Probability & Statistics Workbook Solutions

krista king
MATH

ONE-WAY DATA

■ 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 you 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:

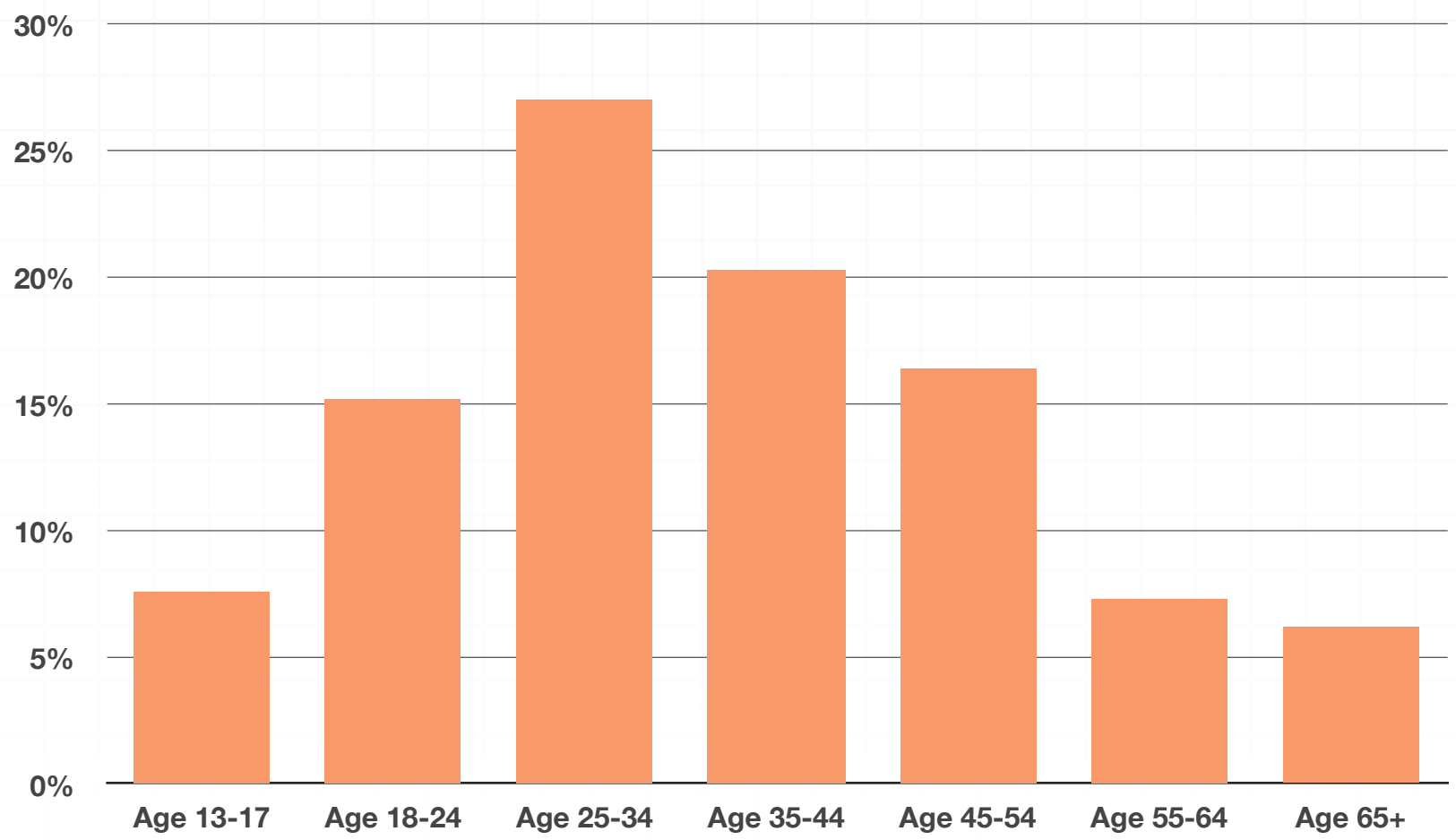
> "In general do you think things have gotten better or worse for our students over the course of the year?"

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?

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.

■ 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

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. Is the data below one-way data? Why or why not?

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

Solution:

This data is not an example of one-way data. In order to know the number of rejected products, you'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?," you'd need to ask something like "How many rejected products were there for worker 1123 during the first shift?" Since you need more than one reference point, this is not one-way data.

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

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 in July?" and I would know right away that the answer was 300.

If the data isn't one-way data, I'd need to answer a question about both categories. For example, to answer a question like "How many scoops were sold?" you 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 one-way data? Why or why not?

Solution:

The experiment does not result in one-way data.

We can think of how the botany student would need to record his data to see whether or not this is an example of one-way data. 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.

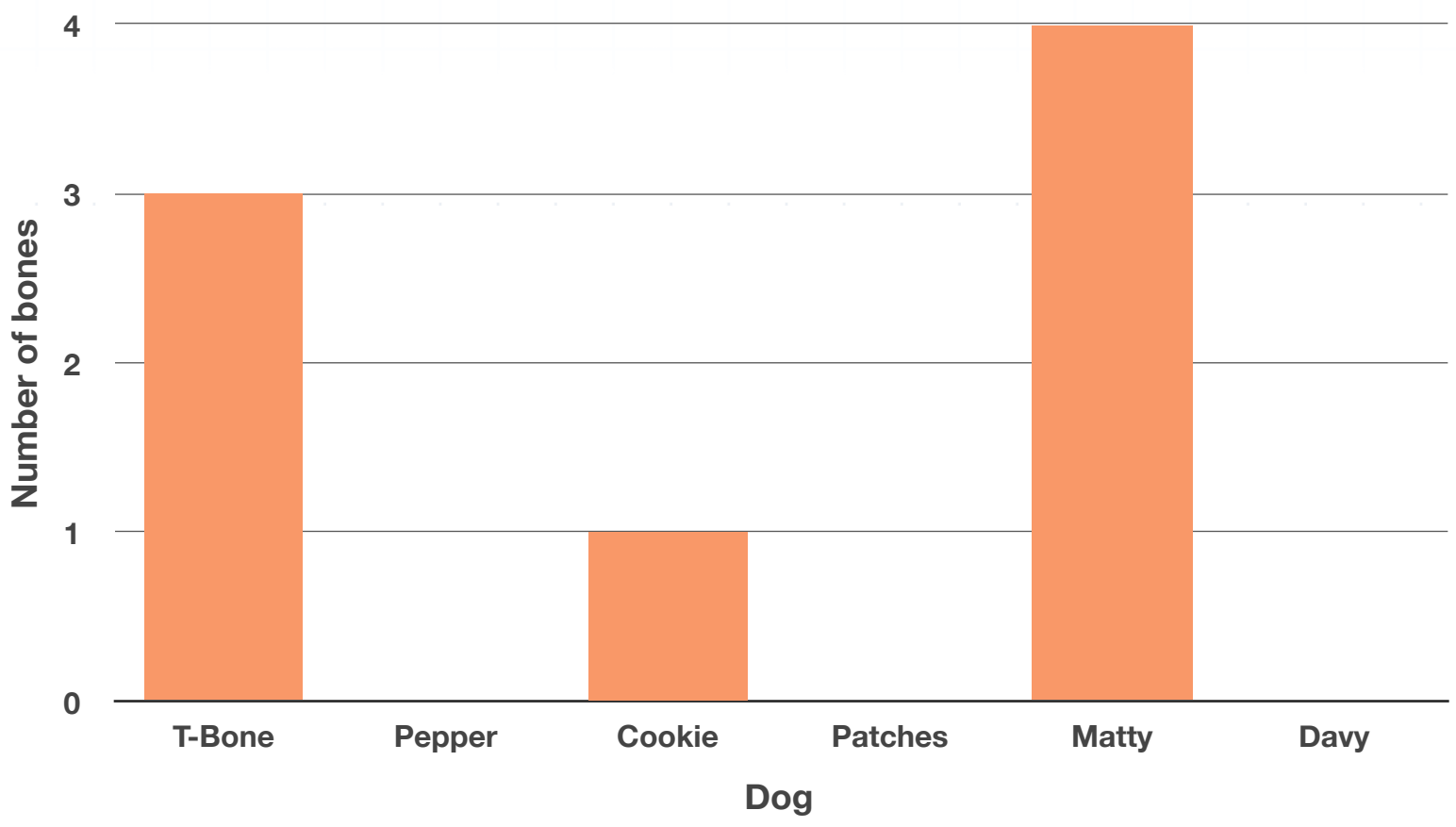
	Check 1	Check 2	Check 3	Check 4	...	Check 15
Plot 1						
Plot 2						
Plot 3						
Plot 4						
...						
Plot 12						

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

BAR GRAPHS AND PIE CHARTS

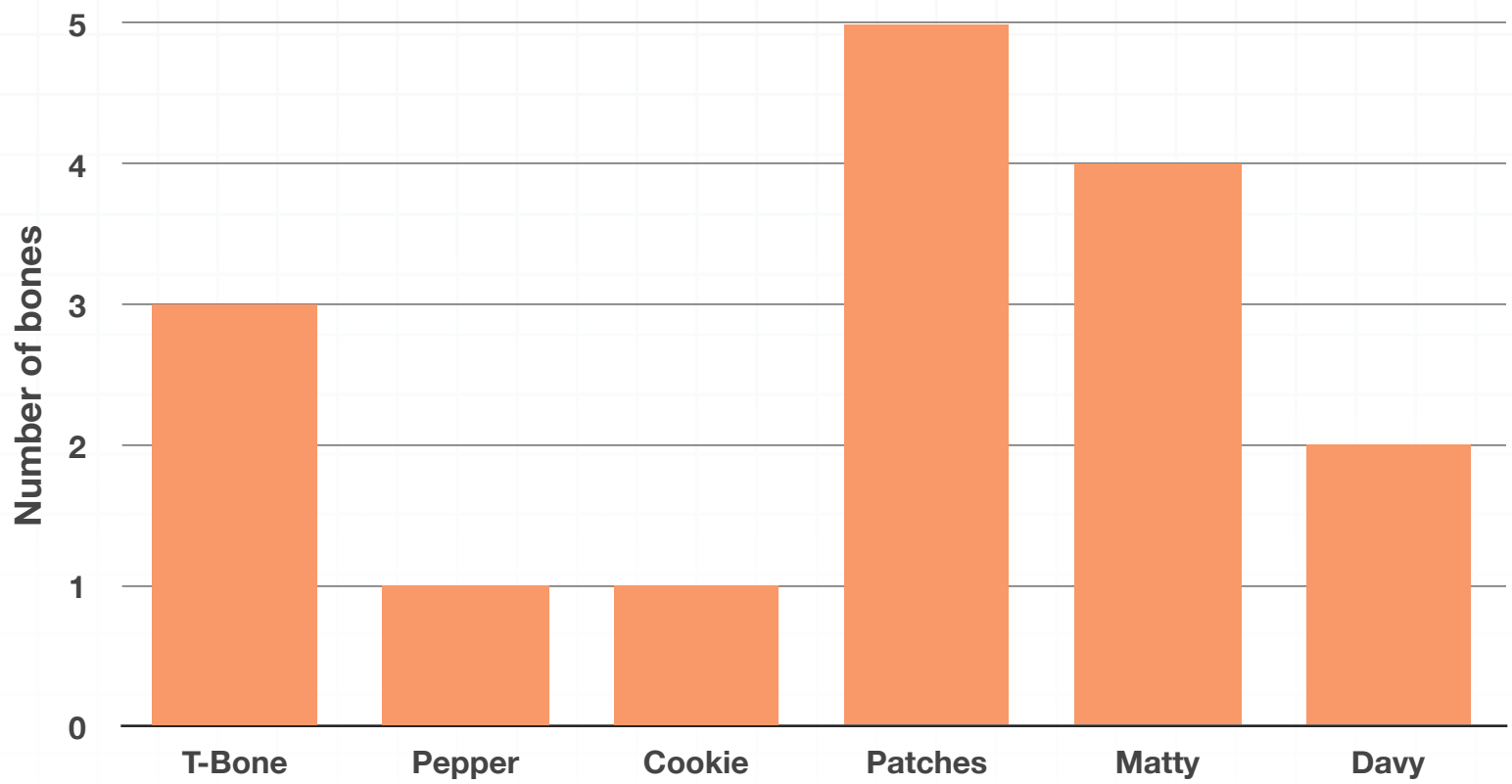
1. Both the bar graph and the table have missing information about the number of bones each dog consumed at doggie daycare. Use the graph and table together to fill in the missing pieces.

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



Solution:

You 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

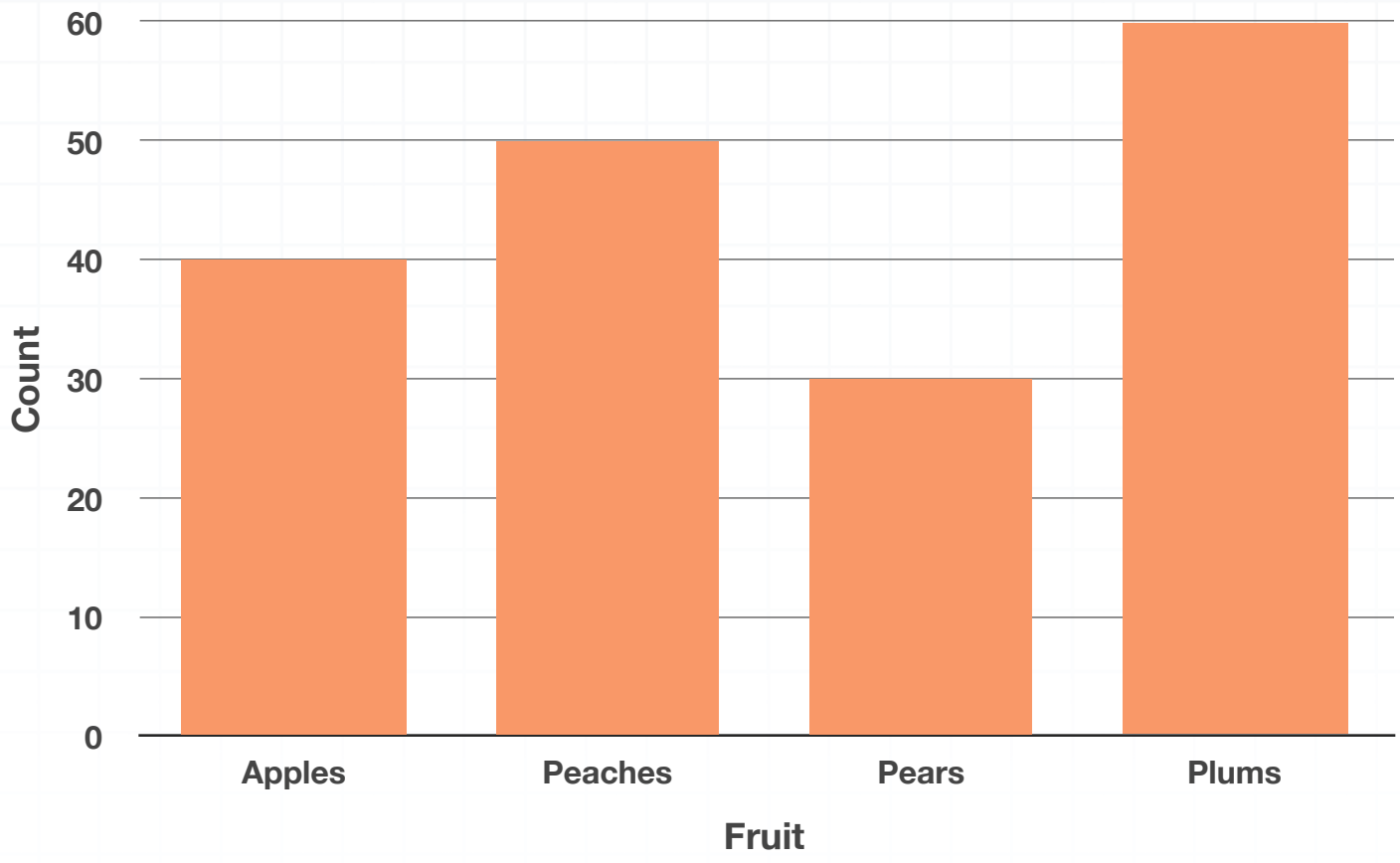


Dog

You 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 you 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: $40 \div 10 = 4$

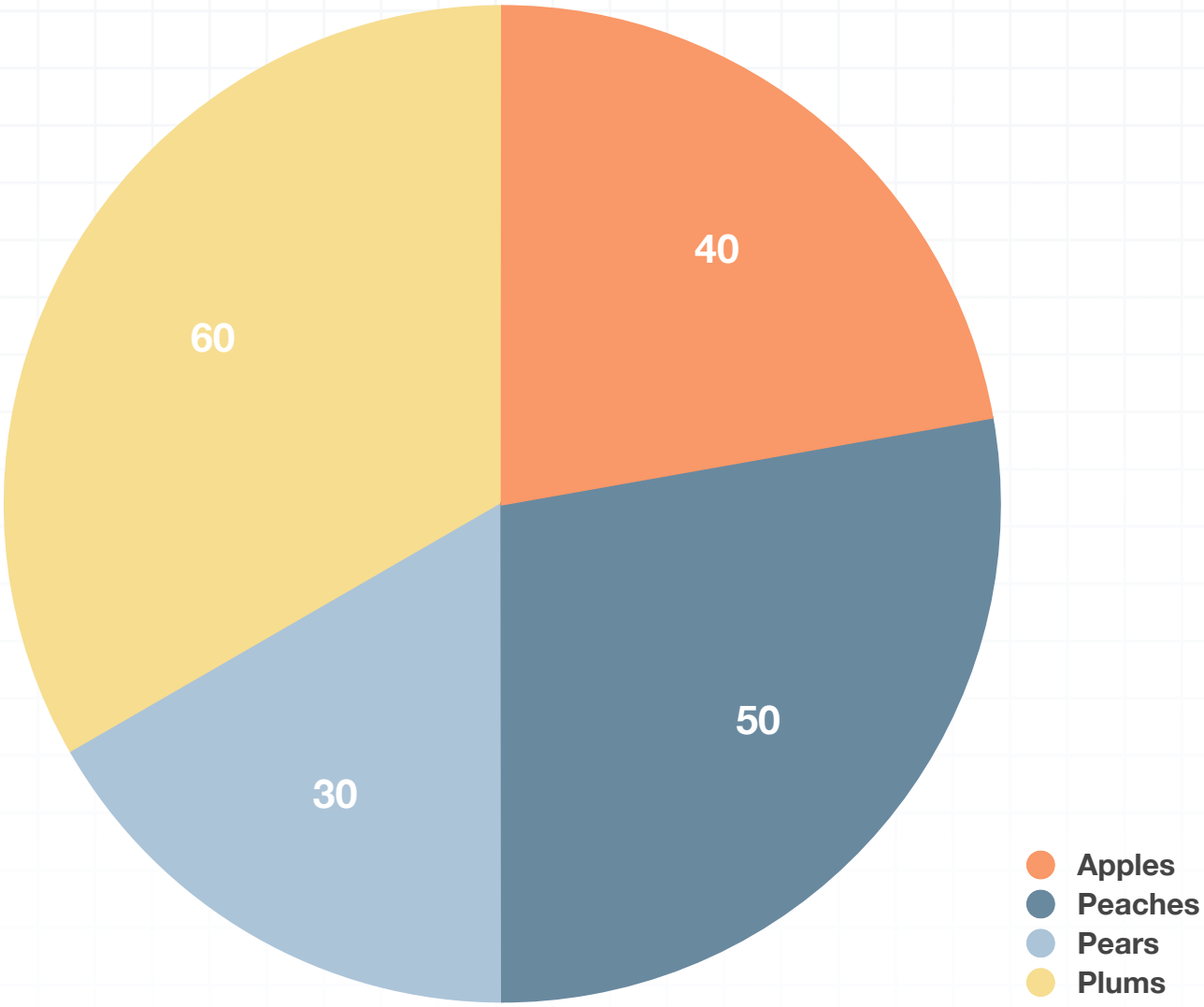
For peaches: $50 \div 10 = 5$

For pears: $30 \div 10 = 3$

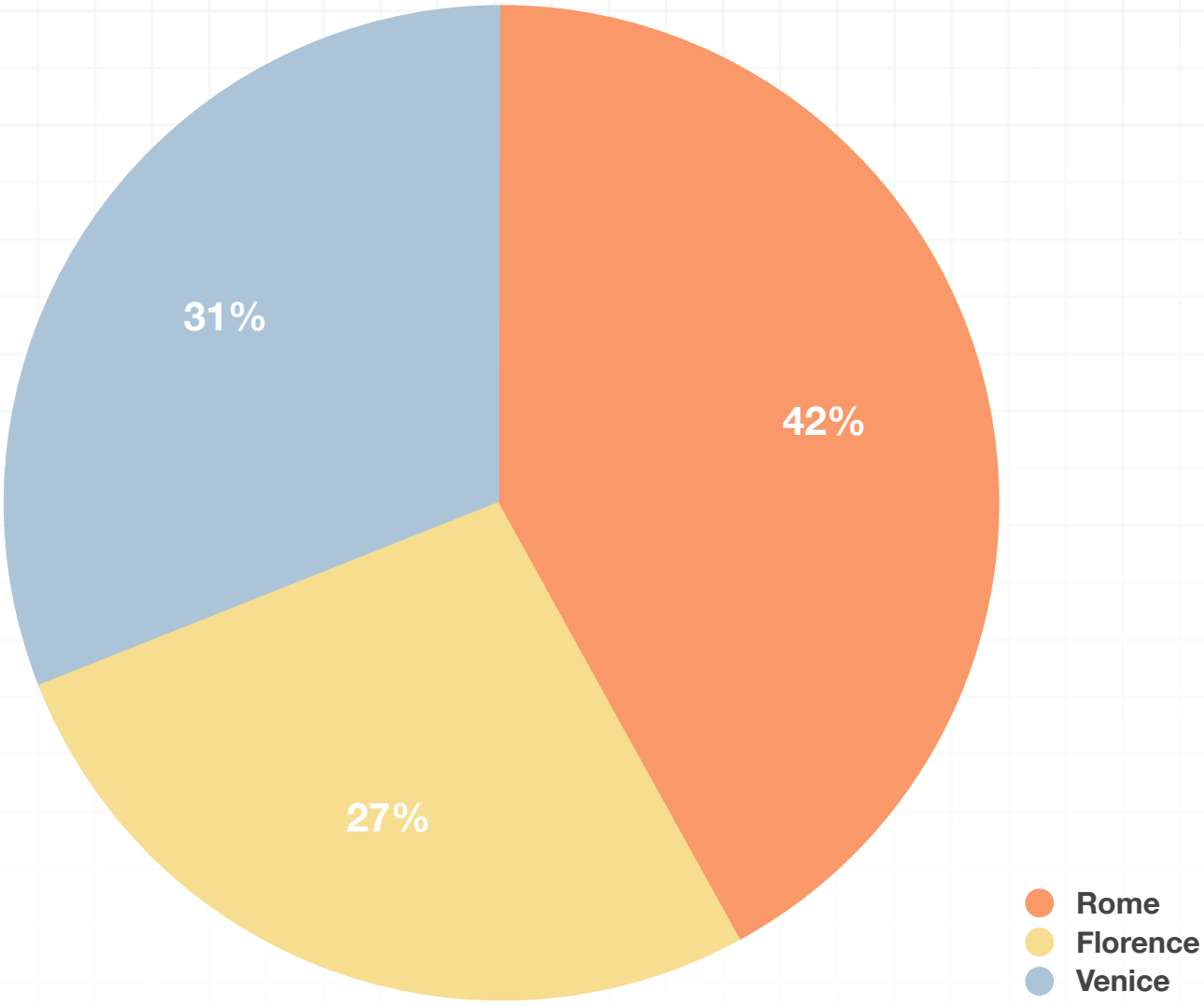
For plums: $60 \div 10 = 6$

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?

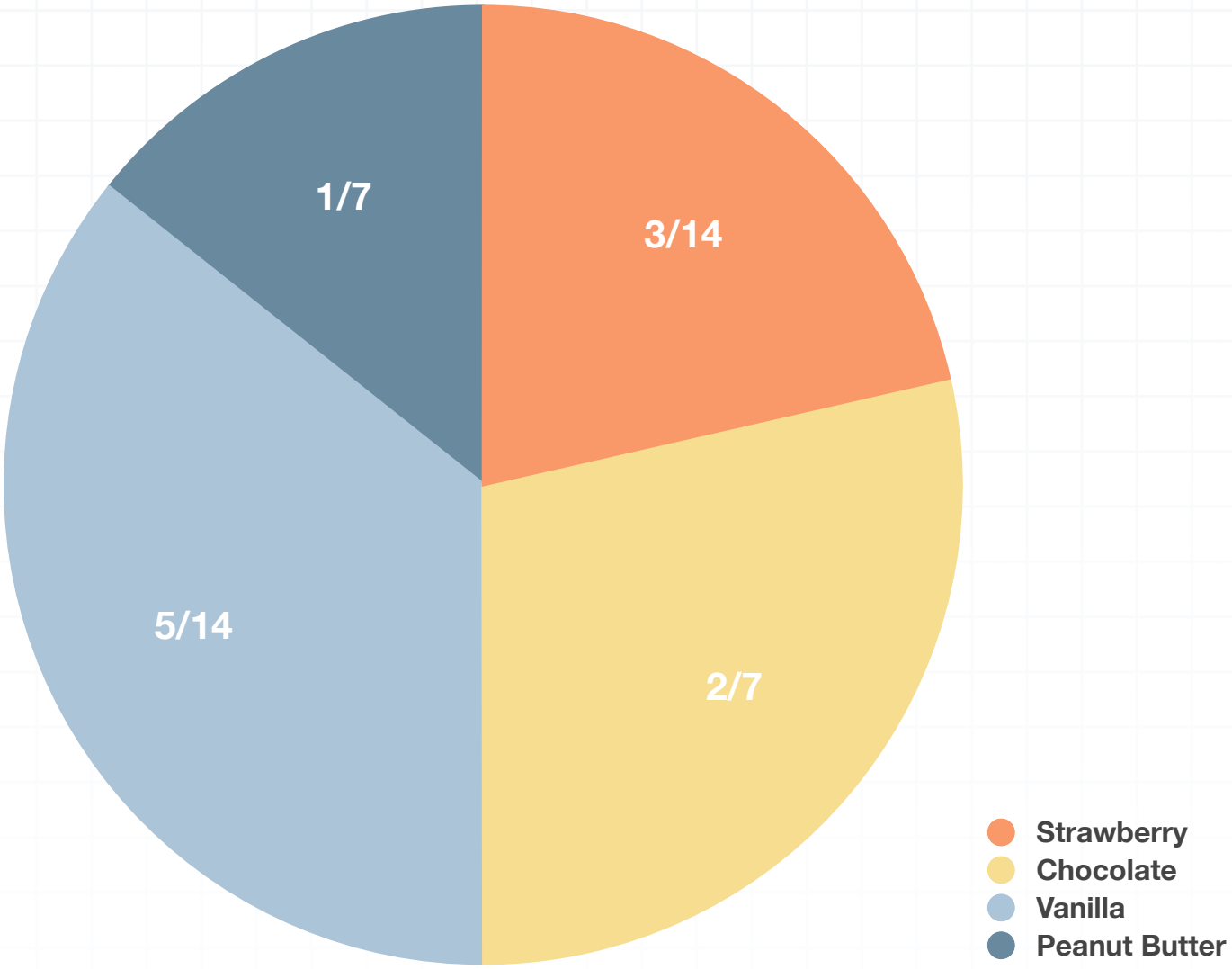


Solution:

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.



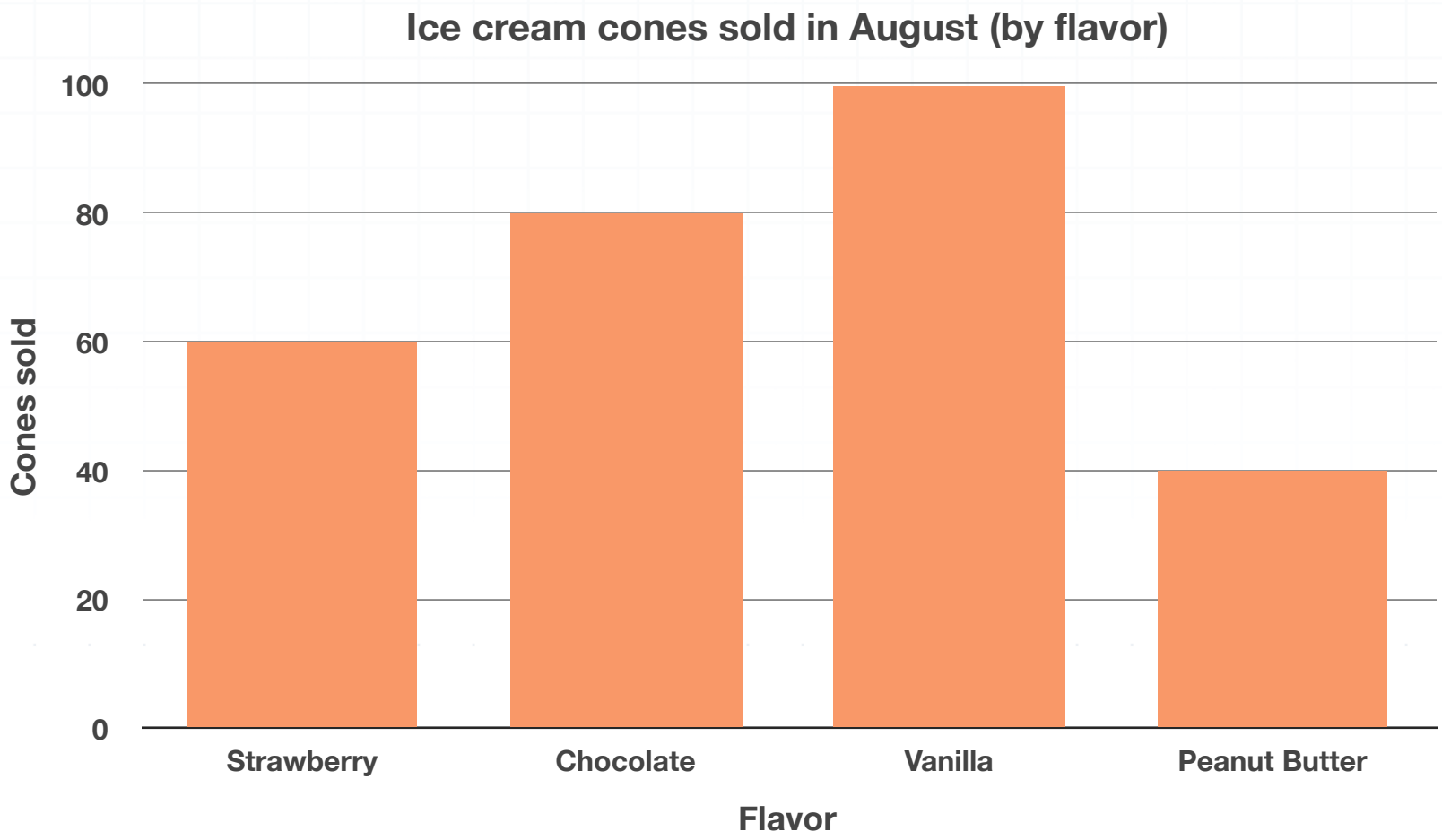
Solution:

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 $3/14$ 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.

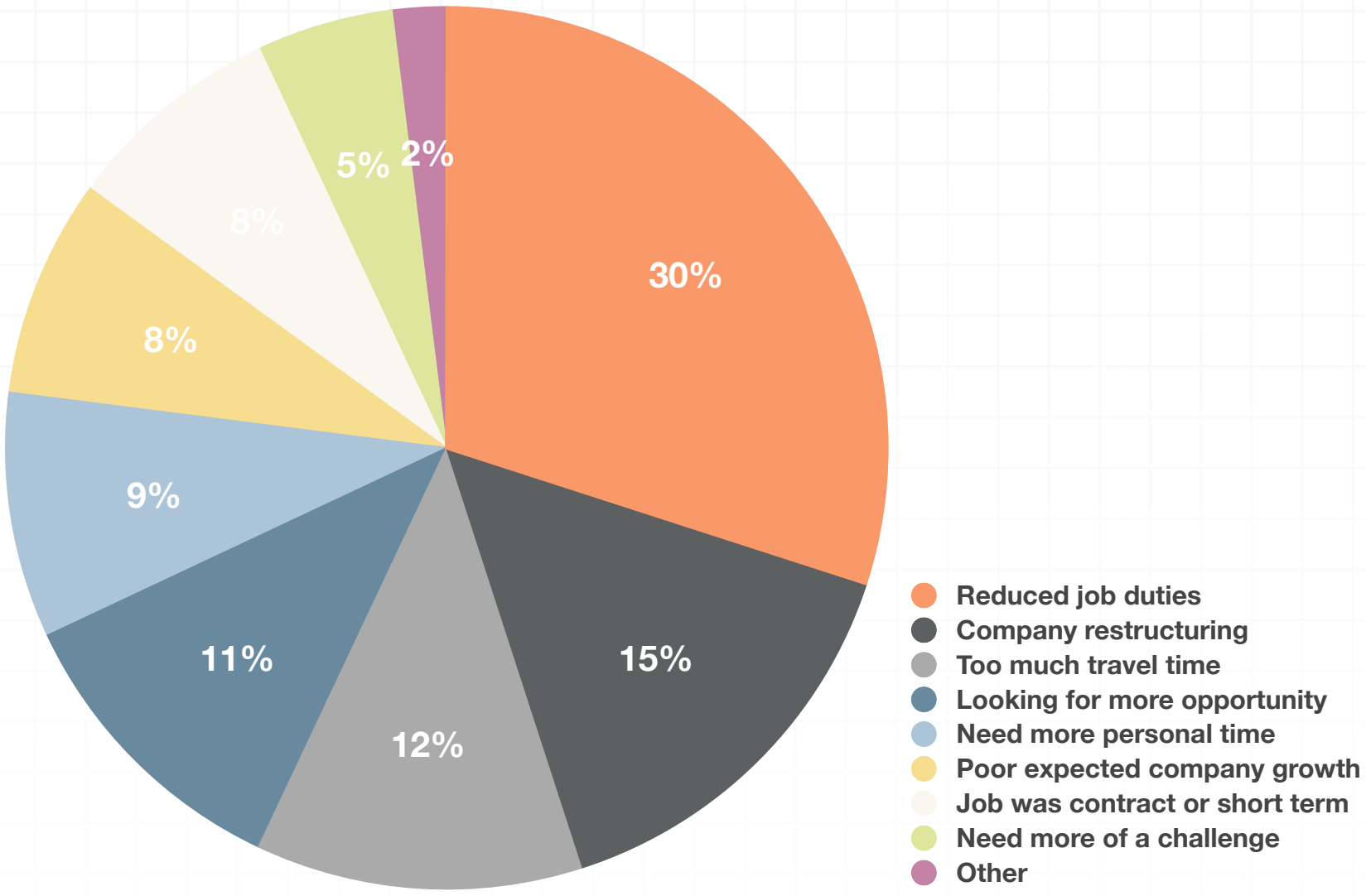


■ 5. A company is analyzing the results from a recent survey about why people left their employment. The results are shown in the data table below. In general, is a bar graph or a pie chart a better choice to display the data? Why?

