

STA1001C

Student

Course Materials

Course Materials
Table of Contents

Chapter 1 – Course Launch, What is Statistics, and Mind Set Activity

Supplement Community Contract..... 3
Lesson 1.1: The Statistical Analysis Process..... 5
Supplement 1.2: Mindset Activity 17
Supplement 1.2: Mindset Questions..... 21
Lesson 1.2: Mindset Follow-up 23
Supplement 1.2: 8 Choices of Successful Students Handout 27

Chapter 2 – Displaying and Summarizing Data

Lesson 2.1: Dotplots, Histograms, and Distributions for Quantitative Data..... 31
Lesson 2.2: Constructing Histograms for Quantitative Data 39
Supplement 2.2 51
Lesson 2.3: Quantifying the Center of a Distribution..... 53
Lesson 2.4: Quantifying Variability Relative to the Median..... 61
Lesson 2.5: Quantifying Variability Relative to the Mean..... 79
Review Chapter 2 91

Chapter 3 – Exploring Bivariate Relationships

Lesson 3.1: Introduction to Scatterplots and Bivariate Relationships..... 103
Supplement 3.1A 113
Supplement 3.1B..... 115
Lesson 3.2: Introduction to Graphing Lines in Statistics 117
Lesson 3.3: Form, Direction, and Strength of the Relationship
 Between Two Measurements..... 127
Lesson 3.4: Introduction to the Correlation Coefficient and Its Properties..... 137
Lesson 3.5: Using Lines to Make Predictions 149
Lesson 3.6: Investigating the Slope and Y-Intercept of the Line of Best Fit..... 161
Lesson 3.7: Least Squares Regression Line as Line of Best Fit..... 171
Lesson 3.8: Using Explained Variation to Measure Fit..... 187
Review Chapter 3 199

Chapter 4 - Investigating Patterns in Bivariate Data

Lesson 4.1: Investigating Patterns in Data	205
Lesson 4.2: Exponential Models.....	217
Review Chapter 4	231

Chapter 5 – Types of Statistical Studies and Producing Data

Lesson 5.1: Research Questions and Types of Statistical Studies	237
Lesson 5.2: Random Sampling.....	251
Supplement 5.2: Random Rectangles	261
Lesson 5.3: Collecting Data by Conducting an Experiment.....	263
Review Chapter 5	279

Chapter 6 – Two-Way Tables with Intro to Probability

Lesson 6.1 A: An Introduction to Two-Way Tables	289
Lesson 6.1 B: An Introduction to Two-Way Tables	291
Lesson 6.2: Marginal, Joint, and Conditional Probabilities from Two-Way Tables	301
Lesson 6.3: Building Two-Way Tables to Calculate Probability	313
Review Chapter 6	329
Acknowledgement	333

Chapter 1

**Course Launch, What is Statistics,
and
Mind Set Activity**

Student Name _____

Community Contract

By signing below, I commit to the following requirements for participation in the course, and acknowledge that I understand the requirements for continued enrollment. Specifically:

- I commit to successfully completing this course with my classmates.
- I commit to helping all of my classmates understand statistics.
- I will come to class every day 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 _____

WITNESS _____

Lesson 1.1

The Statistical Analysis Process

INTRODUCTION

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 discovered 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 earth was the center of the universe. When people started making observations and using measures they discovered that this was incorrect. The earth was not the center of the universe. Data helps us make better conclusions. In this example, data helped us see that the earth goes around the sun, not the other way around.

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 need 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 idea that will help us think about how statistics can help us answer a question that we may have.

- 1 LeBron James is an NBA basketball super star who, in 2008, played for the Cleveland Cavaliers. His team made it to the playoffs that year and faced off against the Boston Celtics. However, sports announcers and fans alike claimed that he did not perform up to his abilities in the playoffs, claiming that he "choked". Imagine you want to determine if this criticism is justified. You know that data involves facts, observations, and measurements about a particular topic or idea. If you wanted to gather data about LeBron's performance, what information might you look for? How could you use data to answer this question?

Lesson 1.1

The Statistical Analysis Process

Statistical analysis is the process of looking at data to learn about something bigger. We can think of the statistical analysis process in 4 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.**

Did LeBron James choke in the 2008 playoffs?

We will now do an activity that will help us learn about the statistical analysis process. In this activity we use the statistical analysis process to investigate a question about whether LeBron choked in the 2008 playoffs.

As you discovered in answering question number 1 above, there are many ways to judge the performance of a basketball player. We will choose to look at his 3 - point shooting percentage and use this to see if his ability to make 3 - point shots decreased in the playoffs. Note, we are using his 3- point shooting percentage to answer the same question we saw above: Did LeBron James choke in the playoffs?

In order to answer our question, we need to collect the necessary data. Specifically we need to know what LeBron's 3-point shooting percentages were in the regular season of 2007-2008 and in the 2008 playoffs.

Lesson 1.1

The Statistical Analysis Process

The data shown below is from www.basketballreference.com.

Regular Season Stats

Click on column header to sort

Year	Age	Team	Lg	G	Min	Pts	PPG	FGM	FGA	FGP	FTM	FTA	FTP	3PM	3PA	3PP	ORB	DRB	TRB	RPG	AST	APG	STL	BLK	TO	
2003-04	19	CLE	NBA	79	3122	1654	20.9	622	1492	.417	347	460	.754	63	217	.290	99	333	432	5.5	465	5.9	130	58	273	
2004-05	20	CLE	NBA	80	3388	2175	27.2	795	1684	.472	477	636	.750	108	308	.351	111	477	588	7.4	577	7.2	177	52	262	
2005-06	21	CLE	NBA	79	3362	2478	31.4	875	1823	.480	601	814	.738	127	379	.335	75	481	556	7.0	521	6.6	123	66	260	
2006-07	22	CLE	NBA	78	3195	2132	27.3	772	1621	.476	489	701	.698	99	310	.319	83	443	526	6.7	470	6.0	125	55	250	
2007-08	23	CLE	NBA	75	3027	2250	30.0	794	1642	.484	549	771	.712	113	359	.315	133	459	592	7.9	539	7.2	138	81	255	
2008-09	24	CLE	NBA	81	3057	2304	28.4	789	1613	.489	594	762	.780	132	384	.344	106	507	613	7.6	587	7.2	137	93	241	
2009-10	25	CLE	NBA	76	2966	2258	29.7	768	1528	.503	593	773	.767	129	387	.333	71	483	554	7.3	651	8.6	125	77	261	
2010-11	26	Mia	NBA	79	3063	2111	26.7	758	1485	.510	503	663	.759	92	279	.330	80	510	590	7.5	554	7.0	124	50	284	
8 Season Totals					627	25180	17362	27.7	6173	12888	.479	4153	5580	.744	863	2623	.329	758	3693	4451	7.1	4364	7.0	1079	532	2086

Playoff Stats

Click on column header to sort

Year	Team	Lg	G	Min	Pts	PPG	FGM	FGA	FGP	FTM	FTA	FTP	3PM	3PA	3PP	ORB	DRB	TRB	RPG	AST	APG	STL	BLK	TO	
2005-06	CLE	NBA	13	604	400	30.8	146	307	.476	87	118	.737	21	63	.333	22	83	105	8.1	76	5.8	18	9	65	
2006-07	CLE	NBA	20	893	501	25.1	166	399	.416	148	196	.755	21	75	.280	26	135	161	8.1	159	8.0	34	10	66	
2007-08	CLE	NBA	13	552	366	28.2	113	275	.411	122	167	.731	18	70	.257	16	86	102	7.8	99	7.6	23	17	54	
2008-09	CLE	NBA	14	580	494	35.3	159	312	.510	149	199	.749	27	81	.333	19	109	128	9.1	102	7.3	23	12	38	
2009-10	CLE	NBA	11	459	320	29.1	106	211	.502	88	120	.733	20	50	.400	15	87	102	9.3	84	7.6	19	20	42	
2010-11	MIA	NBA	21	922	497	23.7	174	373	.466	119	156	.763	30	85	.353	34	142	176	8.4	123	5.9	21	25	66	
Totals				60	2629	1761	29.4	584	1293	.452	506	680	.744	87	289	.301	83	413	496	8.3	436	7.3	98	48	223

Key to Column Abbreviations	
G - Games Played	3PM - Three Pointers Made*
Min - Minutes Played	3PA - Three Pointers Attempted*
MPG - Minutes per Game	3PP - Three Point Percentage*
Pts - Total Points	REB - Total Rebounds
PPG - Points per Game	RPG - Rebounds per Game
FGM - Field Goals Made	AST - Total Assists
FGA - Field Goals Attempted	APG - Assists per Game
FGP - Field Goal Percentage	STL - Steals**
FTM - Free Throws Made	BLK - Blocks**
FTA - Free Throws Attempted	TO - Turnovers**
FTP - Free Throw Percentage	

Lesson 1.1

The Statistical Analysis Process

TRY THESE

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 with data.

- A What question are we trying to answer in our investigation of LeBron’s 3 - point shooting percentages in the regular season and in the playoffs of 2007-2008?

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

- B What information did we get from the website? Are the data related to the question we are trying to answer?

Step 3: Summarize and Analyze Data

- C Use the data shown above to summarize the necessary data in the table below. How might we use this data to decide if his ability to make 3-point shots in the playoffs decreased?

Record the appropriate data below:

LeBron James 2007-2008	3 point shots made	3 point shots attempted	3 point percentage
Regular Season			
Playoffs			

Step 4: Draw a conclusion and communicate the results.

- D Once the data are summarized and analyzed, how can we use this to answer the research question?

Lesson 1.1

The Statistical Analysis Process

NEXT STEPS

As we proceed, we will address the question of *whether LeBron choked in the playoffs*. We have completed the data collecting and summarizing process by researching on the internet and then recording the data we found in a table.

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

- 3 Suppose LeBron did not choke in the playoffs. Would you expect his 3 – point percentage to be different than what it was in the regular season? Explain why you think this.

If LeBron did not choke in the playoffs, about what percentage of 3-point shots would you expect him to make?

- 4 Suppose LeBron did choke in the playoffs. Would you expect the percentage of 3 – point shots he made in the playoffs to be *greater* or *less than* the percentage you wrote above?

How poorly would LeBron had to have performed in order to convince you that he choked in the playoffs, that is, that his 3 - point shooting ability decreased in the playoffs? Give a specific percentage. Tell why you think this.

- 5 Imagine that LeBron’s 3-point percentage in the playoffs was 22%. Does this *prove without a doubt* that LeBron’s performance in the playoffs decreased in the playoffs? If not, give another explanation for why his percentage might have been so different than his percentage in the regular season.

NEXT STEPS

Using simulation to discover what may happen by chance

If LeBron did not choke in the playoffs, we would expect him to perform at close to his regular season percentage, 31.5%. But how far below 31.5% would convince us that he did choke?

We will answer this by simulating LeBron’s 3-point shooting ability in the regular season. From that, we’ll see the typical range of percentages to expect from LeBron when he’s performing at his usual level of ability, in other words, when he’s not choking. This range of percentages is what would likely occur just due to **random (or chance) variation**. We’ll also be able to see what percentages are most likley.

Language Tip
Random variation describes the type of differences we would naturally expect to see between different games.

6 Your instructor has given you a spinner marked to match LeBron’s regular season 3 point percentage: 31.5% of the time he makes it, and 68.5% of the time he misses it. Spin the spinner 70 times (why 70?) and keep track of how many times the spinner lands on “Made It”. Complete the following information.

- Total Spins 70
- Number of “Made Its” :
- Fraction of your spins that “Made It”.
- **Decimal proportion** of your spins that “Made It”.
- **Percentage** of your spins that “Made It”.

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 the fraction of “Made Its” is 21/70, then you divide 70 by 21, or $21 \div 70$. The proportion would be 0.30. To change to a percent we multiply by 100 or move the decimal 2 places to the right. $0.30 = 30\%$.

7 Is your percentage of “Made Its” equal to 31.5%? If not, is it greater than or less than 31.5%?

8 Do you think everyone in the class got the same percentage of “Made Its”? Why or why not?

Lesson 1.1

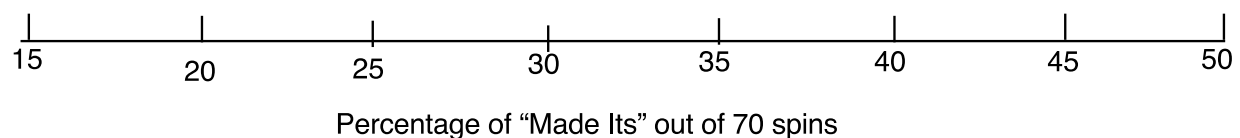
The Statistical Analysis Process

We will gather the class percentages to help us understand what kinds of shooting percentages we could expect if LeBron was shooting with the same ability he displayed in the regular season.

- 9 First, write the *class shooting percentages* in the following table. Add to the table as necessary to accommodate the number of trials your class produced. Each Trial represents 70 shots or 70 spins. *Observed Percentage* means the percentage of “Made Its” in each round of 70 spins produced.

Trial	Observed Percentage
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Second, copy the dotplot your class has made using the class data below:



Lesson 1.1

The Statistical Analysis Process

10 Answer these questions using the **dotplot** you constructed above which shows the observed percentage of “Made Its” for 70 spins, assuming LeBron shot with the same percentage as he did in the regular season, 31.5%.

- A What was the smallest *shooting percentage* observed?

- B What was the largest *shooting percentage* observed?

- C Did the *shooting percentage* differ much from trial to trial?

Use the dotplot produced in the movie to answer the following questions.

11 What does each dot represent in the dotplot?

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

13 What range of percentages is most likely?

14 Look at the dotplot. How can this graph help you decide whether LeBron choked or not?

Lesson 1.1

The Statistical Analysis Process

- 15 Recall that LeBron made 25.7% of his 3-point shots in the 2008 playoffs. Use the dotplot to answer the following question.

If LeBron was shooting in the playoffs with the same percentage as he did in the regular season, would it be unusual for him to have made 25.7% of his 3-point shots? *Hint: Look at your answer to number 13.*

Draw a Conclusion and Communicate the Results.

- 16 So, is there enough evidence for us to conclude that LeBron choked? That is, is there enough evidence for us to conclude that if LeBron was shooting in the playoffs at the same percentage as he was in the regular season, his shooting percentage in the playoffs was sufficiently low enough for us to conclude that his lower percentage was not just due to chance?

Lesson 1.1

The Statistical Analysis Process

NEXT STEPS

The task you have just completed with the LeBron example illustrates the **statistical analysis process**, which we have described in four steps. These are given again below.

Steps in a Statistical Investigation

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.

17 Identify each step of the Statistical Analysis Process for the LeBron investigation below.

Steps in Statistical Analysis	For the LeBron Investigation
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.	

Lesson 1.1

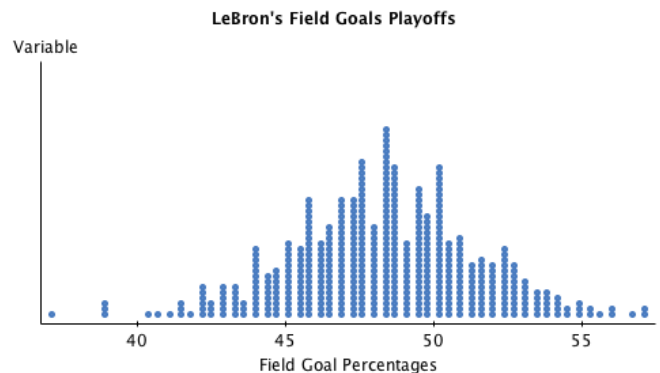
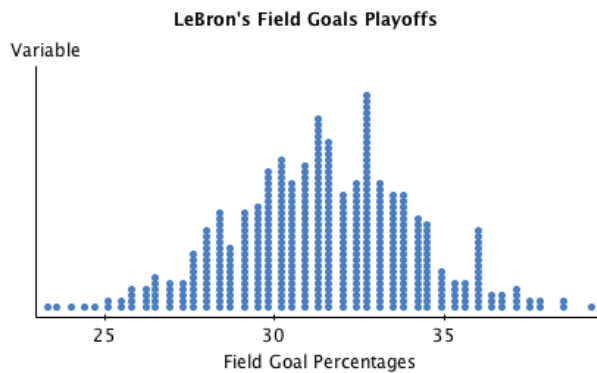
The Statistical Analysis Process

STUDENT NAME _____ DATE _____

TAKE IT HOME

- 1 Considering your answer to question 13, give an example of a 3 – point percentage that would convince you that LeBron DID choke in the playoffs.

- 2 Suppose we wanted to investigate LeBron’s Field Goal percentages, instead of his 3 – point percentages, to decide if he choked in the playoffs. Looking back at our data table, we see his regular season Field Goal percentage is 48.4%. We’ll proceed as before, and simulate this percentage in order to decide what we can expect just from random variation (luck). Note that he attempted 275 field goals in the playoffs. There are two dotplots shown below.



- A Which of the two dotplots would we expect to see if LeBron’s field goal percentage for the regular season was 48.4%? Explain why you made this choice.
-
- B Give an example of a percentage that would convince you LeBron choked in the playoffs, using field goal percentages as our indicator. Explain why you made this choice.
-
- C Look back in the playoffs data table, given at the beginning of this lesson, to find the number of field goals he made and the number he attempted in the 2008 playoffs. Record them below.

Number of Field Goals made _____ Number of Field Goals attempted _____

Lesson 1.1

The Statistical Analysis Process

D Use these values to calculate the percentage of field goals LeBron made in the 2008 playoffs. Show below what you divided. Check your answer in the data table given.

E Looking at the dotplot above and the percentage you calculated in part D, what can we decide? If LeBron was on trial for choking in the playoffs, using Field Goal percentages as our measurement, could we convict LeBron of choking?

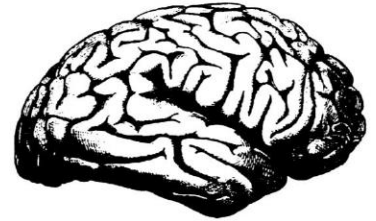
YOU CAN GROW YOUR BRAIN

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. So it's not true that some people are stuck being “not smart” or “not math people.” You can improve your abilities a lot, as long as you practice and use good strategies.

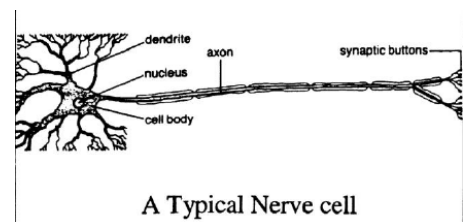


A Section of the Cerebrum nerve fibers (white matter)

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, these tiny connections in the brain actually multiply and get stronger. The more you challenge your mind to learn, the more your brain cells grow.

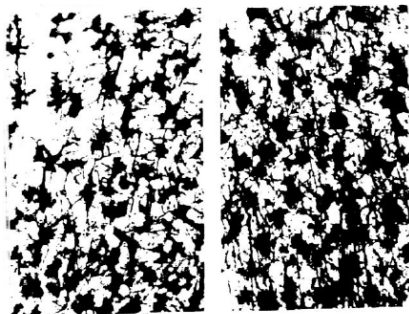
Then, things that you once found very hard or even impossible to do—like speaking a foreign language or doing algebra—become easier. The result is a stronger, smarter brain.



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.

While the animals who lived alone just ate and slept all the time, 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?

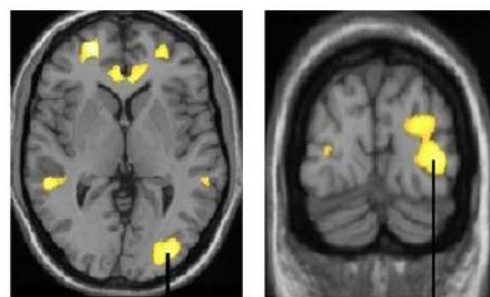
Scientists have recently shown that adults can grow the parts of their brains that control their abilities—like the ability to do math or even to juggle.

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 a long time and got much better at juggling. The other half didn’t practice, and didn’t get better.

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—the visual and motor areas. 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.” But when they learned good strategies for practicing and kept trying, they actually learned and grew their brains.

This can happen because learning causes permanent changes in the brain. The jugglers’ brain cells get larger and grow new connections between them. These new, stronger connections make the juggler’s brain stronger and smarter, just like a weightlifter’s toned muscles.



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

doi:10.1371/journal.pone.0002669.g001

A Formula For Growing Your “Math Brain”: Effort + Good Strategies + Help From Others

Scientists have also found that learning to juggle is a lot like getting better at math. When people learn and practice new ways of doing algebra or statistics, it can grow their brains—even if they haven’t done well in math in the past.

Strengthening the “math” part of your brains usually happens when you try hard on challenging math problems. But it’s not just about effort. You also need to learn skills that let you use your brain in a smarter way.

If you use a bad strategy, you may not learn—even if you try hard. A few people study for math by doing the same set of easy problems and skipping the hard ones, or just re-reading the textbook, because it feels easier. Yet when it comes time to do the test, they don’t do well because they didn’t work on problems that stretched their brains and taught them new things. When this happens, they may even say “I’m just not smart at math.”

But the truth is that everyone can become smarter at math if they practice in the right way. If a weight lifter watched other people exercise all day long, he wouldn't get any stronger. And if someone tried to learn how to juggle by just reading a book about juggling, they wouldn't learn. 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 learn something new, and less when you practice things you already know.

This means that it's not just how much time and effort you put in to studying math, but whether, when you study, you learn something new and hard. To do that, you usually need to use the right strategies. People often learn those good strategies from others, like teachers or students who do well. Luckily, strategies are easy to learn if you get help.

The Truth About “Smart” and “Dumb”

People aren't “smart” or “dumb” at math. 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—because their brain “muscles” have gotten stronger.

This is true even for adults who have struggled for a long time to learn something. Dr. Wittenberg, a scientist from Wake Forest University, 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. They may think they can't do it, or that it's too hard. It does take work to learn, just like becoming stronger physically 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.

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

Mindset 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.

In the article, you learned 3 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."
- 2 Think about an example from your own life. What is something you weren't good at first. Then you practiced it using a good strategy and became really good at it. Write about it and explain how you became good at it.

Mindset Questions

- 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 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:

- 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.
Or 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.)

Lesson 1.2

Mindset Activity and Eight Choices of Successful Students

INTRODUCTION

Thank you for taking the time to complete your online self-assessment. For each of the “8 Choices of Successful Students” (see handout) you have received a score ranging from 0 – 80. The scores for each choice area have been calculated based on your responses and were displayed for you in a table. In total you should have eight scores.

Remember that you completed the assessment by reading the statements and describing if the statement was true about you using the scale from “0: totally false” to “10: totally true.” Score ranges were also displayed to help you interpret your scores. The range of scores for each choice area can be interpreted as follows:

- 0-39 ...an area where your choices will **seldom** keep you on course.
- 40-63 ...an area where your choices will **sometimes** keep you on course.
- 64-80 ...an area where your choices will **usually** keep you on course.

Reflecting on the results of this assessment may help you discover changes that you might want to make in your behaviors and beliefs, changes that will get you on course to success! In order to be able to better reflect on your results, it is helpful to understand how your scores were calculated and how your responses to the questions impacted the scores you received.

TRY THESE

- 1 There were a total of 64 statements on the assessment you took. The statements are separated into eight equal sized groups of statements. Each group of statements is associated with a specific Choice of Successful Students.
 - A How many statements are in each of the eight equal sized groups?

 - B If each of the eight equal sized groups of statements is divided in half to make up a Part A and Part B, how many statements are in each part?

Lesson 1. 2

Mindset Activity and Eight Choices of Successful Students

- 2 In exploring the score calculations for each choice area, you may note that for the statements in Part A of each group “10 is the preferred answer” and that for statements in Part B “0 is the preferred answer.”
 - A What, if anything, does this imply about the behaviors described by the statements in Part A?

 - B What, if anything, does this imply about the behaviors described by the statements in Part B?

- 3 Consider your responses to the previous question (2) and look at Part A and Part B of each choice area.
 - A What are two ways that you can identify areas which would be considered your strengths? (Hint: Think about preferred answers.)

 - B What are two ways that you can identify areas which would be considered in need of improvement? (Hint: Think about answers that are not preferred.)

- 4 Look back at the score range interpretations provided at the beginning of this lesson as you consider the following question.
 - A Based on the score range interpretations, which is better: high scores in a choice area or low scores?

 - B How do the scores for Part A impact your total score for that choice area?

Lesson 1. 2

Mindset Activity and Eight Choices of Successful Students

C What, if anything, does this imply about the behaviors described by the questions asked in Part A?

D How do the scores for Part B impact your total score for that choice area?

E What, if anything, does this imply about the behaviors described by the questions asked in Part B?

8 CHOICES OF SUCCESSFUL STUDENTS

Successful Students . . .	Struggling Students . . .
<p>accept personal responsibility, seeing themselves as the primary cause of their outcomes and experiences.</p>	<p>see themselves as victims, believing that what happens to them is determined primarily by external forces such as fate, luck, and powerful others.</p>
<p>discover self-motivation, finding purpose in their lives by discovering personally meaningful goals and dreams.</p>	<p>have difficulty sustaining motivation, often feeling depressed, frustrated, and/or resentful about a lack of direction in their lives.</p>
<p>master self-management, consistently planning and taking purposeful actions in pursuit of their goals and dreams.</p>	<p>seldom identify specific actions needed to accomplish a desired outcome, and when they do, they tend to procrastinate.</p>
<p>employ interdependence, building mutually supportive relationships that help them achieve their goals and dreams (while helping others do the same).</p>	<p>are solitary, seldom requesting, even rejecting, offers of assistance from those who could help.</p>
Successful Students . . .	Struggling Students . . .
<p>gain self-awareness, consciously employing behaviors, beliefs, and attitudes that keep them on course.</p>	<p>make important choices unconsciously, being directed by self-sabotaging habits and outdated life scripts.</p>
<p>adopt lifelong learning, finding valuable lessons and wisdom in nearly every experience they have.</p>	<p>resist learning new ideas and skills, viewing learning as fearful or boring rather than as mental play.</p>
<p>develop emotional intelligence, effectively managing their emotions in support of their goals and dreams.</p>	<p>live at the mercy of strong emotions such as anger, depression, anxiety, or a need for instant gratification.</p>
<p>believe in themselves, seeing themselves as capable, lovable, and unconditionally worthy human beings.</p>	<p>doubt their competence and personal value, feeling inadequate to create their desired outcomes and experiences.</p>

Chapter 2

Displaying and Summarizing Data

Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

INTRODUCTION

In Chapter 1, you were introduced to 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.**

In this chapter, we will learn statistical tools to summarize and analyze data. We saw in Chapter 1 that variables in a statistical study are characteristics of the subjects in the study. We will now explore **variables** in greater detail.

There are two main types of variables, **quantitative** and **categorical**. Quantitative variables are numerical measures or counts. Examples of quantitative variables include the height and age of a person, or the number of siblings. Categorical variables divide subjects into groups (or categories) based on common values or common characteristics. Political party, favorite sports team, and eye color are examples of categorical variables.

In this lesson, we will summarize and analyze distributions of quantitative variables to investigate a question about professional basketball. A frequency **distribution** of a variable gives two important facts about the variable: 1) all of the values of the variable and 2) how often (or how frequently) the variable takes on the values.

The National Basketball Association (NBA) announced that a new basketball would be used for the 2006–2007 season. Here is the announcement from the NBA about the new ball.
(www.nba.com/news/blackbox_060628.html)

The NBA is introducing a new Official Game Ball for play beginning in the 2006–07 season. The new ball, manufactured by Spalding, features a new design and a new material that together offer better grip, feel, and consistency than the current leather ball. This marks the first change to the ball in over 35 years and only the second in 60 seasons.

Players in the NBA complained about the new ball. The NBA announced that the traditional leather ball would be used again beginning January 1, 2007. (www.washingtonpost.com/wp-dyn/content/article/2006/12/11/AR2006121100898_pf.html)

Washington Wizards guard Gilbert Arenas said the new basketball gets slippery when it comes into contact with even small amounts of sweat. Teammate Antawn Jamison said he had trouble palming the new ball while driving to the basket. Miami Heat center Shaquille O’Neal said it “feels like one of those cheap balls that you buy at the toy store.” Some players, including league MVP Steve Nash, recently began complaining that the new ball was producing small cuts on their hands.

Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

Try These – Part 1

The players' were complaining that the new ball impacted their game performance. If this is true, the new ball would have a **measurable effect** on the games. In this task, you will develop a plan to answer the question, "Did the synthetic ball affect game performance?"

- 1 Discuss the following questions with a neighbor or your group.
 - A What variables could you look at to answer the 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?
 - C How would you collect the data?
 - D If the synthetic ball was affecting the way that player performed during the game, what would you expect to see in the data?

TRY THESE – PART 2

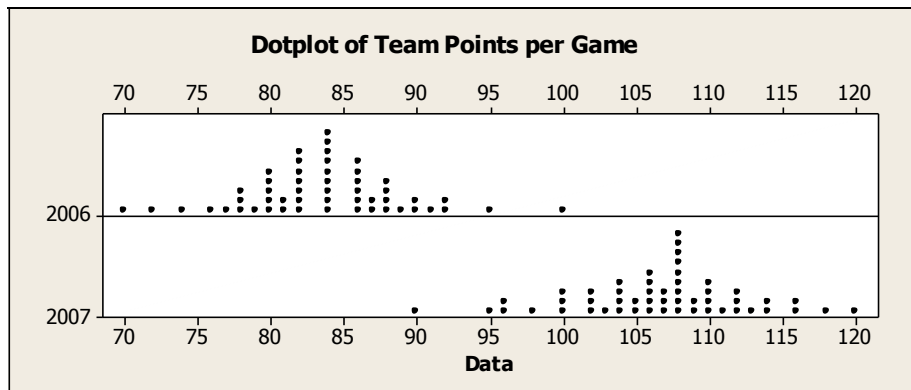
At the end of this lesson, we will analyze real NBA data to determine if the synthetic ball affected the players' performance. We will compare these two variables: 1. the number of points scored by each team for the games played in the last week of 2006 (when the synthetic ball was used), and 2. the number of points scored by each team for the games played in the first week of 2007 (when the traditional leather ball was used). The measure we will use is "points scored per team in each game."

Before we examine real NBA data, we will first examine and compare made-up dotplots. **Dotplots** display all values in a frequency distribution. Each dot represents one value. Dotplots allow us to make quick visual comparisons between two different distributions.

Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

2 The data in the following dotplot are made-up (not real NBA data).



- A What does a dot represent in these dotplots?

- B Which dotplot shows the distribution of points scored per team with the synthetic ball?

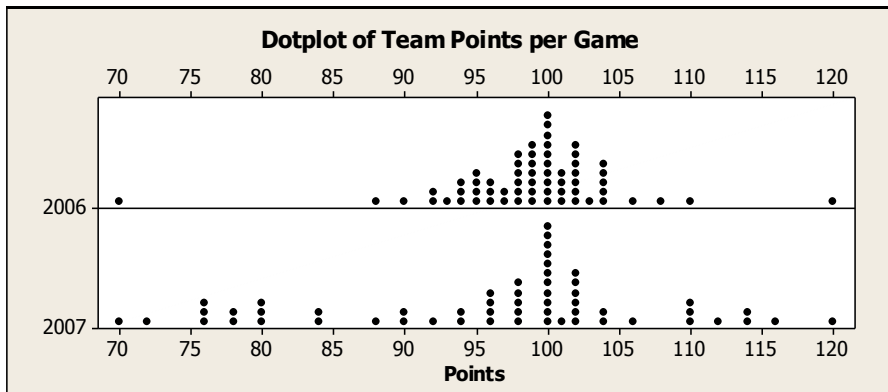
- C What is the most common score in the 2006 data? What is the most common score in the 2007 data?

- D Based on these dotplots, do you think that the synthetic ball affected the points scored by the teams? Tell how the data support your answer.

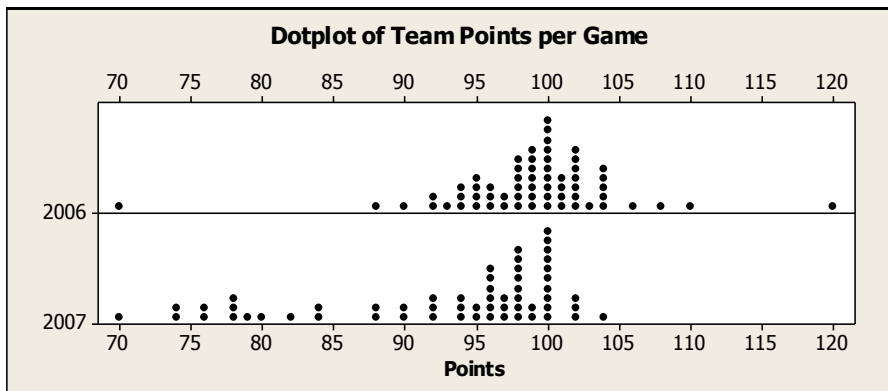
Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

- 3 The following distributions contain made-up data. First, compare the two distributions. To compare the two distributions, describe how they are similar and how they are different. Second, answer this question: Do you think that the synthetic ball affected the points scored by the teams? Tell how the data support your answer.



- 4 The following are more made-up data. Compare the two distributions. Do you think that the synthetic ball affected the points scored by the teams? Tell how the data support your answer.



Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

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.

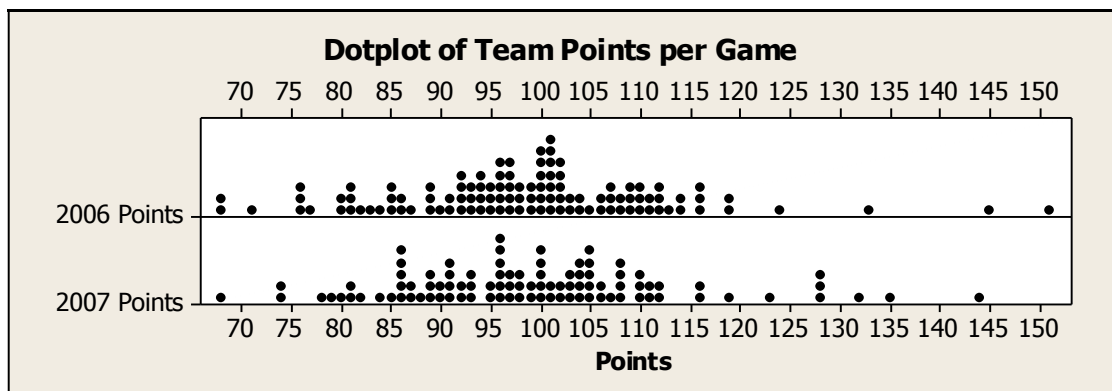
Center of a Distribution: The typical value or value which best represents the distribution.

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

Spread of a Distribution: A description or measure of how much the data values vary around the center. Literally, “how spread out are the data?”

NEXT STEPS

Now we will look at the real NBA data. We have the number of points scored by each team for the games played in last week of 2006, when the synthetic ball was used. For comparison, we have the number of points scored by each team for the games played in the first week of 2007, when the traditional leather ball was used. So, your measure is “points per team scored in each game,” which is called *Points* on the dotplots below.



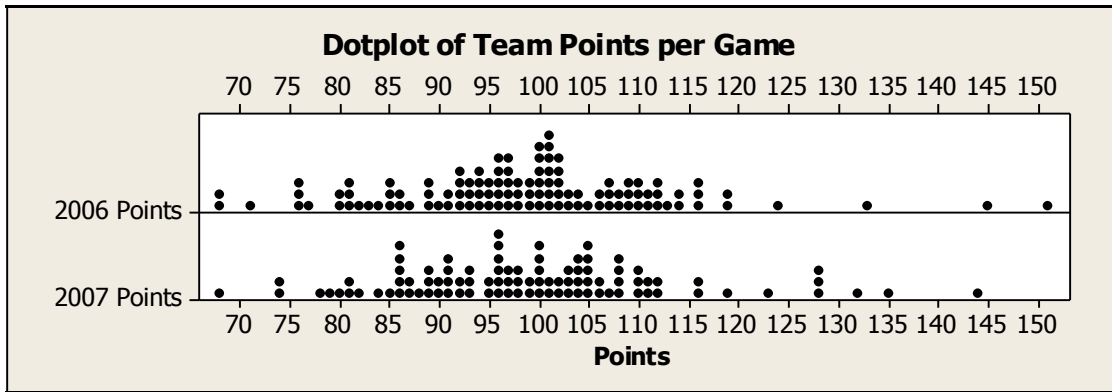
- 5 Do the real data suggest that the synthetic ball affected the number of points scored? Use the distributions to support your answer. (Remember to compare the distributions and explain your answer to the question.)

Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

STUDENT NAME _____ DATE _____

TAKE IT HOME



- 1 One way you can compare distributions is by describing the **center** of the data. The center can be viewed as a typical value that could be used to represent the data.
 - A Pick a typical score to represent the 2006 points and the 2007 points.
 - B Circle that value in each dotplot.
 - C Is the typical score you chose for games played with the synthetic ball *higher, lower, or about the same* as the typical score for games played with the traditional leather ball?

- 2 Another way to compare distributions is to describe the **spread**. The spread is the distance from the lowest value to the highest value. We usually leave out any very high or low numbers. It can also be viewed as the range where most of the values lie.
 - A Give a range where most scores lie for 2006 and for 2007.
 - B Mark the typical range of values in each dotplot.
 - C Does either ball appear to result in more consistent scoring patterns? (Is the typical range you chose noticeably shorter for one dotplot?)

Lesson 2.1

Dotplots, Histograms, and Distributions for Quantitative Data

- 3 In addition to center and spread, you can also use descriptions of **shape**. Shape describes the patterns you see in the data. How would you describe the shape of each dotplot?
- 4 The following paragraphs compare the NBA data for the last week of 2006 and the first week of 2007. Read the paragraphs, and then follow the instructions below.

Our goal was to determine if the synthetic basketball affected NBA game performance. We compared: 1. the number of points scored by each team in games played the last week of 2006 with the synthetic ball to 2. The number of points scored by each team in games played the first week of 2007 with the traditional leather ball. This was an observational study.

We found that the distribution of points scored did not differ much. The typical number of points scored by a team in a game was around 100 points. This was true whether the synthetic ball or the traditional leather ball was used. Of course, there was variability in points scored by different teams during different games. However, typical scores ranged from about 85 points to 110 points for both sets of data. So, scoring was similar with either type of ball. Both balls had scoring patterns that were slightly skewed to the right with a tail made up of 4 to 7 high-scoring games with scores above 120. Again, we viewed this as similar. So, our conclusion is that the synthetic ball did not affect scoring. Of course, other aspects of the game could have been affected by the synthetic ball as the players said. We only looked at the effect of the ball on points scored during one week of play.

These paragraphs are an example of a thorough, or complete, description of a statistical study. In particular, they illustrate how to use descriptions of center, spread, and shape to compare data sets and draw a conclusion.

In the paragraphs above, circle the sentence or sentences that are examples of the following:

- A The research question being investigated.
- B The measure used to define what data are collected.
- C The use of center in the data analysis.
- D The use of spread in the data analysis.
- E The use of shape in the data analysis.
- F The conclusion drawn from the data analysis.

Lesson 2.2

Constructing Histograms for Quantitative Data

INTRODUCTION

Constructing Histograms for a Single Quantitative Data Set

In Lesson 2.1, we analyzed data from the 2006–2007 National Basketball Association (NBA) season when the league changed to a new synthetic basketball. We used dotplots to determine if the synthetic ball affected game performance. We examined whether the change back to the traditional ball seemed to be associated with differences in the distribution of points scored by each team. We analyzed and compared dotplots from two data groups or sets: 1) total points scored by each team in games during the last week of 2006 and 2) total points scored by each team in games during the first week of 2007. In this lesson we will explore additional ways to graphically represent distributions of quantitative variables.

The following tables show visiting team scores for a **sample** of games during the last week of 2006 and the first week of 2007.

Language Tip

A *sample* is a subset of the population that we study to collect or gather data.

Sample of Visiting Team Scores in 2006

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

n = _____

Sample of Visiting Team Scores in 2007

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

n = _____

TRY THESE

- 1 Look over the **data values**. Based only on a visual examination of the data values, without doing any calculations, do you think the visiting team scores between the last week of 2006 and the first week of 2007 are similar or different? Explain your reasoning.

Language Tip

The *data values* are the numbers in the data set.

Lesson 2.2

Constructing Histograms for Quantitative Data

- 2 To help us find patterns within the 2006 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 visiting score in the 2006 data set (the lowest score is 68). The bins are shown in the following table:

Language Tip

A bin is an interval that groups together a range of different data values.

Sample of Visiting Team Scores in 2006

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

For each value in the 2006 data table, determine the bin it falls into. For example, the first visiting 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.

Complete the table with the rest of the values in the 2006 data set. Each time a tally reaches a fifth mark, represent it as a horizontal tally mark (++++).

- 3 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, count the tallies in each bin and write those numbers in the frequency column. These numbers represent the frequencies for each bin.

Language Tip

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

The table we created above is called a **frequency distribution table**. Recall that the **distribution** of a variable gives 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 in this case. The frequency distribution table is one way to show the distribution of a variable.

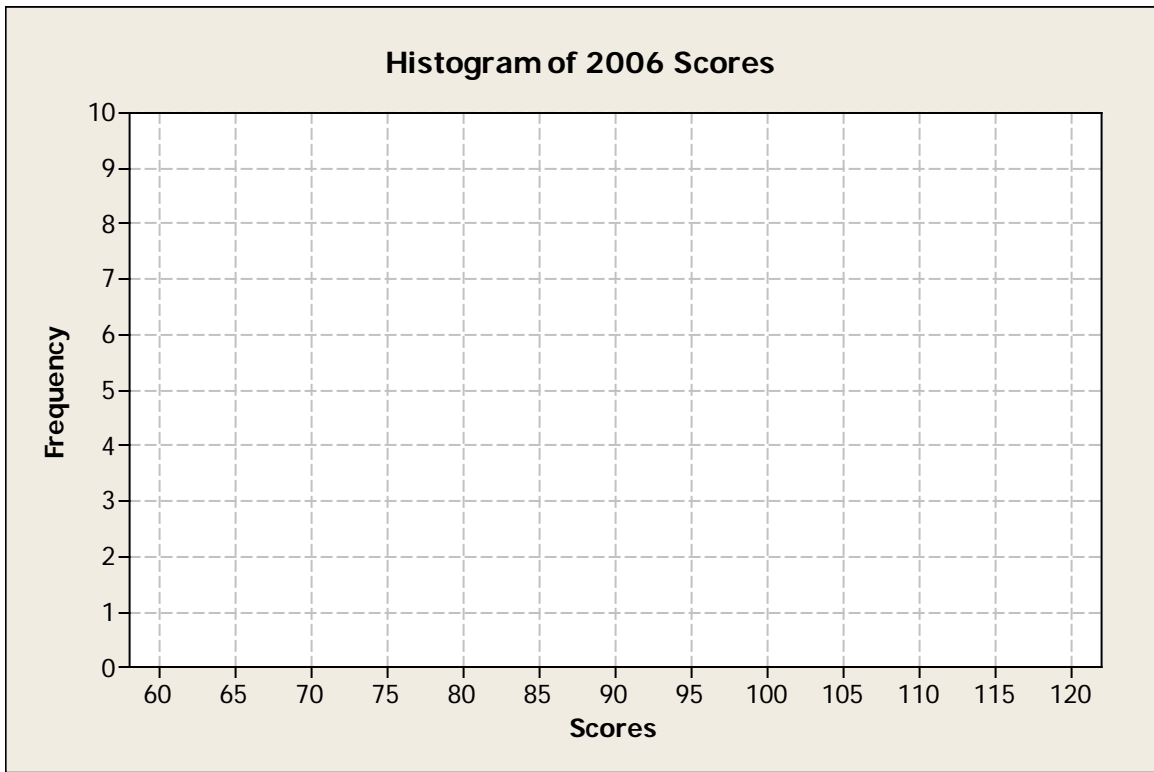
A frequency distribution table contains equal-size bins, together with the frequency of each bin. **The sum of the frequencies always equals the number of values in the data set.**

Lesson 2.2

Constructing Histograms for Quantitative Data

- 4 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.

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



- 5 Look at the bars of the histogram carefully. What is the sum of all of the heights? What does this sum represent?

Lesson 2.2

Constructing Histograms for Quantitative Data

- 6 The **relative frequency** of a bin is the proportion or percentage of data values that are contained in the bin. To calculate the relative frequency, follow these steps:
- *Figure out how many values are in the data set by adding the Frequency column.* There are 25 visiting team scores in 2006 data set.
 - *Divide the frequency of each bin by the number of values in the data set.* In this case, each frequency value is divided by 25.
 - *Write this number into the appropriate space in the column "Relative Frequency."* For example, if a bin has a frequency of 9, the relative frequency is $9/25 = 0.36$ or 36%. You would write 0.36 or 36% under relative frequency.

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

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

- 7 A **relative frequency histogram** displays the relative frequencies for the bins instead of the frequencies. Look at the frequency histogram from Question 4 and the relative frequency histogram that you received from your instructor.

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

- 8 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 2006 season are 90 points or higher?

We use graphs, such as dotplots and histograms, to help us describe the distribution of a variable. Recall from the previous lesson that the 3 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

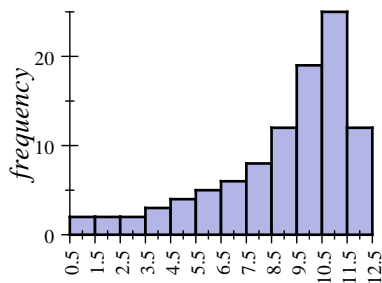
An **outlier** is a number that is separated from the rest of the data set by some distance.

Outliers are extreme values.

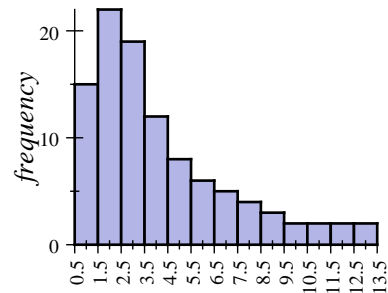
Lesson 2.2

Constructing Histograms for Quantitative Data

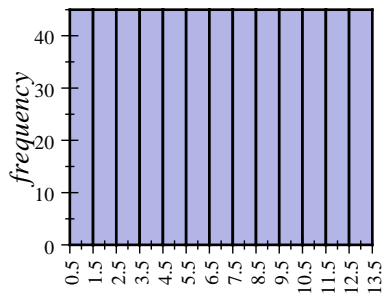
The histograms below display the shapes of different types of distributions.



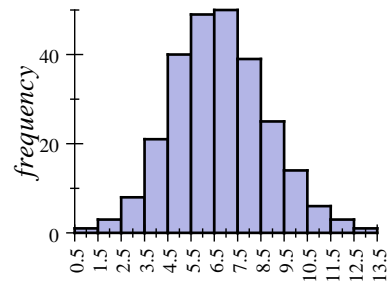
Skewed-Left Distribution



Skewed-Right Distribution



Uniform Distribution



Bell-Shaped Distribution

- 9 Use the frequency histogram of visiting team scores from the 2006 data set 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 visiting team scores from the 2006 data set? Would you say it is skewed-left, skewed-right, bell-shaped, or uniform?
 - D Are there any potential **outliers** in the scores?
 - E Write a paragraph describing the distribution, talking about the center, spread, shape, and outliers in this distribution.

Lesson 2.2

Constructing Histograms for Quantitative Data

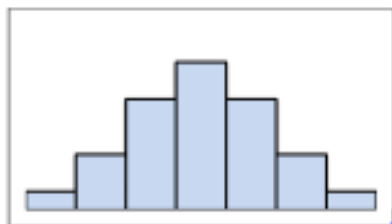
Another important characteristic to note about a histogram is whether the histogram has a single, central hump or several separated humps. These humps are called **modes**. A histogram with one hump, is called **unimodal**; histograms with two humps are called **bimodal**, and those with three or more are called **multimodal**.

Language Tip

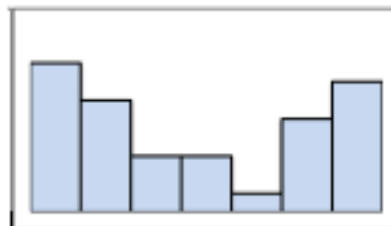
A hump or peak in a histogram is called a *mode*.

There can be more than one mode.

The histograms below display examples of a **unimodal** distribution and a **bimodal** distribution.



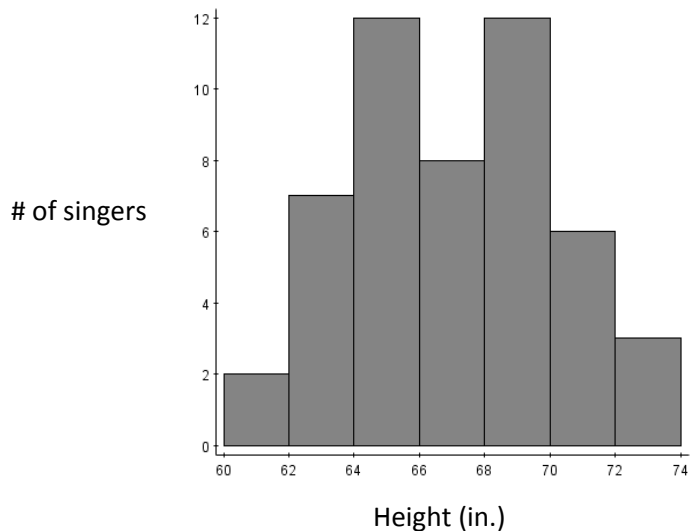
This histogram has one hump or peak.



This histogram has two humps.

Identifying gaps in the distribution may help us see multiple modes. This is important because it can encourage us to notice when data may contain more than one group. These kinds of summaries are meaningless and we should consider separating the data into different groups and summarizing the groups separately.

10 The histogram below shows the heights of some of the singers in a chorus. Can you account for the pattern you see? What may be causing it?



Lesson 2.2

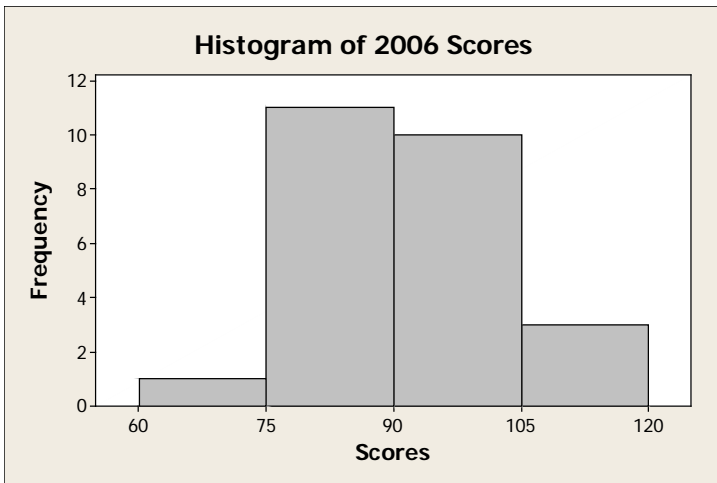
Constructing Histograms for Quantitative Data

STUDENT NAME _____ DATE _____

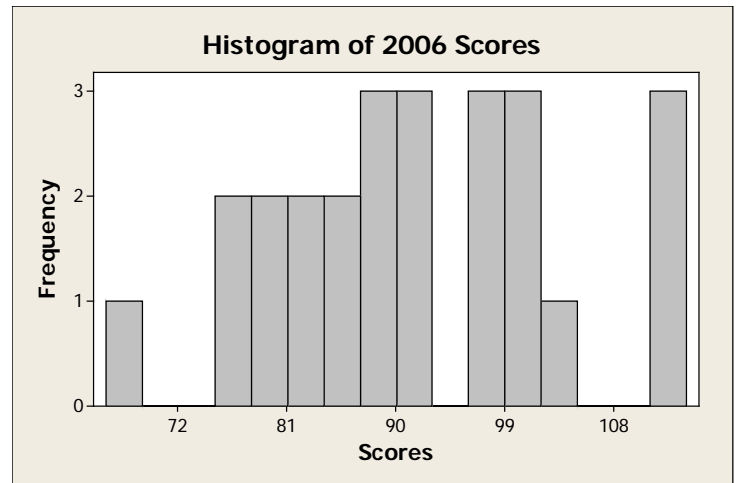
TAKE IT HOME

- 1 We can choose different bin widths when creating frequency distribution tables and histograms. Here are two examples of different bin widths for the visiting team scores from the last week of 2006:

Wide Bin Widths



Narrow Bin Widths



- A What is the bin width for each of the histograms?
- B Compare these histograms to the histogram with bin widths of 10 points that we created in the lesson. Explain the problem with too many or too few bars. Think about looking for shape, center, spread and outliers in the data as you answer the question.

