>
	1	3.1	3.14	-0.04	0.0016
	2	3.4	3.39	0.01	0.0001
	3	3.7	3.64	0.06	0.0036
	4	3.9	3.89	0.01	0.0001
	5	4.1	4.14	-0.04	0.0016
<b>Sum:</b>					<b>0.007</b>

The sum of the squared residuals was 0.007, so RMSE will be

$$\text{RMSE} = \sqrt{\frac{0.007}{5}} \approx 0.0374$$

RMSE is the standard deviation of the residuals, which means that

- 68 % of the data points will be within  $\pm 0.0418$  of the regression line,
- 95 % of the data points will be within  $\pm 2(0.0418)$  of the regression line, and
- 99.7 % of the data points will be within  $\pm 3(0.0418)$  of the regression line.

Since the RMSE we found is a small standard deviation, the data points are going to be more tightly clustered around the line-of-best-fit and the correlation in the data will be stronger.

6. Calculate the RMSE for the data set, given that the least squares line is  $\hat{y} = 0.0028x + 1.2208$ .

x	y
5	1.25
10	1.29
12	1.17
15	1.24
17	1.32

*Solution:*

To find RMSE, we'll use the formula

$$\text{RMSE} = \sqrt{\frac{\sum \text{residuals}^2}{n}}$$

We can calculate the residual sum of squares.

	x	Actual y	Predicted y	e	$e^2$
	5	1.25	1.2348	0.0152	0.00023104
	10	1.29	1.2488	0.0412	0.00169744
	12	1.17	1.2544	-0.0844	0.00712336
	15	1.24	1.2628	-0.0228	0.00051984
	17	1.32	1.2684	0.0516	0.00266256
<b>Sum:</b>					<b>0.01223424</b>

The sum of the squared residuals was 0.01223424, so RMSE will be

$$\text{RMSE} = \sqrt{\frac{0.01223424}{5}} \approx 0.0495$$

## CHI-SQUARE TESTS

- 1. We want to know whether a person's geographic region of the United States affects their preference of cell phone brand. We randomly sample people across the country and ask them about their brand preference. What can we conclude using a chi-square test at 95 % confidence?

	iPhone	Android	Other	Totals
Northeast	72	33	8	113
Southeast	48	26	7	81
Midwest	107	50	10	167
Northwest	59	33	10	102
Southwest	61	27	9	97
Totals	347	169	44	560

*Solution:*

Start by computing expected values.

	iPhone	Android	Other	Totals
Northeast	72 (70.02)	33 (34.10)	8 (8.88)	113
Southeast	48 (50.19)	26 (24.44)	7 (6.36)	81
Midwest	107 (103.48)	50 (50.40)	10 (13.12)	167
Northwest	59 (63.20)	33 (30.78)	10 (8.01)	102
Southwest	61 (60.11)	27 (29.27)	9 (7.62)	97
Totals	347	169	44	560

Now we'll check our sampling conditions. The problem told us that we took a random sample, and all of our expected values are at least 5, so we've met the random sampling and large counts conditions. And even though we're sampling without replacement, 560 is far less than 10% of the US population, so we've met the independence condition as well.

We'll state the null and alternative hypotheses.

$H_0$ : Cell phone brand preference isn't affected by geographic region.

$H_a$ : Cell phone brand preference is affected by geographic region.

Calculate  $\chi^2$ .

$$\begin{aligned} \chi^2 = & \frac{(72 - 70.02)^2}{70.02} + \frac{(33 - 34.10)^2}{34.10} + \frac{(8 - 8.88)^2}{8.88} \\ & + \frac{(48 - 50.19)^2}{50.19} + \frac{(26 - 24.44)^2}{24.44} + \frac{(7 - 6.36)^2}{6.36} \\ & + \frac{(107 - 103.48)^2}{103.48} + \frac{(50 - 50.40)^2}{50.40} + \frac{(10 - 13.12)^2}{13.12} \end{aligned}$$

$$+\frac{(59 - 63.20)^2}{63.20} + \frac{(33 - 30.78)^2}{30.78} + \frac{(10 - 8.01)^2}{8.01}$$

$$+\frac{(61 - 60.11)^2}{60.11} + \frac{(27 - 29.27)^2}{29.27} + \frac{(9 - 7.62)^2}{7.62}$$

$$\chi^2 \approx 0.0560 + 0.0355 + 0.0872$$

$$+0.0956 + 0.0996 + 0.0644$$

$$+0.1197 + 0.0032 + 0.7420$$

$$+0.2791 + 0.1601 + 0.4944$$

$$+0.0132 + 0.1760 + 0.2499$$

$$\chi^2 \approx 2.6759$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (5 - 1)(3 - 1)$$

$$df = (4)(2)$$

$$df = 8$$

With  $df = 8$  and  $\chi^2 \approx 2.6759$ , the  $\chi^2$ -table gives



df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67

We're off the chart on the left, which means we will definitely not exceed the alpha level  $\alpha = 0.05$ . Therefore, we'll fail to reject the null hypothesis, and conclude that geographic region of the country does not affect cell phone brand preference.

