Name: _____ Block: _____ Teacher: _____

Algebra 1

Unit 6 Notes:
Describing Data

DISCLAIMER: We will be using this note packet for Unit 6. You will be responsible for bringing this packet to class EVERYDAY. If you lose it, you will have to print another one yourself. An electronic copy of this packet can be found on my class blog.

KEY STANDARDS

Summarize, represent, and interpret data on a single count or measurement variable.

- **MGSE9-12.S.ID.1** Represent data with plots on the real number line (dot plots, histograms, and box plots). Choose appropriate graphs to be consistent with numerical data: dot plots, histograms, and box plots.

- **MGSE9-12.S.ID.2** Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, mean absolute deviation, standard deviation) of two or more different data sets.

- **MGSE9-12.S.ID.3** Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers). Students will examine graphical representations to determine if data are symmetric, skewed left, or skewed right and how the shape of the data affects descriptive statistics.

Summarize, represent, and interpret data on two categorical and quantitative variables.

- **MGSE9-12.S.ID.5** Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

- **MGSE9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

- **MGSE9-12.S.ID.6a** Decide which type of function is most appropriate by observing graphed data, charted data, or by analysis of context to generate a viable (rough) function to best fit. Use this function to solve problems in context. Emphasize linear, quadratic, and exponential models.

- **MGSE9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

- **MGSE9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

- **MGSE9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient "r" of a linear fit. (For instance, by looking at a scatterplot, students should be able to tell if the correlation coefficient is positive or negative and give a reasonable estimate of the "r" value.) After calculating the line of best fit using technology, students should be able to describe how strong the goodness of fit of the regression is, using "r."

- **MGSE9-12.S.ID.9** Distinguish between correlation and causation.

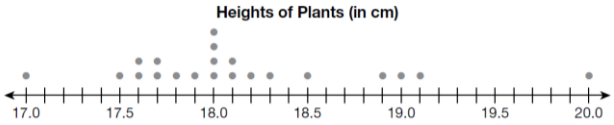
Lesson 1 - Univariate Statistics: Shape, Center, and Spread Shape

Univariate Data – Data involving one variable.

No matter what types of study you choose, it helps to organize your data in a data display. Here are some types of data displays:

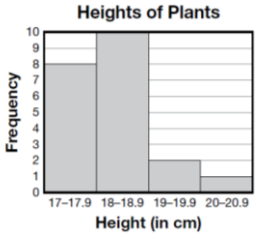
Dot Plot

- Used for numerical data that has relatively few points.
- Dots or x's can be used

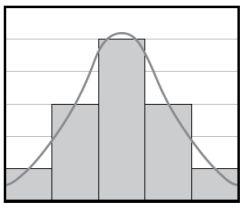


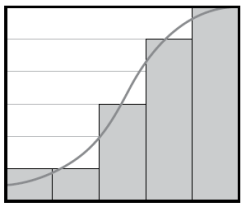
Histogram

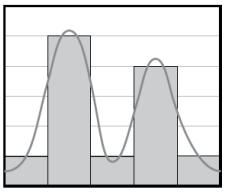
- Groups data points into ranges with **equal** intervals
- Intervals **do not overlap**

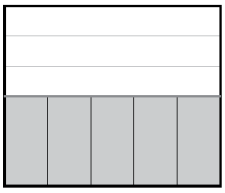


UNDERSTAND You can use the distribution of a data set, or its shape, to interpret it and to compare it to other data sets. Four kinds of distributions are described below.

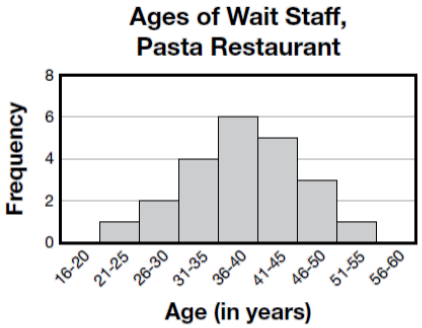

Normal Distribution

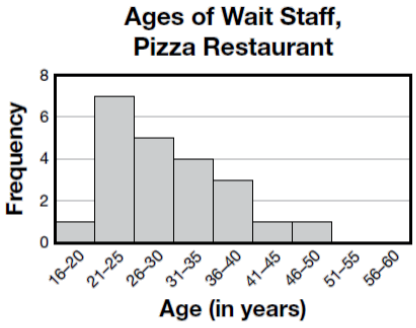

Skewed Distribution


Bimodal Distribution


Uniform Distribution

EXAMPLE B The two histograms below show the ages of wait staff at two restaurants.





