STATWAY® COLLEGE MODULE 1 RECONTEXTUALIZED FOR SOUTH AFRICA

TABLE OF CONTENTS:

Note: Unit 1.5 is not used in the South African version of Statway College; the link above for that unit is to the standard version.

1.1: Setting Course Expectations and Creating Productive Classroom Norms

Statway Cohort Contract

By signing below, I agree to fulfill the following requirements for participation in Statway, and acknowledge that I understand the requirements for continued enrollment.

Specifically:

- I commit to successfully completing Statway with the members of my cohort.
- I commit to helping all of my cohort members understand mathematics and complete Statway.
- I will come to class everyday prepared to participate in all classroom activities.
- I will contribute to creating a productive classroom atmosphere that supports everyone learning.
- I will keep an open mind and a positive attitude, and will be willing to try out new learning strategies and study skills.

Signature: _____ Date: _____

Witness: _____ Date: _____

Preparation 1.2

What is Statistics? Why Do We Study Statistics?

Statistics is about using **data** to answer questions. Data is information that we collect from our world. Data involves facts and observations that we make. Before scientists created statistics and before people used data, they would use opinions and hunches to explain how the world worked. A lot of times these explanations were wrong.

For example, people once believed that the sun revolved around the earth. When people started making observations and using measures, they discovered that this was incorrect. The earth revolves around the sun. Data helps us draw better conclusions.

Data can help us answer many types of questions.

- Students can use data to help pick a college that is best for them.
- Teachers use data to find the best ways to educate their students.
- Medical professionals use data to learn if new treatments actually work.
- Voters use data about their society and planet to create a better democracy.
- Politicians use data to better represent the people who elect them.

In statistics, we gather, summarize, and analyze data to search for answers to our questions.

Let's begin with an exercise that involves using data to answer a question. Remember that data involves facts, observations, and measures about a particular topic or idea.

How significant is your birthday? Many people believe that habits, hobbies, likes, dislikes and many other things are affected by the time of year you were born. One such theory involves *chronotypes*. The term "chronotype" refers to a person's natural tendency to be most active and alert at certain times of the day. People who are most active early in the day are often labeled as morning people, early birds or larks. Those who are at their best in the evening are called night people or owls. Now imagine that you want to answer the following question: Can someone's birth month influence his or her chronotype?

INTRODUCTION

In Collaboration 1.2, you will be introduced to the statistical analysis process and learn how this process enables you to use data to make decisions. How significant is your birthday? Many people believe that habits, hobbies, likes, dislikes and many other things are affected by the time of year you were born. One such theory involves chronotypes. The term "chronotype" refers to a person's natural tendency to be most active and alert at certain times of the day. People who are most active early in the day are often labeled as morning people, early birds or larks. Those who are at their best in the evening are called night people or owls. Now imagine that you want to answer the following question: **Can someone's birth month influence his or her chronotype?**

To answer this question, we could:

- Ask a group of people for their birthdate and ask about their chronotype.
- Examine whether people with certain birth months tend to have certain chronotypes. We are looking for a pattern in the data which shows a clear relationship between birth month and chronotype.

Before you perform this statistical investigation, you will complete an activity to identify your chronotype and check whether or not Chronotype Theory accurately predicts your chronotype based on your birth month.

Identify your Chronotype: Are you a morning person, a night person, or neither?

On a sheet of paper, for each characteristic in the table below, write down the number of the answer that best applies to you (1, 2, or 3). There are a total of 11 characteristics in the table, so you should write down 11 numbers.

Characteristic[1]	1	2	3
Most alert	Around noon	Around 6 PM	Equally alert at noon and 6 PM
Most productive	Morning	Evening	Equally productive in morning and evening
Alarm clock	Don't need it	Need multiple alarms	One alarm is enough
Favorite exercise time	Morning	Evening	Equally likely to exercise in morning and evening
Mood	Declines steadily all day	Rises steadily all day	Same throughout the day
Favorite meal	Breakfast	Dinner	I like breakfast and dinner equally
Coffee use	None or a small amount	Quite a bit	Somewhere in the middle
Class Time	I would prefer a class at 8 AM	I would prefer a class at 7 PM	No preference between a class at 8 AM or 7 PM
Morning behavior	Full of energy	Out of steam	Somewhere in the middle
Evening behavior	Out of steam	Full of energy	Somewhere in the middle
Travel	I get bad jet lag	Adapt quickly to time zone changes	I get a bit of jet lag

[1]Source: Adapted from http://www.nasw.org/users/llamberg/larkowl.htm.

1 How many 1s did you choose?

2 How many 2s did you choose?

3 How many 3s did you choose?

Chronotype Selection

Look at your answers above and use the following to determine your chronotype:

- If you picked more 1s than other choices, this indicates that your chronotype is a morning person.
- If you picked more 2s than other choices, this indicates that your chronotype is a night person.
- If you picked more 3s than other choices, this indicates that your chronotype is neither a morning person nor a night person.

*Tiebreaker. If you have a tie between your two top columns, simply select the result with which you most identify.

Write down your chronotype from this activity on your paper. You will need it for the Collaboration.

Chronotype Theory

Chronotype Theory states that there are three chronotypes and four birth months that correspond to each chronotype. The theory states that being a (a) morning person, (b) night person or (c) neither/both is related to the month when you were born, and predicts your chronotype based on your birth month. The table below shows the predictions for each birth month.

Month	Chronotype	Month	Chronotype
January	Evening	July	Morning
February	Evening	August	Morning
March	Neither/Both	September	Neither/Both
April	Neither/Both	October	Neither/Both
May	Morning	November	Evening
June	Morning	December	Evening

Use the table to identify which chronotype matches your birth month, according to Chronotype Theory. Note that on your paper.

Did the chronotype indicated by the initial activity match the chronotype predicted by Chronotype Theory using your birth month only? Make a note on your paper whether the two chronotypes matched or did not match. You will use all of this information in Collaboration 1.2.

1.2 The Statistical Analysis Process

LEARNING GOALS

By the end of this collaboration, you should understand that:

- Statistical investigations are a cycle of forming a research question, designing a study and collecting relevant data, data analysis, statistical inference, and drawing appropriate conclusions.
- Chance variability can affect the statistical decision-making process.
- Unusual results can provide evidence for a claim about a population proportion.

By the end of this collaboration, you should be able to:

- Identify the steps in the statistical analysis process given a description of a statistical study.
- Interpret data displayed in a dotplot.
- Evaluate the strength of evidence against a claim about a population proportion.

INTRODUCTION

In this collaboration you will be introduced to the statistical analysis process and learn how this process enables you to use data to make decisions.

How significant is your birthday? Many people believe that habits, hobbies, likes, dislikes and many other things are affected by the time of year you were born. One such theory involves *chronotypes*. The term "chronotype" refers to a person's natural tendency to be most active and alert at certain times of the day. People who are most active early in the day are often labeled as morning people, early birds or larks. Those who are at their best in the evening are called night people or owls. In this investigation you will answer the question:

Can someone's birth month influence their chronotype?

CHRONOTYPE THEORY

Chronotype Theory states that there are three chronotypes and four birth months that correspond to each chronotype. The theory states that being a (a) morning person, (b) night person or (c) neither/both is related to the month when you were born, and predicts your chronotype based on your birth month. The table below shows the predictions for each birth month.

Month	**Chronotype**
November, December, January, February	Evening
March, April, September, October	Neither/Both
May, June, July, August	Morning

1A In the preparation for this collaboration, you determined your chronotype based on your personal characteristics. Does the birth-month theory correctly predict your chronotype? Explain.

1B In Preparation 1.2, you were introduced to Chronotype Theory and each student determined whether or not the chronotype they selected for themselves is the same as the chronotype predicted for them by Chronotype Theory.

What fraction of students in your group selected the chronotype that matched the chronotype predicted for them by Chronotype Theory? Express the fraction as a decimal and percentage.

1C Can you use the information collected from your group to decide if Chronotype Theory is reasonable or not reasonable for all college students?

Statistical analysis is the process of looking at data to learn about something bigger. Looking at data allows us to make **generalizations** about **populations** that are large and difficult to understand. We can think of the statistical analysis process in four steps.

Steps in a Statistical Analysis

Step 1: Ask a question that can be answered by collecting data.
Step 2: Decide what to measure and then collect data.
Step 3: Summarize and analyze the data.
Step 4: Draw a conclusion and communicate the results.

A STATISTICAL INVESTIGATION

Consider the following situation. A researcher obtains a random sample of 50 college students that is representative of all college students. Each student in the sample identified their chronotype and determined whether Chronotype Theory correctly predicted the chronotype they selected. Of the 50 students, 26 indicate that they selected the chronotype predicted for them by Chronotype Theory.

With this sample data, we can perform a statistical analysis to investigate if Chronotype Theory is reasonable.

2 Answer each of the following questions about the four steps of the statistical analysis process.

Step 1: Ask a question that can be answered by collecting data.

A What question are we trying to answer in our investigation?

Step 2: Decide what to measure and then collect data.

B What information can we get from each student to answer this question? How are the data related to the question we are trying to answer?

Step 3: Summarize and analyze the data.

C Each student identified their chronotype and compared it to the chronotype predicted from Chronotype Theory for his or her birth month. Some students found that the predicted chronotype matched the one they selected and others did not. How could we *summarize* the results for all students in the class?

D Of the 50 students in the sample, 26 indicate that they selected the chronotype predicted for them by Chronotype Theory. What percentage of students in the sample selected the chronotype that was predicted for them by Chronotype Theory?

(i) 126%

(ii) 50%

(iii) 52%

(iv) 13%

Step 4: Draw a conclusion and communicate the results.

E Given the summary of the data, how can we use the summary to answer the research question? What would we expect to learn from our data? Can we generalize, or **infer**, the results to something bigger than our class?

Language Tip

To *infer* means to use statistical evidence to make a conclusion that applies to a larger group than the sample.

NEXT STEPS

We have outlined how we might approach each of the steps in a statistical analysis in order to determine whether Chronotype Theory is reasonable. We have completed steps 1 and 2 and part of step 3. We have asked a question to be answered with data, collected data, and summarized the data. We now must think carefully about how we can analyze the data to draw a reasonable conclusion.

To analyze our data, we need to think about how the data can be used to answer our question. We will use *probability* to help in this. Talk about the following questions with your group to begin this process.

For each question below, take a minute to think about this on your own first before sharing in your group. Explain your answers.

3 Suppose Chronotype Theory is false and a person's birth month does **not** predict their chronotype. Is it possible that a student in the class would still select the chronotype that Chronotype Theory predicted for them? Explain why you think this.

4 Suppose Chronotype Theory is false and a person's birth month does not predict their chronotype. Out of all the students in the sample, about what percentage of the students should select chronotypes that matched the predictions?

(i) 100%

(ii) 50%

(iii) 33%

(iv) 3%

5 In the random sample of 50 students, 52% selected chronotypes that matched the predictions from Chronotype Theory. Does this result provide evidence in support of, or against, Chronotype Theory? Explain.

NEXT STEPS

Using Probability to Discover What May Happen by Chance

Language Tip

Chance variation **describes the type of differences we would naturally expect to see between different samples.**

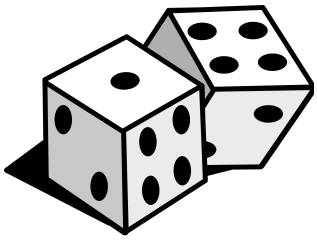
We must use additional tools to assess whether the class result provides evidence in support of, or against, Chronotype Theory. If Chronotype Theory is false and birth month does not predict chronotype, there is a ⅓ chance that a student would select the

chronotype predicted for them solely due to chance. So, based on chance alone, we would expect about ⅓ of the students in the class to select the chronotype that is predicted for their birth month. But how far above 1/3 would the fraction need to be in order to convince us that the opposite is true, that is, that Chronotype Theory is reasonable?

In the next activity, we will answer this question by learning about the values which are most probable. Probable values are those that are most likely to occur by **chance variation**.

If Chronotype Theory is false, we would expect ⅓ of the students in the class to select their predicted chronotype by chance. We will use a 6-sided die to investigate how results vary due to chance when an event has a ⅓ chance of occurring.

In a 6-sided die, each of the six outcomes {1,2,3,4,5,6} are equally likely to occur. The event of rolling a 1 or 2 should occur 33% or ⅓ of the time since this event contains two of the six outcomes that are possible.



6 We will use the following online dice rolling simulator at www.random.org/dice/ to roll a die 50 times and keep track of the number of rolls that produce an outcome of “1” or “2”. In the simulation, select “Roll 50 virtual dice.” The number of rolls is based on the number of students in the random sample. Since there were 50 students in the sample, roll the die 50 times. The outcomes of the 50 die rolls will be visible all at once. Write the fraction and proportion of die rolls that result in a “1” or “2”.

Note: In statistics, a **proportion** is a number between 0 and 1. It represents a portion out of the total. We usually give proportions as decimals or percents. We can calculate a decimal proportion by dividing the **numerator** of a fraction by the **denominator**. For example, if you roll the die 50 times, and the fraction of die rolls that result in a “1” or “2” is (12/50), then you divide 12 by 50. The proportion would be 0.24. To change to a percent we multiply by 100 or move the decimal 2 places to the right. 0.24 = 24%.

7 How does the proportion of die rolls that resulted in a “1” or “2” compare to 0.33?

8 If you repeat the experiment a second time, do you think you will get exactly the same proportion of die rolls resulting in a “1” or “2”? Select the most appropriate answer below.

(i) Yes, since the outcomes of the rolls will be similar each time we roll the die 50 times.

(ii) Yes, since the outcomes of the 50 rolls are based on a random process.

(iii) No, since the outcomes will vary since this is a random process.

(iv) No, since the die has changed over time.

9 Repeat the experiment 5 times, and for each rolling of 50 die, keep track of the proportion of die rolls resulting in a “1” or “2”. Enter in all of the observed proportions from the experiment.

Roll 1
Number of die rolls resulting in "1" or “2”:
Proportion (out of 50):

Roll 2
Number of die rolls resulting in "1" or “2”:
Proportion (out of 50):

Roll 3
Number of die rolls resulting in "1" or “2”:
Proportion (out of 50):

Roll 4
Number of die rolls resulting in "1" or “2”:
Proportion (out of 50):

Roll 5
Number of die rolls resulting in "1" or “2”:
Proportion (out of 50):

10 Your instructor has constructed a **dotplot** which shows the observed proportions of die rolls that resulted in a “1” or “2” for all students in the class. **Note**: If completing this problem online, follow the instructions given online to create your own dotplot.

11 Use the dotplot to answer these questions.

A What was the smallest proportion observed?

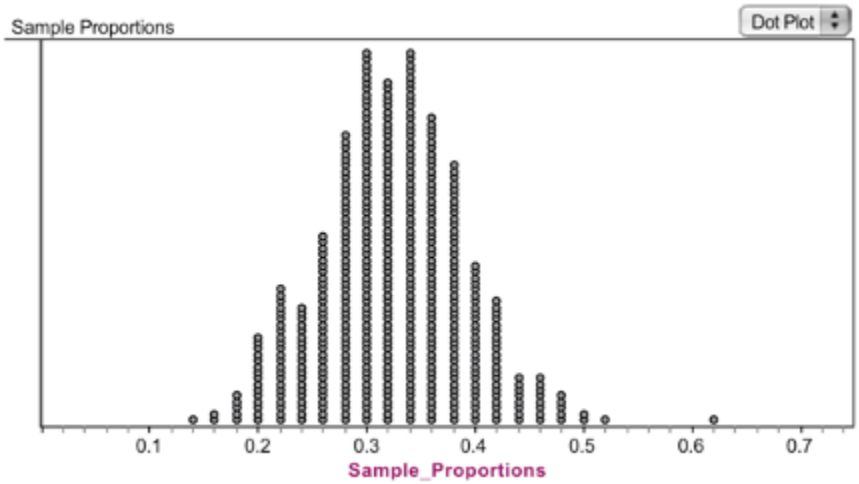
B What was the largest proportion observed?

C Did the proportions differ much from trial to trial?

D Where is the dotplot centered? Why do you think this is?

12 Summarize what the dotplot tells us. What can you learn from this graph?

A **distribution** of data (like the proportion of die rolls resulting in "1" or "2") shows the values that emerge from an experiment and how often each of the values occur. The dotplot which you created is an example of a distribution. To better understand what outcomes are most likely to occur due to chance alone, we can use a computer simulation with a larger number of trials. Below are the results of 500 simulations, where a die was rolled 50 times. Each dot represents the proportion of rolls that resulted in a 1 or 2 out of 50 in a single simulation.



If Chronotype Theory is false, we would expect that ⅓ of the students in the random sample would select the chronotype predicted for them solely by chance. Since a die roll also has a ⅓ chance of resulting in a "1" or "2", the distribution of proportions of die rolls displays the results we could expect to see from the class if Chronotype Theory is false.

13 In the random sample of 50 students, the proportion of students who selected a chronotype for themselves that matched the chronotype predicted for them by Chronotype Theory was 0.52. How does this proportion compare to the values in the dotplot containing 500 sample proportions?

Draw a Conclusion and Communicate the Results

Decision time! It is now time to make an inference. An inference is a decision based on the evidence we have gathered.

14 Does the sample proportion of 0.52 provide convincing evidence that Chronotype Theory is reasonable? Why or why not?

NEXT STEPS

The task you have just completed illustrates the statistical analysis process, which we have described in four steps.

Steps in a Statistical Analysis

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and then collect data.
3. Summarize and analyze the data.
4. Draw a conclusion and communicate the results.

A statistical investigation is an ongoing process. Researchers analyze the results of one study and use the results to generate new research questions. Then they conduct more research. This enables them to think more carefully about how they collect data and identify ways to improve the data collection process.

The table below summarizes what we just completed for the Chronotype investigation.

Steps in Statistical Analysis	For the Chronotype Investigation
1. Ask a question that can be answered by collecting data.	Is Chronotype Theory reasonable? Does a person's birth month predict their chronotype?
2. Decide what to measure and then collect data.	For each student in the class, we collected their birth month, the chronotype they selected for themselves, and the chronotype predicted for them by Chronotype Theory.
3. Summarize and analyze the data.	We summarized the data by determining the proportion of students in the class whose selected chronotype matched their predicted chronotype. We compared the class proportion to the distribution of results that would occur solely due to chance variation.
4. Draw a conclusion and communicate the results.	If the class proportion was similar to results that would occur due to chance, we conclude that Chronotype Theory is false. If the class proportion is much greater than the results that would occur solely due to chance variation, we conclude that Chronotype Theory could be true.

TRY THESE

Consider how you would identify the steps in a statistical analysis in the following study:

A group of researchers studied women who had visited a fertility clinic. The researchers wondered if fewer than half of the women who visit the clinic would want to choose the gender of their future child.[2]

Language Tip

Convincing evidence **is information that provides very strong support for a conclusion. The evidence is beyond what would be expected due to chance variation.**

The researchers mailed a survey to women who had visited the clinic. The survey asked women if they would choose the gender of their future child, if they were able to do so. A total of 561 women responded to the survey. Of these 561 women, 229 said that they wanted to choose the gender of their future child.

The researchers did statistical analysis on the data. Based on their statistical analysis of these data, the researchers concluded that there is **convincing evidence** that *fewer than half* of the women who visit the clinic would choose the gender of a future child. This conclusion is based on the following observation:

If, in reality, at least half of women who visit a fertility clinic would like to choose the gender of a future child, it would be very unusual to observe a percentage as low as 41% in a sample of 561 women who visited the clinic ($229/561 \approx 0.41$).

15 Identify the four steps of the statistical analysis process based on the study description above.

Steps in Statistical Analysis	**Study 1**
1. Ask a question that can be answered by collecting data.	
2. Decide what to measure and then collect data.	

[2]Tarun Jain et al., "Preimplantation Sex Selection Demand and Preferences in an Infertility Population," *Fertility and Sterility* 83, no. 3 (2005): 649-58.

3. Summarize and analyze the data.	
4. Draw a conclusion and communicate the results.	

LET'S SUMMARIZE

Please consider the following key points:

- The statistical analysis process involves the use of data.
- The process involves using evidence from a sample to make an inference about a larger population.
- The process entails asking a question that involves the use of data, collecting data, summarizing and analyzing data, and using data to draw a conclusion.
- To use sample evidence to conclude that a theory about a population is reasonable, we must rule out chance as a plausible explanation for the sample result.

EXERCISE 1.2

The statistical analysis process enables us to use data to make decisions about situations when we only have a limited amount of information. The statistical investigation we conducted in this collaboration involved using data from students in class to assess whether a person's chronotype can be predicted by their birth month. Using data from the class we made an inference about the relationship between birth month and chronotype for all people. Statistical investigations allow us to use data from small samples to make generalizations about much larger populations.

Here, we will use the statistical analysis process to investigate whether a coin is fair.

A student flips a coin ten times and finds that the coin landed "Heads" on eight of the ten flips. Does this result provide convincing evidence that the coin is **not** fair? In other words, does this result provide convincing evidence that the probability that the coin lands "Heads" is **not equal** to the probability that the coin lands "Tails". Answer the questions below to investigate this research question.

1 If a coin is fair (i.e. the probability that the coin lands "Heads" is equal to the probability that the coin lands "Tails"), what is the probability that the coin lands "Heads" on a single coin flip?

(i) $\frac{1}{5}$

(ii) $\frac{1}{4}$

(iii) $\frac{1}{3}$

(iv) $\frac{1}{2}$

2 If a coin is fair and is flipped 20 times, approximately what fraction of the coin flips do you expect will land heads? Choose the best answer below.