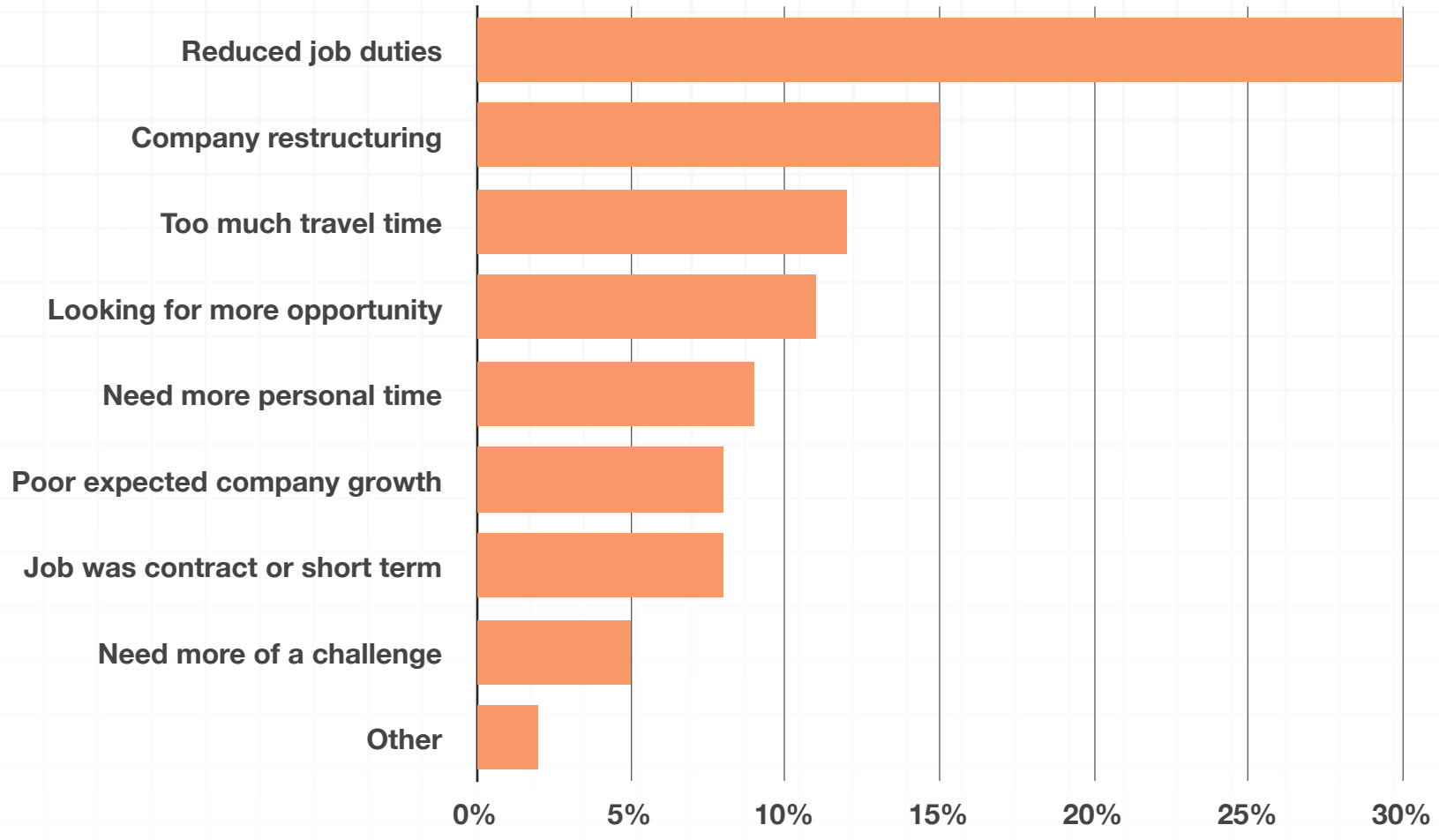
Reasons for leaving job	
Reduced job duties	30%
Company restructuring	15%
Too much travel time	12%
Looking for more opportunity	11%
Need more personal time	9%
Poor expected company growth	8%
Job was contract or short term	8%
Need more of a challenge	5%
Other	2%

Solution:

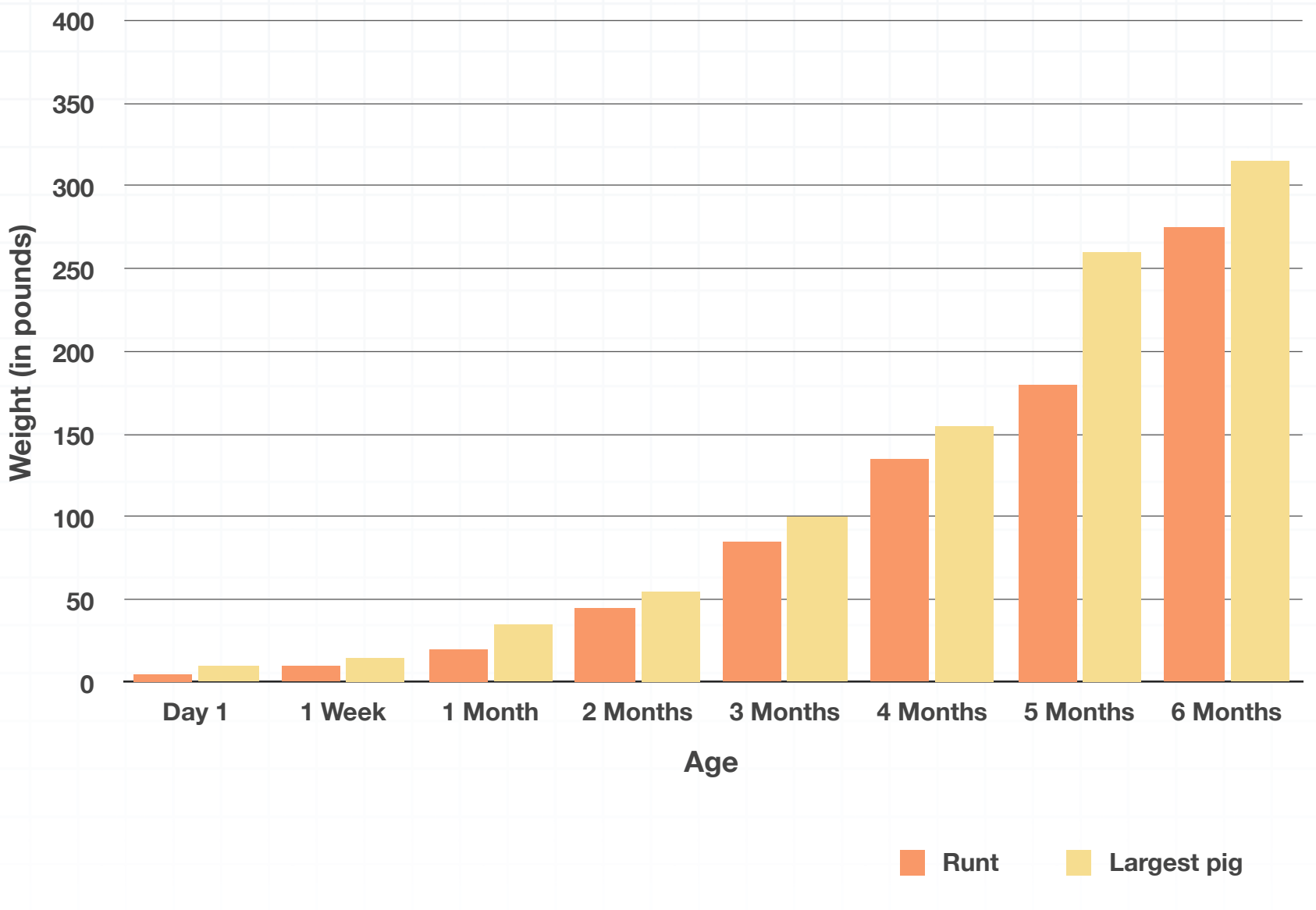
A bar graph is a better choice to display the data because there are so many different categories. A pie chart can get cluttered when there are a lot of categories,



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.

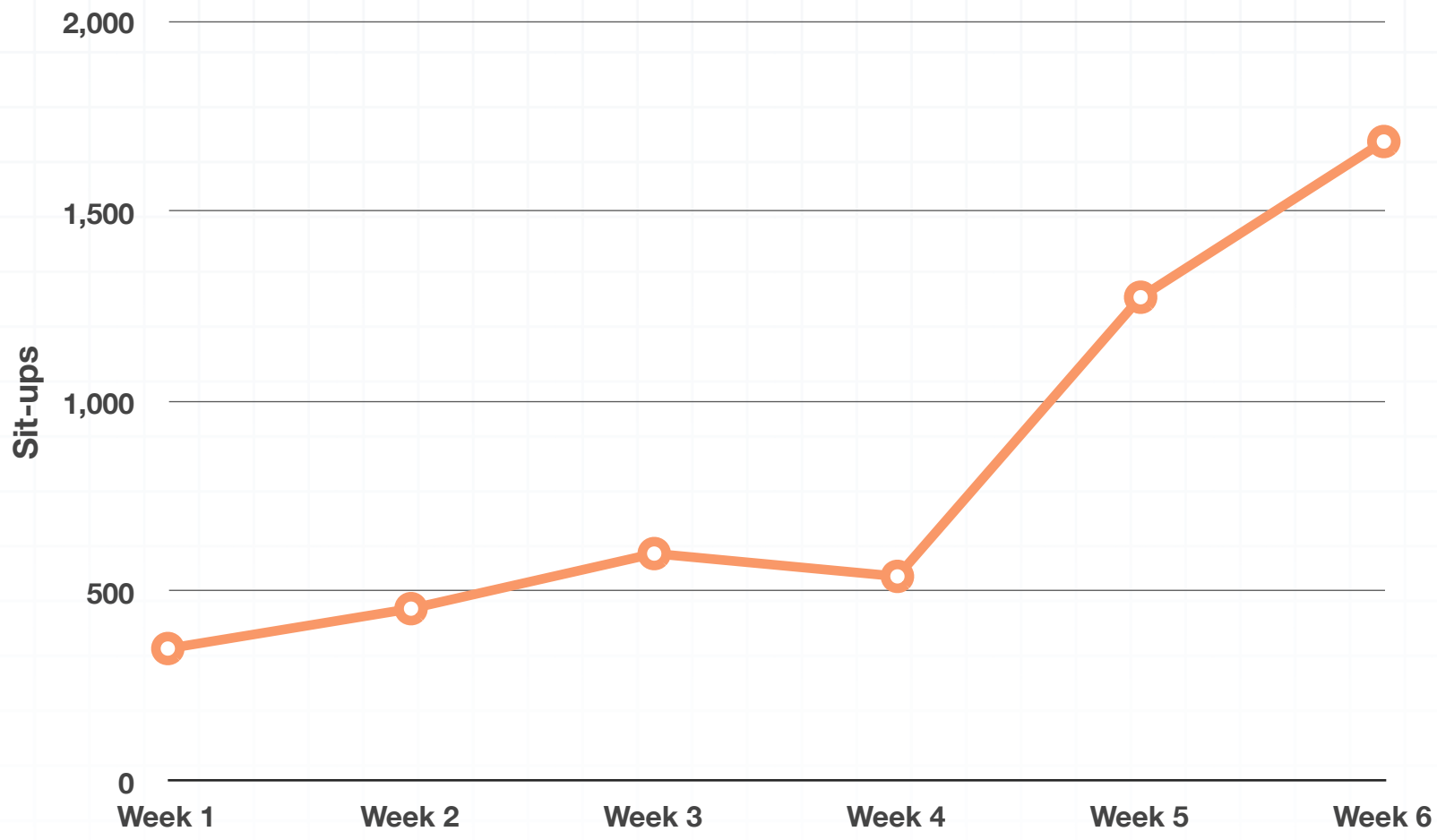
Week	Number of sit-ups
Week 1	350
Week 2	455
Week 3	600
Week 4	540
Week 5	1,275
Week 6	1,685

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.

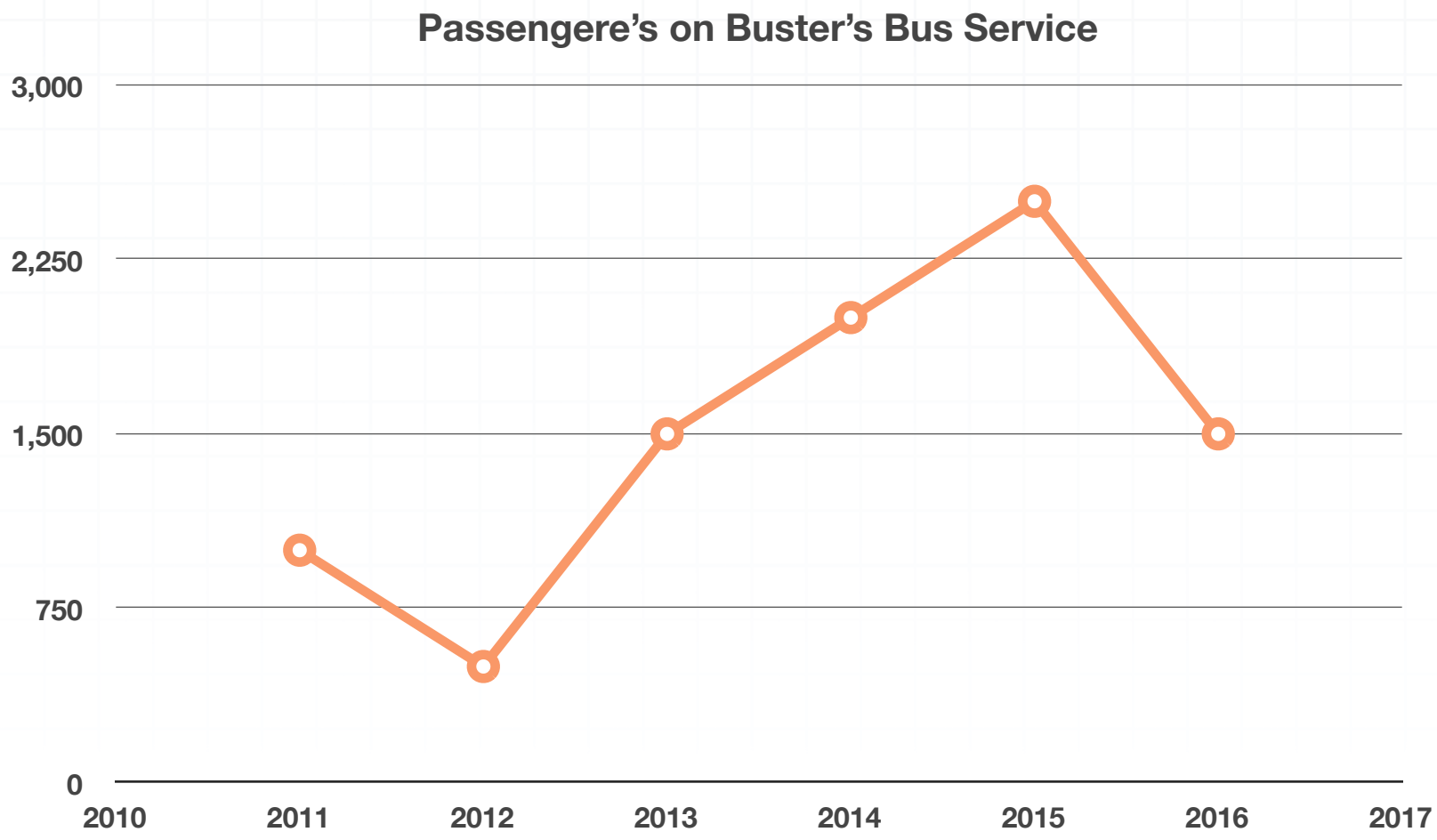


2. The table shows passengers by year for Buster's Bus Service. Create a line graph of the data in the table.

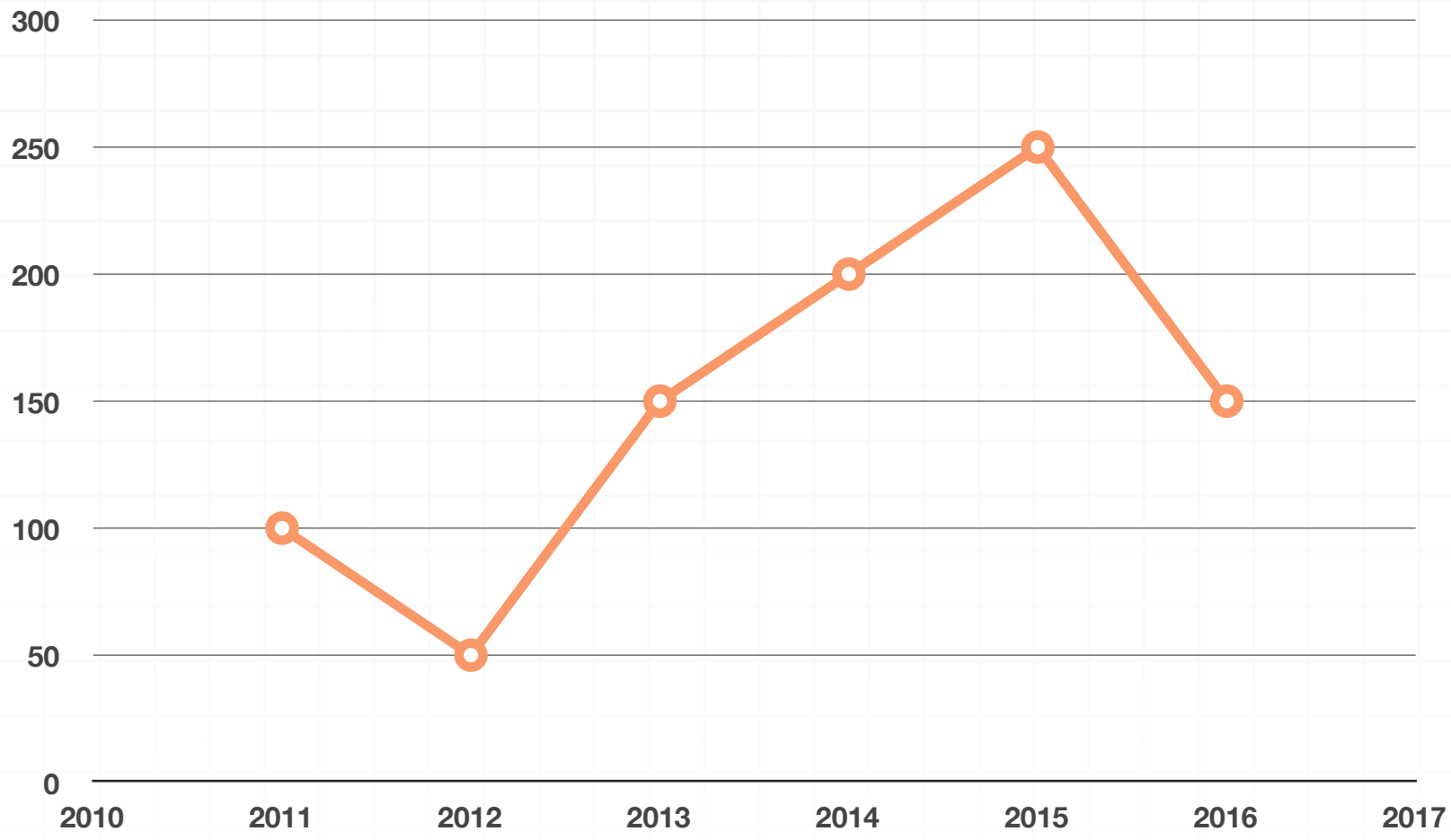
Year	Passengers
2011	1,000
2012	500
2013	1,500
2014	2,000
2015	2,500
2016	1,500

Solution:

Start the line graph at the year 2010 and go by year along the horizontal axis to 2017. Make sure you choose units on the vertical axis that make it easy to graph, as well as read.



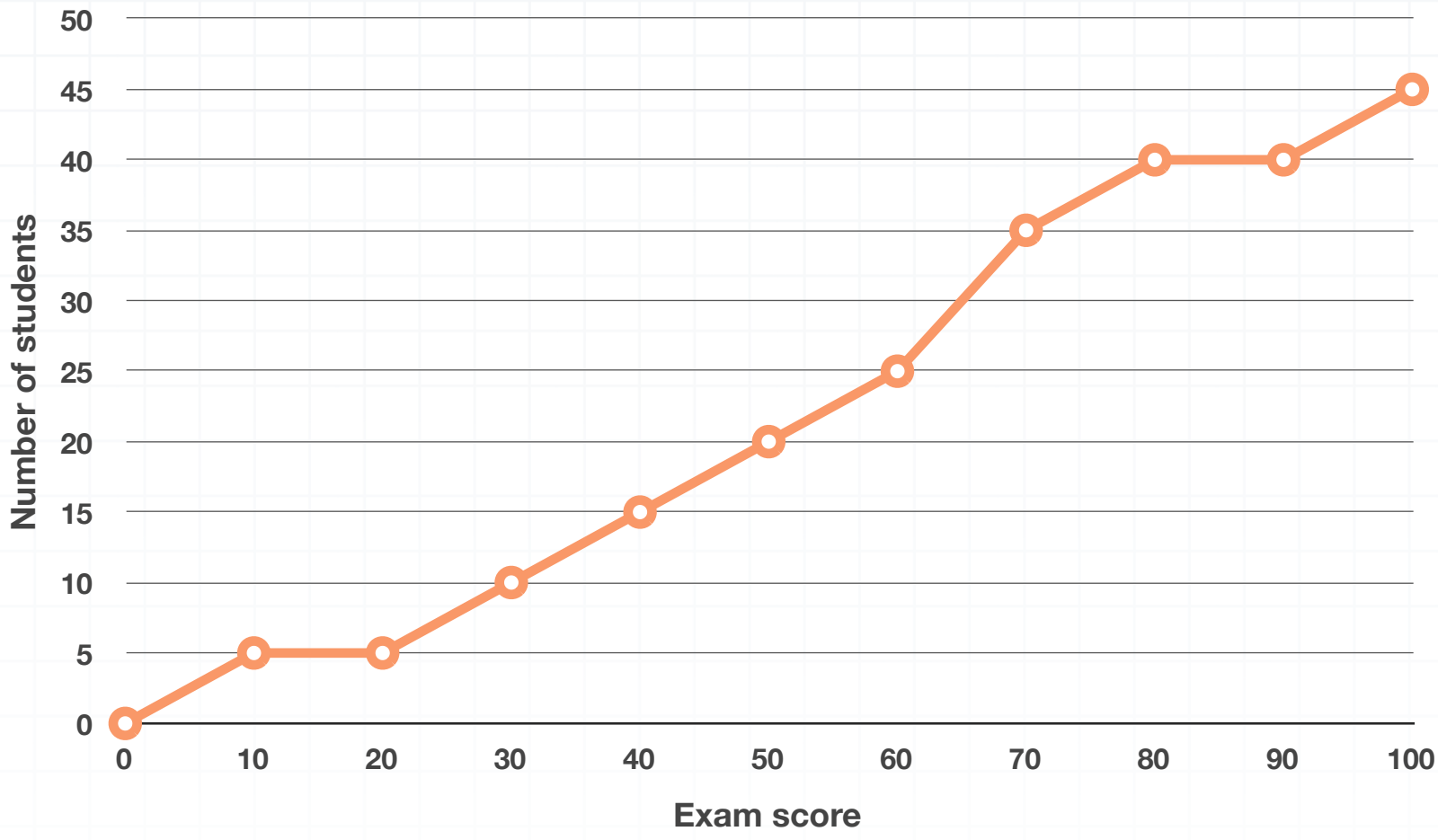
3. Between what two consecutive years was there the largest increase in car sales?



Solution:

The greatest increase in card sales was between 2012 and 2013. If you look at the line graph, you can see that the line increases at the sharpest rate between 2012 and 2013, 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 $90\,\%$?



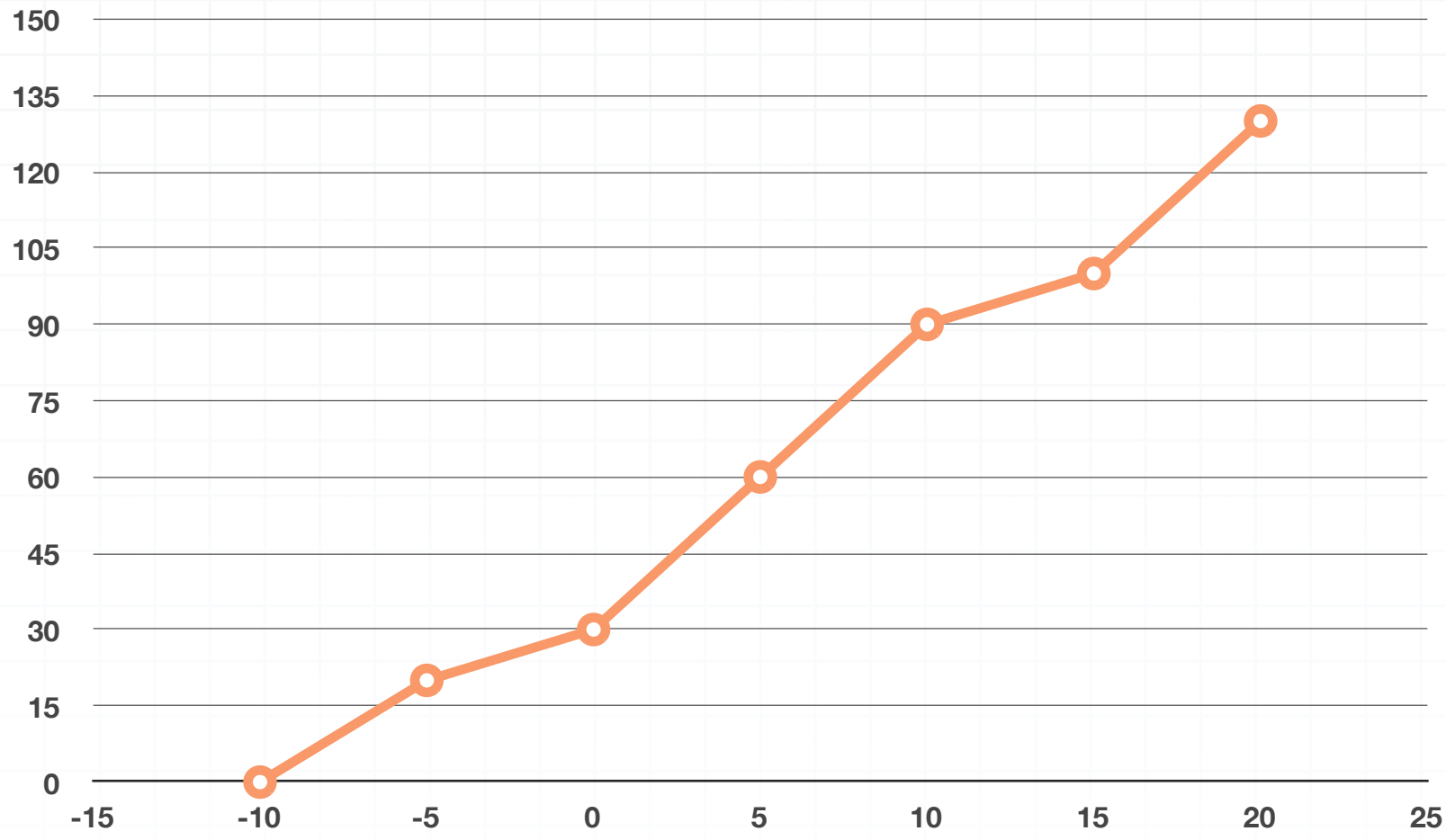
Solution:

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:

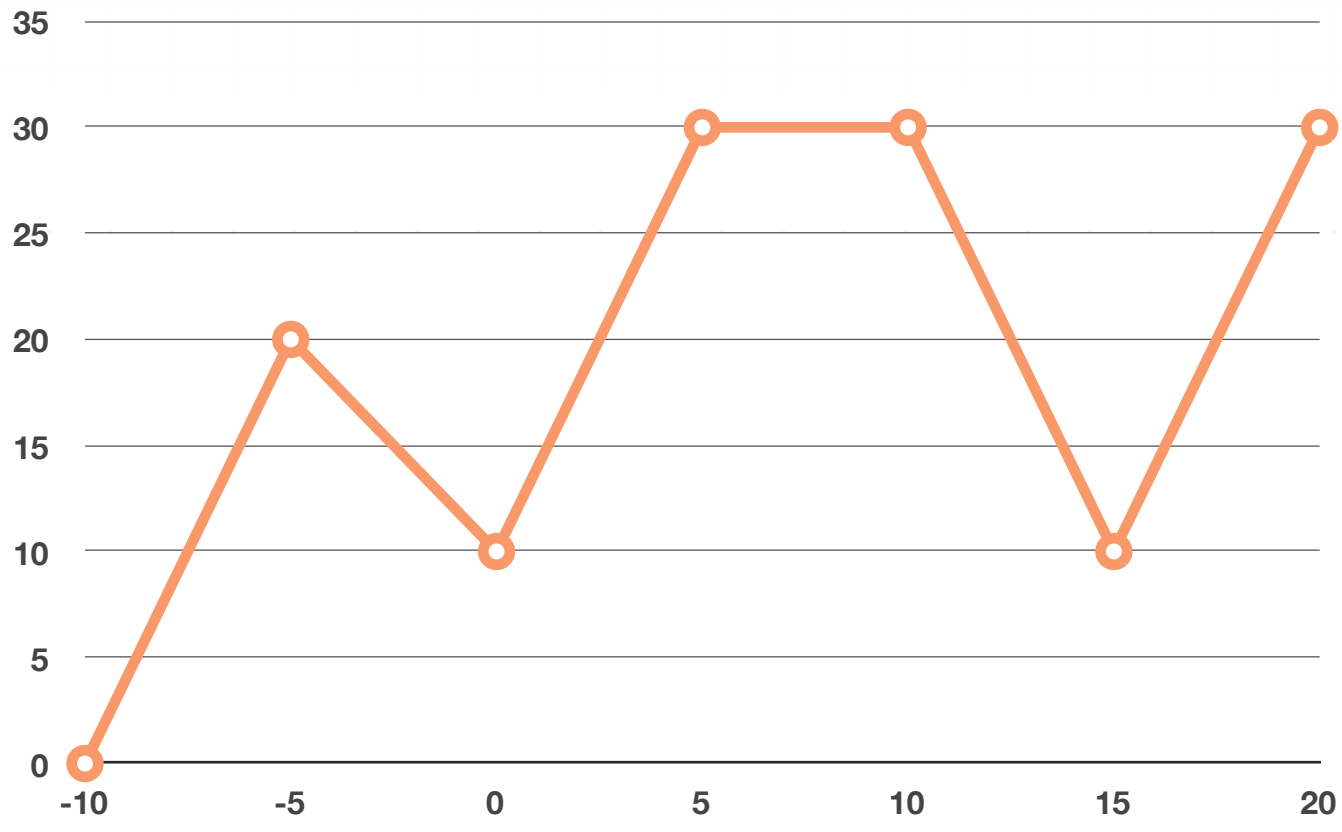
First we can create a table of the information from the ogive. The statistical name for the "total" is the "cumulative frequency."

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

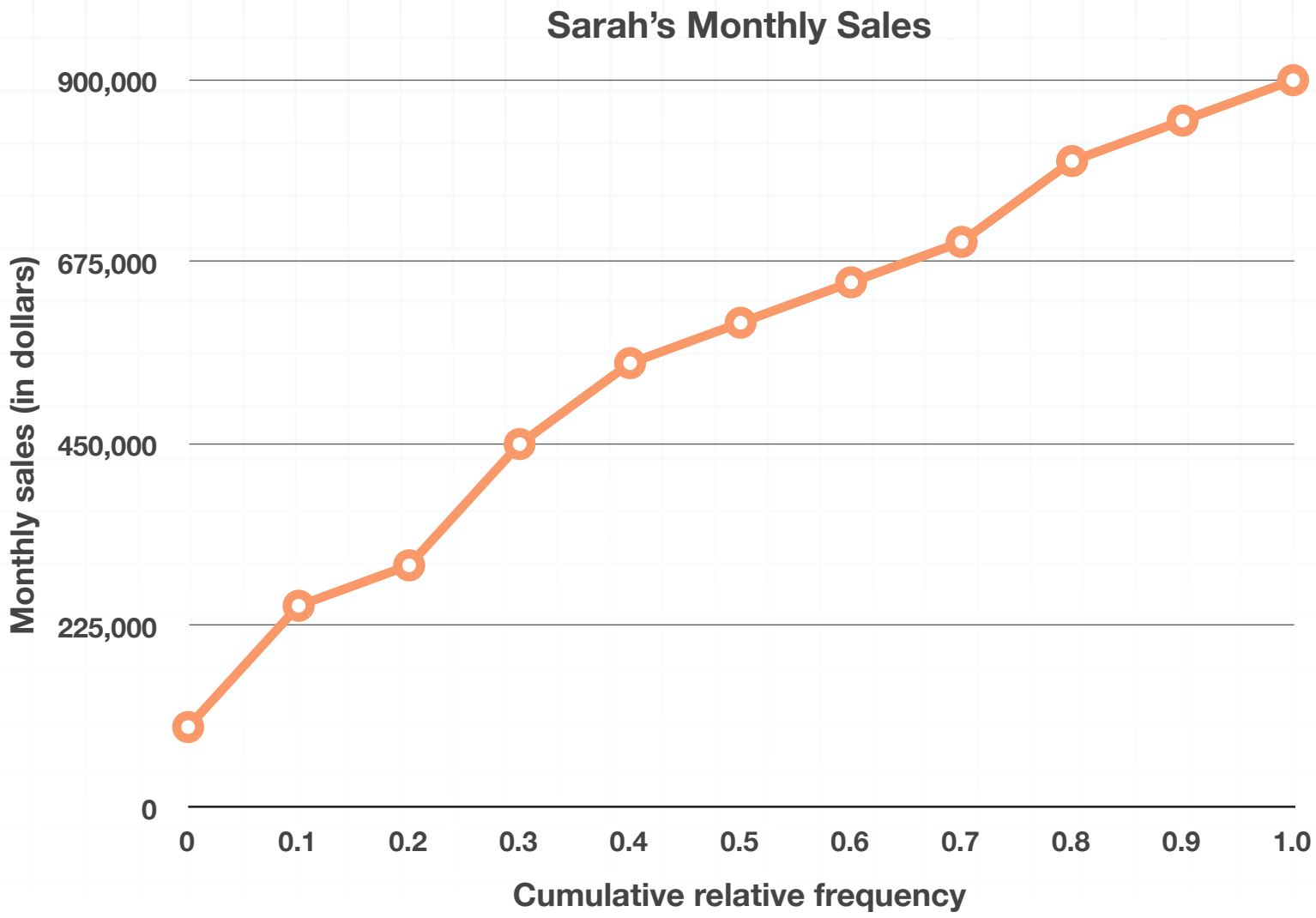
You can use the difference between the cumulative frequencies to find the frequency of each value and create a line graph.