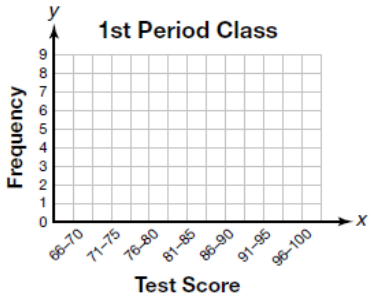
Identify the kind of distribution shown by each histogram. Use the shapes of the data sets to compare them.

For questions 3–5, use the given information. Create a histogram for each data set. Describe the distribution of each data set.

Students in the 1st and 2nd period biology classes took the same test. Their test scores are listed below.

3. 1st period test scores:

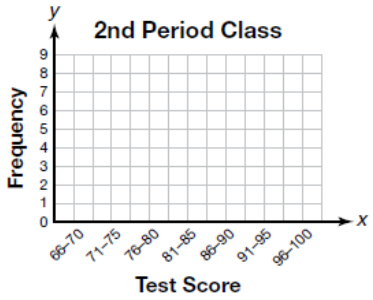
100, 91, 86, 73, 81, 100, 93, 94, 86,
86, 99, 93, 98, 84, 80, 97, 93,
87, 70, 97, 94, 88, 85, 96, 90



Distribution: _____

4. 2nd period test scores:

81, 87, 95, 85, 83, 82, 76, 68, 86,
83, 93, 87, 76, 87, 71, 100, 76,
91, 73, 80, 80, 84, 87, 88, 73



Distribution: _____

5. Compare and contrast the histograms for the biology classes in questions 4 and 5.

_____

_____

Create a dot plot for the given data. Describe the shape of the data.

6. Nathaniel opened 20 peanut shells and recorded the number of peanuts he found in each shell.

3, 2, 0, 1, 5, 2, 1, 2, 3, 1, 2, 2, 1, 2, 2, 3, 2, 3, 1, 2

_____

Fill in each blank with an appropriate word or phrase.

7. A _____ shows data points as dots above a number line.

8. A _____ shows how frequently data occur within certain ranges or intervals.

9. _____ used in a histogram must be equal.

10. A _____ distribution is symmetric and resembles a bell curve.

11. A _____ distribution has two distinct peaks.

12. A _____ distribution has a "tail" that extends more to one side of the graph than the other.

Measures of Center

Measures of Center are used to generalize data sets and identify common values.

Mean	**Definition:** Average of a numerical data set, $\bar{x}$ **Calculation:** Add up all the data values and divide by the number of data values.
Mode	**Definition:** Value that occurs most frequently. There can be no, one, or several modes.
Outlier	Data value that is much greater than or much less than the rest of the data in a data set If an outlier is present, you would use the median to describe the data, NOT the mean!

Example: Below are the scores that Justin earned on his last 8 homework assignments.

80, 95, 0, 90, 95, 80, 85, 90

1. What is his mean/ average homework score?

2. What is his median homework score?

3. Are there any outliers?

Find the mean and median for each data set.

3. 5, 25, 10, 15, 20

median: _____

mean: _____

4. 1, 7, 3, 2, 6

median: _____

mean: _____

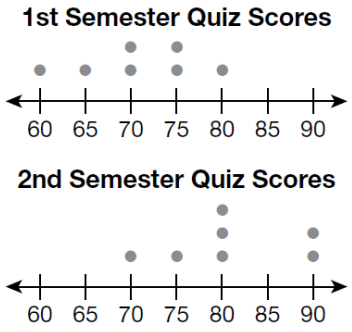
5. 10, 90, 10, 60, 40, 30

median: _____

mean: _____

Find the median for each data set or determine the interval in which the median must fall. Then compare the medians.

