Unit 9, Lesson 1
Connect the Dots

Learning Focus

Represent data using a scatterplot.

Understand the meaning of the correlation coefficient.

Describe the difference between correlation and causation.

After collecting data, how can we tell if there is a relationship between two variables?

Open Up the Math: Launch, Explore, Discuss

One reason to understand functions is that they allow us to model data and then use it to understand the world around us. Data and statistics are used to think about science, medicine, business, sports, and about everything in between. There is a whole field of study focused on collecting and analyzing data, and we will touch on only a little here. Statistics is the practice of asking questions and looking for answers in the data. In the next few lessons, we will be using scatterplots and bivariate data to analyze relationships between variables.

Let's get started by looking at some data. Your teacher will give you some cards, each of which contains a table of data and a scatterplot of the data. Your job is to examine the data and put these cards in order in a way that makes sense to you and that you can justify.

1. Compare each scatterplot with its correlation coefficient. What patterns do you see?

2. Use the data in Set A as a starting point. Keeping the same x-values, modify the y-values to obtain a correlation coefficient as close to 0.75 as you can.

   Record your data here:

2	2.3	3.3	3.7	4.2	4.6	4.5	5	5.5	5.7	6.1	6.4

   What did you have to do with the data to get a greater correlation coefficient?

3. This time, again start with the data in Set A. Keep the same x-values, but this time, modify the y-values to obtain a correlation coefficient as close to 0.25 as you can.

   Record your data here:

2	2.3	3.3	3.7	4.2	4.6	4.5	5	5.5	5.7	6.1	6.4

   What did you have to do with the data to get a correlation coefficient that is closer to 0?

4. One more time: start with the data in Set A. Keep the same x-values, and modify the y-values to obtain a correlation coefficient as close to -0.5 as you can.

   Record your data here:

2	2.3	3.3	3.7	4.2	4.6	4.5	5	5.5	5.7	6.1	6.4

   What did you have to do with the data to get a correlation coefficient that is negative?

5. What aspects of the data does the correlation coefficient appear to describe?

6. On the night before the last math test, Shaniqua held a study group at her house. It was a fun night; they ate a lot of pizza, did math, and laughed a lot. Shaniqua scored better on her test than usual and thought it might be related to pizza. She collected the following data from her friends in the study group:

	Shaniqua	David	Susana	Ruby	Deion	Oscar
Number of Pizza Slices Eaten	2	6	1	4	3	5
% Increase in Test Score	5	9	4	7	6	8

Create a scatterplot of this data, and calculate the correlation coefficient.

Based on this data, would you recommend eating pizza on the night before a test to increase scores? Why or why not?

Ready for More?

1. Describe a situation with two variables that may have a high correlation but are not causally related.

2. What are some reasons that two variables may be highly correlated but may not have a causal relationship?

Takeaways

		It means:	The scatterplot looks like:
The correlation coefficient tells us: the strength of a linear relationship between two variables.	If the correlation coefficient is: 0		
	If the correlation coefficient is: $0 < r < .7$		
	If the correlation coefficient is: $.7 \leq r \leq 1$		
	If the correlation coefficient is: $-.7 < r < 0$		
	If the correlation coefficient is: $-1 \leq r \leq -.7$		

Vocabulary

- **bivariate data**
- **causation**
- **correlation**
- **correlation coefficient**
- observed value
- scatter plot
- **variable (statistics)**

Bold terms are new in this lesson.

Lesson Summary

In this lesson, we have learned about representing two-variable quantitative data with a scatterplot. We have learned that one of the ways we can judge if a line is a good model for the data is by using the correlation coefficient.

Unit 9, Lesson 6
Food for Thought

Learning Focus

Represent data with box plots, dot plots, and histograms.

Analyze data represented in different ways.

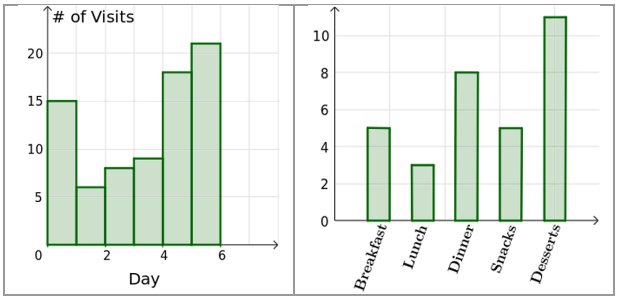
How can we describe differences in data sets? What features should we look for in different representations?

Open Up the Math: Launch, Explore, Discuss

Aaliyah loves to cook and collect recipes. She started her own food blog to share her love of food and passion for cooking. She hopes to build up a following and to eventually get sponsors, so that her blog can become a source of income.