(i) $\frac{0}{20}$

(ii) $\frac{5}{20}$

(iii) $\frac{10}{20}$

(iv) $\frac{20}{20}$

3 If a coin is fair and is flipped 20 times, approximately what proportion of the coin flips do you expect will land *heads*? Choose the best answer below.

(i) 0.40

(ii) 0. 50

(iii) 0.65

(iv) 9.75

4 Using a simulation, you will examine the proportions of *heads* that are likely if a fair coin is flipped 20 times. You will now construct a dotplot of 100 sample proportions based on random trials of 20 coin flips. Follow the instructions below to construct the dotplot in Statkey.

- Open the Statkey simulator by going to http://www.lock5stat.com/StatKey/sampling_1_cat/sampling_1_cat.html
- Click on Edit Proportion and set the proportion to 0.5 (the probability of getting heads).
- Click on the Sample Size (n =) and set the sample size to 20.
- Click Generate 100 samples to generate and plot 100 sample proportions in the dotplot. The number of sample proportions will be shown at the top right of the viewing window. Each dot in the dotplot represents a sample proportion from a random sample of size 20.

What sample proportions are typical when a fair coin is flipped 20 times?

(i) 0.1 to 0.3

(ii) 0.3 to 0.7

(iii) 0.7 to 0.8

(iv) 0.6 to 0.9

5 The simulation displays proportions that are likely when a fair coin is tossed 20 times. The student conducting the experiment got heads 15 times in 20 flips. The proportion of heads was 0.75.

Is this an unusual value? Statisticians often view a value as unusual if it has a 5% chance (or less) of occurring.

Select the statement below that best describes how the proportion 0.75 relates to the sample proportions in the dotplot.

(i) 0.75 is typical because the value occurs more than once in the dotplot.

(ii) 0.75 is unusual because the proportions in the dotplot are not all equal to 0.75.

(iii) 0.75 is unusual because it is greater than 0.5.

(iv) 0.75 is unusual because it happens no more than 5 times out of 100 trials.

6 Which statement below is a reasonable conclusion based on comparing the sample result of 0.80 to the simulated dotplot?

(i) The outcome of Heads" in 10 flips is very unlikely, if the coin is fair. The observed sample result is unlikely to occur by chance alone, so there is evidence to conclude that the coin is fair.

(ii) The outcome of 8 "Heads" in 10 flips is very unlikely, if the coin is fair. This observed sample result is unlikely to occur by chance alone, so there is evidence to conclude that the coin is not fair.

(iii) The outcome of 8 "Heads" in 10 flips is very unlikely, if the coin is fair. Since this outcome is possible, when a coin is fair, we cannot conclude that the coin is not fair.

(iv) Since the coin was only tossed 10 times, we cannot make a decision about whether or not the coin is fair.

Read the following study description and answer the following questions:

Researchers wanted to know if people think a task will be hard to accomplish when the instructions are difficult to read.[3] To answer this question, researchers randomly divided twenty student volunteers into two groups of 10 students each. Researchers gave instructions to each group of students using different fonts (see below). Instructions for one group were written in a large upright font. The other group was given the *same* instructions but in a font that used *hard-to-read italics*. Researchers asked students to read the directions and say how many minutes they thought the task would take. Researchers did this in order to figure out if the fonts used for the instructions made a difference.

This is the easy-to-read upright font that was used in the study.

This is the hard-to-read italic font that was used in the study.

The first group of students, those that read the instructions printed in the easy font, had an average time estimate of 8.23 minutes. The other group, the group that read the instructions in the *hard-to-read italic* font, had an average time estimate of 15.1 minutes.

Researchers concluded that such a large difference between the averages was not likely to have occurred by chance. There was evidence that people think a task will be harder when instructions are difficult to read.

The four steps of the statistical analysis process are listed below:

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and then collect data.
3. Summarize and analyze the data.
4. Draw a conclusion and communicate the results.

[3]Hyunjin Song, "The Effects of Processing Fluency on Judgment and Processing Style: Three Essays on Effort Prediction, Risk Perception, and Distortion Detection" (PhD diss., The University of Michigan, 2009).

7 Which question below is a reasonable research question for this investigation?

(i) Do people like reading in different fonts?

(ii) Do people prefer reading one font to another font?

(iii) Do people think a task will be harder if the instructions for the task are harder to read?

(iv) Do people think that some instructions are easier to follow than other instructions?

8 What variables are used to answer the research question?

(i) Type of font & Amount of time a person thinks a task will take to complete.

(ii) Type of font & Amount of time a task takes to be completed.

(iii) Preferred font & Amount of time a person thinks a task will take to complete.

(iv) Preferred font & Amount of time a task takes to be completed.

9 How are the data summarized?

(i) Researchers compared the total amount of time that the two groups took to complete the tasks.

(ii) Researchers compared the average amount of time that the two groups took to complete the tasks.

(iii) Researchers compared the total amount of time that the two groups estimated that it would take to complete the tasks.

(iv) Researchers compared the average amount of time that the two groups estimated that it would take to complete the tasks.

10 What did the researchers conclude?

(i) The 10 students in the sample who read the instructions in the hard-to-read font took longer to complete the task than the 10 students who read the instructions in the easy-to-read font.

(ii) There is evidence that people will take a longer amount of time to complete a task when the instructions are harder to read.

(iii) The 10 students in the sample who read the instructions in the hard-to-read font think the task will be more difficult than the 10 students who read the instructions in the easy-to-read font.

(iv) There is evidence that people think a task will be harder when the instructions are harder to read.

Read the following study description and answer the following questions:

Healthcare professionals, such as doctors and physiotherapists, advise people to do regular exercise to prevent chronic illnesses, namely heart disease, obesity, and diabetes. Health experts recommend that a person needs to burn 150 to 400 calories per day doing exercise to stay healthy. For various reasons, many people do not engage in regular exercise. However, many activities of daily living, such as gardening, ironing, mopping, and mowing can be seen as forms of exercise, as you also burn calories while performing these tasks.

Researchers conducted a study amongst 150 adults of different body weights who engage in activities of daily living on a regular basis. The aim of the research study was to see if the number of calories burnt during these activities was different across various body weights.

The table below shows the data they collected, including the total number of adults (N) that performed each of the different activities during the research study and the number of calories burnt per activity by participants of various body weights.

The researchers concluded that adults of different body weights burn different amounts of calories while doing these activities.

		Body Weight									
Activity	N	50 kg	54 kg	59 kg	64 kg	68 kg	73 kg	77 kg	82 kg	86 kg	91 kg
		Calories burnt by body weight									
Gardening	61	99	108	117	126	135	144	153	162	171	180
Housework	130	99	108	117	126	135	144	153	162	171	180
Ironing	114	55	60	65	70	75	80	85	90	95	100
Mopping	70	94	102	111	119	128	136	144	153	162	170
Mowing	37	149	162	175	189	202	216	230	243	257	270
Shopping for groceries	146	66	72	78	84	90	96	102	108	114	120
Vacuuming	126	82	90	98	105	112	120	128	135	142	150

As a reminder, the four steps of the statistical analysis process are listed below:

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and then collect data.
3. Summarize and analyze the data.
4. Draw a conclusion and communicate the results.

11 Which question below is a reasonable research question for this investigation?

(i) Do people think that performing activities of daily living burns calories?

(ii) Do activities of daily living burn enough calories to be considered suitable exercise?

(iii) Does a person's body weight influence the amount of calories burnt by performing activities of daily living?

(iv) What is the average amount of calories burnt by people performing different activities of daily living?

12 What data did the researchers collect to answer the research question?

(i) The type of exercise the adults completed.

(ii) The total amount of calories the adults burned through exercising.

(iii) The weight of the adults.

(iv) All of the above.

13 How are the data summarized?

(i) Researchers found the proportion of adults who exercise by means of performing activities of daily living.

(ii) Researchers found the proportion of adults who perform different types of activities of daily living.

(iii) Researchers found the average amount of calories that adults burnt through performing different activities of daily living.

(iv) Researchers found the average amount of calories burnt for adults with different body weights when performing different activities of daily living.

14 What did the researchers conclude?

(i) Performing activities of daily living is the preferred type of exercise for many people.

(ii) Ironing does not appear to burn enough calories, but vacuuming would possibly be enough.

(iii) Adults of different body weights burn different amounts of calories while doing these activities.

(iv) The sample does not include a representative body weight range to make any reasonable conclusions about obese persons.

1.2 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Identify the steps in the statistical analysis process given a description of a statistical study.	
Interpret data displayed in a dotplot.	
Evaluate the strength of evidence against a claim about a population proportion.	

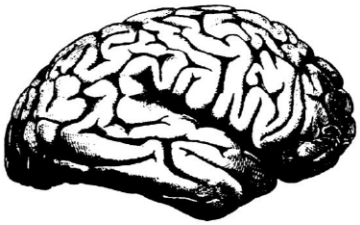
Growth Mindset Activity 1.3

YOU CAN GROW YOUR BRAIN

New Research Shows the Brain Can Be Developed Like a Muscle

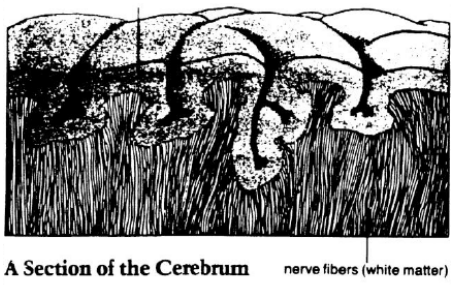
Many people think of the brain as a mystery. We don't often think about what intelligence is or how it works. And when you do think about what intelligence is, you might think that a person is born either smart, average, or dumb—either a "math person" or not—and stays that way for life.

But new research shows that the brain is more like a muscle—it changes and gets stronger when you use it. Scientists have been able to show just how the brain grows and gets stronger when you learn.



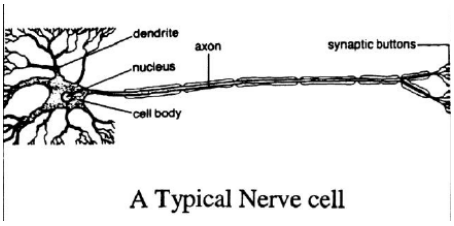
Everyone knows that when you lift weights, your muscles get bigger and you get stronger. A person who can't lift 20 pounds when they start exercising can get strong enough to lift 100 pounds after working out for a long time. That's because muscles become larger and stronger with exercise. And when you stop exercising, the muscles shrink and you get weaker. That's why people say "Use it or lose it!"

But most people don't know that when they practice and learn new things, parts of their brain change and get larger, a lot like the muscles do. This is true even for adults or older teenagers. So it's not true that some people are stuck being "not smart" or "not math people." You can improve your abilities a lot if you practice and use good strategies.



Inside the outside layer of the brain—called the cortex—are billions of tiny nerve cells, called neurons. The nerve cells have branches connecting them to other cells in a complicated network. Communication between these brain cells is what allows us to think and solve problems.

When you learn new things, you actually grow more of these tiny connections and they get stronger. The more you challenge your mind to learn, the more your brain cells grow. Then, things that you once found very hard even impossible to do—like understanding history, doing algebra, or completing complicated paperwork—become easier. The result is a stronger, smarter brain.



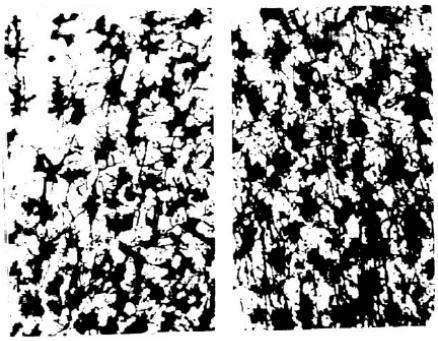
A Typical Nerve cell

or

How Do We Know That The Brain Can Grow Stronger?

Scientists started thinking the human brain could develop and change when they studied adult animals' brains. They found that animals who lived in a challenging environment, with other animals and toys to play with, were different from animals who lived alone in bare cages.

The animals who lived alone just ate and slept all the time, but the ones who lived with different toys and other animals were always active. They spent a lot of time figuring out how to use the toys and how to get along with other animals.



Nerves in brain of animal living in bare cage.

Brain of animal living with other animals and toys.

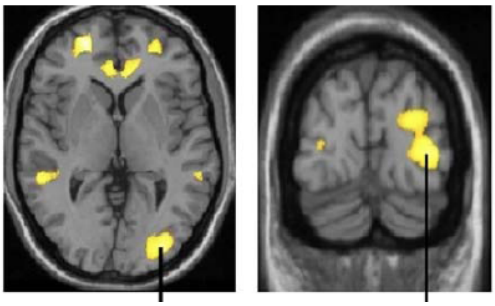
These animals had more connections between the nerve cells in their brains. The connections were bigger and stronger, too. In fact, their whole brains were about 10% heavier than the brains of the animals who lived alone without toys.

The adult animals who were exercising their brains by playing with toys and each other were also "smarter" –they were better at solving problems and learning new things.

Can Adults Grow Their Brains?

Scientists have recently shown that adults can grow the parts of their brains that control their abilities—like the ability to do math or even to juggle.

In one study, scientists found a group of adults who were not jugglers. They taught half how to practice juggling in the right way. These people practiced for long time and got much better at juggling. The other half didn't practice, and didn't get better.



In Yellow: Parts of the brain that grew when adults learned to juggle

doi:10.1371/journal.pone.0002669.g001

Next, the scientists used a brain scanner to compare the brains of the two groups of people. They found that the people who learned how to juggle actually grew the parts of their brains that control juggling skills. Their brains had changed, so they actually had more ability.

This was surprising because these people said before the study that they couldn't juggle—just like some people say they're "not good at math" or "not smart at school." But when they kept trying and learned good strategies for practicing, they actually learned and grew their brains.

Two Parts of Your Brain That Need to Grow:

The "Knowing" Part and the "Know How" Part

The brain doesn't just learn things; it also learns how to *do* things. Scientists have shown that people can grow two types of intelligence: the "knowing" type and the "know how" type. The "knowing" type of intelligence is what most people think about when they are learning new things in science or English or history. But you also need to learn skills that let you use your brain in a smarter way.

How do scientists know that you can grow the "know how" part of your brain? One example comes from the brains of taxi drivers in London, England. The streets of London are like a terrible maze. They are very hard to get around in. Brain scientists found that the longer the taxi drivers spent practicing driving in the streets of London, the more their brains grew. The "hippocampus" area of the brain—which controls memory for how to do everyday things—was actually larger! This means the more they practiced driving the confusing streets of London, the more the "know how" part of their brain grew. Although you might not realize it, we all grow the "know how" part of our brains whenever we learn to do new, hard things.

People often forget that we need to grow the "know how" part of our brains. If you study a long time but use a bad strategy, you may not learn—even if you try very hard. A few students study by doing the same set of easy problems and skipping the hard ones. Or they just re-read the textbook but don't try to challenge the ideas. After all, it's easier that way. But when it comes time to do the test, they don't do well because they didn't do things that grew their brains and taught them new things. When this happens, they may even say "I must not be smart at this."

But brain scientists have shown that a stronger brain doesn't just happen to you because you put in the time. You have to practice in the right way. If a weightlifter watched other people exercise all day long, he wouldn't get any stronger. You actually have to practice the right way—and usually that means the hard way—to get better at something. In fact, scientists have found that the brain grows more when you use new and different strategies. The brain grows less when you stick to things you already know.

How do you grow the "know how" part of your brain? You do it by seeking out help from others, like teachers or older students who do well. Or by going to the academic support center or getting tutoring. And then you practice successful strategies. At first it can be hard—sometimes people feel dumb for not knowing simple things like how to take notes, how to study, or even how to sign up for their classes. Or they feel not smart if they go to a tutor. But it's actually the opposite. Practicing better strategies makes your brain smarter.

The Truth About "Smart" and "Dumb"

People aren't "smart" or "dumb" in school. At first, no one can read or solve equations. But with practice, they can learn to do it. And the more a person learns, the easier it gets to learn new things. This is because the "knowing" and "know how" brain muscles have gotten stronger.

This is true for everyone, even adults or college students. Dr. Wittenberg, the scientist from Wake Forest University who did the research on juggling, said "We used to think adults can't form new brain connections, but now we know that isn't true… The adult brain is like a muscle, and we need to exercise it."

People who don't know this can miss out on the chance to grow a stronger brain. It does take work to learn, just like lifting more weights or becoming a better juggler does. Sometimes it even hurts! But when you feel yourself get better and stronger, you realize that all the work is worth it.

References:

A similar version of this article was written by Lisa Blackwell and can be downloaded from: www.brainology.us/websitemedia/**youcangrowyourintelligence**.pdf

Blackwell, L. A., Trzesniewski, K. H., & Dweck, C. S. (2007). Theories of intelligence and achievement across the junior high school transition: A longitudinal study and an intervention. *Child Development*, *78*, 246-263.

Driemeyer, J., Boyke, J., Gaser, C., Buchel, C., May, A. (2008). Changes in Gray Matter Induced by Learning—Revisited. *PLoS One, 3,* e2669. doi:10.1371/journal.pone.0002669.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, *97*(8), 4398-4403.

Nordqvist, C. (2004, Feb 1). "Juggling makes your brain bigger – New Study." Retrieved from http://www.medicalnewstoday.com/releases/5615.php

Questions: You Can Grow Your Brain

Questions

(1) Think about the article that you just read. What are all the reasons why scientists say that people's math ability can grow and get better with effort and practice?

Please summarize them briefly below.

(2) In the article, you learned three things:

- When you work hard and learn new things, your brain grows new connections and you get smarter.
- The more you challenge yourself, the smarter you will become.
- Smart kids are the kids who have practiced more—they have built up their reading and math "muscles."

Think about an example from your own life. What is something you weren't good at first but then practiced using a good strategy and became really good at it? Write about it and explain how you became good at it.

(3) Not all math students know that the brain can get smarter, even though it may help them have success. And we want to get your help so that we can learn more about how to explain it to them. We're hoping you can explain—in your own words—that the brain gets smarter when people use good strategies and try hard.

Imagine a friend who is struggling in school. This friend used to do pretty well in school but now is having a hard time and is starting to feel dumb. Write a letter to your friend to encourage him or her—tell them about what you just learned about the brain and why they shouldn't be discouraged. For example, you can tell them:

"Dear friend,
I realize that you feel discouraged in your math class because it is becoming a bit challenging for you, but after learning that our brain grows after practicing something hard I strongly encourage you to continue. I have learned that changing our studying strategies, asking for help, using resources, and practicing will help us improve what we do not understand. Life is a challenge and if everything was easy then the whole world would be boring. We have to grow our "know how" parts of our brain which requires a lot of effort. Our brain can become stronger and smarter by practicing what we do not know. So don't give up on your math class; challenge yourself."

Here are some other general ideas you can share with them:

A. How they can get smarter if they work hard and use a good strategy.
B. How they should work hard to build their reading and math muscles.
C. How they are not dumb, they just need to practice using a good strategy.
D. How they can ask the teacher or other students to help them learn better ways to study.
E. Any other tips you have for learning in school and getting smarter.

(Don't worry about writing a perfect final draft. We just want to know how you would say this to another student in your own words.)

1.3 Research Questions and Types of Statistical Studies

LEARNING GOALS

By the end of this collaboration, you should understand that:

- The types of conclusions we can draw from data are based on the methods in which data are collected.
- Statistical methods can be applied to a variety of disciplines.

By the end of this collaboration, you should be able to:

- Identify the research question in a given study.
- Distinguish between an observational study and an experiment.
- Identify the explanatory variable and response variable in an experiment.
- Identify variables of interest and population in an observational study.
- Explain why it is or is not reasonable to generalize conclusions from a particular statistical study.

INTRODUCTION

In Unit 1.2, we studied the four step process used in a statistical analysis. Step 1 in this process is "Ask a question that can be answered by collecting data." You will see in this lesson that understanding the type of research question being asked is very important. The type of research question has an impact on the method we use to collect data.

In a **statistical study**, a **population** is a set of all people or objects that share certain characteristics. A **sample** is a subset of the population used in the study. **Subjects** are the individuals or objects in the study. The subjects are often people, but can be animals, plants, or things. **Variables** are the characteristics of subjects that we study. For example, a variable might be eye color, age, educational level, salary or city. In Lesson 1.2, we asked whether a relationship existed between a person's birth month and chronotype. The population was all people. The sample was the students in our class and the variables were sets of personality traits and birth date groups.

The questions we ask in a statistical study are usually in one of two categories:

1. Research questions about a population.
2. Research questions about the **cause-and-effect relationship** between two variables. Note that a cause-and-effect relationship is a relationship between two *variables* where a known change in the first variable produces a predictable change in the second.

Research Questions about a Population

Type of Research Question	Examples
Make an estimate about the population (often about an average or proportion)	• What is the average amount of sleep university students get at night? • What percent of university students have jobs?
Test a claim about the population (often a claim about an average or proportion)	• Is the average amount of sleep for university students more than seven hours? • Do more than half of university plan to vote in the next election?
Compare two populations (often comparing averages or proportions)	• Do mainstream university students have higher average NBT scores than extended programme students? • Are extended programme university students more likely to receive financial aid than mainstream students?
Investigate a relationship between two variables	• Is there a relationship between the number of hours a full time student works at a job and his or her academic performance? • Are students who drink diet soda more likely to be overweight?

Research Questions about Cause-and-Effect Relationships between Variables

Examples of this type of research question include:

- Does requiring students to do homework in a university class improve test grades?
- Does caffeine reduce the risk of dementia (memory loss associated with old age)?
- Does taking an aspirin daily reduce the risk of heart attacks in adults over the age of 50?