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 cumulative relative frequency plot below. What is the meaning of the circled point?



Solution:

In her fifth month at the company, Sarah had sold $\$550{,}000$ worth of cars. This amounts to $40\,\%$ of her total sales since she's worked at the company.

We know it's Sarah's fifth month at the company because the circle is around the fifth point and we're told that the graph represents the total monthly sales of Sarah's sales to date. The circled point is between $\$500{,}000$ and $\$600{,}000$, so its value in monthly sales must be $\$550{,}000$.

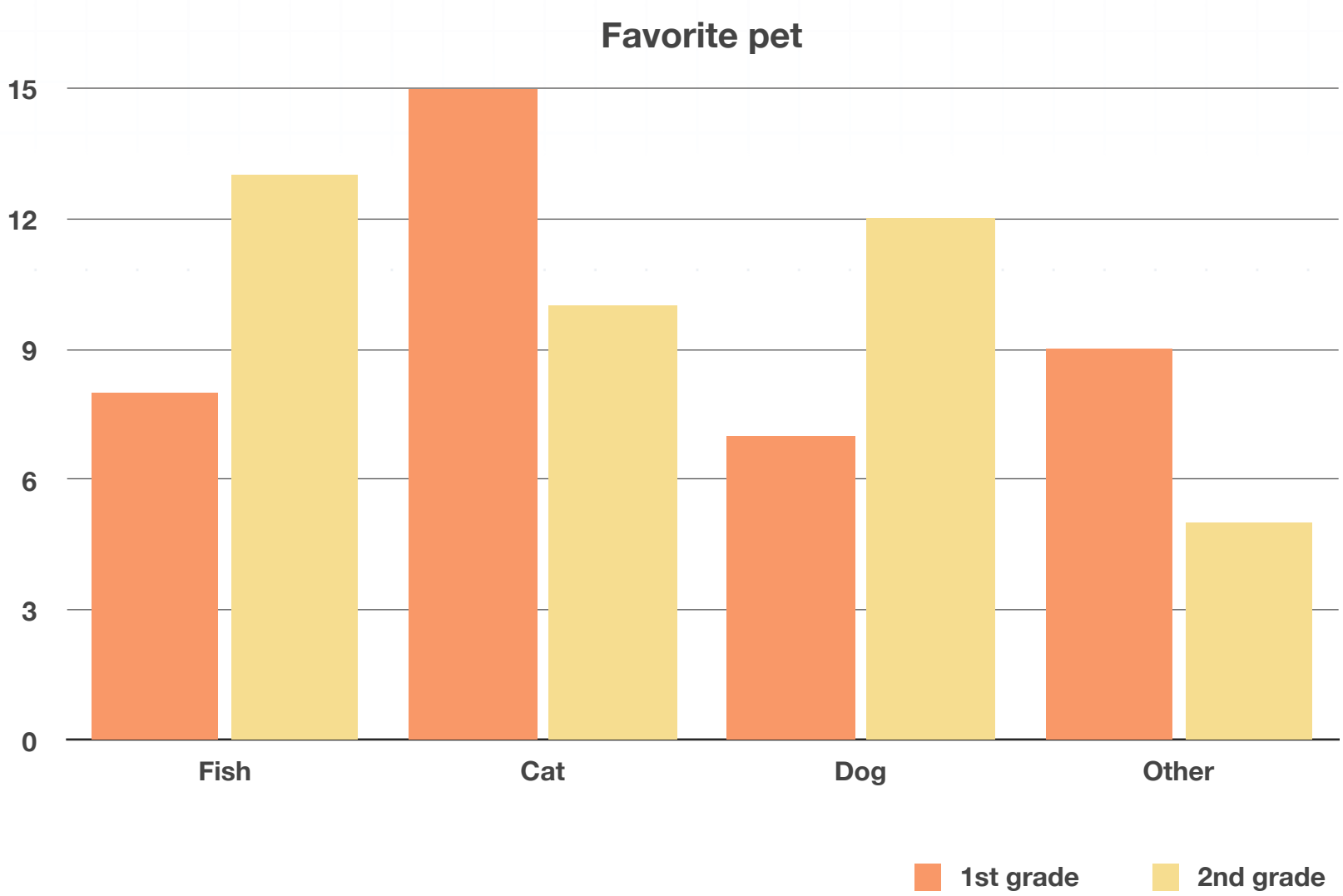
TWO-WAY DATA

■ 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.

What type of data is the pizza parlor collecting, one-way or two-way? 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:

The idea of a row or column variable is that it's how you could 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.

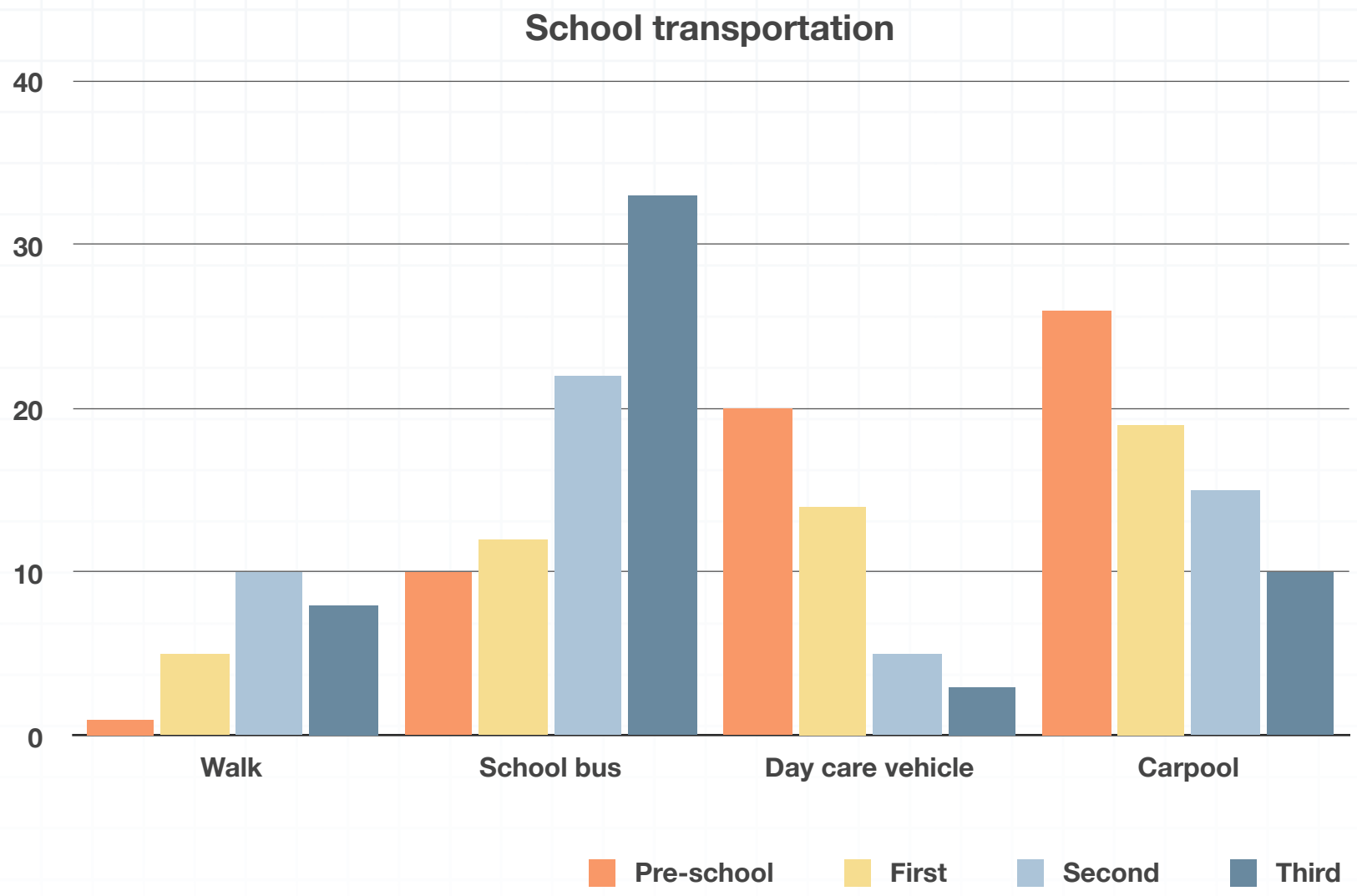
		Method of transportation			
		Walk	School bus	Day care vehicle	Carpool
Grade in school	Pre-school	1	10	20	26
	First	5	12	14	19
	Second	10	22	5	15
	Third	8	33	3	10

■ 4. Which graph would be a better choice to display the data from the two-way table: a comparison bar graph or a comparison line graph? Create your chosen 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. Here we're comparing grades in school, so the comparison bar graph is the best choice. Remember when you create a comparison bar graph you 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. Here are his survey results:

Snack	Yes	Yes	No	No	No	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No
Sleepy	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes

Create a two-way data table for Eric's survey.

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**

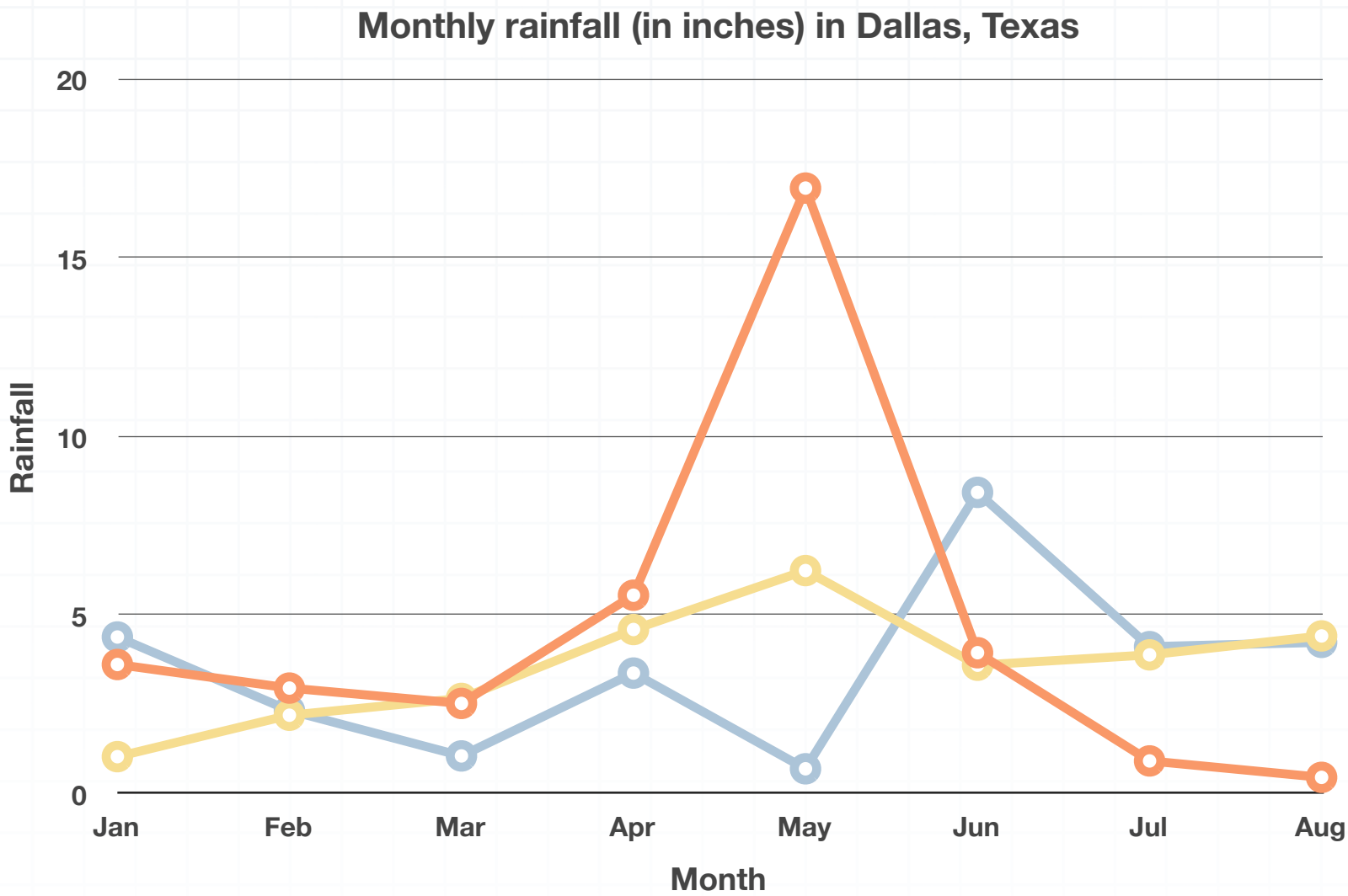
■ 6. Is a comparison line graph an appropriate visual display for the data table, which shows monthly rainfall (in inches) for Dallas, Texas, January -

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

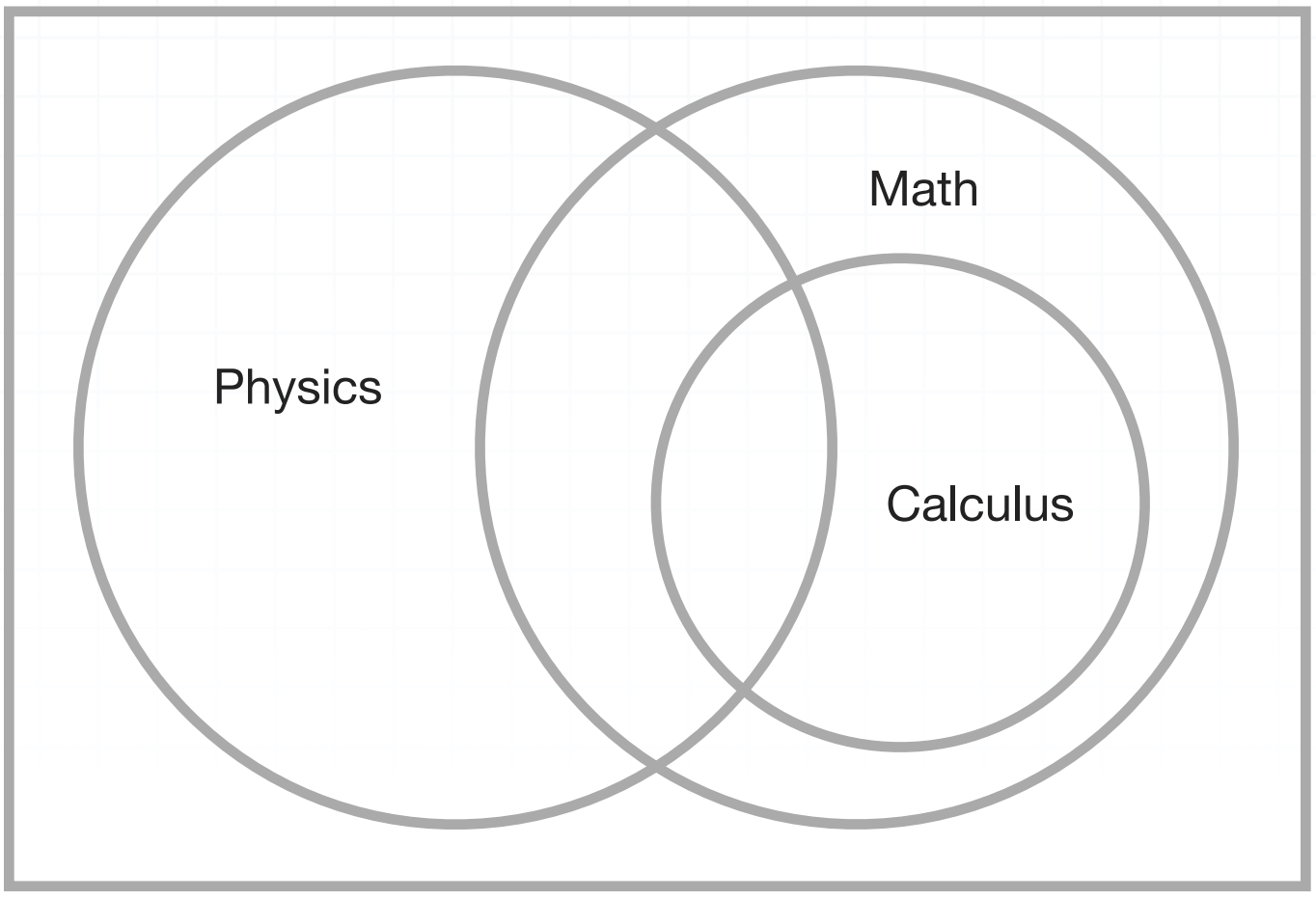
Solution:

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?



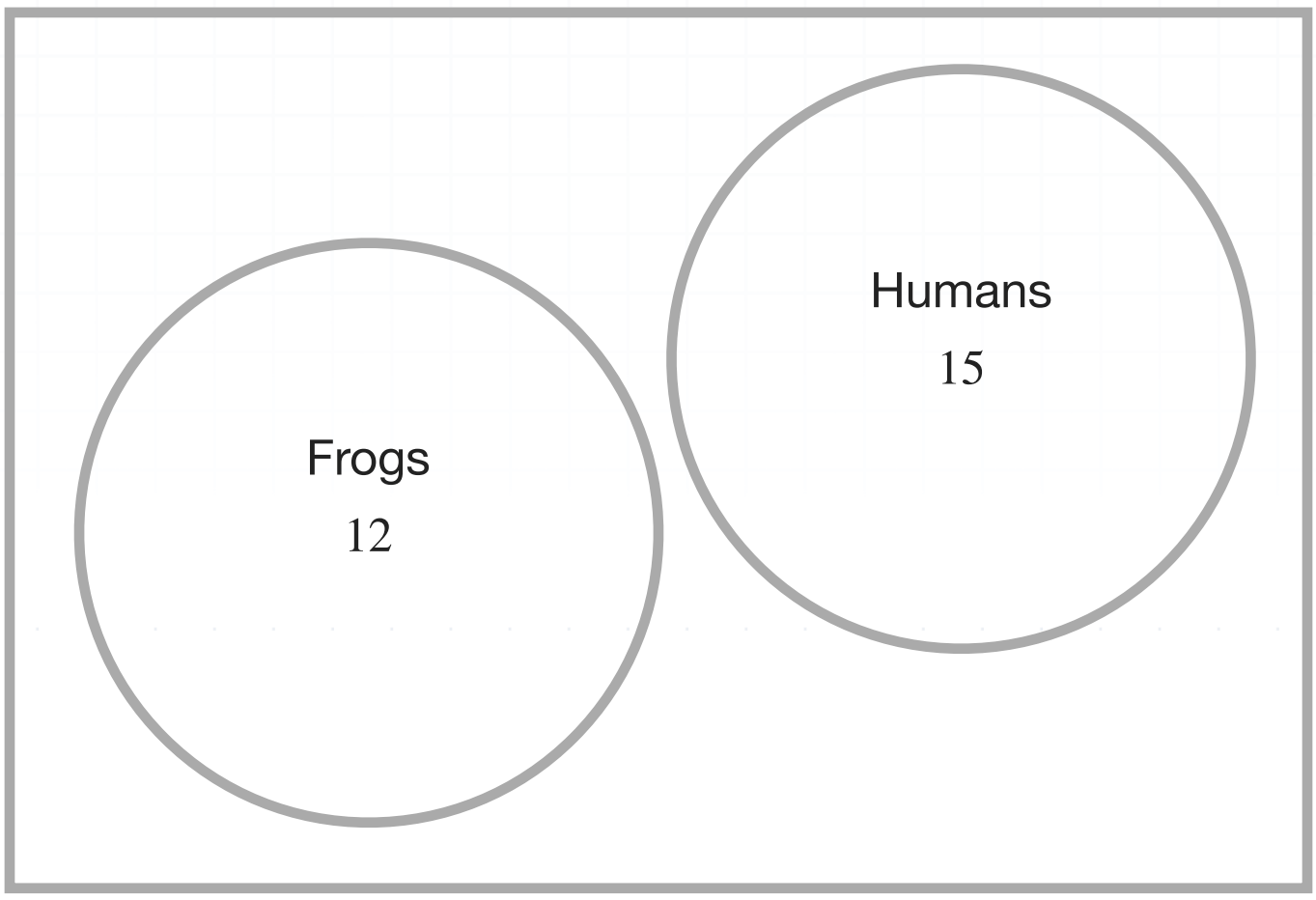
Solution:

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.

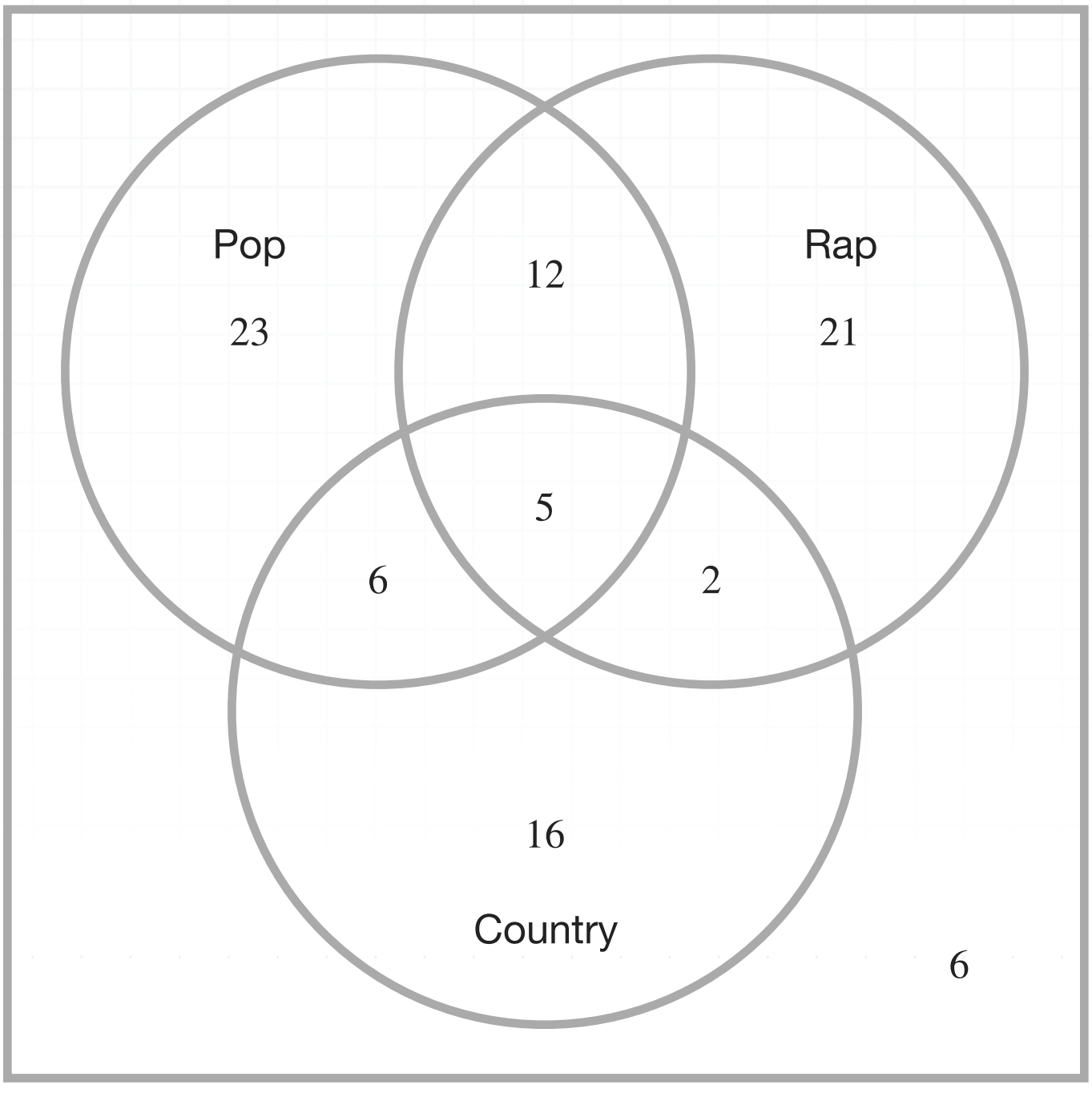
Solution:

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.

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

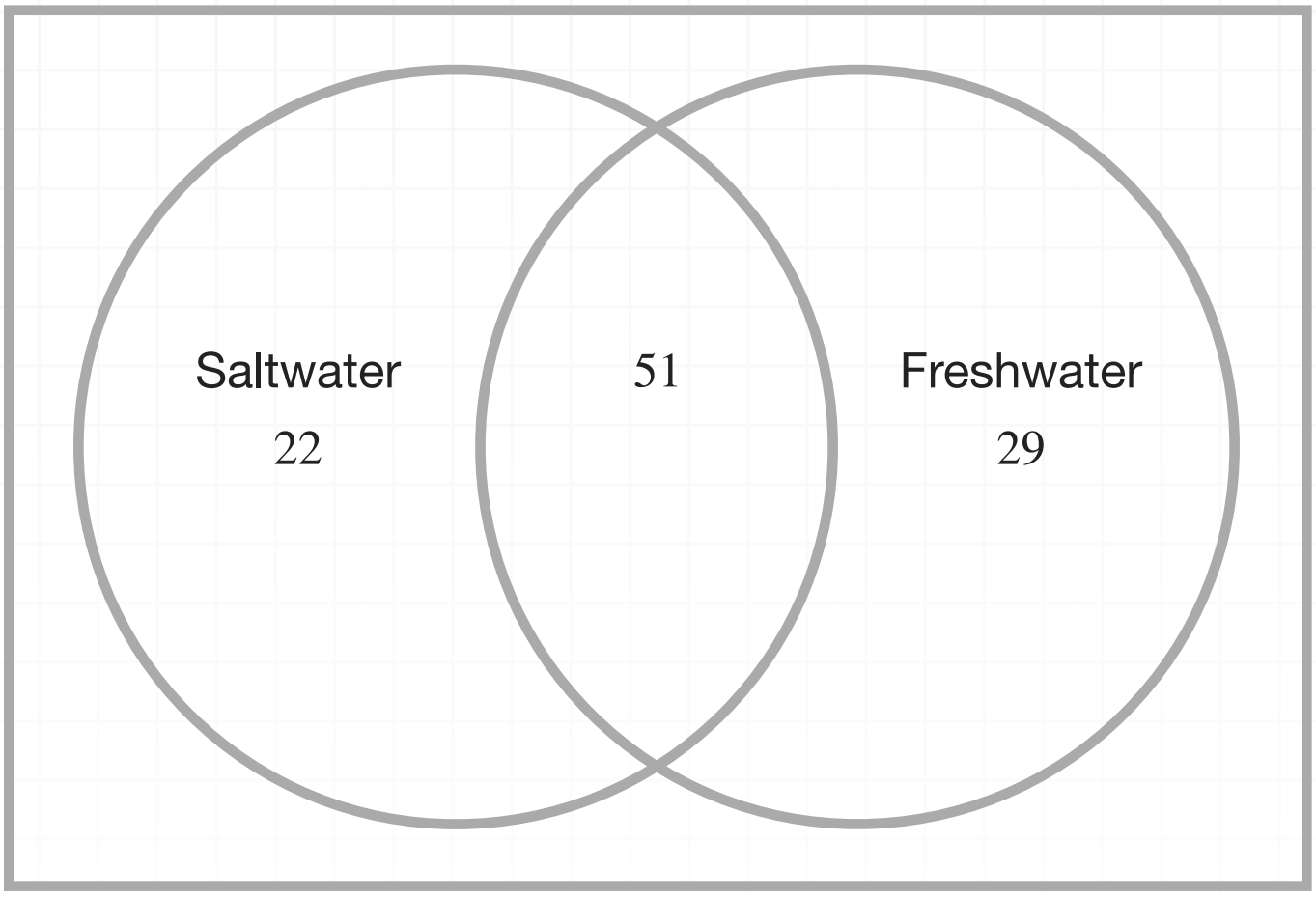
This is the number of students who participated in the survey. The students who did not like Pop music are those who only liked Rap (21), Country (16), Country and Rap (2), or something else (6). That adds to

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

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

$$\frac{45}{91} = 49\,\%$$

■ 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 meat 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 is a combination of fresh and saltwater.



Solution:

$50\,\%$ of the minnows were from brackish water. You can find the total number of minnows caught in the sample by adding $22 + 51 + 29 = 102$. 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 percent 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.

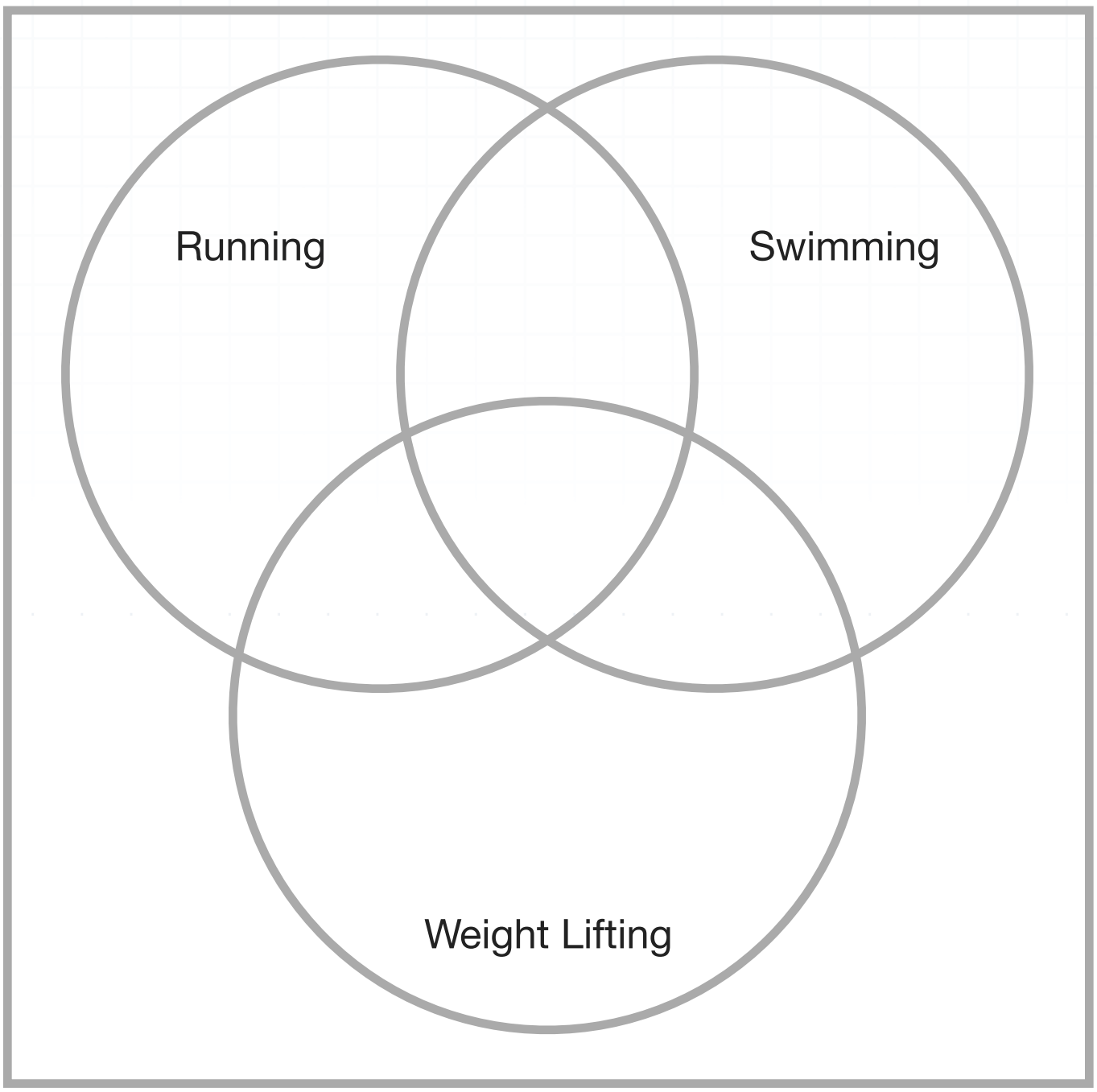
10 people only liked weight lifting.

5 people liked swimming and weight lifting equally.

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

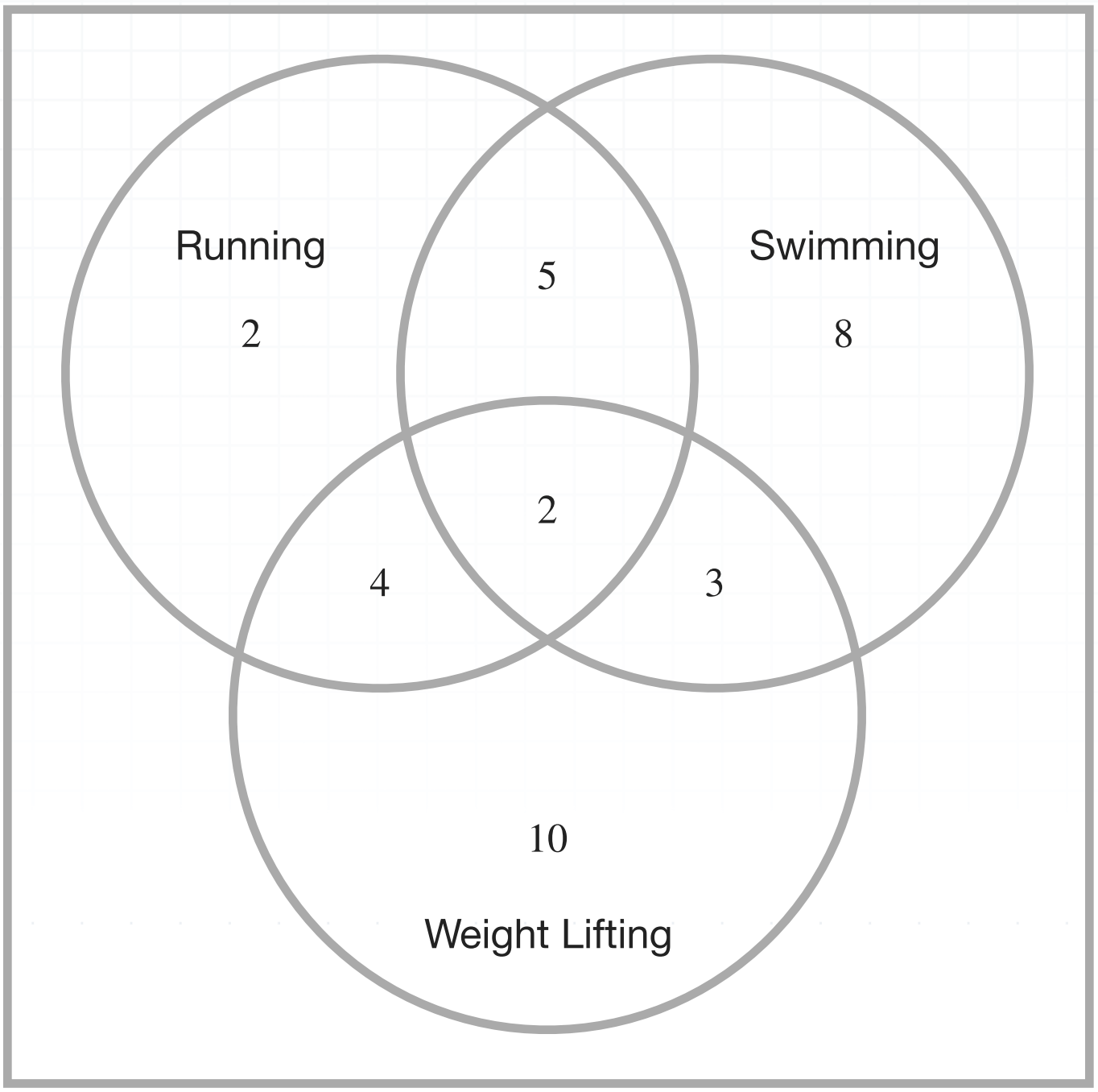
5 people liked running and swimming equally, but not weight lifting.