Lesson 2.2

Constructing Histograms for Quantitative Data

2 Here is a list of the scores for the first week of 2007.

Sample of Visiting Team Scores in 2007

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

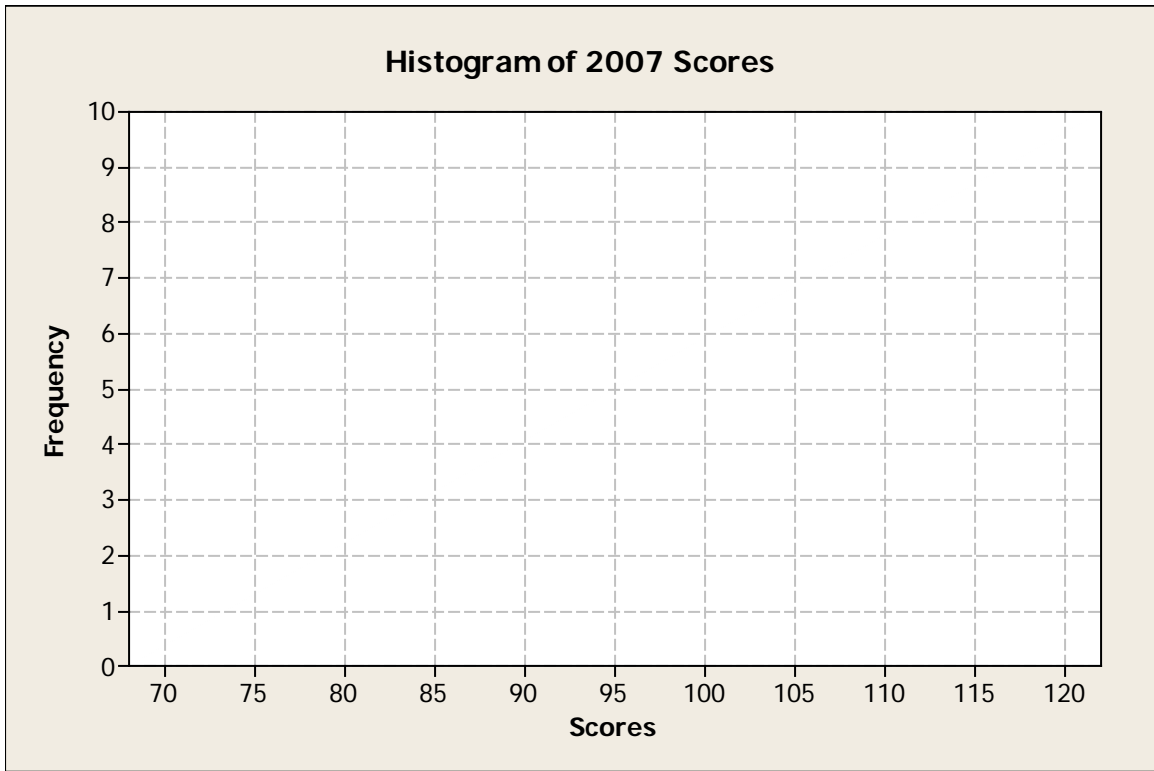
A Complete the table below by finding the frequency and relative frequency of each bin.

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			

Lesson 2.2

Constructing Histograms for Quantitative Data

B Draw the frequency histogram on the graph below.



C In how many games did the visiting team score at least 100 points?

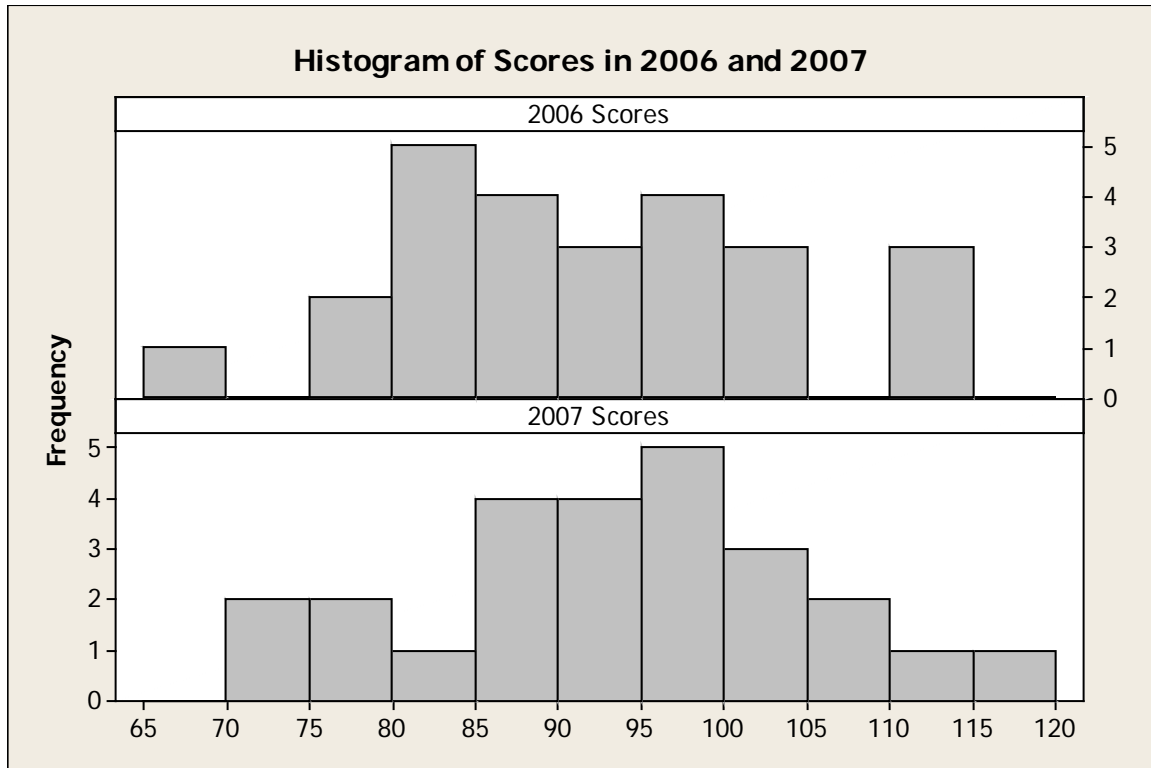
D Explain in your own words what you think the height of the second bar tells us.

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

Lesson 2.2

Constructing Histograms for Quantitative Data

- 3 Now let's compare the scores for the last 25 games of 2006, when the synthetic ball was used, and the first 25 games of 2007, when the traditional ball was used. Here are the histograms for each data set.

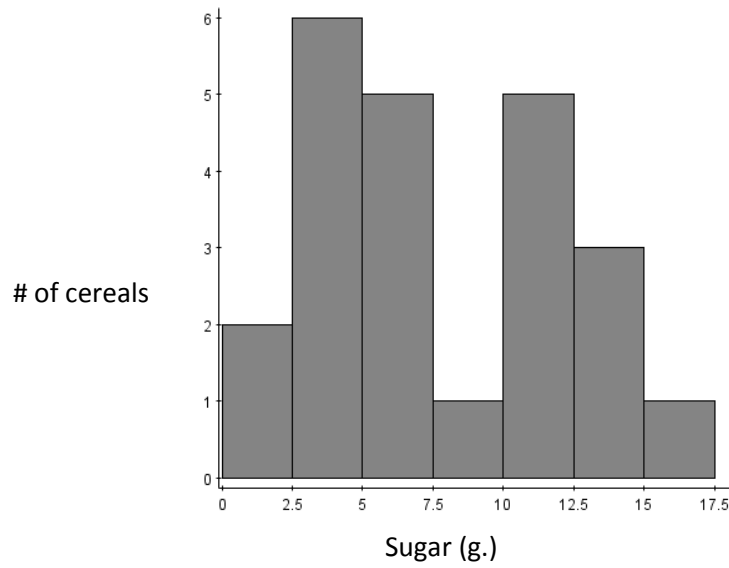


Do the graphs seem to indicate that the new ball (used in 2006) had an impact on the visiting team scores? Make sure to compare the two distributions in terms of center, spread, and shape.

Lesson 2.2

Constructing Histograms for Quantitative Data

4 Examine the histogram you see below. This histogram displays the sugar content of 49 brands of breakfast cereals.

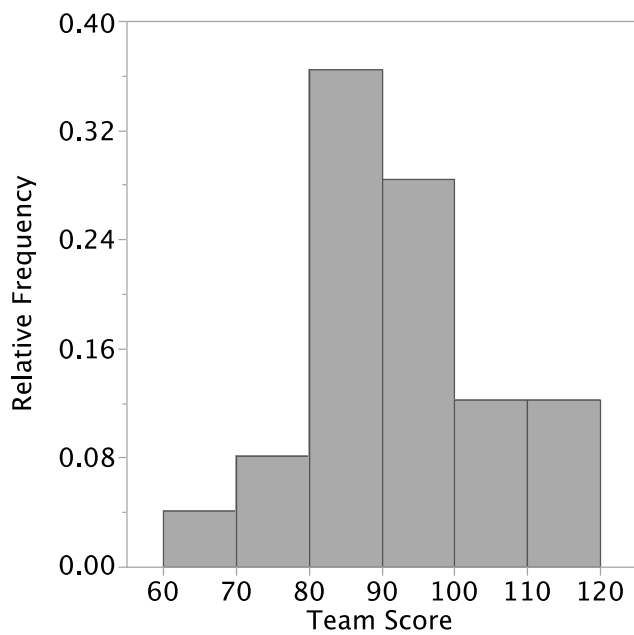
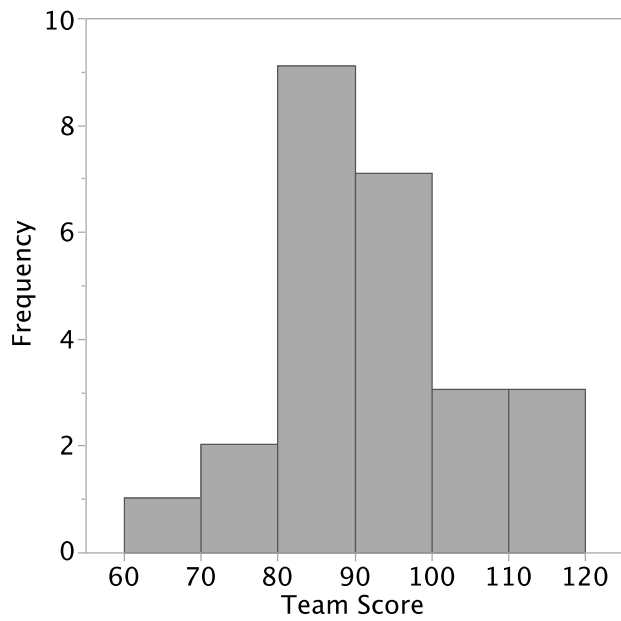


A How would you describe the pattern you see in the shape of this histogram?

B What do you think might account for this shape?

Lesson 2.2 Supplement

Frequency and Relative Frequency Histograms for 25 Visiting Team Scores During the Last Week of 2006



Lesson 2.3

Quantifying the Center of a Distribution

INTRODUCTION

Stimulants are sometimes used to help with weight loss. 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.

Estimating the Center of a Data Set in the Context of Comparisons of Data Sets/Graphics

Here are the 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

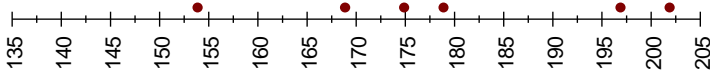
Here are weight gains (in grams) for a treatment group of six adolescent laboratory rats that were given a high daily dose of a stimulant.

137 158 153 147 168 147

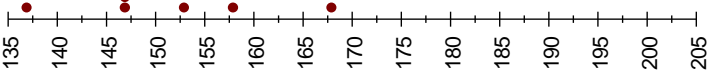
To determine whether there might be an effect on weight gain due to the stimulant, you will determine *representative values* of the two groups, namely the sample mean and sample median.

Below are dotplots for the control and treatment groups of rats.

Control Group Weights



Treatment Group Weights



Language Tip

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

- 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 right?
- 2 Calculate the average weight gain for the rats in the control group.

Lesson 2.3

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

In Statistics, we refer to the common average as the **mean**. The mean is the sum of data values divided by the number of data values (also called the count). 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:

$$\bar{x} = \frac{\text{sum}}{\text{count}} = \frac{\sum x}{n} = \frac{169 + 154 + 179 + 202 + 197 + 175}{6} \approx 179.333 \text{ grams.}$$

- 3 Calculate the **mean** weight gain for the rats given the stimulant. Call this mean \bar{y} .

- 4 Compare the values of \bar{x} and \bar{y} (the sample means for weight gains in the Control and Stimulant 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,

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

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.

- 5 Compute the sample median for the Stimulant group.

- 6 Summarize your results in the table below:

	Mean	Median
Control Group	179.333 g	177 g
Stimulant group		

Lesson 2.3

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

- 7 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.

- 8 We want to compare the two groups of rat weights. One way to do this is by comparing their representative values. Compare the sample mean and median for the control group. Does it matter which of these values you use? Explain your reasoning.

Repeat this analysis for the sample mean and sample median for the Stimulant group.

- 9 Suppose we made a big mistake when we recorded the largest weight gain in the stimulant group. Instead of recording the weight gain as 168, we recorded it as 618. Recalculate the sample mean and sample median for the stimulant group with this new value.

Lesson 2.3

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

When is the mean *not* a good choice?

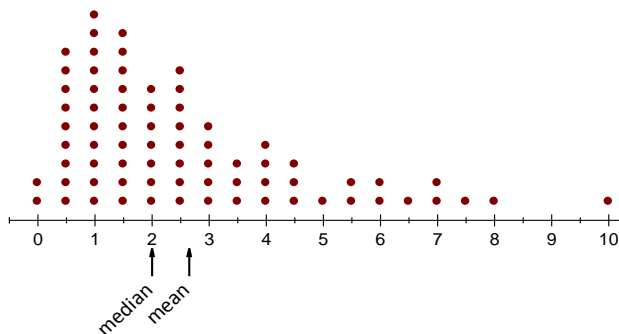
We call the mean and the median *measures of center* or *central tendency*. Measures of center are single values that represent an entire set. You've just seen that the mean can be strongly affected by extreme values. The mean is also strongly influenced by values that cause the distribution to be *skewed*.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

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.

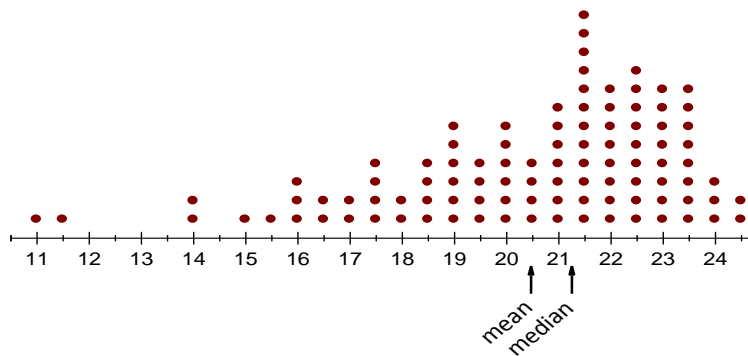
When data are skewed to the right, the mean tends to be greater than the median.

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: $\bar{x} = 2.65$.



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*, with a *longer left tail* of smaller values. The mean ($\bar{x} = 20.47$) is less than the median (21.25).



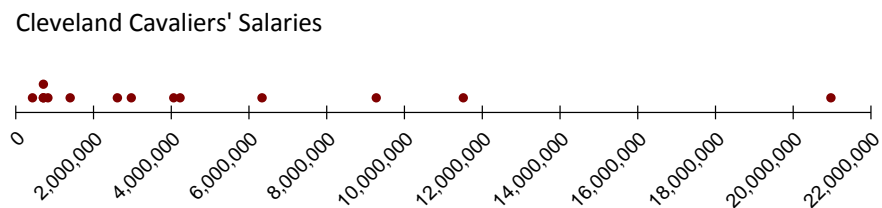
Lesson 2.3

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

- 10 Below are salaries of thirteen players on the Cleveland Cavaliers basketball team during the 2009-2010 season. The 2009-2010 season was LeBron James' last season playing for the Cavaliers. Interestingly, LeBron James was not the highest paid player on the Cavaliers during that season. The highest paid player that season was Shaquille O'Neal.

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

A dotplot of these salaries is given below.



- A Calculate the mean salary for the Cavaliers during 2009-2010.
- B Calculate the median salary. Remember to sort the values first. Since there are an odd number of values, the median is the value in the center of the sorted list.
- 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.

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.

Lesson 2.3

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

STUDENT NAME _____ DATE _____

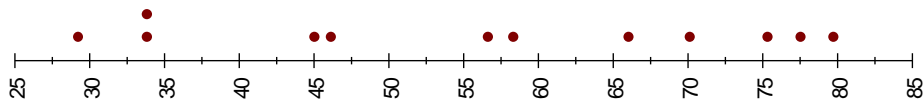
TAKE IT HOME

1 The table below lists the typical monthly temperatures of St. Louis and San Francisco.

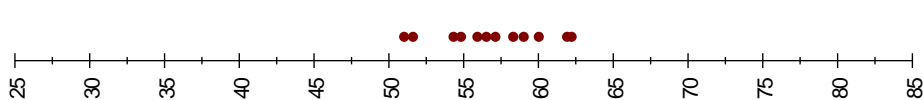
City	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
St. Louis	29.9	33.9	45.1	56.7	66.1	75.4	79.8	77.6	70.2	58.4	46.2	33.9
San Francisco	51.3	54.4	54.9	56.0	56.6	58.4	59.1	60.1	62.3	62.0	57.2	51.7

These data are summarized in dotplots below.

St. Louis Temperatures



San Francisco Temperatures



- A Compute the sample mean of the typical monthly temperatures for both St. Louis and San Francisco. Use the symbols \bar{x} and \bar{y} for these sample means.

- B Compute the sample medians of the monthly temperatures for both St. Louis and San Francisco.

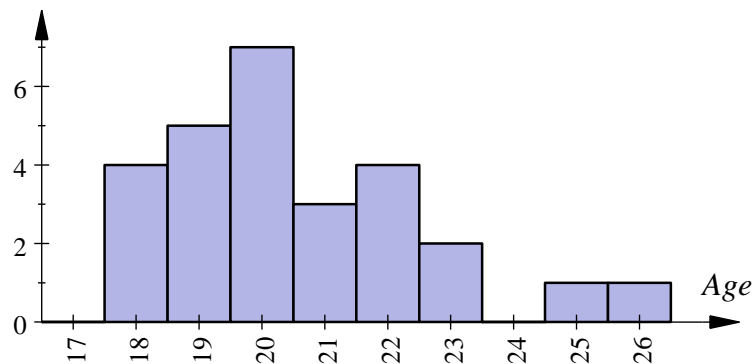
- C Now, write a brief comparison of the sample means and sample medians for the two cities. In a comparison, you should explain what is the same and what is different about the weather in the two cities with regard to their means and medians.

Lesson 2.3

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

- 2 Students in a Statistics course took a test where the highest grade was 100. Most of the students did quite well, with more than half of the students getting A's and B's on the test. Most of the rest of the students scored C's, with just a few D's. Only two students received F's, but their scores were very low, in the 30's.
- A If As are in the 90's, B's in the 80's, Cs in the 70's, D's in the 60's, and F's below 60, what is the shape of the grade distribution?
- B Is the mean likely to be greater than or less than the median? Explain your answer.
- C 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.

- 3 A histogram displaying the distribution for the ages of students in a statistics class is shown below.



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

Lesson 2.4

Quantifying Variability Relative to the Median

INTRODUCTION

In the **Take It Home** section of the last lesson, we looked at temperatures for the cities of St. Louis and San Francisco. We compared typical monthly temperatures for the two cities. These temperatures are presented in the following table.

Typical Monthly Temperatures (°F) for St. Louis and San Francisco

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
St. Louis	29.9	33.9	45.1	56.7	66.1	75.4	79.8	77.6	70.2	58.4	46.2	33.9
San Francisco	51.3	54.4	54.9	56.0	56.6	58.4	59.1	60.1	62.3	62.0	57.2	51.7

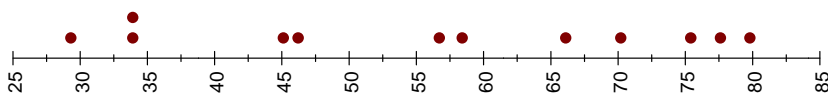
In the last lesson, we examined the center of these distributions by calculating the mean and median temperature for each city. We found these measures to be very similar for the two cities, even though their distributions are quite different. In this lesson, we study the **variability** (or *spread*)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

Language Tip

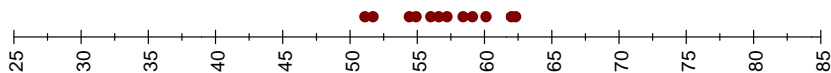
Variability or spread means how widely the data points are scattered.

- The dotplots below show the typical monthly temperatures for St. Louis and San Francisco. Examine the dotplots. Write a sentence that compares the typical monthly temperatures for the two cities. In your answer talk about how the distributions are similar and how they are different.

St. Louis Temperatures



San Francisco Temperatures



Lesson 2.4

Quantifying Variability Relative to the Median

- 2 In previous lessons, 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 St. Louis temperatures. What would it be?
- B What single number would you use to describe the variability in the San Francisco temperatures?
- C The values you chose can be called “measures of temperature variability.” Do your measures of temperature variability for St. Louis and San Francisco reflect 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.

$$\text{range} = \text{maximum} - \text{minimum}.$$

Go back to the temperature data sets for St. Louis and San Francisco. Write the minimum value, the maximum value, and calculate the range in the monthly temperatures for St. Louis. Do the same for San Francisco.

	Maximum	Minimum	Range
St. Louis			
San Francisco			

Which city has the larger temperature range?

Lesson 2.4

Quantifying Variability Relative to the Median

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.

For example imagine we added two extreme values to the San Francisco temperature data, say 28 and 80. Now if we calculated the range for San Francisco it would be about the same as the range for St. Louis, yet the way the temperatures are spread out, or the *distribution* of the data sets, is quite different. The problem is that the range only took into account the smallest and largest values, without taking into account the spread (or variability) of the rest of the data.

We need another way to describe variability – something that takes more than just the smallest and largest data values into account. 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. The name quartiles come from “quarters” or four parts. Think about how you would divide a candy bar into 4 equal pieces. First you would divide it in half, then you would take each half and divide each part into halves again. We’ll do the same with our data set.

Here are the monthly temperatures for St. Louis sorted from lowest to highest.

St. Louis	29.9	33.9	33.9	45.1	46.2	56.7	58.4	66.1	70.2	75.4	77.6	79.8
------------------	------	------	------	------	------	------	------	------	------	------	------	------

To find the quartiles, we’ll divide the data into 4 equal parts. So, just like the candy bar, we’ll divide the data first in half. This is the median of the data set. Since we have an even number of data values, to find the median we’ll need to average the two middle values. The value of the median is $(56.7 + 58.4)/2 = 57.55$ °F.

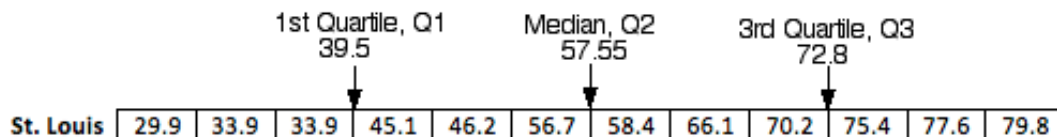
Now we’ll concentrate on the lower half of the data and divide that in half to find the **first quartile, or Q1**. So we want to find the median of the lower half of the data. Again, we have an even number of data values so the median of the lower half of the data is the average of the middle two values in the lower half, or $Q1 = (33.9 + 45.1)/2 = 39.5$ °F.

St. Louis	29.9	33.9	33.9	45.1	46.2	56.7	58.4	66.1	70.2	75.4	77.6	79.8
------------------	------	------	------	------	------	------	------	------	------	------	------	------

To find the **third quartile, or Q3**, we’ll concentrate on the upper half of the data and find the median of the upper half. This gives $Q3 = (70.2 + 75.4)/2 = 72.8$ °F.

St. Louis	29.9	33.9	33.9	45.1	46.2	56.7	58.4	66.1	70.2	75.4	77.6	79.8
------------------	------	------	------	------	------	------	------	------	------	------	------	------

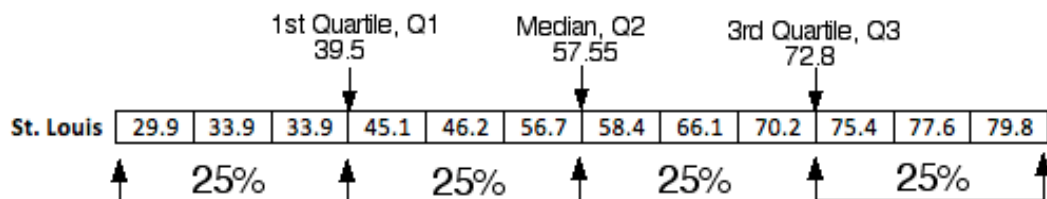
This information is summarized below:



Lesson 2.4

Quantifying Variability Relative to the Median

Note that since the data is divided into four equal parts, 25% of the data falls into each part.



- 4 The minimum, maximum, and three quartiles (the median is considered Q2) form the **five-number summary**. Here is the five-number summary for the monthly temperatures in St. Louis.

St. Louis

Minimum	29.9
Q1 (first quartile)	39.5
Median	57.55
Q3 (third quartile)	72.8
Maximum	79.8

Use the typical monthly temperatures in San Francisco to construct a five-number summary. Record the values below.

San Francisco

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

YOU NEED TO KNOW

If a data set has an *odd* number of values, the median value is not included in the upper or lower half of the data set. That is, first find the median, then ignore it and only look at the lower half to find Q1 and the upper half to find Q3.

- 5 The range from Q1 to Q3 contains the middle 50% of values. The middle 50% of St. Louis temperatures fall between _____ and _____.

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**.

YOU NEED TO KNOW

The formula is $IQR = Q3 - Q1$. The IQR gives the range of the middle 50%.

For the typical monthly temperatures in St. Louis, the IQR is $Q3 - Q1 = 72.8 - 39.50 = 33.3$ °F.

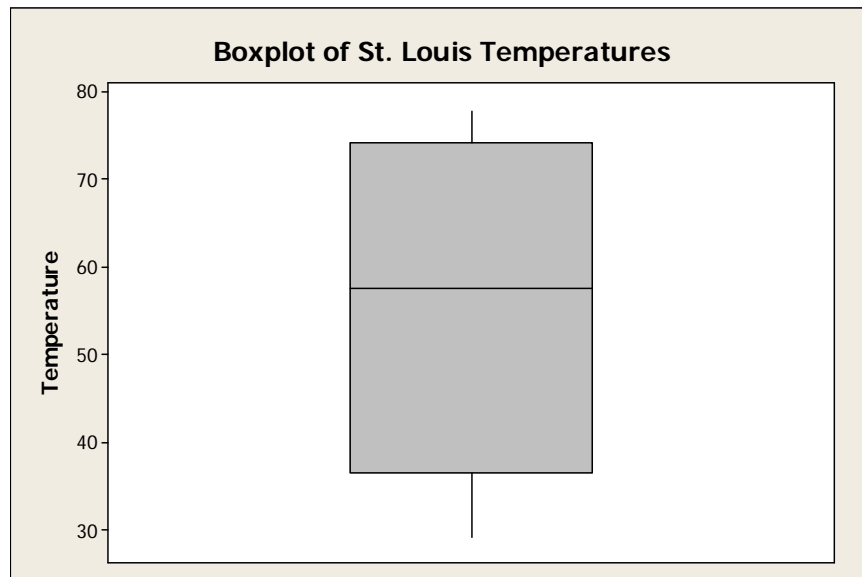
Lesson 2.4

Quantifying Variability Relative to the Median

- 6 Calculate the IQR for the typical monthly temperatures in San Francisco.
- 7 Compare the IQRs for St. Louis and San Francisco. Do these numbers reflect the variability you see in the dotplots given at the beginning of this lesson?

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 St. Louis:



- 8 The five-number summary for the St. Louis temperatures is:
min = 29.9°F, Q1 = 39.5°F, med = 57.55°F, Q3 = 72.8°F, max = 79.8°F.

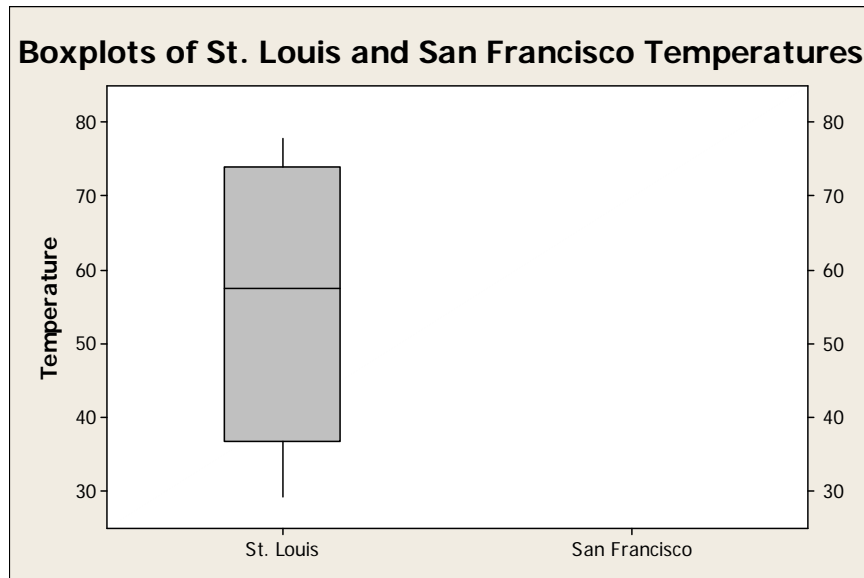
How are these values represented in the boxplot? Label the boxplot above with the name of each number. That is label the boxplot with the Min, Q1, Med, Q3 and Max.

- 9 How are the range and IQR represented in the boxplot?

Lesson 2.4

Quantifying Variability Relative to the Median

- 10 The graph below contains the boxplot for St. Louis. Sketch the boxplot for San Francisco beside it. Use the five-number summary for San Francisco to help you draw the boxplot.



- 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?

Quantifying Variability Relative to the Median

NEXT STEPS

We have mentioned **outliers** previously. An outlier is a value that is much greater than or less than the rest of the data set. But we have not said yet how far away a data value must be from the rest of the data to be considered an outlier.

One common rule for defining outliers is based on the IQR (the interquartile range). It uses the IQR as a “measuring stick”. We say that a data value is an outlier if (on the high side) it falls more than 1.5 IQRs above Q3 or if (on the low side) it falls below 1.5 IQRs below Q1. 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 lower fence, $Q1 - 1.5(IQR)$, or greater than the upper fence, $Q3 + 1.5(IQR)$, is an outlier. See the figure.



12 The ages of the last 27 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

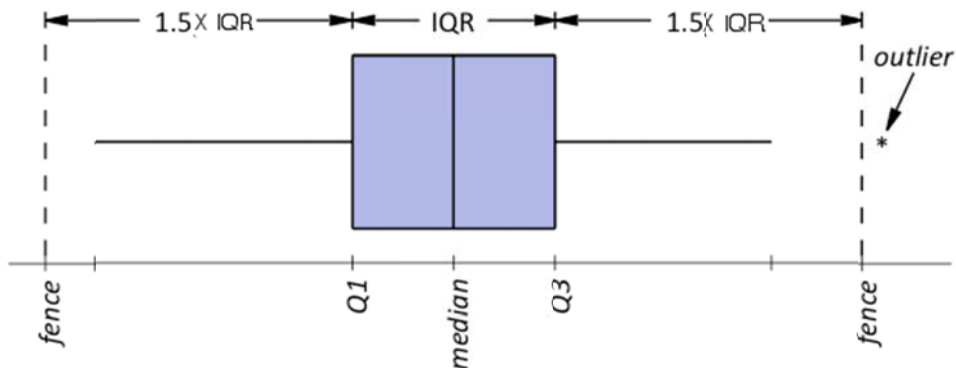
- A The ages are sorted. Find the five-number summary for the data and the IQR.
- B Find the fences for outliers. Use the fences to identify any outliers in the data.

Lesson 2.4

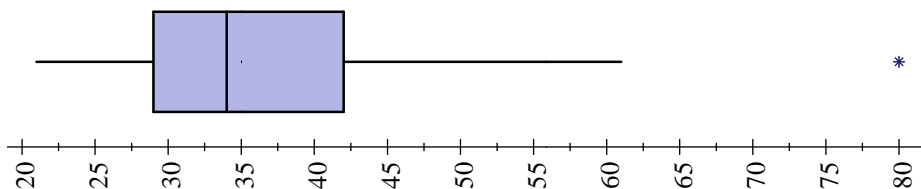
Quantifying Variability Relative to the Median

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 *horizontal* boxplot for a data set that contains an outlier.



The boxplot for the actress's ages is given below.



Lesson 2.4

Quantifying Variability Relative to the Median

SUMMARY

In this lesson, we focused on Two Measures of Variability(or spread):

1. *Range* RANGE = MAXIMUM – MINIMUM
2. *Interquartile range* (IQR) IQR = Q3 – Q1

The larger the IQR, the more variation (or spread) we see in the data.

The range is simple to compute, but it is influenced by extreme observations called outliers. The IQR offers a simple measure of variability. The IQR is not affected much by outliers or skewing. We say, the IQR is *resistant* to outliers.

Also, 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.

$$\begin{aligned}\text{Upper Fence} &= Q3 + 1.5 \cdot \text{IQR} \\ \text{Lower Fence} &= Q1 - 1.5 \cdot \text{IQR}\end{aligned}$$

The rule we use to determine outliers is based on a formula. We may have cases where values that are close to each other are labeled differently – where one value is determined to be an outlier but the other is not.

Boxplots provide a graph with a simple structure that represents center and variability. Boxplots are based on *five number summaries*.

Five Number Summary: Min, Q1, Med, Q3, Max

These five numbers break the data set up into four equal parts, each containing 25% of the data. Therefore, the middle 50% of the data falls between Q1 and Q3.

Lesson 2.4
Quantifying Variability Relative to the Median

STUDENT NAME _____ DATE _____

TAKE IT HOME PART I

- 1 The data below are pulse rates collected from 11 students, measured in beats per minute. Use this data to answer the following questions. Note the data has already been sorted from lowest to highest. You should do these calculations by hand.

62	63	65	65	67	68
72	76	78	82	88	

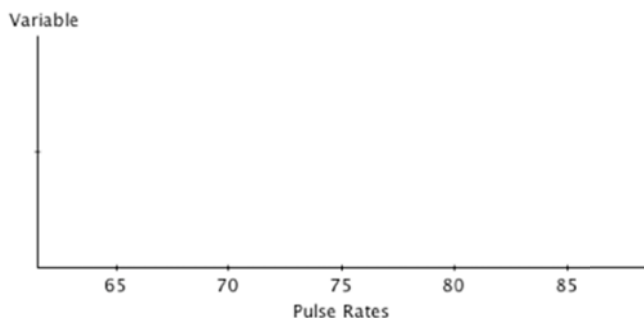
- A Calculate the values for the minimum, maximum, and range for the pulse rates.

- B Calculate the values for the first quartile, the median and the third quartile and report the five number summary. You should do these calculations by hand.

- C Think about what the word “quartiles” means. The quartiles divide the data set into four equal parts. What percentage of the pulse rates are greater than 78 beats per minute?

- D What percentage of the pulse rates are greater than 65 beats per minute?

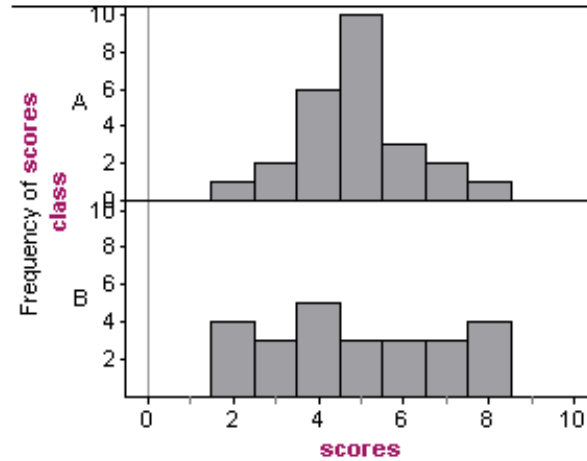
- E Construct a boxplot of this data. You should do this by hand. Use the number line given below and draw the boxplot horizontally.



Lesson 2.4

Quantifying Variability Relative to the Median

- 2 Here are the distributions of quiz scores for two classes. Each class has 25 students. The median for each class is 5.

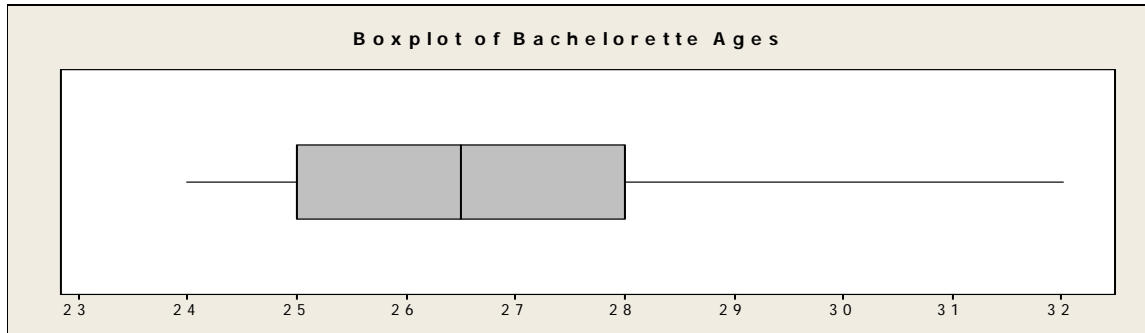


- A Which histogram has the smallest IQR, class A or class B? *Hint: The IQR is the spread in the middle half of the data. The IQR is small when a large portion of the data is close to the median.*
- B Your friend doesn't really understand your answer to part A. Write a few sentences to help her out.

Lesson 2.4

Quantifying Variability Relative to the Median

- 3 The ages of the 30 women on the 15th season (2010) of the television show “The Bachelor” was collected. The box-plot for the ages is shown below.



- A Looking at the box-pot, write down the five number summary for the ages of the 30 women on the TV show “The Bachelor” in the 2010 season.
- Min _____ Q1 _____ Median _____ Q3 _____ Max _____
- B Twenty – five percent of the bachelorettes’ ages are greater than what age?
- C What percentage of the bachelorettes’ ages are between 25 and 28?
- D By looking at the boxplot, can you tell if the distribution of ages is skewed? If it *is* skewed, is it skewed left or right?
- E Considering your answer to part D, which is likely to be larger, the mean or median age? Give a reason for your answer.
- F Fill in the blanks with either “mean” or “median”.
- For data that is skewed to the right, the _____ will be higher than the _____.

Lesson 2.4

Quantifying Variability Relative to the Median

- 4 Imagine that an error was made when the data values for typical monthly temperatures in St. Louis were recorded in the table below. The first two digits for the July temperature were reversed. It was supposed to be 79.8 but was written as 97.8.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Temperature	29.9	33.9	45.1	56.7	66.1	75.4	97.8	77.6	70.2	58.4	46.2	33.9

- A Calculate the range with this new incorrect temperature for July and compare it with the original range you calculated in class, 49.9°F. How does this error affect the range? Does it increase the range, decrease the range, or leave it unchanged?
- B Arrange the data shown in the table above to find the 5-number summary for this data set.
- C The 5-number summary for the original data, as we saw in the lesson was:

min = 29.9°F, Q1 = 39.5°F, med = 57.55°F, Q3 = 72.8°F, max = 79.8°F.

Write the IQR for each data set: the IQR for the correct data set we found in the lesson and the IQR for the data set in the table given above with the incorrect temperature recorded for July.

Correct temperature data set IQR _____ (use the 5-number summary given above)

Incorrect temperature data set IQR _____ (use your answer to part B above)

How does this error affect the IQR? Does it increase the IQR, decrease the IQR, or leave it unchanged?

- D Fill in the blanks by choosing either “is” or “is not”
The IQR _____ affected much by outliers and skewing. The range _____ greatly affected by outliers and skewing.

Lesson 2.4

Quantifying Variability Relative to the Median

STUDENT NAME _____ DATE _____

TAKE IT HOME II

While you're doing this homework, remember to refer to the Summary at the end of the lesson if you need a refresher of terminology or formulas.

- 1 This problem uses data from a previous lesson about variability in basketball scores.

The following tables show visiting team scores for a sample of games during the last week of 2006 (using a synthetic, rubber basketball) and the first week of 2007 (using the traditional leather basketball). You may use the calculator (or Statcrunch) to do this problem.

Scores of Visiting Teams in 2006					Scores of Visiting Teams in 2007				
68	76	77	80	80	95	97	74	100	99
81	82	84	85	87	102	78	89	87	79
89	89	91	92	92	86	88	91	105	105
96	97	98	99	100	74	123	96	91	93
101	103	111	112	114	80	97	104	92	117

- A Calculate the values for the minimum, maximum, and range for the data values in 2006.
- B Calculate the values for the minimum, maximum, and range for the 2007 scores.
- C Report the values of the first and third quartiles (Q1 and Q3) for the scores in 2006. Report the IQR for 2006.
- D Report the values of the first and third quartiles (Q1 and Q3) for the scores in 2007. Report the IQR for 2007.
- E Comparing the IQRs for the two data sets, which has more variability?
- F Calculate the fences for outliers for the 2006 scores. Identify any outliers.

Lesson 2.4

Quantifying Variability Relative to the Median

- G Calculate the fences for outliers for the 2007 scores. Identify any outliers.
- H Draw side-by-side boxplots for the 2006 and 2007 data sets. Remember to plot any outliers using separate points. Look back at problem number 10 in the lesson if you don't remember what a side-by-side boxplot is.
- I Compare the scores from each year based on the graphs. Do the graphs indicate that the synthetic ball, used in 2006, may have made a difference in visiting team scores? Make sure to talk about both center and variability. Refer to the specific graphs to support your conclusion.

Lesson 2.4

Quantifying Variability Relative to the Median

- 2 The following table contains the ages of the 30 women on the 15th season (2010) of the television show “The Bachelor.”

Bachelorette	Age
Alli	24
Ashley H	26
Ashley S	26
Britnee	25
Britt	25
Chantal	28
Cristy	30
Emily	24
J	26
Jackie	27
Jill	28
Keltie	28
Kimberly	27
Lacey	27
Lauren	26

Bachelorette	Age
Lindsay	25
Lisa M	24
Lisa P	27
Madison	25
Marissa	26
Meghan	30
Melissa	32
Michelle	30
Raichel	29
Rebecca	30
Renee	28
Sarah L	25
Sarah P	27
Shawntel	25
Stacey	26

- A Construct a histogram for the ages of the bachelorettes. Describe important features of the graph. You may use the calculator (or Statcrunch) to do this problem.

- B Report the values in the five-number summary for the ages of the bachelorettes.

Lesson 2.4

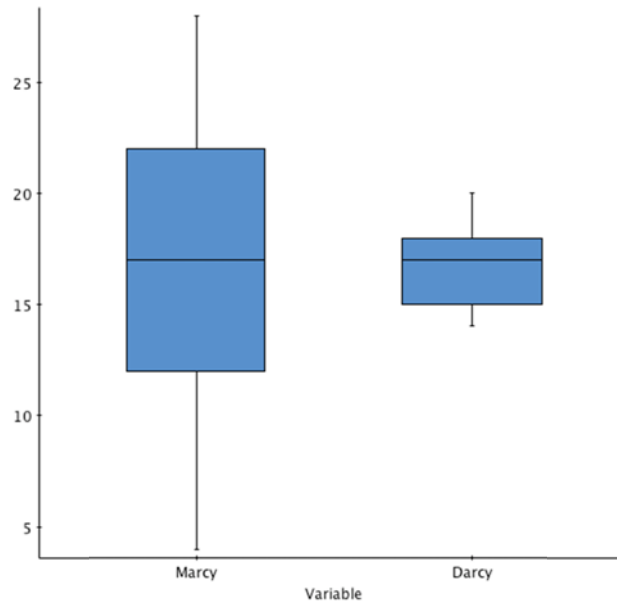
Quantifying Variability Relative to the Median

- C Construct a boxplot for the ages of the bachelorettes. Compare the boxplot you constructed in this question to the histogram you constructed in Question 3A.
- D Which do you think better represents the distribution of the ages of the bachelorettes? Explain your reasoning.

Lesson 2.4

Quantifying Variability Relative to the Median

- 3 The boxplots below show the points scored per game during the regular season for Marcy and Darcy.



- A** Write a few sentences summarizing the similarities and differences in the regular season performance of these two girls. Your summary should mention a comparison of the medians and the IQRs.
- B** The coach can only take one of these two players to the state championship. Which one should he take and why?

Lesson 2.5

Quantifying Variability Relative to the Mean

INTRODUCTION

Two community college students, Amelia and Ben, are taking a psychology course that has five exams and a final exam. Their five exams scores are shown in the table below.

Test	1	2	3	4	5
Amelia	75	98	85	80	62
Ben	65	62	83	92	98

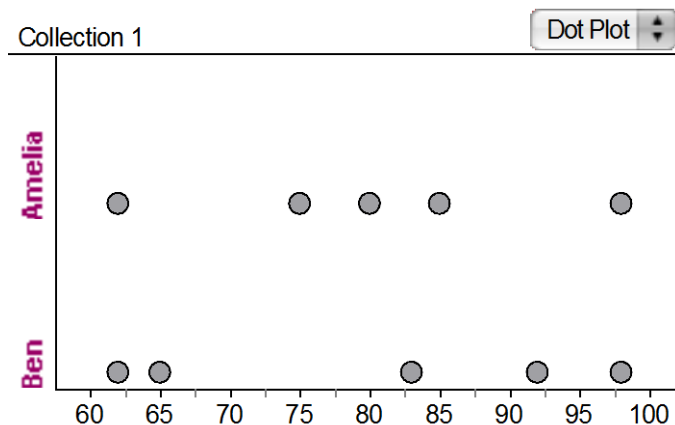
- 1 Which student appears to have the higher average? Check by calculating the average for each student. Recall from section 2.3 that the word **mean** is used in statistics for the common average.

Amelia's mean = _____

Ben's mean = _____

Is this what you guessed?

Even though the mean score was the same for both students, was one student more consistent than the other? Does one of the student's scores vary more than the other? The dot plots for the data are shown below. From the plots, which student appears to have more variation in their test scores?



In the previous lesson you learned to measure variability using the range and also the IQR (interquartile range). In this example both student's scores have the same range but from the plots we probably would not say they have the same variability. In this lesson you will learn about another statistic that is often used to measure the variation or spread in data. The measure is called the standard deviation and it is based on how much the data values vary from the mean. So when the center of a data set is described by the mean the variability will be described by the standard deviation. Now let's see what this means.

Lesson 2.5

Quantifying Variability Relative to the Mean

The mean is the average of all the data values. It is the most commonly used average. As a reminder, the formula for the mean is below:

$$\bar{x} = \frac{\text{sum of data}}{\text{number of data}} = \frac{\sum x}{n}$$

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

Previously, we computed the means of the test scores:

Amelia: $\bar{x} = 80$,

Ben: $\bar{x} = 80$.

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

$$\text{Deviation} = (\text{data value} - \text{mean}) = (x - \bar{x}).$$

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.

- 2 Calculate the deviation from the mean for Amelia on test 1. What does this number tell you about her score on test 1?
- 3 Complete the table below of deviations from the mean for Amelia. Make sure you keep in mind the proper sign for the deviation.

x	$x - \bar{x}$
75	$75 - 80 = -5$
98	
85	
80	
62	

Each data value *deviates* from the mean. In other words, they differ from the mean. Data values with large deviations (they differ from the mean by more) 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.

Lesson 2.5

Quantifying Variability Relative to the Mean

- 4 Looking at the deviations you calculated, ignore the signs of the deviations for a moment. What one number would you use to describe the “typical” amount of deviation from the mean? That is, what is your estimate for the “average” deviation from the mean? Commit to one number – we’ll check back with your estimate after we find out how to calculate this number exactly.

TRY THESE

- 5 A simple way to summarize the total variability is to find the average of the signed deviations. However, if we do this we run into a problem pretty quickly. In order to find the average of the deviations we calculated, we would first find the sum of these deviations. Compute the sum of the deviations in the table for Amelia’s deviations:

x	$x - \bar{x}$
75	-5
98	18
85	5
80	0
62	-18

- 6 We still want to summarize the total variability. What could you do to the deviations to keep them from adding to 0?
- 7 Below is our table of Amelia’s scores and deviations with a new column. In the new column, the deviations are squared. The first new entry is $(-5)^2 = 25$. Complete the column of square deviations. Remember squared numbers are ALWAYS positive, so the squared deviations are positive and the total will not be zero.

x	$x - \bar{x}$	$(x - \bar{x})^2$
75	-5	25
98	18	
85	5	
80	0	
62	-18	

Lesson 2.5

Quantifying Variability Relative to the Mean

Remember we are trying to end up with a number that describes the typical amount that a data value (test score in this example) differs from the mean. So we will now find the “average” of the squared differences. We add the squared differences and divide by $n - 1$ (the number of data values minus 1). Note that we divide by $n - 1$ instead of n , as we would when calculating an average. (The reason for this is subtle. We will not discuss it in this course.)

8 Calculate the sum of the squared deviations. Sum = _____

9 Divide the sum of squared deviations by $n-1$. _____

This number is called the **variance**. Lastly, to “undo” the squaring we did before we will take the square root.

10 Take the square root of the variance. _____

Finally! We have the **standard deviation**. The symbol for the standard deviation is s .

The standard deviation in the example is 13.210.

11 Does this seem like a reasonable number? Recall in question 4 above we asked you to ignore the signs of the deviations and estimate the typical deviation. Let’s do that now. The sizes (ignore sign) of the deviations from the mean were 5, 18, 5, 0 and 18. Is 13.2 a reasonable description of the list of deviations?

12 Look back at your estimate. How does your estimate compare to the calculated standard deviation, 13.21?

Lesson 2.5

Quantifying Variability Relative to the Mean

We can summarize the procedure to calculate the standard deviation with the following steps:

- Step 0 Calculate the mean.
- Step 1 Calculate the deviations from the mean.
- Step 2 Square the deviations
- Step 3 Sum the squared deviations.
- Step 4 Divide the sum by the number of data values minus 1, or $n-1$
- Step 5 Take the square root.

The steps above give a recipe for s using words.

For those who prefer a formula, we can write the formula for the standard deviation as

$$s = \sqrt{\frac{\text{sum of the squared deviations}}{n - 1}}$$

NEXT STEPS

To make sure you understand the process for calculating the standard deviation, s , go back to Ben's test scores and calculate the standard deviation. Make a table showing the deviations and their squares and then show how to finish the calculation of s . After you finish compare answers within your group.

13 Ben's scores are 65,62,83,92,98 and the mean is 80.

x	$x - \bar{x}$	$(x - \bar{x})^2$
65		
62		
83		
92		
98		

Computing the Standard Deviation with Technology

As you can imagine, with larger sets of data, calculating standard deviation is very tedious and subject to careless errors. We almost always use technology to compute the mean and standard deviation.