To answer these research questions, we investigate how one variable responds as another variable is manipulated or changed. An **explanatory variable** is the variable being modified or manipulated in the study. A **response variable** is the output variable which is used to measure the impact of changes to the explanatory variable. An **experiment** involves a change to the explanatory variable.

TRY THESE 1

1 Here is one of the studies we examined in Unit 1.2. As you read the problem, think about the following statistical ideas: (1) the population of interest, (2) the variables being studied, and (3) the type of research question that is being asked.

> A group of researchers studied women who had visited a fertility clinic. The researchers wondered if fewer than half of the women who visit the clinic would want to choose the gender of their future child.[4]
>
> The researchers mailed a survey to women who had visited the clinic. The survey asked women if they would choose the gender of their future child, if they were able to do so. A total of 561 women responded to the survey. Of these 561 women, 229 said that they wanted to choose the gender of their future child.
>
> The researchers did statistical analysis on the data. Based on their statistical analysis of these data, the researchers concluded that there is **convincing evidence** that *fewer than half* of the women who visit the clinic would choose the gender of a future child. This conclusion is based on the following observation:
>
> If, in reality, at least half of women who visit a fertility clinic would like to choose the gender of a future child, it would be very unusual to observe a percentage as low as 41% in a sample of 561 women who visited the clinic (229/561 ≈ 0.41).

A What is the research question being asked?

B Does this study ask a question about a population or about a cause-and-effect relationship between two variables?

C What is the population of interest in this study?

D What variable is being examined for each subject in the study?

[4]Tarun Jain et al., "Preimplantation Sex Selection Demand and Preferences in an Infertility Population," *Fertility and Sterility* 83, no. 3 (2005): 649-58.

2 Here is the other study we examined in Unit 1.2. As you read the problem, think about the following statistical ideas: (1) the population of interest, (2) the variables being studied, and (3) the type of research question that is being asked.

Researchers wanted to know if people think a task will be hard to accomplish when the instructions are difficult to read.[5] To answer this question, researchers randomly divided twenty student volunteers into two groups of 10 students each. Researchers gave instructions to each group of students using different fonts (see below). Instructions for one group were written in a large upright font. The other group was given the *same* instructions but in a font that used *hard-to-read italics*. Researchers asked students to read the directions and say how many minutes they thought the task would take. Researchers did this in order to figure out if the fonts used for the instructions made a difference.

This is the easy-to-read upright font that was used in the study.

This is the hard-to-read italic font that was used in the study.

The first group of students, those that read the instructions printed in the easy font, had an average time estimate of 8.23 minutes. The other group, the group that read the instructions in the *hard-to-read italic* font, had an average time estimate of 15.1 minutes.
Researchers concluded that such a large difference between the averages was not likely to have occurred by chance. There was evidence that people think a task will be harder when instructions are difficult to read.

A What is the research question being asked?

B Does this study ask a question about a population or about a cause-and-effect relationship between two variables?

C What is the explanatory variable?

D What is the response variable?

[5]Hyunjin Song, "The Effects of Processing Fluency on Judgment and Processing Style: Three Essays on Effort Prediction, Risk Perception, and Distortion Detection" (PhD diss., The University of Michigan, 2009).

NEXT STEPS

When we know what type of question a study asks and what the variable(s) are in the study, then we can move on to the second step of a statistical analysis. Step 2 in this four-step process is "decide what to measure and then collect data." There are two main types of studies used to collect data: **observational studies** and **experiments**.

Observational Study

In an *observational study*, researchers observe subjects in a sample to learn about population characteristics. Researchers usually observe a sample of the population, since it is often impossible to obtain information from every member of the population.

Language Tip

A sample is *representative* of the population if it doesn't differ in any important way from the population.

Because the goal of an observational study is typically to learn about the population, it is important that the sample be **representative** of the population.

The first study, about women choosing the gender of their child, is an observational study. The researchers only asked the women about choosing the gender. They were *observing* the women's responses.

Experiment

An *experiment* is used to answer questions about how one variable responds when another is manipulated. In an experiment, researchers observe how a response variable behaves as an explanatory variable is changed. Researchers actively manipulate the explanatory variable.

Because the goal of an experiment is to learn about the effect of the different treatments (changes to explanatory variable) on the response variable, it is important to apply the different treatments to similar samples. Also, the results only apply to a larger population if the sample is representative of the population.

The second study above is an experiment. The researchers changed the font of the instructions for a task to see if it would change the amount of time the students thought the task would take.

Key Difference between an Observational Study and an Experiment

An important difference between these data collection methods is that there is no attempt to influence the results in an observational study. This is different from an experiment. In an experiment, the explanatory variable is manipulated and changes in the response variable are examined.

TRY THESE 2

3 Imagine that our university is having financial problems. The university announces that it will shorten library hours to save money. The library will be closed nights and weekends. Some students think that it is okay to pay a R270 fee to the university to keep the library open nights and weekends.

We are interested in learning about the proportion of students who would pay a R270 fee to keep the library open nights and weekends. To investigate this question, we select a sample of 100 students. We ask each of the students whether he or she agrees with the R270 fee increase to keep the library open nights and weekends.

A Does this study ask a question about a population or about a cause-and-effect relationship between two variables?

B Is this an observational study or an experiment?

If it is an *observational study*, what is the population of interest? What is the question we are asking about the population?

If it is an *experiment*, what is the explanatory variable and what is the response variable?

C It is important to obtain a sample of 100 students that is *representative* of all students at the university. What is an appropriate way of selecting a sample of 100 students from the population that ensures that it is representative? As you answer this question, be sure to think about (1) the population of interest, (2) your sample of 100 students, and (3) the type of research question that is being asked.

4 We are interested in learning whether jogging for longer amounts of time decreases the resting heart rate of university students. We want to see if there is a difference between:

- The resting heart rate of university students that jog for 30 minutes three times a week for six weeks, and
- The resting heart rate of university students that jog for 15 minutes three times a week for six weeks.

To investigate this question, we will use 100 university students who do not currently jog and who have volunteered to participate as subjects in this study. Resting heart rate of each subject will be measured at the start of the study. Fifty of the students will participate in a jogging program where they get together three times a week and jog for 30 minutes. The other 50 students will get together three times a week, but will only jog for 15 minutes. At the end of six weeks, resting heart rate will be measured again.

A What is the research question?

B Does this study ask a question about a population or about a cause-and-effect relationship between two variables?

C Is this an observational study or an experiment?

If it is an *observational study*, what is the population of interest? What is the question we are asking about the population?

If it is an *experiment*, what is the explanatory variable and what is the response variable?

D We need to divide the 100 volunteers into two groups so that there is a "fair" comparison between the 30 minute and 15 minute jogging groups. What would be an appropriate way to divide the 100 volunteers into two groups so that the two groups have similar volunteers?

NEXT STEPS

Drawing Conclusions from Statistical Studies

The fourth step in the statistical process is **drawing a conclusion and communicating the results**. When making a conclusion, researchers extend beyond the data that are observed to explain what they learned from the study.

There are two types of conclusions that might be made from a study. One type of conclusion is "**generalizing from a sample to the population**." Note that researchers always seek to study a **representative sample** of a larger population. When researchers draw this type of conclusion, they are confident that what they observed in the sample is true for the larger population.

The other type of conclusion is "**cause-and-effect**." This conclusion arises from an experiment when a change in a response variable was caused by the manipulation of an explanatory variable. If a researcher manipulates a variable and this change generates an "effect" or response, the researcher can conclude that the change was due to the variable manipulation.

Language Tip

A conclusion is *reasonable* when the study design allows us to make that type of conclusion.

The table below summarizes when each of these types of conclusions is **reasonable**.

Type of Conclusion	Reasonable When
Generalize from sample to population	Observational study is conducted and the sample is representative of the population
Cause-and-effect	Experiment is conducted and groups assigned to experimental conditions are similar

The best way to choose a sample that is representative of the population is to use a **random sample** from the entire population. The best way to ensure similar groups for different experimental conditions is to use random assignment to the experimental groups. In addition, if an experiment uses both random assignment and a random sample from a population, we can draw a cause-and-effect conclusion and apply it to the larger population.

We will see more about these ideas in upcoming lessons, but without a random sample in an observational study or random assignment in an experiment, no conclusions can reliably be drawn.

TRY THESE 3

5 In the USA, the SAT exam is used in admissions decisions by many four-year colleges and universities. This is similar to the use of NBT scores for admission decisions by South African universities. In 2006, The College Board in the USA carried out a study of 6,498 SAT essays that were selected at random from the more than 1.4 million SAT exams taken in the 2005 - 2006 academic year.[6] For this sample of essays, 15% were written in cursive and 85% were printed in block letters. The results showed that the average score for essays written in cursive was higher than the average score for essays that were printed.

A Is this study an observational study or an experiment? Explain.

B Is it reasonable to conclude that writing the essay in cursive was the cause of the higher scores? Explain your answer.

C Is it reasonable to conclude that the average score for essays written in cursive was higher than the average score for essays that were printed for all essays written in this testing period? Explain your answer.

[6]Krista D. Mattern, Wayne Camara, and Jennifer L. Kobrin, "SAT Writing: An Overview of Research and Psychometrics to Date," *collegeboard.org*, January 2007, https://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2007-32-sat-writing-research-psychometrics.pdf.

6 Imagine that a psychologist was interested in finding out whether music affects the ability to remember material that has been read. The psychologist recruited volunteer students who said they liked to study while listening to music and randomly assigned them into two groups. Each group was told to read an essay about the Global Financial Crisis of 2008/2009. One group read the essay in silence. The other group read the essay while music of a style of their choice played in the background. After reading the essay they took a brief test that asked the students to recall details about the essay. The psychologist concluded that students who listened to music while they were reading scored lower than students who read in silence.

A Is this study an observational study or an experiment? Explain.

B What are two possible reasons for why the students who listened to music scored lower than the students who didn't listen to music?

C The psychologist found that the difference was so large that it was unlikely due to chance variation alone. Is it reasonable to conclude that the music was the cause of the lower scores? Explain.

7 An article described a study in which researchers looked at a random sample of 500 publicly accessible web profiles posted by 18-year-olds on a social networking site.[7] The content of each profile was analyzed. Researchers drew many conclusions. One of the conclusions was that profiles that list a sport or hobby have fewer references to "risky behavior." Fewer references to "risky behaviors" means that there were fewer comments on the person's profile about sex, drugs, or violence.

A Is it reasonable to generalize the stated conclusion to all 18-year-olds with a publicly accessible web profile on the social networking site? What aspect of the study supports your answer?

B Not all social networking site users have a publicly accessible profile. Is it reasonable to generalize the conclusion "profiles that list a sport or hobby have fewer references to risky behavior" to all 18-year-old users on the site? Why or why not?

LET'S SUMMARIZE

- We can ask two types of research questions. Each type of research question is answered by a different type of study. Observational studies are used to answer research questions about characteristics of populations or comparing two populations. Experiments are used to answer research questions about cause-and-effect relationships.
- For observational studies, we should avoid making cause-and-effect conclusions, but it is possible to generalize from the sample to the population of interest if the study design incorporated random selection from the population.
- For experiments, it is possible to reach cause-and-effect conclusions if the study design uses random assignment to create the experimental groups.
- If an experiment uses both random assignment to create experimental groups and random selection from some population, it is possible to make cause-and-effect conclusions and to generalize these conclusions to the population.

[7]Megan A. Moreno et al., "Display of Health Risk Behaviors on MySpace by Adolescents: Prevalence and Associations," *Archives of Pediatrics & Adolescent Medicine* 163, no. 1 (2009): 27-34. http://www.ncbi.nlm.nih.gov/pubmed/19124700.

EXERCISE 1.3

"Sweet Potatoes Brighten Your Skin" is the headline of an article that appeared in the magazine *Woman's World* (November 1, 2010). The article concludes that eating sweet potatoes causes skin to be healthier because it reverses age spots, blocks harmful UV rays in sunlight, and protects against skin dryness. Consider the following four hypothetical study designs. For each study design, answer the questions that follow.

Study Design 1: Two hundred students were selected at random from those enrolled at a large university in the Western Cape. Each student in the sample was asked whether he or she ate sweet potatoes more than once in a typical week. A skin specialist rated skin health for each student on a scale of 1 to 10. It was concluded that skin health was significantly better on average for the group that reported eating sweet potatoes more than once a week than it was for the group that did not.

1 A Is the study described above an observational study or an experiment?

B Which phrase/sentence from the study design description helped you determine this?

(i) Two hundred students were selected at random.

(ii) Each student in the sample was asked whether he or she ate sweet potatoes.

(iii) A skin specialist rated skin health for each student on a scale of 1 to 10.

(iv) There is no phrase/sentence in the description that helps determine this.

2 A Did the study use random selection for some population?

B Which phrase/sentence from the study design description helped you determine this?

(i) Two hundred students were selected at random.

(ii) Each student in the sample was asked whether he or she ate sweet potatoes.

(iii) A skin specialist rated skin health for each student on a scale of 1 to 10.

(iv) There is no phrase/sentence in the description that helps determine this.

3 A Did the study use random assignment to experimental groups?

B Which phrase/sentence from the study design description helped you determine this?

(i) Two hundred students were selected at random.

(ii) Each student in the sample was asked whether he or she ate sweet potatoes.

(iii) A skin specialist rated skin health for each student on a scale of 1 to 10.

(iv) There is no phrase/sentence in the description that helps determine this.

4 Is the conclusion "eating sweet potatoes leads to healthier skin" reasonable given the study description?

(i) Yes, this used a large sample and random selection.

(ii) Yes, I already knew eating sweet potatoes is good for your skin.

(iii) No, the sample size is insufficient.

(iv) No, observational studies cannot determine cause-and-effect relationships.

5 Is it reasonable to generalize conclusions from this study to some larger population? If so, what population?

(i) Yes, this study involves a large sample and uses random selection. The sample represents the population, which is all university students in South Africa.

(ii) Yes, this study involves a large sample and uses random selection. The sample represents the population, which is all students enrolled at the university.

(iii) No, a sample of university students cannot represent any larger population.

(iv) No you cannot generalize conclusions from an observational study.

Study Design 2: One hundred people who live in Durban volunteered to participate in a statistical study. The volunteers were divided into two experimental groups based on gender, with females in group 1 and males in group 2. Those in group 1 were asked to eat 200g of sweet potatoes daily for three months. Those in group 2 were asked not to eat any sweet potatoes for three months. At the end of the three months, a skin specialist rated skin health on a scale of 1 to 10 for each of the volunteers. It was concluded that skin health was significantly better on average for group 1 than for group 2.

6 A Is the study described above an observational study or an experiment?

B Which phrase/sentence from the study design description helped you determine this?

(i) Those in group 1 were asked to eat 200g of sweet potatoes daily for three months. Those in group 2 were asked not to eat any sweet potatoes for three months.

(ii) One hundred people who live in Durban volunteered to participate in a statistical study.

(iii) It was concluded that skin health was significantly better on average for group 1 than for group 2.

(iv) There is no phrase/sentence in the description that helps determine this.

7 A Did the study use random selection from some population?

B Which phrase/sentence from the study design description helped you determine this?

(i) One hundred people who live in Durban volunteered to participate in a statistical study.

(ii) The volunteers were divided into two experimental groups based on gender.

(iii) Those in group 1 were asked to eat 200g of sweet potatoes daily for three months.

(iv) There is no phrase/sentence in the description that helps determine this.

8 A Did the study use random assignment to experimental groups?

B Which phrase/sentence from the study design description helped you determine this?

(i) The volunteers were divided into two experimental groups based on gender, with females in group 1 and males in group 2.

(ii) One hundred people who live in Durban volunteered to participate in a statistical study.

(iii) It was concluded that skin health was significantly better on average for group 1 than for group 2.

(iv) A skin specialist rated skin health on a scale of 1 to 10 for each of the volunteers.

9 Is the conclusion "eating sweet potatoes leads to healthier skin" reasonable given the study description?

(i) Yes, this used a large sample and random selection.

(ii) Yes, this was an experiment and cause-and-effect relationships can be determined.

(iii) No, the experiment had a flawed design and should not be used to determine a cause-and-effect relationship.

(iv) No, observational studies cannot determine cause-and-effect relationships.

10 Is it reasonable to generalize conclusions from this study to some larger population? If so, what population?

(i) Yes, this study involves a large sample and uses random selection. The sample represents the population, which is all people who live in Durban.

(ii) Yes, this study involves a large number of subjects and uses random assignment. The subjects represent the population, which is all people in South Africa.

(iii) No, there were not enough subjects in the experiment.

(iv) No, the experiment had a flawed design and is therefore not conclusive.

Study Design 3: One hundred people volunteered to participate in a statistical study. For each volunteer, a coin was tossed in order to place them into a group. If the coin landed head up, the volunteer was assigned to group 1. If the coin landed tail up, the volunteer was assigned to group 2. Those in group 1 were asked to eat 200g of sweet potatoes daily for three months. Those in group 2 were asked not to eat any sweet potatoes for three months. At the end of the three months, a skin specialist rated skin health on a scale of 1 to 10 for each of the volunteers. It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2.

11 A Is the study described above an observational study or an experiment?

B Which phrase/sentence in the study design description helped you determine this?

(i) Those in group 1 were asked to eat 200g of sweet potatoes daily for three months. Those in group 2 were asked not to eat any sweet potatoes for three months.

(ii) A coin was tossed in order to place them into a group.

(iii) If the coin landed tail up, the volunteer was assigned to group 2.

(iv) It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2.

12 A Did the study use random selection from some population?

B Which phrase/sentence in the study design description helped you determine this?

(i) One hundred people volunteered to participate in a statistical study.

(ii) For each volunteer, a coin was tossed in order to place them into a group.

(iii) If the coin landed head up, the volunteer was assigned to group 1.

(iv) If the coin landed tail up, the volunteer was assigned to group 2.

13 A Did the study use random assignment to experimental groups?

B Which phrase/sentence in the study design description helped you determine this?

(i) One hundred people volunteered to participate in a statistical study.

(ii) For each volunteer, a coin was tossed in order to place them into a group.

(iii) Those in group 2 were asked not to eat any sweet potatoes for three months.

(iv) It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2.

14 Is the conclusion "eating sweet potatoes leads to healthier skin" reasonable given the study description?

(i) Yes. While the subjects volunteered to participate in the experiment, random assignment was used and the experiment was well-designed.

(ii) Yes. Group 1 had a higher average than Group 2.

(iii) No, the sample size is insufficient.

(iv) No, we don't know where people in the sample are from.

15 Is it reasonable to generalize conclusions from this study to some larger population? If so, what population?

(i) Yes, this used a large sample and random selection. It represents the larger population.

(ii) Yes, this used a large sample and random assignment.

(iii) No. This study involved volunteers and we have no additional information about the participants. Therefore, we can't comfortably generalize these results to any larger population.

(iv) No, you cannot generalize conclusions from an observational study.

Study Design 4: One hundred students were selected at random from those enrolled at a large university. Each of the selected students was asked to participate in a study and all agreed. For each student, a coin was tossed in order to place them into one of two groups. If the coin landed head up, the student was assigned to group 1. If the coin landed tail up, the student was assigned to group 2. Students in group 1 were asked to eat 200g of sweet potatoes daily for three months. Students in group 2 were asked to not eat any sweet potatoes for three months. At the end of the three months, a skin specialist rated the health of each student's skin on a scale of 1 to 10. The researchers concluded that skin health was significantly better for the students in group 1 than for the students in group 2.

16 A Is the study described above an observational study or an experiment?

B Which phrase/sentence in the study design description helped you determine this?

(i) Students in group 1 were asked to eat 200g of sweet potatoes daily for three months. Students in group 2 were asked to not eat any sweet potatoes for three months.

(ii) Each of the selected students was asked to participate in a study and all agreed.

(iii) One hundred students were selected at random from those enrolled at a large university.

(iv) The researchers concluded that skin health was significantly better for the students in group 1 than for the students in group 2.

17 A Did the study use random selection from some population?

B Which phrase/sentence in the study design description helped you determine this?

(i) Each of the selected students was asked to participate in a study and all agreed.

(ii) One hundred students were selected at random from those enrolled at a large university

(iii) If the coin landed tail up, the student was assigned to group 2.

(iv) At the end of the three months, a skin specialist rated the health of each student's skin on a scale of 1 to 10.

18 A Did the study use random assignment to experimental groups?

B Which phrase/sentence in the study design description helped you determine this?

(i) One hundred students were selected at random from those enrolled at a large university.

(ii) Each of the selected students was asked to participate in a study and all agreed.

(iii) The researchers concluded that skin health was significantly better for the students in group 1 than for the students in group 2.

(iv) For each student, a coin was tossed in order to place them into one of two groups.

19 Is the conclusion "eating sweet potatoes leads to healthier skin" reasonable given the study description?

(i) Yes, this was a well-designed experiment.

(ii) Yes, researchers concluded skin health was significantly higher for Group 1.

(iii) No, these are all university students.

(iv) No, observational studies cannot determine cause-and-effect relationships.

20 Is it reasonable to generalize conclusions from this study to some larger population? If so, what population?

(i) Yes, this study used a large sample, random selection, and random assignment. The sample represents the population, which is all university students in South Africa.

(ii) Yes, this study used a large sample, random selection, and random assignment. The sample represents the population, which is all students enrolled at the university.

(iii) No, a sample of university students from a particular college cannot represent any larger population.

(iv) No. Three months is not a long enough time for the experiment to be conducted.

1.3 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Identify the research question in a given study.	
Distinguish between an observational study and an experiment.	
Identify the explanatory variable and response variable in an experiment.	
Identify variables of interest and population in an observational study.	
Explain why it is or is not reasonable to generalize conclusions from a particular statistical study.	