2 people liked all three equally.



Solution:

From the information we were given, this is the Venn diagram:

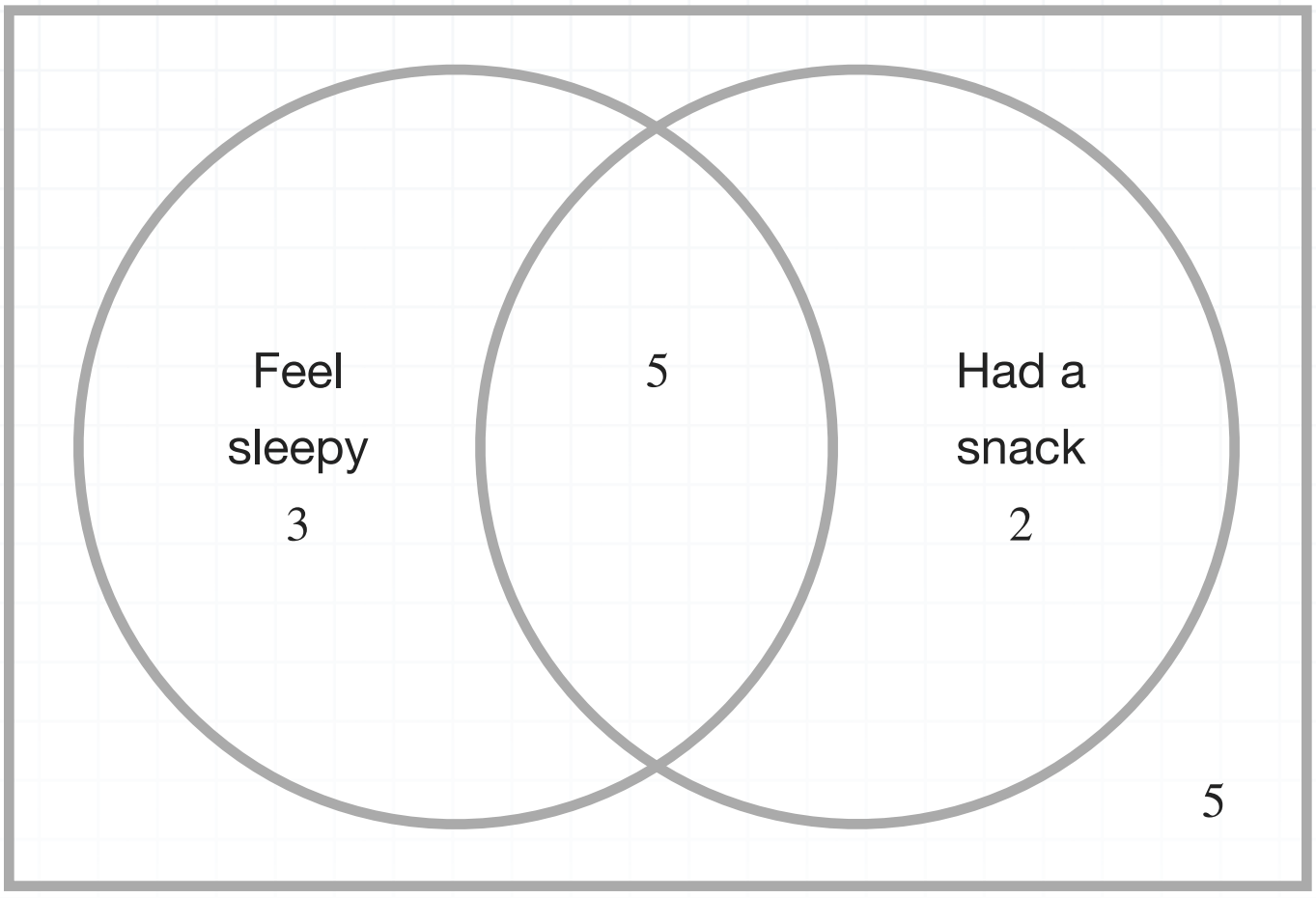


■ 6. Eric creates a survey asking students who ate a snack in the morning between classes if they felt sleepy or not. He organizes his survey results into a two-way data table. Draw a Venn diagram for Eric's survey results.

		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.



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
Total	**64**

Now we'll find the crashes on each day as a percentage of the total. In the table below, the total actually adds to $101\,\%$, due to rounding error, but the highest percentage of accidents is still occurring on Saturday.

Day of the week	Number of crashes
Sunday	13/64=20%
Monday	10/64=17%
Tuesday	8/64=13%
Wednesday	6/64=9%
Thursday	2/64=3%
Friday	11/64=17%
Saturday	14/64=22%
Total	**64/64=100%**

■ 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, did you use a marginal, joint, or conditional distribution?

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

Solution: To answer the question, you need to use a conditional distribution. You'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 you 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**	**10,674**

Solution: Remember that for the marginal distributions, we're just looking at the total column and total row. $10{,}678$ of the $21{,}348$ people in the study took the pain reliever, and $10{,}674$ took the placebo. This means

$$\frac{10{,}678}{21{,}348} \approx 50\,\%$$

took the pain reliever and about

$$\frac{10{,}678}{21{,}348} \approx 50\,\%$$

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

$$\frac{21{,}046}{21{,}348} \approx 99.5\,\%$$

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**	**10,674**

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%**

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 175 people in the study who had minor heartburn, 4 took the pain reliever and 171 took the placebo.

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

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%**

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.

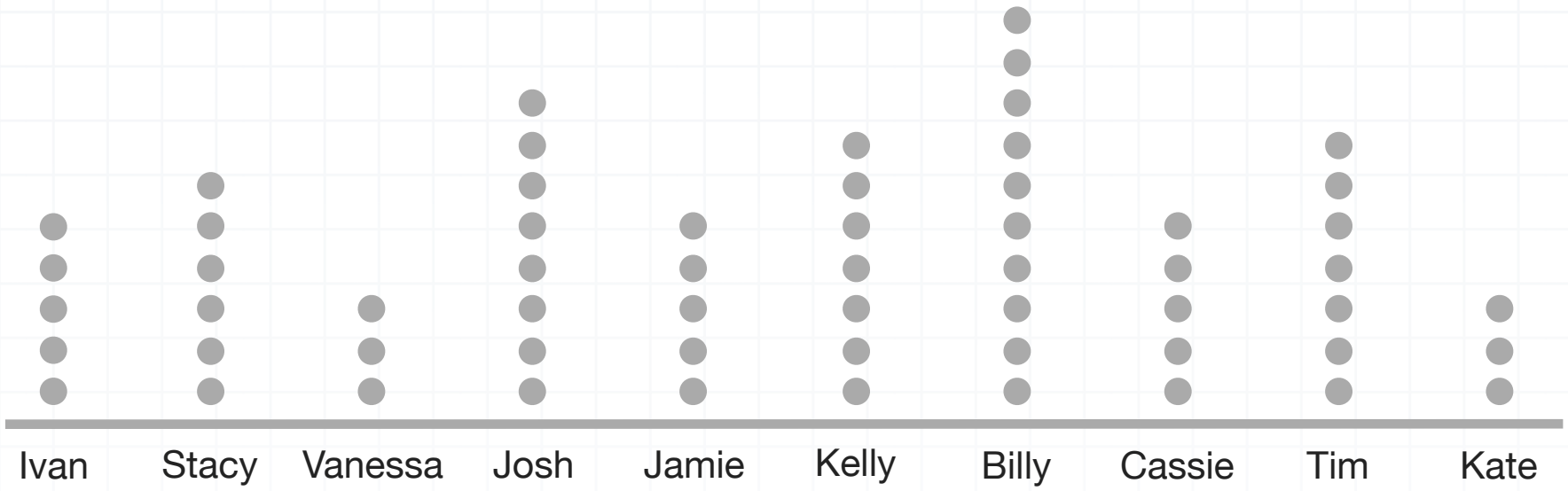
FREQUENCY TABLES AND DOT PLOTS

■ 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.

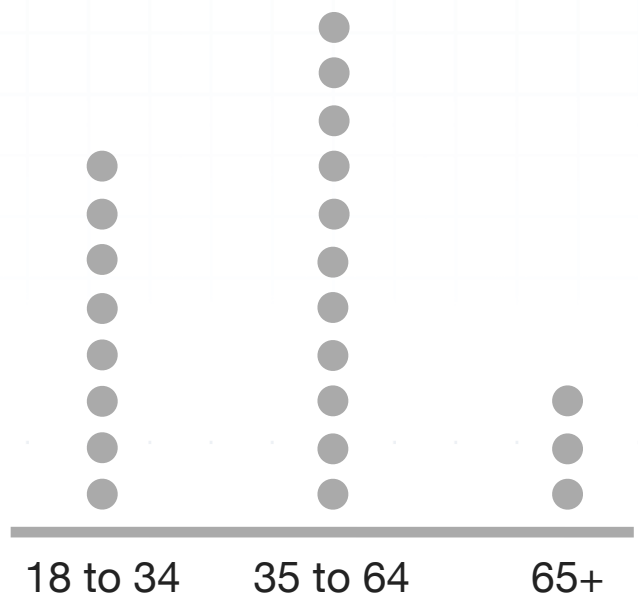
Name	Packets sold
Ivan	5
Stacy	6
Vanessa	3
Josh	8
Jamie	5
Kelly	7
Billy	10
Cassie	5
Tim	7
Kate	3

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	11
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, 1, 2, 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, blue, red, red, red, yellow, orange, orange, green, green, green, black

Solution:

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

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

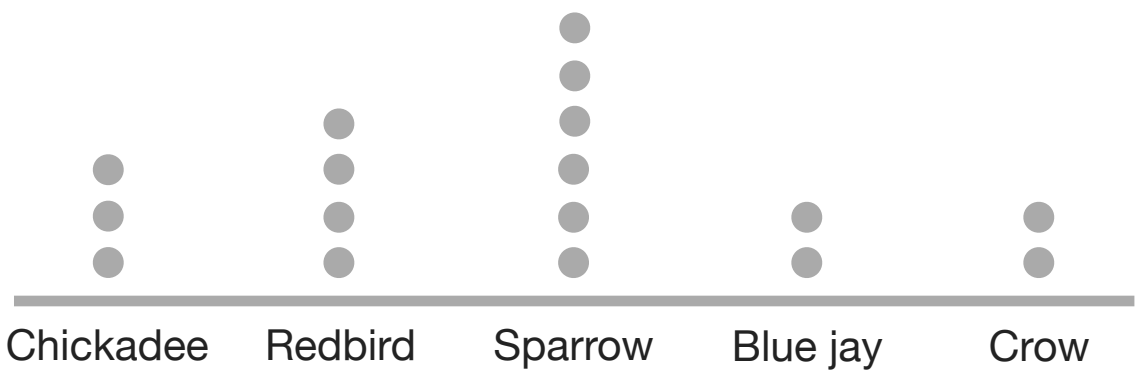
■ 5. Kevin watches birds from his window and records what kind he sees. Create a dot plot from the data.

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

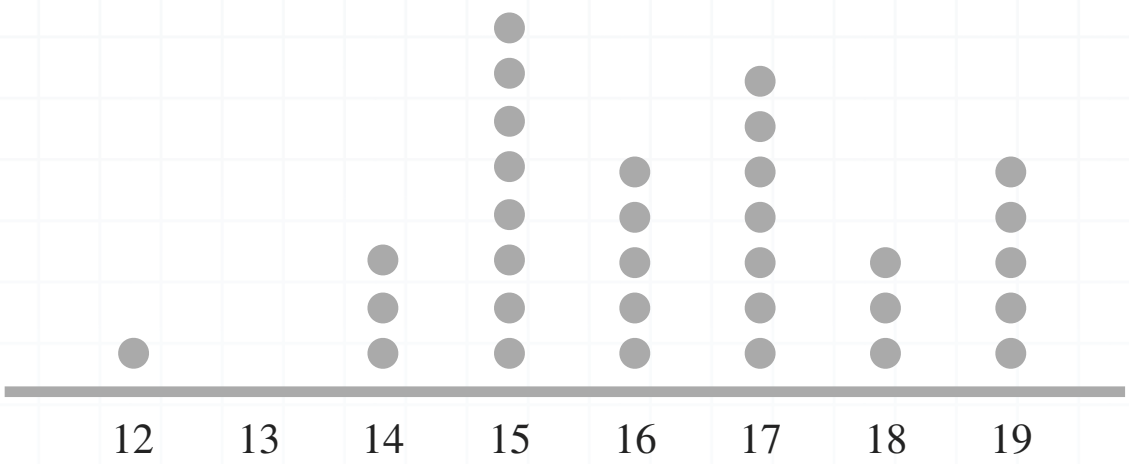
It can help to make a frequency table first. The birds are not grouped according to kind so make sure you count how many of each type you have.

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

Now 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.

$$5 + 7 + 3 = 15$$

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 steam 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

$$1 \mid 0 = 10$$

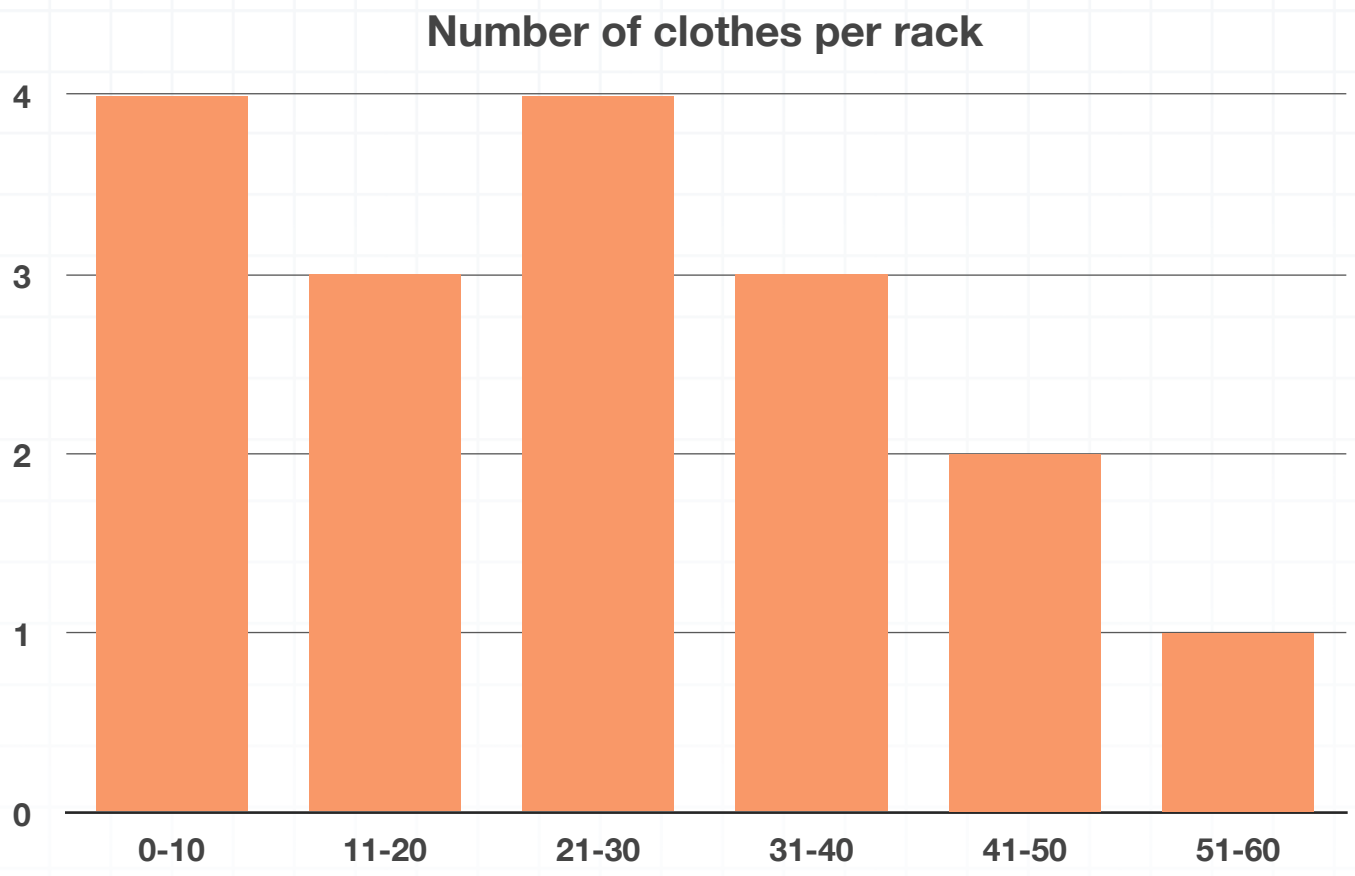
Solution:

The steam and leaf plot counts by 10s along the left-hand side. You can then see how many data points should go into each bucket because we're using buckets of size 10.

In this way, the stems become the buckets and the number of leaves become the frequencies graphed in the histogram.

Buckets	Number of leaves
0-10	0 1 2 8
11-20	8 8 8
21-30	2 6 8 9
31-40	4 4 4
41-50	2 6
51-60	0

Now you can turn this frequency table into a histogram.

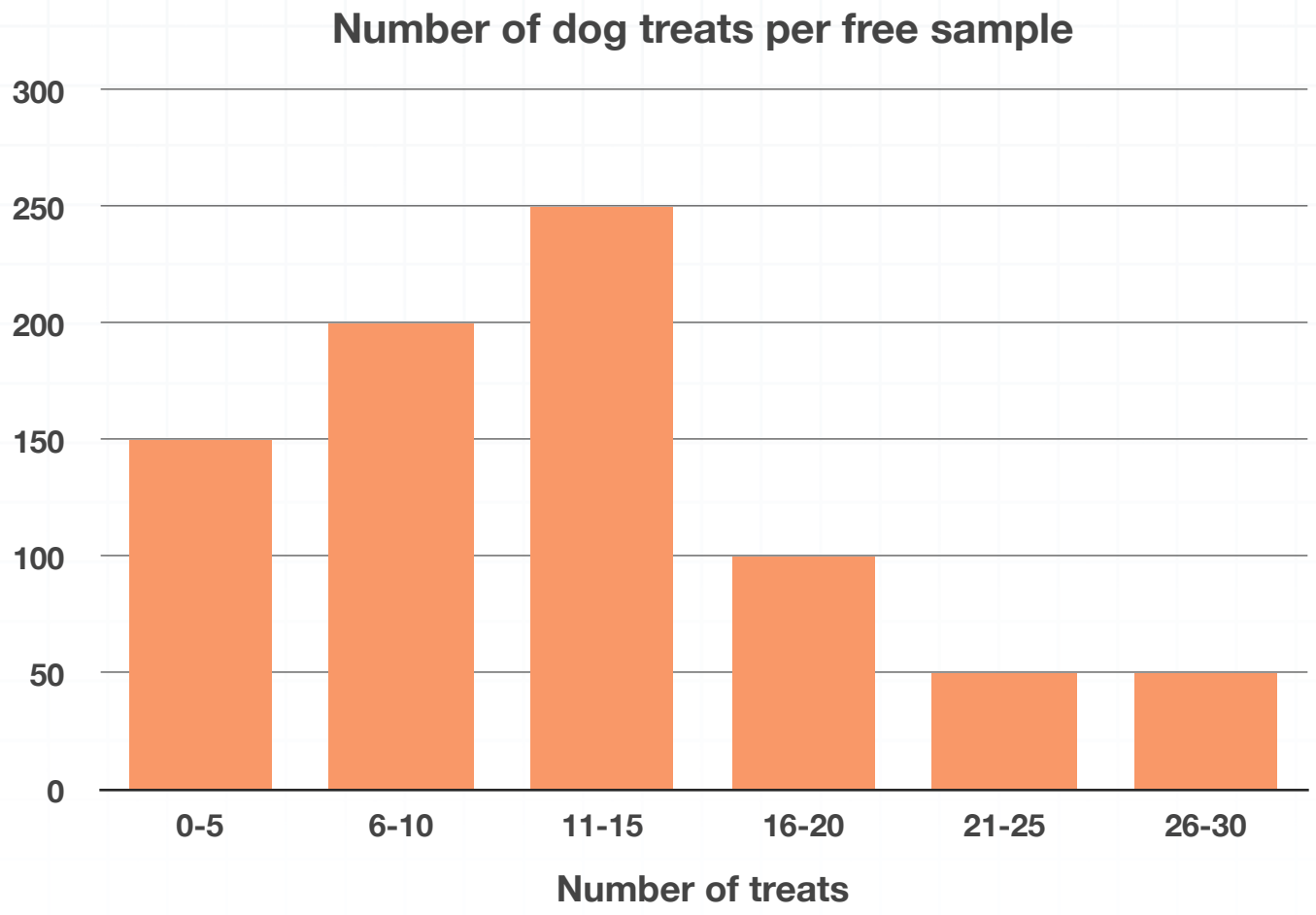


3. Is it possible to create a stem-and-leaf plot from a histogram? Why or why not?

Solution:

You 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.

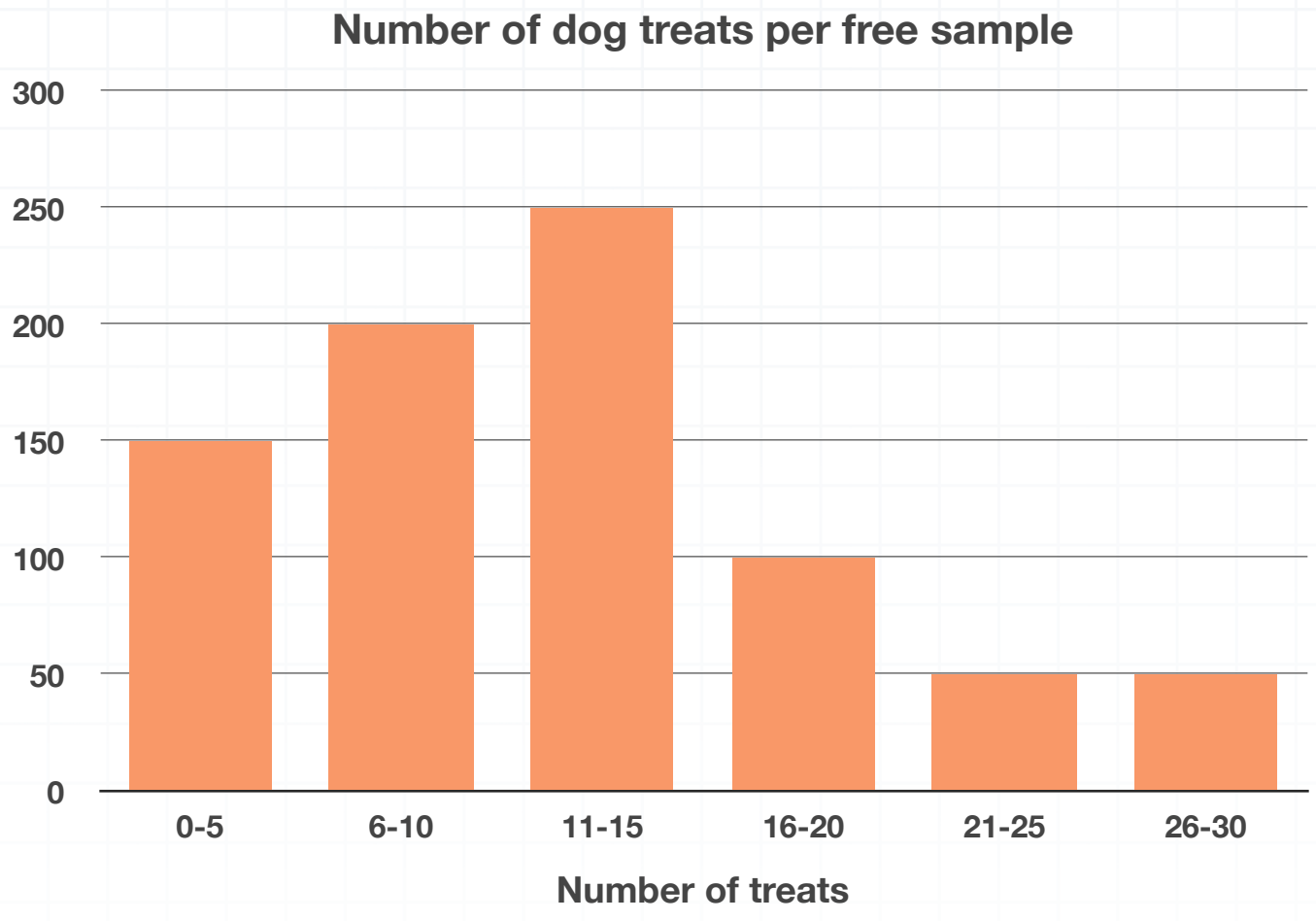
4. A company mails out packets of dog treat samples based on a consumer's previous dog food purchases. How many times did the company mail a packet of $11 - 15$ treats?



Solution:

The bar that is labeled $11 - 15$ 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?



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 your 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 you 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

$$6 \mid 0 = 60$$

CENTRAL TENDENCY: MEAN, MEDIAN, AND MODE

■ 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:

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

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

$$\mu \approx 283.75$$

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.

$$\cancel{62, 64, 69, 70}, 70, 71, \cancel{73, 74, 75, 77}$$

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

$$1 \mid 3 = 13$$

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 you add to the data set that would give you 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.

$$\cancel{1,\ 2,\ 8},\ 13,\ \cancel{20,\ 30,\ 31}$$

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 17, which means we'll need to insert a number larger than 13.

$$\cancel{1,\ 2,\ 8},\ 13,\ \_\_,\ \cancel{20,\ 30,\ 31}$$

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 mean 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: 20

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 20 students who took the test. Let's call Samantha's missing test score s. Then we can set up this equation:

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

$$85 = \frac{s + 3(75) + 5(80) + 4(82) + 6(95) + 1(98)}{20}$$

When we solve for s, we get

$$85(20) = s + 3(75) + 5(80) + 4(82) + 6(95) + 1(98)$$

$$1{,}700 = s + 3(75) + 5(80) + 4(82) + 6(95) + 1(98)$$

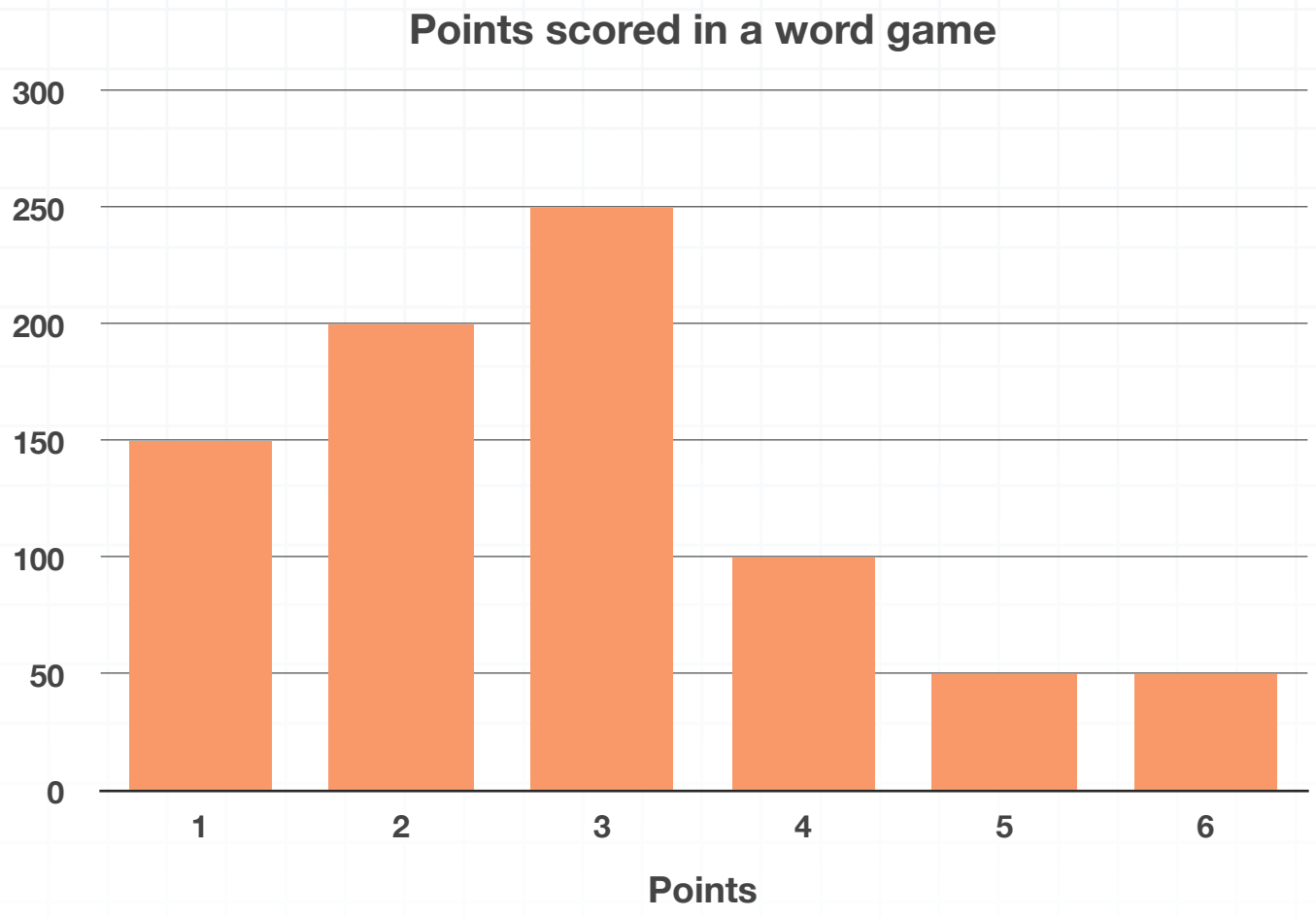
$$1{,}700 = s + 225 + 400 + 328 + 570 + 98$$

$$s = 1{,}700 - 225 - 400 - 328 - 570 - 98$$

$$s = 79$$

Samantha scored a 79 on her test.

6. What is the mode of the data set?



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.

SPREAD: RANGE AND IQR

■ 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.

$$\cancel{0,\ 1,\ 1,\ 1,\ 2},\ 5,\ 5,\ \cancel{7,\ 7,\ 18,\ 24,\ 24}$$

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 $40 - 20 = 23$.

If you are wondering why we didn't really need the frequency side of the table, consider that, since this is a frequency table, the frequency tells us how many times each number appears. The list of data is actually

$$20, 20, 25, 25, 25, 32, 38, 39, 39, 40, 40, 40, 43, 43$$

So the range is still 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.

$$\cancel{2,\ 2,\ 2,\ 2,\ 2,\ 5,\ 5},\ 8,\ 12,\ \cancel{12,\ 13,\ 13,\ 15,\ 15,\ 15,\ 21}$$

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

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

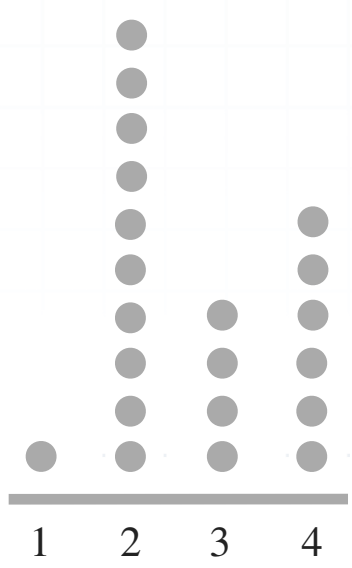
Therefore, the interquartile range is $14 - 2 = 12$.

■ 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 you 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, you 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.

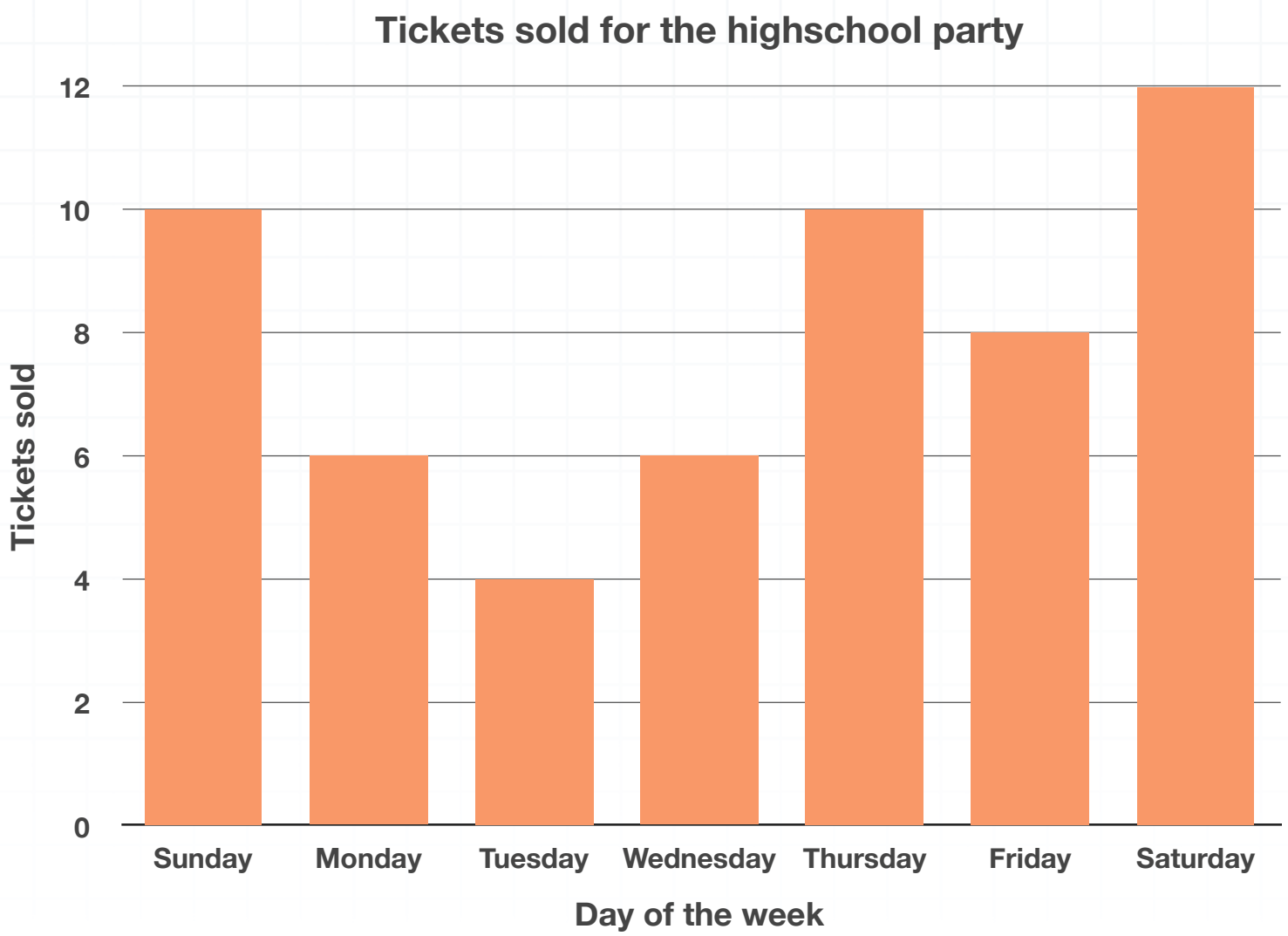
■ 5. The dot plot shows the number of trips to the science museum for a class of 4th graders. What is the range of the data set?