The web service that hosts her site provides some data analytics that she likes to look at to make decisions about future posts.

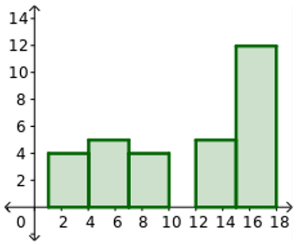
Last week she collected this data. The first set of data shows the number of visits to the site each day, starting with Monday. The second set of data shows how many recipes of each type were downloaded during the week.

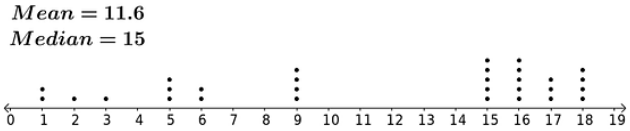


1. Compare the two displays of data. What similarities and differences do you observe?

Aaliyah found that her most popular recipe was for her special, chewy, moist brownies. She checked the analytics and found that during the week, 30 people had checked out the brownie page. The next two diagrams both represent the data for the number of minutes that people spent on the brownie page.

2. Compare the two representations of the data.



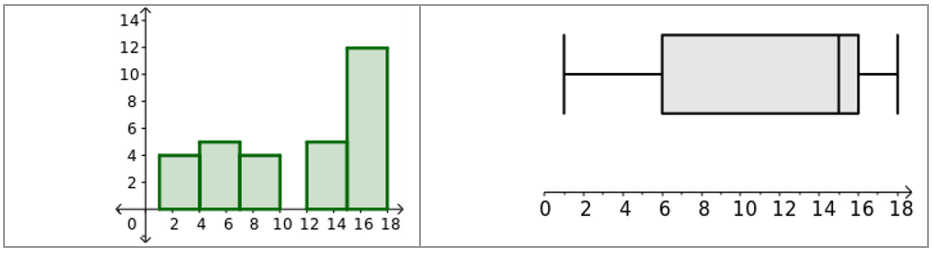


   a. What do you observe about the number of minutes spent on the brownie page?

   b. What aspects of the data are highlighted or obscured by each representation?

Aaliyah's math teacher said that there is one more representation that she should see, the box plot, sometimes called the box and whisker plot.

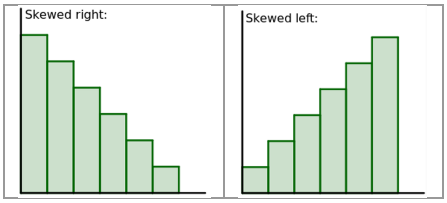
3. Compare the two representations of the data. What aspects of the data are highlighted or obscured by each representation?

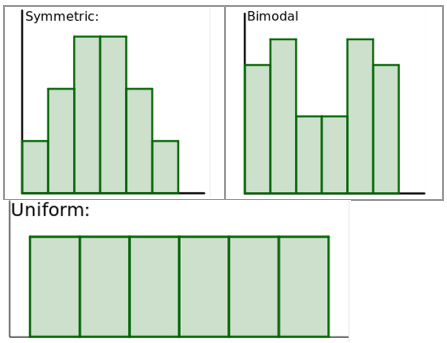


Aaliyah likes histograms and wants to use them more. She asked her math teacher what she should be looking for when she examines a representation of numerical data. Her teacher said that she should look for:

Center: Measures of center include the mean and median.

Shape: There are many possible shapes, but some of the most common are:





Spread: Tells us something about how wide, or disperse, the data are. Measures include the range and the mean absolute deviation. (We'll talk more about spread in the next lesson.)

4. Without using technology, sketch a data distribution with the following characteristics:

   a. A histogram with seven bins that is skewed right.

   b. A box plot with a median of 5 and a range of 15.

   c. A dot plot that is mostly uniform and has an outlier.

   d. A bimodal histogram that is not symmetric.

5. Aaliyah found data for the number of minutes that people spent browsing her recipes blog. Use technology to create a histogram with 5 bins, and describe the center, shape, and spread.

2	3	3	5	5	6	6	6
6	10	10	10	12	12	12	13
13	15	15	15	15	16	16	16
16	16	16	16	17	17	17	18
18	18	18	19	19	20	20	20

Ready for More?

You're surrounded by data in your life! Think of contexts that would produce histograms with the following shapes.