14 Could the standard deviation of a data set ever be negative?

15 Could the standard deviation ever be 0?

Lesson 2.5

Quantifying Variability Relative to the Mean

More Computing the Standard Deviation with Technology

- 16 Here are the typical monthly temperatures from St. Louis. Use technology to calculate the standard deviation for the monthly temperatures in St. Louis.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Temperature	29.9	33.9	45.1	56.7	66.1	75.4	79.8	77.6	70.2	58.4	46.2	33.9

- 17 In Lesson 2.4 homework you examined what happened to the range and IQR when we changed a single data value. We did this for the temperatures in St. Louis. Here are the temperatures again with the mistake in the July temperature.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Temperature	29.9	33.9	45.1	56.7	66.1	75.4	97.8	77.6	70.2	58.4	46.2	33.9

- A How do you think the wrong temperature for July will change the standard deviation? Will it increase, decrease or stay about the same?
- B Use technology to calculate the standard deviation for the monthly temperatures in St. Louis. How does this compare to the correct standard deviation we calculated above?

YOU NEED TO KNOW

Outliers and skewing have a large effect on the standard deviation. Recall this was also the case for the mean. We say 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.

Lesson 2.5

Quantifying Variability Relative to the Mean

SUMMARY

	Center	Variability or Spread
Use when data is roughly symmetric	Mean	Standard deviation
Use when data is skewed or has outliers	Median	Inner Quartile Range, IQR
Quick measurement, but not very useful		Range

We have examined two measures of center (mean and median) and three measures of variability or spread (range, interquartile range, and standard deviation). We use the IQR to describe variability when using the median to describe center. We calculated the standard deviation using deviations from the mean; so we use the standard deviation to describe variability when using the mean to describe center. To decide 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.
- The mean and the standard deviation are both influenced by outliers and strong 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.

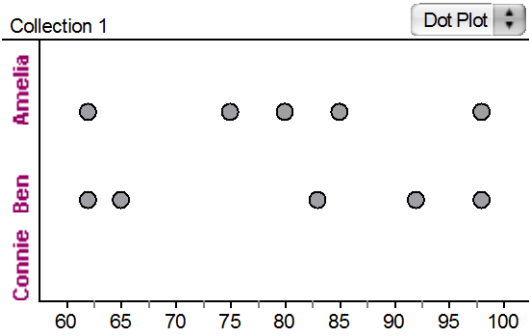
Lesson 2.5
Quantifying Variability Relative to the Mean

STUDENT NAME _____ DATE _____

TAKE IT HOME

1 Ben and Amelia have another friend Connie in the class. Her test scores were 74,77,81,83, and 85.

A Add her data to the dot plot below.



B Show the calculations in finding her mean and standard deviation.

C Explain how her statistics compare with Ben and Amelia. Also explain how these comparisons show in the plot.

Lesson 2.5

Quantifying Variability Relative to the Mean

2 Sally owns a small business with 4 employees. The weekly salaries are 300,300,400, 1000 dollars.

A Calculate the mean and standard deviation of their weekly salaries.

B Suppose that there is an end-of-year bonus of \$200 given to each employee. Calculate the mean and standard deviation of their weekly salaries for the week of the bonus.

C How did the mean change in the bonus week?

How did the standard deviation change in the bonus week?

Explain the changes in terms of how the bonus week affected a dot plot.

Lesson 2.5

Quantifying Variability Relative to the Mean

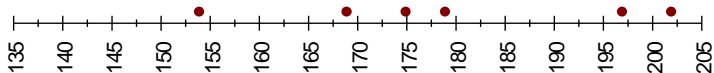
- 3 In Lesson 2.3 we examined the weight gains (in grams) over a one-month period for two samples of laboratory rats. One sample contained six adolescent rats that were given a high daily dose of a stimulant drug. This sample is called the stimulant group. The sample of six normal adolescent laboratory rats was the control group. This group received no treatment.

Here are the weight gains for the two groups:

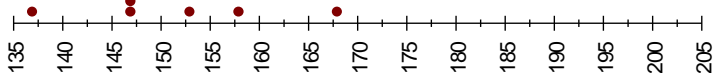
Control Group	169	154	179	202	197	175
Stimulant Group	137	158	153	147	168	147

These values are summarized in the dotplots below.

Control Group Weights



Treatment Group Weights



- A Use technology to compute the means and standard deviations for the weight gains in each group. Be sure to include the units.
- B Write a brief comparison of the standard deviations for the weight gains of the two groups. When making your comparison, think about differences, or variability, between the groups. Also think about whether the standard deviations show that the variability in the distributions is substantially different or not. Explain your reasoning.

Lesson 2.5

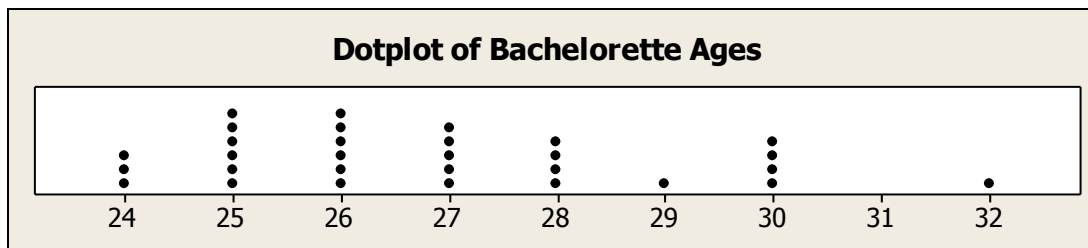
Quantifying Variability Relative to the Mean

- 4 Here are the names and ages for the 30 women on the 2010 season of the television show “The Bachelor.”

Bachelorette	Age
Alli	24
Ashley H	26
Ashley S	26
Britnee	25
Britt	25
Chantal	28
Cristy	30
Emily	24
J	26
Jackie	27
Jill	28
Keltie	28
Kimberly	27
Lacey	27
Lauren	26

Bachelorette	Age
Lindsay	25
Lisa M	24
Lisa P	27
Madison	25
Marissa	26
Meghan	30
Melissa	32
Michelle	30
Raichel	29
Rebecca	30
Renee	28
Sarah L	25
Sarah P	27
Shawntel	25
Stacey	26

The ages are summarized in the dotplot below.



- A The mean age is about 27. Look through the ages and by thinking about “typical deviations” (don’t worry about the sign) from the mean make an estimate of the standard deviation of the ages. (No calculation on paper or using technology is needed.)
- B Use technology to find the standard deviation for the ages of the women on the show. Include units with your answer.
- C How close was your estimate in part A to the actual value in part B?
- D Explain what the standard deviation means in the context of their ages.

Lesson 2.5

Quantifying Variability Relative to the Mean

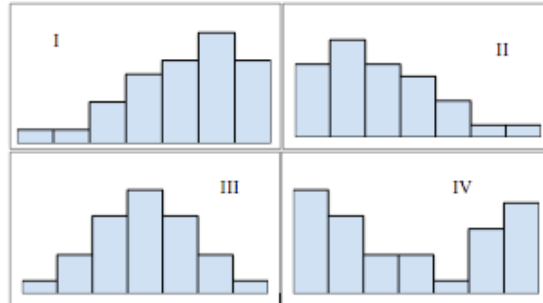
- 5 Below are salaries of thirteen players on the Cleveland Cavaliers basketball team during the 2009-2010 season. This was LeBron James' last season playing for the Cavaliers.

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

- A** Use technology to compute the mean and standard deviation of these values.
- B** Delete LeBron's salary of \$21,000,000 and re-compute the mean and standard deviation. What was the effect on the mean and standard deviation?
- C** What was the effect on the mean and standard deviation? What do you think caused this?

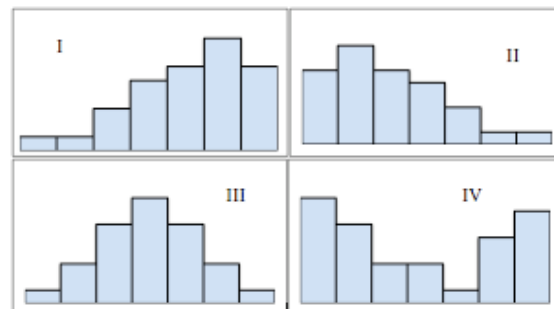
Chapter 2 Review

Assume that the histograms are drawn on the same scale. Which of the histograms has the largest interquartile range (IQR)?



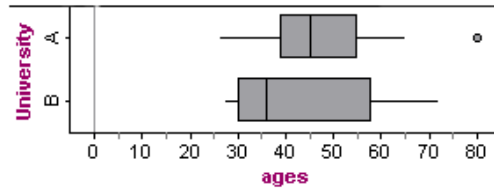
- A. Histogram I
- B. Histogram II
- C. Histogram III
- D. Histogram IV

Assume that the histograms are drawn on the same scale. Which of the histograms has the largest median?



- A. Histogram I
- B. Histogram II
- C. Histogram III
- D. Histogram IV

Which University has a less variability in the ages of their faculty? Pick the statement that gives the best reason.



- A. University A because the IQR is smaller.
- B. University A because the overall range is smaller.
- C. University B because there is no outlier.
- D. University B because the standard deviation is smaller.

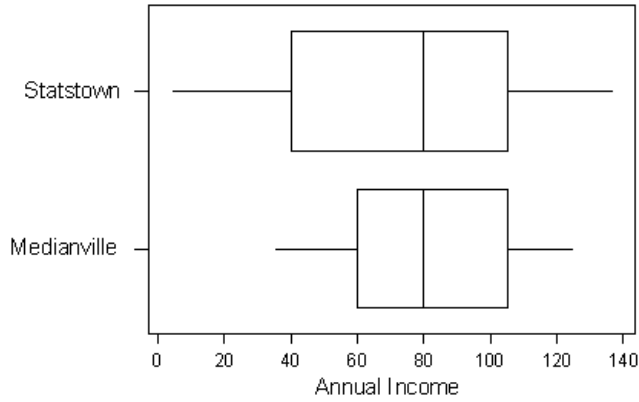
The Boston marathon is a very competitive race. To qualify for the Boston marathon, male runners must have completed a marathon in less than 3 hours and 5 minutes within the last year.

Other marathons, such as the Chicago marathon, have no qualifying times. Anyone is able to run this race, even without completing a different marathon earlier in the year.

Consider the groups of runners of each race: Boston marathon runners and Chicago marathon runners. Which group would most likely have the larger standard deviation?

- A. The Chicago marathon would more likely have a higher standard deviation than the Boston marathon.
- B. The Boston marathon would more likely have a higher standard deviation than the Chicago marathon.
- C. They would likely have the same standard deviation because many of the same people run both races.
- D. There is not enough information to predict the relative sizes of their standard deviations.

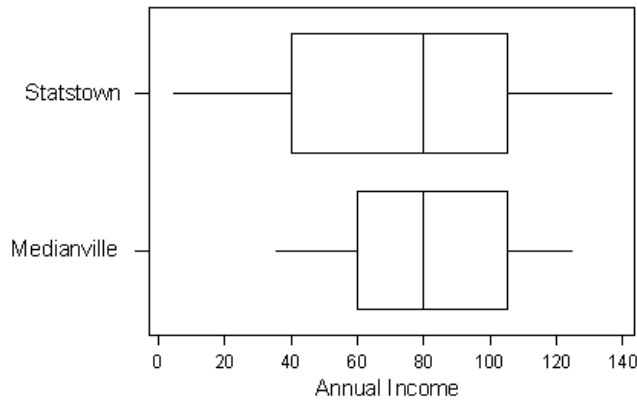
The boxplots below display annual incomes (in thousands of dollars) of households in two cities.



Which city has more households?

- A. Statstown
- B. Medianville
- C. Both cities have the same number of households.
- D. It is impossible to tell from the boxplots.

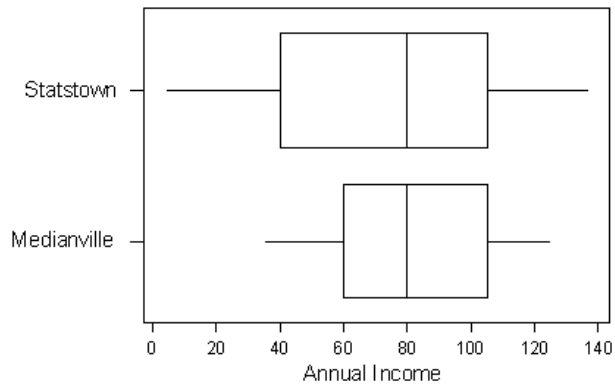
Here again are the boxplots showing annual incomes (in thousands of dollars) of households in two cities.



Which city has greater variability in income?

- A. Statstown
- B. Medianville
- C. Both cities have the same number of households.
- D. It is impossible to tell from the boxplots.

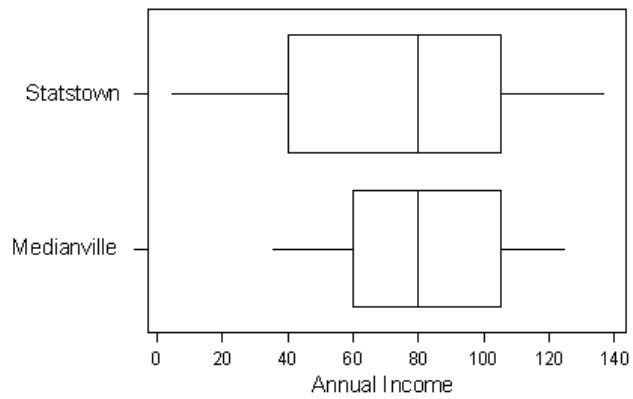
Here again are the boxplots showing annual incomes (in thousands of dollars) of households in two cities.



Which city has a greater percentage of households with annual incomes above \$80,000?

- A. Statstown
- B. Medianville
- C. Both cities have about the same percentage.
- D. It is impossible to tell from the boxplots.

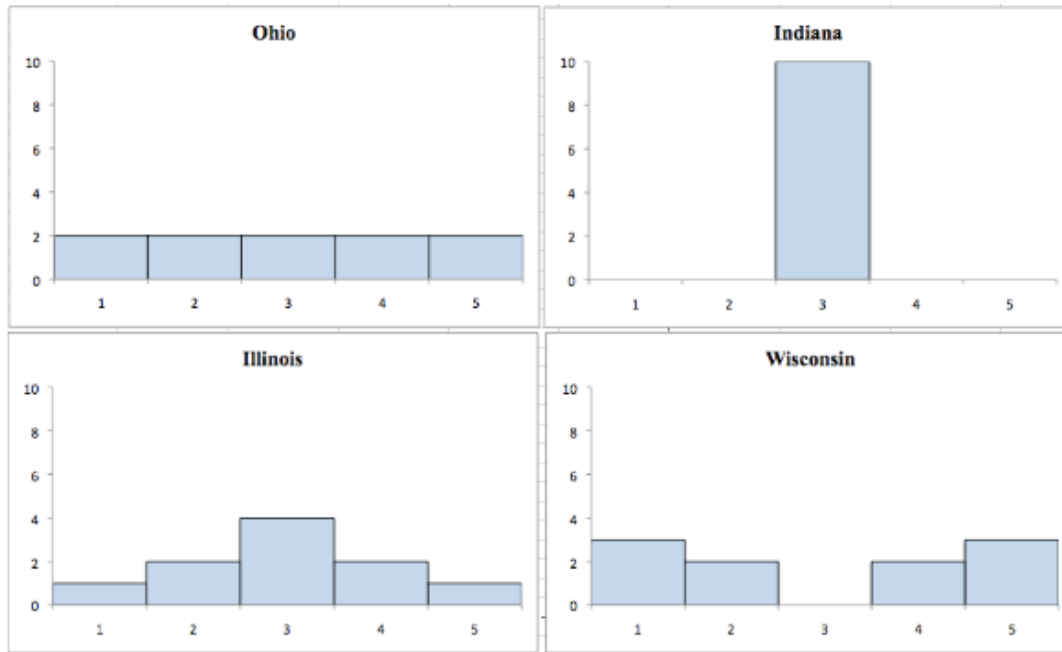
Here again are the boxplots showing annual incomes (in thousands of dollars) for households in two cities.



Which city has a greater percentage of households with annual incomes between \$50,000 and \$80,000?

- A. Statstown
- B. Medianville
- C. Both cities have the same number of households.
- D. It is impossible to tell from the boxplots.

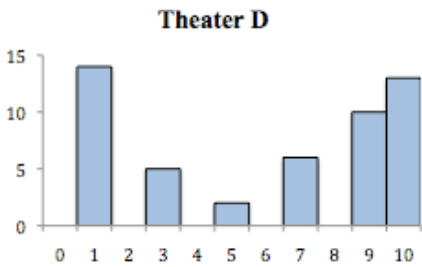
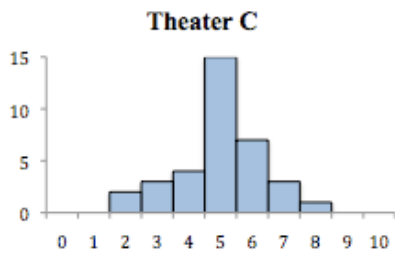
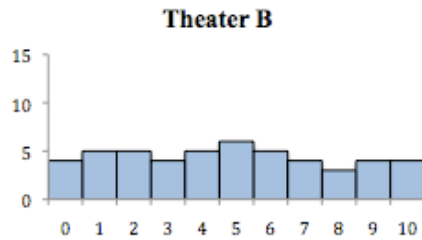
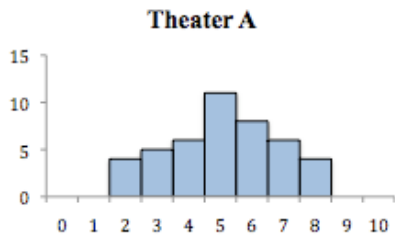
A textbook distributor has 10 employees in each of four midwestern states: Ohio, Indiana, Illinois, and Wisconsin. The variable is the number of unexcused absences in the last year. For each state, the mean number of unexcused absences is 3.



In which state is the standard deviation of unexcused absences zero?

- A. Ohio because all employees had the same number of unexcused absences.
- B. Indiana because all employees had the same number of unexcused absences.
- C. Illinois and Wisconsin because the distributions are symmetric.

Time spent waiting in line for each moviegoer at each of four movie theaters in Downtown San Francisco was measured on a Saturday night. The frequencies for each amount of minutes spent waiting in line are shown on the histograms below.



Consider the variable: time spent waiting in line. Which of the theaters would you expect to have the lowest standard deviation, and why?

- A. Theaters A and C because they look more normal than the others.
- B. Theaters A and C because they have the smallest range.
- C. Theater B because it is roughly uniform.
- D. Theater C because it has the most values close to the mean.
- E. Theater D because there appears to be no pattern to the responses.

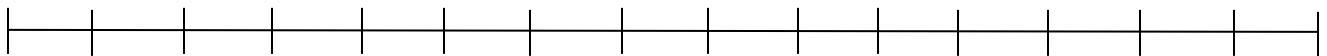
1. You and a friend are talking about collectible items. Your friend likes to collect wooden toys but complains that toy companies don't make them much anymore. In fact, your friend claims that because of this they are often more expensive. She says that you really can't find any wooden toys under \$5. In an effort to investigate your friend's statement, you go to the internet and collect the data below on prices for wooden toys from various toy stores.

Prices of Toys from Various Toy Stores	
Toy	Price
1	\$0.50
2	\$0.65
3	\$0.90
4	\$0.99
5	\$1.12
6	\$1.35
7	\$1.39
8	\$1.45
9	\$1.70
10	\$1.74
11	\$1.85
12	\$1.99
13	\$1.99
14	\$2.00
15	\$2.15
16	\$2.60
17	\$2.85
18	\$3.00
19	\$3.99
20	\$4.20
21	\$4.75
22	\$5.12
23	\$5.81
24	\$6.24
25	\$7.36
26	\$8.69
27	\$9.08
28	\$10.00
29	\$11.50
30	\$11.59
31	\$12.20

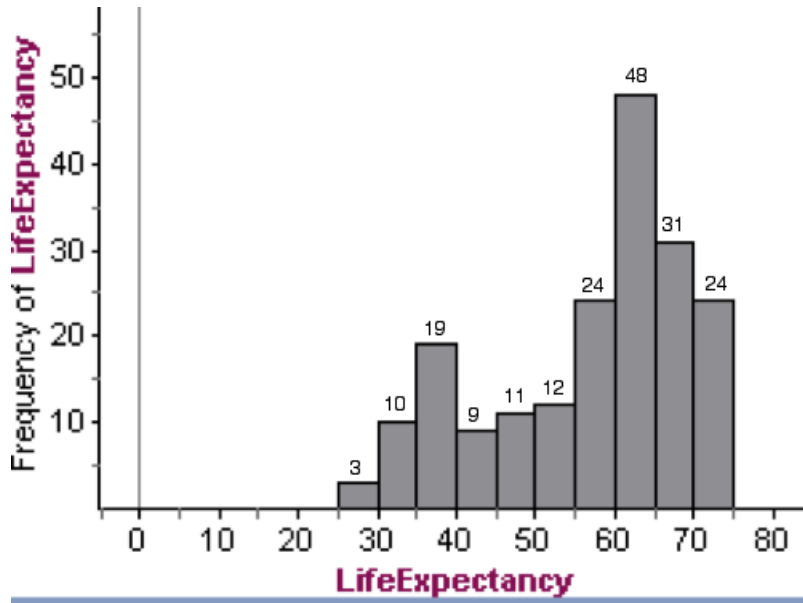
A) Make a dot plot of this data. You may use the number line at the bottom of this page but label the axis to show your scale.

B) Write a few sentences describing the distribution (shape, center and spread).

C) What does the data tell you about your friend's statement? Is it true? False? Explain.

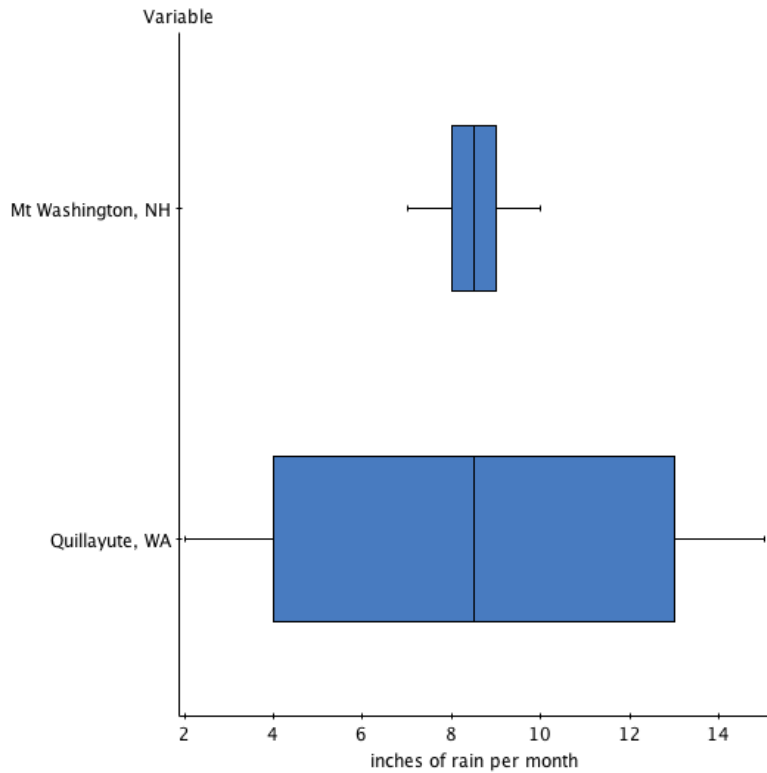


2. Does your life expectancy (number of years you live) vary much from one country to another? The World Health Organization has collected and analyzed data to answer this question (among others). They calculated the life expectancy for each of 191 countries around the world. The histogram below summarizes the distribution of life expectancy in the 191 countries. The numbers given above each bar is the height of the bar.



- A) What does the bar labeled with 3 mean, in terms of life expectancy and number of countries?
- B) The life expectancy in the U.S. was 70. How many countries had life expectancy as long or longer than the U.S.?
- C) What percentage of the 191 countries had life expectancy less than 40 years? Show your work.
- D) Describe the shape of the distribution.

3. Micha lives in Mt. Washington, NH and Quentin lives in Quillayute, WA. The two are friends and use the Internet to Skype quite frequently. Micha says he thinks it's easier to predict the monthly rainfall where he lives in NH than it is where Quentin lives in WA. To settle the question, they collect monthly rainfall data from the National Climatic Data Center's website. The boxplots below show this data for each city.



Use the boxplots to write a few sentences to compare the distributions of rainfall for the two cities. You should point out any similarities and differences. You should also mention the shape (symmetric, skewed left or right), center and spread.

Based on the data they gathered, was Micha right about rainfall prediction in NH versus WA? Explain.

Chapter 3

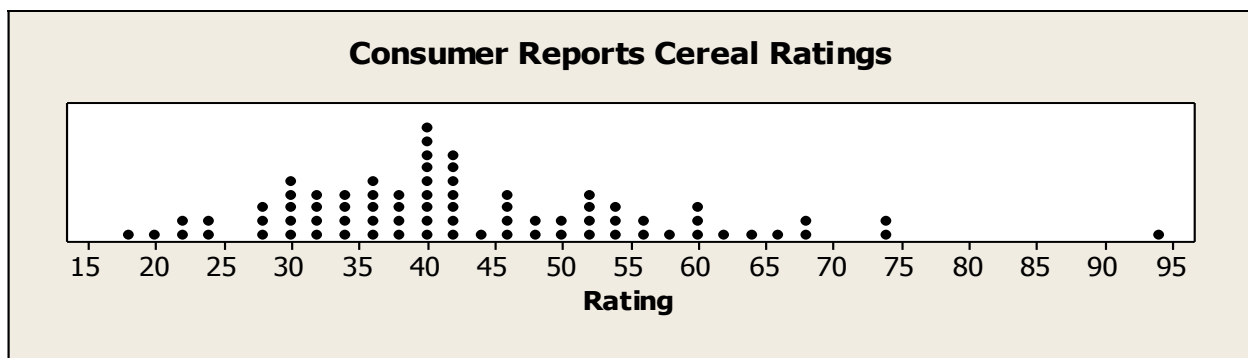
Exploring Bivariate Relationships

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

NEXT STEPS

- 2 *Consumer Reports* magazine reviews and rates many products. It reviews and rates the products to help people make informed buying decisions. The magazine uses its own criteria, or standards, and does all its own product testing. *Consumer Reports* is published by a nonprofit organization called the Consumers Union, whose mission it is to work for a fair, just, and safe marketplace for all consumers. (Retrieved from www.consumerreports.org/cro/aboutus/mission/overview/index.htm) We will now explore the *Consumer Reports* nutritional ratings for 77 breakfast cereals. *Consumer Reports* uses a rating system with a scale of 0 to 100. Here is the distribution of *Consumer Reports* ratings for 77 cereals:



- A What does each dot represent in this distribution?
- B For this distribution, what seems to be an average rating?
- C What is the range for the ratings? How would you describe the variability in ratings?

What we cannot tell from the dotplot is how the cereal ingredients (such as sugar or fat) are related to the ratings. To investigate how two variables are related to each other, we need a new type of graph, called a **scatterplot**. Scatterplots show the relationship between two quantitative variables.

Language Tip

Each point on a *scatterplot* represents two measurements. Scatterplots show the relationship between two quantitative variables.

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

The scatterplots in Supplement 3.1A show the amount of an ingredient in a serving of a cereal and the *Consumer Reports* rating for that cereal. Each graph has 77 points, one for each of the 77 breakfast cereals.

The *Consumer Reports* rating formula is not made public. We do not know which ingredients are used in its rating formula. In this lesson, we will try to identify the more important ingredients for their rating. We will use the data to figure out which ingredients *Consumer Reports* may, or may not, use in their rating formula. The only clues we have are these scatterplots. The first step in this investigation is to answer the following two questions.

Two new cereals are being rated by Consumer Reports. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.

- 3 Predict the *Consumer Reports* rating for the two cereals based on the data in the scatterplots. Tell how you used the scatterplot to help you make your predictions.

- 4 Your prediction is probably more accurate for one of the cereals more than the others. For which one do you think your prediction is more accurate (more likely to be closer to the actual *Consumer Report* rating)? Why?

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

TRY THESE

Reading and Interpreting Scatterplots

We are going to take a short detour from our investigation into which ingredients are the best predictors of *Consumer Reports* ratings. Here, we will work on interpreting scatterplots just to make sure everyone is comfortable reading this type of graph.

- 5 Captain Crunch has the lowest *Consumer Reports* rating of the 77 cereals in the data set. How much fat is in a serving of Captain Crunch?
- 6 In this set of 77 cereals, Product 19 (that's the name of the cereal) has the most sodium in a serving. What is the *Consumer Reports* rating for Product 19?
- 7 All-Bran Extra Fiber is the cereal with the highest rating. How much fat, sugar, protein, and sodium are in a serving of All-Bran Extra Fiber?

NEXT STEPS

Seeing Patterns and Relationships in Scatterplots

Now we will continue our detective work with *Consumer Reports* ratings. We will try to identify ingredients that are good predictors of ratings and ingredients that are not good predictors of ratings. Specifically, we will focus on how patterns in the data help us identify ingredients that are good predictors.

- 8 There are five cereals that have 3 grams of fat in a serving. Estimate the ratings for these five cereals. Why do you think the ratings are not all the same?

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

- 9 Imagine that a cereal has 0 grams of fat in a serving and a rating of 60. The cereal company has decided to increase the amount of fat to 3 grams in a serving. Do you think the *Consumer Reports* rating will most likely increase, decrease or remain about the same? Or do you think that it is impossible to use the scatterplot to predict the impact of this change on the rating? How does the pattern in the data support your decision?

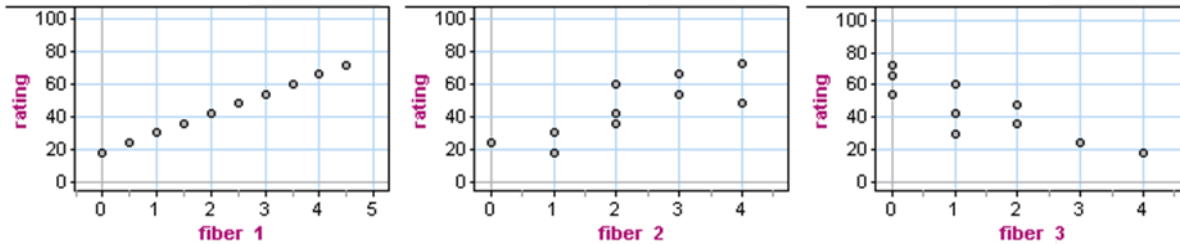
In order to crack the code on the *Consumer Reports* rating we need to understand how each of the ingredients influences the ratings. We will think about whether the ingredients have positive effect(or impact) on the ratings, a negative effect(or impact) on the ratings, and how strong an effect each ingredient has. Use the scatterplots to answer the next few questions.

- 10 Which ingredients have a positive effect on the *Consumer Reports* cereal ratings? How can you tell if an ingredient has a positive effect on rating? Think about the patterns you see in the scatterplots.
- 11 Which ingredients have a negative impact on the *Consumer Reports* cereal ratings?
- 12 Which ingredient seems to have the strongest impact on *Consumer Reports* rating? In answering this question, think about which ingredient would help you make the most accurate prediction for the rating of a cereal. Tell how the patterns in the scatterplots help you make your decision.

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

- 13 Think about how the amount of fiber in a cereal might relate to the *Consumer Reports* rating. Here are three scatterplots with data from 10 imaginary cereals. Which scatterplot displays a pattern similar to what you might see in the actual data? Why? In answering the question think about whether fiber is a healthy or not a healthy ingredient in cereal.



SUMMARY

In this lesson we learned the following facts about relationships between quantitative variables.

- If two quantitative variables are measured for each individual in the sample, we can use a scatterplot to represent the data.
- Each point on the scatterplot represents one individual with measurements of two quantitative variables.
- We look for patterns in scatterplots. One thing we look for is the direction. The direction can be **positive** or **negative**. The association between two variables is **positive** if larger values of *the variable plotted on the horizontal axis* tend to correspond to larger values of *the variable plotted on the vertical axis*. The association between two variables is **negative** if larger values of *the variable plotted on the horizontal axis* tend to correspond to smaller values of *the variable plotted on the vertical axis*.
- We also look for strength in a scatterplot. The association between two variables is considered **strong** when the pattern of the points is very clear. The association is **weak** if there is no clear pattern in the scatterplot. Strong relationships lead to more accurate predictions.

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

STUDENT NAME _____ DATE _____

TAKE IT HOME

- 1 The mean *Consumer Reports* rating for these 77 cereals is 44. What is the largest amount of sugar per serving in a cereal that has a rating above the 44?

- 2 *Consumer Reports* is rating a new cereal. The cereal has 175 milligrams of sodium in a serving. Use the scatterplots to predict the *Consumer Reports* rating for this new cereal.

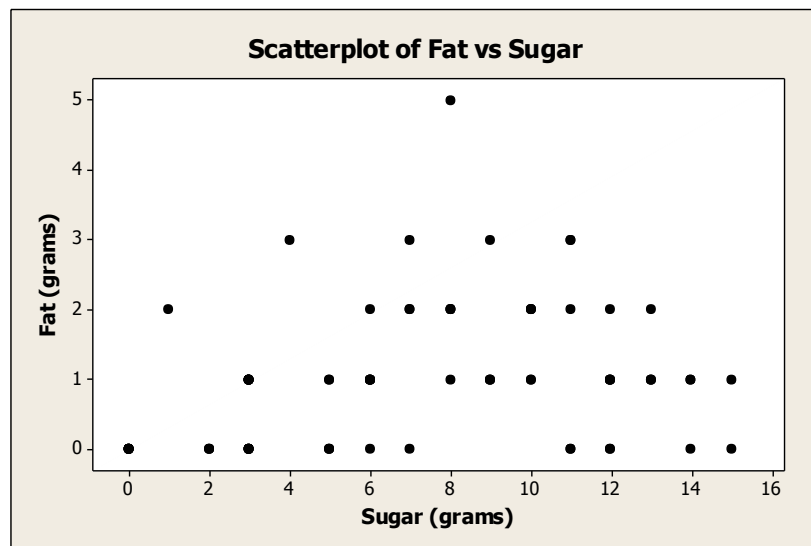
- 3 Does sugar or sodium give more accurate predictions for *Consumer Reports* ratings? Explain how the scatterplot supports your answer.

Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

- 4 A friend says that she only pays attention to sugar amounts. But, when she talks about what she eats, she also is concerned about fat. She believes low levels of sugar in a food indicate the food also has low amounts of fat. She also believes when a food has high levels of sugar it also has high amounts of fat.

The scatterplot below shows the sugar and fat content of cereals. Think about your friend's beliefs about patterns of fat and sugar in food. Does the pattern your friend describes appear to be true for the cereals in the scatterplot? Explain how the scatterplot supports your answer.



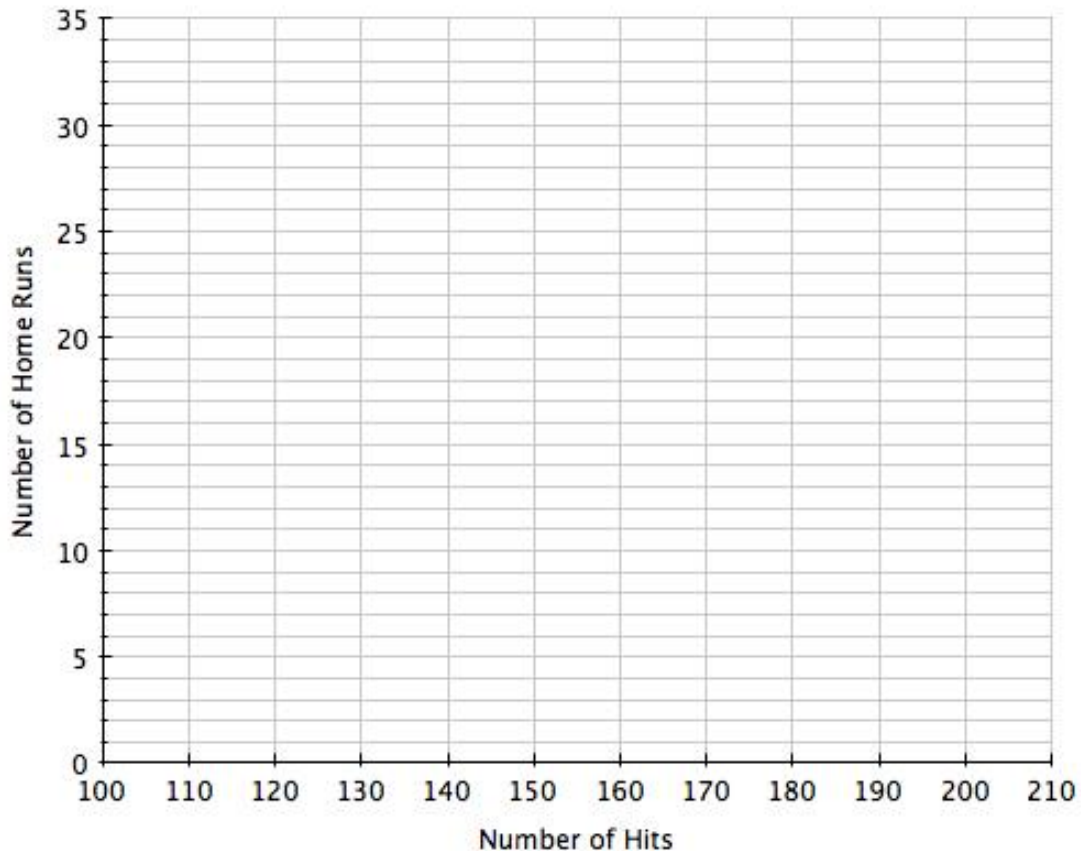
Lesson 3.1

Introduction to Scatterplots and Bivariate Relationships

- 5 If a major league baseball player gets more hits, will the player also get more home runs? We want to examine the relationship between the number of hits and the number of home runs made by professional baseball players. The table below shows a random sample of 10 baseball players from the 2010-2011 season. The table shows the players' number of hits and the number of home runs

Number of Hits	Number of Home Runs
119	7
128	12
109	14
125	18
135	34
111	17
195	21
163	13
207	10
163	31

- A Using this data, draw a scatterplot on the graph below.



Lesson 3.1

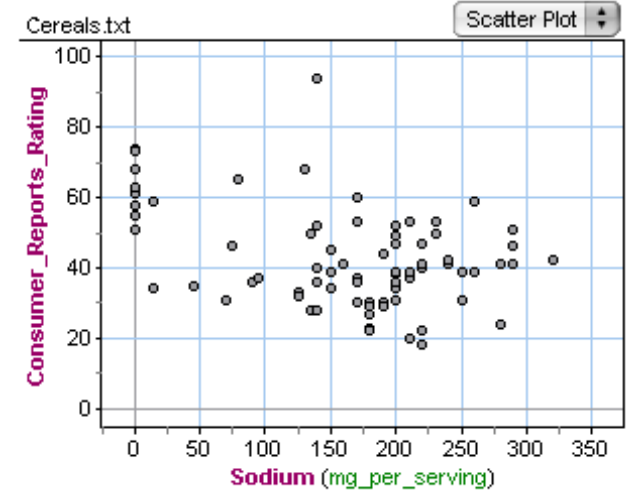
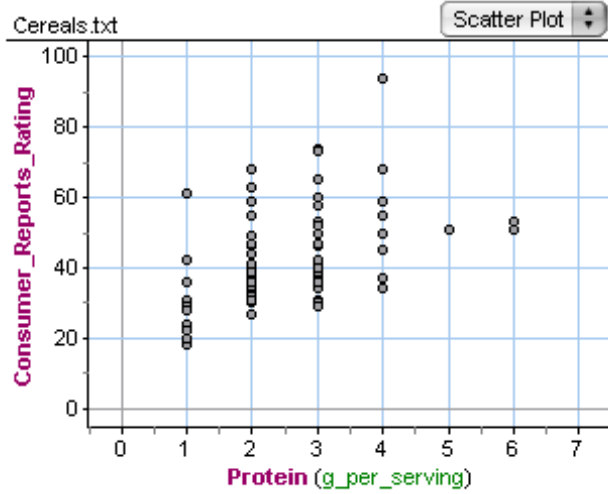
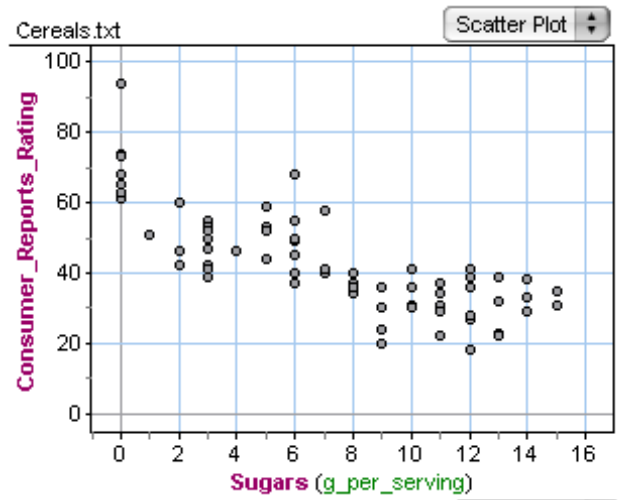
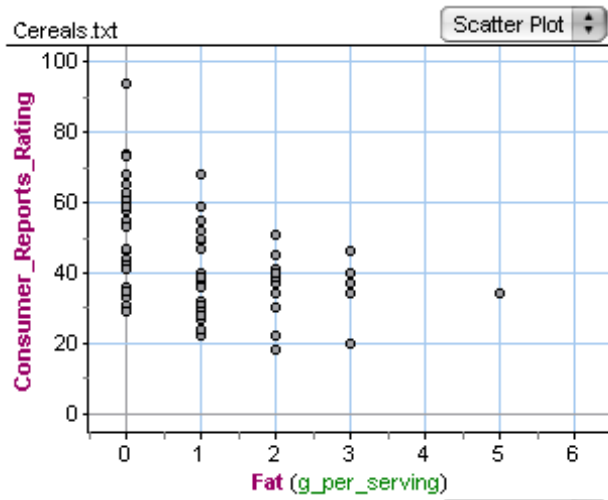
Introduction to Scatterplots and Bivariate Relationships

- B What does each dot in this scatterplot represent?

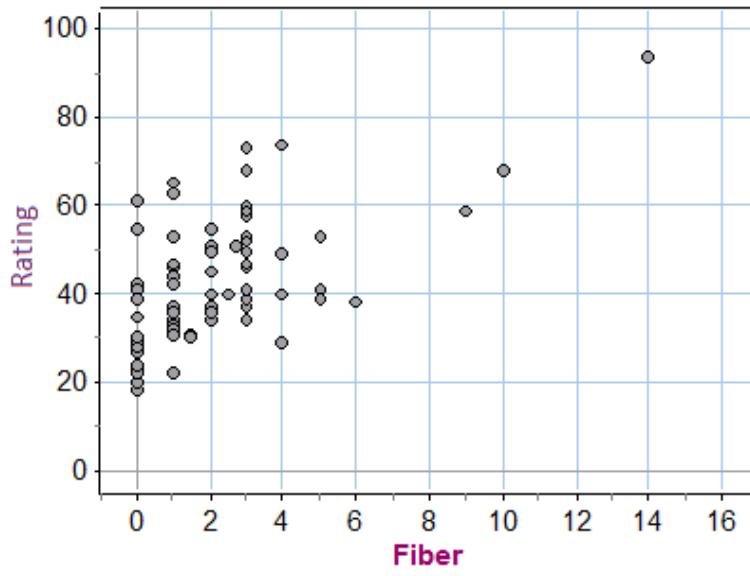
- C Based on the scatterplot, should you use the number of hits a player gets to predict the number of home runs the player will get? Explain how the scatterplot supports your answer.

Lesson 3.1 Supplement A

Scatterplots



Lesson 3.1 Supplement B
Scatterplots



Lesson 3.2

Introduction to Graphing Lines in Statistics

INTRODUCTION

In the previous lesson we learned that if two quantitative variables are measured for each individual in a sample, we can use a scatterplot to represent the data. When we examined the scatterplot, we discovered that each point on the scatterplot represents one individual with measurements of two quantitative variables. If we think one variable influences the other, we call the variable that explains the changes we see the **eXplanatory variable** and the variable that changes, the **response variable**. The eXplanatory variable is on the horizontal or *X*-axis and the response variable is on the vertical or *Y*-axis.

1 Here is a table of data displaying the heights and weights of ninth-grade girls.

A Do you think one variable influences the other?

B Which variable is the eXplanatory variable? In other words, which variable seems to explain the changes you see in the values of the other variable?

C Which variable is the response variable? In other words, which variable seems to change based on the values of the other variable?

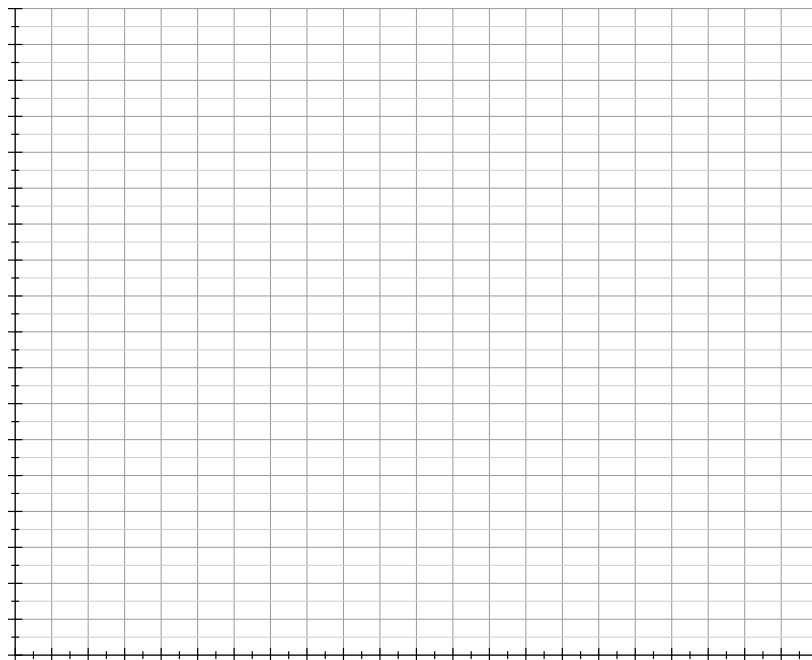
D If the eXplanatory variable is on the horizontal or *X*-axis and the response variable is on the vertical or *Y*-axis, then which axis would each variable be on if we were to make a scatterplot of this data?

Height (in.)	Weight (lbs.)
65	105
65	125
62	110
67	120
69	140
65	135
61	95
67	130

Lesson 3.2

Introduction to Graphing Lines in Statistics

- E Draw a scatterplot for this data. Make sure to label both the x and y axis with the appropriate scale and name.



2 The location of any point in a graph can be described using an **ordered pair**. The ordered pair is written in the form (x, y) , and provides the location of the point along the X-axis and the location of the point along the Y-axis.

- A Pick a point from the scatterplot you have drawn above and give the ordered pair that describes its location.
- B Circle the point $(67, 130)$. What does this point represent in terms of weights and heights?
- C Make-up 3 imaginary points that you think could reasonably be a part of this data set. Plot them and give the ordered pair that describes their location.
- D Explain why you think the points you selected for part C would be “reasonable” observations.

Language Tip

Each point on a scatterplot has a location that can be described using an *ordered pair*. Ordered pairs are written in the form (x, y) .

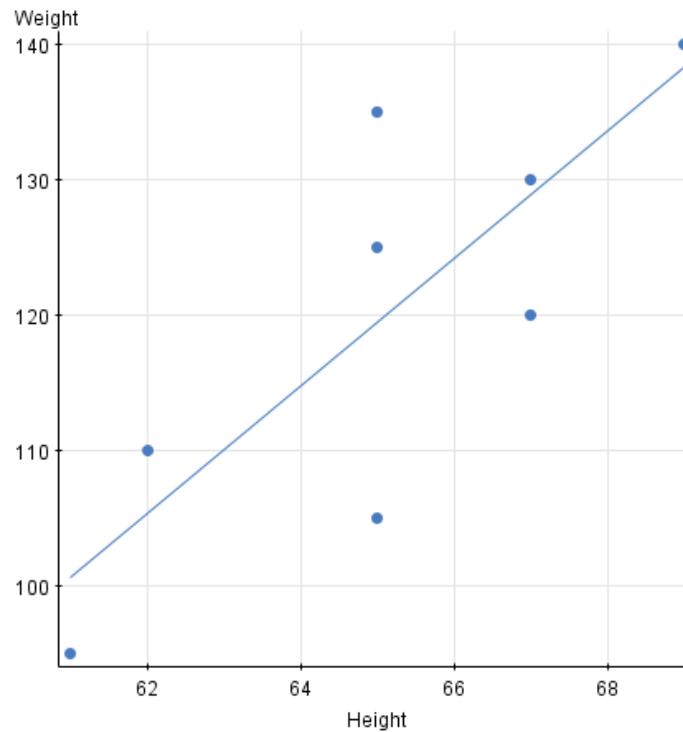
Lesson 3.2

Introduction to Graphing Lines in Statistics

NEXT STEPS

Describing Patterns or Trends in Scatterplots

3 In working with the height and weight data, we noticed a pattern in the relationship between the variables in the scatterplot. Some relationships between the variables in scatterplots can be summarized well by a line and we find that to be true here. Take a look at the scatterplot below summarized by a line.



Do you think this line summarizes the pattern well? Why or why not?

In fact, this line is the best possible summary of the data. We know this because the line shown here is what we call the **line of best fit**. You will learn more about this line, how it is calculated and how it is determined to be “the best” later on in Chapter 3. For now, we simply want to explore how this line helps us to understand what is happening in the picture we see and a little bit about how it can help us make predictions. Here is the equation for the line of best fit:

$$\hat{y} = -186.5 + 4.7(x)$$

or

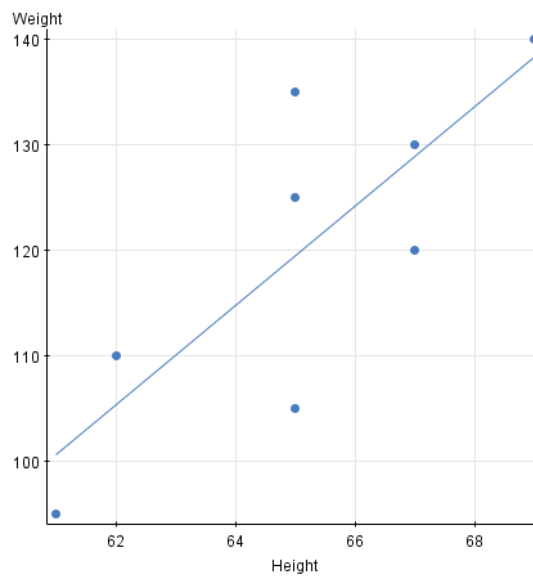
$$\text{Predicted Weight} = -186.5 + 4.7(\text{Height})$$

(Notice that when you use letters to represent variables in the equation of the line, you put a “hat” on the y. The hat is a signal that the variable is a *predicted value*, not an actual observed value.)

Lesson 3.2

Introduction to Graphing Lines in Statistics

- 4 Let's look more closely at our scatterplot and the data we collected.
- A How many observations were collected in the data set? In other words, how many individual ninth-grade girls' heights were recorded?
 - B What were the observed heights that were recorded?
 - C What if we were curious as to what the weight would be of a ninth-grade girl who was 63 inches tall, could you give a reasonable prediction of what you would expect her weight to be? Give your best prediction and explain how you came up with that weight.
 - D Give the height and the corresponding predicted weight you estimated as an ordered pair.
 - E Plot your estimated predicted point on the scatterplot below.



Lesson 3.2

Introduction to Graphing Lines in Statistics

5 Now, let's take another look at the equation for our line of best fit. Let's see if the equation could help us to make a prediction about the weight of a ninth-grade girl that is 63 inches tall.

$$\text{Predicted Weight} = -186.5 + 4.7(\text{Height})$$

- A Looking at the equation again, how do you think we can use it to make a prediction?

- B If we use the equation to make the prediction, what value do we get for the *predicted weight*?

- C Give the height and its corresponding predicted weight as an ordered pair.

- D Plot your predicted point on the scatterplot above.

- E What do you notice about the two points you plotted?