- 2. A beverage company wants to know if gender affects which of their products people prefer. They take a random sample of fewer than 10% of their customers, and ask them in a blind taste test which beverage they prefer. What can the company conclude using a chi-square test at  $\alpha = 0.1$ ?

		Beverage			Totals	
		A	B	C		
Men	35	34	31	100		
	31	33	36	100		
Totals		66	67	67	200	

*Solution:*

Start by computing expected values.

	Beverage			
	A	B	C	Totals
Men	35 (33.0)	34 (33.5)	31 (33.5)	100
Women	31 (33.0)	33 (33.5)	36 (33.5)	100
Totals	66	67	67	200

Now we'll check our sampling conditions. The problem told us that we took a random sample and that we sampled less than 10% of the population, so we've met the random sampling and independence conditions. And all of our expected values are at least 5, so we've met the large counts condition as well.

We'll state the null and alternative hypotheses.

$H_0$ : Gender does not affect beverage preference.

$H_a$ : Gender affects beverage preference.

Calculate  $\chi^2$ .

$$\begin{aligned}\chi^2 = & \frac{(35 - 33)^2}{33} + \frac{(34 - 33.5)^2}{33.5} + \frac{(31 - 33.5)^2}{33.5} \\ & + \frac{(31 - 33)^2}{33} + \frac{(33 - 33.5)^2}{33.5} + \frac{(36 - 33.5)^2}{33.5}\end{aligned}$$

$$\chi^2 = \frac{4}{33} + \frac{0.25}{33.5} + \frac{6.25}{33.5} + \frac{4}{33} + \frac{0.25}{33.5} + \frac{6.25}{33.5}$$

$$\chi^2 \approx 0.1212 + 0.0075 + 0.1866 + 0.1212 + 0.0075 + 0.1866$$

$$\chi^2 = 0.6306$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (2 - 1)(3 - 1)$$

$$df = (1)(2)$$

$$df = 2$$

With  $df = 2$  and  $\chi^2 = 0.6306$ , the  $\chi^2$ -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level  $\alpha = 0.1$ . Therefore, we'll fail to reject the null hypothesis, and conclude that gender does not affect beverage preference.

- 3. A coffee company wants to know whether or not drink and pastry choice are related among their customers. The company randomly sampled fewer than 10 % of their customers, and recorded their drink and pastry orders. What can the restaurant conclude using a chi-square test at 99 % confidence?

	Bagel	Muffin	Totals
Coffee	38	34	72
Tea	25	29	54
Totals	63	63	126

*Solution:*

Start by computing expected values.

	Bagel	Muffin	Totals
Coffee	38 (36)	34 (36)	72
Tea	25 (27)	29 (27)	54
Totals	63	63	126

Now we'll check our sampling conditions. The problem told us that we took a random sample and that we sampled less than 10 % of the population, so we've met the random sampling and independence conditions. And all of our expected values are at least 5, so we've met the large counts condition as well.

We'll state the null and alternative hypotheses.

$H_0$ : Pastry preference isn't affected by beverage preference.

$H_a$ : Pastry preference is affected by beverage preference.

Calculate  $\chi^2$ .

$$\chi^2 = \frac{(38 - 36)^2}{36} + \frac{(34 - 36)^2}{36} + \frac{(25 - 27)^2}{27} + \frac{(29 - 27)^2}{27}$$

$$\chi^2 = \frac{4}{36} + \frac{4}{36} + \frac{4}{27} + \frac{4}{27}$$

$$\chi^2 = \frac{2}{9} + \frac{8}{27}$$

$$\chi^2 \approx 0.52$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (2 - 1)(2 - 1)$$

$$df = (1)(1)$$

$$df = 1$$

With  $df = 1$  and  $\chi^2 = 0.52$ , the  $\chi^2$ -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level  $\alpha = 0.01$ . Therefore, the coffee company will fail to reject the null hypothesis, and conclude that beverage preference does not affect pastry preference.



4. A school district wants to know whether or not GPA is affected by elective preference. They randomly sampled fewer than 10% of their students, and recorded their elective preference and GPA. What can the school district conclude using a chi-square test at  $\alpha = 0.1$ ?

	GPA range				
	<2	2	3	4+	Totals
Music	12	26	31	34	103
Theater	21	22	23	21	87
Art	36	29	29	32	126
Totals	69	77	83	87	316

*Solution:*

Start by computing expected values.

	GPA range				
	<2	2	3	4+	Totals
Music	12 (22.49)	26 (25.10)	31 (27.05)	34 (28.36)	103
Theater	21 (19.00)	22 (21.20)	23 (22.85)	21 (23.95)	87
Art	36 (27.51)	29 (30.70)	29 (33.09)	32 (34.69)	126
Totals	69	77	83	87	316

Now we'll check our sampling conditions. The problem told us that we took a random sample and that we sampled less than 10% of the population, so we've met the random sampling and independence conditions. And all of our expected values are at least 5, so we've met the large counts condition as well.