1. Bimodal:
2. Symmetric and unimodal:
3. Skewed right:

Takeaways

Single-variable quantitative data:

Representations	
Features	

Vocabulary

- **bimodal distribution**
- box and whisker plot (box plot)
- categorical data or categorical variables
- **center (statistics)**
- **distribution of a variable (statistics)**
- dot plot
- histogram
- interquartile range - IQR
- mean
- mean absolute deviation - M.A.D
- measures of central tendency
- median
- mode(s)
- outliers
- quantitative variable
- range (statistics)
- **skewed distribution**
- **spread of a distribution (statistics)**
- standard deviation
- symmetric
- **symmetric distribution**
- **uniform distribution**
- **univariate data**
- **variability**

Bold terms are new in this lesson.

Lesson Summary

In this lesson, we focused on single-variable quantitative data. We compared histograms, dot plots, and box plots to consider what aspects of the data are highlighted in each. We learned to look for the center, the shape, and the spread of the data.

Unit 9, Lesson 7

Bridging the Gap

Learning Focus

Understand standard deviation.

How does standard deviation compare to range as a measure of spread?

Open Up the Math: Launch, Explore, Discuss

Your job in this task is to figure out what the standard deviation measures are and to get a sense of how changes in data affect standard deviation. Your teacher will give you a set of strips that each contain a set of data and a dot plot that represents the data. Follow the instructions for each problem below, and be prepared to discuss your findings.

1.

    a. Use technology to find the standard deviation for each set of data on the strips you have been given.

    b. Arrange the strips with the standard deviations from least to greatest. Write the sets they represent here. Describe any patterns or trends that you notice.

2. Compare data sets 4 and 2. Which one has the greater standard deviation? Why do you think the standard deviation is greater for this data set?

3. Look at data set 6. What impact did the 10 have on the standard deviation? What would happen to the standard deviation if you eliminated this data point? What would happen if you made the 10 a greater number?

4. Compare data sets 6 and 7. What do you notice about the standard deviations for these data sets? Why do you think this is the case?

5. Create your own set of data that has exactly the same standard deviation as data set 2 but a different mean.

6. What does the standard deviation help describe in a distribution?

7. Samantha is entering a bridge-building contest for her math class. The winner of the contest will be the student who can build the bridge that holds the most weight. As she is getting ready to build her bridge, she has narrowed her decision down to three different designs that she could use for her bridge. She has data from bridges that have used the three different bridge designs from previous years, and she is trying to decide which design to go with. The data for the amount of weight, in pounds, for the three different types of designs are given below. Which bridge design should she use? Use both graphs and numerical summaries to back up your decision for the best design.

    Design #1: $850, 912, 980, 1012, 1060, 1407, 1600, 1602, 1680, 1700$

    Design #2: $1380, 1390, 1402, 1425, 1450, 1455, 1477, 1490, 1493, 1500$

    Design #3: $650, 740, 807, 890, 906, 920, 925, 960, 980, 1003, 2257$

Ready for More?

Explain why outliers affect the standard deviation. Do outliers make the standard deviation less or greater?

Takeaways

Standard deviation: _____

- _____.
- _____.
- _____.

Standard deviation and mean:

- _____.
- _____.
- _____.
- _____.
- _____.

Vocabulary

- standard deviation

Bold terms are new in this lesson.

Lesson Summary

In this lesson, we learned about standard deviation, a measure of spread for single-variable quantitative data. The standard deviation is 0 when all the data have the same value.

Unit 9, Lesson 8

Making the Grade

Learning Focus

Compare sets of data using center, spread, and shape.

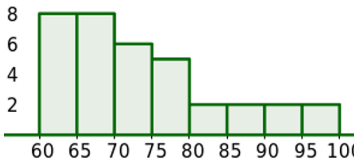
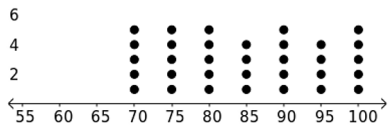
How do I use measures of center and spread together to make decisions about data?

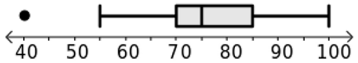
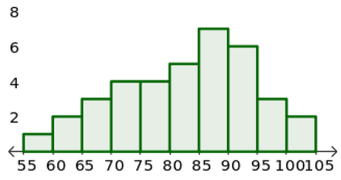
How do I compare two or more data sets that are represented by different plots?

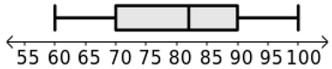
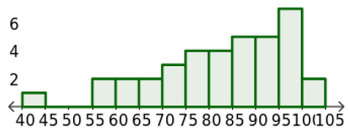
Open Up the Math: Launch, Explore, Discuss

Principal Scanner wants to make his school the best ever. He has offered to buy pizza for the class that shows the best results on the department final at the end of the semester. He tells all the teachers in the department to submit their student test-score data so that he can determine which class did the best. Unfortunately, Principal Scanner underestimated the job he was going to have in determining the best class. To make matters worse, teachers used different representations for the data they submitted for him to evaluate.