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.

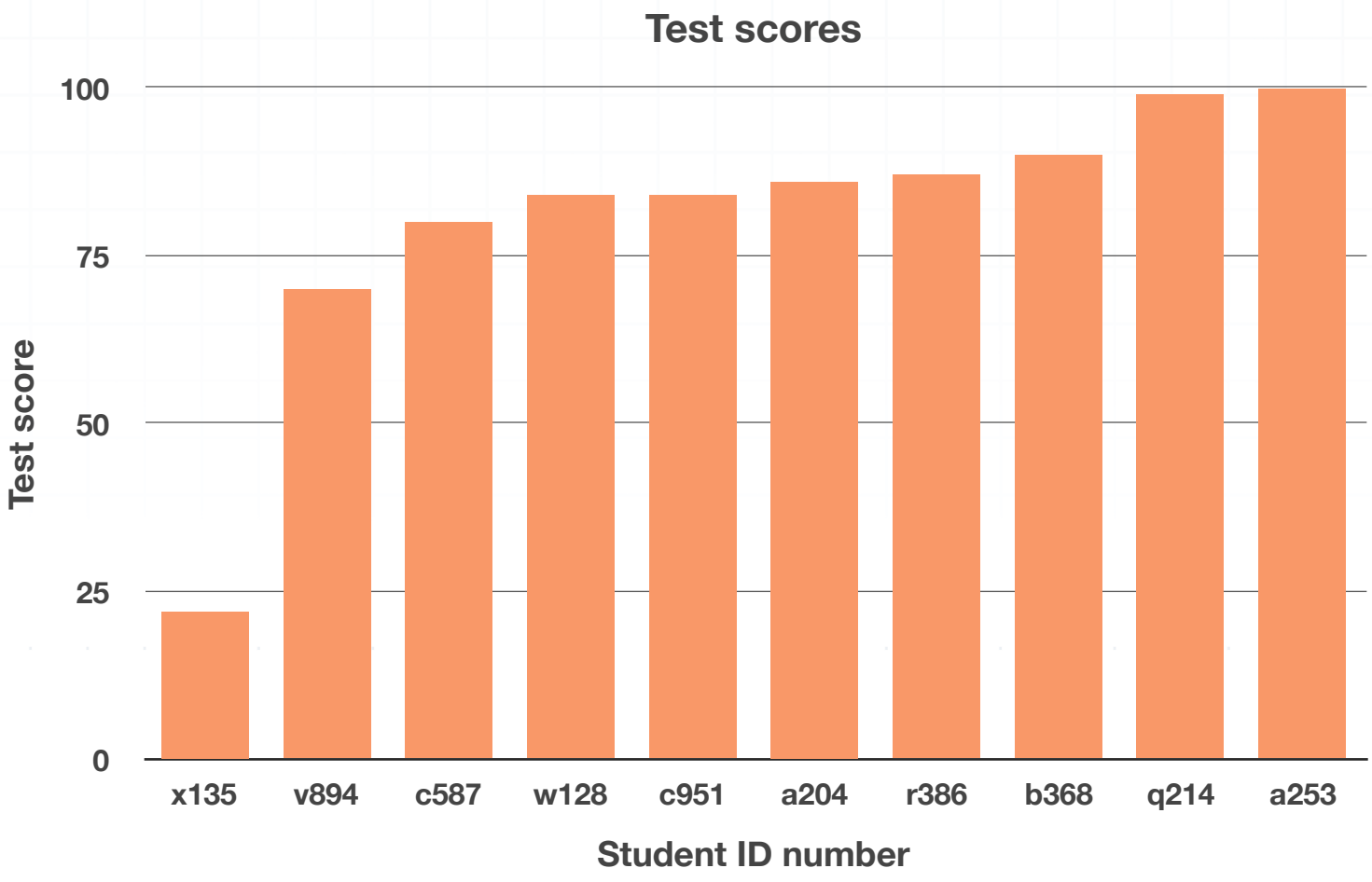
The original mean is 65 and the new mean is $65(1.10) = 71.5$. The original range is 25, and the new range is $25(1.10) = 27.5$.

■ 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 you 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?

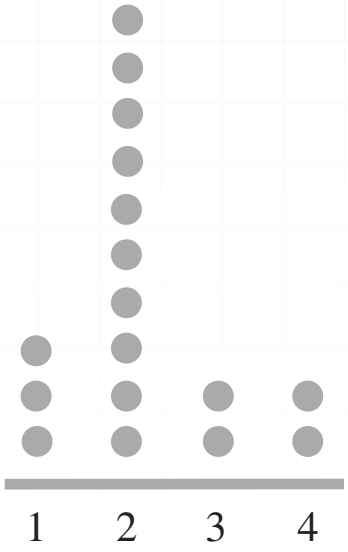


Solution:

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 you about the potential outliers in the data set?



Solution:

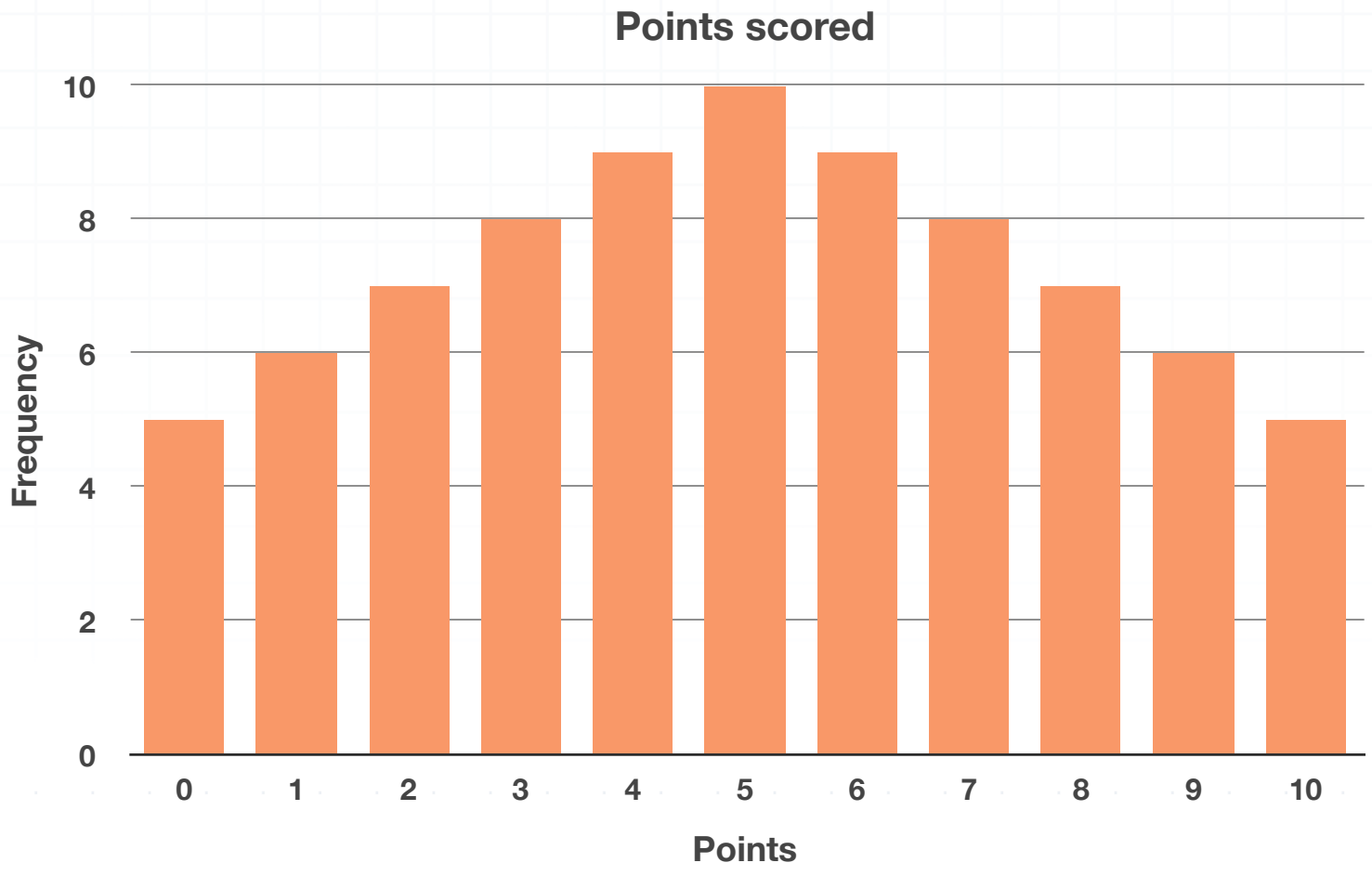
You can calculate the mean and median from the dot plot. The median is 2 and the mean is

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

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 you 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° Celsius

Median: 101° Celsius

Mode: 99° Celsius

Range: 7° Celsius

IQR: 4° 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: $102^\circ(9/5) = 183.6^\circ$ Celsius

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

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

Range: $7^\circ(9/5) = 12.5^\circ$ Celsius

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

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: $183.6^\circ + 32^\circ = 215.6^\circ$ Celsius

Median: $181.8^\circ + 32^\circ = 213.8^\circ$ Celsius

Mode: $178.2^\circ + 32^\circ = 210.2^\circ$ Celsius

Range: 12.5° Celsius

IQR: 7.2° Celsius

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

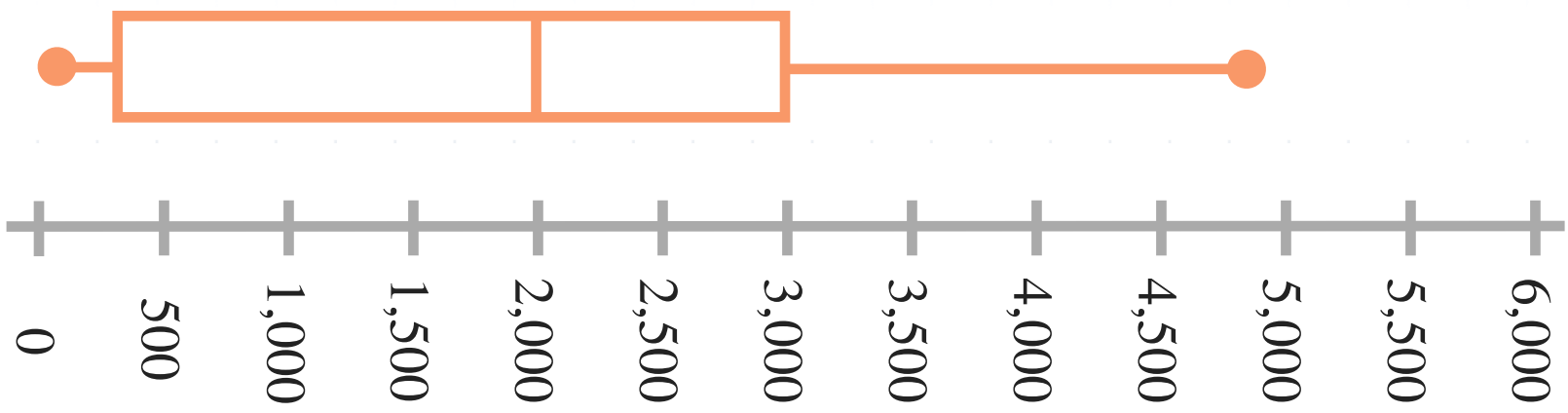
Mean: 1,887.5

Third Quartile: 3,050

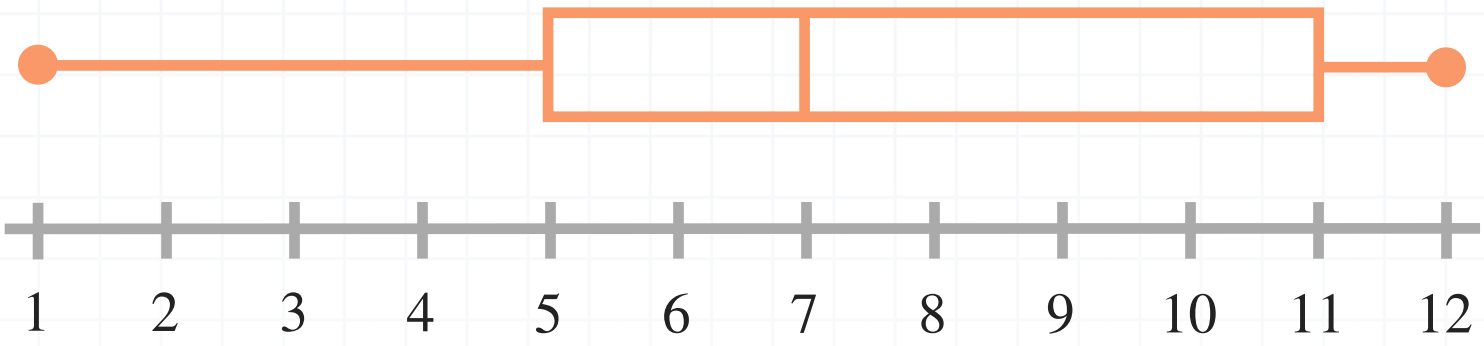
Maximum: 4,800

Solution:

To create a box plot, you 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?

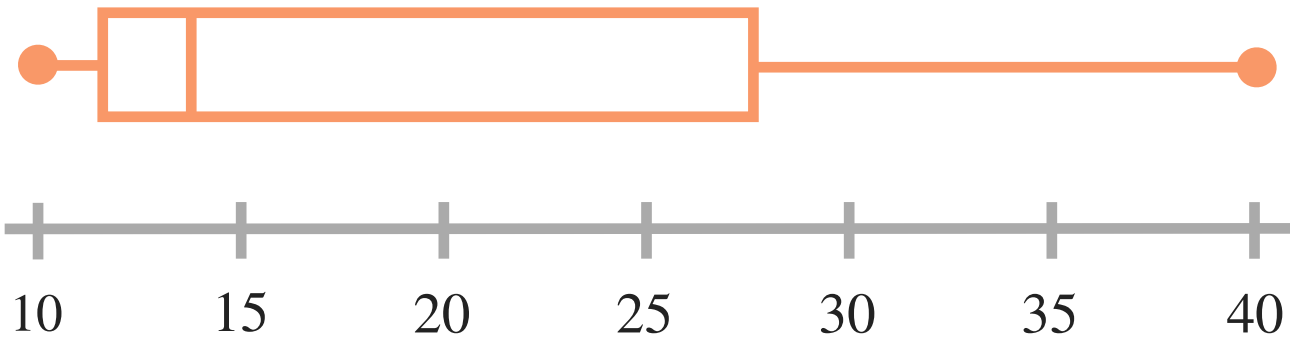


Solution:

The line from 1 to 5 is longer because the numbers in that quartile are more spread. In comparison the numbers in the quartile between 11 and 12 are less spread out.

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 quartile 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 quartile 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: $1 \mid 3 = 13$

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.

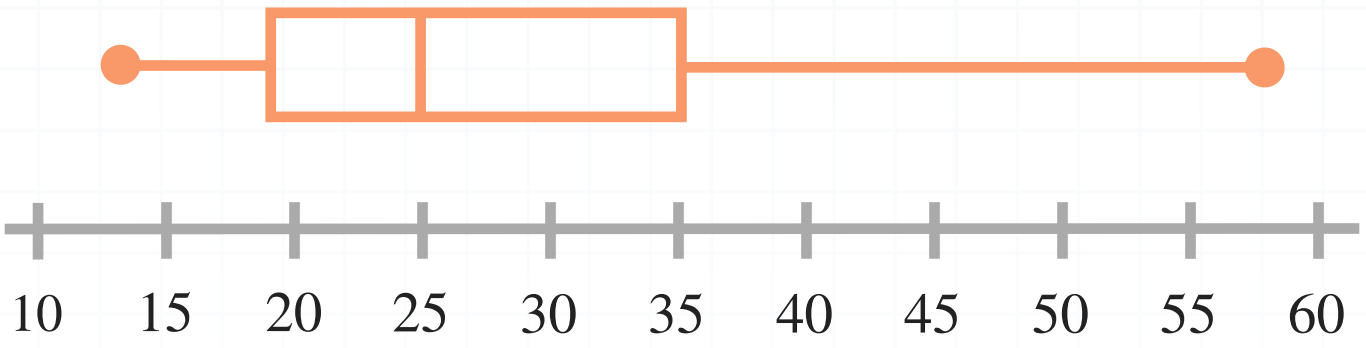
$$13,\ 17,\ 18,\ 21,\ 24,\ 26,\ 35,\ 35,\ 52,\ 56$$

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.



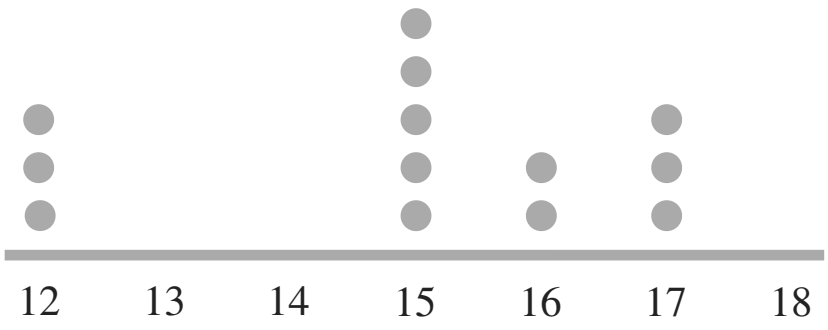
MEAN, VARIANCE, AND STANDARD DEVIATION

■ 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 to the nearest hundredth? Use the sample mean rounded to the nearest hundredth for your calculation.



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

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

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

$$S^2 = \frac{28.9569}{12}$$

$$S^2 \approx 2.39$$

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		
Total		

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)^2=121
110	110-108=2	(2)^2=4
112	112-108=4	(4)^2=16
121	121-108=13	(13)^2=169
110	110-108=2	(2)^2=4
98	98-108=-10	(-10)^2=100
Total		**121+4+16+169+4+100=414**

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

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

$$S \approx 9.067$$

■ 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.

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

$$\sigma^2 = 106$$

The population standard deviation is therefore

$$\sqrt{\sigma^2} = \sqrt{106}$$

$$\sigma = 10.2956$$

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

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

for the data set. 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 another 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 you're measuring. If, for example, you were measuring the price of a soft drink at a state fair, and you found a standard deviation of $5 that is a large standard deviation. It is large because soft drinks usually do not cost very much and this would tell you that you need to hunt around for the best price.

On the other hand, if you were purchasing a specific type of car and you found that the standard deviation was $5 among the dealerships you were considering, that standard deviation would be very small. Small enough, in fact, that it wouldn't matter much where you bought the car because the prices were all 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		
Totals:		**100%**

Solution:

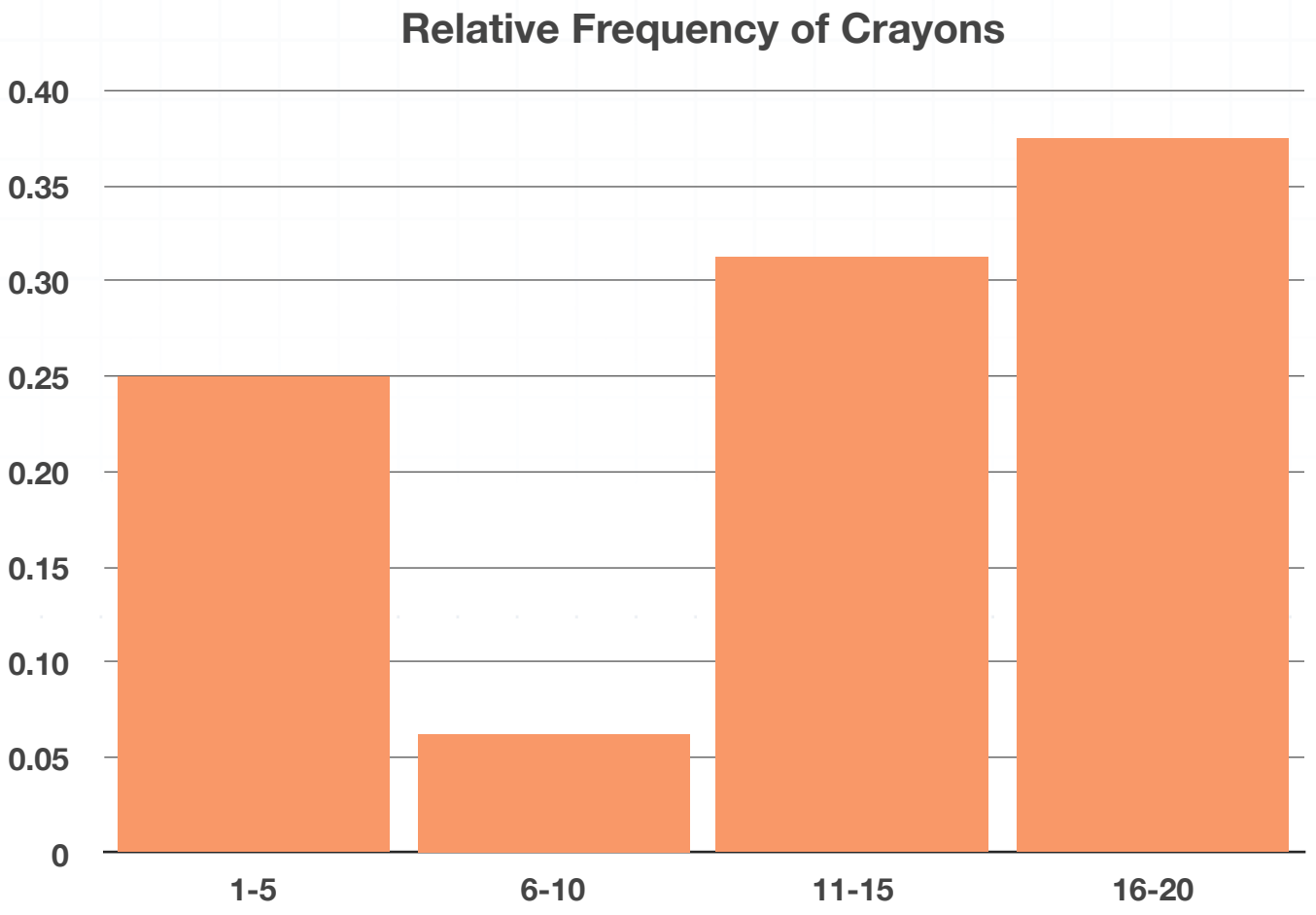
First count the number of items in each frequency interval and add that to the table, as well as calculate the total number of crayons.

Crayons	Frequency	Relative Frequency
1-5	4	
6-10	1	
11-15	5	
16-20	6	
Totals:	**16**	**100%**

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%
Totals:	**16**	**100%**

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.

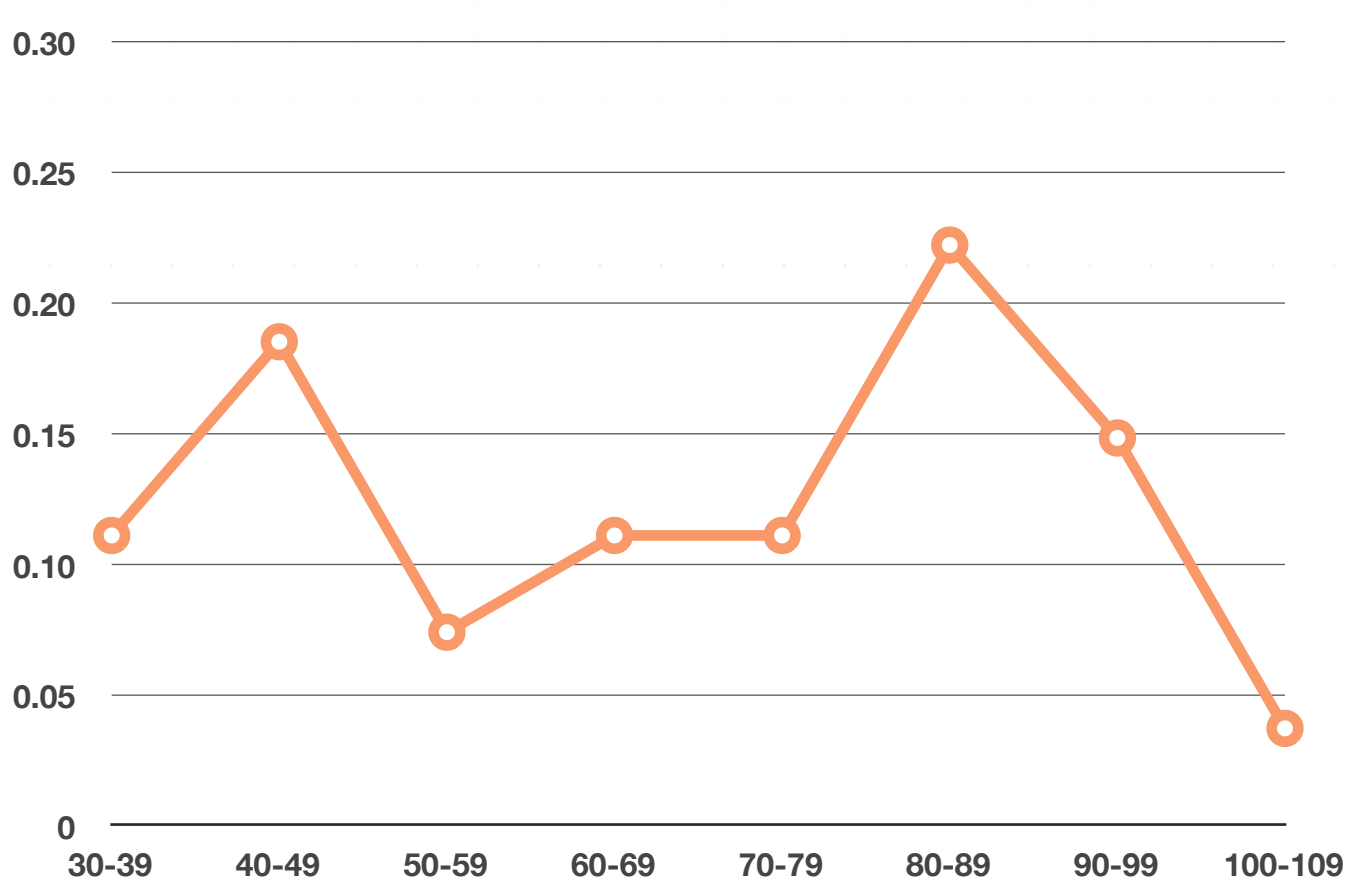
Score	Frequency	Relative Frequency
30-39		
40-49		
50-59		
60-69		
70-79		
80-89		
90-99		
100-109		
Totals:		

Solution:

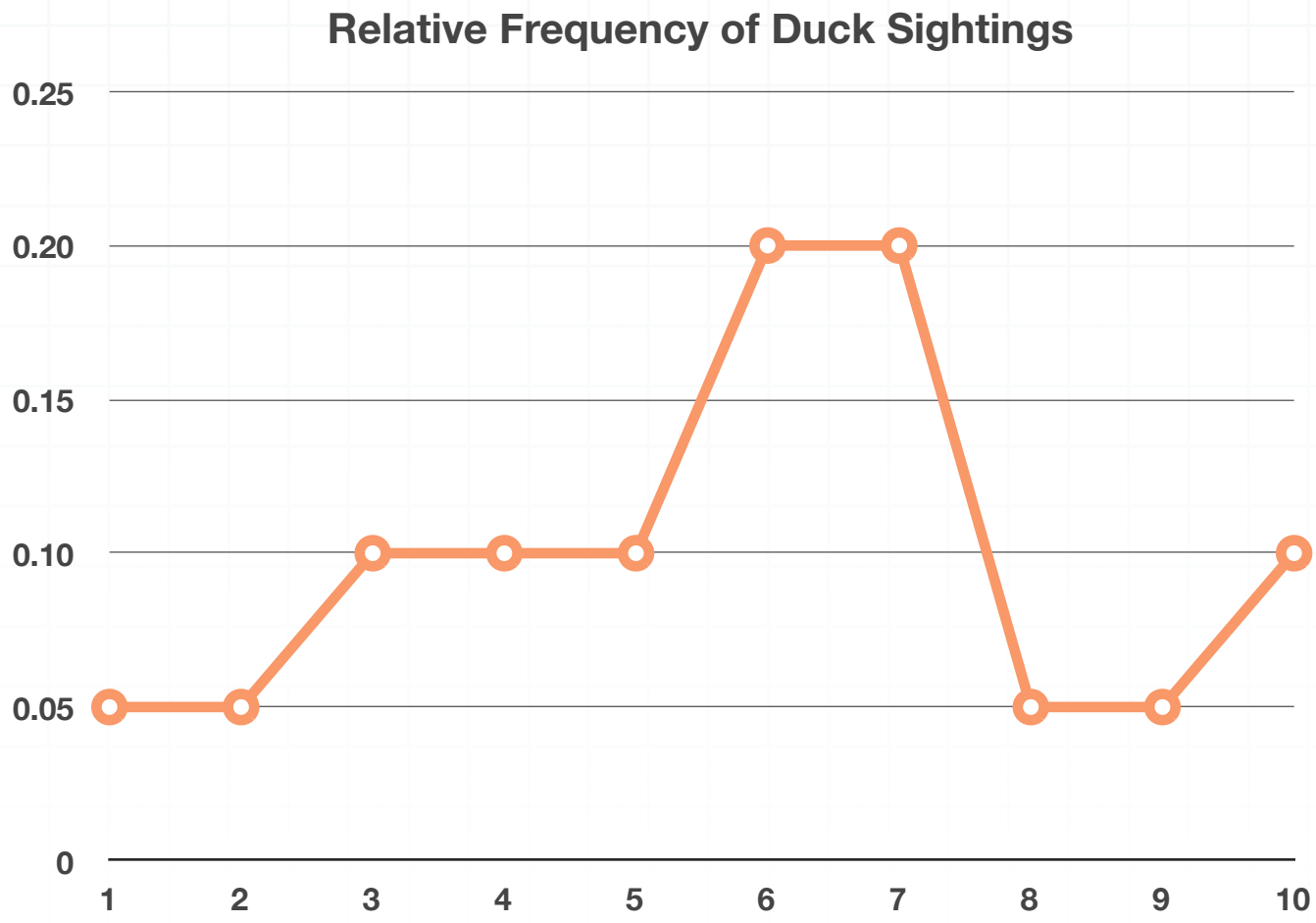
The first step to completing the frequency table is to count the scores in each interval, then use those frequencies and the total number of test scores to calculate the relative frequencies.

Score	Frequency	Relative Frequency
30-39	3	3/27=0.1111
40-49	5	5/27=0.1852
50-59	2	2/27=0.0741
60-69	3	3/27=0.1111
70-79	3	3/27=0.1111
80-89	6	6/27=0.2222
90-99	4	4/27=0.1484
100-109	1	1/27=0.0373
Totals:	**27**	**100%**

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



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?



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.

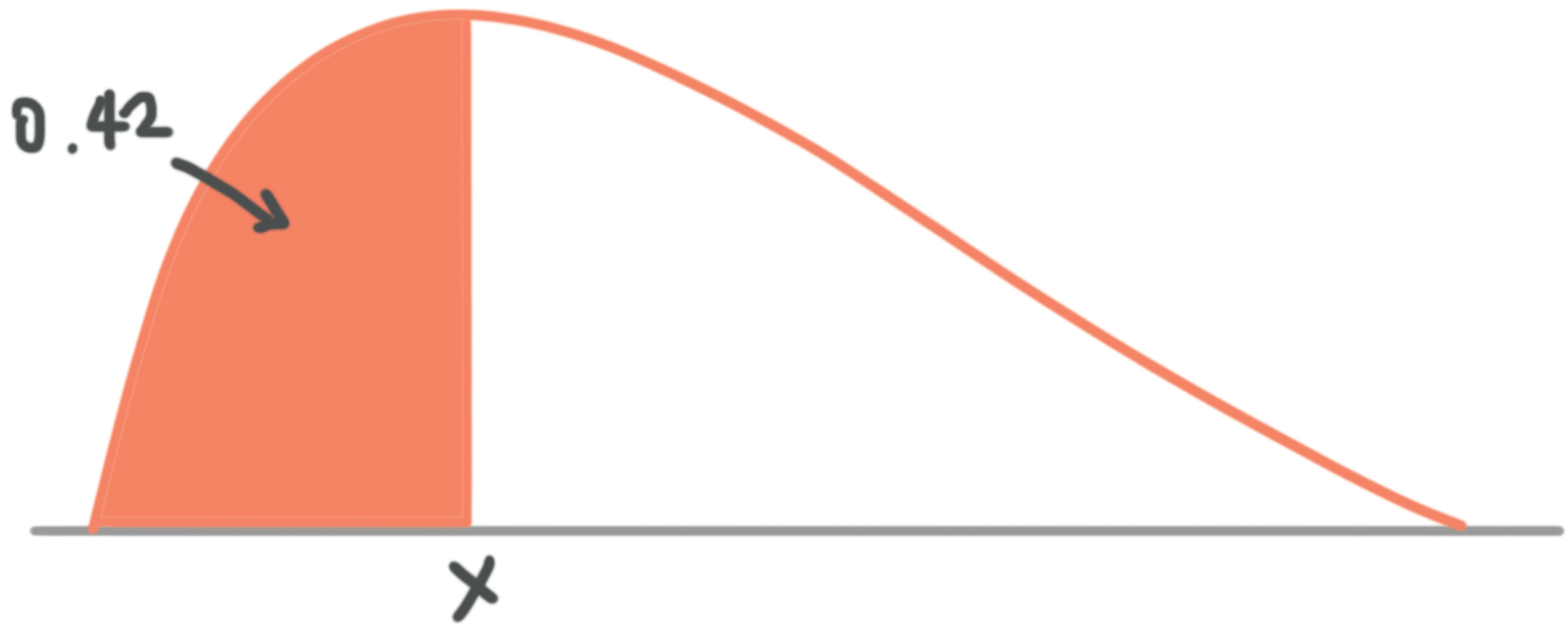
Ducks	Relative Frequency
6	0.20
7	0.20
8	0.05
9	0.05
10	0.10

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?