1.4 Random Sampling & Experiments

LEARNING GOALS

By the end of this collaboration, should understand that:

- Bias can sometimes be introduced by the way the sample is selected.
- The absence of bias is one characteristic of a good sampling plan.
- Random sampling helps protect against selection bias.
- The sample size plays an important role in determining the accuracy of a sample.
- The purpose of random assignment in the design of an experiment is to eliminate the influence of variables that we cannot control or do not know about. The purpose of a control group is for comparison with the treatment groups.
- The purpose of a placebo group is to rule out the placebo effect in an experiment.
- The purpose of blinding is to remove the influence of the participants' prior beliefs and the researchers' prior beliefs.

By the end of this collaboration, you should be able to:

- Describe a method for selecting a simple random sample.
- Given the design of a study or experiment, identify any biases in the design.
- Identify the response variable, explanatory variable and treatments, when given the description of an experiment.
- Describe a process for randomly assigning subjects to treatment groups.
- Explain the benefits of random assignment in the design of an experiment.
- Given the description of an experiment, identify whether random assignment and direct control were part of the experimental design.
- Identify blinding, the control group, and the placebo group in an experiment.
- Explain the importance of blinding, the control group, and the placebo group in an experiment.
- Describe characteristics of a well-designed experiment.

INTRODUCTION

In Unit 1.3, we learned that in order to generalize results from a sample to the population, the sample must be **representative** of the population. A sample is representative if it does not differ in any important way from the population. This collaboration explores sampling techniques that yield representative samples.

1 Suppose our university is thinking of ways to raise money. Many students like parking spaces close to their classes. The administration is thinking of selling reserved parking spaces for R1000. The university wants to know the percentage of students who would support this fee.

One way to find the percentage of students that support this fee would be to conduct a **census**. A census is a **survey** of an entire population. The university would ask every student on campus if she or he would support the fee.

Is this a reasonable plan? Why do you think so?

2 Read the following ways to sample students at our university. For each method:

- Tell whether the method would produce a sample that represents the student population.
- Explain why or why not.

A Choose four 8:00 a.m. classes at random. **Survey** all the students in each class.

B Put a **poll** on the front page of the university website. A poll is an opinion survey. Use the students who answer the question as the sample.

C Talk to students as they enter the Student Center.

3 None of the sampling methods above will produce a sample that is representative of the university's population of students. Suggest a better method to gather a representative sample.

NEXT STEPS

When we sample, our goal is for every population member to have the same chance of being selected. We want to avoid **selection bias**. This is the term used to describe a situation in which a sample differs from the population in some systematic way, so that some individuals are more likely to be selected than others. For example, imagine that we are surveying students and we choose only students entering the library. Not all students go to the library. The students in the sample may not be representative of the population of all students.

One good way to avoid *selection bias* is to select a **simple random sample.** In a simple random sample, all samples (of a given size) have the same chance of being chosen. One way to obtain a simple random sample is to put the name of every individual in the population on a piece of paper and put the names into a hat. Then we draw *n* names to include in our simple random sample of size *n*. In practice, a number is assigned to each individual and then random numbers are used to choose the sample.

Language Tip

Bias **is a systematic (not random) deviation from the true value.**

4 **Biased samples** result when sampling methods tend to leave out certain types of population members. The three previous sampling methods all produced biased samples. There are different types of biased samples.

A One type of biased sample is a **voluntary response sample**. Good samples are chosen by researchers. In a voluntary response sample, the participants are self-selected. In other words, each participant *chooses* to participate.

Which sample below is a voluntary response sample?

(i) Choose four 8:00 a.m. classes at random. Select all the students in these four classes for the sample.

(ii) Put a poll on the front page of the university website. Use the students who answer the question as the sample.

(iii) Talk to students as they enter the Student Center.

B Another biased sample is a **convenience sample**. Convenience sampling does not use random selection. It involves using an easily available or "convenient" group to form a sample. Many samples have convenience sampling problems. Which sample below is the best example of a convenience sample?

(i) Choose four 8:00 a.m. classes at random. Select all the students in these four classes for the sample.

(ii) Put a poll on the front page of the university website. Use the students who answer the question as the sample.

(iii) Talk to students as they enter the Student Center.

5 Suppose our university has 13,000 students. The university has the names and email addresses for all students in its database.

A Suggest a way that the administration could choose a simple random sample of 150 students to survey about the parking fee proposal.

B After the administration has chosen a sample, how could they actually conduct the survey?

C The administrator is told that the university can dedicate additional time and money to conducting this study. Would it make sense for the administrator to double the sample size and survey 300 students? Explain.

NEXT STEPS

Learning from Experiments

Previously, we discussed the four step process in a statistical study. The first step in this process is to ask a question that can be answered by collecting data. In this collaboration we will explore how an experiment could be used to answer the question: "Does listening to music make a difference in how much students remember while studying?"

Experiments often begin with a researcher stating a **hypothesis**. A **hypothesis** is a **conjecture**—an informed guess based on some information you have about the topic you want to study.

6 State a hypothesis about the effect that listening to music has on the amount of material a person remembers when studying. Do you think a person remembers *more*, *less*, or about *the same* as a student who studies in silence? Explain why you think this?

7 Suppose we think that listening to music while studying decreases the amount of information memorized. Design an experiment to test this hypothesis. Explain how you will collect and analyze data to make a decision about this hypothesis. Clearly describe the explanatory variable and the response variable in the experiment.

8 A Suppose a researcher designed the following experiment: In a class of 40 students, 20 were randomly assigned to a control group, and 20 were randomly assigned to a treatment group. Subjects in the control group read a paper for 10 minutes without listening to music, while subjects in the treatment group read the same paper for 10 minutes while listening to music. After reading the paper, all students completed a quiz on the contents of the paper. The quiz scored out of 10 points. The quiz results are shown below.

Quiz Scores: Listening to Music { 5, 6, 4, 5, 4, 6, 5, 4, 3, 4 }

Quiz Scores: Not Listening to Music { 7, 8, 5, 6, 7, 6, 8, 5, 7, 6 }

Find the average score of the students who listened to music.

Find the average score of the students who did not listen to music.

B How do the averages compare?

9 Think about what you know about a well-designed study. What are two possible reasons you might observe a difference between the students who studied while listening to music and the students who studied in silence?

Significance

When a difference between groups of a variable is so large that it is *unlikely* to be caused by chance, we say it is **significant.** In this case a significant difference would mean that we believe that listening to music was the reason for the difference in quiz score averages between the two groups. We do not yet have the tools to determine if the difference between averages is due to music listening or chance variation.

For now, we need to be careful. We shouldn't draw final conclusions yet. We need to rule out other possible causes. Think about other factors, or reasons, that could explain differences in quiz scores. For example, if some students took American History recently, they might get higher scores on the quiz.

Anecdotal Data

We should not draw conclusions based solely on information from a handful of people. For example, suppose one student states that they know from personal experience that they retain more information when they study while listening to music. This evidence may be legitimate, but it is not enough for a conclusion to be made. This is an example of anecdotal evidence. An anecdote, or a personal story, may be persuasive, but it only gives information about one individual.

Another example of anecdotal evidence would be observing a person who smokes a pack of cigarettes every day and does not get cancer. One person does not provide sufficient evidence. It would not be valid to conclude that people who smoke a pack of cigarettes every day do not get cancer.

NEXT STEPS

Direct Control, Random Assignment, Blinding, and the Placebo Effect

The goal of an experiment is to determine the effect of changing a treatment on the response variable. To do this, we want to rule out other possible explanations for differences in responses to treatments. Two common strategies to help with this are **direct control** and **random assignment.**

Direct control means that if you notice that there are other variables, *besides the explanatory variable*, that might affect the response, you try to manage those variables. **Random assignment** helps us create groups that are similar.

TRY THESE

10 Consider the following description of an experiment.[8] In this experiment, the researchers wanted to investigate whether the way people dry their hands after washing them has an effect on how clean their hands are.

> An experiment was conducted to compare bacteria reduction for three different hand drying methods: using paper towels, using a hot air dryer, and evaporation. In this experiment the participants handled uncooked chicken for 45 seconds. The participants then washed their hands with a single squirt of soap for 60 seconds and finally used one of the three hand drying methods. After participants dried their hands, the researchers measured the bacteria count on the participants' hands.

A One variable that might affect the response is the length of time that people handled the raw chicken. The researchers controlled this by having everyone handle the chicken for the same amount of time—45 seconds. Two other variables were controlled in this experiment. What are they?

B A **control group** is a group in an experiment that does not get a treatment. Including a control group in an experiment provides a basis for making comparisons. Is there a control group in this experiment? Explain.

Blinding

Sometimes people already have ideas about whether the treatments in an experiment will be effective. These beliefs might influence the response. When participants in an experiment do not know the type of treatment they are receiving, they are said to be **blinded**. Blinding participants is a way to prevent prior beliefs about the treatment from influencing their response.

[8] A.M. Snelling et al., "Comparative Evaluation of the Hygienic Efficacy of an Ultra-Rapid Hand Dryer vs. Conventional Warm Air Hand Dryers," *Journal of Applied Microbiology* 110, No. 1 (January 2011): 19-26. Details of this experiment were also reported in the online newsletter of the journal *Infectious Disease News* (September 22, 2010).

11 A shoe company wants to compare two different products for making hiking boots waterproof. Researchers create an experiment to compare two different products. In the experiment, researchers randomly give participants hiking boots. The hiking boots were waterproofed using one of the two waterproofing products. Participants then wear the boots on a hike along a mountain stream. They have to cross the stream and walk in the water in several places. After the hike, the boots are left to dry. Then the boots are sent to a lab where a technician evaluates the boots for water damage.

Do you recommend blinding only the participants, only the lab technician, or both the participants and lab technician? Explain your answer.

Placebo Effect

In experiments that use human participants, use of a control group may not be enough to establish whether a treatment really has an effect. Studies have shown that people sometimes respond in a positive way to treatments that have no active ingredients. These non-treatments that have no active ingredients may be colored water or sugar pills, and are called **placebos**. People often report that such non-treatments relieve pain or reduce other symptoms such as dizziness. Thinking that a non-treatment has helped the pain or made you feel better is an example of what is called the **placebo effect**.

Because of the placebo effect, experiments often include a control group that receives no treatment and another control group that receives a placebo. The placebo is identical in appearance (and taste, etc.) to what the people in other experimental groups receive. In such situations, participants should be blinded—we do not want people to know that they are receiving a placebo.

12 A researcher included a control group (with no treatment) and a placebo group in her experiment. How does including both groups allow her to decide when a placebo effect occurs?

13 How does including a placebo group in her experiment allow her to decide whether a particular treatment has a real effect on the response variable?

LET'S SUMMARIZE

Please consider the following key points:

- The goal of an experiment is to determine how changes to an explanatory variable effect a response variable.
- Direct control and random assignment are two strategies that allow us to rule out other possible explanations for changes to the response variable.
- Random assignment ensures that the treatment groups have similar characteristics. This helps to control for variables that we may not know about in advance and enables us to draw cause-and-effect conclusions about the effect of the explanatory variable on the response variable.
- Generally, we only draw cause-and-effect conclusions from experiments and not from observational studies.

EXERCISE 1.4

Collaboration 1.4 examined important characteristics of observational studies and experiments. The questions in Exercise 1.4 will assess your understanding of these concepts.

1 Imagine that you want to learn about the average number of hours, per day, that students at your university spend online. You want to select a simple random sample of 75 students from the full-time students at your university. You have a list of all full-time students, whose names are arranged in alphabetical order. Suppose that there are 5000 students at your university, of which 3500 are full-time.

How would you select a **simple random sample** of 75 students from this population?

(i) Select the first 75 students who you meet on campus for the sample.

(ii) Number all the students from 1 to 5000, use a random number table or calculator to randomly select 75 numbers between 1 and 5000, determine which students correspond to those numbers on the list, and select them for the sample.

(iii) Number all the full-time students from 1 to 3500, use a random number table or calculator to randomly select 75 numbers from1 to 3500, determine which students correspond to those numbers on the list, and select them for the sample.

(iv) Send an email to all 3500 full-time students. Use the first 75 students who respond to your email as the sample.

2 Imagine you want to estimate the average amount of time per week that students at a particular university spend studying. Which method below is an example of a **convenience sample**?

(i) Select 50 students who are enrolled in English literature at the university this semester.

(ii) Select 40 students as they enter the library.

(iii) Select 50 students who participate in student clubs and student activities.

(iv) Select 40 students at random from all students at the university by randomly choosing 40 student identification numbers.

3 Imagine you want to determine the locations on campus where students most prefer to go to study. Which method below is an example of a **voluntary response sample**?

(i) Email all students a link to an online survey. The sample would include all students who respond.

(ii) Survey all first-year students at a mandatory first-year orientation.

(iii) Select 60 students who are studying in the library during one week of the semester.

(iv) Email a random sample of 200 students a link to an online survey. Require their response before they are allowed to enroll in classes for the next term.

4 A researcher at King's College in London found that infomania (information overload) has a temporary, negative effect on intelligence quotient (IQ). Imagine that a group of researchers wished to further test this conclusion. These researchers divided volunteers into two groups. Each subject took an IQ test. One group had to check email and respond to instant messages while they were taking the test. The other group took the IQ test without any distractions. Researchers found that the distracted group's average IQ test score was 10 points lower than the average IQ test score for the group that was not distracted. What is the explanatory variable in this experiment?

(i) IQ test score

(ii) Number of email distractions

(iii) Whether or not the person was distracted with texts and emails

5 What is the response variable in this experiment?

(i) IQ test score

(ii) Number of email distractions

(iii) Whether or not the person was distracted with texts and emails

6 Why would it be good for the researchers to use random assignment to put each volunteer in one of the experimental groups?

(i) Random assignment places each volunteer in the treatment group that they prefer to be in. This helps to ensure that the volunteers will follow directions and participate in the entire experiment.

(ii) Random assignment helps to ensure that the treatment groups have similar characteristics.

(iii) Random assignment helps to ensure that the treatment groups will have different outcomes.

(iv) Random assignment ensures that the volunteers are unaware of which treatment group they are in so their prior beliefs and experiences don't influence the outcomes.

7 Which statement below accurately describes the control group?

(i) The control group was the group of volunteers who were distracted with text and email messaging while taking the IQ test.

(ii) The control group was the group of volunteers who took the IQ test without having any distractions.

(iii) The experiment did not involve a control group.

8 Is it possible for the subjects of this study to be blinded?

(i) Yes, since the volunteers will not know which volunteers are in their group or the other group.

(ii) Yes, since the volunteers will not know whether they are being distracted during the test.

(iii) No, since the volunteers will know that they are taking an IQ test and they will know whether or not they were checking emails and messages.

(iv) No, since the volunteers will know the conditions (distracted, not distracted) under which they took the test.

9 Researchers recently designed an experiment to investigate whether dogs can be trained to recognize cancer by smell. In the experiment, dogs were trained to distinguish between people with and without cancer by sniffing exhaled breath. Dogs were trained to lie down if they detected cancer. After training, the dogs' ability to detect cancer was tested using breath samples from an unfamiliar group of people. According to the article that presented the study's results, "The researchers blinded both [the] dog handlers and [the] experimental observers to the identity of [the] breath samples."[9]

Which statement below explains the meaning of the last sentence of the study?

(i) The dog handlers and experimental observers were unaware of each other.

(ii) The dog handlers and experimental observers were unaware of which dogs received training.

(iii) The dog handlers and experimental observers were unaware of which breath samples came from people who had cancer.

10 Why is blinding important for this experiment?

[9]Michael McCulloch et al., "Diagnostic Accuracy of Canine Scent Detection in Early- and Late-Stage Lung and Breast Cancers," *Integrative Cancer Therapies* 5 (2006): 30. For a summary of the study and its findings, see Linda Goldstein, "'Doctor Dogs' Diagnose Cancer by Sniffing It Out," *Knight Ridder Newspapers*, January 3, 2006, accessed July 5, 2014, http://www.azcentral.com/health/news/articles/0203doctordogs.html.

(i) If the dog handler or the observers knew which breath samples were from people who had cancer they could inadvertently pass some signals on to the dogs and influence the dogs' behavior.

(ii) If the dog handler or the observers knew which breath samples were from people who had cancer they could misinterpret the behavior of the dogs.

(iii) If the dog handler or the observers knew which breath samples were from people who had cancer they could tamper with the breath samples.

11 Which feature of the experiment is an example of direct control?

(i) The breath samples were administered to dogs randomly.

(ii) Dog handlers were randomly assigned to dogs to help dogs participate in the experiment.

(iii) Dog handlers were instructed to allow dogs to have the same amount of time to sniff the breath samples.

(iv) Observers were randomly assigned to observe the dogs' behaviors in the experiment.

1.4 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Identify the response variable, explanatory variable and treatments, when given the description of an experiment.	
Explain the benefits of random assignment in the design of an experiment.	
Explain the importance of blinding, the control group, and the placebo group in an experiment.	
Describe characteristics of a well-designed experiment.	

1.6 Distributions of Quantitative Data: Dotplots & Histograms

LEARNING GOALS

By the end of this collaboration, you should understand that:

- The process of describing distributions includes descriptions of the center, spread, and shape.
- These descriptions are useful for characterizing a distribution and comparing distributions.
- The number of data points within each bin can be used to construct frequency histograms. The proportion of data points within each bin can be used to construct relative frequency histograms.
- Though they look the same at a glance, there is a difference between frequency and relative frequency histograms—and that there are benefits of using each.
- Histograms can be used to characterize a distribution, specifically in terms of center, shape, and spread.

By the end of this collaboration, you should be able to:

- Estimate typical values and typical ranges of values from a dotplot.
- Use dotplots and histograms to compare data sets in terms of center shape and spread.
- Tally data values into bins, report the frequencies for each bin, and compute the relative frequencies for each bin.
- Describe the important characteristics of histograms in context.

INTRODUCTION

In Unit 1.2, you learned about the four steps in a statistical process:

1. **Ask a question that can be answered by collecting data.**
2. **Decide what to measure and then collect data.**
3. **Summarize and analyze the data.**
4. **Draw a conclusion and communicate the results.**

Language Tip
If something is *measurable*, it is large enough to be measured; noticeable; significant.

In this unit, we will learn how to use statistical tools to summarize and analyze data. We saw in previous units that **variables** in a statistical study are characteristics of the subjects in the study. We will now explore variables in greater detail.

Note: Throughout this course, a calculator may help with some problems and activities. When needed, help guides for TI-83/84 calculators are available here: https://carnegiemathpathways.org/go/SWhelpguides

Help for any calculator can usually be found by searching for any type of calculator, along with keywords for what you are trying to do.

In this unit, we will summarize and analyze **frequency distributions** of quantitative variables to investigate a question about sports. A frequency distribution of a variable provides two important facts about the variable:

Language Tip

Frequency **is the number of times a data value, or group of values, occurs.**

1. all of the values of the variable, and
2. how often (or how frequently) the variable takes on the values.

A **quantitative variable** can be measured or counted, and expressed as a number called a **data value**. Another type of variable is **categorical**, which cannot be measured or counted, but can be expressed as membership in a group called a category, such as eye color.

TRY THESE

In baseball, each team has their home field. For instance the New York Yankees have Yankee Stadium, in New York City. Every basketball team has their home court, and every hockey team has their home ice rink. Sports fans are familiar with the notion of "home field advantage" (or home court advantage in basketball, or home ice advantage in hockey). This means the team playing at its home venue has an advantage as a result of playing in a familiar place with supportive fans.

Consider the Golden State Warriors, a basketball team from the National Basketball Association (NBA). During their 2014-2015 regular season, the Golden State Warriors won 39 games and lost only two while playing on their home court. One sports writer wrote:

> *Warriors fans make Oracle Arena roar with their infectious energy and endless enthusiasm. No matter who the Warriors are playing, Dub Nation comes to Oracle and represents every night. They've had 102 straight sellouts, and that streak is likely to continue through this season and beyond*.[10]

But this is just one example. Perhaps the Warriors are an exception, and home court advantage is real for just a few lucky basketball teams. What about the other teams? In this collaboration, we will start by investigating home court advantage for basketball games. Later, we will look at hockey and baseball and soccer games as well.

[10]Maggie Pilloton, "Golden State Warriors Have Ultimate Home Court Advantage," *Golden Gate Sports.com*, January 23, 2015, accessed February 27, 2016, http://goldengatesports.com/2015/01/23/golden-state-warriors-ultimate-home-court-advantage/.

1 What are some possible reasons why the home basketball team might have an advantage over the visiting team?

2 If the home team has a real advantage over the visiting team, there would be a *measurable* effect. In this task, you will discuss the question, "Is there a home court advantage in basketball games?"

A What variables could help you answer this question? Be specific about what you would measure. Think about aspects of the basketball game that you could use as variables.

B Are the variables quantitative or categorical?

3 In the 2014-2015 season, when the Golden State Warriors played at home, they won 39 out of 41 games. When the Golden State Warriors played in other cities (known as "on the road" or "away" games), they won 28 out of 41 games. Calculate the home and away winning proportions as a decimal. Round your answers to two decimal places.

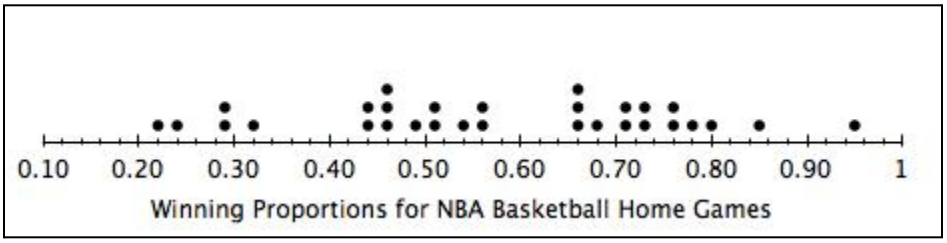
4 Do these results provide evidence of a home court advantage for the Warriors?