We'll state the null and alternative hypotheses.

$H_0$ : Elective choice doesn't affect GPA.

$H_a$ : Elective choice affects GPA.

Calculate  $\chi^2$ .

$$\begin{aligned}\chi^2 = & \frac{(12 - 22.49)^2}{22.49} + \frac{(26 - 25.10)^2}{25.10} + \frac{(31 - 27.05)^2}{27.05} + \frac{(34 - 28.36)^2}{28.36} \\ & + \frac{(21 - 19.00)^2}{19.00} + \frac{(22 - 21.20)^2}{21.20} + \frac{(23 - 22.85)^2}{22.85} + \frac{(21 - 23.95)^2}{23.95} \\ & + \frac{(36 - 27.51)^2}{27.51} + \frac{(29 - 30.70)^2}{30.70} + \frac{(29 - 33.09)^2}{33.09} + \frac{(32 - 34.69)^2}{34.69}\end{aligned}$$

$$\chi^2 \approx 4.89 + 0.03 + 0.58 + 1.12 + 0.21 + 0.03$$

$$+ 0.00 + 0.36 + 2.62 + 0.09 + 0.51 + 0.21$$

$$\chi^2 = 10.65$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (3 - 1)(4 - 1)$$



$$df = (2)(3)$$

$$df = 6$$

With  $df = 6$  and  $\chi^2 = 10.65$ , the  $\chi^2$ -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02

The  $\chi^2$  value just clears  $\alpha = 0.1$ , which means that the school district can reject the null hypothesis and conclude that elective choice affects GPA. If they had set a higher confidence level of 95 % (with  $\alpha = 0.05$ ), they would not have been able to reject the null.

- 5. An airline wants to know if people travel constantly throughout the year, or if travel is more concentrated at specific times. They recorded flights taken each quarter, and recorded them in a table (in hundreds of thousands). What can the airline conclude using a chi-square test at 95 % confidence?

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Flights	3.97	4.58	4.73	5.14	18.42

*Solution:*



With 18.42 (or 18,420,000) total flights, the expected number of flights in each quarter would be  $18.42/4 = 4.605$ .

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Flights	3.97	4.58	4.73	5.14	18.42
Expected	4.605	4.605	4.605	4.605	18.42

We'll state the null and alternative hypotheses.

$H_0$ : Number of flights taken is not affected by quarter.

$H_a$ : Number of flights taken is affected by quarter.

Calculate  $\chi^2$ .

$$\chi^2 = \frac{(3.97 - 4.605)^2}{4.605} + \frac{(4.58 - 4.605)^2}{4.605} + \frac{(4.73 - 4.605)^2}{4.605} + \frac{(5.14 - 4.605)^2}{4.605}$$

$$\chi^2 \approx 0.0876 + 0.0001 + 0.0034 + 0.0622$$

$$\chi^2 = 0.1533$$

The degrees of freedom are  $n - 1 = 4 - 1 = 3$ . With  $df = 3$  and  $\chi^2 = 0.1533$ , the  $\chi^2$ -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00

We're off the chart on the left, which means we will definitely not exceed the alpha level  $\alpha = 0.05$ . Therefore, the airline will fail to reject the null hypothesis, and conclude that number of flights taken is not affected by quarter.

- 6. A sandwich company wants to know how their sales are affected by time of day. They recorded sandwiches sold during each part of the day. What can the sandwich company conclude using a chi-square test at  $\alpha = 0.1$ ?

Time of day	Midday	Afternoon	Evening	Total
Sales	213	208	221	642
Expected	214	214	214	642

*Solution:*

With 642 total sandwiches sold, the expected number of sandwiches in each period would be  $642/3 = 214$ .

Time of day	Midday	Afternoon	Evening	Total
Sales	213	208	221	642
Expected	214	214	214	642

We'll state the null and alternative hypotheses.

$H_0$ : Number of sandwiches sold is not affected by time of day.

$H_a$ : Number of sandwiches sold is affected by time of day.

Calculate  $\chi^2$ .

$$\chi^2 = \frac{(213 - 214)^2}{214} + \frac{(208 - 214)^2}{214} + \frac{(221 - 214)^2}{214}$$

$$\chi^2 = \frac{1}{214} + \frac{36}{214} + \frac{49}{214}$$

$$\chi^2 = \frac{86}{214}$$

$$\chi^2 \approx 0.4019$$

The degrees of freedom are  $n - 1 = 3 - 1 = 2$ . With  $df = 2$  and  $\chi^2 = 0.4019$ , the  $\chi^2$ -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level  $\alpha = 0.1$ . Therefore, the sandwich company will fail to reject the null hypothesis, and conclude that number of sandwiches sold is not affected by time of day.