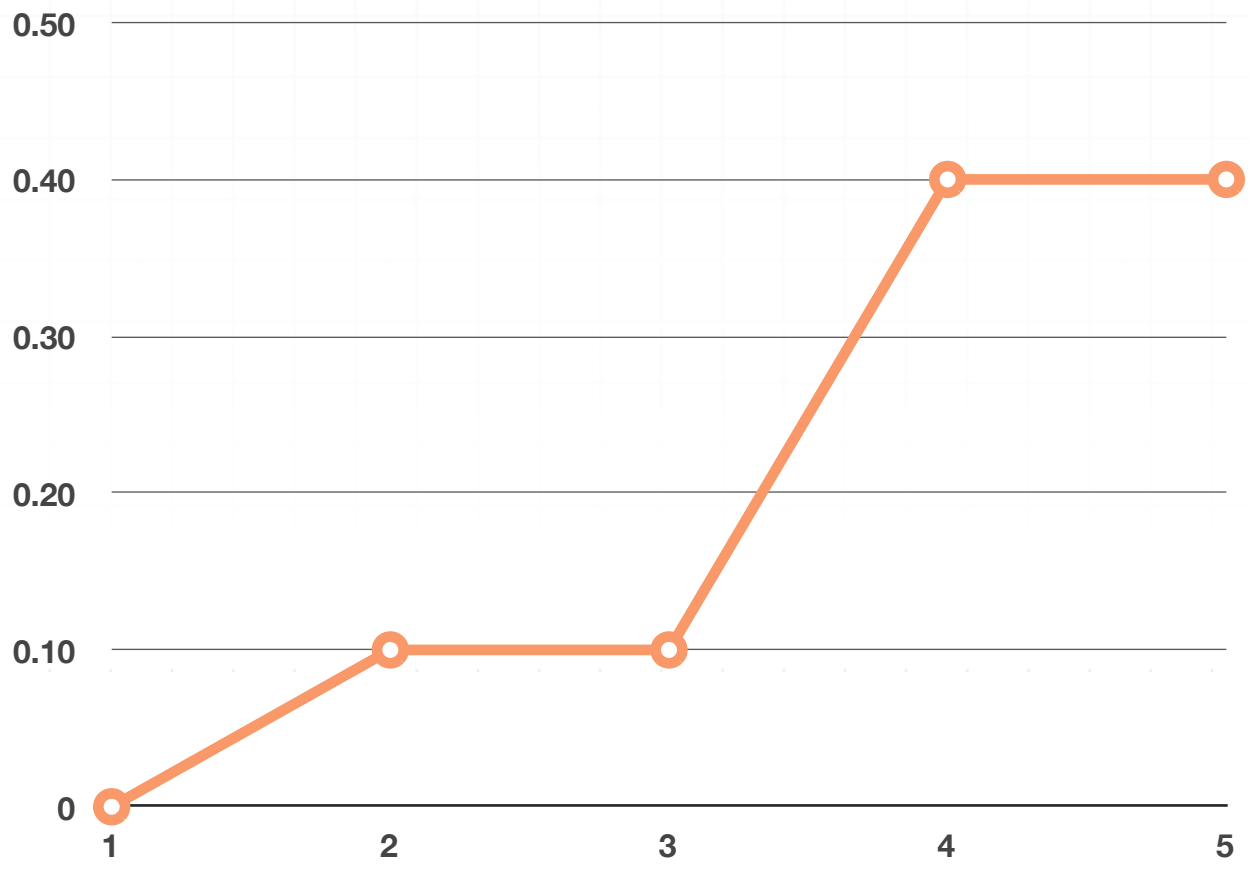
Solution:

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 2 is a triangle, so the area can be found as

$$A = \frac{1}{2}bh = \frac{1}{2}(1)(0.1) = 0.05$$

The area under the curve between 2 and 3 is a rectangle, so the area can be found as

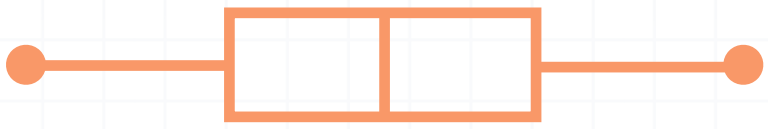
$$A = lw = (1)(0.1) = 0.1$$

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

$$1 - 0.1 - 0.05 = 0.85$$

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:

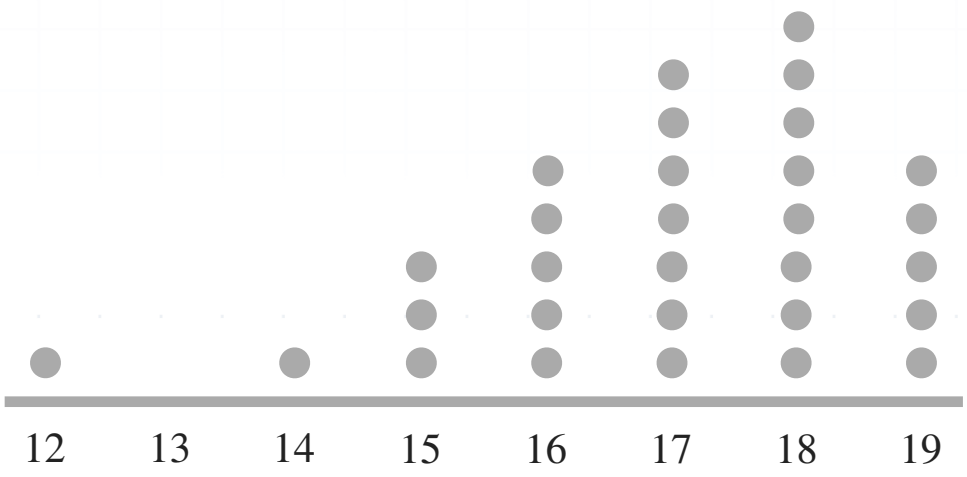
$$Q_3 + 1.5(\text{IQR})$$

$$25 + 1.5(4.5)$$

$$31.75$$

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.



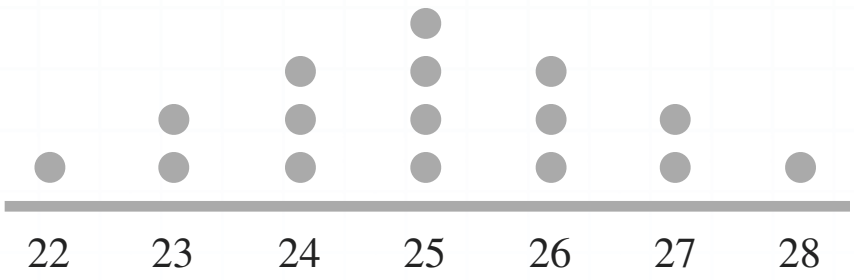
Solution:

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.

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

$$\sigma = 1.5811$$

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:

For this problem we need to find the z-score and then find the percentage from the z-table. We know that the mean is $\mu = 170$ and that the standard deviation is $\sigma = 8$. And the value we're interested in is $x = 154$. So the z-score is

$$z = \frac{154 - 170}{8} = \frac{16}{8} = 2$$

To find the percentage, look up the z-score in the z-table. The amount in the table is 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 and 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

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

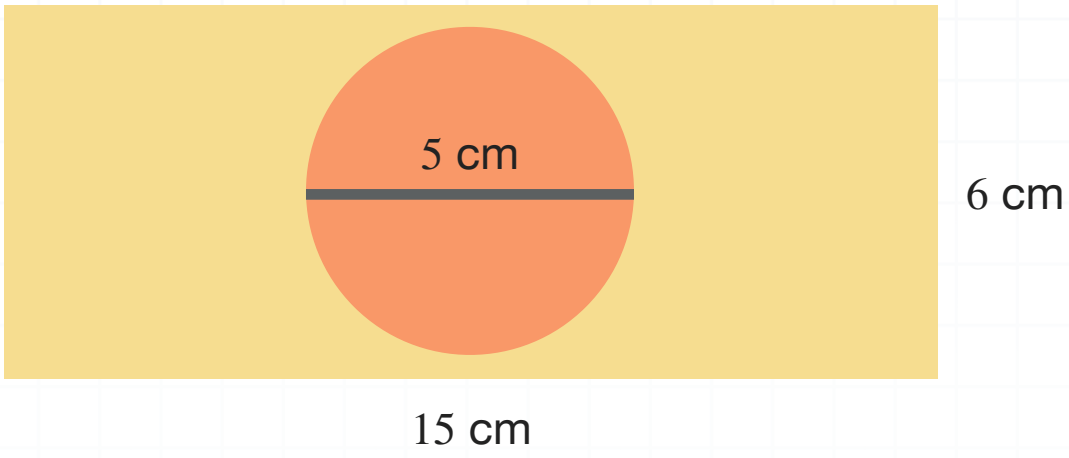
which gives $.9938$ in the z-table. Therefore, we can say that

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

percent of people have an IQ between 120 and 140.

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,

$$A_{\text{circle}} = \pi r^2$$

$$A_{\text{circle}} = \pi(2.5)^2$$

$$A_{\text{circle}} = 19.63 \text{ cm}^2$$

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:

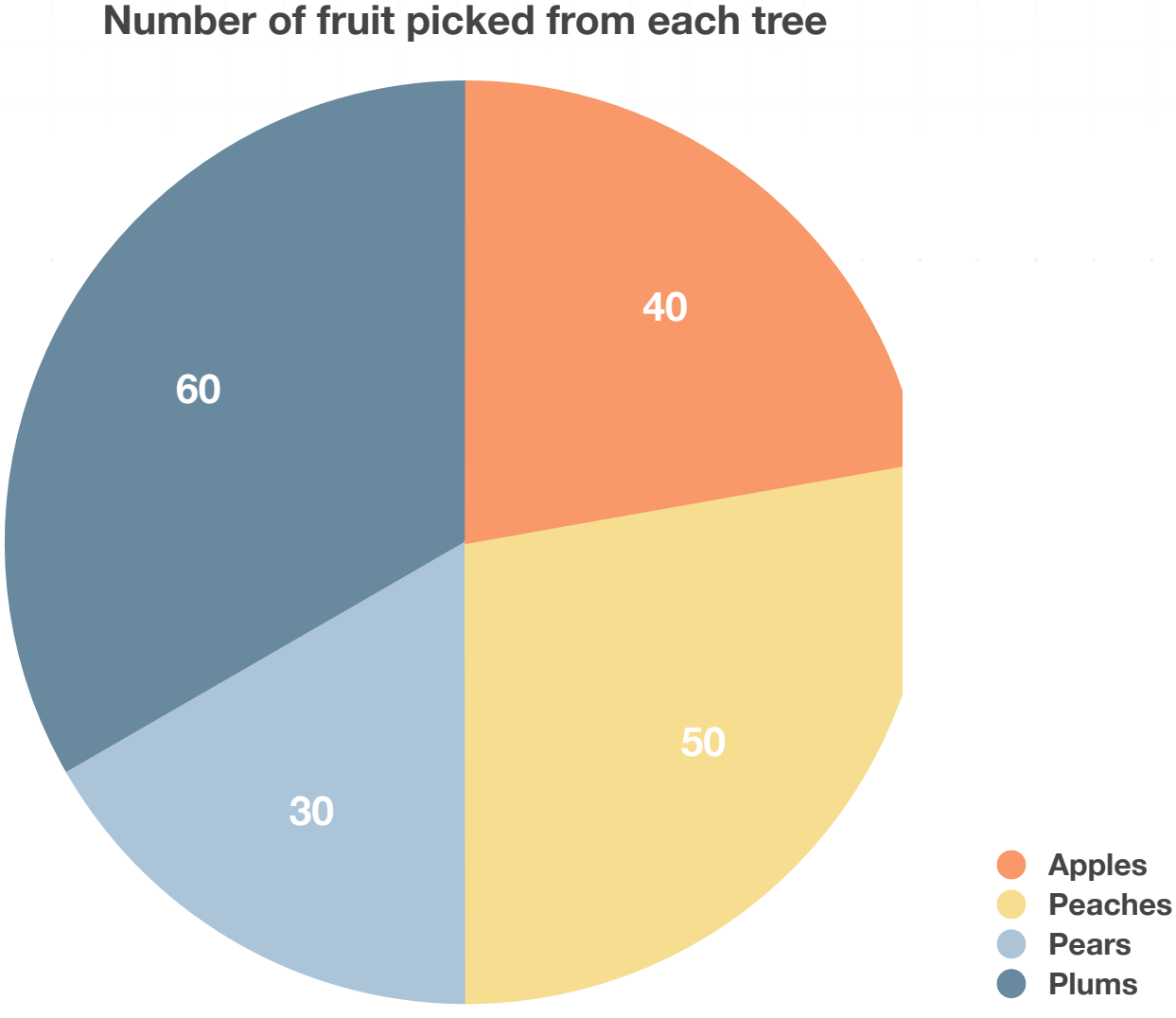
This is an experimental probability because it's based on the results of actual trials. From the table, we can see that we rolled an 11 on the dice 12 times out of the 60 total rolls.

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?



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 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
Total	**57**

Solution:

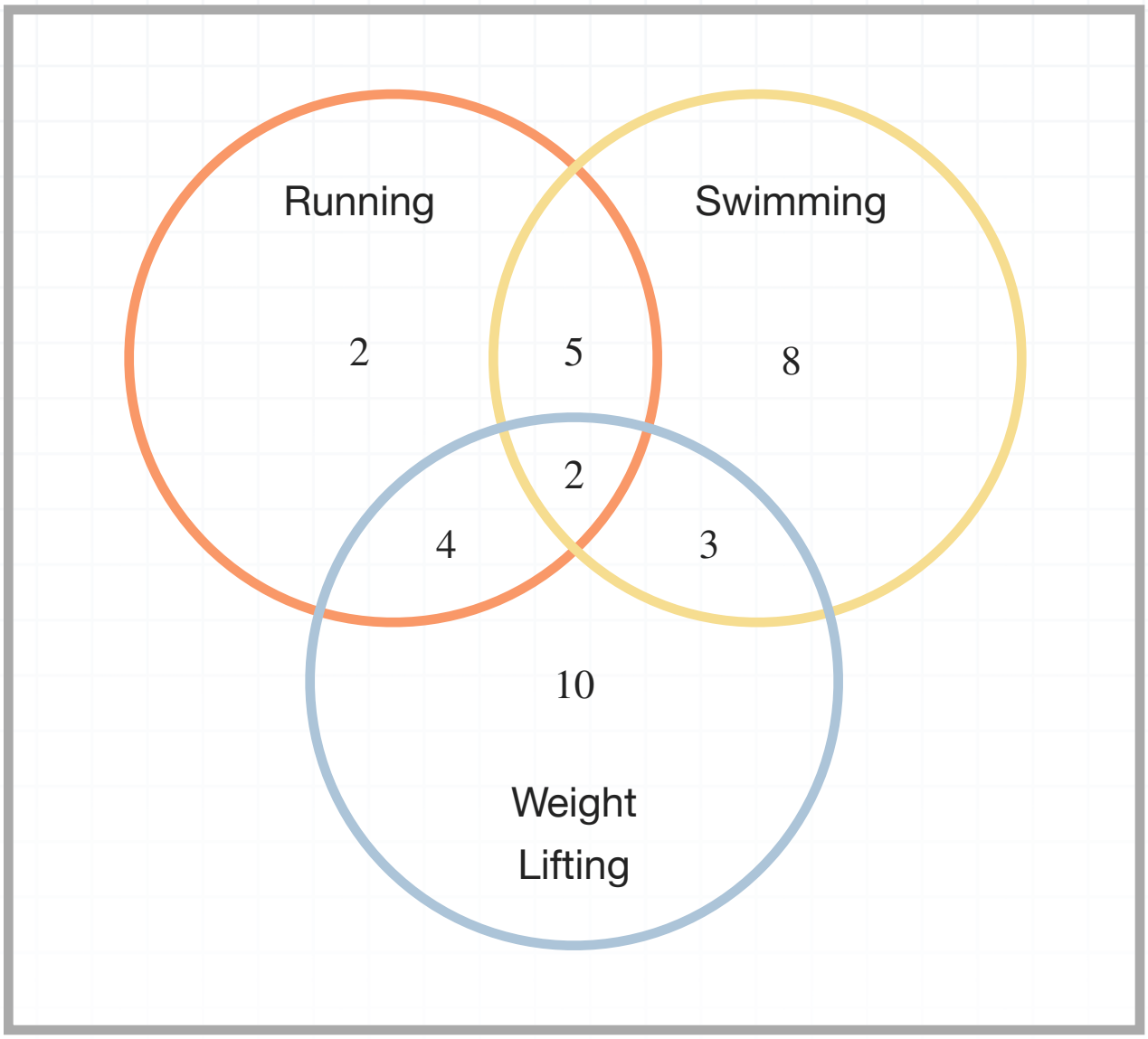
Jamal is not likely to have surveyed everyone who visits the park or everyone who will visit the park in the future. A survey is most often a sample of a larger population, so the results are an experimental probability.

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

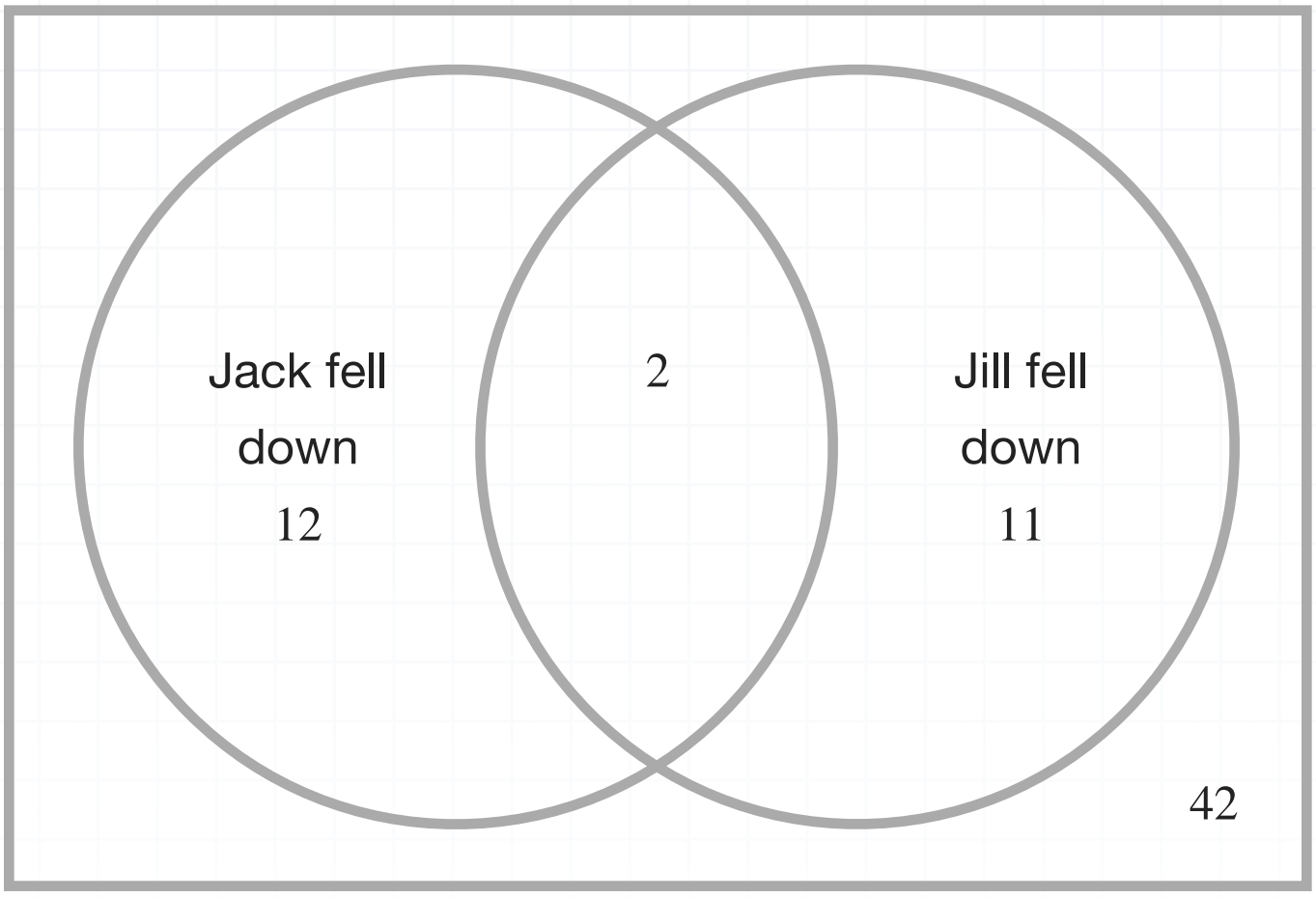
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

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. The Venn diagram shows the number of times Jack and Jill fell when going up the hill. What is the probability that Jack fell down and Jill fell down? What is the probability that Jack fell down or Jill fell down?



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 at least one of them fell down on $12 + 2 + 11 = 25$ trips up the hill. So the probability that either Jack fell down or Jill fell down is

$$P(\text{Jack fell down } \cup \text{ Jill fell down}) = \frac{25}{67}$$

■ 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.

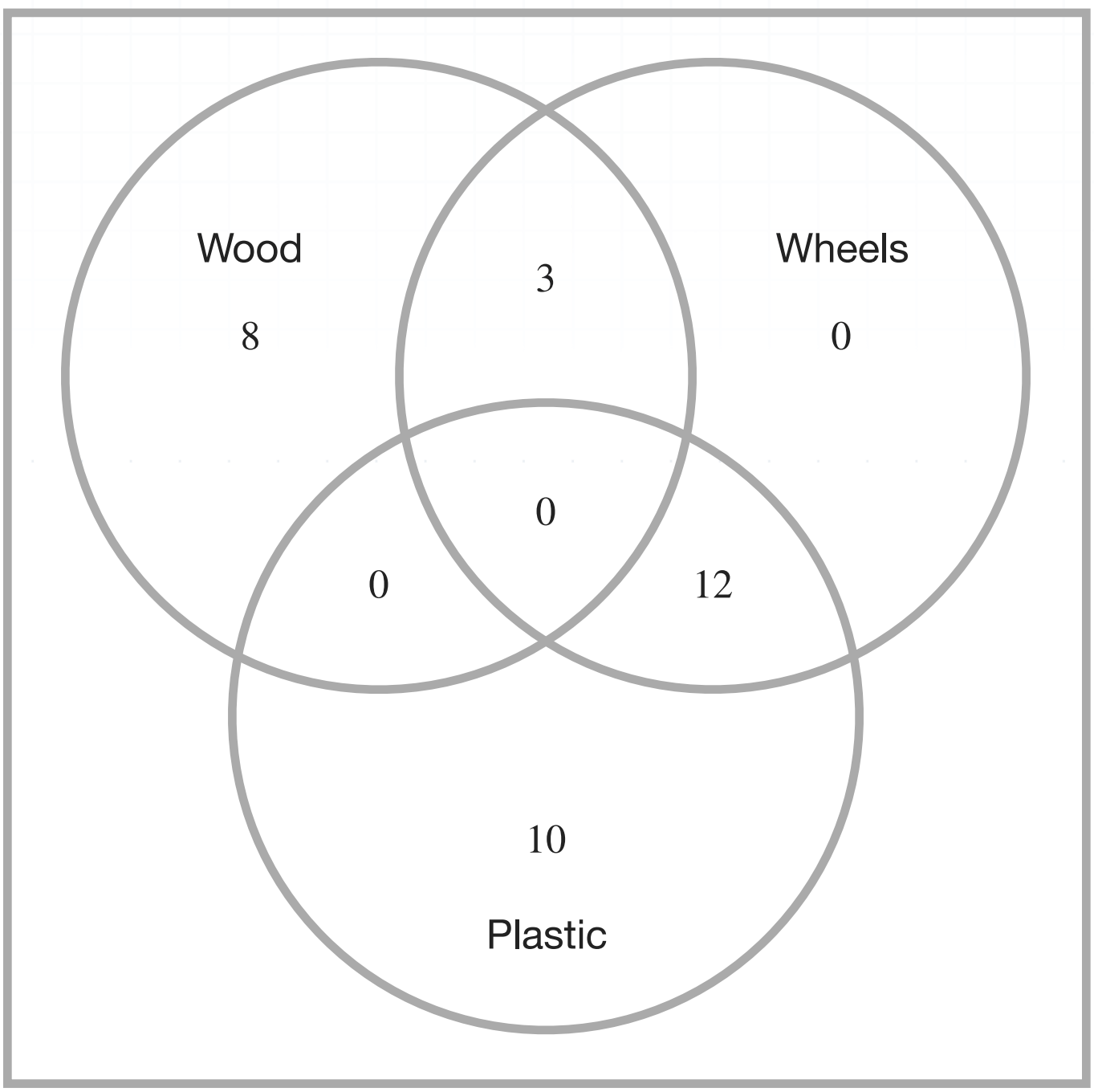
The probability that the purchase of a randomly selected fish is either mostly red or mostly orange is

$$P(\text{mostly red } \cup \text{ mostly orange}) = P(\text{mostly red}) + P(\text{mostly orange})$$

$$P(\text{mostly red} \cup \text{mostly orange}) = P(0.31) + P(0.23)$$

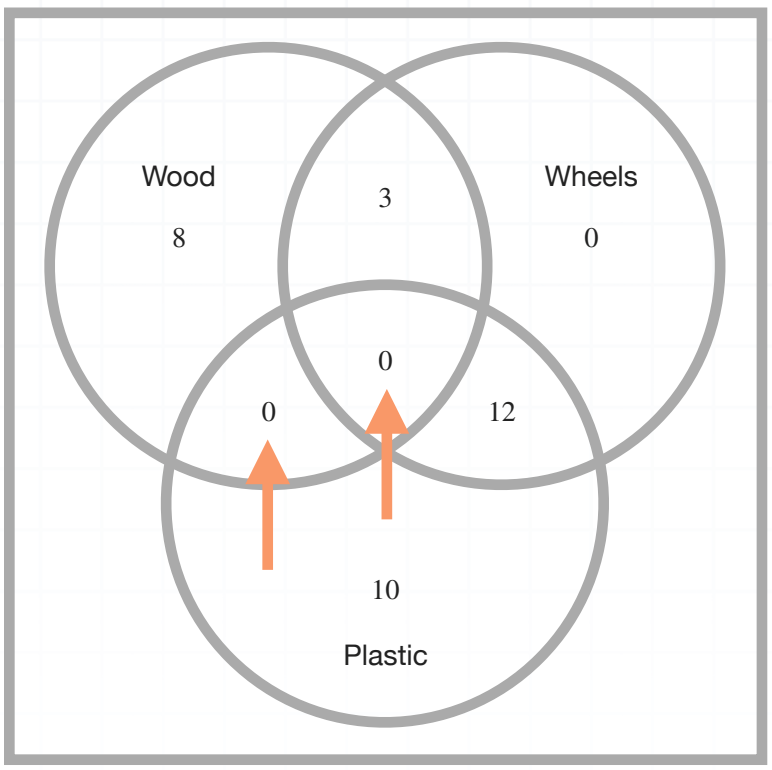
$$P(\text{mostly red} \cup \text{mostly orange}) = 0.54$$

■ 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:

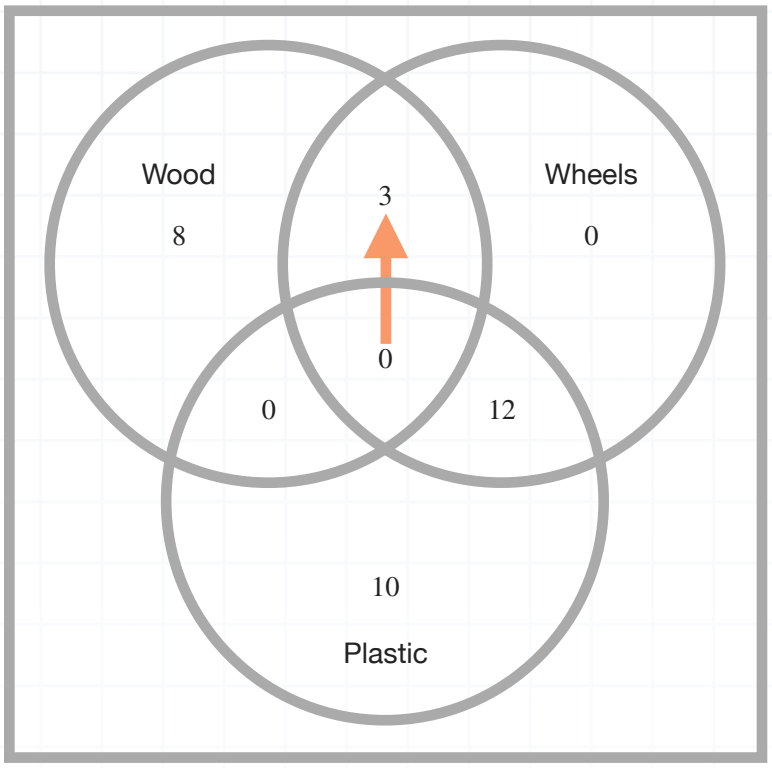
The events “plastic” and “wood” are mutually exclusive, because the intersection between them is 0.



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 taking architecture is 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**	**230**

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 enrolled in architecture is a junior 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**	**230**

The probability that a student is a junior or is taking architecture is given by

$$P(\text{junior} \cup \text{architecture}) = P(\text{junior}) + P(\text{architecture})$$

$$- P(\text{junior} \cap \text{architecture})$$

$$P(\text{junior} \cup \text{architecture}) = \frac{155}{520} + \frac{142}{520} - \frac{45}{520}$$

$$P(\text{junior} \cup \text{architecture}) = \frac{252}{520}$$

$$P(\text{junior} \cup \text{architecture}) = \frac{63}{120}$$

		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**	**230**

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 you 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 $P(\text{heads} \cup 2 \text{ or } 3)$. 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

$$P(\text{heads} \cup 2 \text{ or } 3) = \frac{2}{12} = \frac{1}{6}$$

INDEPENDENT AND DEPENDENT EVENTS AND CONDITIONAL PROBABILITY

■ 1. What is the probability of getting four heads in a row when you 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,

$$P(FFFFFFFFFF) = (0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)(0.6)$$

$$P(FFFFFFFFFF) = (0.6)^{10}$$

$$P(FFFFFFFFFF) \approx 0.006$$

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

$$P(\text{pink opal}) = \frac{2}{19}$$

We can therefore say that the probability of pulling both stones in that specific order (these are dependent events) is

$$P(\text{amethyst} \mid \text{pink opal}) = \frac{1}{4} \cdot \frac{2}{19} = \frac{2}{76} = \frac{1}{38}$$

■ 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 your 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		
		Cube	Rectangular Prism	Total
Block Color	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		
		Cube	Rectangular Prism	Total
Block Color	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 is 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

$$P(\text{cinnamon}) = \frac{12}{21} = \frac{4}{7}$$

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

$$P(A \text{ and } B) = P(A) \cdot P(B\,|\,A)$$

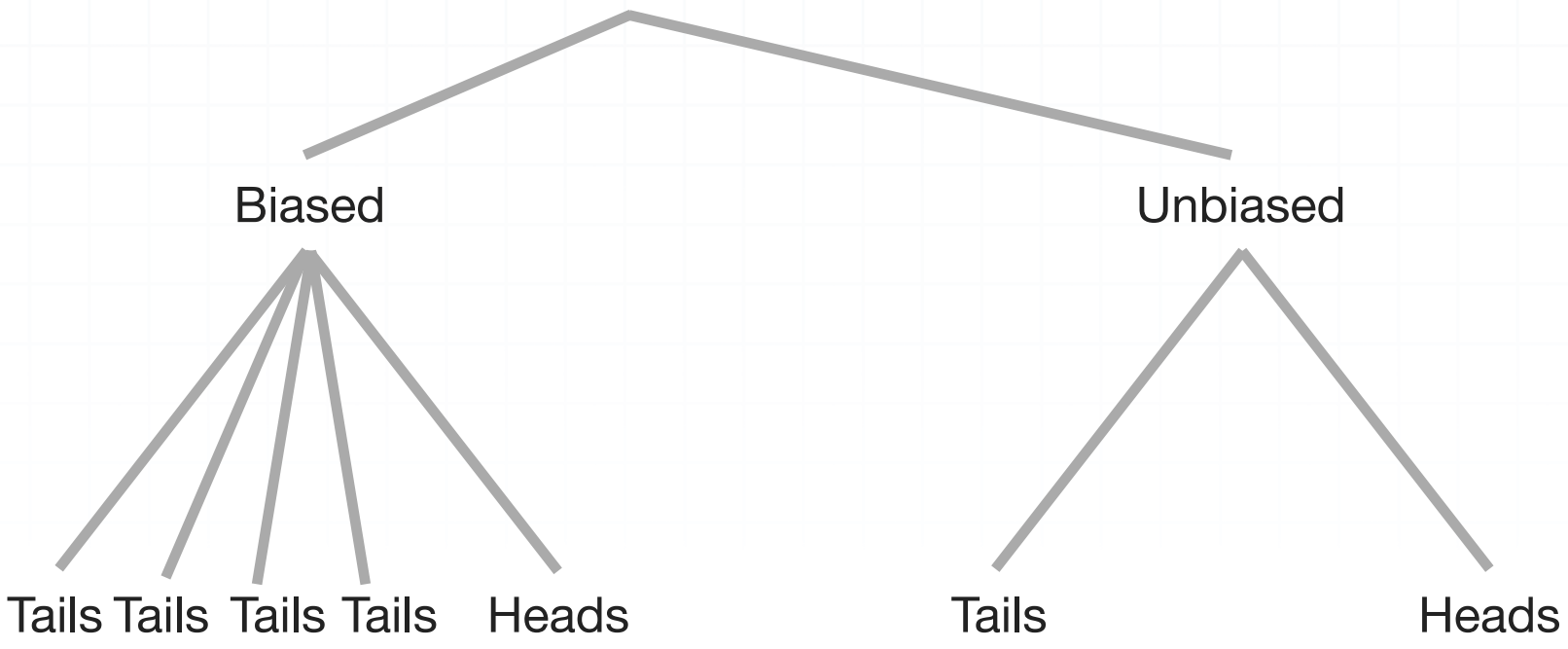
$$P(A \text{ and } B) = \frac{7}{12} \cdot \frac{4}{7}$$

$$P(A \text{ and } B) = \frac{28}{84}$$

$$P(A \text{ and } B) = \frac{1}{3}$$

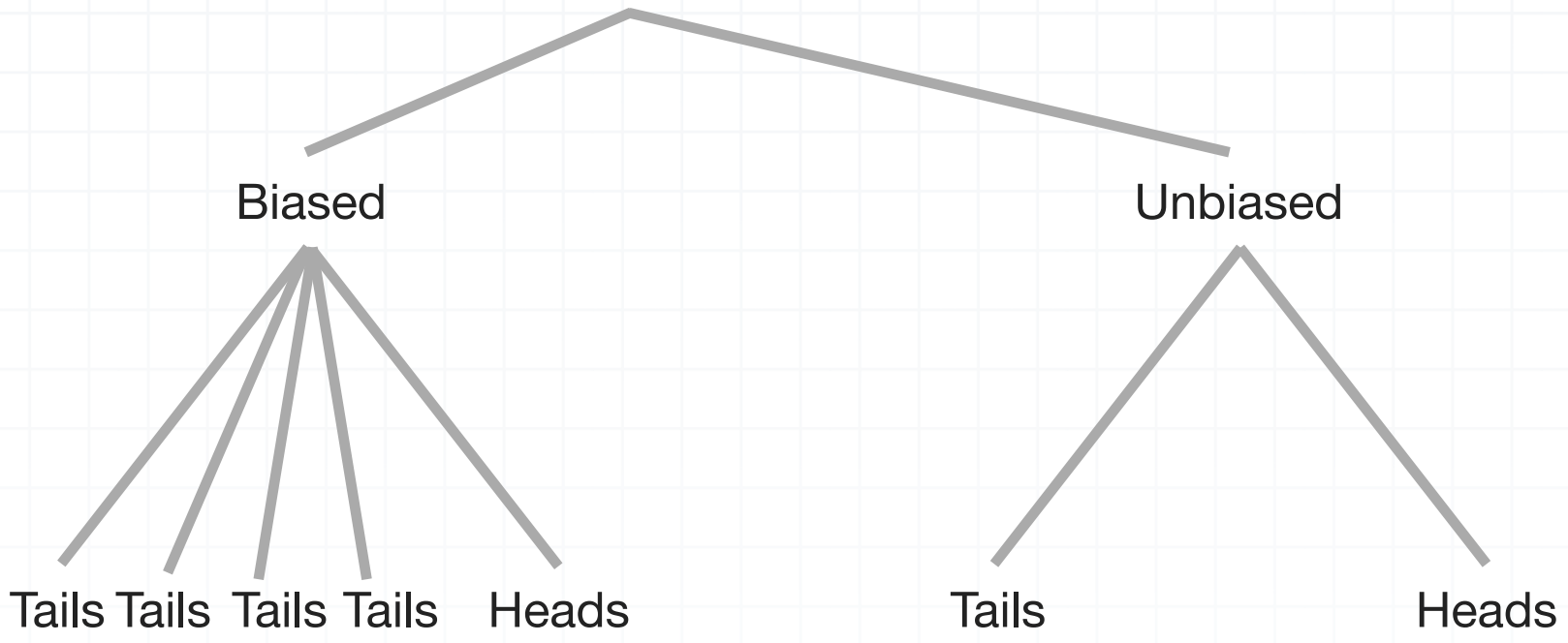
BAYES' THEOREM

■ 1. You have two coins. One is fair and the other one is weighted to land on tails $4/5$ of the time. Without knowing which coin you're choosing, you pick one at random, toss the coin and get tails. What is the probability you flipped the biased coin? Complete the tree diagram to answer the question.