TRY THESE

6 Use the equation of the line of best fit to make a prediction about the weight of a ninth-grade girl that is 67 inches tall.

$$\text{Predicted Weight} = -186.5 + 4.7(\text{Height})$$

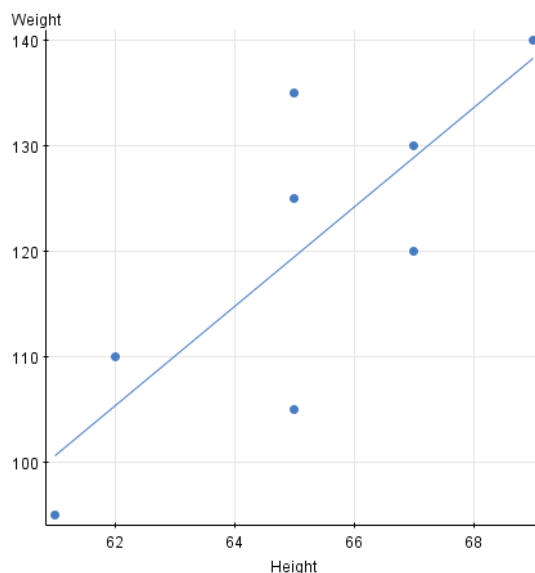
- A What is the predicted weight for this student?

- B Give the height and its corresponding predicted weight as an ordered pair.

Lesson 3.2

Introduction to Graphing Lines in Statistics

C Plot your predicted point on the scatterplot below.



Notice that the predicted point lies on the graph of the line of best fit. There are two other points that correspond to a height of 67 inches. These points represent actual observed values of weight that were recorded for ninth grade girls with a height of 67 inches at the time the data were collected. You can see that these points do NOT lie on the graph of the line of best fit. Although, it is possible for the line of best fit to travel through some of our observed data values in the scatterplot, observed values may or may not be on the line. However, when you use the equation of the line of best fit to find predicted values, the corresponding points will **ALWAYS** be on the line.

SUMMARY

In this lesson we learned the following facts about relationships between quantitative variables.

- Each point on the scatterplot has a location that can be described by an ordered pair. The ordered pair describes the location along the x-axis and the location along the y-axis in the form (x, y) .
- The predictor or eXplanatory variable is on the horizontal axis or X-axis of the scatterplot. The response variable is on the vertical axis or Y-axis of the scatterplot.
- An overall upward or downward trend in the data can be described using a line, and that line has a specific equation.
- The equation of the line can be used to predict values that were not observed in the data and when you do this you create a point that is on the graph of the prediction line.

Lesson 3.2

Introduction to Graphing Lines in Statistics

STUDENT NAME _____ DATE _____

TAKE IT HOME

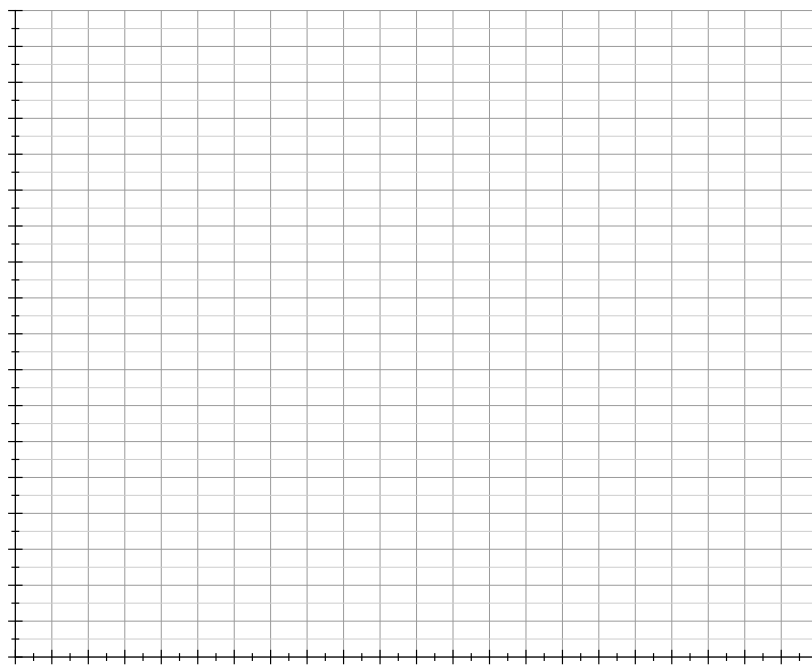
- 1 Data was collected on the fat and calorie content of fast food hamburgers. The number of calories in a hamburger was found to be a good predictor for the total grams of fat the hamburger contained. Below is a table of the data that were collected.

Calories	Total Fat (g)
410	19
580	31
590	34
570	35
640	39
680	39
660	43

A Which is the explanatory variable?

B Which is the predictor variable?

- C Use the data provided in the table to complete the scatterplot below. Be sure to label the x-axis and y-axis with the appropriate name and scale.



Lesson 3.2

Introduction to Graphing Lines in Statistics

- 2 Here is the equation of the line of best fit for the hamburger data.

$$\hat{y} = -15 + 0.08(x)$$

or

$$\text{Predicted Fat} = -15 + 0.08(\text{Calories})$$

- A Use the equation to predict two points that fall within the values provided in the table (for example, when selecting which values to plug in for number of calories you should choose values between about 400 and 700 calories). Write the corresponding predicted points as an ordered pair.
- B Plot your predicted points on the scatterplot above and use them to draw the line of best fit.
- 3 Open your internet web browser and go to the following website:

<http://www.shodor.org/interactivate/activities/SlopeSlider/>

- A Use the first slider (it should be purple) to change the value of the number that sits directly in front of x . What happens as you change this number?
- B What happens when the number is positive?
- C What happens when the number is negative?
- D Use the second slider (it should be green) to change the value of the number that stands alone following the term containing x . What happens as you change this number?

Lesson 3.2

Introduction to Graphing Lines in Statistics

E What happens when the number is positive?

F What happens when the number is negative?

In the equation of the line of best fit there are two very important numbers. The number sitting directly in front of x represents the **slope** of the line. It tells us about the steepness and direction of the line. The number standing alone beside the term containing x represents the **y-intercept**. It is literally the place where the graph crosses the vertical or y -axis. This point sometimes has meaning in the context of our data but not always. You will learn how these numbers give us additional information about the relationship between our explanatory and response variables later on in Chapter 3. For now, it is just useful for you to get an idea of how changing these numbers will impact the picture you see when you look at the graph of the line of best fit. Note that since you can add in any order, $y = (\text{slope})x + yint$ is the same as, $y = yint + (\text{slope})x$.

Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

INTRODUCTION

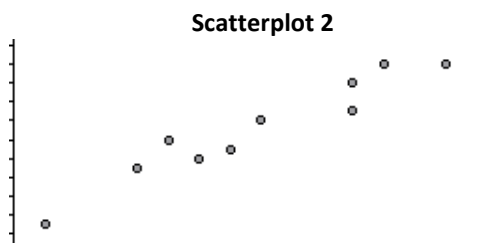
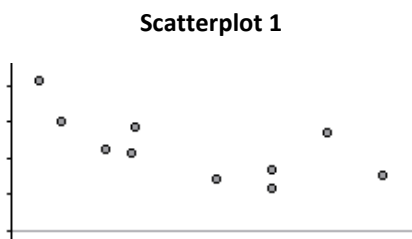
Of the two variables plotted, the **eXplanatory variable** is the one you think *explains* the other variable, which is called the **response variable**. For example, more powerboat registrations might *explain* more manatee deaths. Therefore, powerboat registrations is the explanatory variable and manatee deaths is the response variable. We plot the eXplanatory variable on the X-axis and the response variable on the y-axis.

In the previous lesson we explored the **direction** and **strength** of relationships between quantitative variables. In this lesson, we will compare and contrast several scatterplots and identify the **form** of the relationship between the two variables.

The form of a relationship between two variables provides us information on what happens to the **response variable** as the **explanatory variable** changes.

TRY THESE

- 1 Descriptions “A” and “B”, below describe a set of measurements in a scatterplot. The explanatory variable (x) is represented by the horizontal axis and the response variable (y) is represented by the vertical axis. Match Description A and Description B to a scatterplot, and briefly explain your reasoning.



- A x = city miles per gallons and y = highway miles per gallon for 10 cars

What does a dot represent?

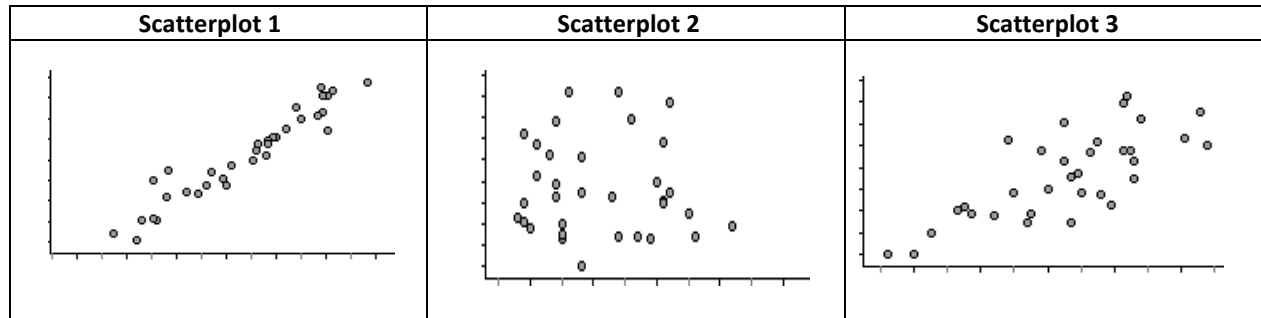
- B x = sodium (milligrams/serving) and y = *Consumer Reports* quality rating for 10 salted peanut butters

What does a dot represent?

Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

2 Match each description (A, B, and C) to a scatterplot. Briefly explain your reasoning.



A The scatterplot shows the height and weight for 34 physically active men: x = height (in), y = weight (pounds).

What does a dot represent?

B The scatterplot shows the distance around the waist (waist circumference) and belt sizes for 34 physically active adult men: x = waist circumference (in), y = belt size (in).

What does a dot represent?

C The scatterplot shows the age and foot size for 34 physically active adult men: x = age (years), y = foot size (in)

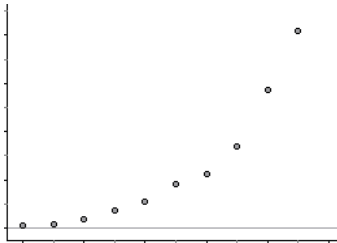
What does a dot represent?

Lesson 3.3

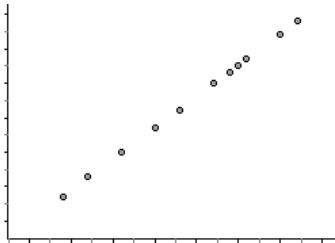
Form, Direction, and Strength of the Relationship between Two Measurements

3 Match each description, A to F, to a scatterplot. Briefly explain your reasoning.

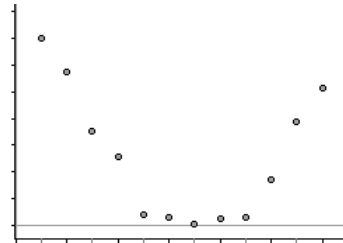
Scatterplot 1



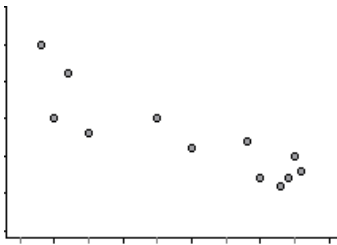
Scatterplot 2



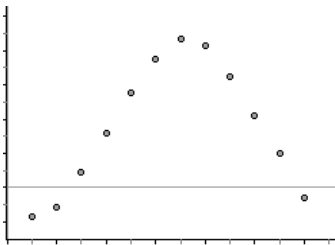
Scatterplot 3



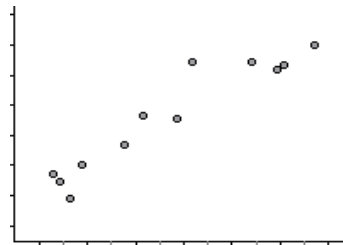
Scatterplot 4



Scatterplot 5



Scatterplot 6



A x = month number (January = 1) and y = rainfall (inches) in Napa, California. Napa has several months of drought each summer.

What does each dot represent?

B x = month number (January = 1) and y = average temperature in Boston, Massachusetts. Boston has cold winters and hot summers.

What does each dot represent?

Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

- C x = year (from 1990) and y = total amount of money in Ali's savings account (\$) at the end of each year. Ali has not taken any money out of the account and it is earning compound interest (the more money in the account, the more interest it earns), not simple interest (same amount of interest earned every year).

What does each dot represent?

- D x = amount of Trail Mix purchased (pounds) and y = cost of Trail Mix (\$).

What does each dot represent?

- E x = test average in a class and y = final exam grade for a class of statistics students

What does each dot represent?

- F x = engine size (in liters) and y = city miles per gallon for a sample of cars. Larger engines tend to use more gas.

What does each dot represent?

Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

SUMMARY

Linear versus Non-linear Form

Some relationships between the variables in scatterplots can be summarized well by a line. Other relationships cannot. We call these nonlinear. If a relationship is nonlinear and shows a curved pattern we call it *curvilinear*. These relationships are better summarized with a curve.

Strong versus Weak Relationships

Some relationships between variables are strong and others are weak. If the relationship is strong, the line or the curve does a good job summarizing the relationship between the measurements. This means the points in the scatterplot would be close to the line or curve that summarizes the data.

Positive versus Negative Association

When an association is positive, larger values of x tend to correspond to larger values of y . When an association is negative, larger values of x tend to correspond to smaller values of y .

In this lesson we had to make judgments about the strength of a relationship by looking at a scatterplot. This can be difficult. In the next lesson we will explore a measure of the strength and direction of a linear relationship between two quantitative variables. This measure is called the *correlation coefficient*.

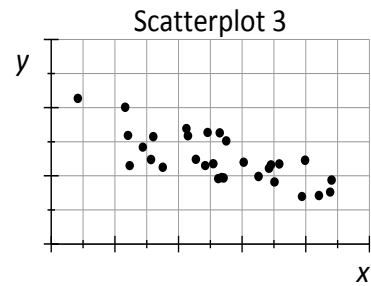
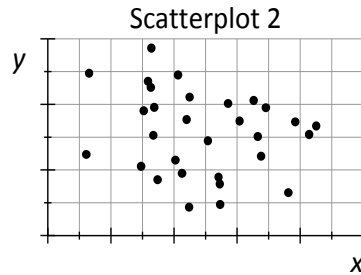
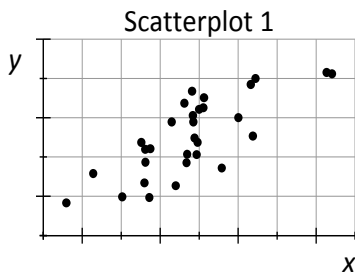
Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

STUDENT NAME _____ DATE _____

TAKE IT HOME

- 1 Match each description (A, B, and C), of a set of measurements, to a scatterplot. Briefly describe your reasoning. Then describe what a dot represents in each graph.



- A x = average outdoor temperature and y = heating costs for a residence for 30 winter days

What does a dot represent?

- B x = height (inches) and y = shoe size for 30 adults

What does a dot represent?

- C x = height (inches) and y = score on an intelligence test for 30 teenagers

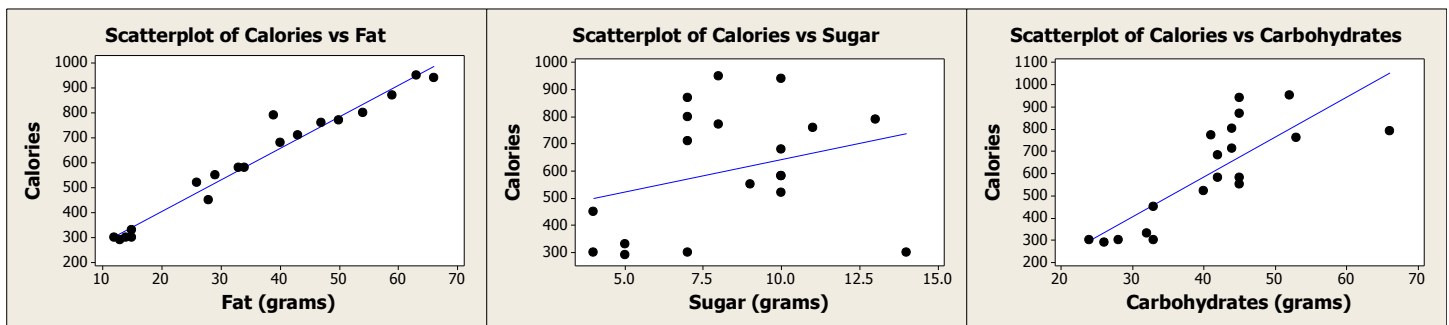
What does a dot represent?

Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

2 Researchers gathered data about the amount of fat, sugar, and carbohydrates in 22 fast food hamburgers. They gathered the information from fast food companies' websites. To keep their measurements consistent, all data was described in grams. Using the companies' websites, the researchers also identified the number of calories for each hamburger. The researchers wanted to know if the amount of calories in the hamburgers depended on how much of each ingredient (fat, sugar and carbohydrates) was in the burger. That is, the researchers wanted to know whether there was a relationship between the amount of fat, sugar, and carbohydrates and the amount of calories in a hamburger. A line has been added to each graph to help you see the patterns more clearly.

A For the first scatterplot, what is the explanatory variable? What is the response variable?



B About how many calories would you predict for a burger that has 20 grams of fat?

C About how many calories would you predict for a hamburger that has 40 grams of carbohydrates?

D Which prediction is likely to be more accurate? Why do you think this?

E Which ingredient has the weakest impact on calories and is therefore a bad predictor for calories? Why do you think this?

Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

- F What does the idea of strength tell you about whether an ingredient is a good predictor of calories?
- G What is the direction of the fat/calories graph? What does the direction of the line tell you about the association between the amount of fat and the calories in fast food hamburgers?

3 Suppose you gathered the following information from students at a local high school:

- GPA (grade point average),
- Average weekly hours spent working at a job,
- Average weekly hours spent doing homework,
- Average hours of sleep a night,
- Hourly wage,
- Height,
- Weight,
- Length of the left foot,
- Age of the oldest child in the student's immediate family,
- Number of children in the student's immediate family,
- Gender,
- Race, and
- Age.

A From this list of variables, choose:

- i Two variables that you think will show a positive linear association,
- ii Two variables you think will show a negative linear association, and
- iii Two variables you think will not show an association in a scatterplot.

You may use the same variable for more than one comparison.

Explain why you chose the two-variable pairs. What was your reasoning for each pair?

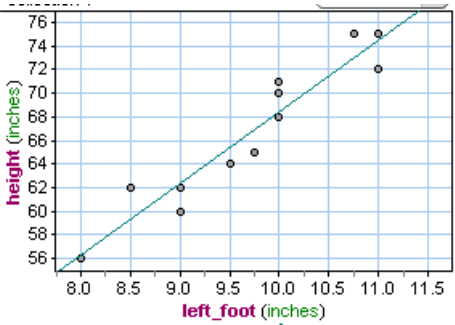
B Sketch a scatterplot on the graph grids provided to illustrate each of the three relationships that you described in part A. Draw three scatterplots showing imaginary data for 12 students to illustrate the association. An example is shown below.

- If there is an association, sketch a line to highlight the association.
- For each scatterplot, label the axes of each graph with the name of the variable.
- Scale the graph, that is, place number measurements on the x and y axes. Make sure that the numbers are realistic for the variable, (the numbers are what you would see if the problem were real and not imaginary).

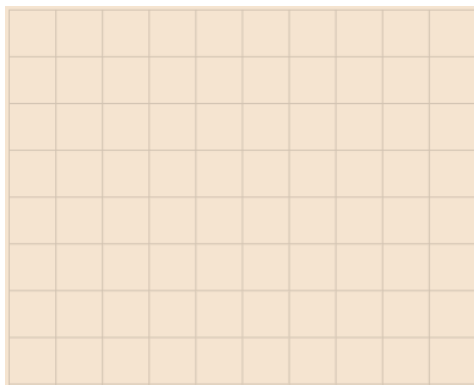
Remember to place the eXplanatory variable on the horizontal axis and place the response variable on the vertical axis.

Lesson 3.3

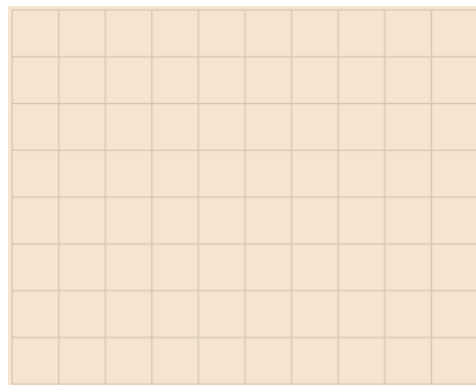
Form, Direction, and Strength of the Relationship between Two Measurements

Sample Solution	What makes this a good answer?
<p>I think there will be a positive linear association between length of the left foot and height. Students with shorter feet tend to be shorter in height and students with longer feet tend to be taller. This might happen because boys and girls are mixed together in the data, and boys tend to be bigger. Also freshmen and seniors could be mixed together, and older students may be bigger.</p>	<p>The answer tells us the two variables and the type of association expected. The explanation says what positive association means for these two variables.</p> <p>An extra special aspect of this answer is the inclusion of the gender and age issues to help the reader understand why the association might be positive for high school students. This is above and beyond what you might expect in an answer, but it shows good thinking about who is described by the data.</p>
	<p>The data has an upward trend, and you see a line that highlights the positive linear association.</p> <p>The measurements are reasonable for the variables. Both axes are labeled with the name of the variable, and units are given. Scales are clearly marked and consistent across the axis.</p> <p>(Students will have hand-drawn sketches, which is fine.)</p>

1.



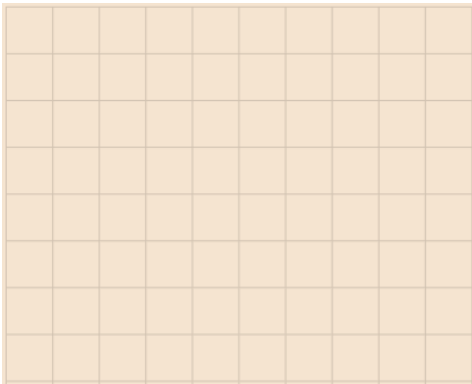
2.



Lesson 3.3

Form, Direction, and Strength of the Relationship between Two Measurements

3.



Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

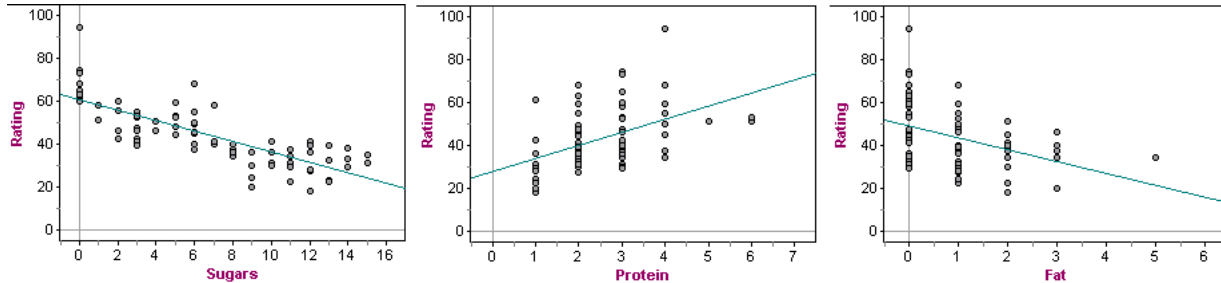
INTRODUCTION

In the last lesson, we studied *bivariate data*. Bivariate data are often displayed in scatterplots. Our goal is to use scatterplots to understand relationships between variables and to make predictions. To accomplish our goal, we must use statistical tools. Statistical tools help us summarize key features in the relationship between two variables.

The scatterplots below show data about breakfast cereals. The scatterplots show relationships between an ingredient of the cereal and the rating. These are the two variables.

A line appears in each scatterplot. The line summarizes the relationship between the ingredient and the rating. The line makes it easier to see the direction of the relationship between the two variables. Recall that the direction of a relationship can be **positive** or **negative**. Two variables have a positive association if *increasing* values of one variable are associated with *increasing* values of the other. Two variables have a negative association if *increasing* values of one variable are associated with *decreasing* values of the other.

For instance, such a line can help us see that sugar and fat are *negatively* associated with ratings. Another line can help us see that protein is positively associated with ratings.



A line can also help you evaluate the *strength* of a relationship. In the scatterplots above, the lines help us evaluate the strength of each relationship between ingredient amounts and ratings. Some ingredients are more strongly related to the rating, like sugar. In the scatterplot with the variable sugar, the dots are fairly close to the line. We say this scatterplot has a linear pattern with not too much scatter around the line. Sugar is more strongly related to the rating. A small amount of scatter around the line means that the ingredient is a good predictor of the rating.

For other ingredients, like fat, the association with rating was not as strong. We see this as *more* scatter about the line. This means that fat is a weaker predictor of ratings.

A relationship is stronger if the points with similar x-values also have similar y-values. In the scatterplots above, cereals with similar amounts of sugar also have similar ratings. Cereals with similar amounts of protein have more variability in their ratings than cereals with similar amounts of sugar. Since it can be hard to judge strength just by looking at a graph, we need a way to measure this variability.

Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

In this lesson, we investigate a measurement of variability in scatterplots. This measure is called the **correlation coefficient**. The correlation coefficient is represented with the letter r . Later, we will learn how to calculate the r -value using our calculators. But for this exercise the correlation coefficients or r -values have been provided for you for each of the scatterplots. The goal in this lesson is to examine the properties of r .

TRY THESE

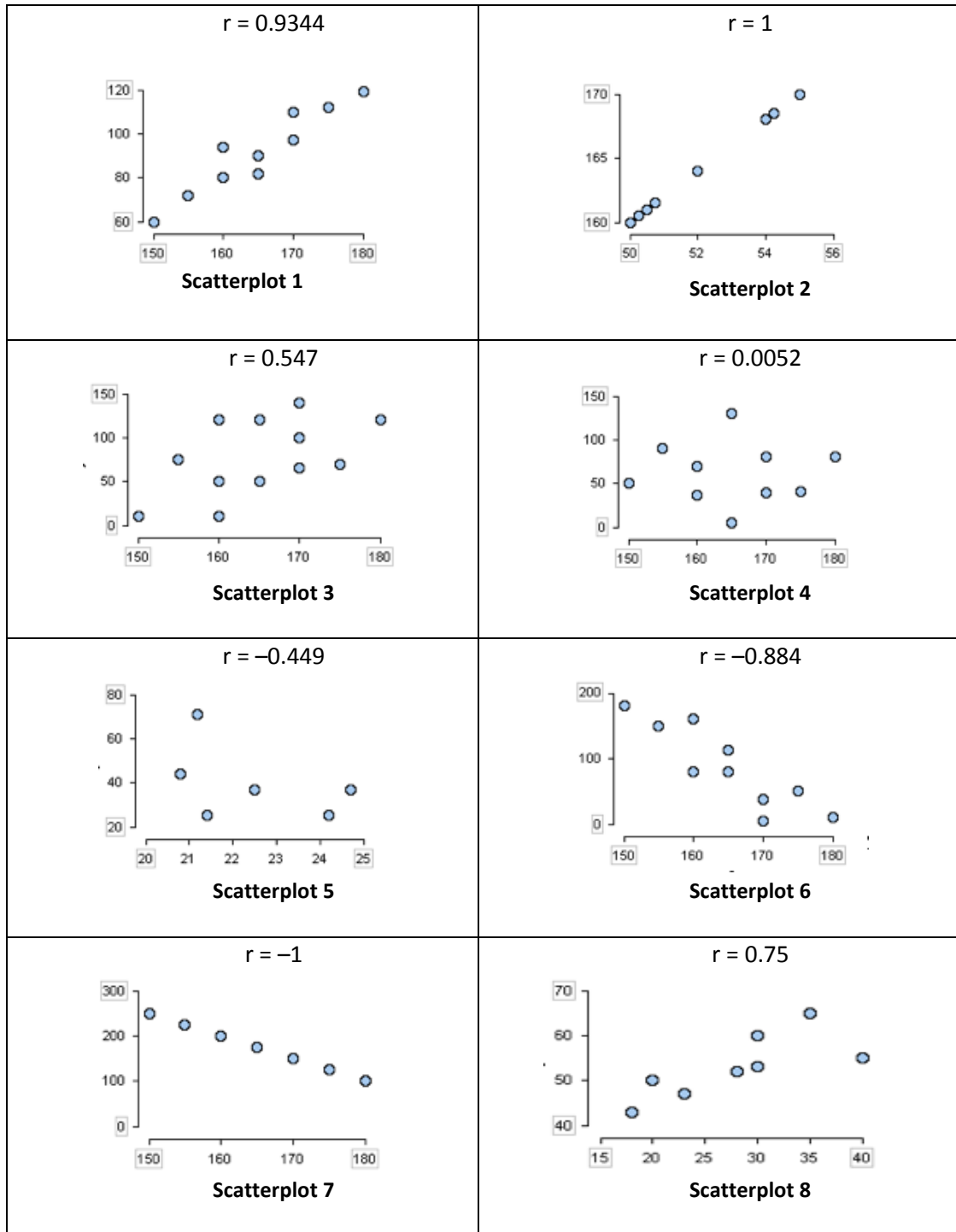
Investigating the Properties of r

For the questions below, please use the graphs that follow on the next page.

- 1 Which of the eight graphs on the following page show:
 - A A positive association between x and y ?
 - B A negative association between x and y ?
 - C No association between x and y ?

Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties



Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

SUMMARY

The following are properties of the correlation coefficient r :

- r measures the direction and strength of a linear association.
- r is positive if there is a positive linear association and negative if there is a negative linear association.
- $-1 \leq r \leq 1$.
- When the association is perfectly linear $r = 1$ or $r = -1$. The closer r is to 1 or -1 , the closer the data are to having a perfect linear association (assuming, of course, that you see linear association in the scatterplot).

TRY THESE

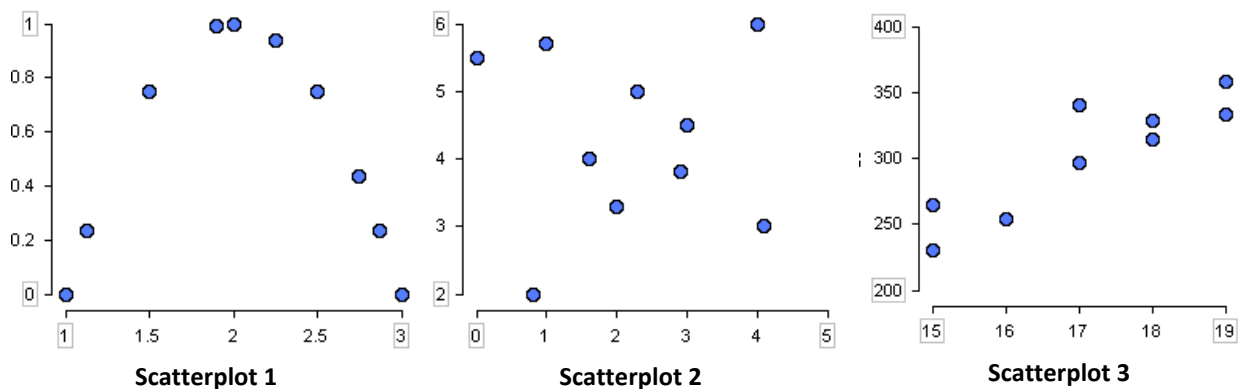
Linear Correlation with Non-linear Scatterplots

4 Based on what you have learned about r , answer the following two questions:

- A Can there be a strong relationship between the variables when r is close to 0?
- B Can variables have a nonlinear relationship when r is close to 1?

5 Examine the following scatterplots.

- A The data in two of these scatterplots have an r -value close to 0. Look over the scatterplots and decide which two have an r -value close to 0. Do not calculate the r -value.



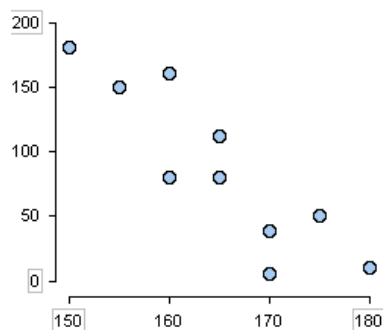
Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

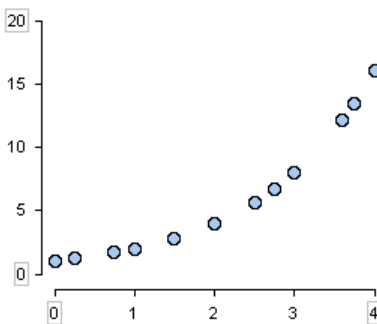
- B Do both scatterplots with an r -value close to zero have *weak associations*? Does this change your answer to Question 4, Part A? Why?

6 Examine the following scatterplots.

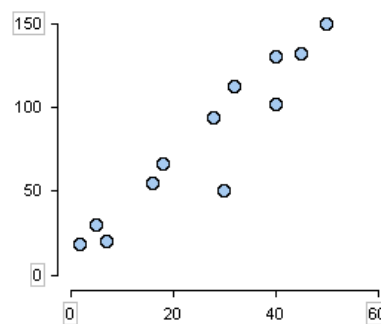
- A The data in two of these scatterplots have an r -value close to 0.94. Determine which two have an r -value close to 0.94. Do not calculate the r -value.



Scatterplot 1



Scatterplot 2



Scatterplot 3

- B Do both scatterplots with an r -value close to 0.94 show *linear associations*? Does this change your answer to question 4, part B? Why?

TRY THESE

Do not make the mistake of observing a high correlation between two variables and jumping to the conclusion that one thing must be **causing** the other. Just because two variables have a high correlation, does not mean that one **causes** the other. Correlation never proves **causation**.

It is easy and fun to construct silly examples of correlations that do not result from causal connections. Here are some examples from John Allen Paulos, a professor at Temple University who is well known for his books on mathematical literacy.

Language Tip

A lurking variable is a variable that is not measured in the study but that impacts the variables in the study. A lurking variable creates an association between the two variables.

Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

7 Read the two scenarios below from *A Mathematician Reads the Newspaper*¹ by Paulo:

Identify these three variables from the scenarios and then fill in the table below.

- a. Explanatory variable
- b. Response variable
- c. Confounding (lurking) variable

Scenario 1: A more elementary [basic] widespread confusion is that between correlation and causation. Studies have shown repeatedly, for example, that children with longer arms reason better than those with shorter arms, but there is no causal connection here. Children with longer arms reason better because they're older!

Scenario 2: Consider a headline that invites us to infer a causal connection: BOTTLED WATER LINKED TO HEALTHIER BABIES. Without further evidence, this invitation should be refused, since affluent parents are more likely both to drink bottled water and to have healthy children; they [wealthy parents] have the stability and wherewithal [money] to offer good food, clothing, shelter, and amenities.

Making a practice of questioning correlations when reading about "links" between this practice and that condition is good statistical hygiene.

	EXplanatory variable	Response variable	Confounding (lurking) variable
Scenario 1			
Scenario 2			

¹Paulos, J.A. (1995). *A mathematician reads the newspaper* (p. 137). New York: Basic Books.

Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

- 8 Work with your group to describe a scenario with two quantitative variables. The variables should be strongly correlated, due to a third confounding (or lurking) variable.

YOU NEED TO KNOW

The correlation coefficient measures association. But correlation is not the same as causation. A strong correlation between two variables is evidence that there is a statistical relationship between the variables, in other words, the variables vary together in a predictable way, but this does not mean one *causes* the other.

Lesson 3.4

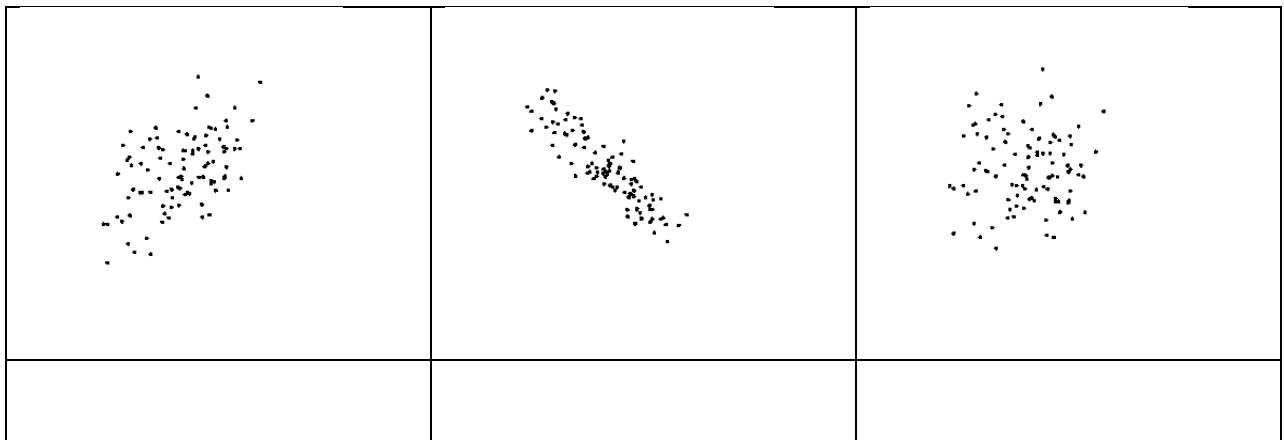
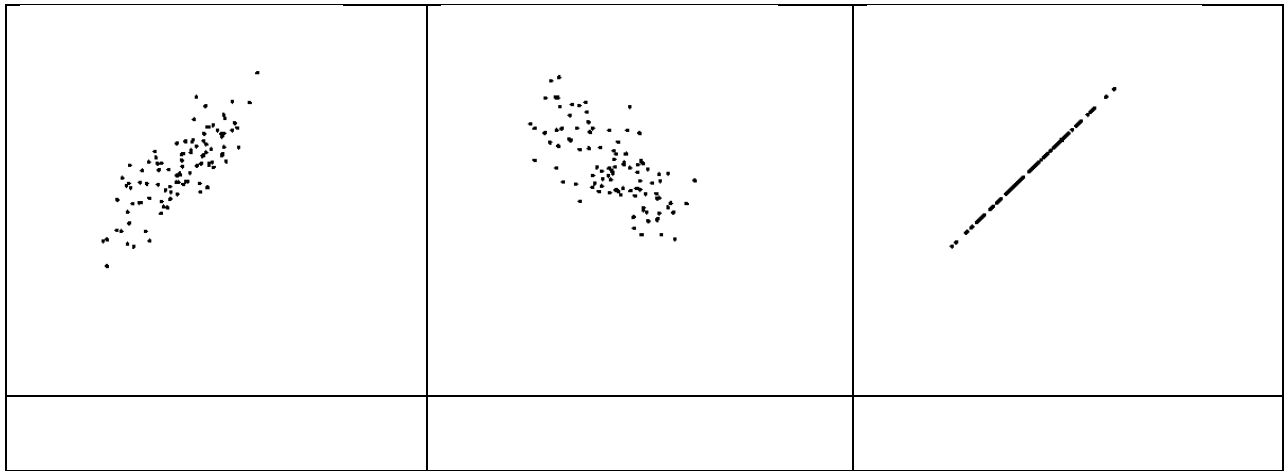
Introduction to the Correlation Coefficient and Its Properties

STUDENT NAME _____ DATE _____

TAKE IT HOME

1 Match the scatterplot to the correct correlation coefficient.

- A. $r = -0.9$ B. $r = -0.7$ C. $r = 0$ D. $r = 0.5$ E. $r = 0.8$ F. $r = 1$



Lesson 3.4

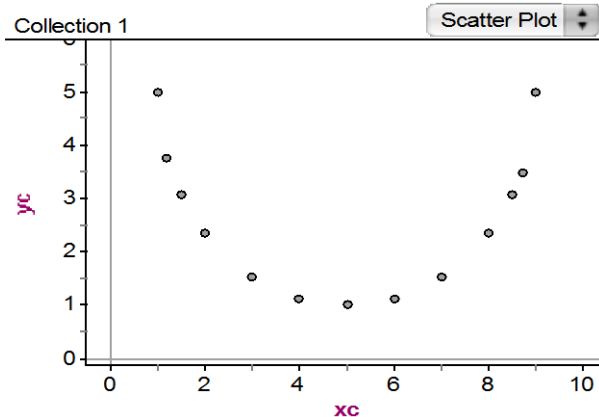
Introduction to the Correlation Coefficient and Its Properties

- 2 Go to <http://istics.net/stat/Correlations>.
 - A Match the values of the correlation coefficient with the corresponding scatterplot. Use what you know about strength and direction of linear associations to complete this task.
 - B Click *Answers* to check your work.
 - C Click *New Plots* for a new set of scatterplots. Just below the plots, the applet keeps a running count of how many correct matches you have made. Continue matching scatterplot and correlation coefficients until you have accumulated (gathered) at least 25 correct matches.

- 3 Go to <http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html>.
 - A Click on the *New Sample* button, which generates a scatterplot.
 - B Type in your guess for the correlation in the box called *Correlation Guess* and hit Enter. The applet then reveals the actual value of the correlation coefficient.

It is not easy to guess the value of the correlation coefficient exactly. If a guess is within 0.1 of the actual value, it is a pretty good guess. (For example, if you guess 0.7 and the actual value is anything between 0.6 and 0.8, you have a pretty good guess.)

- C Click *New Sample* and estimate the correlation as many times as it takes for you to be comfortable with your ability to estimate the value of the correlation coefficient within 0.1.
- 4 Do you see a strong relationship (or pattern) between the variables shown in the scatterplot below?



Lesson 3.4

Introduction to the Correlation Coefficient and Its Properties

- 5 The r – value for the scatterplot shown above is $r = 0$. But, as we learned in the lesson today, scatterplots with r -values close to 1 or -1 have a strong relationship. How do you explain this contradiction with your answer to number 4?
- 6 Below are some incorrect statements. Explain the mistakes.
- A My very low correlation of $r = -0.8$ indicates that there is almost no association between the size of a car engine and the gas mileage it gets.
- B There was a very strong correlation of $r = 1.35$ between what students' test average for the semester and their final exam grade.
- 7 A researcher claims there is a strong correlation between height and reading scores for children in elementary school.
- A Does this mean that taller children are better readers?
- B What is the lurking variable that might explain this strong correlation?
- 8 A survey of the world's nations shows a strong positive correlation between the number of fast food restaurants in the nation and life expectancy in years at birth. So the more fast food restaurants the country has, the greater the life expectancy of its citizens.
- A Does this mean that fast food restaurants are good for your health?
- B What is the lurking variable that might explain this strong correlation?

Lesson 3.5

Using Lines to Make Predictions

INTRODUCTION

Statistical methods are used in *forensics* to identify human remains based on the measurements of bones. In the 1950's, Dr. Mildred Trotter and Dr. Goldine Gleser measured skeletons of people who died in the early 1900's. From these measurements they developed statistical formulas for predicting a person's height based on the lengths of various bones.

Language Tip

***Forensics* is the use of science to investigate crimes.**

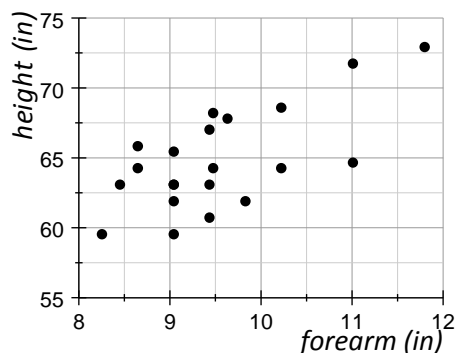
These formulas were first used to identify the remains of U.S. soldiers who died in World War II and were buried in unmarked graves in the Pacific zone. Modern forensic scientists have made adjustments to these formulas to account for the differences in people living now. We will use a process similar to Trotter and Gleser's in this problem.¹

To illustrate the type of data analysis done in forensics, let's see if we can identify a female student based on the length of her forearm. The mystery student has a forearm measurement of 10 inches. She is alive and healthy!

Our task is to determine if the mystery student could be one of three female students. To accomplish our task, we have gathered evidence in the table below. Our evidence lists the heights and ages for three female college students.

	Jane Doe 1	Jane Doe 2	Jane Doe 3
Age	18	23	33
Gender	Female	Female	Female
Height	5 feet, 5 inches	5 feet, 2 inches	6 feet

First, we need data to help us see if there is a relationship between forearm length and height for female students. The scatterplot below is a graph of forearm length versus height for 21 female college students taking Introductory Statistics at Los Medanos College in Pittsburg, California, in 2009.



¹**Note:** For information on the Terry skeleton collection, see <http://anthropology.si.edu/cm/terry.htm>. For a more recent example of how forensic scientists are still building on the work of Trotter and Gleser, see the following: Jantz, R. L. (1993); *Modification of the Trotter and Gleser female stature estimation formulae*, *Journal of Forensic Science*. 38(4), 758–63.

Lesson 3.5

Using Lines to Make Predictions

- 1 Based on the scatterplot, what is a reasonable prediction for the height of the mystery student? Briefly describe or show how you made your prediction.

- 2 There is a lot of variability (or scatter) in the data. This variability can make it difficult to determine if one of these students is the mystery student. One way to help solve the mystery is see if we can eliminate any of the three students. Could any of the three students be eliminated as the mystery student? Explain your reasoning.

Lesson 3.5

Using Lines to Make Predictions

NEXT STEPS

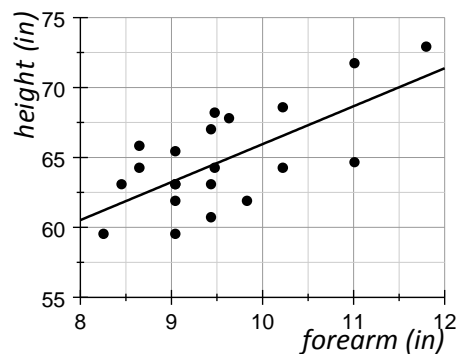
Using a Line to Make Predictions

- 3 The scatterplot has a positive linear association. The correlation is 0.68, which is fairly strong. Since the correlation is fairly strong, it makes sense to use a linear model to summarize the relationship between the forearm and height measurements. In statistics, there is a line that will give us the best description of how height and forearm length are related. We call it the *line of best fit*. We will learn more about how to find this line in future lessons.

Language Tip

The *line of best fit* is the line that represents the center of a linear pattern in a scatterplot.

For now, we will give you the equation of this line.



- A The line of best fit is drawn on the scatterplot above. Use the line of best fit to predict the height of the mystery student.
- B The equation of the line of best fit is:

$$\text{Predicted height} = 2.7 \cdot (\text{forearm length}) + 39.$$

If we use x and y to represent these variables, the equation looks more like something from an algebra class.

$$\hat{y} = 2.7x + 39$$

When we use letters to represent variables in the line of best fit, we put a *hat* on the y and write \hat{y} instead of y . The *hat* is a symbol that tells us that we are making a *prediction*. Results from the equation above do not represent actual observed data values, but rather are our predictions.

Use the equation provided above to predict the height of the mystery person. Recall that the mystery person had a forearm length of 10 inches.

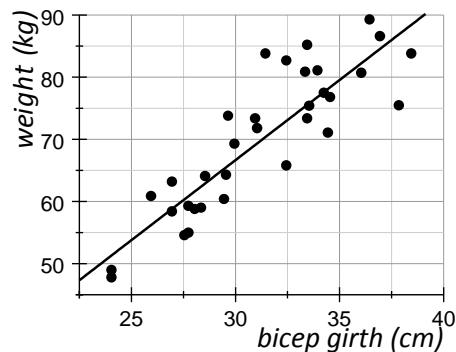
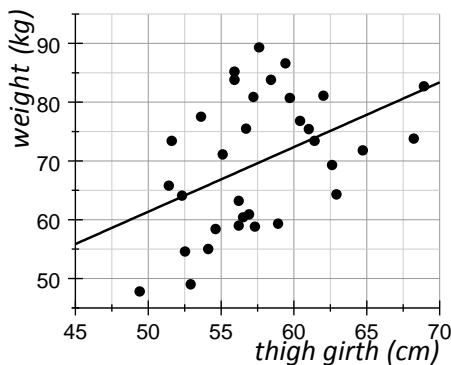
Lesson 3.5

Using Lines to Make Predictions

- C Is the height of Jane Doe 1, Jane Doe 2, or Jane Doe 3 closest to the predicted height of the mystery student given by the line of best fit?

While this does not guarantee that you have correctly identified the mystery student, it suggests that one student's height, together with the 10-inch forearm measurement, fits the linear pattern in the data better than the other students.

- 4 The scatterplots below represent body measurements in centimeters for 34 adults who are physically active.²



- A Based on these data, which do you think is a better predictor of an adult's weight: thigh girth or bicep girth? Why?

- B Adriana has a thigh girth of 57 centimeters and a bicep girth of 25 centimeters. Predict her weight using the graph that you think will give the most accurate prediction. Then plot Adriana's data point on the scatterplot that you used to make her weight prediction.

Language Tip

Girth is the measurement *around* a body part.



²Retrieved from www.amstat.org/publications/jse/v11n2/datasets.heinz.html

Lesson 3.5

Using Lines to Make Predictions

- C The equations of the two lines shown are

$$\text{Predicted weight} = 6.3 + 1.1 \cdot (\text{thigh girth}), \quad \text{Predicted weight} = -10.5 + 2.6 \cdot (\text{bicep girth}).$$

Predict Adriana's weight using the equation that you think predicts weight best.

- D Of course, we do not really know Adriana's weight. How accurate do you think the prediction of Adriana's weight is? Choose the option that is the most reasonable and explain your thinking.

- Very accurate (within a range of plus or minus 1 kilogram).
- Somewhat accurate (within a range of plus or minus 5 kilograms).
- Not very accurate (within a range of plus or minus 10 kilograms).

- 5 In previous lessons, we learned about the concept of correlation. We learned that the correlation coefficient, r , describes the strength and direction of the linear relationship between two quantitative variables. Now we are predicting the value of one variable based on the other.

- A Think about the scatterplots in Question 4. Which has a correlation coefficient closer to 1? Explain your answer.

- B It is important for our predictions to be reliable. How does the correlation coefficient (having an r closer to 1 or -1) relate to the accuracy of the prediction?

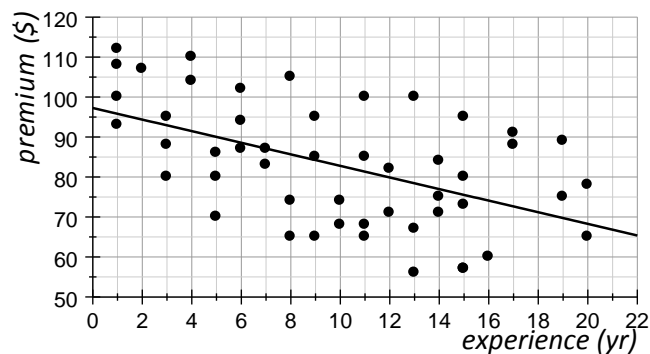
Lesson 3.5

Using Lines to Make Predictions

TRY THESE

Many people pay for car insurance. Sometimes we make monthly payments for the car insurance, which are called premiums. In 2008, a statistics student gathered data on the monthly car insurance premiums that students and faculty at Los Medanos College pay. She explored whether there was a relationship between monthly car insurance premiums to years of driving experience. She found a linear relationship and used statistical methods to get the following model:

$$\text{predicted premium} = 97 - 1.45 \cdot (\text{years of experience})$$



- 6 Predict the monthly car insurance premium paid by someone who has been driving 12 years.
- 7 Give two ways that you can use to make the prediction?
- 8 Use the line of best fit to predict the monthly insurance premium of a driver with 65 years of driving experience.
- 9 Do you think that this is a reliable prediction for such a driver's insurance premium? Explain your answer.
- 10 Why do you think this answer is unreasonable?

Lesson 3.5

Using Lines to Make Predictions

In statistics, *extrapolation* is the process of using a statistical model, like a line, to make predictions that are outside the range of the available data. The problem with extrapolation is that it yields predictions that have no basis in evidence. Because of this, extrapolation is unreliable for predicting values of a response variable.

Language Tip

Extrapolation means using a statistical model to make predictions that are beyond the range of observed values.

- 11 Use the line of best fit to predict the premium of a new driver with 0 years of driving experience.

- 12 Do you think that this is a reliable prediction for a new driver's insurance premium?