So, Principal Scanner has turned the job over to you. You are to compare the two data sets in each problem and give the principal a report, using the sentence frames provided for this purpose. Give each comparison some thoughtful analysis, and use your best statistical language to describe your findings. Somebody's pizza is depending on you!

Data Set I: Williams' Class	**Data Set II: Lemon's Class**
	
Mean $= 73.21, \sigma = 11.95$	Mean $= 84.7, \sigma = 10.3$

Data Set III: Croft's Class	**Data Set IV: Anderson's Class**
	
	Mean $= 80.27, \sigma = 11.6$

Data Set V: Hurlea's Class	**Data Set VI: Jones' Class**
	
	Mean $= 80.27, \sigma = 14.43$

1. Compare data distributions between Anderson's and Williams' classes.

    a. Complete the sentence frames with your findings:

    The data from _____ has a greater _____ than _____because _____.

    The data from _____ has less _____ than _____ because _____.

    The data from _____ is skewed left/right (choose one) while the data from _____ is not because _____.

    b. Which of these two classes do you think did the best? Why?

2. Compare data distributions between Croft's and Hurlea's classes.

    a. Complete the sentence frames with your findings:

    The data from _____ has a greater _____ than _____ because _____.

    The data from _____ has less _____ than _____ because _____.

    The data from _____ contains an outlier while the data from _____. I know this because _____.

    b. Which of these two classes do you think did the best? Why?

3. Compare data distributions between Jones' and Williams' classes.

    a. Complete the sentence frames with your findings.

    The data from _____ has a greater _____ than _____ because _____.

    The data from _____ has less _____ than _____ because _____.

    The data from _____ is skewed left/right (choose one) while the data from _____ is not because _____.

    b. Which of these two classes do you think did the best? Why?

4. Compare data distributions between Lemon's class and that of a teacher of your choice.

    a. Complete the sentence frames with your findings.

    The data from _____ has a greater _____ than _____because _____.

    The data from _____ has less _____ than _____ because _____.

    b. Which of these two classes do you think did the best? Why?

Ready for More?

You now get to make your final recommendation to Principal Scanner. Which class do you choose, and why?

Takeaways

Box plot:

Histogram:

Dot plot:

Lesson Summary

In this lesson, we compared data distributions using shape, center, and spread. We learned to consider the interquartile range and the median when using a box plot and to consider the mean and the standard deviation when using a histogram. We also discussed the effect of outliers on the median, mean, and standard deviation.

Unit 9, Lesson 9
Two-Way Streets

Learning Focus

Organize two-variable categorical data.

Interpret two-way frequency tables.

What information does a two-way table reveal?

What information is highlighted when data is interpreted from relative frequency tables?

Open Up the Math: Launch, Explore, Discuss

Rashid is in charge of determining the upcoming after-school activity. To determine the type of activity, Rashid asked several students whether they prefer to have a dance or to play a game of soccer. As Rashid collected preferences, he organized the data in the following two-way frequency table:

	Girls	Boys	Total
Soccer	14	40	54
Dance	46	6	52
Total	60	46	106

Rashid is feeling unsure of the activity he should choose, based on the data he has collected, and is asking for help. To better understand how the data is displayed, it is useful to know that the outer numbers, located in the margins of the table, represent the total frequency for each row or column of corresponding values and are called marginal frequencies. Values that are part of the inner body of the table are created by the intersection of information from the column and the row, and they are called the joint frequencies.

1. Using the data in the table, construct a viable argument, and explain to Rashid which after-school event he should choose. To support your argument, use the following sentence frame to get started.

   I think the after-school event should be _____ because _____ out of _____ prefer _____.

   This means _____.

Two-way frequency tables allow us to organize categorical data in order to draw conclusions. For each set of data below, create a frequency table. When each frequency table is complete, write three sentences about observations of the data, including any trends or associations in the data.

2. **Data set:** There are 45 total students who like to read books. Of those students, 12 of them like nonfiction and the rest like fiction. Four girls like nonfiction. Twenty boys like fiction.

	Fiction	Nonfiction	Total
Boys			
Girls			
Total			

   Observation 1: _____ out of _____ prefer _____.

   Observation 2:

   Observation 3:

3. **Data set:** 35 seventh graders and 41 eighth graders completed a survey about the amount of time they spend on homework each night. 50 students said they spent more than an hour. 12 eighth graders said they spend less than an hour each night.