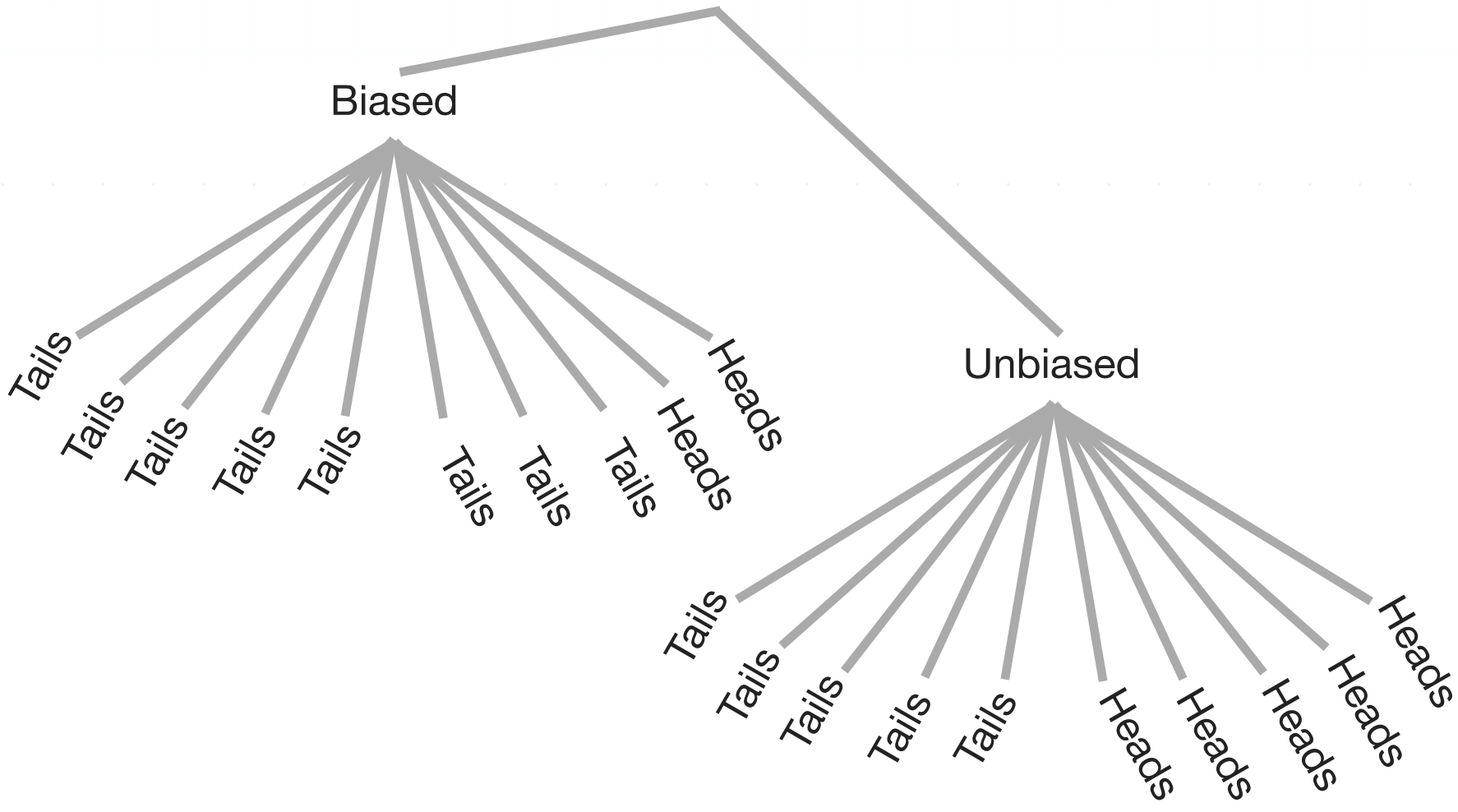


Solution:

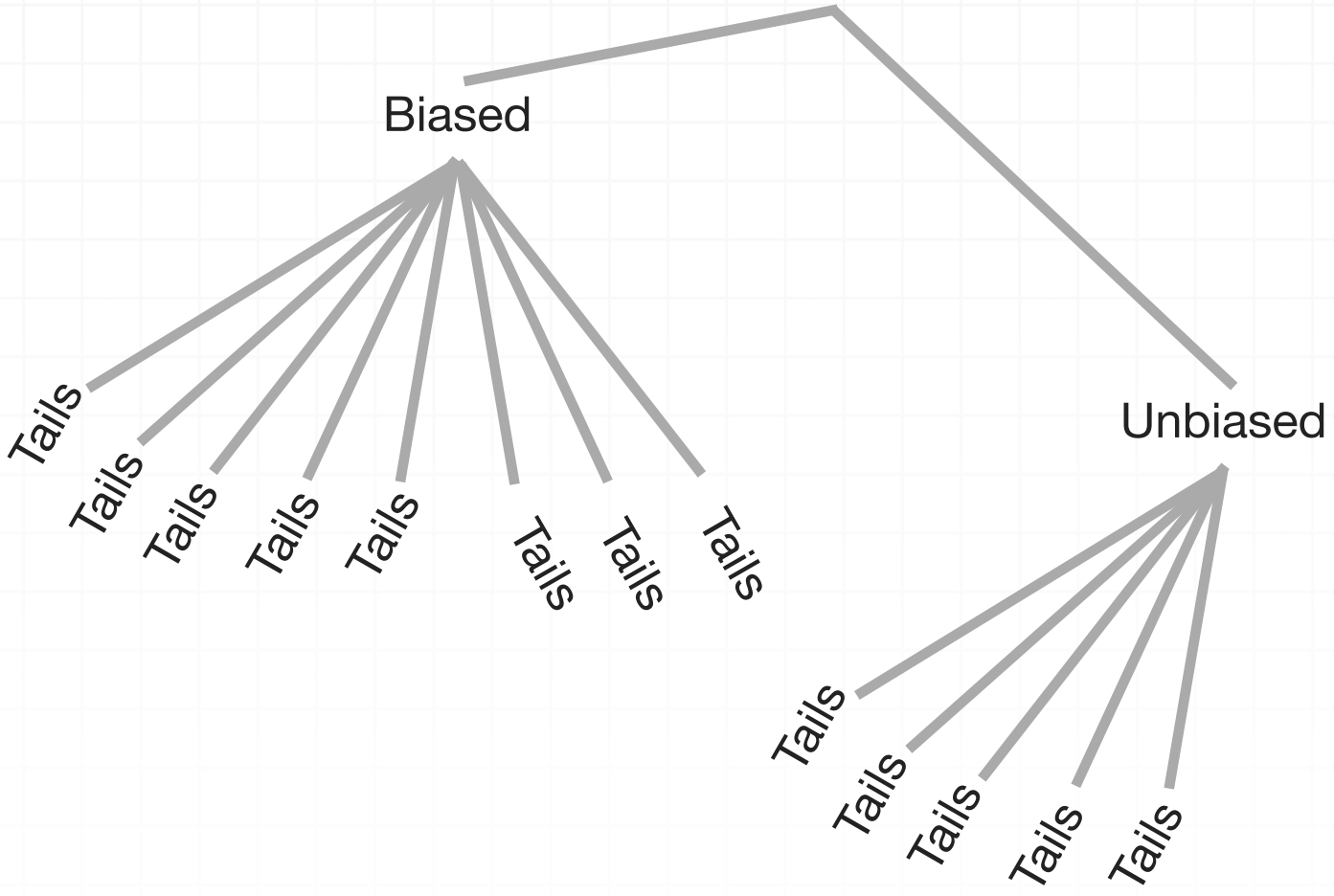
We're looking for the probability that the coin is biased given that we already flipped a tails, so we're looking for $P(\text{biased} \mid \text{tails})$.



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 you tossed the biased coin, knowing that it landed on tails, is $8/13$.

■ 2. You 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 you're choosing, you 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 you rolled the fair die?

$$P(6 \mid \text{fair})$$

$$P(\text{fair})$$

$$P(6)$$

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, you 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 you rolled the fair die?" We're looking for $P(\text{fair} \mid 6)$, and we have everything we need to use Bayes' Theorem.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P(\text{fair} \mid 6) = \frac{P(6 \mid \text{fair}) \cdot P(\text{fair})}{P(6)}$$

$$P(\text{fair} \mid 6) = \frac{\frac{1}{6} \cdot \frac{1}{2}}{\frac{7}{72}}$$

$$P(\text{fair} \mid 6) = \frac{\frac{1}{12}}{\frac{7}{72}} = \frac{1}{12} \cdot \frac{72}{7} = \frac{72}{84} = \frac{6}{7}$$

The probability you rolled the fair die given that you 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. You have two coins. One is fair and the other is weighted to land on tails $3/4$ of the time. Without knowing which coin you're choosing, you pick one at random, toss the coin, and get tails. What's the probability you 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

$$P(A\,|\,B) = \frac{P(B\,|\,A) \cdot P(A)}{P(B)}$$

$$P(A\,|\,B) = \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{5}{8}} = \frac{\frac{3}{8}}{\frac{5}{8}} = \frac{3}{8} \cdot \frac{8}{5} = \frac{3}{5}$$

The probability that you 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 tests positive due to an inaccurate result (a false 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 $P(D\,|\,P)$.

$$P(D\,|\,P) = \frac{(0.90)(0.10)}{0.09 + 0.135}$$

$$P(D\,|\,P) = \frac{0.09}{0.235}$$

$$P(D\,|\,P) \approx 38\,\%$$

This means the probability that an employee who is a drug user will test positive is only about $38\,\%$.

■ 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(DIA)
B	0.45	0.014 P(DIB)

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(DIA)

$$P(A \cap D) = P(D|A) \cdot P(A)$$

$$P(A \cap D) = (0.55)(0.020)$$

$$P(A \cap D) = 0.011$$

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(DIB)

$$P(B \cap D) = P(D|B) \cdot P(B)$$

$$P(B \cap D) = (0.45)(0.014)$$

$$P(B \cap D) = 0.0063$$

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

$$P(B \mid D) = \frac{P(D \mid B) \cdot P(B)}{P(D)}$$

$$P(B \mid D) = \frac{(0.014) \cdot (0.45)}{0.0173}$$

$$P(B \mid D) = \frac{0.0063}{0.0173}$$

$$P(B \mid D) \approx 36\,\%$$

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 $P(X \geq 3)$.

X	1	2	3	4	5
P(X)	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 μ_B and σ_B.

B	0	5	10	15
P(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 μ_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, σ_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:

First, find the probability that the household is a 3-person household.

$$P(X = 3) = 1 - 0.281 - 0.340 - 0.129 - 0.060 - 0.023 - 0.013$$

$$P(X = 3) = 0.154$$

Now, with a complete table, we can find the probability that the household size is at least 4, given that the household size is at least 2.

$$P(\text{at least } 4, \text{ given at least } 2) = \frac{P(\text{size at least } 4)}{P(\text{size at least } 2)}$$

$$P(\text{at least } 4, \text{ given at least } 2) = \frac{0.225}{0.719} \approx 0.313$$

■ 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 $P(R = 0)$.

$$P(R = 0) = \frac{26}{52}\left(\frac{26}{51}\right) = \frac{26}{102}$$

$$P(R = 2) = \frac{26}{52}\left(\frac{25}{51}\right) = \frac{25}{102}$$

Then the probability that one card is red is

$$P(R = 1) = 1 - P(R = 2) - P(R = 0)$$

$$P(R = 1) = 1 - \frac{25}{102} - \frac{26}{102}$$

$$P(R = 1) = \frac{102}{102} - \frac{25}{102} - \frac{26}{102}$$

$$P(R = 1) = \frac{51}{102} = \frac{1}{2}$$

Which means we can build a probability distribution for R.

R	0	1	2
P(R)	26/102	51/102	25/102

■ 5. A local restaurant features a wheel you can spin before you pay your bill. The wheel is split into 8 equal size pieces. One of the sections gives customers a $\$10$ discount on their 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

$$E(X) = \sum XP(X) = 10\left(\frac{1}{8}\right) + 5\left(\frac{2}{8}\right) + 2\left(\frac{3}{8}\right) + 0\left(\frac{2}{8}\right)$$

$$E(X) = \$3.25$$

■ 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 ticket.

$$P(L = 1) = 0.2$$

$$P(L = 2) = (0.2)(0.8) = 0.16$$

$$P(L = 3) = (0.2)(0.8)(0.8) = 0.128$$

$$P(L = 4) = (0.2)(0.8)(0.8)(0.8) = 0.1024$$

$$P(L = 5) = (0.2)(0.8)(0.8)(0.8)(0.8) = 0.08192$$

Then the expected value for the number of tickets he'll buy, L, is

$$E(L) = 1(0.2) + 2(0.16) + 3(0.128) + 4(0.1024) + 5(0.4096)$$

$$E(L) \approx 3.36$$

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 σ_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 σ_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. The class who raises the most money is awarded a pizza party. The school secretary records the amount raised by each class and makes a five-number summary for the data.

Min	Q1	Median	Q3	Max
4.50	15.25	22.00	38.75	95.50

Suppose the school has 45 8th period classrooms with 20 students per classroom. What was the median amount donated per student? With what IQR?

Solution:

We can say that the amount donated per student is

$$\text{Amount per student} = \frac{\text{Amount per class}}{20}$$

Therefore, the median amount donated per student must be

$$\text{Median amount per student} = \frac{\text{Median amount per class}}{20} = \frac{22}{20} \approx 1.10$$

And the IQR of the amount per student will be

$$\text{IQR for amount per student} = \frac{\text{IQR for amount per class}}{20} = \frac{38.75 - 15.25}{20} \approx 1.18$$

So the median amount donated was $\$1.10$ with an IQR of $\$1.18$.

■ 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 $\sigma_S = 18.75$, you 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:

We'll find the expected value of the combination first.

$$E(4A + 2B) = 4E(A) + 2E(B) = 4(6.5) + 2(4.4) = 34.8$$

Then we'll find the standard deviation of the combination by finding the variances. The variances are $SD^2(A) = 1.6^2 = 2.56$ and $SD^2(B) = 2.1^2 = 4.41$. So the standard deviation of the combination is

$$SD(4A + 2B) = \sqrt{4SD^2(A) + 2SD^2(B)} = \sqrt{4(2.56) + 2(4.41)} \approx 7.655$$

■ 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

$$\mu_{Q_1} + \mu_{Q_2} + \mu_{Q_3} + \ldots + \mu_{Q_{40}} = 40(55) = 2{,}200 \approx 36.67 \text{ minutes}$$

and that the variance of the finishing time is

$$\sigma^2_{Q_1} + \sigma^2_{Q_2} + \sigma^2_{Q_3} + \ldots + \sigma^2_{Q_{40}} = 40(12^2) = 40(144) = 5{,}760 \text{ seconds}$$

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 $\sigma_W^2 = 5.76$. 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. You buy a scratch off lottery ticker for $\$1$ at the local gas station. If you get three hearts in a row on your scratch off, the state will pay you $\$500$. Let X be the amount the state pays you and let X have the following probability distribution.

X	$0	$500
P(X)	0.999	0.001

Suppose you buy one of these scratch off tickets every day for a week (7 days). Find the expected value and standard deviation of your total winnings.

Solution:

The expected value of your winnings on any one ticket is

$$E(X) = 0(0.999) + 500(0.001) = \$0.50$$

Find the standard deviation of your 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 you 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

$$SD(W) = \sqrt{(15.80)^2 + (15.80)^2 + \ldots + (15.80)^2}$$

$$SD(W) = \sqrt{7(15.80)^2}$$

$$SD(W) = \sqrt{1{,}747.48}$$

$$SD(W) \approx \$41.80$$

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

$$ {}_nP_k = \frac{n!}{(n-k)!} $$

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 ${}_5P_2$ than ${}_5C_2$?

Solution:

We'll calculate both values, then find the difference.

$$ {}_5P_2 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 5 \cdot 4 = 20 $$

$$ {}_5C_2 = \frac{5!}{2!(5-2)!} = \frac{5!}{2! \cdot 3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} = \frac{5 \cdot 4}{2} = 10 $$

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.

$$ {}_nP_k = \frac{n!}{(n-k)!} = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 5 \cdot 4 = 20$$

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.

$$ {}_nP_k = \frac{n!}{(n-k)!} = \frac{7!}{(7-7)!} = \frac{7!}{0!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} = 5{,}040 $$

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.

$$ {}_nC_k \cdot {}_nC_k = {}_{14}C_3 \cdot {}_{16}C_3 = 364 \cdot 560 = 203{,}840 $$

BINOMIAL RANDOM VARIABLES

■ 1. You 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. You randomly select students from your school until you 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 $p = 0.08$.

$$P(X = 4) = \binom{12}{4}(0.08)^4(1 - 0.08)^8$$

$$P(X = 4) = (495)(0.08)^4(1 - 0.08)^8$$

$$P(X = 4) = 0.0104$$

■ 4. Let Y be the number of times you roll a 1 on a fair 6-sided die if you do 10 trials. Fill in the following probability distribution for Y, rounding each probability to 4 decimal places.

Y	0	1	2	3	4	5	6	7	8	9	10
P(Y)											

Solution:

With $n = 10$, $p = 1/6$, and $k = 0, 1, 2, 3, \ldots 10$, 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

Y	0	1	2	3	4	5	6	7	8	9	10
P(Y)	0.1615	0.3230	0.2907	0.1550	0.0543	0.1300	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.

$$P(X = 3) = \binom{8}{3}(0.3)^3(1 - 0.3)^5$$

$$P(X = 3) = (56)(0.3)^3(1 - 0.3)^5$$

$$P(X = 3) = 0.2541$$

"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

$$P(X > 7) = P(X = 8) + P(X = 9) + \ldots + P(X = 15)$$

which is the same as

$$1 - P(X \leq 7)$$

$$1 - 0.6535$$

$$0.3465$$

■ 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

$$P(X < 8) = P(X \leq 7) = P(X = 0) + P(X = 1) + \ldots + P(X = 7)$$

$$P(X < 8) = 0.3308$$

■ 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

$$\sigma_X^2 = Var(X) = (200)(0.68)(1 - 0.68) = 43.53$$

which means the standard deviation is

$$\sigma_X = SD(X) = \sqrt{43.53} \approx 6.597 \text{ households}$$

■ 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. You roll a fair die 6 times. What is the probability you'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 $X \sim B(6,0.5)$.

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$P(X \leq 3) = 0.6563$$

■ 6. You 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 you roll a sum of 7. Since there are 36 possible rolls when you roll two die, and 6 of them result in a sum of 7, 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 six 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

$$1 - P(X \leq 5)$$

$$1 - 0.9976$$

$$0.0024$$

or about a $.24\,\%$ 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." You buy 5 spins for $\$3.00$. If you land on "winner" on any of your 5 spins, you 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." You buy 5 spins for $\$3.00$. If you land on "winner" on any of your 5 spins, you 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." You buy 5 spins for $\$3.00$. If you land on "winner" on any of your 5 spins, you 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."

You buy 5 spins for $\$3.00$. If you land on "winner" on any of your 5 spins, you 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

$$P(\text{winner}) = p = \frac{1}{4} = 0.25$$

$$P(\text{non-winner}) = 1 - p = 1 - 0.25 = 0.75$$

Therefore, the probability of no winners in 5 spins is

$$P(\text{no winners in } 5 \text{ spins}) = (0.75)^5 = 0.2373$$

■ 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." You buy 5 spins for $\$3.00$. If you land on "winner" on any of your 5 spins, you 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. Your goal is to learn about the percentage of students with high ACT scores. You 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, you could define a specific range of ACT scores as "failures" and another range as

"successes." For instance, you 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 you now have an experiment in which you're using Bernoulli trials.

GEOMETRIC RANDOM VARIABLES

■ 1. You toss a coin until you 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. You randomly select students from your school until you 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. You 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 you have a large population. You're selecting students until you 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. Is the data one-way or two-way data? 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 is one-way data. 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 activities for 50 adults. Is the data one-way data? 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. Is the data one-way or two-way data? Is the study observational or experimental?

Solution:

The data is two-way data 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. 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 a simple random sample. Patrons are randomly selected with no regard to groups or clusters.

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 satisfy normality?

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 of normality. 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. You want to verify this claim, so you randomly select 120 customers to see if they order a sugar cone. Is this a normal sampling?

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 than 10 expected successes and failures.

$$120(0.05) = 6 \ngeq 10$$

$$120(0.95) = 114 \geq 10$$

The sample space doesn't meet the conditions of normality 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. Every 10th customer or group is recorded as a family, and defined as a group with children under 12 or not. They find that 45 families are recorded and only 20 are not part of a family with children under 12. That day there were 650 visitors or groups. What is the standard deviation for the sample?

Solution:

To verify normality, 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:

$$\frac{65}{650} = 0.1 = 10\,\% \leq 10\,\%$$

And 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 of normality, so we'll identify the sample size $n = 65$ and the population proportion as

$$p = \frac{45}{65} = 0.69$$

Now we can calculate standard deviation.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.69(1-0.69)}{65}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.69(0.31)}{65}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2139}{65}}$$

$$\sigma_{\hat{p}} = 0.057365$$

■ 4. A pizza shop finds that $80\,\%$ of the 75 randomly selected pizzas ordered during the week have pepperoni. What is the standard deviation for the sample if the pizza shop has a total of $1{,}000$ pizzas ordered during the week?

Solution:

To verify normality, 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 of normality, so we'll identify the sample size $n = 75$ and the population proportion as $p = 0.8$. Now we can calculate standard deviation.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.8(1-0.8)}{75}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.8(0.2)}{75}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.16}{75}}$$

$$\sigma_{\hat{p}} = 0.046188$$

■ 5. A hospital conducts a survey on a particular day and finds that 10 patients of 30 randomly selected have high blood pressure. There were 325 patients in the hospital that day. What is the standard deviation for the sample?

Solution:

To verify normality, 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:

$$\frac{30}{325} = 0.0923 = 9.23\,\% \leq 10\,\%$$

And there are more than 10 expected successes and failures.

$$30(0.33) = 10 \geq 10$$

$$30(0.67) = 20 \geq 10$$

We've met the conditions of normality, so we'll identify the sample size $n = 30$ and the population proportion as

$$p = \frac{10}{30} = 0.33$$

Now we can calculate standard deviation.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.33(1-0.33)}{30}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.33(0.67)}{30}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2211}{30}}$$

$$\sigma_{\hat{p}} = 0.085849$$

■ 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 first-born children. What is the probability that their results are within $2\,\%$ of the first study's claim?

Solution:

To verify normality, 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 of normality. The original study found the population proportion to be $p = 72\,\%$. So the standard deviation of our sample will be

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.72(0.28)}{175}} \approx 0.0339$$

We need to find the probability that our results are within $10\,\%$ 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:

$$\frac{0.02}{0.0339} \approx 0.59$$

This means we want to know the probability of $P(-0.59 < z < 0.59)$. Using a z-table, -0.59 gives us 0.2776 and 0.59 gives us 0.7244, so the probability is

$$P(-0.59 < z < 0.59) = 0.7244 - 0.2776$$

$$P(-0.59 < z < 0.59) = 0.4468$$

There's a $44.68\,\%$ chance that our sample proportion will fall within $2\,\%$ of the first study's claim.

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?

Solution:

Find the standard deviation of the sampling distribution using $\sigma = 3$ and $n = 9$.

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

$$\sigma_{\bar{x}} = \frac{3}{\sqrt{9}}$$

$$\sigma_{\bar{x}} = 1$$

■ 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.

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

$$\sigma_{\bar{x}} = \frac{5}{\sqrt{80}}$$

$$\sigma_{\bar{x}} = 0.559$$

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.

$$\frac{29 - 30}{0.559} = \frac{-1}{0.559} = -1.79$$

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.

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

$$\sigma_{\bar{x}} = \frac{0.5}{\sqrt{100}}$$

$$\sigma_{\bar{x}} = 0.05$$

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.

$$\frac{12.5 - 12.6}{0.05} = \frac{0.1}{0.05} = 2$$

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.

$$P(z > 2) = 1 - 0.9772$$

$$P(z > 2) = 0.228$$

$$P(z > 2) = 22.8\,\%$$

There's a $22.8\,\%$ chance that our sample mean will be greater than 12.6 ounces.

■ 5. A hospital finds that the average body temperature of their patients is $98.4°$, with a standard deviation of $0.6°$, 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 $0.2°$ 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 the 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.

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

$$\sigma_{\bar{x}} = \frac{0.6}{\sqrt{30}}$$

$$\sigma_{\bar{x}} = 0.1095$$

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.

$$\frac{0.2}{0.1095} = 1.83$$

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 an $93.28\,\%$ chance that our sample mean will fall within $0.2°$ of the population mean of $98.4°$.

■ 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.

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

$$\sigma_{\bar{x}} = \frac{0.2}{\sqrt{50}}$$

$$\sigma_{\bar{x}} = 0.02828$$

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.

$$\frac{0.05}{0.02828} = 1.77$$

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.

ALTERNATIVE AND NULL 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. State if the test is a one- or two-tailed test.

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 μ.

We're looking at a one-sided test because we're only interested if whether the new option preforms better than the current one (not better and worse). Since we're looking at how much better it will perform, we use the greater than symbol in our hypothesis statement.

$$H_0 : p = 0.85$$

$$H_a : p > 0.85$$

■ 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. They set up their hypothesis statements as

$$H_0 : \mu = 3.5$$

$$H_a : \mu < 3.5$$

What are the Type I and Type II errors in this study? What could be the consequences of each type of error in a published report of the study?

Solution:

A Type I error is rejecting the null hypothesis when the null hypothesis is true. In this case the alternative hypothesis is that it takes a smoker less than 3.5 attempts to quit smoking in order to be successful, which means a Type I error would be claiming that the average number of attempts to quit before a smoker is successful is less than 3.5 attempts, when in actuality it would takes 3.5 attempts or more before a smoker is successful. If the study is published with Type I error included, then smokers who need more than 3.5 attempts to quit may become discouraged and give up after reading the article, even though they might still be on the road to success.

A Type II error is failing to reject the null hypothesis when the null hypothesis is false. Failing to reject the null hypothesis would bring the researcher to the false conclusion that there was not enough evidence to prove that it takes smokers less than 3.5 attempts to quit. If the study is published with Type II error included, smokers who read the article may

take more attempts to quit than necessary because they believe they're still in the boundaries of what is normal.

■ 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 parts wires 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.

Decide if the hypothesis statement would require a population proportion or a population mean, then set up the statistical hypothesis statements for the situation. State if the test is a one- or two-tailed test.

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, so the factory needs to perform a two-tailed test.

$$H_0 : \mu = \frac{1}{16}$$

$$H_a : \mu \neq \frac{1}{16}$$

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. The clothing company used its sample data to test the hypothesis statements.

$$H_0 : \mu = 17$$

$$H_a : \mu > 17$$

Did the clothing company reject or fail to reject the null hypothesis? If their conclusion is incorrect, what type of error are they making (Type I or Type II) and what are the consequences of making that error?

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 $17\,\%$. If we look at the hypothesis statements,

$$H_0 : \mu = 17$$

$$H_a : \mu > 17$$

this claim means they have rejected the null hypothesis, because the null hypothesis is that the average sales are not greater than $17\,\%$.

If they have made an error, they have rejected the null hypothesis when it's actually true. That means the error here would by Type I error. The consequences of this error might be that the company produces many clothes in lime green and polka dots that would fail to produce the expected increase in sales. The company would lose money and waste resources due to this type of error.

■ 5. 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 parts wires 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.

In the context of this situation, describe the Type I and Type II errors and the consequences of each. Based on the consequences, should you choose an α-level of 0.10 or 0.01?

Solution:

This was a two-tailed test with the following hypothesis statements:

$$H_0 : \mu = \frac{1}{16}$$ the machine doesn't need recalibration

$$H_a : \mu \neq \frac{1}{16}$$ the machine needs recalibration

Remember that a Type I error is rejecting the null hypothesis when it's true. Here the alternative hypothesis would lead the factory to believe that the machine needs recalibration when it actually doesn't. The consequences of a Type I error would lead to recalibrating a machine that didn't need it, which is a minor consequence.

On the other hand, remember that Type II error means you fail to reject the null hypothesis even though there's enough evidence to do so. Failing to reject the null hypothesis would lead the factory to believe that the lead wires for the pace makers were being produced correctly, when in fact, they weren't within the proper specifications. An error of this type could lead to faulty parts and engine problems for the cars that receive the parts. This means that a Type II error is more serious than a Type I error.

In this case the Type I error is not the concern so we want to use the higher significance level. That means we could choose a significance level of $\alpha = 0.10$. If we had chosen to use the smaller significance level, that would increase the Type II error which we surely want to avoid in this case.

■ 6. A new heartworm test is being developed to help improve the accuracy of detecting heartworms in dogs. The heartworm test was given to 60 dogs known to have heartworms, and it correctly identified 59 as positive for heartworms.

The test was also given to 64 dogs already known to be heartworm-free. Out of the 64 heartworm-free dogs, the test correctly identified 58 as heartworm-free.

The hypothesis statements for the study were:

H_0 : the dog has heartworms

H_a : the dog doesn't have heartworms

Define the Type I and Type II errors. Estimate the probabilities of each error, α and β, based on the study.

Solution:

Remember that a Type I error is rejecting the null hypothesis when it's actually true. Here the alternative hypothesis is that the dog does not have heartworms. This means a Type I error would misdiagnose a sick dog as healthy.

According to the study, only 1 of the 60 sick dogs was diagnosed incorrectly, since 59 of them tested positive for heartworms.

$$\alpha = \frac{1}{60} = 0.0167 \approx 1.7\,\%$$

On the other hand, remember that Type II error means you fail to reject the null hypothesis even though there's enough evidence to do so. Failing to reject the null hypothesis would lead to claiming a healthy dog was sick.

In the study $64 - 58 = 6$ dogs were healthy but diagnosed as sick. So that is 6 misdiagnosed dogs out of the 64 healthy dogs.

$$\beta = \frac{6}{64} = 0.09375 \approx 9.3\,\%$$

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 high schoolers had an average test score of 129 points with a sample standard deviation of 26.8. Choose and calculate the appropriate test statistic for the data. Choose and calculate the appropriate test statistic for the data.

Solution:

The sample is comparing average scores, and we're given the sample standard deviation. The sample size is large enough at 256 high schoolers that we can assume the distribution is approximately normal.

This means the population characteristic is a population mean with an unknown standard deviation (since we have the sample standard deviation and not the population standard deviation). In the case of a population mean with an unknown standard deviation, we use a t-test statistic with the following formula:

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

Let's state the values and calculate the test statistic. The sample mean is the average test score of the local high schoolers, $\bar{x} = 129$. The

hypothesized value is the test score of the state. Notice this is the value that would appear in the hypothesis statements if we wrote them.

$$H_0 : \mu = 106$$

$$H_a : \mu > 106$$

The hypothesized value is 106 and the sample standard deviation is given as $s = 26.8$. The sample size is the number of students from the high school, $n = 256$. Now we can calculate our t-test statistic:

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{129 - 106}{\frac{26.8}{\sqrt{256}}} \approx 13.73$$

■ 2. A dietician 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.

Solution:

The sample is comparing the average number of calories, and we're given the sample standard deviation. The sample size is large enough at 35 meals that we can assume the distribution is approximately normal. This means the population characteristic is a population mean with an unknown

standard deviation (since we have the sample standard deviation and not the population standard deviation). In the case of a population mean with an unknown standard deviation we use a t-test statistic with the following formula:

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

Let's state the values and calculate the test statistic. The sample mean is the average calories in the sample of restaurant meals, $\bar{x} = 1{,}250$. The hypothesized value is the average number of calories in a meal at the national level. Notice this is the value that would appear in the hypothesis statements if we wrote them.

$$H_0 : \mu = 1{,}500$$

$$H_a : \mu < 1{,}500$$

The hypothesized value is $1{,}500$ and the sample standard deviation is given as $s = 350.2$. The sample size is the number of students from the high school, $n = 35$. Now we can calculate our t-test statistic:

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{1{,}250 - 1{,}500}{\frac{350.2}{\sqrt{35}}} \approx -4.22$$

■ 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 for their dogs while they were away on vacation. The study would like to make a

conclusion that the majority of dog owners use a kennel run by their veterinary office when the owners go on vacation. Choose and calculate the appropriate test statistic for the data.

Solution:

The sample is comparing the proportion of dog owners who use a kennel run by their veterinary office to board their dogs while they are away on vacation. 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 n is 768 and the population proportion p is

$$\frac{567}{768} \approx 0.738$$

Therefore,

$$np = (768)(0.738) = 567 \geq 10$$

$$n(1-p) = (768)(1-0.738) = 201 \geq 10$$

This means we can use the hypothesis test for our population proportion. The test statistic will be the z-test statistic with the formula for a population proportion.