- 13 Both predictions above are examples of extrapolation, but one is much more likely to be inaccurate. Which extrapolation is most likely to be inaccurate, and why?

SUMMARY

- When a line is a good summary of a statistical relationship, the line can be used to predict the value of a response variable for a given value of the explanatory variable (the predictor).
- Predictions are more accurate when the relationship is strong. Since the correlation coefficient, r , is a measure of strength (and direction) of the relationship between two variables, the closer the value of r is to 1 or -1 the more accurate your predictions will be.
- Making predictions based on extrapolating outside the range of the data can be risky and should be done with caution.

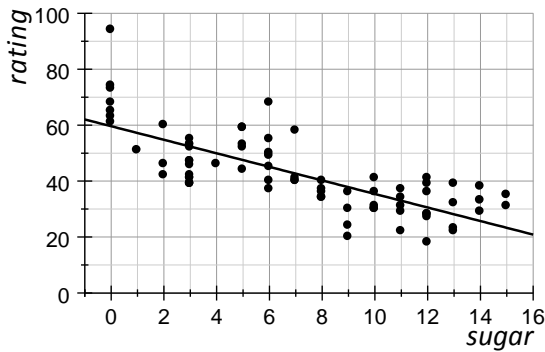
Lesson 3.5

Using Lines to Make Predictions

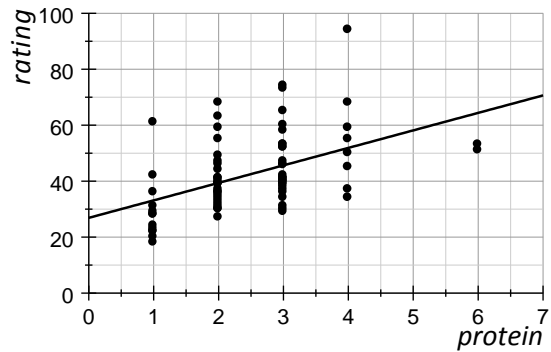
STUDENT NAME _____ DATE _____

TAKE IT HOME

- 1 For this problem, we are returning to the data set for the 77 breakfast cereals we investigated at the beginning of Chapter 3.



$$\text{Predicted rating} = 60 - 2.43 \cdot (\text{sugars})$$



$$\text{Predicted rating} = 28 + 5.96 \cdot (\text{protein})$$

Two new cereals are being rated by *Consumer Reports*. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.

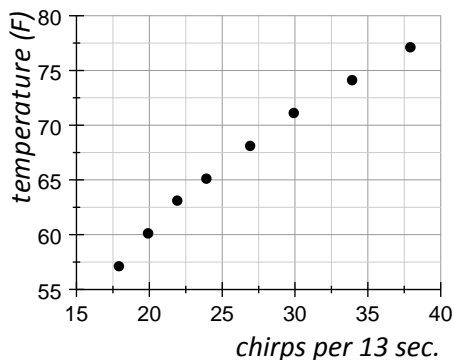
- A Using the lines of best fit, predict the *Consumer Reports* rating for the two cereals.
- B Approximate the value of the correlation coefficient, r , for each of the scatterplots above.
- C For which cereal do you think your prediction is probably more accurate? That is, for which cereal do you think your prediction is likely be closer to the actual *Consumer Reports* rating)? Why?

Lesson 3.5

Using Lines to Make Predictions

- 2 Can we predict the temperature based on how fast crickets chirp?

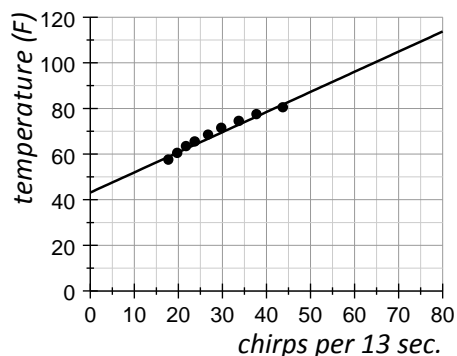
Crickets chirp by rubbing their wings together. According to scientist Tom Walker, crickets are good thermometers because their chirp rate is related to temperature. The snowy tree cricket chirps at a rate that is slow enough to count.



A The scatterplot above is a graph of data from the June 1995 issue of *Outside* magazine. Use the scatterplot to predict the temperature outside when the snowy tree crickets are chirping at a rate of 40 chirps every 13 seconds.

- B How accurate do you think your prediction is? There are 3 options below. Choose the option that is most reasonable and briefly explain your thinking.
- Very accurate (within a range of plus or minus 1 degree).
 - Somewhat accurate (within a range of plus or minus 5 degrees).
 - Not very accurate (within a range of plus or minus 10 degrees).

C This is the same data graphed over a wider field of view, like zooming out on a photograph. The window has been enlarged by expanding both axes.



Lesson 3.5

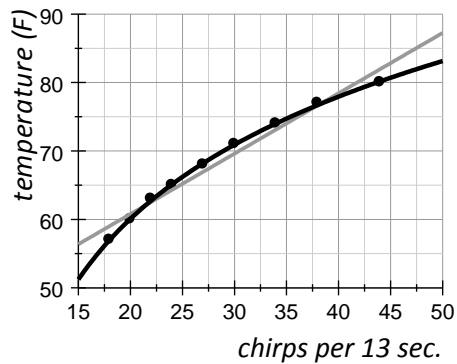
Using Lines to Make Predictions

The line pictured is the best-fit line:

$$\text{Predicted temperature} = 0.88 \cdot (\text{chirp rate}) + 43$$

For some chirp rates, this line gives very accurate predictions of the temperature. However, the pattern of this data is actually slightly curved. Illustrate the concept of extrapolation by picking a point on the line that gives unreliable estimates. Explain how this point illustrates the concept of extrapolation (gives meaningless or unreasonable results).

- D For chirp rates above 45, a nonlinear model might give more accurate predictions. One possible nonlinear model is shown with the linear model in the graph below.



Predict the temperature for a chirp rate of 50 per 13 seconds using both the linear model and the curved model. What is the difference in temperatures predicted by the two models?

Lesson 3.5

Using Lines to Make Predictions

- 3 We learned that a variable that is used to predict the value of another variable is called an **explanatory variable**. The other variable, whose values we are predicting, is called the **response variable**.
- A The introductory problem in this lesson has forearm lengths and heights for 21 female college students. In this situation, which is the explanatory variable?
- B The cereal data has the amount of sugar in a serving and the *Consumer Reports* rating. In this situation, which is the explanatory variable?
- C When graphing bivariate data do we put the explanatory variable on the horizontal axis or vertical axis?
- D A group of students use technology to find a best-fit line. They used measurements of temperature (°F) and the chirp rate of the snowy tree cricket (measured in number of chirps in 13 seconds). However, some students use *temperature* as the explanatory variable, and others use *chirp rate* as the explanatory variable. Which of the two equations below treats temperature as the explanatory variable?

$$\text{Predicted temperature} = 0.88 \cdot (\text{chirp rate}) + 43$$

$$\text{Predicted chirp rate} = 1.1 \cdot (\text{temperature}) - 47$$

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

TRY THESE

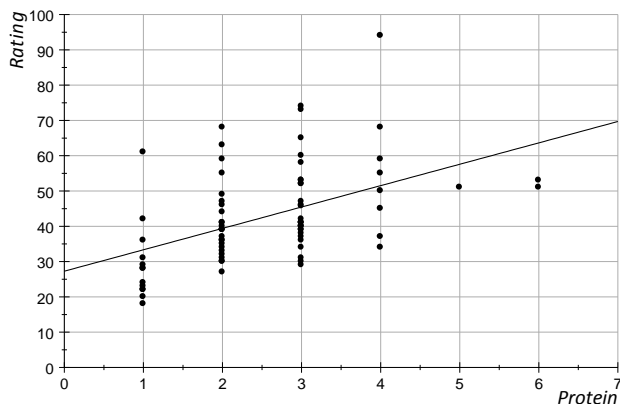
Investigating the Line of Best Fit Equation

- 1 In the previous lesson we learned about using a line that summarizes the data to predict the response variable, or y -value. In our lessons, we have talked about cereal ingredients and their *Consumer Reports* ratings. For this lesson, we will return to that data. Below is the equation of best fit. For now, we are giving you the equation of this line, but in the next lesson we will learn how to find this equation using our calculator.

$$\text{Predicted rating} = 28 + 6 \cdot \text{protein}$$

We can use this line to predict a cereal's rating based on the amount of protein (in grams) in a serving.

- A Use the equation of the line given above – not the graph below - to predict the rating for a cereal containing 2 grams of protein in a serving.
- B The scatterplot of the amount of protein in the cereals, and their ratings, is given below. The graph also shows the line of best fit.



There are two cereals with 6 grams of protein in a serving. Circle them on the graph. Is the predicted rating, from the line of best fit, too high or too low for these cereals? Use the graph to help you answer the question.

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

- C The correlation coefficient is $r = 0.48$. Which of these statements would you use to describe how well the protein, in the cereal, predicts *Consumer Reports Magazine's* rating? Use the graph and think about what the correlation coefficient tells you about the strength of the linear relationship.
- Predicted *Consumer Reports* ratings based on protein amounts are very accurate (residuals within a few rating points would be typical).
 - Predicted *Consumer Reports* ratings based on protein amounts are not very accurate (residuals as large as 10 rating points would not be surprising).

- 2 Now let's focus on understanding the numbers in the equation of the line.

- A Below, again, is the equation of the best fit line. On the right is a table of amounts of protein and the predicted cereal ratings. Use the line of best fit equation,

$$\text{Predicted rating} = 28 + 6 \cdot \text{protein}$$

to fill in the rest of the table.

$x = \text{protein}$ (g/serving)	$\hat{y} = \text{predicted}$ rating
0	
1	34
2	40
3	
4	52
5	

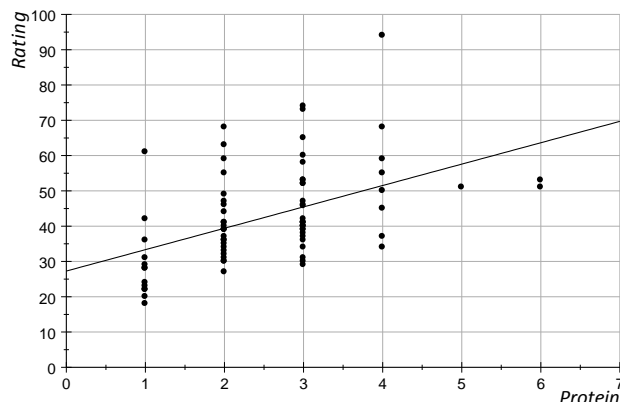
- B Each line in the table represents a point on the line given by the equation:
 $\text{Predicted rating} = 28 + 6 \cdot \text{protein}$. Each point can be written as an ordered pair (x,y) . Use the table to fill in the blanks below.

$(0, \underline{\hspace{1cm}})$

$(2, \underline{\hspace{1cm}})$

$(\underline{\hspace{1cm}}, 58)$

Now, plot these three points on the scatterplot below. Should your points be on the line, or just close to the line?



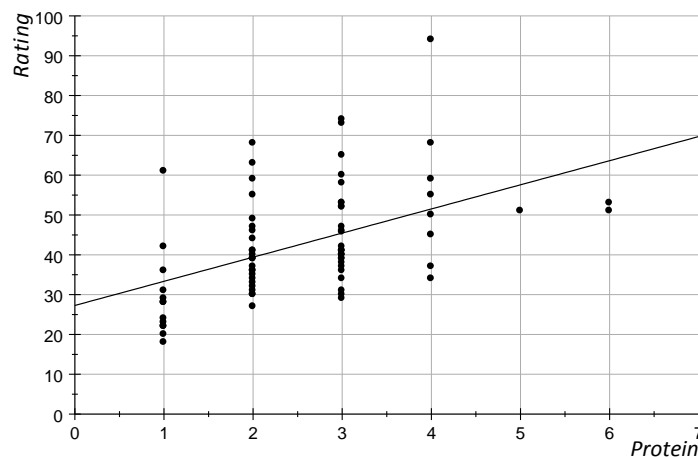
Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

- C Let's focus on the numbers in the given line equation: $\text{Predicted rating} = 28 + 6 \cdot \text{protein}$. First, what does the 28 tell us about the amount of protein in cereals and *Consumer Reports* ratings? Be as specific as you can.
- D Looking at the scatterplot and the line shown, what is special about the 28?
- E Recalling the equation of the line again, $\text{Predicted rating} = 28 + 6 \cdot \text{protein}$, what does the 6 tell us about the relationship between the amount of protein in cereals and *Consumer Reports* ratings? Be as specific as you can. It may be helpful to look back at the table you made using the equation of the line.

$x = \text{protein}$ (g/serving)	$\hat{y} = \text{predicted}$ rating
0	28
1	34
2	40
3	46
4	52
5	58

- F Below is the scatterplot of protein and ratings again. We learned above that the 28 is where the line crosses or "intercepts" the y-axis. In fact, it is called the **y-intercept**. But, can you "see" the 6 on the line? Use your answer for part E to help you find 6. This number, 6, is called the **slope** of the line.



Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

YOU NEED TO KNOW

Notice that the line equation has the form:

$$\text{Predicted } y = a + b \cdot x \text{ or } \text{Predicted } y = b \cdot x + a$$

The numbers a and b in the line equation have the following properties:

- a (the number that stands alone on the right-side of the equation) is called the **y-intercept** of the line. The y-intercept of a line is the point at which the line crosses the y-axis. a is also called the **initial value** because it is the predicted value of the response variable when the explanatory variable is 0.
- b (the number multiplied by the x or explanatory variable) is the **slope** of the line. b describes the change in the response variable when the explanatory variable increases by one unit.

You can interpret a line equation as:

$$\hat{y} = a + b \cdot x$$
$$\hat{y} = (\text{Initial Value}) + (\text{Slope}) \cdot x$$

- 3 Below is the line of best fit for predicting *Consumer Reports* ratings of cereals, based on the amount of sugar (in grams) in a serving. Recall, we are providing this equation for you now and in the next lesson, we'll find out to calculate this for ourselves. The table shows the amounts of sugar, in grams, per serving, and the predicted ratings that were calculated from the regression line.

$$\text{Predicted rating} = 60 - 2.4 \cdot \text{sugar}$$

$$\hat{y} = 60 - 2.4x$$

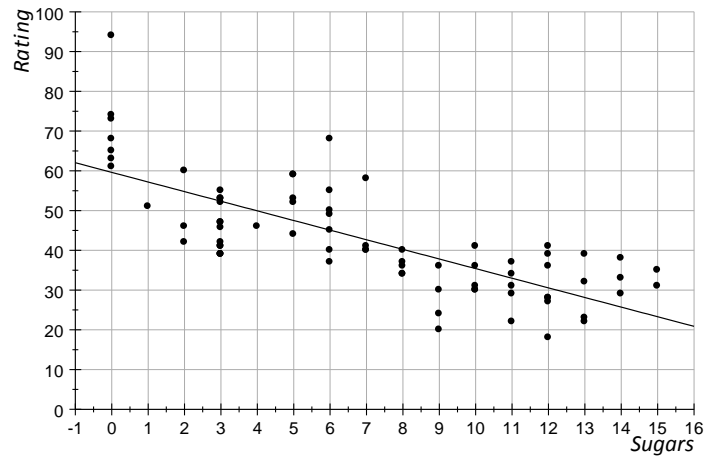
$x = \text{sugar}$ (g/serving)	$\hat{y} = \text{predicted}$ rating
0	
1	57.6
2	55.2
3	
4	
5	48

- A Use the equation above to fill in the missing values in the table.

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

- B Plot any three points you choose in the table on the scatterplot shown below.



- C What does the 60 tell us about the relationship between the amounts of sugar in cereals and *Consumer Report's* ratings? Use the table of values to help you answer the question.

Fill in the blanks:

60 is the predicted _____ when a cereal has _____ grams of sugar per serving.

Use this as a model sentence when asked to interpret the meaning of the ***y*-intercept**.

- D What does the -2.4 tell us about the relationship between the amounts of sugar in cereals and *Consumer Report's* ratings? Why is the word “decreases” in the sentence below? Use the table of values to help you answer the questions.

Fill in the blanks:

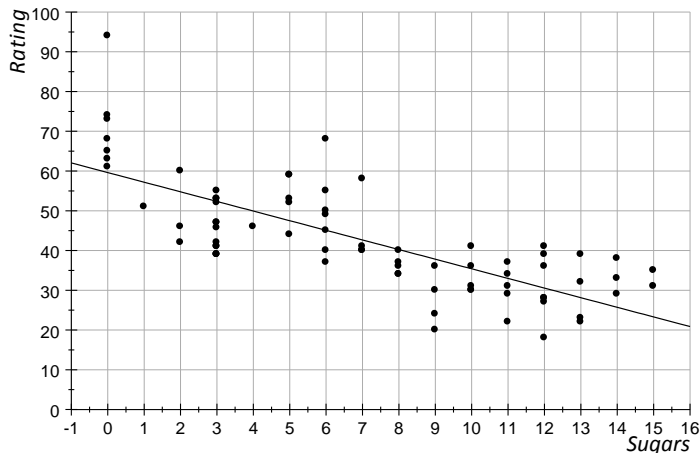
2.4 is the amount that a predicted _____ decreases by when the _____ is increased by _____ gram per serving.

Use this as a model sentence when asked to interpret the meaning of the ***slope***.

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

- E The scatterplot of sugar and the predicted ratings is given below. How can you see the 60 and -2.4 demonstrated in the graph of the line equation? Use your answers above to help you.

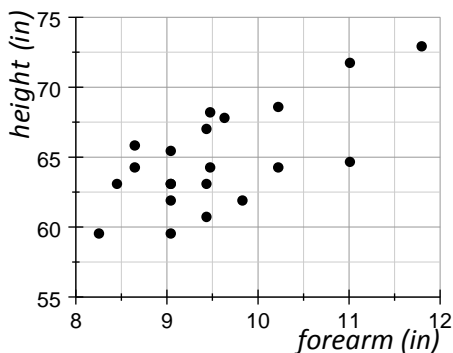


TRY THESE

Let's return to another data set we studied recently. In this data set, we were given the forearm and height measurements for 21 female college students. These students were taking Introductory Statistics, in 2009, at Los Medanos College in Pittsburg, California. The equation of the best fit line is:

$$\text{Predicted height} = 39 + 2.7 \cdot \text{forearm length}$$

$$\hat{y} = 39 + 2.7x$$



- 4 What do the numbers 2.7 and 39 in the equation tell you? Refer to the fill in the blank sentences above for some guidance.
- 5 Does the initial value or y-intercept make sense in the context of this data? Why or why not?
- 6 If a female college student's forearm length increased by 4 inches, how would her predicted height change?

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

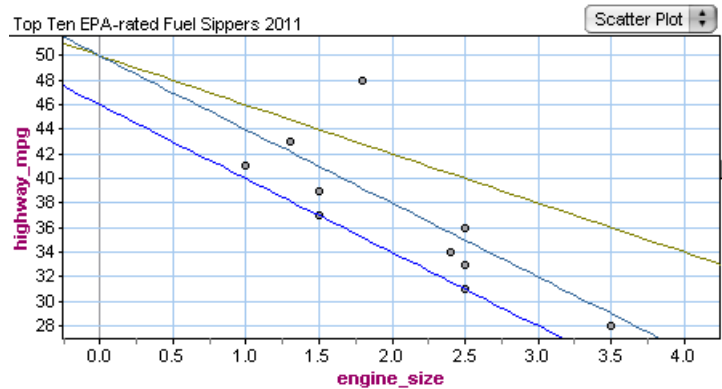
NEXT STEPS

Identifying the Best Fit Line

- 7 The Environmental Protection Agency picks the 10 most fuel-efficient cars each year. Below is a scatterplot of the highway miles per gallon and the engine size (measured in liters) for the EPA's top 10 for 2011. Cars with bigger engines usually get lower miles per gallon. (Source: www.fueleconomy.gov)

The equations of the three lines in the graph are:

- a. $\hat{y} = 46 - 6x$
- b. $\hat{y} = 50 - 6x$
- c. $\hat{y} = 50 - 4x$



- A Label each line with the letter of the correct equation (*a*, *b*, or *c*).
- B One of the lines is the *line of best fit*. Write the equation for the line of best fit, below. Briefly explain how you made your decision.
- C What is the slope of the line that you chose? What does the slope tell you about the relationship between engine size and highway miles per gallon?
- D What is the y-intercept of the line that you chose? Interpret the y-intercept in the context of this problem. Explain why your interpretation does or doesn't make sense for this problem.

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

8 Which of the following equations matches the lines A and B?

a. $y = 4 - 2x$

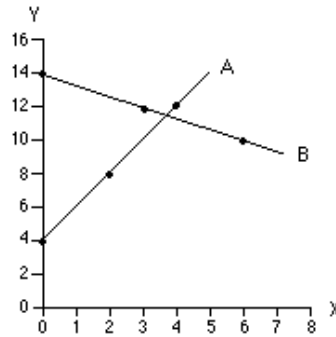
d. $y = 4 + 2x$

b. $y = (2/3)x + 14$

e. $y = 14 - (3/2)x$

c. $y = 1/2x + 4$

f. $y = 14 - (2/3)x$



SUMMARY

- There are two important numbers in the equation that defines the line of best fit: slope and y-intercept.
- The slope is the number attached to the explanatory variable or x-term, and the y-intercept is the number that stands alone.
- The slope is described by a ratio of the change in y (the response variable) relative to the change in x (the explanatory variable) and can be interpreted as the predicted change in y associated with a one unit increase in x .
- The y-intercept is the y -value when x (or the value of the explanatory variable) = 0. Frequently, this point does not have meaning in the context of the data because $x = 0$ lies outside the range of the data. The initial value is still a necessary part of the equation of the line.

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

STUDENT NAME _____ DATE _____

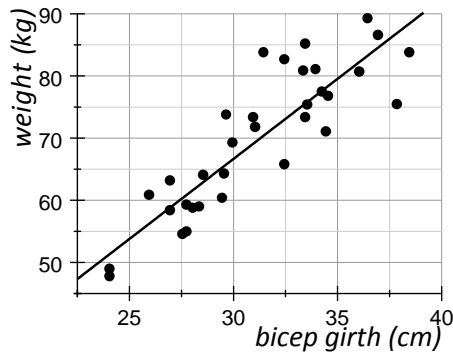
TAKE IT HOME

- 1 Based on data from 34 adults who exercise regularly, the best fit line for the relationship between bicep girth (the length in centimeters around the upper arm) and the person's weight is

$$\text{Predicted weight} = 2.6 \cdot \text{bicep girth} - 10.5$$

$$\hat{y} = 2.6x - 10.5$$

where predicted weight is measured in kilograms and bicep girth is measured in centimeters.



- A Complete the table below using the given line equation. Use the equation NOT the graph of the line.

$$\text{Predicted weight} = 2.6 \cdot \text{bicep girth} - 10.5$$

$x = \text{bicep girth}$ (in cm)	$\hat{y} = \text{predicted weight}$ (in kg)
0	
20	
25	
33	
45	

- B In the line equation, when $x = 0$, the initial value is -10.5 . Does this number, -10.5 , have meaning in this scenario? Why or why not?
- C Explain the meaning of the slope of the line equation. Tell how the slope relates a person's bicep girth to his or her predicted weight. Refer back to the fill in the blank sentences in the lesson for a guide.
- D If the bicep girth of an adult increased by 3 cm, how would their predicted weight change?

Lesson 3.6

Investigating the Slope and Y-intercept of the Line of Best Fit

- 2 With the following applet, you will investigate how outliers impact the line equation. This link is given in the Blackboard course, so you don't have to type it in.

<http://www.stat.berkeley.edu/~stark/Java/Html/Correlation.htm>

Click (turn off) the "SD Line" and "Graph of Ave". Click (turn on) the "Regression Line" and "Use Added Points". You can add points anywhere by clicking. Change the value of n to 10 to see a more dramatic effect when adding (clicking) outliers.

- A Add points to the scatterplot that are close to data points shown. Describe what happens to the line of best fit or regression line.
- B A data point is an *outlier* if it is far away from the other data points. Add points to the scatterplot that are outliers relative to the other data points. In other words, add points that are far away from the other data points. Describe what happens to the best fit line or regression line.
- C What happens to the correlation coefficient, or r -value, when you add points that are outliers? Explain your understanding of why this happens.

Lesson 3.7

Least Squares Regression Line as Line of Best Fit

INTRODUCTION

Comparing Lines for Predicting Textbook Costs

A *response variable* is one whose value we can estimate based on the value of another variable. This other variable is called the *explanatory variable*. When a linear pattern exists in a scatterplot representing these variables, we can use a *line of best fit* to represent this pattern.

How do we identify the line of best fit? We often use technology to find the equation of this line, but what makes it a *line of best fit*? In this lesson, we investigate this question with the goal of explaining what we mean by *best-fit*.

- 1 Below are suggested prices for 12 popular introductory statistics textbooks from different publishing companies.

\$119.00	\$151.95	\$122.00	\$158.95	\$107.95	\$122.00
\$149.10	\$166.15	\$191.00	\$150.67	\$182.00	\$150.67

The table below gives the five-number summary, mean, and standard deviation for the price data.

Min	Q1	Median	Q3	Max	Mean	Standard Deviation
\$107.95	\$122.00	\$150.67	\$162.55	\$191.00	\$147.62	\$25.73

- A Imagine someone asks you about how much an introductory statistics textbook costs. What would you tell them? Explain your reasoning.
- B These textbooks do not all cost the same. What are some variables that might help us predict the cost of an introductory statistics textbook?

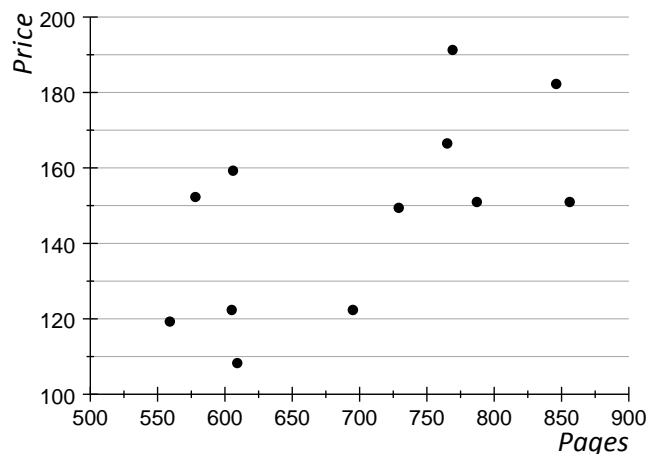
Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- C The number of pages in the textbook is one variable we could use to predict price. In the table below, the number of pages in each book is listed together with the price of the book.

Pages	560	579	606	607	610	696	730	766	770	788	847	857
Price	119.00	151.95	122.00	158.95	107.95	122.00	149.10	166.15	191.00	150.67	182.00	150.67

The scatterplot shows the relationship between pages and price for these 12 textbooks.



The data have a somewhat linear form and the correlation coefficient is 0.62. It makes sense to use a line to summarize the relationship.

Draw a line that you think is a good summary of the relationship between these two variables. Use your line to estimate the price of a 650-page textbook.

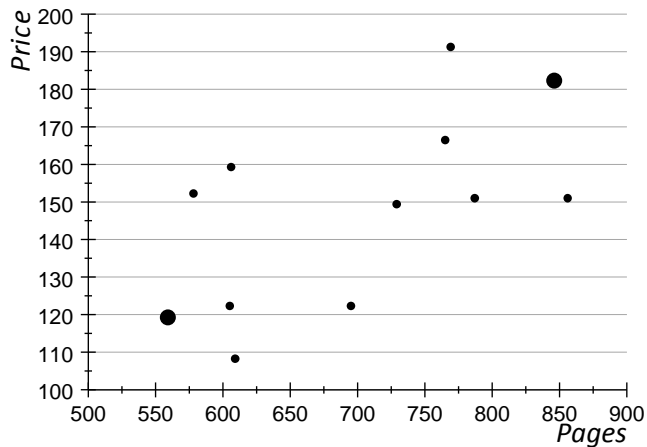
predicted price = _____

- 2 There is no limit to the lines that we *could* draw. How might we decide on the *best* line? Is it necessary to have the line go through one or more of the data points?

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- 3 One way to create a line would be to choose two of the data points that seem to represent the center of the linear pattern. The scatterplot is graphed again below, but this time we have chosen two points, (560, 119.00) and (847, 182.00) that seem to lie in the center of the linear pattern. We have highlighted these points with *large* dots.



Use a straight-edge to draw a line through the (larger) chosen points. Use this line to predict the price of a textbook with 650 pages.

predicted price = _____

How closely does your estimate match your prediction from Question 1, Part C?

Draw the point representing the predicted price of a textbook with 650 pages on the line you drew above.

- 4 The equation of the line in Question 3 is:

$$\text{Predicted price} = 0.22 \cdot \text{pages} - 3.92$$

Use this equation to predict the price of the 650 page textbook.

predicted price = _____

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

The equation above is a first attempt to fit a line to a linear pattern. But, we want to know how well the line fits the data points. To do this we must develop a way to measure how well the line *fits* the data points. In Statistics, a line is a good fit if it is close to as many of the points as possible. But to compare lines we need a more precise measure.

To measure fit, we summarize the distances from the points to the line. We specifically look at vertical distance from each point to the line. The **y values on the line** represent **predicted price**. To find the distances, we subtract the predicted price from the actual price. This type of deviation is known as a **residual** or **prediction error**. The formula is given below.

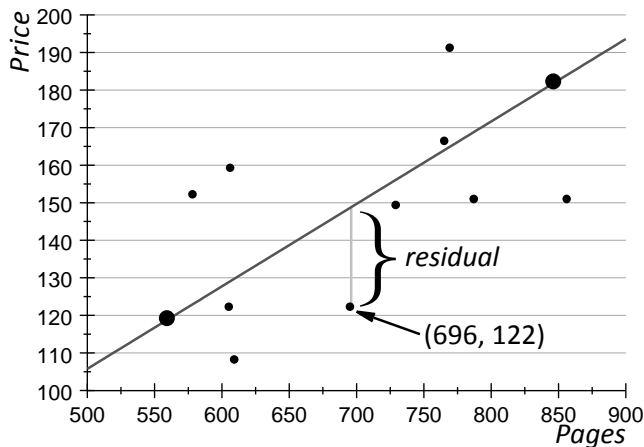
$$residual = data\ y\text{-value} - predicted\ y\text{-value}$$

comes from the actual data point

calculated from the line equation

If this seems familiar it is because this is similar to the first step in calculating the standard deviation. In that case we found the deviation of the data from the mean. In this case we find the deviation of the data from the line.

In the graph below, the residual is labeled for a single point: (696, 122). This data point represents an actual observed data point: a book with 696 pages, costs \$122. The **residual** is the vertical distance between the actual price and the predicted price on the line. Remember the predicted price comes from the line equation.



Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- 5 Use the equation of the line to find the predicted price of the 696 page book.

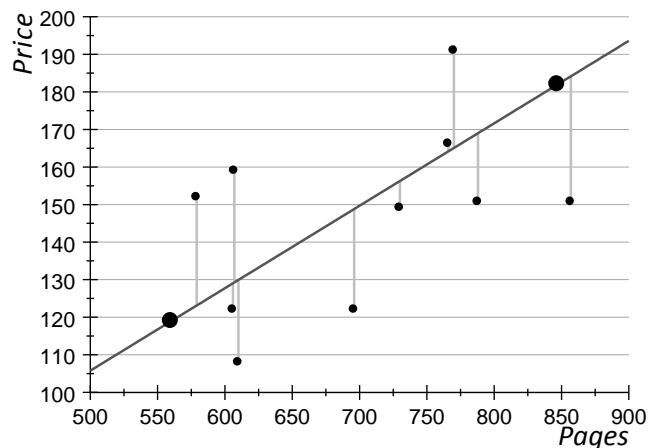
$$\begin{aligned}
 \text{Predicted price} &= 0.22 \cdot \text{pages} - 3.92 \\
 &= 0.22 \times \underline{\hspace{2cm}} - 3.92 \\
 &= \underline{\hspace{2cm}}
 \end{aligned}$$

- 6 The residual is the difference between the *actual* price of the 696 page book and the *predicted* price (from the equation of the line). Compute the residual for this book below.

$$\begin{aligned}
 \text{residual} &= \text{actual price} - \text{predicted price} \\
 &= \underline{\hspace{2cm}}
 \end{aligned}$$

- 7 The residual you just computed is negative. What does this tell us about the location of the actual data point, in relation to the line?

The scatterplot below displays the residuals for all the actual data points using vertical line segments. Note that the two data points we used to draw our line have residuals equal to 0 since they are on the line.



We want to summarize the residuals for *all* points. We'd like to have a single number that describes the residuals. As we think of such a number we must remember that large values of the residual indicate that the data point is far from the line, while small values of the residual indicate that it is near.

We could simply average all the residuals to get a single number that represents the fit. But because some of the residuals are positive and others are negative, they tend to cancel out. And for the line of best fit, they always average to zero. We saw this with the deviations from the mean when calculating a standard deviation. We can solve this problem by squaring the residuals (just as we did when

Lesson 3.7


Least Squares Regression Line as the Line of Best Fit


calculating the standard deviation) and then summing them up. We use the sum of squared residuals as our measure of fit. This is denoted by SSE (sum of squares of error)

- 8 If we use the sum of squared residuals as a measure of *fit*, does a smaller or larger sum indicate a better fit? Explain your answer.

- 9 Compute the missing predicted prices, residuals and their squared residuals in the table below. The column of “Pages” and “Actual Price” comes from the original data set we were working with. It is given back in question 1C. Then add the squared residuals. Recall the formula for calculating residuals:

$$\text{residual} = \text{data } y\text{-value} - \text{predicted } y\text{-value}$$


 comes from the
actual data point


 calculated from
the line equation

For this problem, the line equation we are using is: $\text{Predicted price} = 0.22 \cdot \text{pages} - 3.92$

<i>Pages</i>	<i>Actual Price</i>	<i>Predicted Price</i>	<i>Residual</i>	<i>Residual²</i>
560	119.00	119.28	-0.28	0.08
579	151.95	123.46	28.49	811.68
606	122.00	129.40	-7.40	54.76
607	158.95	129.62	29.33	860.25
610	107.95	130.28	-22.33	498.63
696	122.00	149.20	-27.20	739.84
730	149.10	156.68	-7.58	57.46
766	166.15	164.60	1.55	2.40
770	191.00			
788	150.67			
847	182.00			
857	150.67			
Sum of squared residuals:				

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

10 Looking at the Squared Residual column, which of the points is closest to the line? Which point is farthest from the line?

11 We are calculating the sum of the squared residuals. We will need to add up all the numbers in the “Squared Residual” column. To save you some time so that you don’t have to type all of the squared residuals into the calculator, the sum of the numbers that were originally shown in the table is 3025.1. So to find the total of the entire column, take this number and add in the squared residuals for the data points you found. What is the total sum of all the squared residuals in this example?

Total sum of squared residuals is: $3025.1 + \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$

In summary we have decided to use SSE as a measure of how well a line fits the data. We have chosen a particular line and calculated $SSE = 5181.46$. The line looks like a good fit but is it the best we could have done? This process of finding SSE is long and tedious. We will turn to technology to determine if we can find a line with a smaller SSE.

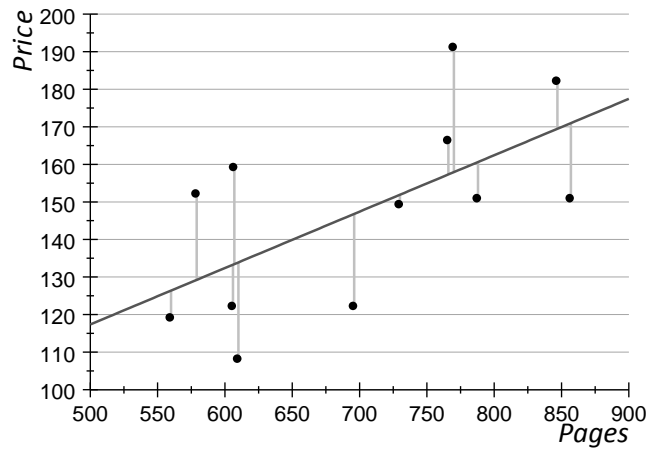
The line of best fit (also called the regression line or prediction line) is the line that gives the smallest SSE. We have estimated it well using trial and error with the computer but in practice there are formulas for the coefficients (a and b) that have been derived mathematically. These formulas have been programmed into the computer and calculator and we will use them to get the equation of the line of best fit. We will see how to use the calculator and Statcrunch to do that soon but for now the equation of the line of best fit will be given to us. The equation of the line of best fit (regression line) is

$$\text{Price} = 0.15 * \text{Pages} + 42.32$$

The line of best fit is plotted in the scatterplot below, along with depictions of the residuals. It is hard to see visually that this fit is better than the first line. In fact this line does not pass through **any** of the points in the scatterplot.

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit



- 12 Use the equation to predict the price for the book with 696 pages. Recall the equation was:
$$\text{Price} = 0.15 * \text{Pages} + 42.32$$
- 13 Compute the residual for this book below. Recall: $\text{Residual} = \text{Actual price} - \text{Predicted price}$

YOU NEED TO KNOW

To measure variation of points from a line, we added the squared residuals. The line of best fit is a *least squares line* because it minimizes this sum of squares. This means that compared to other lines, the line of best fit gives us the smallest sum of squared residuals. **By this measure, no other line fits the data better.**

The line of best fit has several names. It is the *least squares* line or the prediction line. Sometimes we call it the least squares *regression* line, or simply abbreviate it as the *LSR line*.

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

NEXT STEPS

Thinking about Residuals

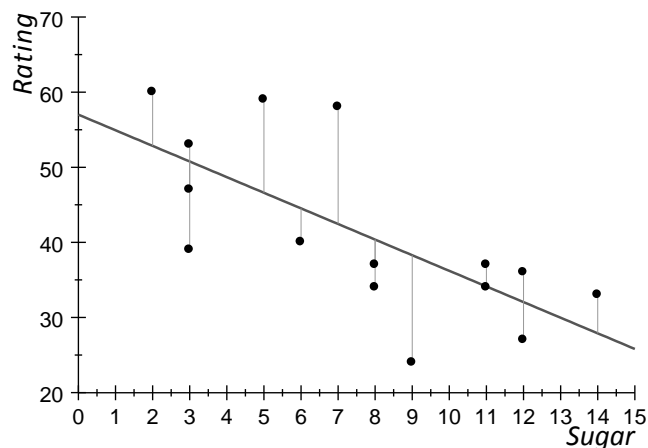
In the following activity, we use technology to compute a line of best fit. We will also think some more about residuals. Recall that the response variable is graphed on the y-axis. The value of a residual is

$$\text{residual} = (\text{data } y\text{-value}) - (\text{predicted } y\text{-value}).$$

So the data y-value comes from the actual data set and the predicted y-value comes from the line equation.

- 14 In this activity, we return to the relationship between sugar in breakfast cereals and their Consumer Reports ratings. In the table below are the grams of sugar per serving in 15 randomly selected cereals and each cereal's Consumer Reports rating. The scatterplot below the table includes the line of best fit, and the residuals.

Sugar (grams)	6	8	7	11	11	3	3	12	12	2	14	8	9	5	3
Rating	40	37	58	34	37	39	53	27	36	60	33	34	24	59	47



- A Use technology to compute the correlation coefficient and line of best fit.
- B Use the equation of the line of best fit to predict the rating of a cereal with 6 grams of sugar.
- C What is the residual for the cereal in the sample that had 6 grams of sugar?

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- D What does the sign of a residual tell you about the location of its actual point relative to the line of best fit?

- E Find and circle the data point on the graph with the largest *negative* residual. Use the table of values to determine the coordinates of this point.

- F Use the equation of the line of best fit to predict the rating for this cereal.

- G Compute the residual for this point.

- H Find and circle the actual data point on the graph with the largest *positive* residual. Write the point's coordinates and compute the residual for this point.

- I What does the sign of a residual tell you about the location of its point relative to the line of best fit?

- J Is there a line that has a smaller SSE than the line we used above? Explain.

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

STUDENT NAME _____ DATE _____

TAKE IT HOME PART 1

- 1 Sign into Statcrunch at <http://www.statcrunch.com>. Follow the instructions below to run the *Regression by Eye* applet. Here is a link to a video to show you how to run the applet: <http://screencast.com/t/7nWiAcl2hZm> Note: This link is also posted in Blackboard.

Instructions

- i. Click on *My StatCrunch* near the top of the page.
- ii. Click on *Open StatCrunch*.
- iii. Click on the blue StatCrunch button above the table, right above the word "Row".
- iv. Choose Applets from the drop down menu and select "Regression by eye".
- v. Click on "Create Applet" in the bottom right of the window that opened.

The applet should open with a scatterplot drawn, using randomly generated data, and a green line. Also, notice the table at the bottom of the window. You can move the line by clicking on the endpoints of the line and moving them around. Notice that the Sum of Squared Errors (SSE) is recorded in the table located at the bottom of the window. The slope and y- intercept of the equation of your line is also recorded there.

- A As you move the line around, your goal is to find the line of best fit, that is, the line with the smallest SSE. Record your results in the table below.

	Equation of Line $Y = \text{slope} *x + \text{intercept}$	SSE
Your line: User	$Y = \underline{\hspace{1cm}}x + \underline{\hspace{1cm}}$	

- B If you click on one of the data points, you can remove it from the set by dragging into the trash. Remove a few data points. Record your new line and SSE.

	Equation of Line $Y = \text{slope} *x + \text{intercept}$	SSE
Your line: User	$Y = \underline{\hspace{1cm}}x + \underline{\hspace{1cm}}$	

- C What effect does this have on the SSE you found in part A. Why does it have that effect?

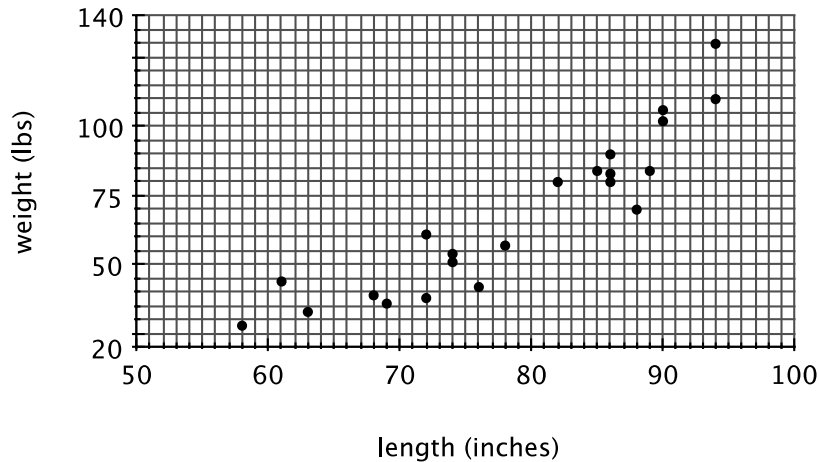
- D By clicking on the graph, you can add a point to the data set. Click to add a point far away from the rest of the data points. This is an outlier. What effect does this have the SSE? Explain why it would have that effect.

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- 2 The scatterplot below shows the lengths (inches) and weights (pounds) of a sample of alligators. One equation for a line that represents the data is

$$\text{Predicted weight} = -123 + 2.4 \cdot \text{length}$$



- A Use the equation given to find the predicted weight of an alligator that is 66 inches long. Plot this point on the scatterplot above.
- B Use the line equation given, $\text{Predicted weight} = -123 + 2.4 \cdot \text{length}$, to find the predicted weight of an alligator that is 92 inches long. Plot this point on the scatterplot.
- C Use a straight edge to draw the line connecting these two points on the graph.
- D One of the alligators in the sample measured 88 inches long and weighed 70 pounds. Circle this data point on the scatterplot.
- E Find the residual for this data point. Look back in the lesson to find the formula to calculate a residual.
- F Draw the vertical line on the scatterplot that represents this residual. Look back at the graph that went with question number 7 in the lesson.

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

STUDENT NAME _____ DATE _____

TAKE IT HOME PART 2

1 Here is data collected from students at Los Medanos College in 2009. The variable, *credits*, is the number of credits each student took that semester. The variable, *textbooks*, is the amount students spent on the textbooks that were required for their courses that semester. The credits and textbooks data come from student reports on a survey.

Credits	Textbooks
3	120.25
4	65.95
9	465.00
12	430.00
14	396.50
16	475.00
8	208.00
1	5.00
6	49.10
15	685.00
9	220.00
4	172.00
12	302.00
12	460.12
12	530.00

A Use technology to find the least squares regression (LSR) line. (Think carefully about which variable is *eXplanatory*.)

B Use the LSR line equation given,
 $predicted\ textbooks = -42.914 + 38.158 \cdot credits$
to predict the amount spent on textbooks for a student taking 12 units.

C Calculate the residual for the data highlighted in the table.

D Explain what the sign of the residual you calculated tells you about the data value.

Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- 2 A statistics class collected the following data for the class: height (inches) and navel height (inches).
- A Which is the explanatory variable?
- B Arif, Bob, and Catherine each found a linear equation to fit the data. They are shown in the table below, along with the associated SSE for the line they found.

	Line Equation	SSE
Arif	$\hat{y} = 9.4 + 0.45x$	556.3
Bob	$\hat{y} = 8.9 + 0.32x$	823.2
Catherine	$\hat{y} = 10.1 + 0.42x$	678.5

Is it possible to determine whose line fit the data best? If so, whose line best fit the data?

- C If you were able to determine whose line fit the data best, how were you able to tell? If you are unable to determine whose line is best, explain why.

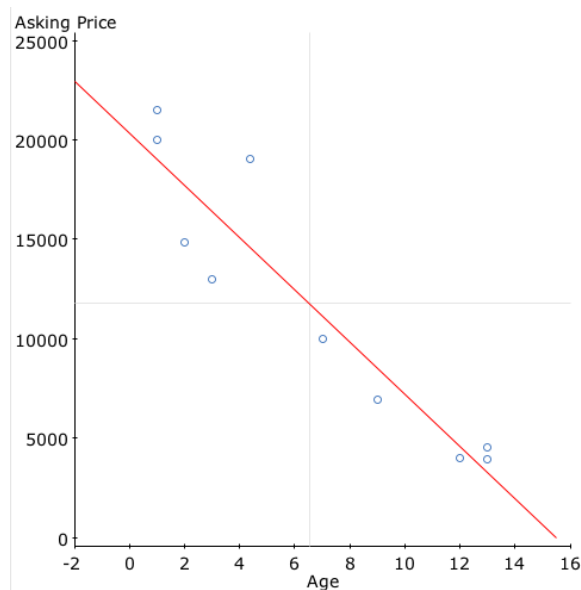
Lesson 3.7

Least Squares Regression Line as the Line of Best Fit

- 3 The following table shows data that was collected from Craig’s List. It shows the age and the asking price of various Toyota Camry’s.

Age	3	13	1	2	12	1	7	13	9
Asking Price	12995	3899	21495	14822	4000	19995	9995	4500	6900

The scatterplot for this data is shown below along with the Least Squares Regression Line (LSR).



- A Use your calculator to find the equation of the LSR line. Fill in the blanks below. There is a link in the Blackboard course showing you how to do this. Look in the Help Videos. Also, here is the link: <http://screencast.com/t/b47bEmVCE>

Predicted asking price = _____ + _____ • age

- B Estimate the correlation coefficient by choosing an r - value from the following list. Then explain your choice. If you don’t remember what the correlation coefficient is, go back to lesson 3.1.3 and review.

i) $r = 0.3$ ii) $r = 0.9$ iii) $r = -0.3$ iv) $r = -0.9$

- C Imagine you see a 4 year old Toyota Camry for sale and the asking price is \$12,000. Is this a good price for the car? Explain how you know this by discussing the scatterplot and the regression line.

Lesson 3.8 (OPTIONAL)

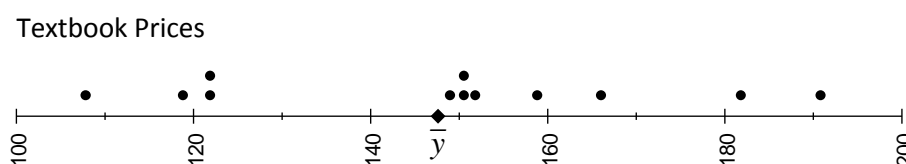
Using Explained Variation to Measure Fit

INTRODUCTION

In earlier lessons we saw data for the prices for 12 popular elementary textbooks.

\$119.00	\$151.95	\$122.00	\$158.95	\$107.95	\$122.00
\$149.10	\$166.15	\$191.00	\$150.67	\$182.00	\$150.67

A dotplot for the prices is below.



The mean book price is $\bar{y} = \$147.60$. This mean is plotted on the dotplot axis above.

There is a lot of variability in this elementary textbook price data set. In fact, none of the points are equal to the mean. The Standard Deviation helps us describe this variability. Remember that the standard deviation is an average of deviations from the mean.

- 1 The second most expensive elementary textbook costs \$182.00. We can calculate its **total** deviation from the mean just as we did to compute standard deviation.

Language Tip

The **total deviation** for a point on a scatterplot is the difference between the **y-value** and the mean of all y-values.

Total deviation from the mean: $y - \bar{y} = \underline{\hspace{2cm}}$

The standard deviation describes variability, but it doesn't *explain* it. How can we explain the variability? So far we haven't considered *why* individual values deviate from the mean. There are often reasons that *explain* the variability, in this case, the reason why some textbooks are more expensive than average and others are less. In the next few problems, we will think about why there is variability in the price of textbooks. We will try to *explain* the variability; that is, we will look for *explanatory* variables.

What is it that makes a \$182 text so expensive? Could the price be explained by quality? What about demand? In earlier lessons, we learned that number of pages in a textbook is related to the price of the textbook. The explanatory variable, *pages*, can be used to explain *some* of the deviation above the mean price.

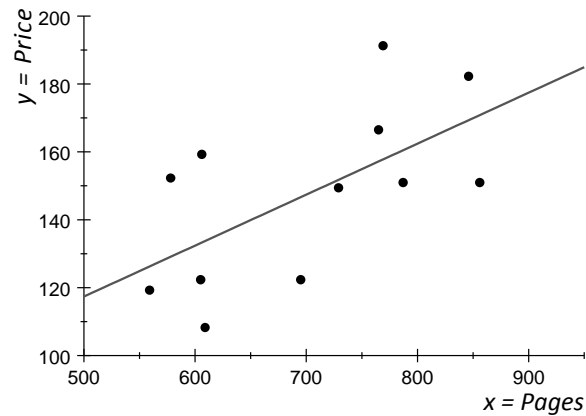
Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

In the table below, the number of pages in each book is listed with the corresponding price.

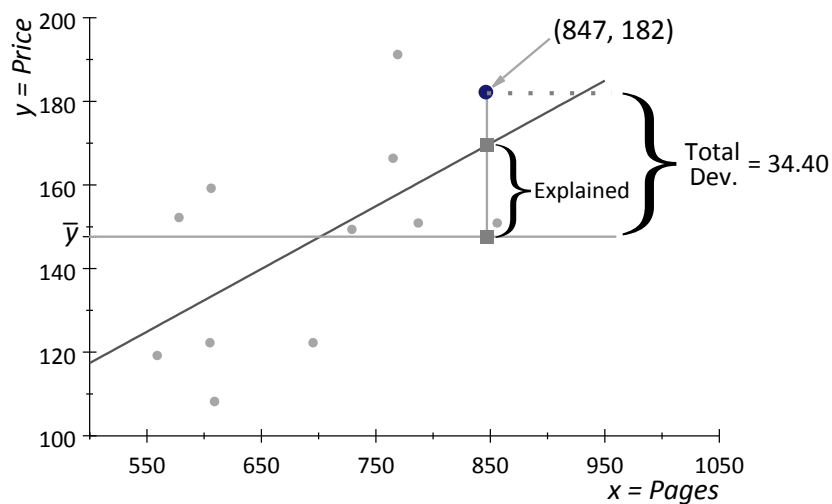
x = Pages	560	579	606	607	610	696	730	766	770	788	847	857
y = Price	119.00	151.95	122.00	158.95	107.95	122.00	149.10	166.15	191.00	150.67	182.00	150.67

As the pages in a book increases, price tends to increase as well. The line of best fit is plotted with a scatterplot of the data points below.