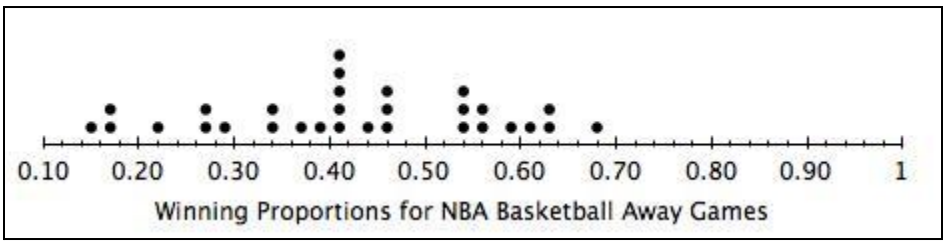
Comparing Distributions

Dotplots allow us to make quick visual comparisons between two different frequency distributions. In this case, we will compare two variables for all NBA teams in the 2014-2015 season:

- Home game winning proportion
- Away game winning proportion

5 Let's look at dotplots for all NBA teams:





A What does a dot represent in these dotplots?

B The home-game winning proportion of the Golden State Warriors was 0.95, and the away-game winning proportion of the Golden State Warriors was 0.68. Circle these dots on the dotplots above.

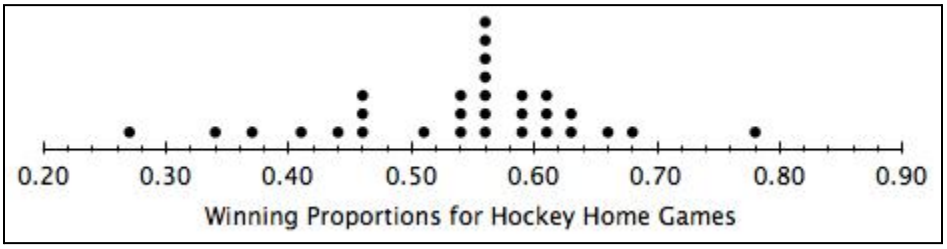
C Enough with the Warriors, because this is not a typical team. Let's look for overall trends now. Let's start with the home games. If you had to choose one number that represents a *typical* winning proportion of all the basketball teams, which would you pick? Write the proportion as a decimal rounded to two decimal places.

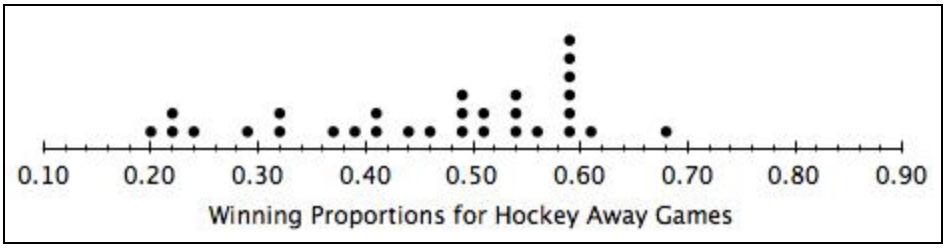
D Out of the away-games, if you have to choose one number that represents a typical winning proportion, which would you pick?

E Based on these dotplots, do you think that, in general, teams have a higher winning proportion when playing at home? How do the data support your answer?

F Does this data suggest that there is more variability in the winning proportion for home games, compared to away games? Variability refers to how spread out values are in a distribution. Explain your answer.

6 Now, let's look at the home game winning proportions in hockey, to see if there is a "home ice advantage":





A Do you think that, in general, the winning proportion for home games is greater than for away games? Explain.

B What other differences do you notice between the home and away game dotplots for hockey?

YOU NEED TO KNOW

Frequency Distribution of a Variable: A table or graph containing all values of a variable and how often the variable takes on each value.

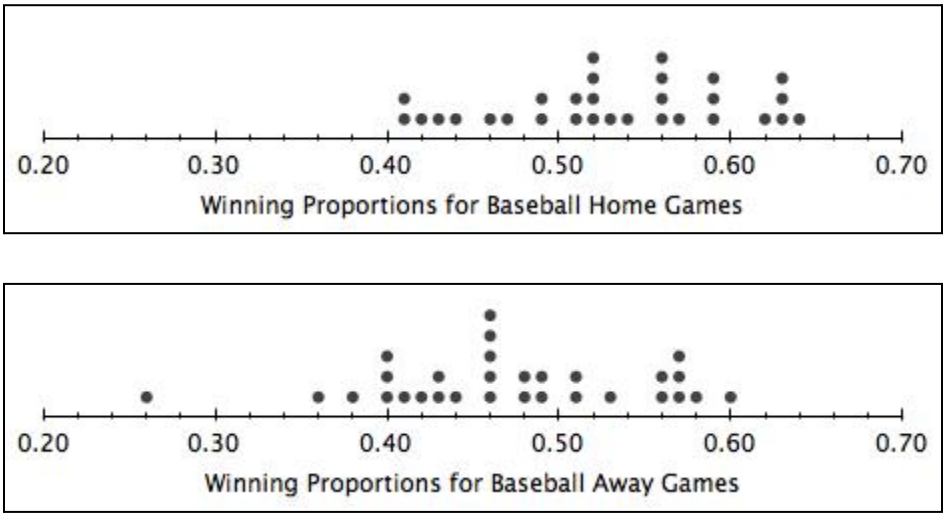
Center of a Distribution: The typical value or value which best represents the distribution.

Shape of a Distribution: A description of the overall pattern of the distribution.

Spread of a Distribution: The difference between the lowest value and the highest value.

TRY THESE 2

7 Below are dotplots showing the distribution of winning proportions of home and away games for baseball during the 2014 regular season.



Compare the centers, shapes and spreads of the distributions of winning proportions of home and road games. When describing the center of a distribution, specify one number that represents the typical value in the dotplot.

When considering "shape", one thing to think about is where the data is concentrated. Is most data to the right (on the low end)? In the center? To the left (on the high end)? Is data spread out evenly throughout?

	Home	Away
Shape		
Center		
Spread		

8 Based on these comparisons, would you say there is a home field advantage in baseball?

NEXT STEPS

Constructing Histograms for a Single Quantitative Data Set

We just analyzed data from the National Basketball Association (NBA). We used dotplots to determine if teams had a higher winning percentage at home or away.

We will now explore additional ways to graphically represent distributions of quantitative variables. We will continue using NBA data from the same season as above. However, we will now look at the *number of points* teams scored during home games and away games.

The following tables show how many points were scored by the home team and away team for a **sample** of games during the 2014-2015 season.

Sample of Scores from Away Teams

68	76	96	77	85
80	87	99	80	112
100	89	101	81	98
82	91	111	97	89
84	92	114	103	92

Sample of Scores from Home Teams

95	97	74	100	99
102	78	89	87	79
86	88	91	105	105
74	118	96	91	93
80	97	104	92	111

TRY THESE

9 To help us find patterns within the away teams data set, we will group data values into **bins** of equal width. Each bin is an interval that allows ten different data values (60 to 69 points, 70 to 79 points, etc.). The first bin starts a little lower than the lowest away score (the lowest score is 68). The bins are shown in the following table:

Language Tip
A *bin* is an interval that groups different data values.

Sample of Scores from Away Teams

Bin	Tally	Frequency
60-69	I	
70-79		
80-89		
90-99		
100-109		
110-119		

For each value in the away teams data table, determine the bin it falls into. For example, the first away score is 68, so it belongs in the bin with range 60 to 69. A **tally mark** (|) has been placed in the Tally column next to the bin 60 to 69. This tally represents the value of 68.

Language Tip

Tally marks **are used for counting. Each mark in a category represents a single occurrence (so ~~𝍸~~ II stands for 7 occurrences).**

Complete the table with the rest of the values in the away teams data set. Each time a tally reaches a fifth mark, represent it as a horizontal tally mark (~~𝍸~~).

The *frequency* is the number of data values contained in each bin. This is equal to the number of tally marks. In the table above, fill in the frequency for each bin. Do this by counting the tallies in each row and writing those numbers in the frequency column. Then compute the *relative frequency* by dividing the frequency by the sample size of 25.

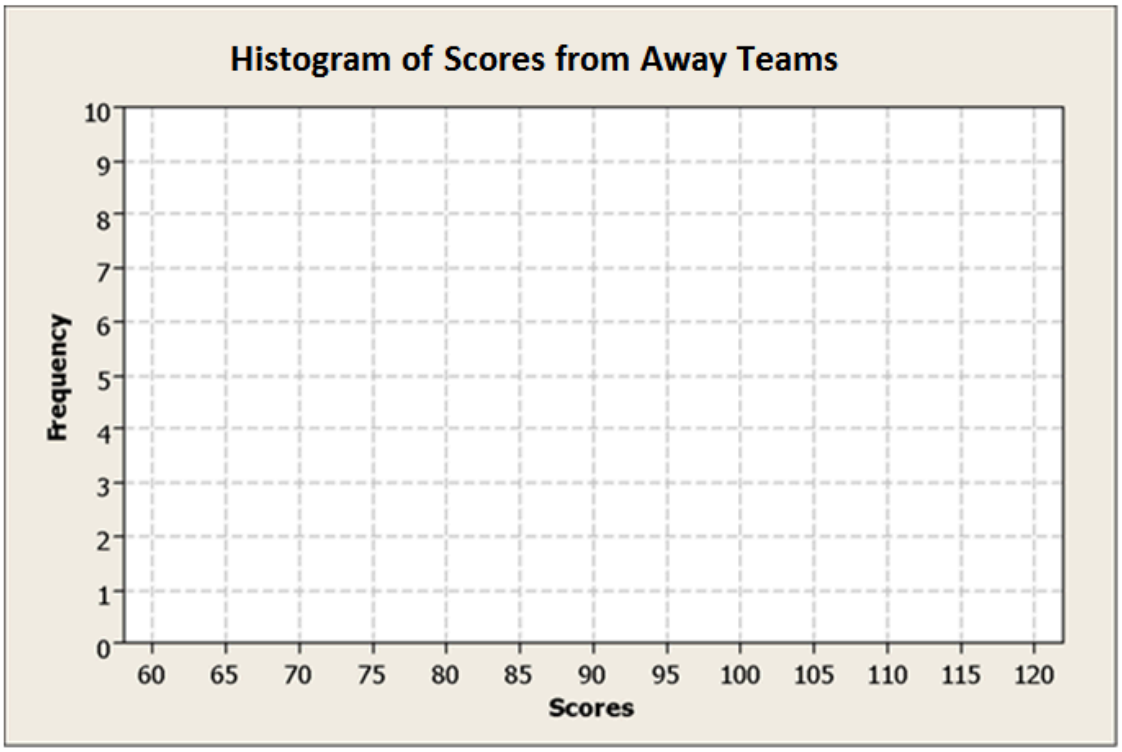
The table we created above is called a *frequency distribution table*. As we learned earlier, the *distribution* of a variable shows the possible values of the variable and how often the variable takes on each of those values. In our example, the variable was points scored. The frequency distribution table is one way to show the distribution of a variable.

A frequency distribution table contains equal-size bins, together with the frequency of each bin. The sum of the frequencies always equals the number of values in the data set.

Language Tip

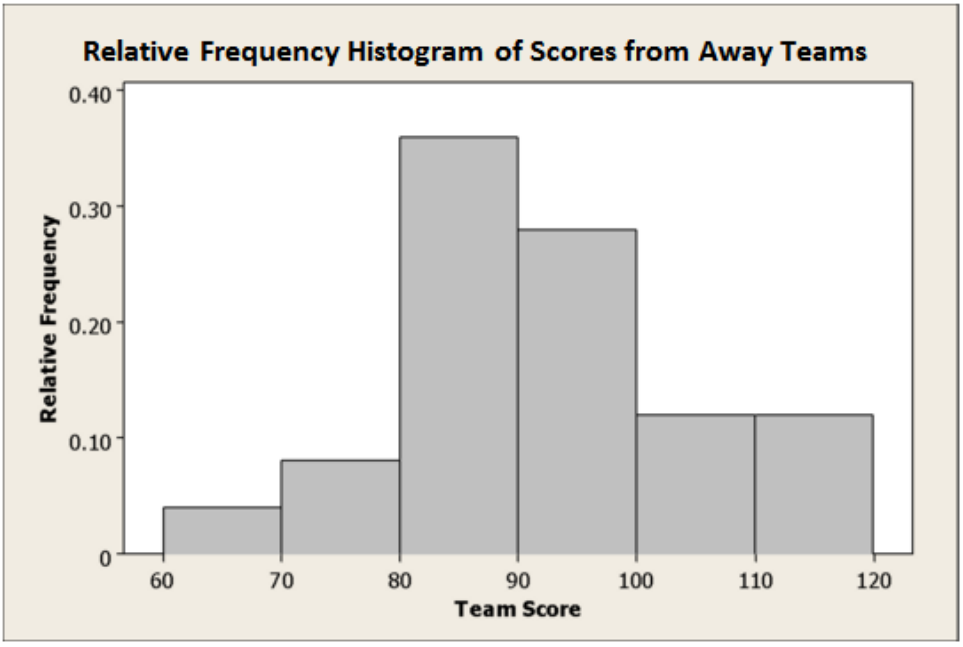
A *histogram* is a graph that uses bars to represent the bins in a frequency distribution table.

Frequency distributions can be displayed using a **histogram**. Each bin in the frequency distribution is represented by a vertical bar in the histogram. The height of the bar is the frequency of the bin. Draw the bars for each bin on the graph below.



10 Look at the bars of the histogram carefully. What is the sum of all of the frequencies? What does this sum represent?

11 A *relative frequency histogram* displays the relative frequencies for the bins instead of the frequencies. Look at your frequency histogram that you created and the relative frequency histogram below.



Do you notice any similarities or differences between the two histograms?

12 Compute and enter the relative frequencies by dividing the count for each bin by the total, 25. The 60 - 69 bin has been completed as an example.

Bin	Frequency	Relative Frequency
60 - 69	1	1/25 = 0.04
70 - 79		
80 - 89		
90 - 99		
100 - 109		
110 - 119		

13 Relative frequencies help us see what percentage of data values are in a range of values. What percentage of data values in the sample from the away scores are 90 points or higher?

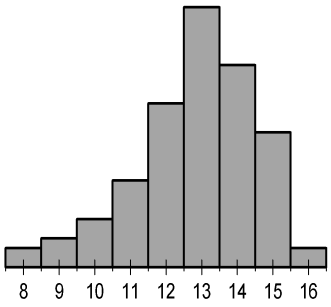
We use graphs, such as dotplots and histograms, to help us describe the distribution of a variable. Recall that three key characteristics we look at for a variable are center, spread and shape. The graph can also help us find potential **outliers**. An outlier is a data value much higher or lower than most of the other values.

Language Tip

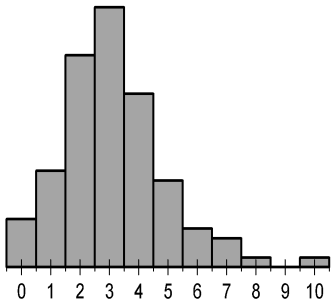
An *outlier* is a number that is separated from the rest of the data set by some distance.

Outliers are *extreme values*.

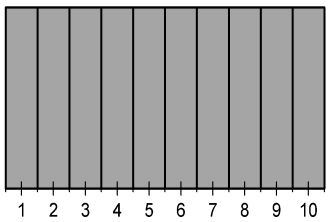
The histograms below display the shapes of different types of distributions. A **skewed** distribution is a *distribution* of a *quantitative variable* that is not symmetric about its center. A **uniform distribution** is a distribution of a variable in which every *data value* is equally likely to occur.



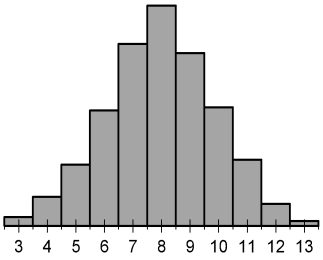
Skewed-Left Distribution



Skewed-Right Distribution



Uniform Distribution



Bell-Shaped Distribution

14 Use the frequency histogram of scores from away teams to answer the following questions:

A Estimate the **center** of the distribution.

B Describe the **spread** of the distribution.

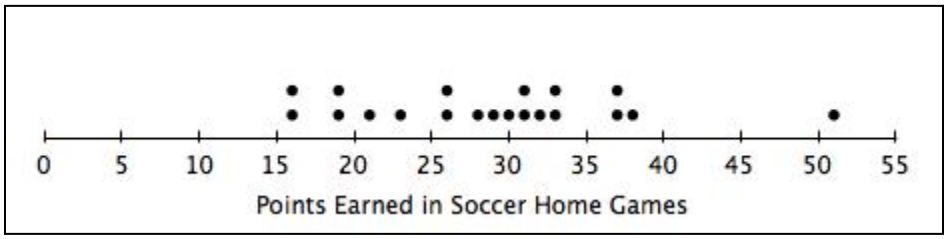
C How would you describe the **shape** of the distribution of scores from away teams? Would you say it is skewed-left, skewed-right, bell-shaped, or uniform?

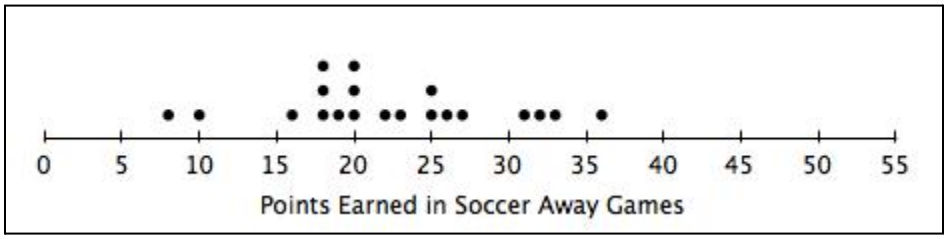
LET'S SUMMARIZE

- Center of a Distribution: The typical value or value which best represents the distribution.
- Shape of a Distribution: A description of the overall pattern of the distribution.
- Spread of a Distribution: The difference between the lowest value and the highest value.
- We can identify a data set's shape using modes and symmetry, find centers using the mean, median, and mode, calculate and compare spreads by identifying the IQR, and calculate and compare spreads using standard deviation.

EXERCISE 1.6

During the 2014–2015 soccer season, 20 teams competed with each other in the Serie A soccer league of the Italian Football Federation. Each team played 38 games, half at home and half away. A winning team earns three points, a losing team does not earn any points, and each team earns one point in the case of a draw. The dotplots below show the total home game points and total away game points of the 20 teams in Serie A in 2014–15.[11] Use these dotplots to help you answer the questions below.





1 Juventus, the Serie A champion in 2014-2015, won 16 games and had 3 draws in home games (without any losses). Compute the total points of Juventus in home games based on the point scheme described above and circle the dot that corresponds to that number of points in the dotplot above.

2 Another way to compare distributions is to compare the spreads of the distributions. The spread of a distribution describes the variability of values in the distribution. One way of computing the spread is to find the difference between the highest and lowest value. Another way is to focus on smaller areas, where most scores are concentrated. Find the minimum, maximum, and range for home games and away games.

[11]Source: The UEFA (European Football Association) website, http://www.uefa.com/.

3 You may notice that extreme values, like 51 on the home games dotplot, can make the range significantly larger. Focus only on areas where most data is concentrated. Provide a smaller range of typical values.

4 Did the range for the home game increase or decrease by excluding the most extreme score?

5 Let's return to the data from the National Basketball Association (NBA) we analyzed during the collaboration. Here is the list of scores from the sample of home teams. **Note**: If doing this problem online, follow the instructions given online to create your histogram.