10. The dot plots show Kyla's Spanish quiz scores during the 1st and 2nd semesters.

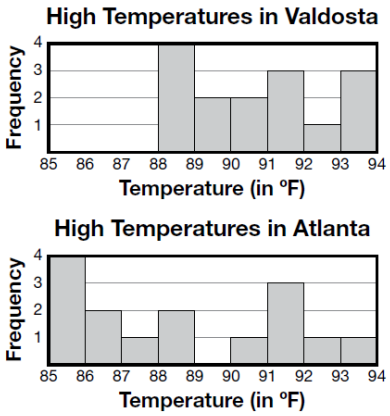


median score, 1st semester: _____

median score, 2nd semester: _____

Comparison: _____

11. The histograms show the daily high temperatures in two cities.



median high temperature, Valdosta: _____

median high temperature, Atlanta: _____

Find the mean for each data set. Then compare the means.

12. The tables show the number of ads that were sold by the actors and stage-crew members working on a school play.

Actor	Rajiv	Amy	Penny	Leonard	Adriel
Ads Sold	4	4	5	6	7

Crew Member	Tina	Ben	Ronny	Irene	Cris
Ads Sold	6	7	8	9	9

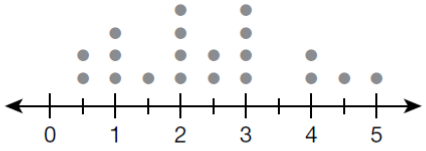
mean number sold, actors: _____

mean number sold, crew members: _____

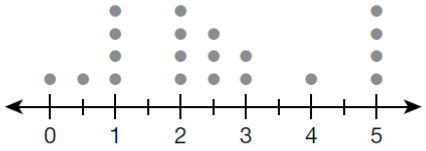
Comparison: _____

13. The dot plots show the number of hours of television watched yesterday by students from two homerooms.

Hours of Television Watched, Room 101



Hours of Television Watched, Room 102



mean number of hours, Room 101: _____

mean number of hours, Room 102: _____

Comparison: _____

---

14. Which statement accurately compares the average weight of a puppy from the 2nd litter to the average weight of a puppy from the 1st litter?

Weights of Puppies (in ounces)

1st Litter	$3\frac{1}{2}$, 4, 4, $4\frac{1}{2}$
2nd Litter	$4\frac{1}{2}$, 5, 7, $7\frac{1}{2}$

A. The average weight is about the same for both litters.

B. The average weight of a puppy from the 2nd litter is about $\frac{1}{2}$ as great.

C. The average weight of a puppy from the 2nd litter is about $1\frac{1}{2}$ times as great.

D. The average weight of a puppy from the 2nd litter is about $2\frac{1}{2}$ times as great.

15. To compare two shipments, five packages from each shipment were chosen at random and weighed. Which measure or measures of center would be best to use if you wanted to compare the weight of a typical package from each shipment?

Weights of Packages (in pounds)

1st Shipment	2, 4, 6, 8, 10
2nd Shipment	3, 3, 5, 8, 50

A. Median would be the best measure of center.

B. Mean would be the best measure of center.

C. Median and mean would both be equally good measures of center.

D. Neither the mean nor the median would be a good measure of center.

Measures of Spread

Measures of Spread describe the "diversity" of the values in a data set. Measures of spread are used to help explain whether data values are very similar or very different.

Range	• Range = Biggest # - Smallest #
Mean Absolute Deviation (MAD)	• Indicates how spread out or variable data are. • Measures how the data points in a set vary from the mean, $\bar{x}$ The formula for mean absolute deviation is: $\frac{\sum_{i=1}^{N} \left\| x_i - \bar{x} \right\|}{N}$ X_1 = data value $\bar{x}$ = mean $\sum$ = sum N = number of data values **Calculation:** - Find the mean of the set of numbers - Subtract each number in the set by the mean and take the absolute value of each new number (new number will be positive) - Find the sum of the new numbers and divide by the number of data values

Example: Calculate the MAD of this data set: 5, 8, 9, 11, 12.

Calculate the mean, $\bar{x}$.

$$\bar{x} = \frac{5 + 8 + 9 + 11 + 12}{5} = \frac{45}{5} = 9$$

Find the absolute deviation of each data point from the mean. Use a table to organize your work.