			Total
More than one hour			
Less than one hour			
Total			

Observation 1: _____ out of _____ do _____.

Observation 2:

Observation 3:

Rachel and her mother have an ongoing disagreement about whether Rachel spends too much time on her phone and sends too many texts. To bolster their arguments, Rachel collected data from among her friends and her mother collected data from among her friends. We will learn later that this not a good sampling method, but that is a story for another day. Rachel thought about the data she and her mom collected for the average number of texts a person sends each day and started thinking that perhaps a two-way table of the data they collected would help convince her mom that she does not send an excessive number of texts for a teenager. The table separates the data by age (teenager and adult) and by the average number of texts sent (more than 100 per day or less than 100 per day).

	Average is more than 100 texts sent per day	Average is less than 100 texts sent per day	Total
Teenager	20	4	24
Adult	2	22	24
Total	22	26	48

4. Write two observation statements from this two-way table.

To provide further evidence, Rachel decided to do some research. She found that only 43% of people with phones send over 100 texts per day. She was disappointed that the research data did not support her case and confused because it did not seem to match what she found in her survey.

5. What questions do these statistics raise for you? What data should Rachel look for to support her case?

After looking more closely at the data, Rachel found other percentages within the same data that seemed more in line with the data she collected from her teenage friends.

6. How might Rachel use the data in the two-way table to find percentages that would be useful for her case?

Once Rachel realized there are a lot of ways to look at a set of data in a two-way table, she was motivated to learn about relative frequency tables and conditional frequencies. When the data is written as a percent, this is called a **relative frequency table**. In this situation, the "inner" values represent a percent and are called conditional frequencies. The conditional values in a **relative frequency table** can be calculated as percentages of one of the following:

- The whole table (relative frequency of table)
- The rows (relative frequency of rows)
- The columns (relative frequency of column)

Since Rachel wanted to emphasize that a person's age makes a difference in the number of texts sent, the first thing she decided to do was to focus on the **row** of values, so she could write conditional statements about the number of texts a person is likely to send based on their age. This is called a **relative frequency of row** table.

7. Fill in the percentage of teenagers for each of the conditional frequencies in the highlighted row:

	Average is more than 100 texts sent per day	Average is less than 100 texts sent per day	Total
% of Teenagers	20 _ %	4 _ %	24 100%
% of Adults	2 8%	22 92%	24 100%
% of Total	22 46%	26 54%	48 100%

Since the **percentages** that were created focus on **row** values, all conditional observations are specific to the information in the row. Complete the following sentence for the **relative frequency of row**:

8. Complete:

   Of all teenagers in the survey, _ % average more than 100 texts per day.

9. Write another statement based on the **relative frequency of row**:

Below is the **relative frequency of column,** using the same data. This time, all of the percentages are calculated using the data in the column.

	Average is more than 100 texts sent per day	Average is less than 100 texts sent per day	Total
% of Teenagers	20 91%	4 15%	24 50%
% of Adults	2 9%	22 85%	24 50%
% of Total	22 100%	26 100%	48 100%

10. Write two conditional statements using the **relative frequency of column**.

This data represents the **relative frequency of whole table**:

	Average is more than 100 texts sent per day	Average is less than 100 texts sent per day	Total
% of Teenagers	20 42%	4 8%	24 50%
% of Adults	2 4%	22 46%	24 50%
% of Total	22 46%	26 54%	48 100%

11. Create two conditional distribution statements for the **relative frequency of whole table**.

12. What information is highlighted when data is interpreted from **relative frequency tables**?

Ready for More?

Create a relative frequency of column table from the two-way table in problem 3.

			Total
% More than one hour			
% Less than one hour			
% Total			

Takeaways

Two-way Frequency Table:

	Girls	Boys	Total
Soccer	14	40	54
Dance	46	6	52
Total	60	46	106

Relative Frequency Table:

	Average is more than 100 texts sent per day	Average is less than 100 texts sent per day	Total
% of Teenagers	20 42%	4 8%	24 50%
% of Adults	2 4%	22 46%	24 50%
% of Total	22 46%	26 54%	48 100%

Vocabulary

- **conditional frequency**
- **joint events**
- **marginal frequency**
- **relative frequency table (statistics)**
- **two-way frequency and two-way relative frequency table**
- **two-way table**

Bold terms are new in this lesson.

Lesson Summary

In this lesson, we learned about two-way tables and relative frequency tables. Both types of tables are used for two-variable categorical data. Two-way tables show data counts, and relative frequency tables show percentages.