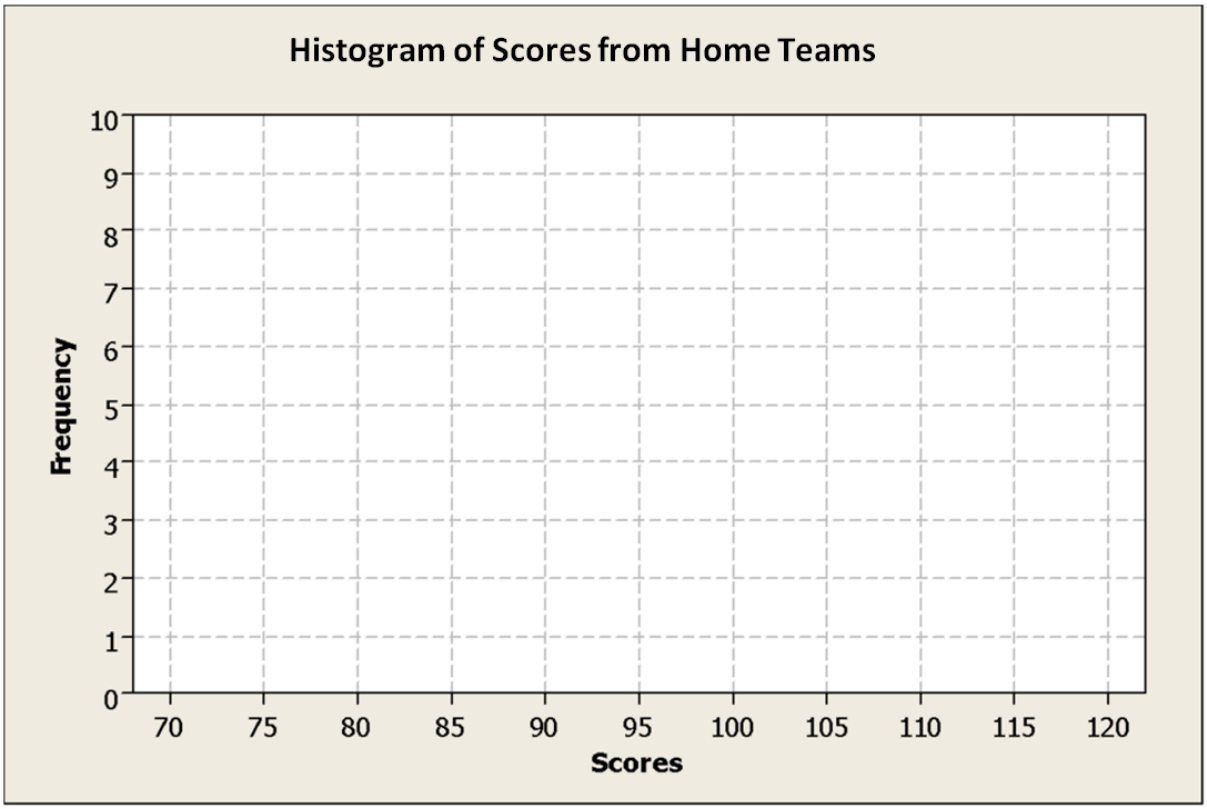
Sample of Scores from Home Teams

95	97	74	100	99
102	78	89	87	79
86	88	91	105	105
74	118	96	91	93
80	97	104	92	111

Complete the table below by finding the frequency and relative frequency of each bin.

Bin	Tally	Frequency	Relative Frequency
70-74			
75-79			
80-84			
85-89			
90-94			
95-99			
100-104			
105-109			
110-114			
115-119			

Draw the frequency histogram on the graph below.

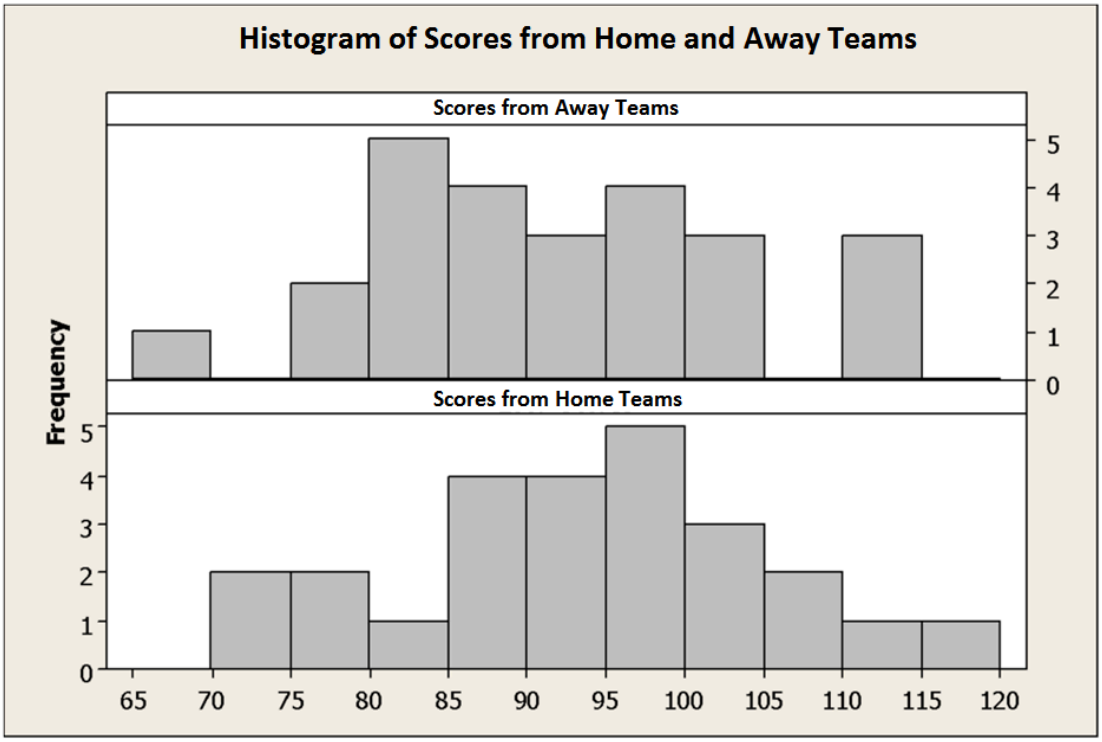


6 In how many games did the home team score at least 100 points?

7 Explain in your own words what you think the height of the second bar tells us.

8 Describe the center, shape, and spread of the frequency distribution.

9 Now let's compare the scores for away teams and the scores for the home teams in this sample. Below are the histograms for each data set:



Do the graphs seem to indicate that playing at home has an impact on teams' scores? Make sure to compare the two distributions in terms of center, spread, and shape.

1.6 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Estimate typical values and typical ranges of values from a dotplot.	
Use dotplots and histograms to compare data sets in terms of center shape and spread.	
Tally data values into bins, report the frequencies for each bin, and compute the relative frequencies for each bin.	
Describe the important characteristics of histograms in context.	

1.7 Quantifying the Center of a Distribution—Sample Mean and Sample Median

LEARNING GOALS

By the end of this collaboration, you should understand that:

- The center of a data set is an important and commonly studied characteristic.
- There is more than one way to quantify the center. Two commonly used measures are the mean and median.
- The characteristics of a data set affect which measure of center (median or mean) is most appropriate.

By the end of this collaboration, you should be able to:

- Approximate the values of the sample mean and sample median by visually examining a graph.
- Compute the values of the sample mean and sample median by hand or using technology.
- Compute the value of the sample mean from a frequency distribution table.
- Recognize the advantages and disadvantages of using the sample mean or sample median in different circumstances (i.e., given the characteristics of a given data set).

INTRODUCTION

Stimulants are sometimes used to help with weight loss but some can lead to weight gain. Some stimulants, like nicotine, can be dangerous. Others, like caffeine, are safer. In this activity, we examine the effect of a stimulant on the weight gains of a Treatment group of rats. These are compared to a Control group of rats who receive no stimulant treatment.

TRY THESE 1

Estimating the Center of a Data Set

Suppose we observe the following weight gains (in grams) for a Control group of six normal adolescent laboratory rats over a one-month period.

169 154 179 202 197 175

Now suppose we observe the following weight gains (in grams) for a Treatment group of six adolescent laboratory rats that were given a high daily dose of a stimulant:

137 158 153 147 168 147

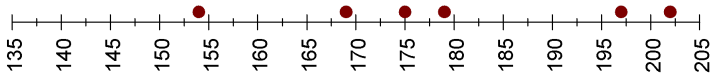
To determine whether there might be an effect on weight gain due to the stimulant, you will determine **representative values** of the two groups, namely the **sample mean** and **sample median**. We will explore the sample mean and sample median in the next few questions.

Below are dotplots for the Control and Treatment groups of rats.

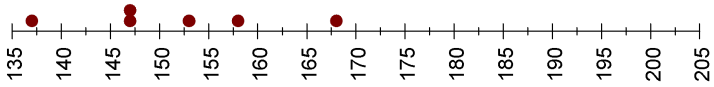
Language Tip

When we say *representative value*, we are talking about a central value that we use to represent a data set.

Control Group Weights



Treatment Group Weights



1 Imagine the dotplot as a scale that can tip left, tip right or stay balanced on both sides. Where do you think that the Control group's dotplot balances? That is, where on the number line would you set a balance point so that the distribution does not tilt to the left or to the right?

In statistics we refer to the average as the **mean**. The mean is the sum of data values divided by the number of data values. We usually calculate the mean using technology. Most often, we calculate the mean of a sample and refer to it as the **sample mean**. The symbol for the sample mean is $\bar{x}$ pronounced *x-bar*. For the Control group:

$$\bar{x} = \frac{\sum x}{n} = \frac{169 + 154 + 179 + 202 + 197 + 175}{6} \approx 179.3\ grams$$

Note: *n* is a variable we use throughout the course. It represents the number of items in a data set. Here, *n* = 6 because there are six rats in the Control group.

2 Calculate the mean weight gain for the rats in the Treatment group (the rats that were given the stimulant). Call this mean $\bar{y}$. Round your answer to the nearest tenth (one decimal place).

3 Compare the values of $\bar{x}$ and $\bar{y}$ (the sample means for weight gains in the Control and Treatment groups). Which sample mean is larger? Is the difference between the sample means enough to make you believe that the stimulant might have affected weight gain in adolescent rats? Explain your answer.

A **sample median** is the middle of a sorted list of data values. The following is the process for finding the sample median, applied to the Control group values:

- First, sort the data values in order from smallest to largest:

  154 169 175 179 197 202

- Notice that the middle of this list falls between 175 and 179:

  154 169 **175** **179** 197 202

This means that the sample median for the Control group is the average of the two values beside it, 175 and 179. That is,

$$\textit{median} = \frac{(175 + 179)}{2} = 177 \text{ grams}$$

Note: If there are an odd number of values, the median is the value in the middle of the sorted list. We do not use a symbol for the sample median.

4 Compute the sample median for the Treatment group and then summarize your results in the table below:

	Mean	**Median**
Control Group	179.3 g	177 g
Treatment Group		

5 Compare the median weight gains of the two groups. Which median is larger? Is the difference between the medians enough to make you believe that the stimulant might have affected weight gain in adolescent rats? Explain your answer.

6 Suppose we made a big mistake when we recorded the largest weight gain in the Treatment group. Instead of recording the weight gain as 168, we recorded it as 618.

137 158 153 147 **618** 147

A Recalculate the sample mean and sample median for the Treatment group with this new value. Enter your answer rounded to one decimal place.

B Recalculate the sample median for the Treatment group with this new value.

C How did this change impact the mean and median?

NEXT STEPS

When is the mean *not* a good choice?

Language Tip
When we say that a distribution is *skewed*, we mean that it is not symmetric, and one side has a longer *tail* of values.

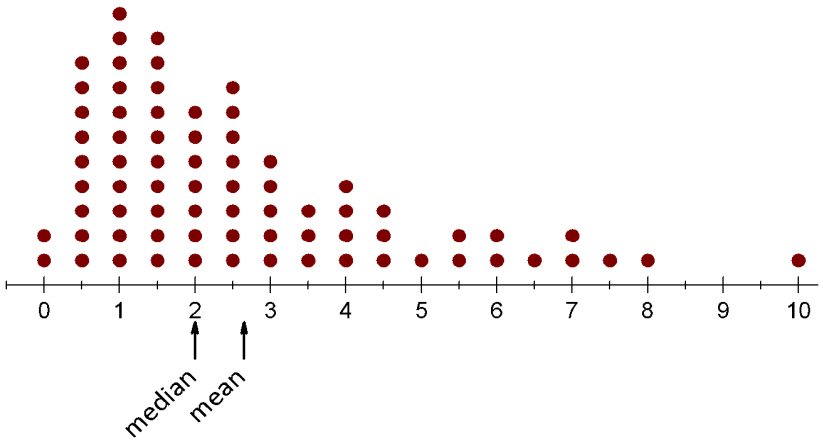
We call the mean and the median **measures of center.**

Measures of center are single values that represent an entire set. You've just seen that the mean can be strongly affected by extreme values. The mean is also strongly influenced by values that cause the distribution to be **skewed**. When a distribution is skewed, the mean is pulled toward the longer **tail**.

When data are **skewed to the right**, the mean tends to be greater than the median.

The values represented by the dotplot below are *skewed to the right*. Skewed to the right means that the longer *tail* is on the right. The median of these values is 2, but the skew pulls the mean to the right, and it is larger: $\bar{x}$= 2.65.

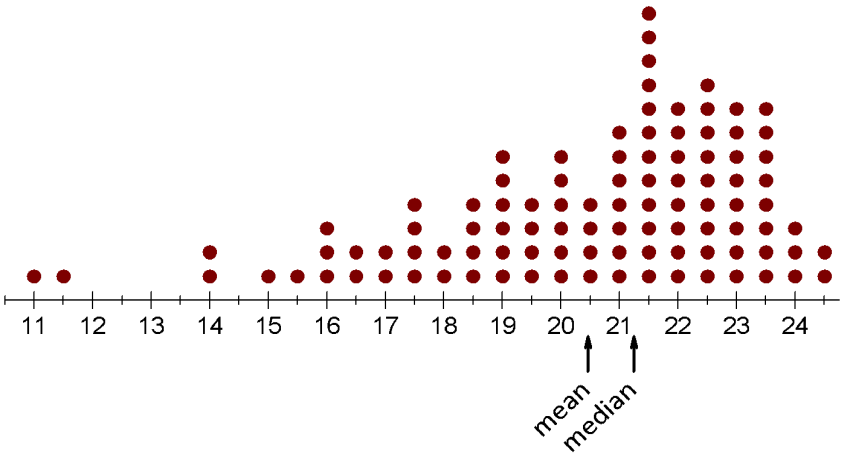
Skewed to the Right



When data are **skewed to the left**, the mean tends to be less than the median.

The distribution represented below is *skewed to the left*, so the longer *tail* is on the left. The mean ($\bar{x}$= 20.47) is less than the median (21.25).

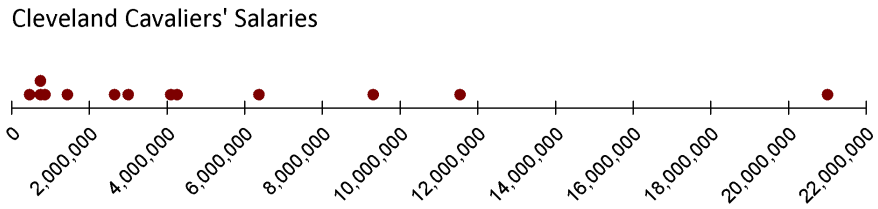
Skewed to the Left



7 Below are salaries of thirteen players on the Cleveland Cavaliers basketball team during the 2009–2010 season.[12] The 2009–2010 season was LeBron James' last season playing for the Cavaliers (until his later return to the team). Interestingly, LeBron James was the second highest paid player on the Cavaliers during that season. The highest paid player that season was Shaquille O'Neal.

$458,000	$4,089,000
$736,000	$4,254,000
$736,000	$6,364,000
$855,000	$9,300,000
$1,429,000	$11,541,000
$2,644,000	$21,000,000
$3,000,000	

A dotplot of these salaries is given below.



A Calculate the mean salary for the Cavaliers during 2009–2010. Round the mean to the nearest dollar.

B Calculate the median salary. Remember to sort the values first. Since there is an odd number of values, the median is the value in the center of the sorted list.

[12]Salary information can be found at http://espn.go.com/nba/salaries/_/year/2010/page/1. The Cavaliers roster for the start of the 2009–10 season can be found at http://www.basketball-reference.com/teams/CLE/2010.html.

C Would the *mean* or the *median* give a value that is most representative of the Cleveland Cavalier players' earnings in 2009–2010? Give your reasoning.

8 How does Shaquille O'Neal's salary impact the mean and median? We can examine this question by computing the mean and median salary without Shaquille O'Neal's salary included:

$458,000	$4,089,000
$736,000	$4,254,000
$736,000	$6,364,000
$855,000	$9,300,000
$1,429,000	$11,541,000
$2,644,000	~~$21,000,000~~
$3,000,000	

A Calculate the mean with Shaquille O'Neal's salary removed. Round the mean to the nearest dollar.

B Calculate the median salary with Shaquille O'Neal's salary removed.

C What effect did removing Shaquille O'Neal's salary have on the mean and median?

YOU NEED TO KNOW

The median is the best measure of center for data that are skewed. This is because, unlike the mean, extreme values which cause skewing have little impact on the median. The median is resistant to the extreme values but the mean is *not* resistant to these extreme values.

LET'S SUMMARIZE

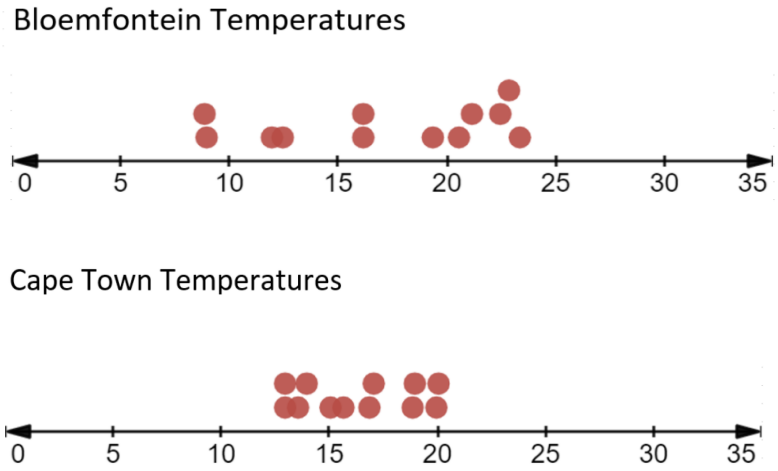
- The mean and median provide us values that are representative of an entire data set.
- The mean is calculated using all of the values in a data set.
- The median is based on the number, or numbers, in the middle of a data set.
- The median is the best measure of center for data that are skewed. This is because, unlike the mean, extreme values which cause skewing have little impact on the median.
- The median is resistant to the extreme values but the mean is not resistant to these extreme values.

EXERCISE 1.7

The table below lists average monthly temperatures in Bloemfontein and Cape Town.[13]

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Bloemfontein	23.4	22.5	20.6	16.2	12.5	9	8.9	12	16.2	19.4	21.2	22.9
Cape Town	20	20.1	18.9	16.9	15.1	13.6	13	13	14	15.7	17.1	19

These data are summarized in dotplots below.



1 Compute the sample mean of the typical monthly temperatures for Bloemfontein. Use the symbol $\bar{x}$ for Bloemfontein's' sample mean. Round your answer to one decimal place.

2 Compute the sample mean of the typical monthly temperatures for Cape Town. Use the symbol $\bar{y}$ for Cape Town's sample mean. Round your answer to one decimal place.

[13]Bloemfontein weather by month.. Source:
https://en.climate-data.org/africa/south-africa/free-state/bloemfontein-394/, accessed April 6 2021. Cape Town weather by month. Source:
https://en.climate-data.org/africa/south-africa/western-cape/cape-town-788/, accessed April 6 2021.

3 Compute the sample median of the monthly temperatures for Bloemfontein. Round your answer to one decimal place.

4 Compute the sample median of the monthly temperatures for Cape Town.

5 Which of the statements below best compares the temperatures of the two cities?

(i) The two cities are totally different in regards to temperature.

(ii) The mean and median of Bloemfontein are slightly higher, but the most extreme temperatures are the same.

(iii) The mean and median of Bloemfontein are slightly higher and the most extreme temperatures are quite different.

6 Students in a Statistics course took a test where the highest score was 100. Most of the students did quite well, with more than half of the students getting As and Bs on the test. Most of the rest of the students earned Cs, with just a few Ds. Only two students received Fs, but their scores were very low, in the 30s.

If As are for scores in the 90s, Bs for the 80s, Cs for the 70s, Ds for the 60s, and Fs for scores below 60, what is the shape of the distribution of scores? You may want to draw a quick picture based on the description above.

7 Is the mean likely to be greater than or less than the median? Explain your answer.

8 Students frequently want to know the "average" test score. What measure of center should the professor share with the class, the mean or the median? Think about which measure will give a better representation of typical student scores. Give a reason for your answer.

A frequency distribution for the ages of students in a Statway® class is given below.

***x* = age**	***f* = frequency**
18	4
19	5
20	7
21	3
22	4
23	2
24	0
25	1
26	1

Rather than finding the mean by adding up every single value and dividing the sum by the total number of values, we can instead find the mean of this *grouped data* by using the frequencies. If we multiply *each x* by its frequency, *f*, and add the answers, we get the sum of all values. We can determine the sample size by adding the frequencies. Therefore, the formula for the sample mean is:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

The table below has a 3rd column which contains the product of frequency and age. The value "72" in the first row represents the sum of the ages of the four 18-year olds.

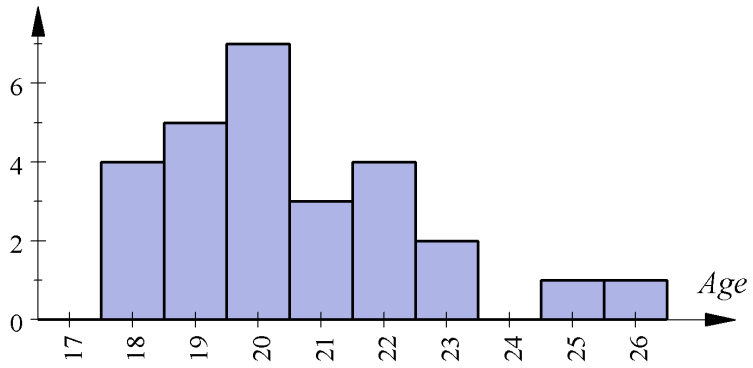
***x* = age**	***f* = frequency**	***f* · *x***
18	4	72
19	5	95
20	7	140
21	3	63
22	4	88
23	2	46
24	0	0
25	1	25
26	1	26

9 How many students are in the class?

10 What is the sum of all the ages?

11 Compute the mean age of the students. Round your answer to one decimal place. Use the formula for finding the sample mean which is displayed above.

12 A bar chart of the data is shown below.



Is the mean likely to be greater than or less than the median? Explain your answer.