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

Let's state the values and calculate the test statistic. The sample proportion is the number of dog owners who board their dogs at the veterinary office, divided by the total number of dog owners in the survey.

$$\hat{p} = \frac{567}{768} \approx 0.738$$

Let's think about how to find the hypothesized value since it's not clearly stated in the problem. The wording in the problem states: "The study would like to make a conclusion that the majority of dog owners use a kennel run by their veterinary office when the owners go on vacation." A majority is anything above half, or in the case of a population proportion, $50\,\%$. This means the null hypothesis should state that the population proportion is equal to $50\,\%$. Then the alternative hypothesis that we're trying to prove is that the population proportion is greater than $50\,\%$. That would make it a majority.

$$H_0 : p = 0.5$$

$$H_0 : p > 0.5$$

The hypothesized value is 0.50, and the sample size is the number of dog owners in the survey, $n = 768$. Now we can calculate our z-test statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}} = \frac{0.7383 - 0.5}{\sqrt{\frac{(0.5)(1 - 0.5)}{768}}} \approx 13.21$$

4. A school board wants to support opening a new day care center. They look at local random sample of 500 households with children under preschool age. 243 of the households were using a family member to care for their children that were under preschool age. The school board wants to determine if less than half of the households are now using a family member to care for their children 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:

This is a population proportion because the study is interested in the proportion of households who use a family member to care for their children under preschool age. According to the problem: “The school board wants to determine if less than half of the households are now using a family member to care for their children at a statistically significant level.” This means we’re interested in looking at half of the households or $50\,\%$. Our hypothesis statements in words would be:

H_0 : half of the households use a family member for childcare

H_a : less than half of the households use a family member for childcare

and in symbols:

$$H_0 : p = 0.5$$

$$H_a : p < 0.5$$

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 500 households. The proportion of interest is the 243 out of the 500 households, so $n = 500$ and $p = 243/500 = 0.486$.

The sample size is a simple random sample of 500 households and it meets the conditions for a population proportion,

$$np = (500)(0.486) = 423 \geq 10$$

$$n(1-p) = (500)(1-0.486) = 257 \geq 10$$

so the distribution is approximately normal. This is a lower-tailed test because the alternative hypothesis uses the less than sign.

This is a population proportion so we will calculate a z-test statistic with the population proportion formula.

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

Let's state the values and calculate the test statistic. The sample proportion is the proportion of households who use a family member to care for their children under preschool age. According to the study that's 243 out of 500 households. That gives us:

$$\hat{p} = \frac{243}{500} = 0.486$$

We can get the hypothesized value from the hypothesis statements we made earlier, $H_0 : p = 0.5$ and $H_a : p < 0.5$. The sample size is the number of households in the survey, $n = 500$. Now we can calculate our z-test statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}} = \frac{0.486 - 0.50}{\sqrt{\frac{(0.5)(1 - 0.5)}{500}}} \approx -0.6261$$

The test statistic is somewhere on the lower-tail of the z-curve.

■ 5. The highest allowable amount of bromate in drinking water is 0.0100 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 mg/L^2 of bromate with a sample standard deviation of 0.0025. 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:

This is a population mean with an unknown standard deviation because the study is interested in comparing the average amount of bromate in drinking water and the population standard deviation is unknown. We're given the sample standard deviation in the problem.

According to the problem: "The highest allowable amount of bromate in drinking water is 0.0100 mg/L^2. This means we're interested in looking at the pair of hypothesis statements that could show us if the amount in the city's water is greater than this.

H_0 : the amount of bromate in the city's drinking water is equal to 0.0100 mg/L^2.

H_a : the amount of bromate in the city's drinking water is greater than 0.0100 mg/L^2.

and in symbols:

$$H_0 : \mu = 0.0100$$

$$H_a : \mu > 0.0100$$

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.

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

Let's state the values and calculate the test statistic. The sample mean is the average level of bromate in the samples, $\bar{x} = 0.0102$. The hypothesized value is the amount of allowable bromate in the drinking water. We can see this in our hypothesis statements.

$$H_0 : \mu = 0.0100$$

$$H_a : \mu > 0.0100$$

The sample standard deviation is given as $s = 0.0025$, and the sample size is the number of water samples, $n = 50$. Now we can calculate our t-test statistic:

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{0.0102 - 0.010}{\frac{0.0025}{\sqrt{50}}} \approx 0.5658$$

The test statistic is somewhere on the upper tail of the t-curve.

■ 6. A farmer reads a study that states: The average weight of a day-old chick upon hatching is 38.60 grams with a population standard deviation of 5.7 grams. The farmer wants to see if her day-old chicks are within the average quoted in the study. She takes a simple random sample of 60 of her day-old chicks and finds their average weight is 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:

This is a population mean with a known standard deviation because the farmer is interested in comparing her day-old chicks to the mean and population standard deviation given in the article. A side note about this: it's very rare to have an actual population standard deviation in a real statistics problem, so you should be skeptical when you see one.

According to the problem: "The farmer wants to see if her day-old chicks are within the average quoted in the study." This means we're interested in the pair of hypothesis statements that could show us whether or not her chicks are close to the quoted mean. In other words, the hypothesis statements are

H_0 : the day-old chicks are within 38.60 grams

H_a : the day old chicks are not within 38.60 grams

and in symbols:

$$H_0 : \mu = 38.60$$

$$H_a : \mu \neq 38.60$$

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 "not equal to" 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.

$$z = \frac{\bar{x} - \text{hypothesized value}}{\frac{\sigma}{\sqrt{n}}}$$

Let's state the values and calculate the test statistic. The sample mean is the average weight of the farmer's day-old chicks, 39.1 grams. The hypothesized value is the mean from the study that we used in the hypothesis statements.

$$H_0 : \mu = 38.60$$

$$H_a : \mu \neq 38.60$$

The population standard deviation is given as $\sigma = 5.7$. The sample size is the number of chicks in the farmer's sample, $n = 60$. Now we can calculate our z-test statistic:

$$z = \frac{\bar{x} - \text{hypothesized value}}{\frac{\sigma}{\sqrt{n}}} = \frac{39.10 - 38.60}{\sqrt{60}} \approx 0.6795$$

Since this is a two-tailed test, we look at both the positive and negative test statistics, $z \approx \pm 0.6795$. The test statistics are somewhere on both ends of the z-curve.

P-VALUES

■ 1. A private university is conducting a statistical test to determine whether or not the percentage of students who live on its campus is above the national average of $64\,\%$. They've calculated the test statistic as equal to 1.40. Set up the hypothesis statements and determine the type of test, then find the p-value.

Solution:

Let's set up the hypothesis statements to determine the type of test.

The private university wants to know if their percentage of students who live on campus is above the national average in a statistically significant way. Since the university is looked at a proportion on the students who live on campus we need to use a population proportion.

The hypothesis statements would be:

H_0 : the proportion of students who live on campus equals $64\,\%$

H_a : the proportion of students who live on campus is greater than $64\,\%$

and in symbols:

$$H_0 : p = 0.64$$

$$H_a : p > 0.64$$

From our hypothesis statements we can see that this is an upper tailed test because the alternative hypothesis uses the greater than sign. Since this is a population proportion, we'll look up the test statistic of 1.40 in the z -table.

The value in the table is $.9192$. This is an upper tailed test, which means we want the area outside of $.9192$. Remember that the total area under the curve adds to 1, so we can find the area we want by subtracting.

$$p\text{-value: } 1 - .9192 = .0808$$

■ 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 the sample of the $9{,}411$ births at the hospital to calculate a test statistic of -1.6. Set up the hypothesis statements and determine the type of test, then find the p-value.

Solution:

Let's set up the hypothesis statements to determine the type of test.

The hospital wants to know information about the average length of a pregnancy. They are comparing their data to a population with a known population standard deviation. They want to know if the average length of

pregnancy differs, so we need to use the "not equal to symbol" in our hypothesis statements. The hypothesis statements would be:

H_0 : the average length of pregnancy at the hospital is 283.6

H_a : the average length of pregnancy at the hospital is different from 283.6

and in symbols:

$$H_0 : \mu = 283.6$$

$$H_a : \mu \neq 283.6$$

From our hypothesis statements we can see that this is a two-tailed test because we used the "not equal to" sign in the alternative hypothesis.

For this test we have a population mean with a known population standard deviation, so the test statistic is a z-score. We're told in the problem that the test statistic is equal to -1.6. Since this is a two-tailed test, we'll need to double the area we find.

For the lower tail, $z = -1.6$ gives an area of $.0548$. Now to calculate our p-value, we multiply this by 2:

p-value: $2(.0548) = .1096$

The reason we multiply by 2 is that we're interested in the area in both the upper and lower tails of the standard normal curve. Remember the curve is symmetric, so if we find the area in one tail, we know the area in the other.

■ 3. 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 61 water samples in random locations around the city and used the data to calculate a test statistic of 0.57. 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:

Let's set up the hypothesis statements to determine the type of test. According to the problem: "The highest allowable amount of bromate in drinking water is $0.0100\ (\text{mg/L})^2$." This means we're interested in looking at the pair of hypothesis statements that could show us if the amount in the city's water is greater than the allowable amount.

H_0 : the amount of bromate in the city's drinking water is equal to $0.0100\ (\text{mg/L})^2$

H_a : the amount of bromate in the city's drinking water is greater than $0.0100\ (\text{mg/L})^2$

and in symbols:

$$H_0 : \mu = 0.0100$$

$$H_a : \mu > 0.0100$$

This is an upper-tailed test because the alternative hypothesis uses the greater than sign.

For this test, we have a population mean with an unknown population standard deviation, so the test statistic is a t-test.

To look up a t-test statistic, we'll also need to know the degrees of freedom from the problem. We know the study included 61 samples. The degrees of freedom are equal to $n - 1$, so for this study we have $61 - 1 = 60$ degrees of freedom.

We are told in the problem that the test statistic is equal to 0.57. To use the table, we'll need to round to the nearest tenth, so we'll use a test statistic of $0.57 \approx 0.6$.

Now we look up where our test statistic and degrees of freedom intersect. The diagram on the table shows us that this is the area in the upper tail, so we have our p-value.

p-value: $.275$

■ 4. A company produces red 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

$$H_0 : \mu = 15$$

$$H_a : \mu < 15$$

and using a sample of 14 observations from a random sample, she calculates a test statistic of -3.2. She assumes the distribution is

approximately normal. What would be the conclusions of her hypothesis test at significance levels of $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$?

Solution:

To come up with a conclusion to a hypothesis test, we compare the p-value of the test to the significance level. If the p-value is less than the significance level, we reject the null hypothesis. If the p-value is greater than the significance level, we fail to reject the null hypothesis.

This means to answer the question we need to calculate the p-value.

The painter is comparing a small sample of data that has an unknown population standard deviation. This means we need to use the t-test to look up our p-value. (We can use the t-test because she's assuming the population is normally distributed.) The test is a lower-tailed test because we have a "less than" sign in our alternative hypothesis statement. This means we're interested in the lower tail of the t-curve.

To look up the p-value for a t-test statistic, we'll also need to know the degrees of freedom from the problem. We know the study included 14 observations. The degrees of freedom are equal to $n - 1$, so for this study we have $14 - 1 = 13$ degrees of freedom. We're told in the problem that the test statistic is equal to -3.2.

Most of the time when you use a t-table, it'll only include the upper tails. To find the lower tails, remember the t-table is symmetric. If we look up positive 3.2, then we'll know the area in the lower tail at -3.2.

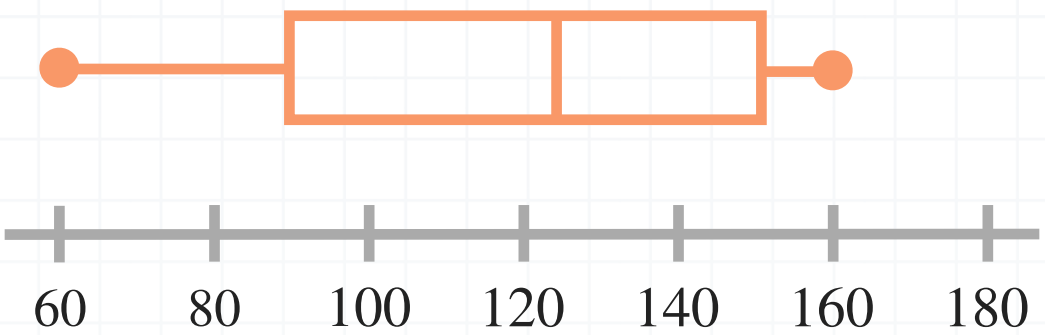
Now we look up where our positive test statistic and degrees of freedom intersect. The p-value is $.003$. Now let's compare this to each of our significance levels.

For $\alpha = .05$, we can say $.003 < .05$, so the p-value is less than the significance level and we can reject the null hypothesis and make a claim. There is enough evidence to suggest that the paint glows for less time than was advertised at the significance level of $.05$.

For $\alpha = .01$, we can say $.003 < .01$, so the p-value is less than the significance level and we can reject the null hypothesis and make a claim. There is enough evidence to suggest that the paint glows for less time than was advertised at the significance level of $.01$.

For $\alpha = .001$, we can say $.003 > .001$, so the p-value is greater than the significance level. We fail to reject the null hypothesis, and we can't make a claim. There is not enough evidence to suggest that the paint glows for less time than was advertised at the significance level of $.001$.

■ 5. A manager reads an article that reports the average wasted time by an employee is 125 minutes every day. She takes a small random sample of 16 employees and monitors their wasted time. She calculates the following based on her observations: the average wasted time for her employees is 122 minutes with a standard deviation of 28.7. She wants to know if 122 minutes is below average at a significance level of $\alpha = .05$. She creates a box plot of her data for "wasted time at work" to check for normality:

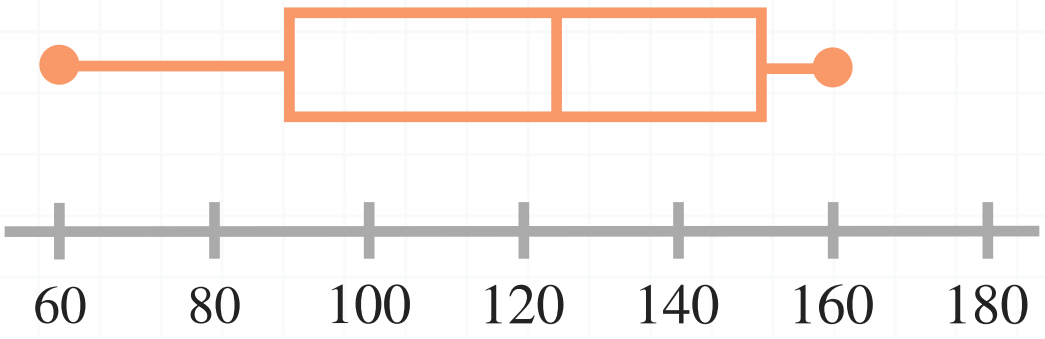


1. State the population characteristic and whether you should use a t-test or z-test statistic.
2. Check that the conditions for performing the statistical test are met.
3. Set up the hypothesis statements.
4. State the type of test: upper-tailed, lower-tailed or two-tailed.
5. Calculate the test statistic 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. We're given the box plot of the data.



We can see that the quartiles are almost evenly distributed, so we can say the distribution is approximately normal. We're also told in the problem: "She takes a small random sample of 16 employees..." so we know the data came from a simple random sample. We have met the conditions we need to continue with the hypothesis test.

The manager wants to know if 122 minutes is below average. We're comparing 122 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 : the wasted time is equal to 125 minutes

H_a : the wasted time is less than 125 minutes

and in symbols:

$$H_0 : \mu = 125$$

$$H_a : \mu < 125$$

This is a lower tailed test because of the less than sign in the alternative hypothesis.

This is a population mean with a small sample size and an unknown population standard deviation, so we'll calculate a t-test statistic with the population mean formula.

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

Let's state the values and calculate the test statistic. The sample mean is the average time wasted by the manager's employees, $\bar{x} = 122$. The hypothesized value is the mean from the study that we used in the hypothesis statements, 125, and the sample standard deviation is given as $s = 28.7$. The sample size is the number of employees in the manager's sample, $n = 16$, so we can now calculate the z-test statistic:

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{122 - 125}{\frac{28.7}{\sqrt{16}}} \approx -0.4181$$

The next step is to find our 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 $t \approx -0.4181$, but to use the table, we'll need to round to the nearest tenth, so we'll use $t = -0.4$. We're looking for the area in the lower tail, but the table will give us the area in the upper tail when $t = 0.4$. 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 $p = .312$.

Now we need to use our p-value and the significance level of the test to draw a statistical conclusion. With $p = .312$ and a significance level of $\alpha = .05$, we can say $.312 > .05$. The p-value is greater than the significance level, so we fail to reject the null hypothesis, and we conclude that there is not enough evidence to conclude that the manager's employees waste less time than the average rate of 125 minutes per day at the significance level of $\alpha = .05$.

■ 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 leading cardiologists were attending an academic conference.

Is there enough evidence to conclude that the percentage of patients who die when the lead cardiologists are away is any different than when they are at the hospital? Use a significance level of $\alpha = .05$ and also $\alpha = .01$.

1. State the population characteristic and whether you should use a t-test or z-test statistic.

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. For a population proportion, you use the z-test statistic.

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:

$$np = (3{,}153)(.167) = 527 \geq 10$$

$$n(1-p) = (3{,}153)(1-.167) = 2{,}626 \geq 10$$

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 cardiologists were attending an academic conference differs from when they were not away. This means we need to use the "not equal to sign" in our hypothesis statement.

$H_0 : 15.3\,\%$ of patients admitted to the hospital with a heart attack die within 30 days

H_a : the percentage of patients who are admitted to the hospital with a heart attack and die within 30 days is different than $15.3\,\%$

and in symbols:

$$H_0 : p = .153$$

$$H_a : p \neq .153$$

This is a two-tailed test because of the not equal to sign in the alternative hypothesis. Since we're dealing with a population proportion, we'll calculate a z-test statistic with the population proportion formula.

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

Let's state the values and calculate the test statistic. The sample proportion is the $16.7\,\%$ of patients who went to the hospital with a heart attack and died within 30 days of admission when the leading cardiologists were attending an academic conference. So $\hat{p} = .167$. And we can get the hypothesized value from the hypothesis statements we made earlier, and say that the hypothesized value is 0.153.

The sample size is the number of patients who went to the hospital with a heart attack died within 30 days of admission when the leading cardiologists were attending an academic conference, which is $n = 3{,}153$. Now we can calculate the p-value.

$$z = \frac{0.167 - 0.153}{\sqrt{\frac{(.153)(1 - .153)}{3{,}153}}} \approx 2.1837$$

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. The table goes to the nearest hundredth so we can round the z-score to $z \approx 2.18$.

From the table, we read a value of $.9854$. 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 the value from the table from 1.

$$1 - .9854 = .0146$$

Now to calculate the p-value, we multiply the upper tail by 2.

$$p\text{-value: } 2(.0146) = .0292$$

Next, we use the p-value and the significance level to make a statistical claim. First let's look at $\alpha = .05$. We know that $.0292 < .05$, so since the p-value is less than the significance level, we can reject the null hypothesis and make a statistical claim.

There is 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 cardiologists were attending an academic conference is different than when the leading cardiologists are present, at a statistical significance level of $\alpha = .05$.

On the other hand, now let's consider $\alpha = .01$. We know that $.0292 > .01$, so since the p-value is greater than the significance level, we can't reject the null hypothesis and make a statistical claim.

There is not 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 cardiologists were attending an academic conference is different than when the leading cardiologists are present, at a statistical significance level of $\alpha = .01$.

CONFIDENCE INTERVALS OF A POPULATION MEAN

■ 1. 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.

$$\bar{x} = \frac{11.5 + 18.5}{2} = \frac{30}{2} = 15$$

■ 2. 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 small bags of chocolate candy. From the sample she found an average of 17 pieces of candy per bag with a standard deviation of 2.030.

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 amount of chocolate candy 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 one-sample t confidence interval.

To calculate the confidence interval we use the formula

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Let's set up the values we need for the calculation. We know that $\bar{x} = 17$ is the sample mean and that $s = 2.030$ is the sample standard deviation. We also know the sample size, $n = 20$.

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 table is 2.093.

Now we can compute the confidence interval.

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

$$(a, b) = 17 \pm 2.093 \cdot \frac{2.03}{\sqrt{20}}$$

$$(a, b) = 17 \pm .9501$$

$$(a, b) = (17 - .9501, 17 + .9501)$$

$$(a, b) = (16.0499, 17.9501)$$

The confidence interval means: Based on the sample we are $95\,\%$ confident that the average number of chocolates per bag is between 16.0499 and 17.9501 pieces.

■ 3. 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 you would get a better idea of your estimate for 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.

$$(a, b) = 17 \pm .9002$$

Here you can see that you're adding and subtracting a smaller amount when the sample size is doubled. This would make the confidence interval narrower, which means you would get a better idea of your estimate for the population mean.

■ 4. A magazine took a random sample of 540 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 540 people and an unknown population standard deviation. The sample size is large enough to make the distribution approximately normal. This means we need to use a one-sample t confidence interval.

To calculate the confidence interval we use the formula:

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

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 $n = 540$ and that t^* is the test statistic. We look this up in the t-table. We need to use the confidence level and the degrees of freedom. The confidence level is $98\,\%$, and the degrees of freedom is $n - 1 = 540 - 1 = 539$. Since most t-tables only show 100 and $1{,}000$ degrees of freedom, we'll round 539 up to $1{,}000$. Then the value we find in the t-table is $t^* = 2.539$. Now we can compute the confidence interval.

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

$$(a, b) = 44.78 \pm 2.539 \cdot \frac{18.10}{\sqrt{540}}$$

$$(a, b) = 44.78 \pm 1.98$$

$$(a, b) = (42.80,46.76)$$

The confidence interval means: Based on the sample, we're $98\,\%$ confident that the average amount spent on Easter baskets was between $\$42.80$ and $\$46.76$.

■ 5. The national average of calories served in a restaurant meal is $1{,}500$ calories per meal from a sample of 31 randomly selected samples. For illustrative purposes, say that the a population standard deviation of the calories in a restaurant meal is 350.2. Construct and interpret a $90\,\%$ confidence interval for the mean number of calories in a restaurant meal.

Solution:

We're told in the problem that the data is from a simple random sample. We have a large sample size of 31 people and an known population standard deviation. The sample size is large enough to make the distribution approximately normal. This means we can use a one sample z confidence interval.

To calculate the confidence interval we use the formula:

$$(a, b) = \bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

Let's set up the values we need for the calculation. We know the sample mean is $\bar{x} = 1{,}500$ and the population standard deviation is $\sigma = 350.2$. We also know the sample size, $n = 31$. The test statistic is z^* and the confidence level is $95\,\%$.

We're going to need to use what we know about a z-curve to figure out our test statistic. We know that the area under the curve is $100\,\%$. Let's look at what the area in both of the tails would be. We know we want a confidence level of $95\,\%$ and that this is the area in between the two tails. The area in both tails is then

$$100\% - 95\,\% = 5\,\%$$

The area in one tail is

$$\frac{5\,\%}{2} = 2.5\,\% = .025$$

We can look up $.025$ in the lower tail area to find z^*. The value we read from the z-table is $z^* = \pm 1.96$. Now we can calculate the confidence interval.

$$(a, b) = 1{,}500 \pm 1.96 \cdot \frac{350.2}{\sqrt{31}}$$

$$(a, b) = 1{,}500 \pm 123.28$$

$$(a, b) = (1{,}376.72, 1{,}623.28)$$

The confidence interval means: Based on the sample, we are $95\,\%$ confident that the average amount of calories in a restaurant meal was between $1{,}376.72$ and $1{,}623.28$ calories.

■ 6. A bus that travels from Kansas City to Denver had the following travel times, in hours, on 11 randomly selected bus trips:

11.7	12.0	11.75	11.5
12.25	11.5	12.0	11.5
11.25	11.25	11.75	

Construct and interpret a $95\,\%$ confidence interval for the mean bus trip time in hours from Kansas City to Denver.

Solution:

We're told in the problem that the data is from a simple random sample (because the bus trips were randomly selected). We have a small sample size of 11 bus trips and an unknown population standard deviation. We need to check to see that the population is approximately normal.

If we were to make a box plot of the data, we can see that the data is fairly evenly distributed across quartiles, so the distribution is approximately normal. Since we have a small sample size and unknown population standard deviation but an approximately normal distribution, we need to use a one-sample t confidence interval.

To calculate the confidence interval, we use the formula

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Let's set up the values we need for the calculation. We know from the problem that the sample size is $n = 11$. We can use the data to find the sample mean, $\bar{x} \approx 11.677$. The sample standard deviation we also calculate from the data set, and we get

$$s \approx \sqrt{.100682} \approx .317$$

We need to find the test statistic t^*, which we'll look up in the t-table. We need to use the confidence level of $95\,\%$ and the degrees of freedom $n - 1 = 11 - 1 = 10$. From the table, we find a value of $t^* = 2.228$.

Now we can compute the confidence interval.

$$(a, b) = 11.677 \pm 2.228 \cdot \frac{.317}{\sqrt{11}}$$

$$(a, b) = 11.677 \pm .213$$

$$(a, b) = (11.464, 11.890)$$

The confidence interval means: Based on the sample, we are $95\,\%$ confident that the average bus trip from Kansas City to Denver takes between 11.464 and 11.890 hours.

CONFIDENCE INTERVALS OF A POPULATION PROPORTION

■ 1. A court case questioning the use of a drug dog named Bentley, due to his low success rate of correctly alerting to drugs, was in the U.S. Court of Appeals (U.S. vs. Bentley). In Bentley's time on the job, it's estimated that he correctly identified drugs $59\,\%$ of the time. How many different trials should they put Bentley through to show that this is his actual success rate at a $95\,\%$ confidence level with a margin of error of $.05$?

Solution:

You can use the formula

$$n = \hat{p}(1-\hat{p})\left(\frac{z^*}{B}\right)^2$$

to find the correct sample size (number of trials) for the case. Bentley's success rate is:

$$\hat{p} = 59\,\% = 0.59$$

The confidence level is $95\,\%$ and the test statistic for this confidence level is $z^* = 1.96$. The margin of error is $\beta = .05$. Now we can find the number of trials we need for Bentley.

$$n = 0.59(1-0.59)\left(\frac{1.96}{0.05}\right)^2$$

$$n \approx 371.71$$

Since we need more than 371 trials to test Bentley, we have to round up to 372 trials, so we can say $n = 372$.

■ 2. 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 to have a $90\,\%$ confidence level with a margin of error of $.08$?

Solution:

You can use the formula

$$n = \hat{p}(1 - \hat{p})\left(\frac{z^*}{B}\right)^2$$

to find the correct sample size (number of trials) for this scenario. 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 $z^* = 1.645$. The margin of error is $\beta = .04$. Now we can find the number of juniors Sarah needs to include in her simple random sample.

$$n = 0.65(1 - 0.65)\left(\frac{1.645}{.08}\right)^2$$

$$n \approx 96.19$$

Since we need more than 96 juniors for the sample, we have to round up to 97 juniors, so we can say $n = 97$.

■ 3. 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 $.05$?

Solution:

You can use the formula

$$n = \hat{p}(1 - \hat{p})\left(\frac{z^*}{B}\right)^2$$

to find the correct sample size (number of trials) for this scenario. The confidence level is $95\,\%$ and the test statistic for this confidence level is $z^* = 1.96$. The margin of error is $\beta = .05$. Now we can find the number of physicians we need to include in the simple random sample.

$$n = 0.10(1 - 0.10)\left(\frac{1.96}{.05}\right)^2$$

$$n \approx 138.2976$$

Since we need more than 138 physicians for the sample, we have to round up to 139 physicians, so we can say $n = 139$.

■ 4. A study shows that $78\,\%$ of patients who try a new medication for migraines feel better within 30 minutes of taking the medicine. If the study involved 120 patients, construct and interpret a $95\,\%$ confidence interval for the proportion of patients who feel better within 30 minutes of taking the medicine.

Solution:

We want to use the large-sample confidence interval for a population proportion:

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

First, we need to check that the sample is large enough. This means we need to check that the proportion of "successes" $n\hat{p}$ and the proportion of "failures" $n(1-\hat{p})$ are both greater than 10. With $n = 120$ and $\hat{p} = .78$, we get

$$n\hat{p} = 120(.78) = 93.6 \geq 10$$

$$n(1-\hat{p}) = 120(1-.78) = 26.4 \geq 10$$

We know that the sample proportion is $\hat{p} = .78$, and that the confidence level is $95\,\%$. The test statistic for this confidence level is $z^* = 1.96$. The sample size is $n = 120$. Now you can calculate the interval.

$$(a, b) = .78 \pm 1.96 \sqrt{\frac{.78(1-.78)}{120}}$$

$$(a, b) = .78 \pm .074$$

$$(a, b) = (.706,.854)$$

This means that we're $95\,\%$ confident that the proportion of patients who feel better within 30 minutes after taking the medicine is between $70.6\,\%$ and $85.4\,\%$.

■ 5. A study shows that 243 out of 500 randomly selected households were using a family member to care for their children who were under preschool age. Construct and interpret a $90\,\%$ confidence interval for the proportion of households using a family member to care for children under preschool age.

Solution:

We want to use the large-sample confidence interval for a population proportion:

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Let's calculate the population proportion.

$$\hat{p} = \frac{243}{500} = .486$$

Now we need to check that the sample is large enough. This means we need to check that the proportion of "successes" $n\hat{p}$ and the proportion of

"failures" $n(1-\hat{p})$ are both greater than 10. We know that $n = 500$ and that $\hat{p} = .486$, so

$$n\hat{p} = 500(.486) = 243 \geq 10$$

$$n(1-\hat{p}) = 500(1-.486) = 257 \geq 10$$

Next, we need to write down our numbers to evaluate the interval:

$$(a,b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The rate is $\hat{p} = .486$ and the confidence level is $95\,\%$. The test statistic for this confidence level is $z^* = 1.645$, and the sample size is $n = 500$. Now we can calculate the interval.

$$(a,b) = .486 \pm 1.645 \sqrt{\frac{.486(1-.486)}{500}}$$

$$(a,b) = .486 \pm .0368$$

$$(a,b) = (.4492,.5228)$$

This means that we're $90\,\%$ confident that the proportion of households using a family member to care for children under preschool age is between $44.92\,\%$ and $52.28\,\%$.

■ 6. 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 percent of Americans who make weekend plans based on the weather.

Solution:

We want to use the large-sample confidence interval for a population proportion:

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Now we need to check that the sample is large enough. This means we need to check that the proportion of "successes" $n\hat{p}$ and the proportion of "failures" $n(1-\hat{p})$ are both greater than 10. We know that $n = 648$ and that $\hat{p} = .47$, so

$$n\hat{p} = 648(.47) = 304.56 \geq 10$$

$$n(1-\hat{p}) = 648(1-.47) = 343.44 \geq 10$$

Next, we need to write down our numbers to evaluate the interval:

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

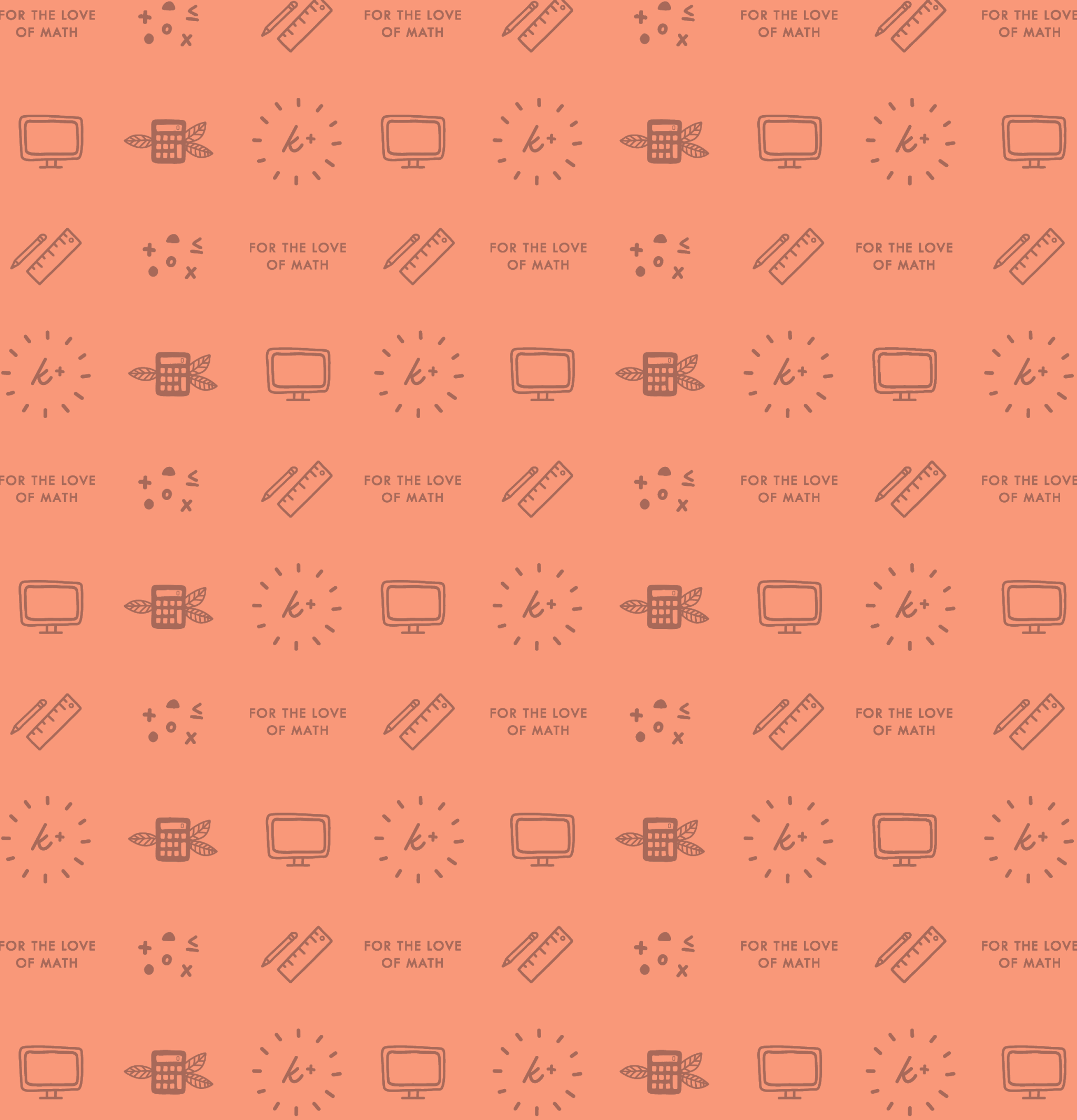
The rate is $\hat{p} = .47$ and the confidence level is $99\,\%$. The test statistic for this confidence level is $z^* = 2.576$ and the sample size is $n = 648$. Let's calculate the interval.

$$(a, b) = .47 \pm 2.576 \sqrt{\frac{.47(1-.47)}{648}}$$

$$(a, b) = .47 \pm .0505$$

$$(a, b) = (.4195,.5205)$$

This means that we're $99\,\%$ confident that the percentage of Americans who make weekend plans based on the weather is between $41.95\,\%$ and $52.05\,\%$.


FOR THE LOVE
OF MATH