When we look at the scatterplot, the pattern is roughly linear. So we can say that *some* of the deviation from the mean price is explained by pages through a linear model.

The \$182 textbook has 847 pages. In question 1 we saw that the \$182 textbook has a total deviation from the mean of \$34.40. In the figure below, the mean, $\bar{y} = 147.60$, is represented as a horizontal line. A vertical line represents the total deviation of the \$182 textbook from the mean price.



Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

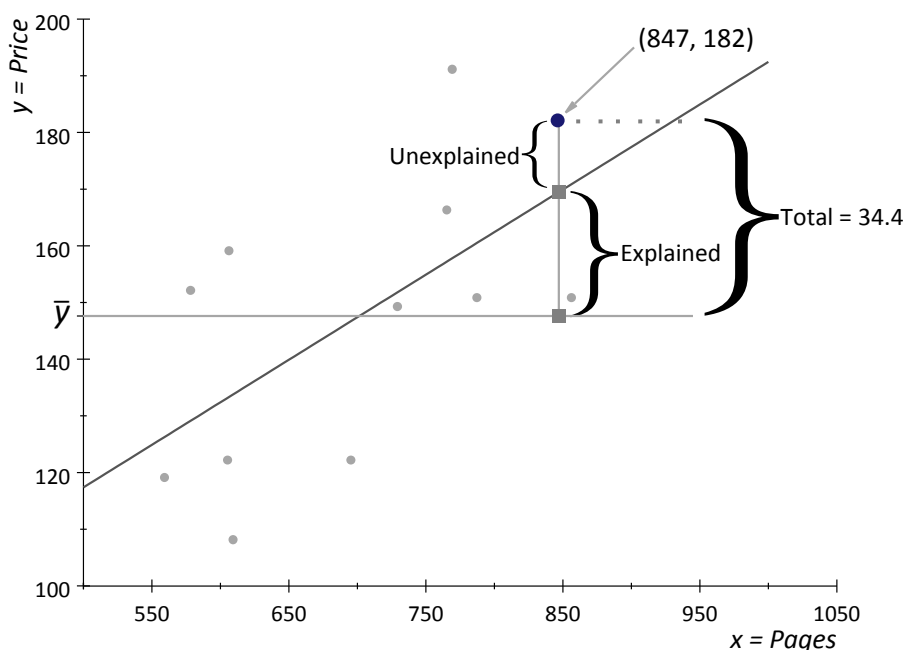
The part of total deviation that is **explained** by the number of *pages in the textbook* is the *deviation from the mean predicted by the line*. In the figure above, it is represented by the line segment identified as *Explained*.

Language Tip

The **explained deviation** is the portion of the total deviation that is **predicted** by the explanatory variable.

- 2 Look at the scatterplot above. Visually estimate the *proportion* of the *total* deviation for our \$182 textbook that is explained by the number of pages. Express the proportion as a number between zero and one, and as a percent. (What fraction of the total deviation is explained? Then turn your fraction into a decimal and a percent.)

In the image below, the total deviation from the mean for the \$182 textbook is plotted again as a vertical line. This deviation is broken into two parts - the part *explained* by *pages* through the regression line, and the part that is *unexplained* by this.



- 3 For the \$182 textbook, the amount of total deviation (34.4) that is *explained* by the number of *pages* through the regression line is equal to 21.9. What proportion of the total deviation is this? Write your answer as a proportion and as a percent.

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

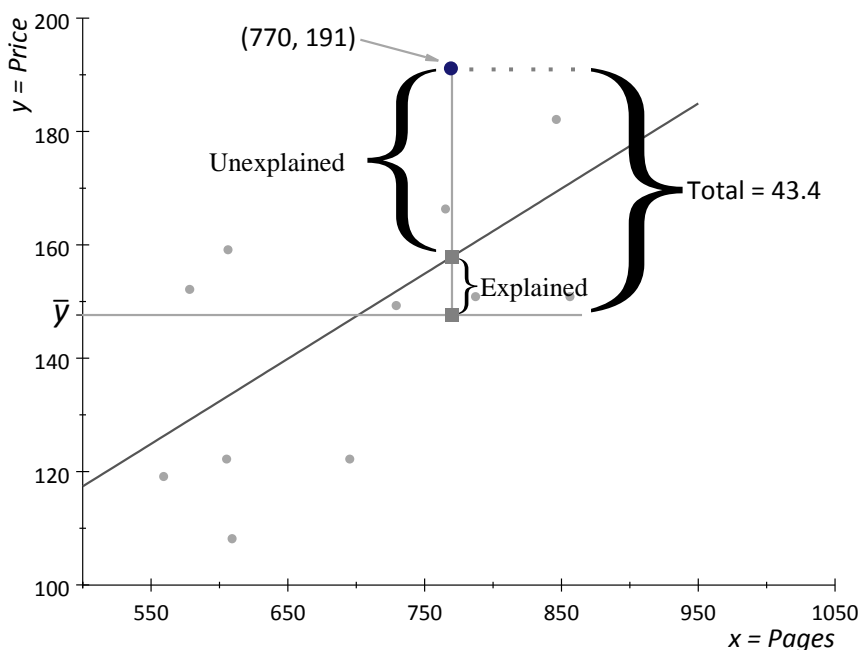
- 4 We have studied the part of total deviation that is *unexplained* by the number of *pages* through the regression line. We can compute this unexplained deviation by subtracting a predicted y -value from an observed y -value. What is another name for result when the predicted y -value is subtracted from the observed y -value?

Language Tip

The *unexplained deviation* is the difference between the value of y predicted by the regression line and the actual value of y , also called the *residual*.

- 5 The most expensive book in the data set costs \$191. What is the total deviation of this value from the mean price?
- 6 Look at the scatterplot image below. Do you think that the number of *pages* does better or worse at explaining the total deviation above the mean for the \$191 text, compared to the \$182 text?

The image below breaks down the total deviation from the mean for the \$191 text into explained and unexplained parts.



- 7 Look at the scatterplot above. Visually estimate the *proportion* of the *total* deviation for our \$191 textbook that is explained by the number of *pages*. Express the proportion as a number between zero and one, and as a percent. (What fraction of the total deviation is explained? Then turn your fraction into a decimal and a percent.)

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

- 8 For the \$191 textbook, the amount of total deviation (43.4) that is *explained* by the number of *pages* through the regression line is equal to 10.2. What proportion of the total deviation is this? Write your answer as a proportion and as a percent.
- 9 Does your answer (to question 8) support or not support your answer from Question 6?

YOU NEED TO KNOW

The **total** deviation of a variable, y , from the mean, \bar{y} , is $y - \bar{y}$. The **explained deviation** is the part of the total deviation that is predicted by the explanatory variable. If the relationship between x and y is not perfect, then each point has a residual. This residual is the part of total deviation that is **unexplained** by the explanatory variable. Together, the explained and unexplained deviations make up the total deviation of a variable.

$$\textit{explained deviation} + \textit{unexplained deviation} = \textit{total deviation}$$

The *proportion* of total deviation that is **explained** by x is:

$$\frac{\textit{explained deviation}}{\textit{total deviation}}$$

The *proportion* of total deviation that is **unexplained** by x is:

$$\frac{\textit{unexplained deviation}}{\textit{total deviation}}$$

The *sum of the proportions* of total deviation which are explained and unexplained by x is equal to one.

$$\frac{\textit{explained deviation}}{\textit{total deviation}} + \frac{\textit{unexplained deviation}}{\textit{total deviation}} = 1$$

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

NEXT STEPS

In this lesson so far, we have used the proportion of total deviation that is explained by the explanatory variable as a *measure of fit* for several individual points. However, in order to use the proportion of total deviation that is explained by the explanatory variable as a *measure of fit* for all of the data, we need to expand this idea so that it includes *all of the points* in the textbook data set.

There is a long and a short road to doing this calculation . . . To measure variability for *all* data points, it is useless to add the deviations, because we have already discovered that deviations will always add to zero. In addition, the sum of the *explained* parts of the deviation will always be zero as well. To find the proportion of total variation, we need to sum the *squares* of the deviations. That is the long road. However, there is a surprising relationship between the proportion of total variability that is explained by the explanatory variable and r , the correlation coefficient. This is the short road!

If we compute the proportion of the total variation in textbook price that is explained by the number of *pages* with the LSR line by taking the long road we find that it is approximately 0.383. This tells us that about 38% of the total variability among textbook prices is explainable by number of *pages* through the linear relationship model.

10 Let see what happens when we take the short road.

- A For the textbook data, $r = 0.619$. Compute r^2 .

$$r^2 = \underline{\hspace{2cm}}$$

- B How does your answer compare to the calculation of the proportion of total variability that is explained by the number of *pages* with the LSR line by taking the long road?

Since we have found that the proportion of total variability that can be explained by *number of pages* is 0.383 (the value of r^2) we can make the statement that **38.3% of total variability in *textbook price* is explained by *number of pages*.**

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

YOU NEED TO KNOW

The square of the linear correlation coefficient, r^2 , is the proportion of total variability in the response variable, y , that is explained by the explanatory variable, x , through the line of best fit. This value is also known as the *coefficient of determination*.

$$r^2 = r * r$$

Since r is always between -1 and 1 so, the value of r^2 is always between 0 and 1 .

When explaining the value of r^2 in the context of our problem we use the following sentence:

(r^2 as a percentage) % of total variability in the (response variable, y) is explained by the (eXplanatory variable, x) .

EXAMPLE: 38.3% of total variability in the *textbook price* is explained by the *number of pages*.

We saw earlier with the individual points (like each individual textbook), that the sum of the explained and unexplained proportions of total deviation is 1 . The sum of the explained and unexplained proportions of total deviation is also 1 when we use sums of squared deviations to compute the proportions for *all points*.

$$\left(\begin{array}{l} \text{proportion of} \\ \text{total variability} \\ \text{explained by } x \end{array} \right) + \left(\begin{array}{l} \text{proportion of} \\ \text{total variability} \\ \text{unexplained by } x \end{array} \right) = 1$$

- 11 When we use the number of pages to explain the price for statistics textbooks, what proportion of total variation is *unexplained* by the LSR line? (Hint: the explained and unexplained proportions add up to 1 .)

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

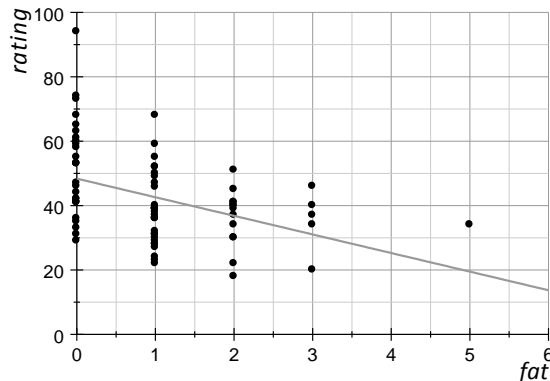
TRY THESE

The last few problems bring together several ideas discussed in this lesson, and in previous lessons. Among these ideas are:

- The coefficient of determination, r^2 , gives the proportion of total variability that is explained by x through the linear LSR model.
- The proportion of total variability that is unexplained by x through the LSR model is $1 - r^2$.
- The residual for a data point (x, y) is $y - \hat{y}$.
- When the residual for a data point is positive, the point is above the line of best fit.
- When the residual for a data point is negative, the point is below the line of best fit.

Use these ideas to help you solve the next few problems.

- 12 Consumer Reports magazine rated 76 breakfast cereals. The scatterplot below gives the amount of fat (in grams) for those cereals and the Consumer Reports rating for each of the 76 breakfast cereals.



- A The correlation coefficient for these data points is $r = -0.41$. What proportion of total variability in the y -values (the ratings) is *explained* by the cereals' fat content through the linear regression model?
- B What proportion of total variation is *unexplained* by the fat content through the linear LSR model?
- C The cereal with the most fat has a rating of 34. What is the fat content in a serving of that cereal?

fat, $x =$ _____

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

- D The point representing the cereal with the most fat is above the line of best fit. Use this fact to decide if the residual is positive or negative for this cereal.
- 13 The equation for the line of best fit is: $\hat{y} = -5.8x + 48.4$.
- A What is the predicted Consumer Reports rating for the cereal with the most fat?
- B Compute the residual for this cereal.
- C The cereal with the lowest rating had an 18. What is the fat content for a serving of this cereal?
- D The cereal with the lowest rating is represented by a point that is below the line of best fit. Use this fact to determine if the residual for this cereal is positive or negative.
- E Use the LSR equation to find the residual for this cereal.

SUMMARY

- A residual is the difference between an observed value and corresponding predicted value of the response variable.
- Positive residuals indicate that the response value is higher than the predicted response value, while negative residuals indicate that the response value is lower than the predicted response value.
- The total deviation of each y -coordinate from the mean, \bar{y} , is composed of two parts: the part explained by the model and the residual (the part unexplained by the model).
- The proportion of the total variability in y (*the response variable*) that can be explained by the x (*the explanatory variable*) through the regression model (LSR line) is given by r^2 .

Lesson 3.8 (OPTIONAL)

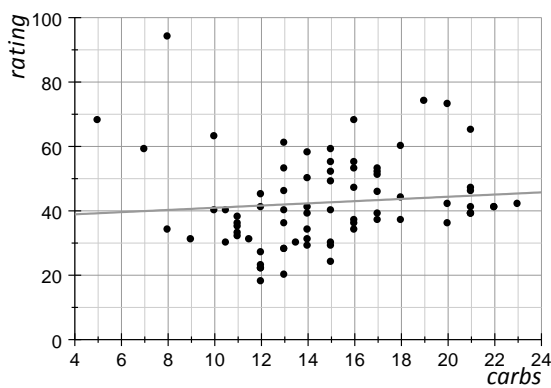
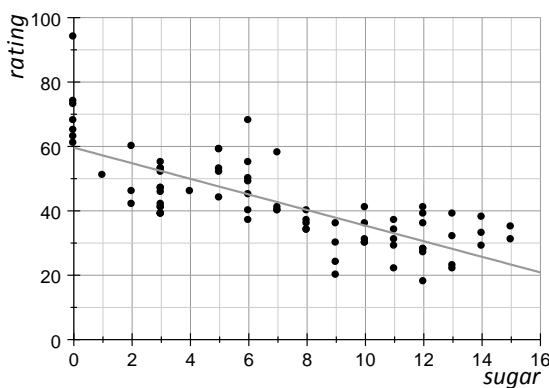
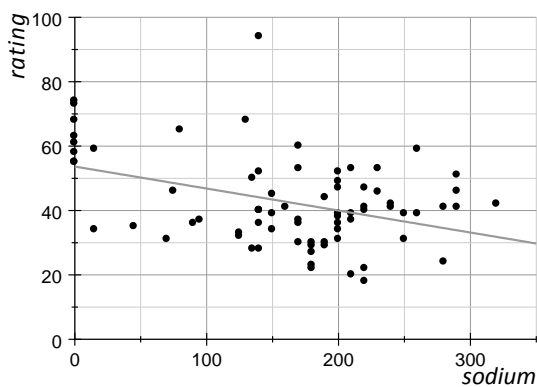
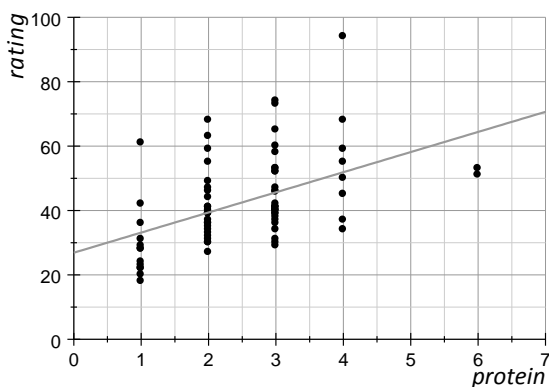
Using Explained Variation to Measure Fit

STUDENT NAME _____

DATE _____

TAKE IT HOME

The four scatterplots, below, give the relationships between the Consumer Reports ratings and four different ingredients in 76 different breakfast cereals. The ingredients are: Protein, Sodium (salt), Sugar, and Carbs (Carbohydrates).



- 1 One of the scatterplots above has a correlation coefficient of $r = 0.09$.
 - A Which of the scatterplots is it?
 - B Compute r^2 for this correlation coefficient.
 - C Now, looking at the scatterplot you identified in part A, interpret the value of r^2 in the context of that data.

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

- 2 One of the four scatterplots above has a correlation coefficient of $r = 0.47$.
- A Which one of the scatterplots is it?

 - B The cereal with the highest rating in the scatterplot you identified in part A has $y = 94$. What is the corresponding value of x ?

 - C The line of best fit for the data set for this scatterplot is: $\hat{y} = 6.3x + 26.9$. Compute the value of \hat{y} for the cereal with a rating of 94.

 - D Is the residual for the cereal with a rating of $y = 94$ positive or negative?

 - E Calculate the residual for this point.

 - F What proportion of the total variability of y coordinates can be explained by x through the LSR line for the data in this scatterplot?

Lesson 3.8 (OPTIONAL)

Using Explained Variation to Measure Fit

- 3 One of the four scatterplots above has a correlation coefficient of $r = -0.75$.
- A Which scatterplot is it?

 - B What is the value of x for the cereal with the lowest rating of $y = 18$ in the scatterplot you identified in part A?

 - C By looking at the point with a rating of 18 in this scatterplot, is the residual positive or negative for this value?

 - D The line of best fit for the data in this scatterplot is: $\hat{y} = -2.4x + 59.6$. What is the residual for the point with a rating of 18?

 - E Compute and interpret the *coefficient of determination*.
- 4 One of the four scatterplots above has a correlation coefficient of $r = -0.40$.
- A Look at the point representing the cereal with the lowest rating in this scatterplot. Is the residual positive or negative?

 - B Estimate this residual.

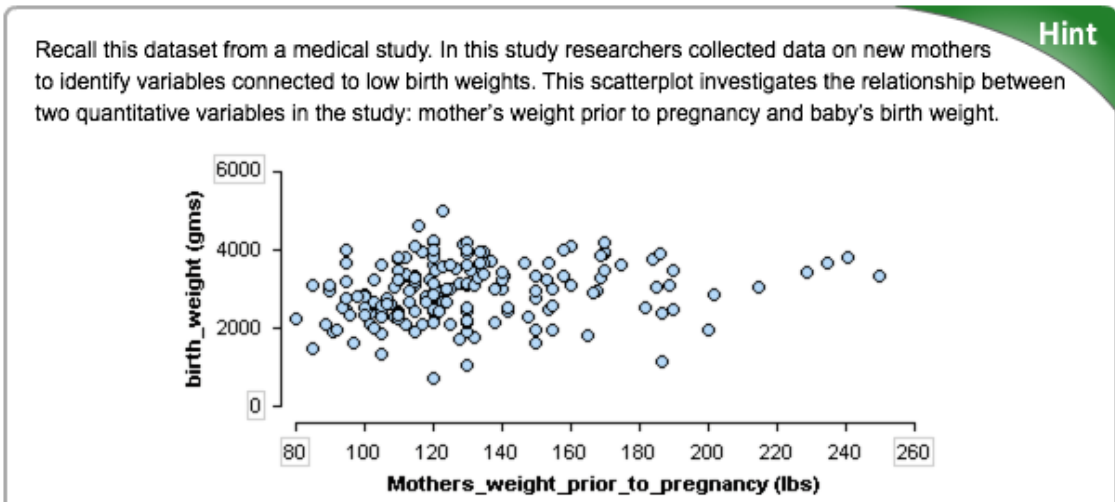
 - C Is the residual for the cereal with the highest rating in this scatterplot positive or negative?

 - D Estimate this residual.

 - E What proportion of total variability can be explained by the LSR line?

 - F What proportion of total variability is unexplained by the LSR line?

Chapter 3 Review

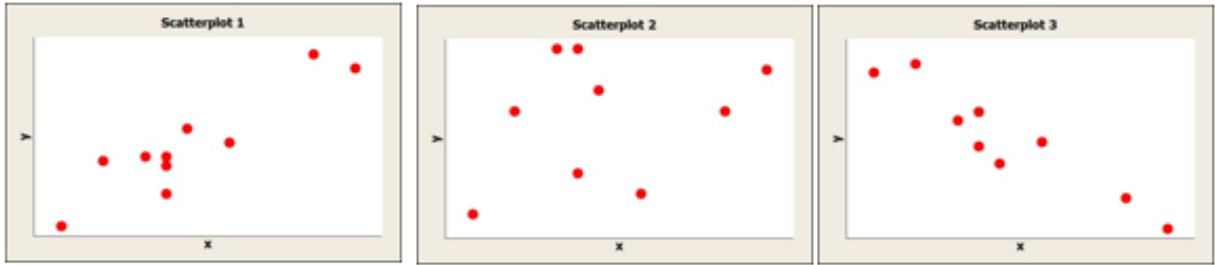


1. Each dot represents
a) a new mother b) a mother's weight c) a baby's birth weight
2. What is the explanatory variable?
a) mother's weight prior to pregnancy b) baby's birth weight
3. What is the approximate pre-pregnancy weight of mother with the heaviest baby?
a) 80 b) 124 c) 250

Each research question below describes a relationship between two quantitative variables. Which variable is the explanatory variable and therefore be plotted on the x – axis?

4. Is the sales price of a townhouse in San Francisco related to the number of square feet in the townhouse?
 sales price square footage either
5. For women aged 25-35 what is the relationship between their annual salary and the number of years of education? salary years of education either

6. Match each description of a set of measurements (A, B, and C) to a scatterplot.



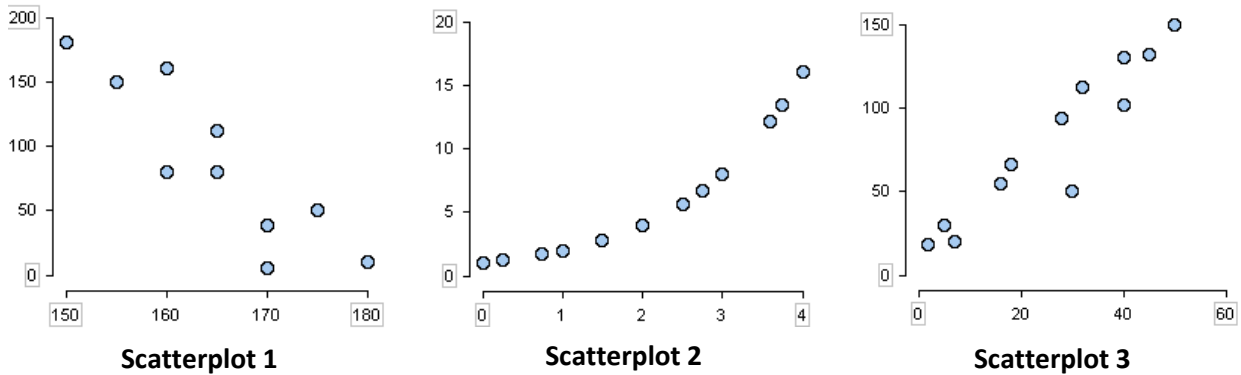
- A x = average outdoor temperature and y = heating costs for a residence for 10 winter days
- B x = height (inches) and y = shoe size for 10 adults
- C x = height (inches) and y = score on an intelligence test for 10 teenagers

7. Match the r -values with the plots above.

- A $r = 0.2$
- B $r = -0.8$
- C $r = 0.7$

Paulos also writes a column for ABCNews.com called *Who's Counting?* In his February 1, 2001, column, Paulos discusses the idea that correlation does not imply causation. He points out that the consumption of hot chocolate is negatively correlated with crime rate. Obviously, drinking more hot chocolate does not lower the crime rate.

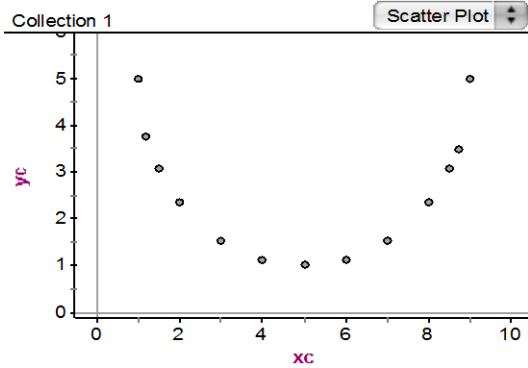
8. According to this scenario, which of the scatterplots below would match this situation?



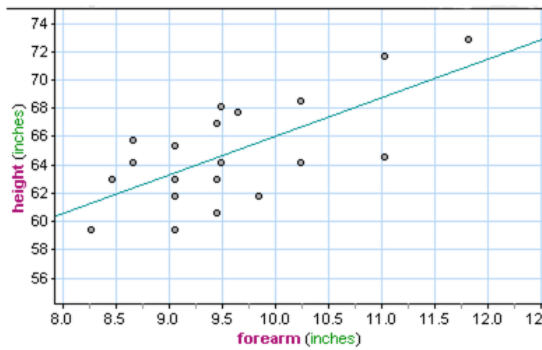
9. Identify a reasonable lurking variable in this scenario.

- a) There is no lurking variable. This makes perfect sense: the more hot chocolate consumed, the less crime committed.
- b) Drinking hot chocolate makes people happy.
- c) People consume more hot chocolate in the colder months. Crime rates are lower in the colder months.

10. Would you expect the r – value for this scatterplot to be closer to 1, 0, or -1? Explain.



11. The scatterplot below shows the forearm length and the height of 21 female college students.



The least squares regression equation is $\hat{y} = 39 + 2.7x$.

- A) Use the equation to predict the height of a female college student who has a forearm that is 11 inches. Show what you calculate.
- B) Plot the point on the scatterplot that corresponds to your calculation. Mark it with the letter A.
- C) What is the y -intercept in the equation?
- D) Does it have a meaning in the context of this problem? Explain.
- E) Which would be an extrapolation? Predicting height based on a forearm length of
 - A) 10 inches
 - B) 11 inches
 - C) 13 inches

12.

For a statistics project a student gathered data on monthly car insurance premiums paid by students and faculty at a California community college in 2008. She found a fairly strong linear relationship between driving experience and monthly car insurance premiums.

Here is the equation of the line:

$$\text{Predicted monthly car insurance premium} = 97 - 1.45 * \text{years of driving experience.}$$

- A) Use the equation of the line to predict the monthly car insurance premium for 12 years of driving experience.

- B) Identify the slope in this problem and write a sentence explaining what it means in the context of this problem.

13. The following data show the miles traveled and the standard nonpeak (reduced) fare amount needed for travel from the Metro Center station stop to nine other Metro stations.¹

Station	Miles	Fare
Pentagon	2.98	\$1.95
Virginia Square-GMU	4.47	\$2.40
Congress Heights	4.66	\$2.45
Medical Center	8.15	\$3.40
Branch Ave	9.02	\$3.65
West Falls Church-VT/UVA	9.29	\$3.70
New Carrollton	10.23	\$3.95
Greenbelt	11.49	\$4.30
Shady Grove	17.44	\$5.00

- 1. Enter this data into your calculator and report the LSR line and the r – value.
- 2. Calculate the predicted fare for Branch Ave.
- 3. Calculate the residual for Branch Ave.
- 4. The residual for Shady Grove is negative. This means that the predicted fare is an (over, under) estimate for the actual fare a rider pays.

Chapter 4

Investigating Patterns in Bivariate Data

Lesson 4.1

Investigating Patterns in Data

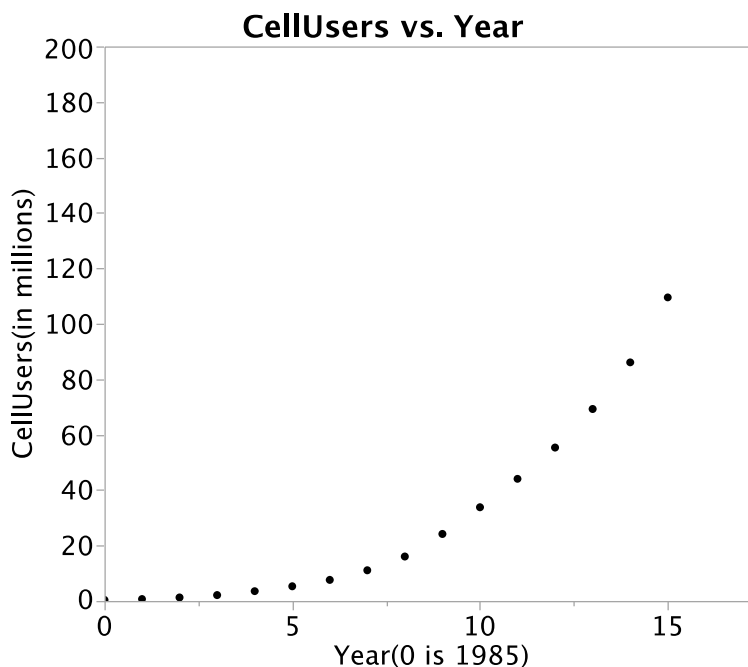
INTRODUCTION

Despite the fact that the first cell phones, introduced to the general public in 1983, had a talk time of half an hour and took 10 hours to recharge, the use of cell phones in the US grew very quickly. In fact, you might have heard the phrase “it has grown exponentially”. Or, maybe you have heard the expression, “That YouTube video went *viral*.” These two phrases actually mean the same thing, mathematically. Let’s take a look at some data and see what that looks like graphically. In the next lesson, we will see what it means to grow (or decay) exponentially.

The data in the following table and graph give the number of cell phone subscribers in the US from the year 1985 to 2000. <http://www.infoplease.com/ipa/A0933563.html>

- 1 You will notice in the graph that on the horizontal scale, we graphed the years after 1985, not the year numbers directly. For example in 1996, there were 44.043 million subscribers so instead of graphing the point (1996, 44.043) we graphed the point (11, 44.043) since 1996 is 11 years after 1985. Therefore, we are using the year 1985 as our “zero” year. Fill in the missing values in the table for the years since 1985.

Year	Years since 1985	Subscribers in millions
1985		0.34
1986	1	0.682
1987	2	1.231
1988	3	2.069
1989	4	3.509
1990		5.283
1991	6	7.557
1992	7	11.033
1993	8	16.009
1994	9	24.134
1995	10	33.759
1996	11	44.043
1997	12	55.312
1998		69.209
1999	14	86.047
2000	15	109.478



Lesson 4.1

Investigating Patterns in Data

- 2 In Chapter 3, you learned to describe data in a scatter plot by discussing the direction, form and strength of the data. Do that for this data set.
 - A Direction
 - B Form
 - C Strength

- 3
 - A One of the points on the graph has coordinates (7, 11.033). Find this point on the graph above, mark it with an X and then label the point with the (x, y) coordinates.
 - B Now, write a sentence explaining what this point means in terms of cell phone subscribers.
 - C During which year did the number of cell phone subscribers reach 50 million?

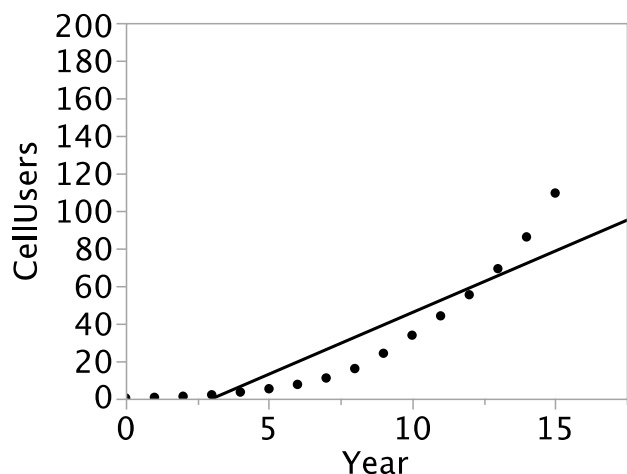
- 4
 - A Use the table to calculate the increase in the number of subscribers between the year 1990 and 1995.
 - B Do this again, for the years 1995 to 2000.
 - C Considering your answers to A and B, did the number of subscribers grow at a constant (the same) rate over the years shown in the graph?
 - D As we learned in Chapter 3, when we use a line to model data, the slope of the linear regression equation represents the rate of change of y and that rate is constant. That is, the slope between any two points on the line is the same. Therefore, should we use a linear regression equation to fit this data? Why?

- 5 Using the graph, what do you predict the number of cell phone users to be in the year 2002?

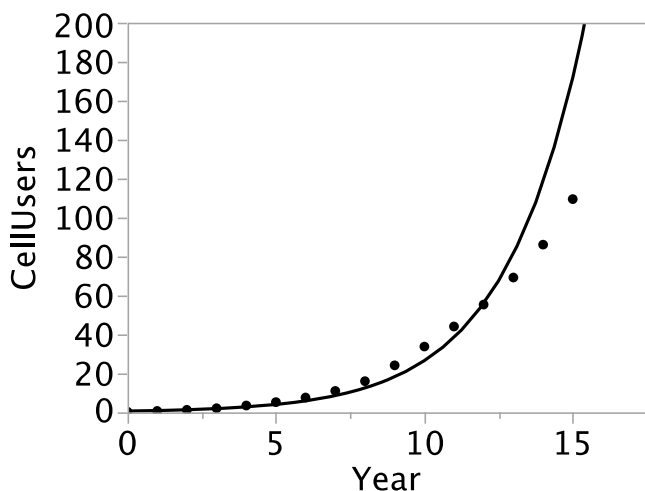
Lesson 4.1

Investigating Patterns in Data

We have established that the data is not linear. To further reinforce this idea, look at the figure on the right. It shows the data with along with the least squares regression line. We can see that the line does not match the pattern of the data. Also, note that the data is first all above the line, then below the line and then all above the line again. While we usually expect data not to fall exactly on the linear regression line, we don't want to see this pattern of scatter about the line. The data should appear to be "randomly" scattered above and below the line.



In this figure, we have fit the data with another kind of regression model called *exponential regression*. The curve fits very well at the beginning and does pretty good in the middle, but doesn't do very well as we get into the next century. Even so, it does fit better than the linear model.



- 6 A Using the graph of the exponential regression curve (not the actual data points), estimate the number of cell phone subscribers in the year 1995.

Lesson 4.1

Investigating Patterns in Data

7 The equation of the exponential regression equation for this data is $\hat{y} = 0.631(1.453)^x$, where \hat{y} is the estimated number of cell phone subscribers in millions and x is the years since 1985. Notice that it looks a lot different than the linear equation $\hat{y} = ax + b$ we are familiar with. In fact, the general form for an exponential equation is $y = a \cdot b^x$. We'll take a closer look at the meaning of a and b in this equation in the next lessons. We will also learn how to find this equation using the calculator.

A Use the equation given to find what this model predicts for the number of cell phone users in 1995. Remember that x is not the year 1995 it is the years *since* 1985. Also, when you use your calculator use the \wedge key to raise to the power.

B Compare the value in the table given for 1995, the value you estimated using the graph in question 6, and the value you found from the equation in part A.

Table _____ Graphical Estimate _____ Equation _____

C Which of these is the actual data value?

D Which of these estimates, the graphical estimate or the estimate from the equation, is the most accurate estimate of this value?

8 A Recall the *Residual = Data value – Predicted value*. Use the values you found in the previous question to calculate the residual for 1995.

B Is the predicted value using the exponential regression equation an over or under estimate for the actual number of cell phone users in 1995?

Lesson 4.1

Investigating Patterns in Data

9 A Using the exponential equation, $\hat{y} = 0.631(1.453)^x$, predict the number of subscribers in 1985.

B Do you see this value in the equation?

C Recall we can find the y - intercept by finding the y value when $x = 0$. Therefore, since we get this value by letting $x = 0$ in the equation, this is the y-intercept.

The coordinates of this point are $(0, \underline{\hspace{2cm}})$.

Mark this point on the graph given above problem 6.

10 This is no coincidence. The number in front of the parentheses, in this case 0.631, is the y-intercept. That will ALWAYS be true. Can you explain why? Think about what happens to any value when you raise it to the 0 power. We will look into this more later.

11 Fill in the blanks below to write a sentence in context describing what this number means. Use this as a model when you are asked to write sentences about this special number.

In the initial year, which was _____, the number of _____ was _____ million, according to the exponential regression equation.

Next Steps

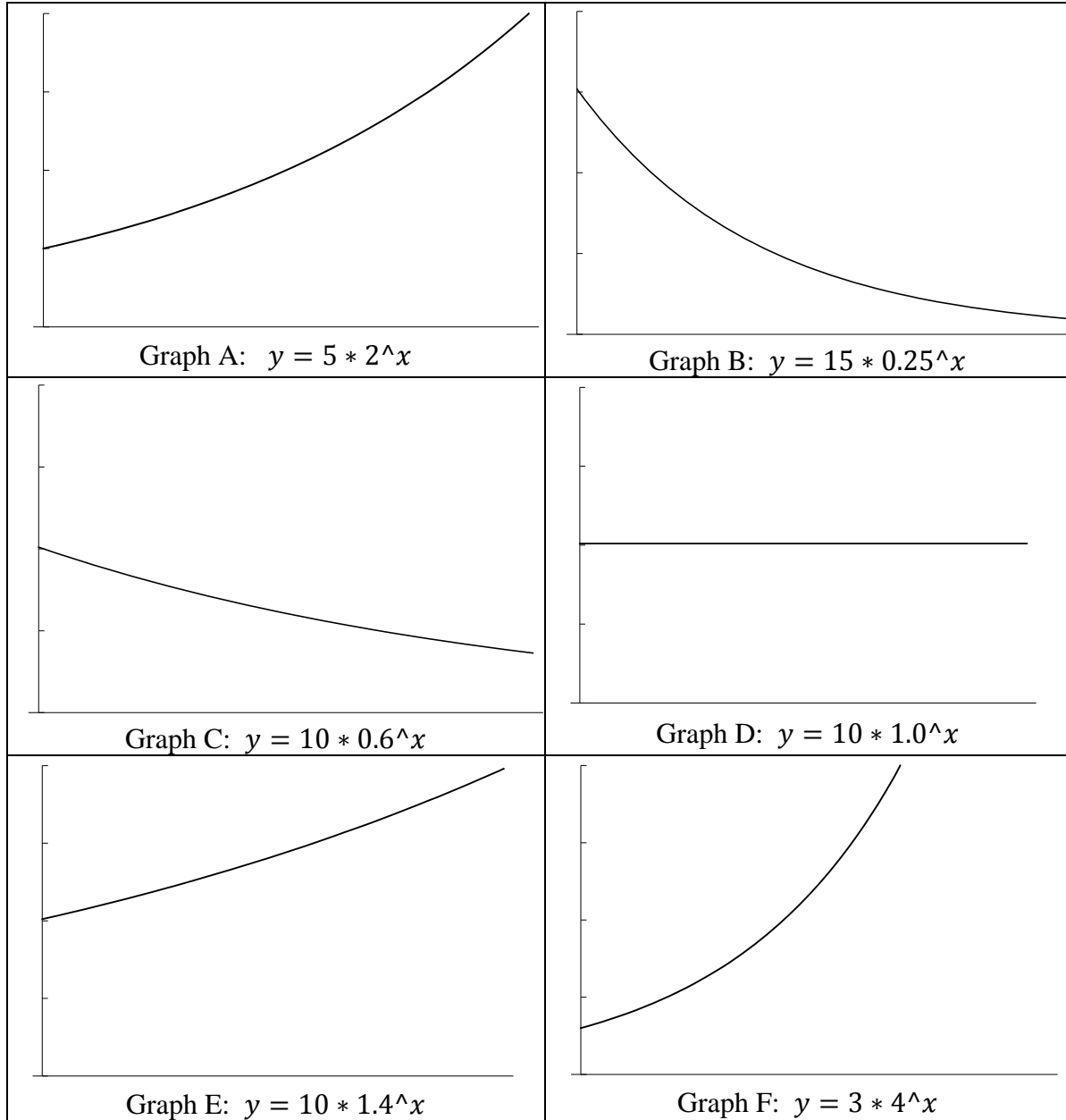
The general form of the exponential equation is $y = a \cdot b^x$. Recall in the linear regression equation, $\hat{y} = ax + b$, a and b each affected the graph in different ways. This is true in the exponential equation as well. But before we look closer at the a and b in the exponential equation, let's take a moment and recall what the a and b represent in the linear equation.

12 In the linear regression equation, $\hat{y} = ax + b$, describe what the a and b tell you about the graph of the line.

Lesson 4.1

Investigating Patterns in Data

Now we're ready to dive into the exponential equation, $y = a \cdot b^x$. Examine the 6 graphs below. All are exponential graphs of the form $y = a \cdot b^x$



13 A In which graph does y increase?

B What are the values of b for those graphs?

Lesson 4.1

Investigating Patterns in Data

- C In which graphs does y decrease?

 - D What are the values of b for these graphs?

 - E What is special about the graph when $b = 1$?

 - F What do you think is the relationship between b in the equation and the graph?
- 14 Now take a look back at the six graphs and the y -intercepts of all the graphs. What is the relationship between a in the equation and the graph?
- 15 What is the value of x at the y -intercept of any graph?
- 16 How do you get the value of y at the y -intercept using the equation? Plug that value into the general equation $y = a \cdot b^x$.

Lesson 4.1

Investigating Patterns in Data

YOU NEED TO KNOW

An exponential model has the general form: $\hat{y} = a \cdot b^x$

The numbers a and b in the exponential model have the following key properties:

- a is the **y-intercept** of the model. It is the y-value when $x = 0$. It is also the **initial value**.
 - If $b > 1$, the graph increases and b is called the **growth factor**, and the model represents **exponential growth**.
 - If $b < 1$, the graph decreases and b is the **decay factor**, and the model represents **exponential decay**.
-

TRY THESE

- 17 The concentration of a drug injected into the bloodstream can be modeled by the exponential equation $\hat{y} = 1.2(0.8)^x$ where x is the time since the injection in hours and y is the concentration of the drug in the bloodstream measured in milligrams per milliliter.
- A What is the value of b in this model?
- B Based on your answer to part A, is this exponential growth or exponential decay?
- C Explain why this makes sense in terms of drugs in the bloodstream.
- D What is the value of a in this model? Is this number measured in hours or concentration of drug in the bloodstream measured in milligrams per milliliter?
- E What is the value of x , that corresponds to the value you found in part D?
- F Looking back at the sentence you wrote in question 11, write a sentence describing what the value of a means in the context of this problem.

Lesson 4.1

Investigating Patterns in Data

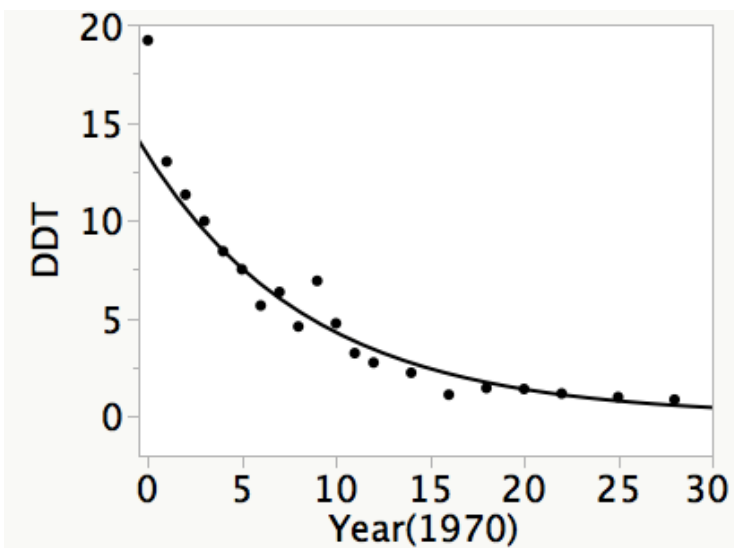
STUDENT NAME _____

DATE _____

TAKE IT HOME

- The chemical pesticide DDT was widely used in the US and other countries in the 1940's as an agricultural insecticide as well as to control diseases carried by insects such as malaria and typhus. It was eventually banned in the US in 1972 because of its danger to both humans and wildlife, however, its effects can still be found in the environment. The data shown below in the table and in the scatterplot gives the concentration of DDT in trout from Lake Michigan. The data was collected over the years from 1970 to 1998. The concentration is measured in parts per million, or ppm. Use the table, graph and equation shown below to answer the following questions.

Year	Year since 1970	DDT Conc in ppm
1970		19.19
1971	1	13
1972	2	11.31
1973	3	9.96
1974	4	8.42
1975	5	7.5
1976	6	5.65
1977	7	6.34
1978		4.58
1979	9	6.91
1980	10	4.74
1981	11	3.22
1982	12	2.74
1984	14	2.22
1986	16	1.1
1988	18	1.44
1990	20	1.39
1992		1.16
1995	25	0.98
1998	28	0.85



$$\hat{y} = 13.3(0.893)^x$$

- Fill in the missing values in the “Years Since 1970” column in the table above.
- Looking at the data in the scatterplot, should we use a linear model?

Lesson 4.1

Investigating Patterns in Data

- C Using the data in the table, calculate the difference in the amount of DDT found in the trout over the five - year period from 1972 to 1977.

 - D Using the data in the table, calculate the difference in the amount of DDT found in the trout over the five - year period from 1990 to 1995.

 - E Considering your answers to C and D above, does the data change at a constant rate?

 - F Explain why your answer to E further justifies why we should not use a line to fit the data.
- 2
- A This problem also refers to the data given above. What is the actual concentration of DDT in the trout in 1970?

 - B What does the exponential regression equation predict for the year 1970?

 - C Calculate the residual for 1970.

 - D Are there any other years that have larger residuals? If so, list them. If you can't tell, explain why.

 - E If 0.448 parts per million is considered the "safe" level to eat trout from Lake Michigan, using the equation, use trial and error to estimate the year in which it would be safe to have a trout dinner.

Lesson 4.1

Investigating Patterns in Data

- 3 A This problem also refers to the data given above. What is the value of b in the exponential equation given?
- B Fill in the blanks using the words below the statement:
"Because _____ is _____ than the number _____, I know the data represents exponential _____."
- a b 0 1 growth decay
- C What is the value of a in the exponential equation given?
- D Write the xy - coordinates that correspond to this value and then mark that point on the graph above and label that point with those xy – coordinates.
- (0, _____)
- E Write a sentence in the context of the problem describing what the value of a means in the context of the problem.

Lesson 4.1

Investigating Patterns in Data

- 4 David Sifry's, on his website <http://www.sifry.com>, keeps track of the "Blogosphere". Among other data, he has tracked the number of blogs worldwide since January 2003.
- A Would you expect this to be linear data? That is, do you think the number of blogs on the internet is growing at a constant rate? Explain your answer.
- B If you fit this data using exponential regression, would you expect b to be bigger or smaller than 1?
- C Using the data given on the website, we can model this data with the exponential regression equation $\hat{y} = 0.21(5.6)^x$, where y is measure in millions of blogs and x is the number of years since January 2003. According to this model, how many blogs were there in January 2003?
- D According to this model, how many blogs were there in June 2004? Hint: Since you need to write June 2004 as years since Jan 2003, x will be a decimal.
- E The equation given above was generated using data gathered over the 1.5 years from January 2003 to June 2004. Use the equation to predict the number of blogs in April 2007 ($x = 4.3$)
- F The actual number of blogs in April 2007 was reported to be 72 million. Considering your answer from part E, why do we have such a big error in the prediction?

Lesson 4.2

Exponential Models

INTRODUCTION

In lesson 4.1 we examined exponential curves. In that lesson, the exponential curves modeled cell phone subscriptions over time. We also used exponential curves to model the amount of DDT in trout from Lake Michigan. We learned that there are two forms of exponential curves – exponential decay and exponential growth. The amount of DDT found in trout, since it was banned in 1972, decreased exponentially so the model was an example of *exponential decay*. Cell phone subscriptions grew exponentially so that was an example of *exponential growth*. In this lesson, we will learn how to find the exponential regression equation using the TI graphing calculator and we will discover what makes some quantities grow exponentially while others grow linearly.

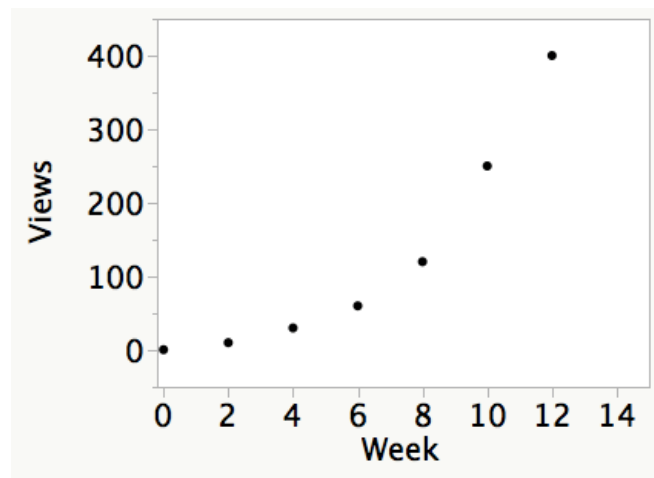
Viral Videos

You have probably all heard of or perhaps viewed a YouTube video that has gone “viral”. A video that has gone *viral* is a video that becomes extremely popular through the process of people sharing the video on websites, Facebook, email, Twitter, etc. One example of such a video is the “Gangnam Style” music video by the South Korean musician Psy, <http://www.youtube.com/watch?v=CH1XGdu-hzQ>. The music video went viral in August 2012. By December 21, 2012 it had been viewed over 1 billion times and, as of this writing (June 1, 2014), it has been watched just over 2 billion times and is YouTube’s most watched video.

Let’s take a look at the data for the number of views on YouTube over the first several weeks of it being posted. The video was first posted on July 15, 2012. We will treat that as our initial data value. The data in the table below shows the number of YouTube views, in millions, by week starting from July 15, 2012.

connect.icrossing.co.uk/gangnam-style

Date	Week Number	Views in Millions
07/15/2012	0	0.5
07/29/2012	2	10
08/12/2012	4	30
08/26/2012	6	60
09/09/2012	8	120
09/23/2012	10	250
10/07/2012	12	400



Lesson 4.2

Exponential Models

Looking at the scatter plot, the data is clearly not linear, and, considering how the video was spread, we wouldn't expect it to be. We'll have more on this idea later. For now, let's enter the data into the calculator and find the exponential regression equation. The process is the same as what we did to find a linear regression equation, except we will choose exponential regression instead of linear regression in the final step.

- Enter the week number into List 1 and the number of views in List 2. Then press the Stat button, go over to the Calc tab, scroll down and select ExpReg for Exponential regression. Press enter and you should see the exponential equation, $y = a * b^x$, with the values for a and b below it. Report the values for a and b here. Round each to three decimal places.

$a =$ _____

$b =$ _____

- Now use those values to write the exponential equation for this data.

$$\hat{y} = \text{_____} (\text{_____})^x$$

- Enter this equation into your calculator and graph the equation. Use the same viewing window as the scatterplot shown above.

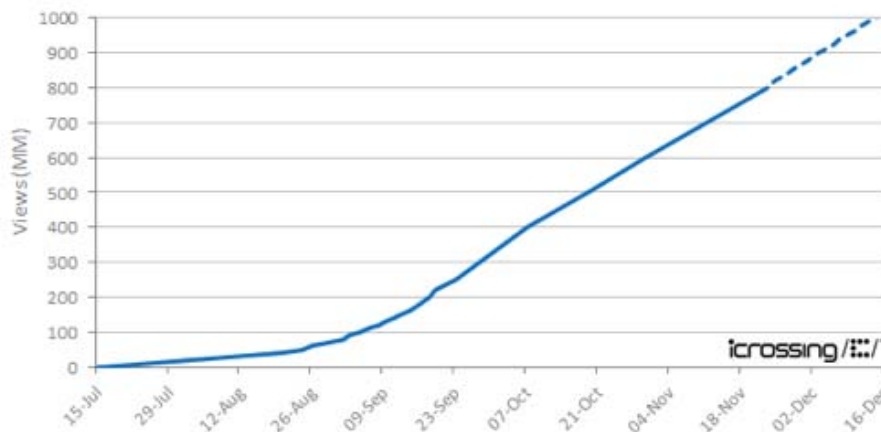
- Using the equation you found, $\hat{y} = 1.901 (1.645)^x$, answer the following questions.

- How many views does the model predict for July 15th?
- How many views does this model predict for December 24, 2012? On this date, there were 23 weeks since July 15th. Remember this is in millions of views.
- December 24, 2012 is the approximate date that YouTube reports the video actually had 1 billion views. One billion views is the same as 1000 million views. How accurate was the exponential regression model? Report the difference in the predicted value and the actual value.
- What went wrong?