Data Point, x	Deviation from Mean, $x - \bar{x}$	Absolute Deviation from Mean, $\lvert x - \bar{x} \rvert$
5	$5 - 9 = -4$	$\lvert -4 \rvert = 4$
8	$8 - 9 = -1$	$\lvert -1 \rvert = 1$
9	$9 - 9 = 0$	$\lvert 0 \rvert = 0$
11	$11 - 9 = 2$	$\lvert 2 \rvert = 2$
12	$12 - 9 = 3$	$\lvert 3 \rvert = 3$

Calculate the mean of the absolute deviations.

$$\text{MAD} = \frac{4 + 1 + 0 + 2 + 3}{5} = \frac{10}{5} = 2$$

The mean absolute deviation is 2.

Use the data below for questions 1–4.

Heights (in inches) of Starting Players, Girls' Basketball Team	Heights (in inches) of Starting Players, Boys' Basketball Team
64, 66, 66, 68, 71	67, 67, 69, 70, 72

1. Calculate the mean and MAD of the heights of starting players for the girls' team. Use the table. Show your work.

$\bar{x}$ = _____

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
64		
66		
66		
68		
71		

MAD = _____

2. Calculate the mean and MAD of the heights of starting players for the boys' team. Use the table. Show your work.

$\bar{x}$ = _____

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
67		
67		
69		
70		
72		

MAD = _____

3. On average, which team has taller starting players? Use the means you calculated above and the dot plots on the right.

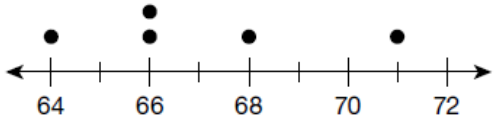
_____

4. On which team are the heights of the starting players more variable? Use the MADs you calculated above and the dot plots on the right.

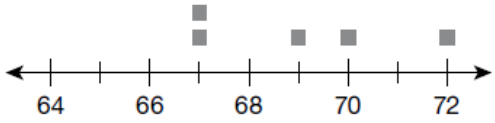
_____

Height of Basketball Players

Girls' Basketball Team



Boys' Basketball Team



Conduct the required calculations for the data sets below.

5. Set M: \$8.50, \$8.50, \$10.00

Set N: \$5.60, \$7.40, \$8.00

HINT: The mean is the sum of all the numbers in a set divided by the number of numbers in the set.

mean of Set M = _____ mean of Set N = _____

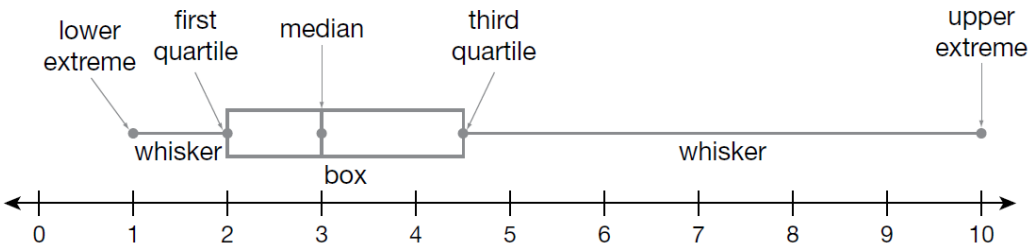
6. Set P: 54, 58, 90, 191, 142

Set Q: 96, 117, 107, 97, 118

MAD of Set P = _____ MAD of Set Q = _____

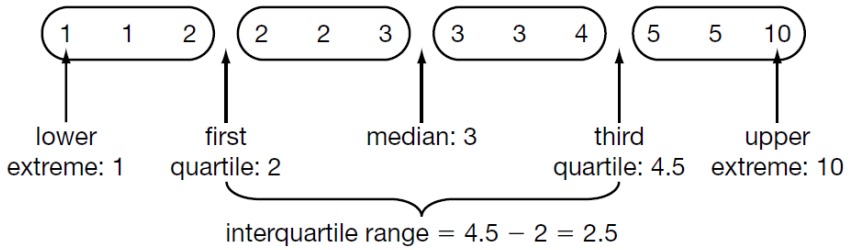
Lesson 2 - Box (and Whisker) Plots

A **box plot** (sometimes called a box-and-whisker plot) is an excellent way to display the extremes, quartiles, and median of a data set.