1.7 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Approximate the values of the sample mean and sample median by visually examining a graph.	
Compute the values of the sample mean and sample median by hand or using technology.	
Compute the value of the sample mean from a frequency distribution table.	
Recognize the advantages and disadvantages of using the sample mean or sample median in different circumstances (i.e., given the characteristics of a given data set).	

1.8 Quantifying Variability Relative to the Median

LEARNING GOALS

By the end of this collaboration, you should understand that:

- Quartiles are dividers that separate a sorted data set into four equal parts.
- The five-number summary is a simple way to summarize data in order to capture multiple characteristics.
- Boxplots are a graphical representation of the five-number summary.

By the end of this collaboration, you should be able to:

- Compute the values in the five-number summary by hand or using technology.
- Compute the interquartile range (IQR).
- Construct boxplots, including side-by-side boxplots.
- Use the IQR to identify outliers.

INTRODUCTION

In the **Exercise** of the previous collaboration, we looked at temperatures for the cities of Bloemfontein and Cape Town. We compared typical monthly temperatures for the two cities. These temperatures are presented in the following table.

Typical Monthly Temperatures (°C) for Bloemfontein and Cape Town

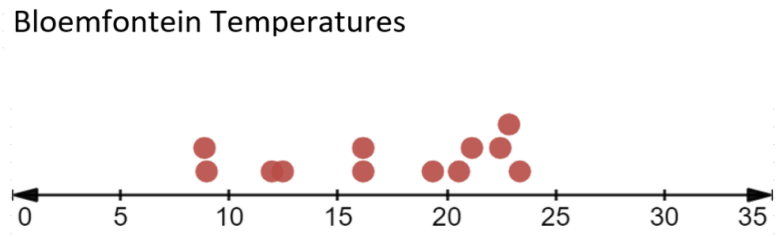
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Bloemfontein	23.4	22.5	20.6	16.2	12.5	9	8.9	12	16.2	19.4	21.2	22.9
Cape Town	20	20.1	18.9	16.9	15.1	13.6	13	13	14	15.7	17.1	19

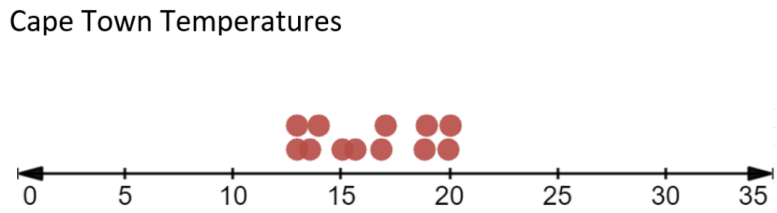
In our previous collaboration, we examined the *center* of these distributions by calculating the mean and median temperature for each city. We found these measures to be very similar for the two cities, even though their distributions are quite different. In this collaboration, we study the **variability** (or **spread**) in the data sets.

Language Tip
Variability **or** ***spread*** **means how widely the data points are scattered.**

VARIABILITY

1 The dotplots below show the typical monthly temperatures for Bloemfontein and Cape Town. Examine the dotplots. Record a few sentences that compare the typical monthly temperatures for the two cities. In your answer, mention how the distributions are similar and how they are different.





2 In previous collaborations, we summarized quantitative data in *dotplots* and *histograms.* We summarized the *center* of a distribution using the *sample mean* and *median*.

Describing the center of a distribution is important, but it is not enough. Look at the temperatures for the two cities. The centers of the distributions are similar. But one distribution has more *variability* than the other distribution.

Sometimes particular values can help us describe variability in the data sets.

A Think of *a single number* to describe the variability in the Bloemfontein temperatures. What would it be?

B What single number would you use to describe the variability in the Cape Town temperatures?

C The values you chose can be thought of as "measures of variability in temperature." Do your measures of variability in temperature for Bloemfontein and Cape Town show the differences in their distributions? Which sample has greater variability?

NEXT STEPS

3 One way to represent the variability in data is with the **range**. The range is the difference between the *maximum* and *minimum* data values. That is, the range is the *largest* value minus the *smallest* value.

$$range = maximum - minimum$$

A With reference to the table of "Typical Monthly Temperatures for Bloemfontein and Cape Town," find the minimum value, the maximum value, and calculate the range in monthly temperatures for Bloemfontein and Cape Town. Complete the table below.

Typical Monthly Temperatures (°C) for Bloemfontein and Cape Town

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Bloemfontein	23.4	22.5	20.6	16.2	12.5	9	8.9	12	16.2	19.4	21.2	22.9
Cape Town	20	20.1	18.9	16.9	15.1	13.6	13	13	14	15.7	17.1	19

	Bloemfontein	Cape Town
Minimum		
Maximum		
Range		

B Which city has the larger temperature range?

One problem with the range is that it is influenced by *outliers*. That is, the range always uses the most *extreme* data values. Extreme values do not represent the data well. Another problem is that the range depends only on *two* values. The rest of the values are ignored.

Another way to describe variability is with **quartiles**. When a data set is sorted from the lowest values to the highest, the quartiles are the *dividers* that separate the data set into four equal parts.

Here are the sorted monthly temperatures for Bloemfontein:

Bloemfontein	8.9	9	12	12.5	16.2	16.2	19.4	20.6	21.2	22.5	22.9	23.4

To find the quartiles, we divide the data into four equal parts. Since there are 12 observations for Bloemfontein, each of the four parts contains 12/4 = 3 observations. The dividers in the following table separate the four equal parts.

Bloemfontein	8.9	9	12	12.5	16.2	16.2	19.4	20.6	21.2	22.5	22.9	23.4

The *median* is the middle divider, and is the *second quartile*. The value of the median is (16.2 + 19.4)/2 = 17.8 °C.

The first divider occurs between the values 12 and 12.5. This is the **first quartile**, or ***Q1***. In this example

Q1 = (12 + 12.5)/2 = 12.25 °C.

Also, look at just the lower half of the values (16.2 and below in this example). The first quartile (*Q1*) represents the median of the lower half of values in the data set.

The third divider occurs between the values of 21.2 and 22.5 °C. This **third quartile** is ***Q3*** = (21.2 + 22.5)/2 = 21.9 °C. The third quartile represents the median of the upper half of values in the data set.

The minimum, maximum, and three quartiles form the **five-number summary**. Here is the five-number summary for the monthly temperatures in Bloemfontein.

Bloemfontein

Minimum	8.9
Q1 (first quartile)	12.25
Median	17.8
Q3 (third quartile)	21.9
Maximum	23.4

4 Use the typical monthly temperatures in Cape Town to construct a five-number summary. Record the values below.

Cape Town	13	13	13.6	14	15.1	15.7	16.9	17.1	18.9	19	20	20.1

Cape Town

Minimum	
Q1 (first quartile)	
Median	
Q3 (third quartile)	
Maximum	

YOU NEED TO KNOW

Sometimes a data set has an *odd* number of values. When this happens, the middle value is not included in the upper or lower half of the data set.

5 Another way to describe variability in a data set is to find the distance between the first and third quartiles (Q1 and Q3). This distance is the **interquartile range**, abbreviated as **IQR**. The formula is:

IQR = Q3 – Q1

The IQR gives the *range* of the middle 50%.

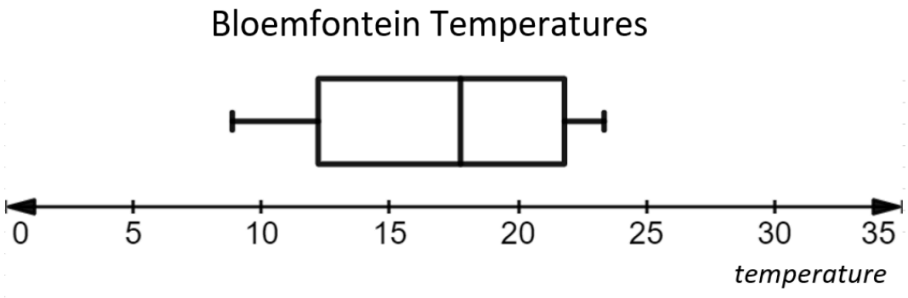
For the typical monthly temperatures in Bloemfontein, the IQR is Q3 – Q1 = 21.9 – 12.25 = 9.65 °C.

Calculate the IQR for the typical monthly temperatures in Cape Town. Round your answer to 2 decimal places.

6 Compare the IQRs for the typical monthly temperatures in Bloemfontein and Cape Town. Consider also what you observed by looking at the dotplots for the two cities in Question 1. Comment on how the measures of IQR and range describe differences in variability of the temperatures in these two cities.

The values in a five-number summary can be represented in a graph called a **boxplot** (sometimes called a *box and whiskers plot*).

Here is a boxplot for the typical monthly temperatures for Bloemfontein:



7 The five-number summary for the Bloemfontein temperatures is:

Bloemfontein

Minimum	8.9
Q1 (first quartile)	12.25
Median	17.8
Q3 (third quartile)	21.9
Maximum	23.4

How are these values represented in the boxplot? Label the boxplot above with the name of each number.

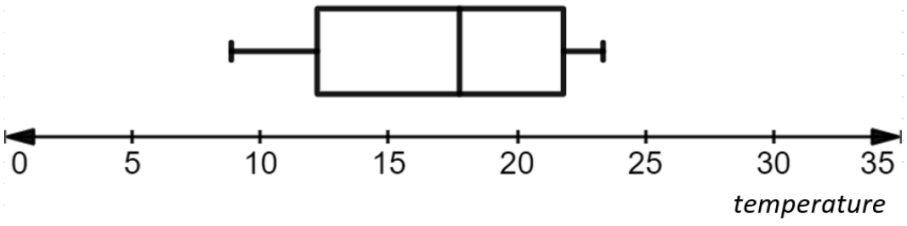
8 How is the range represented in the boxplot?

9 How is the IQR represented in the boxplot?

10 The graph below contains the boxplot for Bloemfontein. Sketch the boxplot for Cape Town above it. Use the five-number summary for Cape Town to help you draw the boxplot. **Note**: If completing this problem online, follow the instructions given online to create the boxplot.

Cape Town Temperatures

Bloemfontein Temperatures



11 Use the boxplots to compare temperature variation for the two cities. Think about the comparisons you made in earlier questions. Does your conclusion agree with these comparisons?

NEXT STEPS

We have mentioned **outliers** previously. An outlier is a value that is much greater than or less than the rest of the values in the data set.

One common rule for defining outliers is based on the IQR (the interquartile range). Outliers are more than 1.5 IQRs outward from Q1 or Q3. That is, outliers are outside of the range that extends from

$$Q1 - 1.5(IQR) \text{ to } Q3 + 1.5(IQR)$$

These values are called the **fences** for outliers. Any value less than the first fence, Q1 − 1.5(IQR), or greater than the second fence, Q3 + 1.5(IQR), is an outlier.

12 The ages of 27 recent Academy Award winners for Best Actress are given below.

21	25	26	26	28	29	29	30	32
33	33	33	33	34	35	35	36	38
39	41	42	45	49	49	61	61	80

A Complete the table below.

Ages of Academy Award winners

Minimum	
Q1 (first quartile)	
Median	
Q3 (third quartile)	
Maximum	

B What is the IQR for the ages of Academy Award winners?

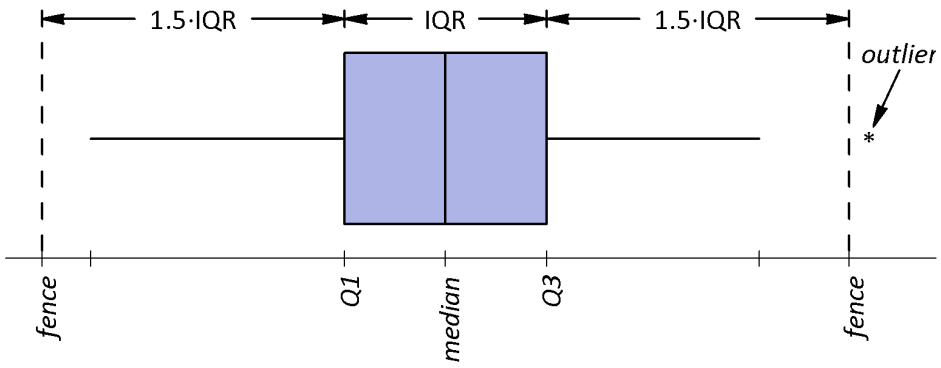
C Give the lower fence for outliers. Use the figures provided in the solution to part A. Round to 1 decimal place.

D Give the upper fence for outliers. Use the figures provided in the solution to part A. Round your answer to 1 decimal place.

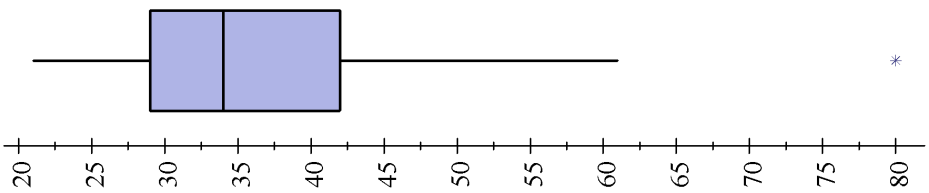
E What value is an outlier in the data set? Use the figures provided in the solution to part A.

When a data set contains outliers, we draw its boxplot differently. To draw a boxplot for a data set that contains outliers, *draw the line from the box to the last data value that is not an outlier*. Then draw the outliers as separate points.

Here is a boxplot for a data set that contains an outlier.



The boxplot for the actresses' ages is given below.



LET'S SUMMARIZE

- In this collaboration, we focused on two measures of variability: the *range* and the *interquartile range* (IQR).

- The range is simple to compute, but it is influenced by extreme observations (outliers). The IQR offers an alternative simple measure of variability. The IQR is not affected much by outliers and skewing. We say that the IQR is *resistant* to outliers.
- In addition to being resistant to outliers, the IQR allows us to determine whether high or low values are *outliers*. We use the IQR and quartiles Q1 and Q3 to determine *fences*. We call any values that are beyond the fences outliers.
- The rule we use to find fences and identify outliers is somewhat arbitrary. This means that the difference between regular data values and outliers is defined by a mathematical formula rather than anything about the data itself. Because the rule is arbitrary, it is possible that values that are close to each other may be labeled differently—so that one value is labelled an outlier but the other is not. We can use this rule as a guide, but sometimes we must also use our best judgment to determine which values are outliers.
- *Boxplots* provide a graph with a simple structure that represents center and variability. Boxplots are based on *five-number summaries*. They allow us to visualize important features of data sets.

EXERCISE 1.8

This problem uses data from Collaboration 1.6.

The following tables show how many points were scored by the home team and away team for a **sample** of games during the 2014-2015 season. The scores from away teams are sorted (from lowest to highest), but the scores from home teams are not sorted. With 25 unsorted values, it would benefit you to sort the list on a separate piece of paper. Double-check that you have 25 values before finding the 5-Number Summary.

Sample of Scores from Away Teams

68	76	77	80	80
81	82	84	85	87
89	89	91	92	92
96	97	98	99	100
101	103	111	112	114

Sample of Scores from Home Teams

95	97	74	100	99
102	78	89	87	79
86	88	91	105	105
74	123	96	91	93
80	97	104	92	111

1 Calculate the five-number summary for the data values from away teams and the home teams.

Away Teams

Minimum	
Q1 (first quartile)	
Median	
Q3 (third quartile)	
Maximum	

Home Teams

Minimum	
Q1 (first quartile)	
Median	
Q3 (third quartile)	
Maximum	

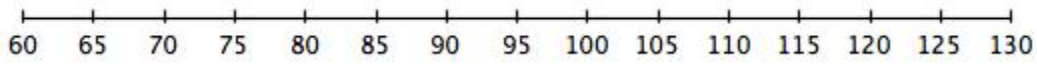
2 Calculate the range for the scores from the away and home teams. It may be helpful to sort the scores first.

3 Report the IQR for the scores from away teams and the home teams.

4 Calculate the fences for outliers for the scores from away teams. Round to 1 decimal place.

5 Calculate the fences for outliers for the scores from home teams. Round to two decimal places.

6 Draw stacked boxplots for the scores from away teams and home teams. Remember to plot any outliers using separate points. Compare the scores based on the graphs. Do the graphs indicate that playing at home can lead to higher scores? Make sure to talk about both center and variability. Refer to the specific graphs to support your conclusion.



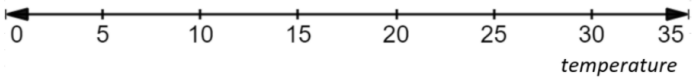
7 Imagine that an error was made when the data values for typical monthly temperatures in Bloemfontein were recorded in the table below. The first two digits for the January temperature were reversed. It was supposed to be 23.4 but was written as 32.4.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Bloemfontein	**32.4**	22.5	20.6	16.2	12.5	9	8.9	12	16.2	19.4	21.2	22.9

A Calculate the range with this new incorrect value and compare it with the original range you calculated earlier. How does this error affect the range?

B How does this error affect the IQR? Does it increase the IQR, decrease the IQR, or leave it unchanged? (**Note**: The IQR is not affected much by outliers and skewing. The range is greatly affected by outliers and skewing.)

C How does this error affect the boxplot for the data? Explain by sketching the boxplot for the data with the mistake and comparing it to the boxplot for the correct data set that you created in class.



8 The following table contains the ages of 43 women when they won their first Grand Slam title in singles tennis during the Open Era.

Athlete	Age
Nancy Richey	25
Billie Jean King	24
Virginia Wade	23
Margaret Court	26
E. Goolagong Cawley	19
Chris Evert	19
Sue Barker	20
Kerry Reid	29
Mima Jaušovec	20
Virginia Ruzici	23
Martina Navratilova	21
Tracy Austin	16
Barbara Jordan	22
Hana Mandlíková	18

Athlete	Age
Steffi Graf	17
A. Sánchez Vicario	17
Monica Seles	16
Gabriela Sabatini	20
Conchita Martínez	22
Mary Pierce	20
Martina Hingis	16
Iva Majoli	19
Jana Novotná	29
Lindsay Davenport	22
Serena Williams	17
Venus Williams	20
Jennifer Capriati	24
Justine Henin	21

Athlete	Age
Anastasia Myskina	22
Maria Sharapova	17
Svetlana Kuznetsova	19
Kim Clijsters	22
Amélie Mauresmo	26
Ana Ivanovic	20
Francesca Schiavone	29
Li Na	29
Petra Kvitová	21
Samantha Stosur	27
Victoria Azarenka	22
Marion Bartoli	28
Flavia Pennetta	33
Angelique Kerber	28
Garbiñe Muguruza	22

Here are those ages, sorted least-to-greatest:

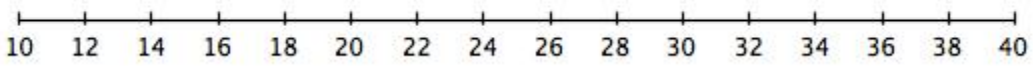
16	16	16	17	17	17	17	18	19	19
19	19	20	20	20	20	20	20	21	21
21	22	22	22	22	22	22	22	23	23
24	24	25	26	26	27	28	28	29	29
29	29	33							

A Report the values in the five-number summary for the ages of the Grand Slam title winners.

Ages of Grand Slam Title Winners

Minimum	
Q1	
Median	
Q3	
Maximum	

B Construct a boxplot for the ages of the Grand Slam title winners.



C Think about what "quartiles" means. The quartiles divide the data set into four equal parts. What proportion of these Grand Slam title winners were older than 25?

D Based on what you see in the boxplot of Grand Slam title winners' ages, do you think the mean or median age is likely to be larger? Give a reason for your answer.

E Calculate the mean of the ages. Compare it to the median. Was your answer to Question D correct?

F What range of values represents the middle 50% of the values?

G Are there any outliers in this data set? Use the IQR to justify your answer.

1.8 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Compute the values in the five-number summary by hand or using technology.	
Compute the interquartile range (IQR).	
Construct boxplots, including side-by-side boxplots.	
Use the IQR to identify outliers.	

1.9 Quantifying Variability Relative to the Mean

LEARNING GOALS

By the end of this collaboration, you should understand that

- Distributions of quantitative data exhibit variability and therefore measures of variability must be used to effectively describe distributions.
- The standard deviation of a data set is a measure of the average deviation of the data values from the mean.

By the end of this collaboration, you should be able to

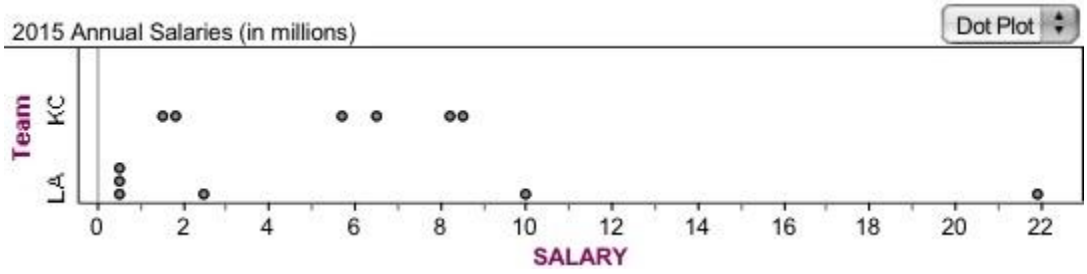
- Compute the sample standard deviation by hand and using technology.
- Interpret the standard deviation in context.

INTRODUCTION

The table and stacked dotplot below display the 2015 salaries (in millions) of a sample of players from the Kansas City Royals (KC) and Los Angeles Dodgers (LAD).