Lesson 4.2

Exponential Models

Actually, if you look at the data past the middle of October, it becomes quite linear. Beware of Extrapolation, especially in exponential data! You can run into big problems because exponential equations grow very, very quickly so moving outside the data range can produce big errors quickly.



NEXT STEPS

So far we have learned that not all data is linear. We have looked at data that grows exponentially and decays exponentially. We know how to enter the data into the calculator and find the exponential regression equation, $\hat{y} = a * b^x$. Furthermore, we know a is the y-intercept, or the initial value. Also, we know that b controls whether the graph goes up or down.

What we would like to do next is to look deeper into the parameters a and b . How does b relate to how fast the graph grows or decays? Also, what is the difference between linear growth and exponential growth? That is, why do some things grow linearly, and others grow exponentially?

Let's consider a specific example in order to gain some insight into these questions. Note that the following examples do not involve fitting an equation to data as we have done in the previous examples. In the following examples, you will be asked to write the exponential equation that describes a certain situation given in words, but no data points are given.

Suppose Camilla has a rich Aunt Sofia. When Camilla was born her Aunt Sofia put \$5 in a piggy bank for her. On each birthday, Aunt Sofia doubles the amount of money in Camilla's piggy bank. Camilla does not take money out of the piggy bank nor does she (or anyone else) put any money into the piggy bank.

- 3 How much money will Camilla have in the piggy bank when she enters kindergarten at age 5?

Lesson 4.2

Exponential Models

- 4 Let's look at the calculations you did to answer the question above by putting the information in a table. Do you see a pattern? Continue filling out the table for Camilla's 4th and 5th birthday.

Camilla's Age	Calculations	Total \$ in Piggy Bank
0 (Born)	5	5
1	$5 \cdot 2$	10
2	$5 \cdot 2 \cdot 2 = 5 \cdot 2^2$	20
3	$5 \cdot 2 \cdot 2 \cdot 2 = 5 \cdot 2^3$	40
4		
5		

- 5 When Camilla turns 18, will she have enough money to go to any college of her choice such as: Harvard with annual tuition of \$39,000, MIT with annual tuition of \$43,000 or Princeton University with annual tuition of \$53,000?
- 6 If Camilla doesn't use the money for college and is thinking about an early retirement at age 30 instead, how much money will she have to retire with?

The formula you have been using to calculate how much money Camilla has in her account is $A = 5 \cdot 2^x$. This is the same form we have seen in the exponential regression equation. The 2 in the formula is called the *growth factor*, or in algebra, this is called the *base*. It is a 2 because the amount of money *doubled* every year. The *growth factor* is similar to the *slope* in linear equations in that it controls how fast the output variable grows(or decays).

The 5 in the formula is the y- intercept, or the initial y- value.

Lesson 4.2

Exponential Models

NEXT STEPS

- 7 Imagine the number of people infected with a certain disease triples every month and in April 10 people had the disease. Let D be the number of people with the disease and t be the number of months since April. So April is our initial month.
- A What is the growth factor? What is the initial number of people with the disease?
- B Write the exponential formula that models this situation.
- C Use your exponential formula to calculate the number of people with the disease in December, 8 months later.
- 8 The number of transistors on a computer chip is a measure of computing power. The number of transistors has grown by a factor of 1.54 every year since 1984 when there were 68 transistors. Let N be the number of transistors on a chip and t be the number of years since 1984.
- A What is the growth factor? What is the initial number of transistors?
- B Write the exponential formula that models this situation.
- C Use your calculator to draw a graph of this equation. Set your graphing window to go from $x = 0$ to $x = 18$ and $y = 0$ to $y = 70000$. Set x scale to 5 and y scale to 10000

Lesson 4.2

Exponential Models

- D Use your exponential formula to calculate the number of transistors on a chip in the year 2000.
- 9 Suppose that a patient in the hospital is given a drug for pain relief. He is given an initial injection of 100 milligrams. Each hour the amount of drug in the blood is $\frac{3}{4}$ of what it was the previous hour. Let A be the milligrams of drug in the blood and t be the number of hours since the injection.
- A Is the amount of drug in the blood growing or decaying?
- B What is the decay factor? What is the initial amount of drug in the blood?
- C Write the exponential formula that models this situation.
- D Use your calculator to draw a graph of this equation. Set your graphing window to go from $x = 0$ to $x = 10$ and $y = 0$ to $y = 110$. Set x scale to 1 and y scale to 10.
- E Use your exponential formula to calculate the amount of drug in the blood 4 hours after the injection.

Lesson 4.2

Exponential Models

NEXT STEPS

Now that we know how to write exponential formulas to model exponential behavior, let's do some comparisons with linear growth situations.

- 10 Consider the two types of growth: linear and exponential. Here are two equations. $y = 3x + 5$ and $y = 5 \cdot 3^x$. Which one is the linear equation and which is the exponential?
- 11 For the linear equation, what is the slope? What is the y -intercept?
- 12 For the exponential equation, what is the growth factor? What is the y -intercept?
- 13 Finish filling out the table below, comparing linear and exponential growth, using these two equations.

x	$y = 3x + 5$	$y = 5 \cdot 3^x$
0	5	5
1	8	15
2	11	45
3		
4		
5		

- 14 One thing you should notice is that the exponential equation is growing much faster than the linear equation. But another thing we should note is, in each case, how do you get from one line to the next line? Fill in the blanks below:

In the linear equation we _____ 3 to get to the next line and in the exponential equation we _____ by 3 to get to the next line.

Lesson 4.2

Exponential Models

- 15 Imagine a certain company's sales could be modeled by the exponential equation $S = 5(1.2)^x$, where S is the company's sales in millions of dollars and x is in years since 2000. Suppose their sales in a certain year were about 15 million dollars. How much were their sales in the *next* year?
- 16 How would your answer to the above question change if the company's sales were modeled instead by the linear equation $S = 1.2x + 5$? That is, suppose their sales in a certain year were about 15 million dollars. How much were their sales in the *next* year?

The next examples will give you some practice in distinguishing between situations that grow exponentially and those that grow linearly.

17 Situation A:

Imagine that I tell a rumor I heard to 2 people and the next hour each of those 2 people tell another 2 people and in the next hour all of those people each tell another 2 people and so on and so on....

Situation B:

Imagine that I tell a rumor I heard to 2 people but those 2 people don't tell ANYONE else the rumor. The next hour I tell 2 more people the rumor, but again, everyone is keeping the rumor to themselves and so and so on.....

Which of these situations describe linear growth and which describes exponential growth?

18 Situation A:

Imagine that initially 5 people eat hot dogs from a hot dog stand on the corner. They each come down with food poisoning from eating the hot dogs, and each hour, 4 more people eat a hot dog and come down with food poisoning and so on and so on.

Situation B:

Imagine that at some initial time, 5 people have an extremely infectious disease. Each hour they come into contact with 4 people and spread the disease to those 4 people. In turn, each of those 4 people infect another 4 and so on and so on.

Which of these situations describe linear growth and which describes exponential growth?

Lesson 4.2

Exponential Models

- 19 Let's wrap this lesson up by reflecting back on our opening example of the "viral video". Explain why this is an example of exponential growth instead of linear growth. Give some specific examples of how the video might spread, similar to an infectious or "viral" disease.
- 20 Can you give another example of something else that grows exponentially?

SUMMARY

- We have learned how to find the exponential regression equation using your calculator.
- We have learned how to write exponential formulas given a starting value and a growth (or decay) factor.
- We have seen how very quickly exponential equations can grow.
- We have compared linear and exponential growth and learned that linear growth is additive while exponential growth is multiplicative.
- We have learned what makes some situations grow linearly and some grow exponentially.

Lesson 4.2

Exponential Models

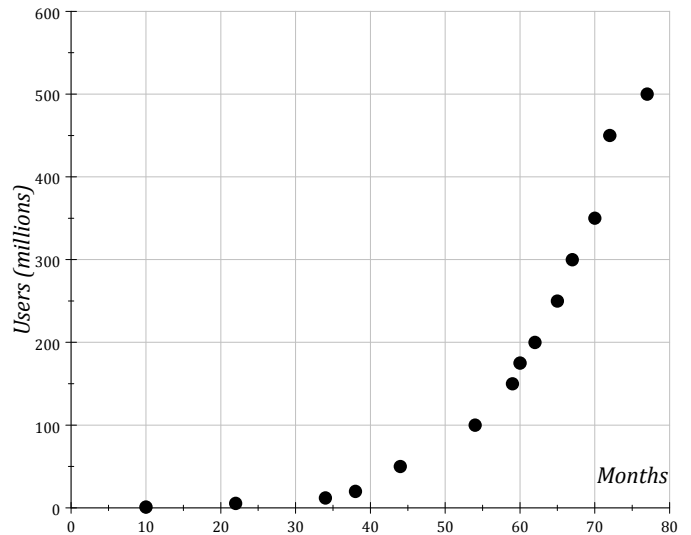
STUDENT NAME _____ DATE _____

TAKE IT HOME

- Facebook, started in February 2004, is an online social networking community. From 2004 to 2010, its membership grew rapidly as its popularity increased. As a measure of how fast it is growing, Facebook tracks the number of members. It uses information on its users to create new business opportunities.

Below are membership counts for some months between December 2004 and July 2010. The membership counts represent the number of people who are Facebook users. The numbers are given in millions, so 5.5 means 5,500,000.

Month-Yr	x (Months since Feb-04)	Facebook users (millions)
Dec-04	10	1
Dec-05	22	5.5
Dec-06	34	12
Apr-07	38	20
Oct-07	44	50
Aug-08	54	100
Jan-09	59	150
Feb-09	60	175
Apr-09	62	200
Jul-09	65	250
Sep-09	67	300
Dec-09	70	350
Feb-10	72	450
Jul-10	77	500



- A Identify which is the explanatory variable and which is the response variable for this data.

Months since Feb-04 _____

Number of Facebook users _____

- B Do you expect b in the exponential regression equation, $y = a * b^x$, to be bigger or smaller than 1? Why?

Lesson 4.2

Exponential Models

- C Enter the data into the calculator to find the exponential regression equation. Enter the months since Feb-04 (column 2 in the table above) in List 1 and the number of Facebook users in millions (column 2) in List 2.

$$a = \underline{\hspace{2cm}}$$

$$b = \underline{\hspace{2cm}}$$

$$\hat{y} = \underline{\hspace{2cm}}$$

- D When did Facebook's membership actually exceed (grow larger than) 200 million users?
- E The exponential regression equation you should have found to model this data is $\hat{y} = 0.6(1.1)^x$. Write a sentence describing what the 0.6 means in the context of the problem.
- F Use this equation to predict the Facebook membership in September 2007 (month 43).
- G Use the exponential regression equation to predict the Facebook membership in January 2014 (month 129).
- H Which of these two predictions do you think is more reliable or are they equally reliable and why?
- I Using the exponential regression equation, $\hat{y} = 0.6(1.1)^x$, if you knew the Facebook membership in a certain month was 175 million, according to the regression equation, approximately how many Facebook memberships would there be in the next month? You can assume we are not extrapolating.

Lesson 4.2

Exponential Models

- 2 The population of Americans who are age 65 and older is projected to increase rapidly over the next few decades. According to the 2010 Census, there were approximately 39 million Americans age 65 and older in 2010. Statisticians have projected that this population will increase to 89 million in 2050. Assuming that this population grows exponentially, the number of Americans age 65 years or older (in millions) can be modeled by

$$\hat{y} = 39(1.021)^x$$

where x is the number of years after 2010.

- A Write a sentence explaining what the 39 in the exponential regression equation means in the context of the problem.
- B What is the predicted number of Americans age 65 years or older in 2012? Round your answer to two decimal places.
- C What is the predicted number of Americans age 65 years or older in 2030? Round your answer to two decimal places.
- D What is the growth factor?
- 3 A research biologist is growing bacteria in a Petri dish in his lab. He has found that every hour the number of bacteria in the dish triples. The more bacteria there are in the dish, the more bacteria are produced. There are 200 bacteria in the Petri dish initially. Write a formula to model this situation. Let N be the number of bacteria in the dish and t be the number of hours since the bacteria were put in the dish.

Lesson 4.2

Exponential Models

- 4 The number of people in a small rural town has been decreasing such that after every year, starting 2011, there were only $\frac{5}{6}$ of the number of people who were in the town the previous year. Assuming there were 20000 people in the town in 2011, write a formula to model this situation. Let P be the number of people, in thousands, in the town and t be the number of years since 2011.
- 5 For the following situations, decide if you should use a linear or exponential model.

Situation A

Harry has a job selling used cars. Typically he sells 3 used cars a week. Let C be the total number of cars he has sold and w is the number of weeks he has been working.

Linear or Exponential? Justify your answer.

Situation B

Debbie receives an email from one of her friends with directions to read it and pass it on to 5 of your friends – or, you will be cursed! Let F be the number of people who have received the email after w weeks.

Linear or Exponential? Justify your answer.

Summary

An exponential model has the general form: $\hat{y} = a \cdot b^x$

The numbers ***a*** and ***b*** in the exponential model have the following key properties:

- ***a*** is the **y-intercept** of the model. It is the y-value when $x = 0$. It is also the **initial value**.
 - If $b > 1$, the graph increases and ***b*** is called the **growth factor**, and the model represents **exponential growth**.
 - If $b < 1$, the graph decreases and ***b*** is the **decay factor**, and the model represents **exponential decay**.
-

- We have learned how to find the exponential regression equation using your calculator.
 - We have learned how to write exponential formulas given a starting value and a growth (or decay) factor.
 - We have seen how very quickly exponential equations can grow.
 - We have compared linear and exponential growth and learned that linear growth is **additive** while exponential growth is **multiplicative**.
 - We have learned what makes some situations grow linearly and some grow exponentially.
-

1. Which statement about the general exponential equation $y = 600(1.05)^t$ is **FALSE**? (Assume t is time in years, with $t = 0$ in 1950.)
 - A After 1950, each year the y-value is 1.05 times greater than the previous year.
 - B The initial amount of 600 is increasing at a rate of 1.05% each year after 1950.
 - C When $t = 1$, y is 105% of its original value, 600.
 - D The initial amount of 600 is increasing at a rate of 5% each year after 1950.

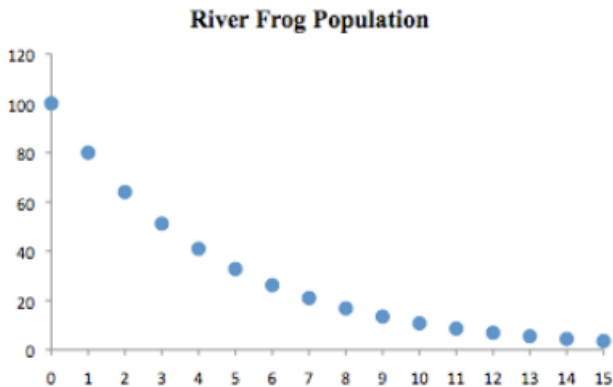
2. Which statement about the general exponential equation $y = 600(0.85)^t$ is *FALSE*?

- A The initial amount of 600 is decaying at a rate of 15%.
- B The initial amount of 600 has a decay factor of 0.85.
- C When $t = 1$, y is 85% of its original value, 600.
- D The initial amount of 600 is decaying at a rate of 85%.

3. The data below shows the population in Orange County, Florida for the years 1900 through 2000. Find the exponential regression equation and use it to predict Orange County's population for the year 2020.

Year since 1900	US Census for Orange County, Florida
0	11,374
10	19,107
20	19,890
30	49,737
40	70,074
50	114,950
60	263,540
70	344,311
80	471,016
90	677,491
100	1,145,956

4. *River Frogs*: Use the information and graph below to answer the question. A non-native species of snake appeared in a large southern swamp in 1995. Shortly thereafter, scientists noticed that a particular species of river frog began to decline exponentially. They suspected that the snakes were eating the frogs at an alarming rate. The scientists made an exponential model to predict the decline in the frog population. The points plotted below come from their exponential model. Note that t is measured in years, the value $t = 0$ corresponds to 1995, and y is the predicted number of remaining frogs in thousands.



Which of the following values could represent the size of the frog population for the year 2005, as shown in the graph above?

- A 32,800
- B 3,500
- C 10,700
- D 100,000

Which of the following formulas could represent the exponential decay shown in the graph above?

- A $y = 100 (0.80)^t$
- B $y = 80 (0.80)^t$
- C $y = 100 (1.20)^t$
- D $y = 100 - 0.80^t$

What was the first year that the frog population fell below 60,000?

A 1995

B 1996

C 1998

D 2003

Chapter 5

Types of Statistical Studies and Producing Data

Lesson 5.1

Research Questions and Types of Statistical Studies

INTRODUCTION

In Lesson 1.1, we studied the four step process used in many statistical investigations. 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.

We will start by introducing some new vocabulary. A **population**, in a statistical study, is a set of all people or objects that share certain characteristics. A **sample** is a subset of the population used in a statistical 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 a statistical study, we usually ask one of the following types of questions:

- research questions about a population
- research questions about the cause and effect relationship between two variables

Here are some examples of each kind.

Research Questions about a Population

Type of Research Question	Example
Make an estimate about the population (often about an average or proportion)	What is the average amount of sleep community college students get at night? What percent of community college students have jobs?
Testing a claim about the population (often a claim about an average or proportion)	Is the average amount of sleep for community college students more than 7 hours? Do more than half of community college students plan to vote in November?
Compare 2 populations (often comparing averages or proportions)	Do freshmen university students have higher average GPA's than freshmen community college students? Are community colleges students more likely to receive financial aid than university students?
Investigate a relationship between 2 variables	Is there a relationship between the number of hours a full time student works at a job and their GPA? Are students who drink diet soda more likely to be overweight?

Lesson 5.1

Research Questions and Types of Statistical Studies

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 college class improve test grades?
- Does caffeine reduce the risk of dementia (memory loss associated with old age)?
- Does taking aspirin daily reduce the risk of heart attacks in adults over 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 As you read the scenario below, think about the following statistical ideas: (1) the population of interest, (2) the variables being studied, (3) and the type of research question that is being asked.

A group of researchers studied women who visit a fertility clinic. The researchers wondered if less than half of women who visit the clinic would want to choose the gender of their future child.

They 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. Five hundred sixty one 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 a statistical analysis of the data. In the randomly chosen sample of 561 women from the fertility clinic 229, or $229/561 \approx 41\%$, said they would want to choose the sex of a future child. Based on this randomly chosen sample of 561, the researchers concluded that there is convincing evidence that *less than half (50%)* of **all** women who visit the fertility clinic would choose the sex of their child. They drew this conclusion since it would be unusual to see a sample of 41%, if in reality, more than 50% would like to choose the gender.

- 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?

Lesson 5.1

Research Questions and Types of Statistical Studies

- C What is the population of interest in this study?
 - D What variable is being examined for each subject in the study?
- 2 As you read the problem, think about the following statistical ideas: (1) the population of interest, (2) the variables being studied, (3) and 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. 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?

Research Questions and Types of Statistical Studies

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 investigation. 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.

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. A sample is representative if it contains individuals who are similar to the individuals in the entire 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 did not influence their choice or try to manipulate it. They were *observing* the women’s responses.

Language Tip - *representative*

For example, if we were interested in the average height of all males, we would not use an NBA basketball team as a *representative* sample of all males.

Experiment

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

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

Language Tip

The word *manipulate* means the experimenter assigns subjects participating in the experiment to do certain things.

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, we manipulate the values of an **explanatory variable**, and then observe the corresponding values of a **response variable**.

Lesson 5.1

Research Questions and Types of Statistical Studies

TRY THESE

Read the two statistical studies below and answer the following questions.

- 3 Imagine that our college is having financial problems. The college 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 \$20 fee to the college to keep the library open nights and weekends.

We are interested in learning about the proportion of students who would pay a \$20 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 \$20 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 Suppose that we collect data by asking 100 students who are entering the library whether they would pay the fee. Would this sample be representative of all students on campus? Why or why not?
- D Suppose that we collect data by asking 100 students who we find hanging out in the school gym if they would pay the fee. Would this sample be representative of all students on campus? Why or why not?

Lesson 5.1

Research Questions and Types of Statistical Studies

- E It is important to obtain a sample of 100 students that is **representative** of students at the college. In parts C and D, you read about some ways to select 100 students. Now think of a better way to select 100 students than the ideas in parts C and D. Why do you think your way is better? When you answer the 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 college students. We want to see if there is a difference between:
- The resting heart rate of college students that jog for 30 minutes three times a week for six weeks, and
 - The resting heart rate of college students that jog for 15 minutes three times a week for six weeks.

To investigate this question, we will use 100 college 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?

Lesson 5.1

Research Questions and Types of Statistical Studies

- D Imagine that we create the two groups for this study according to age. We group the 50 youngest volunteers in the 30 minute jogging group. We group the 50 oldest volunteers in the 15 minute jogging group. Is this a good idea? Why or why not?

- E Imagine that we create the two groups for this study according to weight. We group the 50 volunteers that weigh the most into the 30 minute jogging group. We group the 50 volunteers that weigh the least into the 15 minute jogging group. Is this a good idea? Why or why not?

- F 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 a better way to divide the 100 volunteers into two groups? Why would your way be better than to divide the volunteers by age (like in part “d”) and weight (like in part “e”)?

Research Questions and Types of Statistical Studies

INTRODUCTION

Drawing Conclusions from Statistical Studies

The fourth step in the statistical process is **drawing a conclusion**. 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 best way to choose a sample that is representative of the population is to choose a **random sample** from the entire 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 manipulation that was done to the explanatory variable.

The best way to ensure there are no pre-existing differences between the experimental groups for different experimental conditions is to use **random assignment** to the experimental groups.

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 randomly chosen from the population of interest.
Cause-and-effect	Experiment is conducted and subjects are randomly assigned to the experimental groups.

Notice that both kinds of research questions use randomization in their design. When we generalize from a sample to a population, we can only generalize to the population if we **choose a random sample** to represent the whole population.

However, when we draw a cause-and-effect conclusion, as in an experiment, the participants in the experiment are often *not* randomly chosen: they are volunteers. The randomness comes from **randomly assigning** the volunteers to different experimental groups. So we can still make cause-and effect conclusions with volunteers as long as the volunteers are **randomly assigned** 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.

Lesson 5.1

Research Questions and Types of Statistical Studies

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.

In Summary:

- For **Observational Studies**, we should NEVER make 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 of interest.
- For **Experiments**, it is possible to reach cause-and-effect conclusions if the study design uses **random assignment** to create the experimental groups, even if the subjects were volunteers.
- 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.

What does the phrase “reasonable conclusion” mean?

The following questions and the questions in the Take It Home all contain the phrase “reasonable conclusion”. It is important that you understand what this phrase means in this setting. The question is NOT asking about your opinion based on your personal experience relative to the study being described. In these questions, what the phrase “reasonable conclusion” refers to is the design of the study or experiment and whether the criteria described in the summary above is being followed. That is, considering the design of the study, is it appropriate to draw a cause- and-effect conclusion? If it is an observational study, was the sample randomly chosen? If it was an experiment, were the subjects randomly assigned?

- 5 Recall the study of women who visit a fertility clinic in Question 1. Considering the design of this study, is it reasonable to conclude that less than half of **all** women who visit the fertility clinic would choose the sex of their child? Explain why this is or is not a reasonable conclusion.
- 6 Think about the study about font style and the amount of time the reader thinks the task will take in Question 2. Is it reasonable to conclude that a font which is more difficult to read will **cause** a reader to think the task described in the instructions will take longer? Explain why this is or is not a reasonable conclusion.

Lesson 5.1

Research Questions and Types of Statistical Studies

- 7 The SAT exam is used in admissions decisions by many four-year colleges and universities. In 2006, The College Board 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. 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?
- B If this is an observational study, did the researchers use random selection? If it is an experiment, did they use random assignment?
- C Is it reasonable to conclude that writing the essay in cursive was the **cause** of the higher scores? Explain your answer? Recall what the phrase “reasonable” means. If you need to, refer back to the lesson where it describes what this phrase refers to.
- 8 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 Pearl Harbor and the U.S. entry into World War II. 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 If this is an observational study, did the researchers use random selection? If it is an experiment, did they use random assignment?
- C If this is an observational study, what is the population? If it is an experiment, what are the explanatory and response variables?
- D 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.

Lesson 5.1

Research Questions and Types of Statistical Studies

STUDENT NAME _____ DATE _____

TAKE IT HOME PART 1

- 1 In 2002 the journal *Science* reported that a study of women in Finland indicated that having sons shortened the life spans of mothers by about 34 weeks per son, but that daughters helped to lengthen the mothers' lives. The data came from church records over the years from 1640 – 1870. Is this an observational study or an experiment?

- 2 A dentist wonders if having a TV installed on the ceiling above the dentist's chair will reduce patient's anxiety during routine teeth cleanings. He asks some of his patients to volunteer to participate in this study. Those patients who volunteered with morning appointments will have their teeth cleaned in the room with the TV on the ceiling. Patients with afternoon appointments will have their teeth cleaned a room with no TV. After all appointments, he has the patients fill out a survey about their experience to measure the amount of anxiety they felt during their teeth cleaning.
 - A Is this an observational study or an experiment?

 - B What is the explanatory variable?

 - C What is the response variable?

 - D Is this a good way to assign the patients to the groups?

Lesson 5.1

Research Questions and Types of Statistical Studies

- 3 The Valencia student government would like to know if students at Valencia think the cafeteria on campus charges too much for their food. In order to answer this question they decide to set up a table outside of the cafeteria between the hours of 11 AM to 1PM on Tuesday and Thursday this week and ask students who drop by the table if they think the cafeteria's prices are too high.
 - A Is this an observational study or an experiment?
 - B What is the population of interest?
 - C Do you think this method of collecting data will produce a representative sample? Explain your answer.
- 4 Explain why it would be unethical to design an experiment that would show smoking 2 packs of cigarettes a day causes lung cancer. Unethical means going against what is socially or morally accepted as right.
- 5 It is a common belief that a full moon causes people to behave strangely and, as a result, police stations and emergency rooms are busier than usual. Imagine how you might go about investigating this belief. Would you use an experiment or an observational study? Why?

Lesson 5.1

Research Questions and Types of Statistical Studies

STUDENT NAME _____ DATE _____

TAKE IT HOME PART 2

- 1 The April 20, 2009 issue of the magazine *Sports Illustrated* reported the win-loss record for the Oklahoma City Thunder, a professional basketball team for the 2008-2009 season. This record was actually worse for home games that were sold out (3 wins and 15 losses) than for home games that were not sold out (12 wins and 11 losses).
 - A Based on the design of the study, is it reasonable to conclude that a sell-out crowd is the **cause** of the team's poor performance at sold-out home games? Can you think of another explanation for why the win-loss record might be worse for sold out games than for games that are not sold out?
 - B Did random selection or random assignment play any role in the collection of these data?

- 2 One hundred people volunteered to participate in a statistical study. The volunteers were divided into two experimental groups based on gender, with women in group 1 and men in group 2. Those in group 1 were asked to eat 6 ounces of sweet potatoes daily for 3 months. Those in group 2 were asked not to eat any sweet potatoes for 3 months. At the end of the 3 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.
 - A Is the study described an observational study or an experiment? Explain your answer.
 - B Did the study use random assignment to experimental groups? If so, explain what the study did to randomly assign students.
 - C Is the conclusion "eating sweet potatoes leads to (causes) healthier skin" reasonable given the study description? Explain your answer.

Lesson 5.1

Research Questions and Types of Statistical Studies

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 6 ounces of sweet potatoes daily for 3 months. Those in group 2 were asked not to eat any sweet potatoes for 3 months. At the end of the 3 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.

A Is the study described an observational study or an experiment? Explain your answer.

B Did the study use random assignment to experimental groups? If so, explain what the study did to randomly assign students.

C Is the conclusion “eating sweet potatoes leads to (causes) healthier skin” reasonable given the study description? Explain your answer.

4 Does exercise help with insomnia (those who can’t sleep) sufferers? Forty insomnia sufferers agreed to participate a month long study. The researchers put 40 slips of paper in a bag, 20 with the word *Exercise* and 20 with the words *No Exercise* printed on the paper. The exercise group was assigned to an exercise program for a month. After a month, the study found the exercise group was able to sleep better than the no exercise group.

A Is the study described an observational study or an experiment? Explain your answer.

B Did the study use random assignment to experimental groups? If so, explain what the study did to randomly assign students.

C Is the conclusion “exercise leads to (causes) better sleep among insomnia sufferers” reasonable given the study description? Explain your answer.

Lesson 5.2
Random Sampling

INTRODUCTION

- 1 We would like to estimate the average area of all 100 rectangles on the page. To find the exact average area of all 100 rectangles we could find the areas of each of the 100 rectangles and then add all 100 areas up and divide by 100 to find the average. This would be a very tedious job! Can you think of another idea that would give us an *estimate* of the average area of all the rectangles?

- 2 One way to estimate the average area of all the rectangles would be to choose a sample. We will choose a sample of 5 rectangles and calculate the average area of those 5 rectangles. In this problem, what is the **population**? What is the **sample**?

NEXT STEPS

- 3 Choose 5 rectangles from the paper. Calculate the areas of each of the rectangles and record the areas below:

Rectangle Number					
Area					

- 4 Find the average area of the 5 rectangles you recorded.

- 5 Recall our goal was to estimate the average of all the rectangles on the page. So at this point, what is your best guess as to the average of all the rectangles on the page?

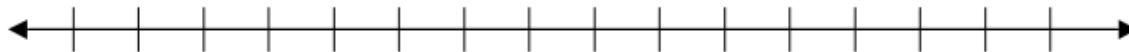
Lesson 5.2

Random Sampling

- 6 There are likely several different answers to question 5. Why?

- 7 How might we get a better estimate of the average of all the rectangles on the page?

- 8 Make a dot plot of the class's sample averages.



- 9 Describe the shape.

- 10 Approximate the center of the distribution.

- 11 Use your calculator and the STAT menu to find the average of all the samples.

Lesson 5.2

Random Sampling

NEXT STEPS

In a truly random sample, each subject in the population has the same chance of being chosen. So let's use the 10-sided die to choose a random sample **of size 5**.

- 12 Use the 10-sided die to choose a random sample of 5 rectangles. Calculate the areas of each of the rectangles and record the areas below:

Rectangle Number					
Area					

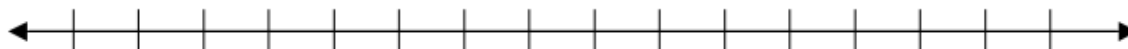
Language Tips

Often, we refer to a “random sample of size n .” Here, n means the number of subjects or individuals in the sample and is called the **sample size**.

- 13 Find the average area of the 5 rectangles you recorded.
- 14 Recall the average area of the population of rectangles is 7.42. How close is your sample average to the average of all the rectangles?

Let's combine the class's data to see if we did better estimating the average of all the rectangles with our random sample.

- 15 Make a dot plot of the class's sample averages using random sampling. Use the same scale as you did on the other dot plot.



Lesson 5.2

Random Sampling

- 16 Compare the shape of this dot plot with the first one you made.

- 17 Approximate the center of the distribution.

- 18 Use your calculator and the STAT menu to find the average of all the random samples.

- 19 So which method, the “visual” sampling method or the random sampling method, produced the best estimate of the area of all the rectangles?

NEXT STEPS

When we sample, our goal is for every population member to have the same chance of being selected.

As we discovered, one good way to do this 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.

- 20 **Biased samples** result when sampling methods tend to leave out certain types of population members.

Suppose our college 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 \$100. The college wants to know the percentage of students who would support this fee. Here are some possible sampling methods:

- 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 college website. A poll is an opinion survey. Use the students who answer the question as the sample.

- C Survey students as they enter the Student Center.

Tell whether each method would produce a sample that was **representative** of the entire student population. If you think it will not, explain how the sample would be different.

Lesson 5.2

Random Sampling

21 The three sampling methods listed above would all produce biased samples. However, 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 from (20) above is a voluntary response sample?

- 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 from (20) above is the best example of a convenience sample?

22 Suppose our college has 13,000 students. The college has the names and email addresses for all students in its database. Suggest a way that the administration could choose a simple random sample of 150 students to survey about the parking fee proposal. After the administration has chosen a sample, how could they actually conduct the survey?

23 When a researcher does not have a list of population members, it can be difficult or even impossible to get a simple random sample. In such cases researchers must still try to get a representative sample.

Suppose that a writer for the student newspaper wants to survey students about the parking fee proposal. The writer does not have access to the college’s student database. What is a method that the writer might use to get a sample that is representative of the student population?

Lesson 5.2

Random Sampling

NEXT STEPS

Making the leap from a small sample to the entire population is impossible without understanding Statistics. We need to be able to stretch beyond the data at hand to the world at large. To make that stretch, we need three big ideas.

24 Imagine you are having friends over for dinner and you are wondering how the vegetable soup you are cooking is going to taste to your company.

A How can you determine whether the soup meets your standards?

B You may have come up with a few different ideas for part A. One of the simplest ideas is to taste the soup, but how much would you have to taste in order to know if the soup was any good?

It should be enough for you to simply taste a spoonful or two to decide whether the soup meets your standards. You certainly would not need to consume the whole pot of soup! You trust that the taste will *represent* the flavor of the entire pot. The idea behind your tasting is that a small sample, if selected properly can represent the entire population.

BIG IDEA #1: Examine a Part of the Whole – This is an idea you have already been introduced to, selecting a sample to represent the population. However, as we have begun to learn in this lesson, this can be more difficult than it sounds. We have to be careful when selecting our sampling method in order to reduce the chances of selecting a biased sample.

C Suppose you decide the soup is a little bland and you needed to add some salt to the pot. What will happen if after adding the salt you sample from the top of the pot without stirring it first? Or, if you sample from the bottom of the pot before you stir?

D Suppose a friend is helping you and while you weren't looking, they added some peas to the soup. If you skim a little soup off the top of the pot without stirring, will you even know the peas are in there?

If you sample from the pot without stirring you will get a misleading idea about the whole pot of soup. By stirring the soup, you *randomize* the amount of salt throughout the pot and move the peas throughout the pot even though you didn't know they were there. This will give you "a little bit of everything" in your taste of the soup, making each taste more typical of the whole pot.

Lesson 5.2

Random Sampling

BIG IDEA #2: Randomize – We can't "stir" people, but we can select them at random. Randomizing protects us from the influences of *all* the features of our population by making sure that, on average, the sample looks like the rest of the population.

E A friend that tasted your soup at dinner really liked it and asked you to make the soup for a large banquet that she was organizing. Of course in making the soup for the banquet, you need to use a much larger pot! As you prepare to "sample" your soup to see if it is ready to be served, how much will you need to taste?

Even though you have a much bigger pot than before, you do not need a bigger spoon to decide how the soup tastes. The same-size spoonful is probably enough to make a decision about the soup, no matter how large the pot.

BIG IDEA #3: It's the Sample Size – The *fraction* of the population that you have sampled doesn't matter. It is the size of the sample itself that is important. A small drop of soup probably isn't enough for you know what the soup tastes like, but once you have found a spoon size that gives you a taste typical of the pot, you do not need to increase the spoon size simply because you are sampling from a bigger pot.

SUMMARY

- We want our sample to be **representative** of the population so that the results from the sample are accurate.
- **Randomness** in choosing the sample is the key to making sure the sample is representative of the population.
- For simple random sampling, precision depends on sample size but not population size (as long as the population size is large).

Lesson 5.2

Random Sampling

STUDENT NAME _____ DATE _____

TAKE IT HOME

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

How would you select a simple random sample of 75 students from this population? Describe your process.

- 2 You want to estimate the average amount of time, per week, that students at a particular college spend studying. Which of the following sampling methods do you think would be best? For each method, explain why you did or did not select it as the recommended method.
 - Method A: Select 50 students at random from the students at the college.
 - Method B: Select 100 students as they enter the library.
 - Method C: Select 200 students at random from the students at the college.
 - Method D: Select the 300 students enrolled in English literature at the college this semester.

Lesson 5.2

Random Sampling

- 3 In California, a new initiative was proposed. The initiative wants to have people vote on whether the option “none of the above” should be added to the list of candidates running for a political office. This would mean that voters in California could pick “none of the above” when they did not want to vote for any of the candidates. More than half of voters in California would have to vote “yes” for the initiative to become law.

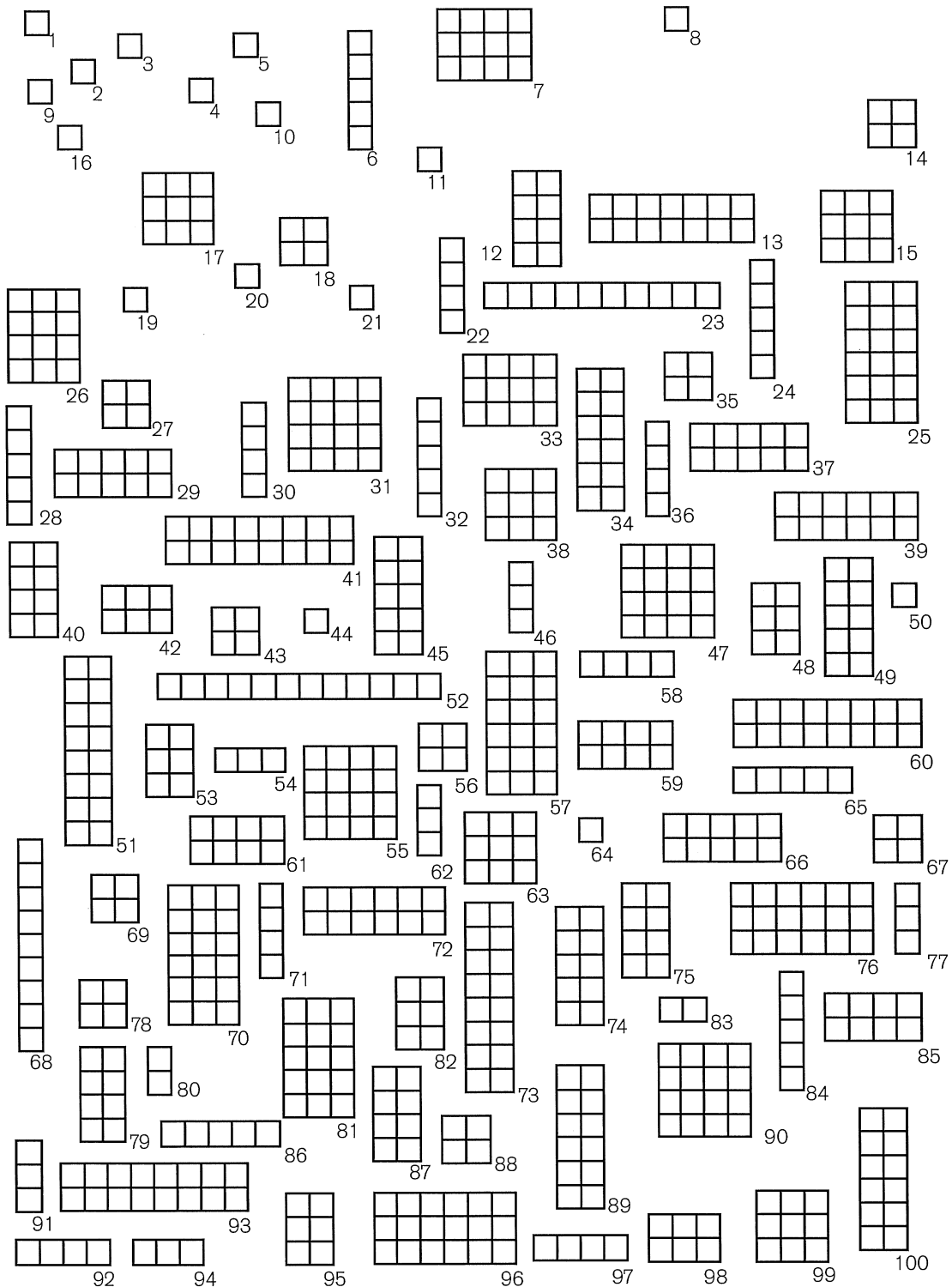
A newspaper conducted a poll to see if more than half of people would vote “yes.” The results of the poll were:

- 55% were against the initiative and would vote “no”
- 45% were for the initiative and would vote “yes.”

A spokesperson for the initiative questioned this result. He said a random sample of 1,000 registered voters was not large enough because there are about 23.5 million registered voters in California. The spokesperson said that 1,000 voters could not possibly represent all voters in California. (Associated Press, January 30, 2000)

- A Is his criticism valid? Explain why or why not. Think about the type of sampling method that was used, the sample size and the population.
- B Would the criticism be valid if this had been a national initiative and 1,000 people were randomly selected from all registered voters in America?

Random Rectangles



You will explore this population of 100 rectangles in the following activities:

- Sampling Rectangles
- Sample Size
- Sampling Methods

Lesson 5.3

Collecting Data by Conducting an Experiment

INTRODUCTION

In Lesson 5.1, you learned the following:

- The primary goal of an **experiment** is to provide evidence for a cause-and-effect relationship between two variables.
- In an experiment, we manipulate the values of an **explanatory variable**, and then observe the corresponding values of a **response variable**.
- The response variable usually changes under different conditions. We want to learn whether an observed change is the result of the changes in the explanatory variable.
- To determine if the observed difference is the result of varying the explanatory variable, we must rule out **chance variation**.
- Even after chance variation is ruled out, we must think about other possible explanations for differences in the response variable. To accomplish this, the experiment must be planned carefully.

TRY THESE

Conducting an Experiment

Today we are going to conduct an experiment. In past lessons we discussed the four step process in a statistical study. The first step in this process was to *ask a question that can be answered by collecting data*. In today's experiment we ask the following question: "Does listening to music make a difference in how much students remember while studying?"

In order to prepare for our experiment we first need to write a **hypothesis**. A **hypothesis** is a conjecture – an informed guess based on some information you have about the subject. For example, an informed guess about music and studying may be based on the fact that you and your friends listen to music when you study and it seems to help you remember things better.

We *test* our hypothesis by collecting and analyzing data.

Your answers to the following three questions will help you as you write your own hypothesis.

- 1 Think about your own study habits. Do you prefer to study in silence or while listening to music? Write your answer and one reason why you prefer studying this way.

Lesson 5.3

Collecting Data by Conducting an Experiment

- 2 Imagine that you are asked to memorize a list of 20 words. You study the words for 90 seconds and try to remember as many as possible. What effect might listening to music have on how much you can remember?

- 3 Think about other students who listen to music while studying. Do you think they remember *more*, *less*, or about *the same* as students who study in silence? Why do you think this?

Your answer to question 3 is your hypothesis. This is your best guess of the answer to your research question. It is “informed” by your prior knowledge and experiences. We can test your hypothesis by gathering data.

- 4 You will now design an experiment to test your hypothesis. You must collect data and make a decision about your hypothesis. Write down what you will do to test your hypothesis.

To test your hypothesis, we will now gather data. Your instructor will provide you with a list of 20 words. Some students will study the words while listening to music and the rest of you will study in silence. You will have 90 seconds to study the words and then the instructor will take the list of words from you. You will then have 90 seconds to write down (in any order and spelling doesn't count) as many of the words as you can remember.

Lesson 5.3

Collecting Data by Conducting an Experiment

NEXT STEPS

You will now grade the papers as your instructor reads the correct words aloud. Once this is done, answer the questions below.

- 5 Who had the highest score in your group? Did that person study in silence or while listening to music?

- 6 Did the person in your group with the lowest score study in silence or while listening to music?

- 7 Think about the study method, silence or with music, of the person in your group with the highest score. Are you convinced that the method they used is the best one? Why or why not?

We should not make conclusions based on the person who got the highest score. That result represents only one person. One result is not enough evidence. This is an example of **anecdotal evidence**. An anecdote, or 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.

- 8 We want to see if there is a difference between the students who studied while listening to music and the students who studied in silence. What could we do with the memory quiz scores from all students in our class to find out?

Lesson 5.3

Collecting Data by Conducting an Experiment

TRY THESE

- 9 Compute the average score of the students who studied in silence, as well as the average score of students who listened to music. Describe any reasons for the difference between the scores of those who studied with music and those who studied in silence.

- 10 Even if the music really makes no difference, would you expect the averages for the two groups to be exactly the same?

- 11 If you don't expect them to be exactly the same, what might cause the averages not to be the same?

NEXT STEPS

When a difference 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 scores between the two groups. We do not yet have the tools to determine if the difference between quiz score averages is due to music listening or chance variation.

For now, we need to be careful. We shouldn't make 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 don't have English as their first language, they might get lower scores on the quiz.

A Bit of Terminology

Before we move on, let's take a look at some terminology used when describing experiments.

When planning an experiment, it is helpful to think of the research question in the following form:

What is the effect of *A* on *B*?

In such questions, *A* represents the **explanatory variable** and *B* represents the **response variable**. For the music listening experiment, we could say:

What is the effect of *listening to music* on *memory retention*?

Lesson 5.3

Collecting Data by Conducting an Experiment

The first part of the question identifies the *explanatory variable* which we manipulate – we change it intentionally. The explanatory variable is *reading conditions*. The reading conditions are: *music* and *silence*.

The second part of the question identifies the *response variable* that we observe, looking for differences.

For the music listening experiment, the response variable is memory retention which we measure by observing quiz scores.

As another example, we could conduct an experiment to investigate the effect of fertilizer on tomato plants. That is, does fertilizer affect the number of tomatoes a plant produces? In this experiment, the fertilizer is the explanatory variable. The response variable is the number of tomatoes the plant produces. The fertilizer is applied and we test for differences in the number of tomatoes produced by the plant.

As with the music listening experiment, we might choose to compare the number of tomatoes produced for two groups of similar tomato plants, where one group of plants gets no fertilizer and the other group of plants gets fertilized.

INTRODUCTION

Direct Control

The goal of an experiment is to determine the effect of changing a treatment on the response variable. In our experiment, the treatment we changed was the music and we wanted to see the effect of that on memorization skills. But were there other things beside the music that might have affected the quiz grade?

Therefore, we want to rule out other possible explanations for the observed 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 control those variables by making them as similar as possible across the groups.

Suppose a student listens to music that she or he *enjoys* as opposed to music he/she doesn't and can easily tune out. Could music enjoyment itself affect the ability to remember? One way to control this is to allow participants to choose the music they enjoy so everyone who is listening to music is listening to music they enjoy instead of everyone listening to the same music. Some might enjoy it, some may not, and this might affect their ability to remember. We want everything to be as similar as possible.

Lesson 5.3

Collecting Data by Conducting an Experiment

TRY THESE

Consider the following description of an experiment. In this experiment, the researchers want 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 the amount of bacteria present after using 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.¹

12 What is the explanatory variable in this experiment? How many treatments were there in this experiment?

13 List the treatments in this experiment.

14 What is the response variable in this experiment?

15 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?

Before starting a study, we can often identify variables that might affect the response. Sometimes, these variables are not easily controlled. For example, in our music listening experiment, some students might be better at remembering. That is, regardless of the music or silence, they just have better memories. It is hard to know who those students are before the experiment.

In cases like this, when we can't control for certain differences, (like memory ability in our music example) we use **random assignment**. Random assignment helps to spread the difference out over both groups. So some who have good memorization skills will be randomly assigned to the silence group and some will be in the music group. Random assignment helps us create groups that are similar.

¹ *Infectious Disease News*. (September 2010).

Lesson 5.3

Collecting Data by Conducting an Experiment

NEXT STEPS

When possible an experiment should include a control group. A **control group** is a group that does not get a treatment. Including a control group in an experiment provides a basis for making comparisons.

16 What was the control group in our music listening experiment?

17 Why was it important to include a control group in this experiment?

Sometimes it is not possible to include a control group. In Lesson 5.1 we studied an experiment in which students were given the same instructions in two different fonts. One font was easy to read and the other was difficult to read. The purpose of the experiment was to determine if the font had an effect on how long the students thought the task would take. There would have been no way to include a control group in that experiment.

Blinding

Sometimes people already have ideas about whether the treatments in an experiment will be effective. These beliefs might influence the response.

18 You are planning an experiment to compare the effects that two different doses of medication will have on the pain that patients have after knee surgery. Each patient is randomly assigned to one of two groups. One group receives a small dose of the medication after surgery and the second group receives a larger dose after surgery. Both groups of patients are asked to rate their level of pain one hour later.

Imagine that the patients know their dosage level. That is, they know that they are receiving either a small dose of pain medication or a higher dose of pain medication. Do you think this might affect how they rate their pain level after one hour? Write the reasons why you think this.

19 Think about your answer to question 18 above. What is one strategy that you can use to make sure that participants' prior beliefs do not influence their 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.

Lesson 5.3

Collecting Data by Conducting an Experiment

- 20 Would it have been possible to implement *blinding* in the music listening experiment in class? Explain your answer.

Another way blinding can be used in an experiment, is to plan the experiment so that the person who is measuring the response doesn't know which group is getting the treatment and which group isn't.

For example, a person who grades student essays on the Scholastic Aptitude Test (SAT) often needs to use her or his judgment to decide between two scores. In some cases, an essay might fall between a score of 4 and 5 on the grading scale. The grader must decide between these two scores.

Suppose you want to conduct an experiment to compare essay scores for two groups of students:

- Students who participated in an 8-hour SAT review course
- Students who participated in a 40-hour SAT review course.

Suppose a grader needs to choose between two close scores (like between a score of 4 and a score of 5). If she knows which review course the student took, she might be influenced in one direction or the other. That is, she may give a score based on her prior beliefs or opinions about the review courses that students took. This problem could be eliminated if the researchers implement blinding for the graders.

When possible, it is usually a good idea to consider blinding both the participants and the person measuring the response. We call such experiments **double blind**.

- 21 A group of students is preparing to take the Critical Reading section of the SAT exam. Researchers are interested in investigating whether the length of a review course has an effect on exam scores. Researchers randomly assign participants to one of two groups. One group attends an 8-hour review course and the second group attends a 40-hour review course. Is it possible to blind the students participating in this experiment? Explain your answer.

Lesson 5.3

Collecting Data by Conducting an Experiment

- 22 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 that 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 judges how much water damage was done to the boots.

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

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. Comparing the placebo control group to the no treatment group allows them to see whether the treatment had an effect *over and above* that of the placebo. That is, if there is a significant difference in the response variable between the control group who receives no treatment and the control group who received the placebo, then the placebo effect occurred. 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.

- 23 A researcher included a control group (with no treatment) and a placebo control group in her experiment. How does including both groups allow her to decide when a placebo effect occurs?
- 24 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?

Lesson 5.3

Collecting Data by Conducting an Experiment

SUMMARY

Term	Explanation	Example
Explanatory variable	It is what you think is responsible for some measured change.	Listening to music in the classroom experiment.
Response variable	This is the variable that you are measuring the change in due to a change in explanatory variable.	Number of words recalled in classroom experiment.
Direct Control	This is what researchers do to try to keep experimental groups as similar as possible.	In the hands drying experiment, having everyone handle the chicken for the same amount of time.
Random assignment	Researchers randomly assign subjects to groups to make groups as similar as possible, after they have directly controlled everything they can. "Control what you can and randomization will take care of those variables you can't control."	In our classroom experiment, we randomly assigned students to music or no music because we can't control who were the students with a better memory. So we try to spread them out evenly among the groups by randomly assigning them.
Treatments	These are the things that are assigned to each group.	Music and no music
Control Group	This is the group that does not get a treatment or gets a placebo so that we have something to compare our responses to.	No music group.
Placebo	This is a treatment that is known to have no effect such as a sugar pill.	Placebo is a sugar pill or a vaccine with no medicine in it.
Placebo effect	This describes the effect when a response is reported even when a placebo is given.	If a patient reports pain relief after taking a placebo.
Blinding	This means that the subject and/or the researcher (or technician) does not know what treatment they are receiving.	The group who receives the placebo, should not know they are receiving a sugar pill. Therefore they are blinded.
Double Blind	This means that both the subjects and the technician do not know what treatment is being applied.	If the technician who is asking "Is your pain reduced by taking this pill?" doesn't know the patient received a placebo and the patient doesn't know they got a placebo.