2015 Salaries (in millions)

Kansas City Royals	1.5	1.8	5.7	6.5	8.2	8.5
Los Angeles Dodgers	0.5	0.5	0.5	2.5	10	21.9



In this collaboration, we develop a measure of variability that depends on the *mean*. The mean is the average of all the data values, so all data values are used to compute it. As a reminder, the formula for the mean is below:

$$\bar{x} = \frac{\Sigma x}{n}$$

We previously measured variability in data sets using the *range* and the *interquartile range* (*IQR*). Recall that variability refers to the spread of the values in the data set.

The range is the difference between the maximum and minimum values. Because the range uses only these two data values, it is sensitive to outliers.

The IQR uses only two values as well, but it is not sensitive to outliers. It is not sensitive to outliers because it is the difference between the third and first quartiles. We find these quartiles by determining the *median*, so the IQR depends on the median.

Our goal now is to develop a measure of variability that depends on the **mean** and uses *all* of the data values.

1 A Compute the mean of each sample. Take a minute to try this on your own first before sharing in your group. Round your answers to the nearest tenth.

B How do they compare?

Language Tip

Deviations from the mean **are the distances from the mean. Data values that are above the mean have positive deviations. Data values that are below the mean have negative deviations.**

An important measure of variability for individual data values is **deviation from the mean**. Deviations from the mean are calculated with the formula,

$$deviation = (data\ value - mean) = x - \bar{x}$$

2 Calculate the deviation from the mean for each data value in the Kansas City Royals' sample. Complete the table below.

Kansas City Royals

Value	Deviation
1.5	
1.8	
5.7	
6.5	
8.2	
8.5	

3 Calculate the deviation from the mean for each data value in the Los Angeles Dodgers' sample. Complete the table below.

Los Angeles Dodgers

Value	Deviation
0.5	
0.5	
0.5	
2.5	
10	
21.9	

4 A Which data set has larger deviations from the mean?

B What does this tell us about which data set has a larger spread?

NEXT STEPS

Each data value *deviates* from the mean. In other words, each value differs from the mean. Data values with large deviations (i.e., those that differ farther from the mean) contribute more to the variability in the data set. Values with small deviations do not contribute as much to the total variability. To measure total variability, we need a way to summarize these deviations.

5 How can we summarize the deviations to obtain a single number that describes the typical deviation from the mean for all values in a data set? Take a minute to consider this on your own first before sharing in your group.

The **standard deviation** is a measure of variability that describes the average deviation from the mean for all values in a data set. To find the standard deviation of a sample, we (1) square each deviation, (2) find the sum of the squared deviations, (3) find the average of the squared deviations, and (4) take the square root of the average of the squared deviations.

6 Your instructor will assign you one of the samples to analyze. For the sample that you are assigned, complete the table below.

A Enter in the deviations and compute the squared deviations. For the squared deviations, keep two places after the decimal.

Kansas City Royals

Value	Deviation	$(\text{Deviation})^2$
1.5		
1.8		
5.7		
6.5		
8.2		
8.5		

Los Angeles Dodgers

Value	Deviation	$(\text{Deviation})^2$
0.5		
0.5		
0.5		
2.5		

10		
21.9		

B Compute the sum of the squared deviations for your sample. Round your answer to two decimal places.

C The sum of the squared deviations is one way to represent the variability in a distribution. But it is not a commonly used measure. A more commonly used measure is the **variance**. To calculate the variance, we divide the sum of the squared deviations by $n - 1$, , where n is the **sample size**. So the sample variance is an average of the squared deviations.

Below is the formula for the sample variance, denoted by s^2 :

$$s^2 = \frac{sum\ of\ the\ squared\ deviations}{n-1}$$

Compute the sample variance. Keep three places after the decimal. (We are keeping all digits until we are done with the computation to avoid rounding error.)

D The sample **standard deviation**, denoted by *s,* is the square root of variance. Calculate the standard deviation for the sample you were assigned. We're finished computing now, so it's okay to round to two places after the decimal. Interpret this result. Include units.

E Which data value in your sample had the greatest impact on the standard deviation?

7 Assuming the samples are representative of the players' salaries on each team, which team had more variability?

NEXT STEPS

The following questions focus on properties of the standard deviation. The standard deviation is a measure of the average deviation of data values from the mean.

8 Could the standard deviation of a data set ever be negative? Explain.

9 Could the standard deviation ever be 0? Explain.

Computing the Standard Deviation with Technology

As you have seen, calculating standard deviation has many steps and can take a long time. We usually use technology to compute the mean and standard deviation.

10 In Unit 1.8, you examined what happened to the range and IQR when we changed a single data value. We did this for the temperatures in Bloemfontein. Let's examine what happens to the standard deviation when a single data value is changed in the Kansas City Royals sample. Here is the Kansas City Royals sample with a mistake in the first salary. The correct value is 1.5.

2015 Salaries (in millions)

Kansas City Royals	15.5	1.8	5.7	6.5	8.2	8.5

A How do you think the wrong salary will change the standard deviation? Explain.

B Use technology to calculate the standard deviation for the sample of salaries (including the erroneous value). Round to two places after the decimal.

C How does the sample standard deviation, computed using the *incorrect* value, compare to the correct sample standard deviation? What does this tell us?

YOU NEED TO KNOW

Outliers and skewing have a large effect on the standard deviation. Recall that this was also the case for the mean. We say that the standard deviation is *not resistant* to the effects of outliers and skewing. Very large or small values can have a large impact on the standard deviation.

LET'S SUMMARIZE

We have examined two measures of center (mean and median) and three measures of variability or spread (range, interquartile range, and standard deviation). We calculated the IQR using the median. We calculated the standard deviation using the mean. In deciding which measures of center and spread to use, we need to remember two things:

- Mean and standard deviation go together. Median and IQR go together.

- Both the mean and standard deviation are influenced by outliers and skew.

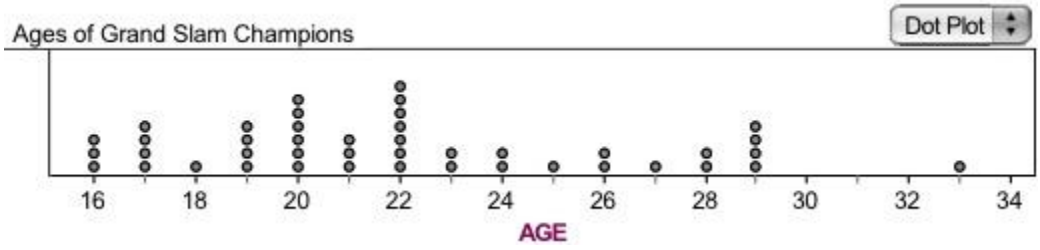
When the data are skewed or contain outliers, we usually use the median and IQR to summarize the data. When the data are reasonably symmetric we use the mean and standard deviation. In addition, these summary values are never enough. We should always look at a graph as well. This can be a dotplot, histogram, or boxplot.

EXERCISE 1.9

1 The following table contains the ages of 43 women when they won their first Grand Slam title in singles tennis during the Open Era (these data were introduced in Unit 1.8).

Athlete	Age	Athlete	Age	Athlete	Age
Nancy Richey	25	Steffi Graf	17	Anastasia Myskina	22
Billie Jean King	24	A. Sánchez Vicario	17	Maria Sharapova	17
Virginia Wade	23	Monica Seles	16	Svetlana Kuznetsova	19
Margaret Court	26	Gabriela Sabatini	20	Kim Clijsters	22
E. Goolagong Cawley	19	Conchita Martínez	22	Amélie Mauresmo	26
Chris Evert	19	Mary Pierce	20	Ana Ivanovic	20
Sue Barker	20	Martina Hingis	16	Francesca Schiavone	29
Kerry Reid	29	Iva Majoli	19	Li Na	29
Mima Jaušovec	20	Jana Novotná	29	Petra Kvitová	21
Virginia Ruzici	23	Lindsay Davenport	22	Samantha Stosur	27
Martina Navratilova	21	Serena Williams	17	Victoria Azarenka	22
Tracy Austin	16	Venus Williams	20	Marion Bartoli	28
Barbara Jordan	22	Jennifer Capriati	24	Flavia Pennetta	33
Hana Mandlíková	18	Justine Henin	21	Angelique Kerber	28
				Garbiñe Muguruza	22

The ages are summarized in the dotplot below.



1 The mean of these ages is 22.1 years. Use technology to compute the standard deviation for the athletes' ages. Round to one decimal place. Include units with your answer.

2 Explain what the standard deviation means in the context of their ages.

3 Which data value in the sample has the greatest impact on the standard deviation?

4 What is the deviation from the mean for this data value? Round to one decimal place.

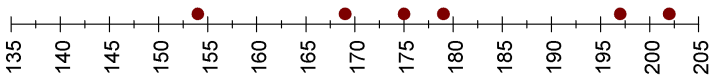
In Unit 1.7, we examined the weight gains (in grams) over a one-month period for two samples of laboratory rats. One sample contained six adolescent rats that were given a high daily dose of a stimulant drug. This sample was called the Treatment group. The sample of six normal adolescent laboratory rats was the Control group. This group received no treatment.

Here are the weight gains for the two groups:

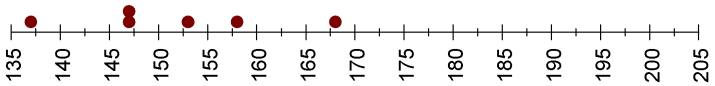
Control Group	169	154	179	202	197	175
Treatment Group	137	158	153	147	168	147

These values are summarized in the dotplots below.

Control Group Weights



Treatment Group Weights



5 Use technology to compute the standard deviation for the Control group. Round your answer to two decimal places. Be sure to include the units.

6 Use technology to compute the standard deviation for the Treatment group. Round your answer to two decimal places. Be sure to include the units.

7 What do the standard deviations for the weight gains tell us about the distributions of the two groups? When making your comparison, think about differences in variability between the groups.

(i) The weight gains of the Control group are greater than the weight gains of the Treatment group.

(ii) The weight gains of the Control group have more variability than the weight gains of the Treatment group.

(iii) The weight gains of the Control group are less than the weight gains of the Treatment group.

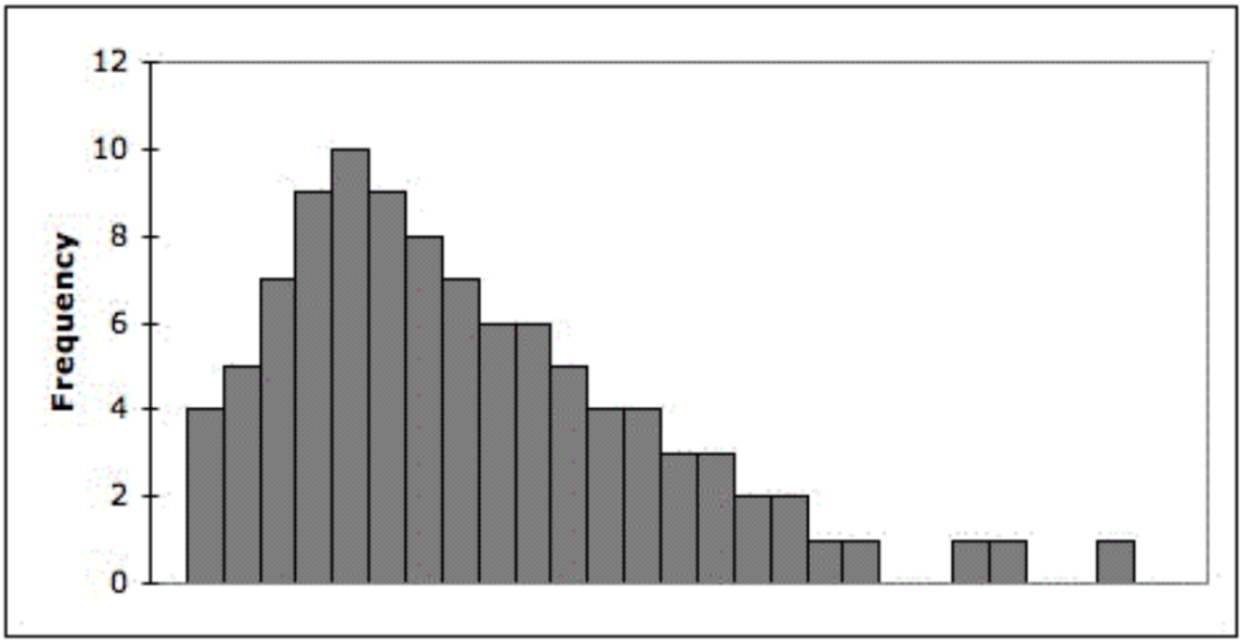
(iv) The weight gains of the Control group have less variability than the weight gains of the Treatment group.

8 Which of the following two quantitative data sets would you expect to have the larger standard deviation?

(i) The collection of heights, measured in inches, of 1st- through 6th-grade students in an elementary school.

(ii) The collection of heights, measured in inches, of all 12th-grade students in a high school.

9 Consider a quantitative data set with histogram shown below.



Which would be a better tool for measuring center and spread?

(i) Mean and Standard Deviation

(ii) Median and IQR

10 A university statistics class conducted a survey of how students spend their money. They asked 25 students to estimate how much money they typically spend each week on fast food. They determined that the mean amount spent on fast food is $31.52 with a standard deviation of $21.60. Later they realized that a value entered as $3 should have been $30. They recalculate the mean and standard deviation. The mean is now $32.60. Which of the following is true about the standard deviation?

(i) The standard deviation will increase, because we have increased the value of a data point.

(ii) The standard deviation will stay the same, because the standard deviation is not affected by a change in a single measurement.

(iii) The standard deviation will decrease, because this change moved a data point closer to the mean.

11 Suppose that 20 students take a quiz worth 80 points. The standard deviation of the scores is 1 point. Which of the following gives the most reasonable description of the distribution of quiz scores?

(i) All of the individual scores are one point apart.

(ii) The difference between the highest and lowest score is 1 point.

(iii) The IQR is 2 points.

(iv) A typical quiz score is about 1 point from the mean.

1.9 Monitor (survey)

Monitor your progress on learning the objectives for this unit. If you identify any objectives you need to review, go back through the unit's activities. Your responses are not graded, but will be available for your instructor to see.

Rate how confident you are on a scale of 1–5 (1 = not confident and 5 = very confident).

Skill or Concept: I can …	**Rating from 1 to 5**
Compute the sample standard deviation by hand and using technology.	
Interpret the standard deviation in context.	

STRESS REAPPRAISAL

As you might expect, taking a math test can be a very stressful experience. Before taking the Module 1 exam, we are going to complete a short reading exercise designed to help you perform well.

In the following pages, you will be presented with scientific research. Because scientific research articles can be long, complex, and difficult to understand we had students like you summarize the research articles discussed below.

The research presented in the following materials was conducted with people with anxiety disorders to help them cope with feelings of overwhelming anxiety. More specifically, materials provide you with information about how stress can help improve performance.

While you read each summary, please try to think about how stress can help you. Then, after each one, you will be asked to write a short summary using your own words describing how the information presented can help maximize your performance on the Module 1 exam. The information you write here will be included in messages to future Quantway and Statway students designed to provide students with the confidence to thrive in testing situations.

The following summary is of Jamieson and Mendes' 2010 study that appeared in the Journal of Experimental Psychology:

> In stressful situations, people experience changes in their body. They might experience these changes as "unsettled feelings" or "butterflies in their stomach," and conclude that they are nervous. However, the bodily changes that happen during stress can be good, not just bad. For instance, scientists have found that feelings of "butterflies" or nervousness indicate that the body is gathering resources to meet the demands of the situation. In other words, the body needs energy to perform and stress helps deliver this energy to where it is needed.
>
> Stress can be "good" or "bad," and the type of stress we experience depends on our perceptions and beliefs. To demonstrate how the body and mind work together, imagine you are a skier staring down a steep, icy slope with no other way off the mountain than going down this dangerous trail. Regardless of whether you like skiing, this situation would cause stress. Expert skiers may experience the stress as "excitement" if they believe they can handle the difficult trail, whereas novices are more likely to experience the stress as "fear" if the difficulty of the trail exceeds their skill level. Thus, the skier's response (excitement vs. fear) depends on how they perceive the situation and their body's responses.
>
> The negative consequences of stress are the direct result of people's beliefs that feeling anxious will make them perform poorly. However, our research from people with anxiety disorders indicates that stress does not hurt performance, but can actually help because our brain releases chemicals that help us think quickly. So, during the test today, try and view stress as a coping tool.

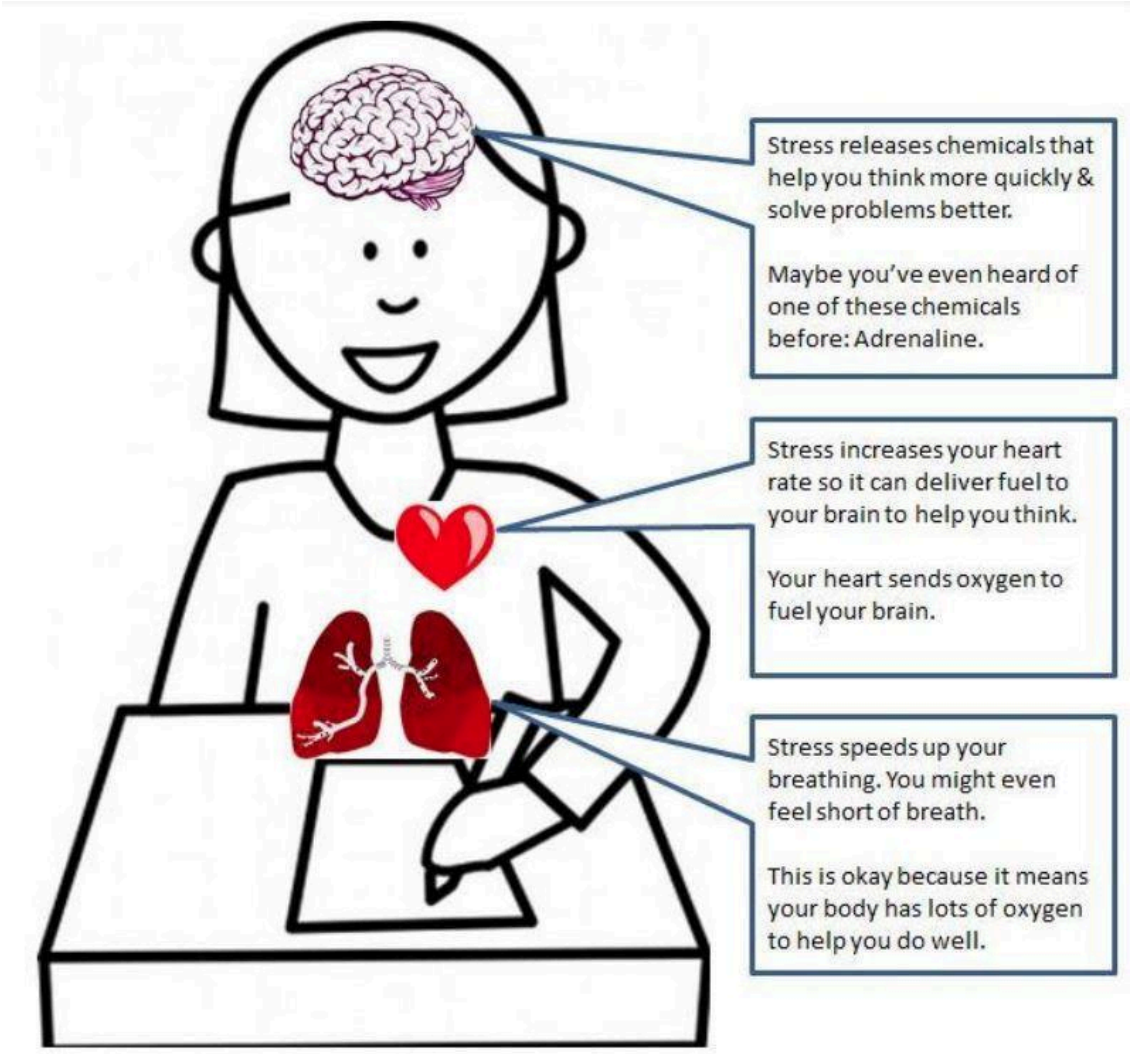
1 In your own words please briefly describe how this information can help you perform well on the Module 1 exam:

The following summary is of Nock et al.'s (2011) study that appeared in the Journal of Clinical Psychology:

> Stress is a normal physical reaction that helps you face the challenges in your life. It is not harmful. In fact, if we did not have stress reactions we could not survive. If stress is helpful, then why do most people see it as a negative experience?
>
> Research from patients coping with anxiety disorders indicates that negative reactions to stressful situations like taking an exam in class or going on a job interview are the result of how we think about stress (also known as "cognitive appraisals"). When the "fight or flight" system activates, our brain searches for potential sources of harm. However, in modern society there is often no physical threat. When no explanation can be found, the search turns inwards and the brain invents explanations such as, "There must be something wrong with me." Nothing could be further from the truth.
>
> During stressful situations remember that your body's responses are good. Increased heart rate, sweating, and heavy breathing all help deliver oxygen where it is needed. Remember, if you find yourself experiencing stress during tests, focus on the benefits of your body's responses.

2 In your own words please briefly describe how this information can help you perform well on the Module 1 exam:

The following is an illustrative diagram that shows the biological changes that happen when we experience stress. Please take a minute to note where the changes occur and how these help us do well.



3 In your own words please briefly describe how this information can help you perform well on the Module 1 exam:

Great job! You have finished the reading exercise.

Remember during the Module 1 exam, we ask that you try to remind yourself that your body's responses to the stressful testing situation will help you to perform well.

Good luck!