Lesson 5.3

Collecting Data by Conducting an Experiment

STUDENT NAME _____ DATE _____

TAKE IT HOME PART 1

- 1 A study done by researchers at Kings College in London found that infomania (information overload) has a temporary, negative effect on intelligence quotient (IQ). In the experiment, the researchers divided volunteers into two groups. Each subject took an IQ test. One group had to check e-mail and respond to instant messages while they were taking the test. The other group took the IQ test without any distractions.
 - A What is the response variable in this experiment?

 - B What is the explanatory variable?

 - C What are the treatment values (the individual treatments)?

 - D Imagine that distractions, such as responding to email messages, don't really have any effect on IQ test scores. Would you still expect the scores between the two groups to be exactly the same? Why or why not? Recall this idea was discussed in today's lesson. Go back and look at your answers to questions 10 and 11.

- 2 A medical researcher believes that giving surgery patients high dosages of vitamin E will help their incisions heal faster. To find out, he designs an experiment in which he will give one group 15 mg, another group 20 mg and a third group a placebo (fake pill) with no vitamin E.
 - A What is the response variable in this experiment?

 - B What is the explanatory variable?

 - C What are the treatment values (the individual treatments)?

Lesson 5.3

Collecting Data by Conducting an Experiment

- 3 Just as many older people suffer from arthritis and other joint problems, older dogs do too. Suppose a research veterinarian wants to test a joint treatment pill for dogs. He designs an experiment to test the new drug. In the experiment he randomly assigns 20 Golden Retrievers (a breed of dog) to two treatment groups. All the Golden Retrievers are 8 years old and approximately the same weight. At the end of the study he tests the dogs for signs of swelling, pain and stiffness.
- A What is the response variable in this experiment?
- B What is the explanatory variable?
- C In this experiment the vet attempted to directly control the influence of three variables that could affect the dogs' swelling, pain and stiffness. One of those variables was the breed of the dog by using dogs of the same breed. Name two more.

Lesson 5.3

Collecting Data by Conducting an Experiment

STUDENT NAME _____ DATE _____

TAKE IT HOME PART 2

- 1 A study done by researchers at Kings College in London found that infomania (information overload) has a temporary, negative effect on intelligence quotient (IQ). In the experiment, the researchers divided volunteers into two groups. Each subject took an IQ test. One group had to check e-mail 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.
 - A Explain why it would be good for the researchers to use random assignment to put each volunteer in one of the experimental groups. Why should the researchers do this rather than letting the volunteers decide which group they wanted to be in?
 - B Identify the control group in this experiment.
 - C Is it possible for the subjects to be blinded? Explain your answer.

- 2 Suppose a farmer wishes to evaluate a new fertilizer. She uses the new fertilizer on one field of crops (A), while using her current fertilizer on another field of crops (B). The irrigation (watering) system on field A has recently been repaired and provides adequate water to all of the crops, while the system on field B will not be repaired until next season. She concludes that the new fertilizer is far superior since crop A grew so much better than crop B.
 - A Is the farmer's conclusion valid?
 - B What is wrong with the design of the farmer's experiment?

Lesson 5.3

Collecting Data by Conducting an Experiment

- 3 Do eating fruits and vegetables help prevent colon cancer? These foods are rich in “antioxidants” found in vitamins like A, C and E. A medical experiment is done in which 500 physically active, white males aged 55-60 years were randomly divided into 4 groups: daily dose of vitamin A, daily dose of vitamin C, daily dose of vitamin E, and daily dose of a placebo. After two years the researchers found no significant difference in colon cancer among the 4 groups.
- A What is the response variable in this experiment?
- B What is the explanatory variable?
- C What are the treatment values (the individual treatments)?
- D In the experiment described, the researchers chose to control a number of variables that might have influenced the recommended treatment. How did they do this?
- E Is there a control group? Why is it important to have a control group?
- 4 A recent newspaper article, “Doctor Dogs Diagnose Cancer by Sniffing It Out,” described an experiment about cancer. Researchers designed an experiment that investigated 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. The article states, “The researchers blinded both the dog handlers and the experimental observers to the identity of the breath samples.”
- A Explain what was meant by the last sentence.
- B Explain why this blinding is important.

Lesson 5.3

Collecting Data by Conducting an Experiment

- 5 Researchers have concluded that viewing and discussing art can lead to lower blood pressure. This was based on an experiment in which 20 elderly women gathered once a week to discuss different works of art. The study also included a group of 20 elderly women who met once a week to have dinner and have casual conversation on various topics. At the end of four months, the women in the art group were found to have lower blood pressure than the women in the casual conversation group.
- A Why is it important to determine whether the researchers randomly assigned the women in the study to one of the two groups?
- B Why do you think the researchers included the casual conversation group in this study?
- C Imagine that the experiment also included a third group of women who did not meet in groups at all. The researchers randomly assigned participants to one of the three groups – an art group, a dinner group, and a group that didn't meet at all. Discuss if and how this improves the study.

CHAPTER 5 REVIEW

Fill in the blanks with the values:

Cause-and - effect	does	experiment	explanatory	observational study
response	does not	population	sample	random
convenience	confounding variables	Voluntary response	blinded	randomly
placebo			double blinded	

There are two types of studies in statistics.

An _____ selects individuals from a population and measures variables of interest. We conduct observational studies to investigate questions about a _____ or about an association between two variables. An observational study alone _____ provide convincing evidence of a cause-and-effect relationship.

An _____ intentionally changes one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a _____ relationship between two variables. The variable that the researcher changes is called the _____ variable and the variable that is measured to see if there was an effect is called the _____ variable.

In an observational study we draw a conclusion about the population based on a _____. To get a sample that is representative of the population we take a _____ sample.

Observational Studies:

- No cause-and-effect conclusions – ever!
- Can only generalize from the sample to the population if there was **random selection** from the population.

Experiments:

- Cause-and-effect conclusions are OK – IF the experiment uses **random assignment** to place the participants in experimental groups.
- Can only generalize from the sample to the population if there was **ALSO random selection** from the population.
- If the experiment used both random selection of participants **AND** random assignment to groups **THEN** you can draw cause and effect conclusions **AND** generalize the conclusions to the population.

Study 1:

Many students listen to music while studying. Does listening to music improve learning? Students in a Statistics class decide to investigate this question. They write more specific research questions related to the topic of music and learning. Then they design the following two studies:

Study 1

Specific research question: When we compare students who study with music to students who study in a quiet environment, which group gives higher ratings for understanding what they studied?

To investigate this question the instructor asks her students who listens to music and who doesn't when studying. Based on their answers to that question she divides the class into two groups: (1) those who listen to music when they study and (2) those who do not listen to music when they study. The students keep a journal for a week. Each time they study, they record the following information: on a scale of 1-10, a rating of how well they understood what they studied: 1 = "no understanding," 10 = "excellent understanding."

- A) Is this an observational study or experiment? Explain your answer.

- B) Can we make any cause and effect conclusions?

- C) Were the students randomly selected to participate? Can we generalize ANY conclusions to the population?

- D) Suppose that the students who listened to music rated their understanding lower. Can we conclude that listening to music while studying decreases understanding? Explain why or why not.

Study 2:

Specific research question: Does listening to music improve students' ability to quickly identify information?

To investigate this question the instructor uses word-search puzzles. She divides the class into two groups. She assigns students on one side of the room do a word puzzle for 3 minutes while listening to music on an iPod. She assigns students on the other side of the room do a word puzzle for 3 minutes without music. The instructor calculates the average number of words found by each group.

- A) Is this an observational study or experiment? Explain your answer.

- B) Was randomization used to assign the students to the groups? Can we draw any cause and effect conclusions? Explain.

- C) Imagine she randomly assigns the students to the 2 groups and the students who listened to music had a lower average number of words found. Can we conclude that listening to music while doing a word search caused these students to do worse on word-search puzzles?

Study 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 study design:

Two hundred students were selected at random from those enrolled at a large college in California. 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.

A) Is the study described an observational study or an experiment? Explain your answer.

B) If it was an observational study, did the study use random selection from some population? If it was an experiment, did the study use random assignment to experimental groups?

C) Is the conclusion “eating sweet potatoes causes healthier skin” appropriate given the study description? Explain your answer.

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

Study 4:

Evaluate the design of the following experiment. Consider the features: **use of a placebo, random assignment, use of a control group and blinding**. Identify the features that are missing or that could be used more effectively **AND** describe how those features could be improved to make the results more reliable.

Researchers are testing a new drug – Drug X – that is supposed to reduce fever and relieve aches and pains. The researchers have 150 volunteers willing to participate in the experiment who have reported that they have a fever. The researchers assign 50 to take 325 mg of the drug, another 50 to take 500 mg of the drug and the last 50 to take 660 mg of the drug. They assign volunteers to the groups by measuring their temperatures and assigning those with the 50 highest fevers to the 660 mg group, those with next 50 highest fevers to the 500 mg group, and so on. Three hours after taking the drug, the researchers compare the change in body temperature between the several treatment groups as well as record the level of aches and pains. The participants were aware of how treatments were assigned and so were the lab technicians who were administering the drugs and recording the results.

Review for Multiple Choice

The National Health and Nutrition Examination Survey (NHANES) is a large and on-going series of studies that investigate questions about the health and nutrition of adults and children in the United States. The survey is unique because it combines interviews and physical examinations.

Studies that are part of NHANES are what kind of studies?

- A. observational studies
- B. experiments

In June 2011 CBS News and the *New York Times* reported the results of a poll about problems and priorities for the U.S. In the poll 979 U.S. adults answered the question, “What do you think is the most important problem facing this country today?” From a list of options, 53% of those polled said “economy/jobs.”

This is an observational study designed to answer a question about a population. Who is the population?

- A. the 979 U.S. adults polled
- B. U.S. adults who watch CBS News or read the *New York Times*
- C. the 53% who answered “economy/jobs”
- D. all U.S. adults

When conducting a survey, it is important to use a random sample because random samples:

- A. help us make cause-and-effect connections
- B. are an example of a census
- C. systematically favor the correct response over every other response on each question
- D. avoid bias and are representative of the population

A recent NHANES study compared death rates for people with health insurance and people without health insurance. This study included anyone who completed the survey between 1986 and 1994. The study tracked these people through 2000. The study attempted to control for factors such as education, income, smoking, drinking, and obesity.

The *American Journal of Public Health* published an article about this NHANES study. The following headlines appeared in the news based on this article. Which headline is **NOT** an appropriate summary of this study:

- A. Lack of insurance to blame for almost 45,000 deaths (ABC News)
- B. 45,000 American deaths associated with lack of insurance (CNN)
- C. No Health Insurance, Higher Death Risk: 45,000 U.S. Deaths Per Year May Be Linked to Lack of Health Insurance (WebMD)

Which of the following statements is **NOT** true?

- A. We can use results from an observational study to test a claim about a population.
- B. We can use results from an observational study to establish an association between two variables.
- C. When an observational study has a large sample and follows people for many years, we can use the results to establish a cause-and-effect relationship between two variables.
- D. When many observational studies together meet specific criteria, the results can provide varying degrees of evidence for a cause-and-effect relationship between two variables. But we should be cautious in interpreting such results.
- E. A well-designed experiment is the only legitimate way to establish a cause-and-effect relationship between two variables.

Suppose that you want to estimate the proportion of students at your college that attend at least one of the college's sports events.

Which sampling plan will produce the most reliable results?

- A. Select 50 students at random from the college.
- B. Select 100 students at random from students attending one of the college's football games.
- C. Select 200 students at random from the list of student athletes.
- D. Select 300 students who purchase tickets to at least one of the college's sports events.
- E. Either (A) or (B) or (C). All are equally representative of the student population at the college because they are random samples.

Newborn Brain Damage

Researchers studied 208 infants whose brains were temporarily deprived of oxygen due to complications at birth. When researchers detected oxygen deprivation, they randomly assigned babies to either usual care or to a whole-body cooling group. The goal was to see whether reducing body temperature for three days after birth increased the rate of survival without brain damage.

What is the explanatory variable?

- A. infants whose brains are temporarily deprived of oxygen
- B. usual care or whole-body cooling
- C. survival without brain damage

Newborn Brain Damage

Researchers studied 208 infants whose brains were temporarily deprived of oxygen due to complications at birth. When researchers detected oxygen deprivation, they randomly assigned babies to either usual care or to a whole-body cooling group. The goal was to see whether reducing body temperature for three days after birth increased the rate of survival without brain damage.

What is the response variable?

- A. infants whose brains are temporarily deprived of oxygen
- B. usual care or whole-body cooling
- C. survival without brain damage

Newborn Brain Damage

Researchers studied 208 infants whose brains were temporarily deprived of oxygen due to complications at birth. When researchers detected oxygen deprivation, they randomly assigned babies to either usual care or to a whole-body cooling group. The goal was to see whether reducing body temperature for three days after birth increased the rate of survival without brain damage.

Which of the following are **NOT** used in the design of this experiment?

- A. random assignment
- B. control group
- C. double-blinding

Researchers randomly divide participants into groups. Each group takes a different amount of omega-3 fatty acid supplements daily for a month. One group receives a placebo. The researchers measure the impact on cholesterol levels in the blood.

What is the purpose random assignment in this experiment?

- A. to produce treatment groups with similar characteristics
- B. to ensure that all people with high cholesterol have an equal chance of being selected for the study
- C. to increase the accuracy of the research results and prevent skewness in the data

Chapter 6

Two-Way Tables with Intro to Probability

Lesson 6.1A

An Introduction to Two-Way Tables

INTRODUCTION

In Chapters 3 and 4 we focused on relationships between two quantitative variables. In Chapter 6, we turn our attention to relationships between two categorical variables. Categorical variables divide individuals into categories in which order doesn't matter. Examples include gender, ethnicity, and state of birth.

Suppose you want to learn whether male or female students are more likely to drink soda on a daily basis. The variables for each student are gender and whether they drink soda on a daily basis.

TRY THESE

- 1 The explanatory variable in the research question above is gender and the categories for this variable are male and female. What is the response variable and what are its categories?

Lesson 6.1 A

An Introduction to Two-Way Tables

The following table contains data that were collected from a website survey. To gather the data, researchers surveyed students about their soda drinking habits and asked the students to identify their gender.

Your job is to determine if there is a relationship between gender and soda consumption.

Table 1: Gender and Soda Consumption

Drink Soda?	Gender
No	Female
Yes	Male
Yes	Female
No	Male
No	Male
No	Female
No	Male
No	Male
No	Male
No	Male
Yes	Male
Yes	Female
Yes	Male

Drink Soda?	Gender
Yes	Male
No	Female
Yes	Male
Yes	Female
Yes	Male
Yes	Female
Yes	Male
Yes	Female
Yes	Male
Yes	Male
Yes	Male
No	Female
Yes	Male
Yes	Male

Drink Soda?	Gender
Yes	Male
Yes	Male
Yes	Female
Yes	Female
Yes	Female
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Female
Yes	Female
No	Female
Yes	Male

Drink Soda?	Gender
Yes	Female
No	Female
Yes	Female
Yes	Female
Yes	Female
No	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Female
Yes	Female
Yes	Male
Yes	Male

- Can you tell by looking at the data if there is a relationship between the two variables? Explain why you can or can't tell if there is a relationship between the variables. What information might be helpful to you?

- Try to summarize these data in a way that will make it easier for you to compare the males and females.

Lesson 6.1B

An Introduction to Two-Way Tables

Table 1: Gender and Soda Consumption

Drink Soda?	Gender
No	Female
Yes	Male
Yes	Female
No	Male
No	Male
No	Female
No	Male
No	Male
No	Male
Yes	Male
Yes	Female
Yes	Male

Drink Soda?	Gender
Yes	Male
No	Female
Yes	Male
Yes	Female
Yes	Male
Yes	Female
Yes	Male
Yes	Male
Yes	Male
Yes	Male
No	Female
Yes	Male
Yes	Male

Drink Soda?	Gender
Yes	Male
Yes	Male
Yes	Female
Yes	Female
Yes	Female
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Female
Yes	Female
No	Female
Yes	Male

Drink Soda?	Gender
Yes	Female
No	Female
Yes	Female
Yes	Female
No	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Male
Yes	Female
Yes	Female
Yes	Male
Yes	Male

Summarizing Two Variable Categorical Data with Tallies

One way to display data for two categorical variables is to use a **two-way table**.

Language Tip
Tally marks are used for counting. Each mark in a category represents a single occurrence (so ### || stands for 7 occurrences).

- Look at the data in Table 1 above. Look at the number of people in each of the four categories (drink soda- yes/no; male, female). Fill in the two-way table below by using tally marks to count the number of people in each of the four categories.

Drink Soda Daily?	Female	Male
Yes		
No		

Lesson 6.1B

An Introduction to Two-Way Tables

- 5 Now, count the tally marks for each category. Enter the number (or *frequency*) into the two-way table below. The first value is entered for you. After you have entered each frequency into the two-way table, below, compute totals for each row and column. Enter the *grand total* in the lower right cell.

Drink Soda Daily?	Female	Male	Totals
Yes	12		
No			
Totals:			

- 6 Give two ways to calculate the grand total at the bottom right.
- 7 In this problem, you are going to make up your own values for the two-way table. Fill in the two-way table below in a way that shows that the males and female are *equally likely* to drink soda on a daily basis.

Drink Soda Daily?	Female	Male	Totals
Yes			
No			
Totals	18	30	48

- 8 We are interested in whether there is a difference between daily soda consumption for males and females. Fill in the two-way table below in a way that shows that the females are *more likely* to drink soda on a daily basis than the males. Make up your own values.

Drink Soda Daily?	Female	Male	Totals
Yes			
No			
Totals	18	30	48

Lesson 6.1B

An Introduction to Two-Way Tables

- 9 Now, make up a new set of values, showing that males are more likely to drink soda. Fill in the two-way table below in a way that shows that the males are *more likely* to drink soda on a daily basis than the females.

Drink Soda Daily?	Female	Male	Totals
Yes			
No			
Totals	18	30	48

NEXT STEPS

We need to look for even better ways to summarize two variable categorical data. In the next few problems, we are going to use the results from a different survey. A larger survey about soda consumption was conducted on the Stat Crunch website. Visitors to the website were asked whether they drank soda daily. Those who said they drink soda were also asked whether they drank regular soda or diet soda. Some of the data from the survey are displayed in the two-way table below.

Table 2: Larger survey sample

Type of Soda	Gender		Total
	Male	Female	
Regular	96		168
Diet	36	58	
None		39	
Total	167		

Lesson 6.1B

An Introduction to Two-Way Tables

TRY THESE

10 Recall what we learned in Chapter 5 about survey questions and generalizations. Think about how the survey was conducted. Can we apply the survey results to a larger population? Why or why not?

11 In the survey, what is the explanatory variable and what is the response variable?

12 Fill in the missing cells in Table 2.

Type of Soda	Gender		Total
	Male	Female	
Regular	96		168
Diet	36	58	
None		39	
Total	167		

13 How many people responded to the survey?

14 How many of the females who were surveyed drink diet soda?

We want to know if gender matters in soda drinking habits. In this case, do males have different soda drinking habits from females? The data in Table 2 indicates that males and females have similar totals, but it will be easier for us to make comparisons if we use percentages.

15 What percent of the females who were surveyed drink diet soda on a regular basis?

Lesson 6.1B

An Introduction to Two-Way Tables

16 What percent of the males who were surveyed drink diet soda on a regular basis?

17 Is the proportion of those students who prefer diet soda higher for the females or males?

Since percentages are more helpful in comparisons, we often change the frequencies in a two way table into percentages. Since gender is the explanatory variable we calculate the percentages based on each gender. In this case the explanatory variable is the column variable so we only look at the column totals. We can ignore the totals for each row.

Consider the table below.

Type of Soda	Gender	
	Male	Female
Regular		
Diet	21.6%	34.3%
None		
Total		

We want to compare the distributions for the males and females. These are called **conditional distributions** because the percentages are calculated for each condition; a percentage is shown for males, and a percentage is shown for females. The percentages in the table are those we calculated above. Notice also that there is no total column on the right. This is because we are only interested in the distribution of soda drinking percentages within each gender category.

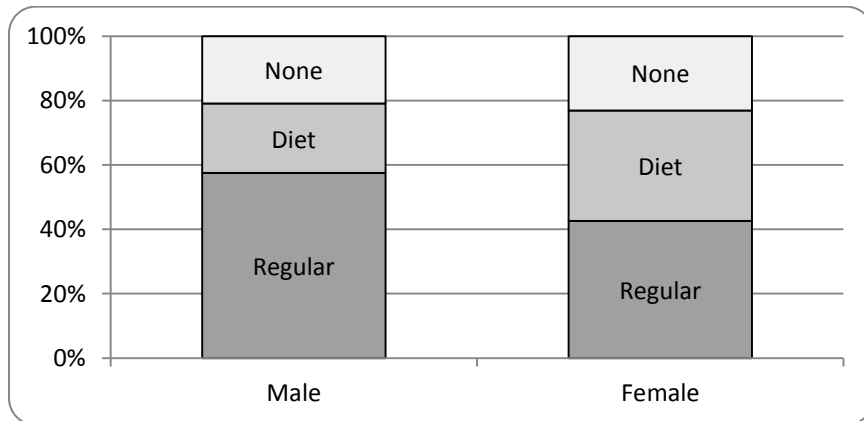
18 Calculate the remaining percentages. You need to calculate the percent of males who drink regular soda and who don't drink soda, and the same percentages for the females. The totals at the bottom are the sums of the percentages in the male column and the female column.

19 Are the column totals exactly 100%? If not, why?

Lesson 6.1B

An Introduction to Two-Way Tables

20 Using the conditional distribution you completed above and the **stacked bar graph** below, write a brief summary comparing soda consumption for the males and females and this survey. Your summary should compare regular, diet and none for the males and females.



Lesson 6.1B

An Introduction to Two-Way Tables

STUDENT NAME _____ DATE _____

TAKE IT HOME

- 1 The General Social Survey is conducted every two years by the National Opinion Research Center at the University of Chicago. They sample a large number of Americans and ask many different questions. In 2009, two of the questions they asked were “what is the highest level of education you have achieved?” and “do you agree or disagree that homosexual couples should be allowed to legally marry?” Table 3 (below) is a two-way table that summarizes sample responses (education level and whether the people that were surveyed think that homosexual couples should be allowed to get married).

Table 3: National Opinion Research Survey

Highest Level of Education	Homosexual Couples Should Be Allowed To Legally Marry		
	Agree	Disagree	Total
Less Than HS	54	115	
High School Diploma	257	344	
Community College	46		98
Bachelor’s Degree		87	
Graduate Degree	77		
Total	528		1169

- A Fill in the missing cells the table. Make sure to write the value at the top of each cell, leaving space underneath the number as you will be asked to calculate percentages later in this question.

Lesson 6.1B

An Introduction to Two-Way Tables

- B Look at the value for number of people who responded whose highest education level is a graduate degree. How many of those respondents, disagree that homosexual couples should be allowed to marry?
- C Which variable makes the most sense to use as the explanatory variable? Which variable should be used as the response variable? Why?
- D The explanatory variable is the row variable (the different values are the rows). So we **calculate the percentages based on the total for each row**. Ignore the column totals! Determine the conditional percentages for education level. Write the percentages underneath the frequencies in each box in the table.
- E Now that you have analyzed the data from Table 3, summarize what the information in this data set tells you about how a respondent's education level is related to whether the person thinks that homosexual couples should be allowed to marry.

Lesson 6.1B

An Introduction to Two-Way Tables

- 2 Many states now collect data on traffic stops regarding the race of the driver. Data from 2533 traffic stops in Cincinnati report the race of the driver and whether the traffic stop resulted in a search of the vehicle. This problem was adapted from *Stats, Modeling the World*, by Bock, Velleman, and Deveau. The two-way table below shows the data collected.

Vehicle Searched	Race			Total
	Black	White	Other	
No	787	594	27	1408
Yes	813	293	19	1125
Total	1600	887	46	2533

Analyze the data from the table to answer the question: Is race related to whether a traffic stop results in a vehicle search? Note: You will have to do steps similar to the ones you did in question 1 to answer the question posed in part E of question 1. HINT: *Race is the explanatory variable so the percentages you calculate will be based on column totals.* Show all your calculations.

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

INTRODUCTION

In the last lesson we constructed a two-way table to represent the relationship between two categorical variables. This lesson introduces three types of probabilities that can be determined from two-way tables: *marginal probabilities, joint probabilities, and conditional probabilities.*

U.S. Census Data

The United States Census Bureau collects large amounts of data on the American population. The two-way table below contains information about the population's marital status in 2010. The table gives the current marital status by gender for US citizens ages 18 and older.

Each of the numbers is in millions. So, 32.4 means 32,400,000. Each number is rounded to the nearest 100,000. Use the two-way table to answer the following questions.

		Gender		Total
		Male	Female	
Marital Status	Never Married	33.7	27.8	61.5
	Married	64.4	65.1	129.5
	Divorced	10.0	13.7	23.7
	Widowed	3.0	11.4	14.4
	Total	111.1	118.0	229.1

- 1 What proportion of American adults are male?

- 2 What proportion of American adults are currently married?

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

YOU NEED TO KNOW

We calculated the previous two proportions using only numbers in the *margins* (the grey part of the table) of the table. These proportions are called **marginal proportions**.

Up to this point, we have only been talking about proportions and percentages. But now we would like to consider how these ideas relate to *probability*.

- 3 What does the word probability mean? Can you give an example?

Now that you have an idea about what the word *probability* means, consider the next question.

- 4 The calculation you performed in Question 2 indicates that 56.5% of American adults are married. If we choose one American adult at random from this population, what is the probability that this adult is married?

The *proportion* of American adults who are married is equal to the *probability or chance* of randomly selecting an American adult who is married. This probability is an example of a **marginal probability**. Marginal probabilities are calculated using only the totals in the margins (the grey parts) of a two-way table.

A common practice is to use P to represent probability. Thus the probability that a randomly selected American adult is married can be written: **$P(\text{married})$** . This way of writing the probability is called the **probability notation**.

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

Use the two-way table to find the following marginal probabilities:

		Gender		Total
		Male	Female	
Marital Status	Never Married	33.7	27.8	61.5
	Married	64.4	65.1	129.5
	Divorced	10.0	13.7	23.7
	Widowed	3.0	11.4	14.4
	Total	111.1	118.0	229.1

- 5 If we choose one American adult at random, find $P(\text{widowed})$.
- 6 If we choose one American adult at random, find $P(\text{female})$.

YOU NEED TO KNOW

Marginal probabilities involve only one of the categorical variables. We often need to calculate probabilities that involve a combination of both categorical variables. A probability that focuses on a combination of two categorical variables is called a **joint probability**.

- 7 What is the probability that a randomly chosen adult is male and has never been married? We write this probability as $P(\text{male and never married})$.
- 8 If an American adult is chosen at random, find $P(\text{female and married})$.
- 9 If an American adult is chosen at random, find $P(\text{male and divorced})$.

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

NEXT STEPS

Marginal probabilities and joint probabilities show us the likelihood that a randomly selected American adult is (a) in a particular category, or (b) in a combination of categories. These probabilities provide valuable information. However, they don't allow us to determine if there is a *relationship* between the categories. In this case, they do not help us figure out if there is a relationship between gender and marital status.

Suppose we want to know whether the likelihood that a person is widowed depends on his or her gender. Or, does being divorced affect the likelihood that a person is male or female? To answer these questions we will find **conditional probabilities**.

- 10 Let's explore whether the gender of an American adult affects the probability that he or she is widowed (their husband or wife has died). We will first find the probability that a randomly chosen female is widowed.

Be careful about understanding what this question is asking. We start with a randomly chosen female. So our starting point is that the adult chosen is female. Once the condition (person is female) is set, we focus only on the female population. This also determines our denominator in the probability calculation. That is, the total number of females is our denominator. In terms of the two-way table, that means the only numbers we will be using are in the **female** column. If you have a highlighter, go back to the table and highlight the female column. Now answer the following questions.

		Gender		Total
		Male	Female	
Marital Status	Never Married	33.7	27.8	61.5
	Married	64.4	65.1	129.5
	Divorced	10.0	13.7	23.7
	Widowed	3.0	11.4	14.4
	Total	111.1	118.0	229.1

- A How many females are in the population?
- B How many of the females are widowed?
- C What proportion of female American adults is widowed?
- D Determine the probability that a randomly chosen female is a widow.

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

- 11 Now let's find the probability that a randomly chosen male is a widower (his wife is no longer living). Our starting point is that the adult chosen is male. The total number of males determines our denominator. So we are only looking at the male column.

		Gender		Total
		Male	Female	
Marital Status	Never Married	33.7	27.8	61.5
	Married	64.4	65.1	129.5
	Divorced	10.0	13.7	23.7
	Widowed	3.0	11.4	14.4
	Total	111.1	118.0	229.1

- A How many males are in the population?
- B How many of the males are widowed?
- C What proportion of male American adults is widowed?
- D What is the probability that a randomly chosen male is widowed?

- 12 How does gender impact the likelihood that an American adult is widowed?

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

YOU NEED TO KNOW

The probabilities you just calculated are called **conditional probabilities**. A *conditional probability* is a probability that is based on a given condition or situation. In the previous problem we found the probability that a randomly chosen adult is widowed *given* that the adult is male. This means that we assume the adult chosen is male, so we only look at the males. In this probability, the condition is that the American adult is male. Symbolically we write:

$$P(\text{widowed} \mid \text{male})$$

The vertical line replaces the word “given”. When we calculate conditional probabilities, the *denominator is always the total of the given condition*.

TRY THESE

- 13 The newspaper headline “Drinking Coffee Reduces the Risk of Dementia by 65%” summarizes the findings of a study described in the paper, “Caffeine as a Protective Factor in Dementia and Alzheimer’s Disease.”¹ (Note: Dementia is loss of or damage of mental abilities, such as memory and the ability to understand and reason, among other symptoms.)

The study followed 1,409 adults for 21 years. During that time, 61 adults developed dementia. The researchers classified the research participants into three categories based on how much coffee they drank in a typical day: Low (0 to 2 cups per day), Medium (3 to 5 cups per day) and High (6 or more cups per day).

Information in the paper was used to construct the following data table:

	Low coffee consumption	Medium coffee consumption	High coffee consumption	Total
Developed dementia	20	20	21	61
Did not develop dementia	204	622	522	1,348
Total	224	642	543	1,409

- A What is the probability that a randomly chosen adult in the study developed dementia? Is this a marginal, joint, or conditional probability?

¹Eskelinen, M. H., & Kivipelto, M. (2010). Caffeine as a protective factor in dementia and Alzheimer’s disease. *Journal of Alzheimer’s Disease*, 20, 167–174.

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

- B What is the probability that a randomly chosen adult drank coffee at the medium coffee consumption and did not develop dementia? Is this a marginal, joint, or conditional probability?
- C The number of people who developed dementia in each of the three different coffee consumption groups was very close (20, 20, and 21). Does this mean that the risk of developing dementia is about the same for each group? Explain your reasoning.
- D We want to determine how the level of coffee consumption impacted whether the subjects developed dementia. We first calculate the following conditional probabilities. Remember the condition after the “given” vertical bar tells you which column or row to look at in the table. That column or row total will be the denominator in the probability fraction you calculate.

P(developed dementia low coffee consumption)	
P(developed dementia medium coffee consumption)	
P(developed dementia high coffee consumption)	

- E Which group had the lowest rate of developing dementia?

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

SUMMARY

Definition	Example (Refer to data table given below)
Marginal probabilities focus on a single categorical variable. The numerator comes from the margin (category total) of the two-way table and the denominator is the grand total.	$P(male) = \frac{111.1}{229.1} = \frac{\text{Margin Value}}{\text{Grand Total}}$
Joint Probabilities: numerator is a cell value (within a table) and the denominator is the grand total	$P(male \text{ and } married) = \frac{64.4}{229.1} = \frac{\text{Cell Value}}{\text{Grand Total}}$
Conditional Probabilities: numerator a cell value (within a table) and the denominator is either a row total or column total	$P(married male) = \frac{64.4}{111.1} = \frac{\text{Cell Value}}{\text{Column Total}}$

		Gender		Total
		Male	Female	
Marital Status	Never Married	33.7	27.8	61.5
	Married	64.4	65.1	129.5
	Divorced	10.0	13.7	23.7
	Widowed	3.0	11.4	14.4
	Total	111.1	118.0	229.1

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

STUDENT NAME _____ DATE _____

TAKE IT HOME

- 1 Polio is a very severe illness that can cause paralysis in its victims. Many people who had polio in the early part of the 1900’s became paralyzed and were unable to move their legs or other limbs. The most famous Polio victim was Franklin D. Roosevelt, who was the US President from 1933-1945.

In 1954, researchers did a randomized experiment on the effectiveness of a vaccine to prevent Polio. The real vaccine was given to 200,745 children and a fake vaccine was given to 201,229 children. The results showed that 33 of the children given the real vaccine developed polio, while 115 of the children given the fake vaccine developed polio.

- A Was this an experiment or an observational study?
- B What was the response variable?
- C What were the treatments?
- D What was the control group?
- E Complete the two-way table below that compares vaccine type and whether or not the children developed polio.

	Type of Vaccine		
Polio	Real	Fake	Totals
Developed Polio			
Did not Develop Polio			
Totals			

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

- F Find the joint probability that a child was given the real vaccine and developed polio? Use probability notation and give the answer as a decimal. Recall from the lesson summary that a joint probability is calculated by $\frac{\text{Cell Value}}{\text{Grand Total}}$.
- G Calculate the conditional probability that a child who was given the real vaccine developed polio. Use probability notation and give the answer as a decimal. Recall from the lesson summary that a conditional probability is calculated by $\frac{\text{Cell Value}}{\text{Column or Row Total}}$. Notice that while the numerators are the same in this question and the previous, the denominators are quite different.
- H Calculate the conditional probability that a child who was given the fake vaccine developed polio. Use probability notation and give the answer as a decimal.
- I Which children had a greater rate of developing polio?

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

- 2 A local animal shelter reports that it currently has 32 dogs and 28 cats available for adoption. Twenty of the dogs and 12 of the cats are male.
- A There are two variables and each variable has two categories. What are they?
The two variables are gender and type of animal. Give the two categories for each of the two variables. So you should have 4 answers: 2 categories for gender and 2 for type of animal.
- B Use the information given to make a two way table.
- C If an animal is selected at random, what is the probability that it will be male, given that it is a cat?
- D If an animal is selected at random, what is the probability that it will be a cat, given that it is a male?
- E If an animal is selected at random, what is the probability that it will be a male cat?

Lesson 6.2

Marginal, Joint, and Conditional Probabilities from Two-Way Tables

- 3 Imagine that two treatments, Treatment A and Treatment B, have been recommended for treating a particular medical condition. Both treatments have two potential side effects: severe skin rash and blurred vision.

The report of a recent study of side effects included the following information:

- 75 people who received Treatment A and 150 people who received Treatment B experienced severe skin rash.
- 5 people who received Treatment A and 11 people who received Treatment B experienced blurred vision.

The study had 650 participants. 200 of them received Treatment A. The rest of the group received Treatment B. Which treatment would you recommend and why? Construct a two way table and show your work in calculating the probabilities you use to answer this question. When you construct the table, keep in mind that not all the participants experienced a side effect. Therefore, the category “side effect” would have three values: skin rash, blurred vision and none.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

PART I: DAY ONE

INTRODUCTION

Buttoning Up Probability

Now that you have a basic understanding of probability, to help you visualize the outcomes you will use bags of buttons with different characteristics. Keep in mind that the buttons could represent similar classifications; for example, a shipment of new iPods that are 8 GB or 16 GB and are either white or black.

TRY THESE – PART 1

- 1 First, organize the data in the table. Carefully remove buttons from the bag and separate them into the listed categories. Fill in the cells in the table with the counts from your bag of buttons.

	Black	Blue	Red	Total
Two holes				
Four holes				
Total				

- 2 Split your group into two teams. Have Team 1 work the buttons as directed and have Team 2 use the table to perform the required calculation. Once you have completed your calculations for parts A – D, compare your answers.
 - A Find the probability that a randomly selected button is blue. Separate out the blue buttons and find the probability.
 - B Find the probability that a randomly selected button has four holes. Separate out the four-holed buttons and find the probability.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

- C Find the probability that a randomly selected button is blue *and* has four holes.

- D Find the probability that a randomly selected four-holed button is blue.

- E Compare the two probabilities in Questions 2C and 2D. How do they differ?

TRY THESE – PART 2

- 3 Now switch teams, Team 1 use the table to perform the required calculation and Team 2 work with the buttons as directed. After you have completed your calculations, compare your answers.
 - A Find the probability that a randomly selected button is black. Separate out the black buttons and find the probability.

 - B Find the probability that a randomly selected button has two holes. Separate out the buttons with two holes and find the probability.

 - C Find the probability that a randomly selected button is black *and* has two holes.

 - D Find the probability that a randomly selected black button has two holes.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

- E Compare the two probabilities in Questions 3C and 3D. How do they differ?

NEXT STEPS

Recall from Lesson 6.2 that we discussed three different types of probabilities. See the table below.

Review the probabilities you calculated in **TRY THESE PART I & II**. Identify the type of probability you were asked to calculate for each exercise A – D.

Definition
Marginal Probabilities: comparing column or row total to grand total; values in the margins
Joint Probabilities: compare a cell value (within a table) to the grand total
Conditional Probabilities: compare a cell value to a row total or column total

Also recall that:

Probabilities are stated “P(outcome)”	P(married)
---------------------------------------	-------------------

- 4 Using the table of buttons data:
1. Create your own question for each type of probability. Practice stating the probability using the notation above.
 2. Do the required calculations to find the probability.

A Marginal Probability

B Joint Probability

C Conditional Probability

Lesson 6.3

Building Two-Way Tables to Calculate Probability

NEXT STEPS

Oftentimes when interpreting probability, you need to think very carefully about how the outcome is defined. A simple change in wording can have a large impact on the estimated probability.

What Is the Question?

The National Basketball Association (NBA) consists of 30 professional male basketball teams. Consider the following:

- a) The probability that a randomly selected man who is over 6-foot tall is playing in the NBA (National Basketball Association)
 - b) The probability that a randomly selected NBA player is over 6-foot tall
- 5 Are these two probabilities about the same? If not, which do you think is larger? Tell why you think so.
- 6 These are good examples of conditional probabilities. What is the condition in each of these cases?

Building Two-Way Tables to Calculate Probability

PART II: DAY TWO

INTRODUCTION

Evaluating Drug Screening

Many companies require job applicants to take drug-screening tests prior to employment. Such tests are not always accurate. Sometimes a test is positive even when a person is not using drugs. This is called a *false positive*, because the positive result is incorrect. Sometimes the test is negative for a person who is using drugs. This is called a *false negative*, because the negative result is incorrect and the person is using drugs. When new tests are introduced, it is important to evaluate how likely these errors are. Companies use statistics to figure out how likely false positives and false negatives are.

There are many different types of drug-screening tests. In a study, scientists compared three different tests for illegal drugs.¹ To analyze each test, the scientists used a large number of blood samples. Before they analyzed the tests, they had to know the drug-use status of the people who gave blood samples. They had to know which blood samples came from people who used drugs and which came from people who did not use drugs (their drug-use status).

The scientists had to know the drug-use status to estimate the false-positive and false-negative rates. This helped them figure out how likely false positives and false negatives are for each test. The false-positive and false-negative rates for these three methods are shown in the following table. The false-positive rate is the percentage of non-drug users who test positive for drug use. The false-negative rate is the percentage of drug users who test negative for drug use.

Test	False-Positive Percentage (Not a drug user but the test is positive)	False-Negative Percentage (Is a drug user but the test is negative)
Test 1	37.4	10.3
Test 2	36.4	49.3
Test 3	35.7	4.8

TRY THESE

- 7 None of the three tests would be considered acceptable for employment drug screening. Why do you think this is so?

¹Lynch, K. L., Breaud, A. R., Vandenberghe, H., Wu, A. H. B., & Clarke, W. (2010). Performance evaluation of three liquid chromatography mass spectrometry methods for broad-spectrum drug screening. *ClinicaChimicaActa*, 411, 1474–1481.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

- 8 If you had to recommend one test for employment drug screening, which would you select? Why did you choose this test?

For the following questions we focus on Test 3, which has a false-positive rate of 35.7% and a false-negative rate of 4.8%.

Test	False-Positive Percentage (Not a drug user but the test is positive)	False-Negative Percentage (Is a drug user but the test is negative)
Test 1	37.4	10.3
Test 2	36.4	49.3
Test 3	35.7	4.8

- 9 If this test is used on people who actually use illegal drugs, what percentage of the tests will correctly show that they use drugs?
- 10 If this test is employed on people who do not use drugs, what percentage of the tests will correctly indicate that they do not use drugs?
- 11 Imagine 1% of the people sent for a screening test actually use illegal drugs. The employer does not know this. The employer screens these people with Test 3. You know there will be some false positives as well as some real positives.
- A Suppose 100 people test positive. Give your best guess for the number of these 100 people who are actually illegal drug users.
- B Your answer to part A was some number out of 100. Write this as a fraction and then as a decimal. This is your estimate of the probability that *if a person tests positive, he or she is actually a drug user*.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

In this exercise you are trying to estimate a *conditional probability*. You are trying to estimate the probability that someone is a drug user given that they test positive. Remember that a conditional probability is a probability based on a given condition. In the previous sentence, the word “given” gives you the hint that “they test positive” is the condition. The condition is that the 100 people tested positive, but we don’t know for sure if they are really drug users. So, we need to evaluate the percentage of these people who are actually drug users.

Let’s see how good your estimate was. One way to evaluate the percentage of people testing positive who are actually drug users is to construct a two-way table. Imagine a hypothetical *population* of 10,000 people who will be screened using Test 3.

- 12 Suppose that 1% of these people are actually drug users. How many drug users are there in the hypothetical population of 10,000? How many nonusers are there? Enter these numbers in the two-way table below. Remember that $1\% = 1/100 = 0.01$.

	Positive test	Negative test	Total
Drug user			
Not a drug user			
Total			10,000

- 13 Think about the people who are *not* drug users. Of those, 35.7% of them test positive even though they are not drug users (false positive).

About how many of the non-drug users test positive? Remember that if you are counting people, you must use a whole number. Enter this number in the appropriate location in the table.

- 14 Consider those who are drug users. Of those, 4.8% of them falsely test negative. About how many of these people test negative? Enter this number in the appropriate location in the table.

- 15 Use addition and subtraction to fill in the remaining entries and row totals for the table.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

16 *If a person tests positive, what is the probability he or she is actually a drug user? (Hint: Is this a joint, marginal, or conditional probability?) Is this surprising? Was your guess from Question 11 close to the actual probability?*

17 If a person tests negative, what is the probability they are actually not a drug user?

18 Next, let's take a look at what happens if the percentage of the population who are drug users is 10% rather than 1%.

Again, consider a population of 10,000 people and assume that 10% of these people are actual drug users. How many drug users are there in the population? How many nonusers are there? Enter these numbers as the row totals in the table below.

	Positive test	Negative test	Total
Drug user			
Not a drug user			
Total			10,000

19 Take a look at those people who are not drug users. If 35.7% of them test positive, how many of them test positive? Enter this number in the appropriate cell in the table.

20 Consider those who are drug users. If 4.8% of them test negative, how many negatives are there? Enter this number in the appropriate cell in the table.

21 Enter the remaining values and totals in the table.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

- 22 If a person tests positive, what is the probability they are actually a drug user? Does this surprise you?
- 23 If a person tests negative, what is the probability they are actually not a drug user?

Cancer Screening

Interpreting the results of medical tests can be difficult, even for doctors. In a study to see whether doctors are correct in their interpretations, a researcher studied 160 doctors. The researcher asked a question that is similar to the following question.²

Suppose mammograms (x-ray examinations of human breasts) are used to screen for breast cancer in a particular city. The following information is known.

- 1% of the women in the city have breast cancer.
- If a woman has breast cancer, the probability that the mammogram is positive is 0.90. (This rate of correct positives is called *sensitivity*.)
- If a woman does not have breast cancer, the probability that the mammogram is positive is 0.09 (this is the false-positive rate).

TRY THESE

- 24 A woman has a mammogram, and the result is positive. She wants to know the probability that she has breast cancer in light of the positive mammogram. Which of the following is the best answer?
- A The probability that she has breast cancer is about 81%.
 - B Out of 100 women with a positive mammogram, about 90 have breast cancer.
 - C Out of 100 women with a positive mammogram, about 9 have breast cancer.
 - D The probability that she has breast cancer is about 1%.

Surprisingly, only 21% of the doctors in a study answered a question like this correctly.

²Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8, 53–96.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

Working with a partner, use the table approach to determine the best answer to this question. Say you begin with a population of 10,000 women getting a mammogram.

Notice that even though there are a large number of false positives, the probability that a woman has breast cancer is still greater if you know she has tested positive than if you do not know the result of the mammogram.

WRAP UP

Our intuition, hunches, or guesses about probability and risk are often not very accurate. This is especially true when dealing with conditional probability. The activities in this lesson were designed to help you think about probability and risk in a systematic way. The technique of translating probability information into a “hypothetical 10,000” table provides a powerful tool for computing and interpreting probabilities.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

STUDENT NAME _____ DATE _____

TAKE IT HOME – FOR PART I: DAY ONE

1 In its monthly report, the local animal shelter states that it currently has 24 dogs and 18 cats available for adoption. Eight of the dogs and 6 of the cats are male.

A Use the information in the report to help you create a two-way table. Ask yourself: What are the two types of pets and what are the two categories for gender?

B Use the table to find the probability that a randomly selected pet is a dog *and* is female. Give the final probability as a decimal.

C Use the table to find the probability that a randomly selected pet is a cat, *given* that it is a female. Give the final probability as a decimal.

D Which pet, if randomly selected, has a greater probability of being a male?

Lesson 6.3

Building Two-Way Tables to Calculate Probability

2 A public university is interested in addressing the issue of binge drinking on their campus. They collect some preliminary data on frequent binge drinking behavior and gender by conducting a random survey of 17,096 male and female students. The report that resulted from the survey included the following information:

- Of the 7180 males surveyed, 1630 replied “yes” when asked if they were frequent binge drinkers.
- Of the 9916 females surveyed, 8232 reported that they were not frequent binge drinkers.

A Use the information in the report to help you create a two-way table. Ask yourself: What are the two responses to the survey question and what are the two categories for gender?

B Use the table to find the probability that a randomly selected student is a frequent binge drinker. Give the final probability as a decimal.

C Use the table to find the probability that a randomly selected student is a female, *given* that the student is a frequent binge drinker. Give the final probability as a decimal.

D Use the table to find the probability that a randomly selected student is a male, *given* that the student is a frequent binge drinker. Give the final probability as a decimal.

E Based on your answers to C and D, which gender represents the greater proportion of frequent binge drinkers?

Lesson 6.3

Building Two-Way Tables to Calculate Probability

F For which gender does frequent binge drinking seem to be a bigger problem? (In other words, for which gender is the probability of being a frequent binge drinker greater?)

G How does your answer to part E compare to your answer to part F? How do they differ?

Lesson 6.3

Building Two-Way Tables to Calculate Probability

STUDENT NAME _____ DATE _____

TAKE IT HOME – FOR PART II: DAY TWO

- 1 Consider the breast cancer example from the lesson. Recall that 90% of the women who tested positive had breast cancer; while 9% of the women who tested positive did not have cancer. What if the percentage of women who had breast cancer was 10%? Use the table approach to determine the probability that a person who tests positive actually has breast cancer. Assume a population of 10,000 women.

Lesson 6.3

Building Two-Way Tables to Calculate Probability

- 2 Blood donors are usually tested for HIV. The blood is screened for the safety of the blood supply and for the benefit of the donor. The test, called an ELISA, tests positive 97.5% of the time if the donor actually has HIV. If the donor does not have HIV, the ELISA test correctly indicates that the person does not have the disease 92.6% of the time.

According to the Centers for Disease Control (CDC) about 0.2% of college students have HIV. After a blood drive at a college, the lab calls and tells a student that the test results indicated he has HIV.

- A Use what you learned in this lesson to find the probability that the student actually has HIV. Because the percentage of college students with HIV is so small, it may be better to use 100,000 as your total population.

- B Considering the answers to questions 1 and 2, should you be more concerned about a positive test for a rare disease, like HIV, or a common disease, like breast cancer?

Lesson 6.3

Building Two-Way Tables to Calculate Probability

- 3 In the game of tennis, one player hits the ball to the opponent to start a game. This first hit of the ball is called a serve. A player gets two attempts to serve the ball correctly. If the first serve lands in the court, the serve is played until one player wins the point. If the first serve is out, the server can then hit a second serve, and that serve is played until one player wins the point.

For this question, imagine a tennis player named David. David wins 76% of the points when his first serve lands in the court. So, his probability of winning the point if the first serve lands in is 0.76.

David wins 43% of the points in which his first serve does not land in and he has to try again.

Also, suppose David's first serve lands in for 64% of his serves.

Use this information to complete the hypothetical 10,000 table below and to determine the probability that this player wins a point when he serves. (Hint: Because there are many percentages here, read each sentence carefully to not get confused.)

	Wins point	Does not win point	Total
First serve in			
First serve not in			
Total			10,000

Chapter 6 Review

0 categorical 1 two-way tables proportions decimal percentages probabilities

We looked at a way to display data for two _____ variables. We constructed _____ to represent the relationship between two categorical variables. This allows us to see in a simple way the number of individuals or objects who fall into each combination of categories. At first we used _____ and _____ to explore the data. Then we learned about three types of _____ that can be determined from two-way tables: *marginal probabilities, joint probabilities, and conditional probabilities* (see the table below). In general, probability refers to the likelihood that something will occur and is expressed in _____ or fraction form. Probability will always be a value between ___ and ___.

Definition	Example (Refer to data table given below)
Marginal probabilities focus on a single categorical variable. The numerator comes from the margin (category total) of the two-way table and the denominator is the grand total.	$P(\text{male}) = \frac{111.1}{229.1} = \frac{\text{Margin Value}}{\text{Grand Total}}$
Joint Probabilities: numerator is a cell value (within a table) and the denominator is the grand total	$P(\text{male and married}) = \frac{64.4}{229.1} = \frac{\text{Cell Value}}{\text{Grand Total}}$
Conditional Probabilities: numerator a cell value (within a table) and the denominator is either a row total or column total	$P(\text{married} \text{male}) = \frac{64.4}{111.1} = \frac{\text{Cell Value}}{\text{Row or Column Total}}$

Here again is data from 311 customers who purchased a Honda Civic.

	Hybrid Honda Civic	Standard-engine Honda Civic	Row totals
Male	77	117	194
Female	34	83	117
Column totals	111	200	311

What does the data suggest about the relationship between gender and engine type?

- A. Women are more likely to purchase a Honda Civic with a standard engine than men.
- B. Women are less likely to purchase a Honda Civic with a standard engine than men.
- C. Women and men are equally likely to purchase a Honda Civic with a standard engine.

Here again is data from 311 customers who purchased a Honda Civic.

	Hybrid Honda Civic	Standard-engine Honda Civic	Row totals
Male	77	117	194
Female	34	83	117
Column totals	111	200	311

Which of the following probability statements is found with the computation $117 / 200$?

- A. $P(\text{male and standard engine})$
- B. $P(\text{standard engine} \mid \text{male})$
- C. $P(\text{male} \mid \text{standard engine})$
- D. $P(\text{female})$

Here are the results of a survey that students conducted at a mall. The students conducted this survey as part of a statistics project to determine if younger adults are more likely to have tattoos.

	At least one tattoo	No tattoo	Row Totals
Age 18 - 29	170	320	490
Age 30 - 50	60	445	505
Column Totals	230	765	995

We randomly select a person who responded to this survey. Which calculation gives the probability that the person has a tattoo?

- A. 765 out of 995
- B. 230 out of 995
- C. 230 out of 765
- D. 60 out of 505

If we randomly select a person in the sample who is 30 to 50 years old, what is the probability that this person has a tattoo?

- A. 0.12
- B. 0.26
- C. 0.06
- D. 0.46

If we randomly select a person in the sample, what is the probability that this person is 30 to 50 years old and has a tattoo?

- A. 0.12
- B. 0.26
- C. 0.06
- D. 0.46

+++++

The original version of this material is from STATWAY™, A Pathway Through College Statistics, which is a product of a Carnegie Networked Improvement Community that seeks to advance student success and The Charles A. Dana Center at The University of Texas at Austin under sponsorship of the Carnegie Foundation for the Advancement of Teaching. This material and all subsequent versions is a result from the continuous improvement efforts of Valencia College.