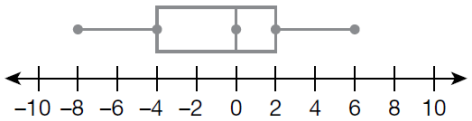
- The box contains the middle 50% of data (bounded by Q_1 and Q_2)
- Left whisker contains lower 25% of data
- Right whisker contains 25% of data

The diagram below shows the **lower extreme, upper extreme,** quartiles, and median of a data set, as well as the interquartile range. It also helps illustrate how the median and quartiles divide a data set into four discrete sets of data.



- Median: M – divided the data into 2 halves
- Lower Quartile: Q_1 – median of the lower half
- Upper Quartile: Q_3 – median of the upper half
- Interquartile Range: IQR = $Q_3 - Q_1$

Use the box plot for questions 1–5.



1. What is the median? _____

HINT The median is the middle value in the set.

2. What is the lower extreme? _____

3. What is the upper extreme? _____

4. What is the first quartile? _____

5. What is the third quartile? _____

One Variable Statistics – Univariate Data

Steps for finding the median, Q1, Q2, min, max using Technology

1. Press [data] and enter list of numbers into L1 (first column)
2. Once your data is entered into a list, Press [2nd] [data] to get [stat]
3. Highlight [1: 1-Var Stats] and then press enter
   - xDATA: L1 (should be selected)
   - FRQ: ONE (should be selected)
   - Highlight [CALC] and then press [enter]

4. You should see a list of calculations. Here are the ones that we will use:

1: n =	(number of values you entered in each list)
2: $\overline{x}$ =	(mean)
7: minx =	(minimum or lower extreme)
8: Q1 =	(Lower quartile)
9: Med =	(Median)
A: Q3 =	(Upper quartile)

Find the median (M), first quartile (Q_1), and third quartile (Q_3) of the data.

6. 1, 2, 3, 5, 7, 9, 10

M = _____

Q_1 = _____

Q_3 = _____

REMEMBER The median divides the data set into two halves.

7. 10, 12, 12, 15, 17, 19, 21, 25

M = _____

Q_1 = _____

Q_3 = _____

8. −2, −1, 2, 3, 4, 6, 7, 7, 9

M = _____

Q_1 = _____

Q_3 = _____

9. 25, 35, 40, 45, 45, 50, 60, 65, 75, 95

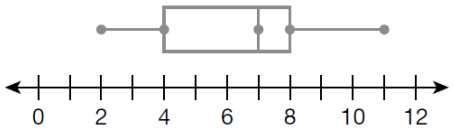
M = _____

Q_1 = _____

Q_3 = _____

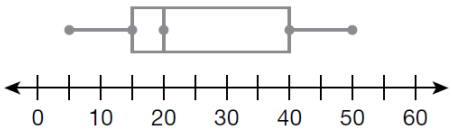
Calculate the interquartile range of the data.

16.



IQR = _____

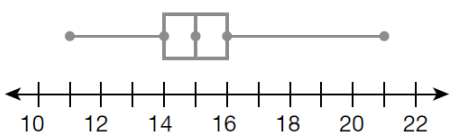
17.



IQR = _____

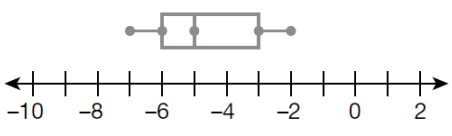
IQR = $Q_3 - Q_1$

18.



IQR = _____

19.

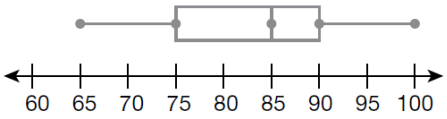


IQR = _____

Choose the best answer.

20. The box plot shows the test scores earned by students in a biology class. Which statement about the test scores is **not** true?

Biology Test Scores



A. The scores ranged from 65 to 100.

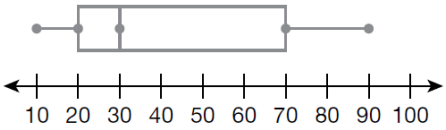
B. The median score earned was an 85.

C. 25% of students scored less than 75 points on the test.

D. 50% of students had scores that ranged from 75 to 85 points.

21. The box plot shows the prices of 20 skirts for sale at a boutique. Which statement about the prices is true?

Skirt Prices (in dollars)



A. The highest-priced skirt costs $100.

B. The median price of a skirt is $70.

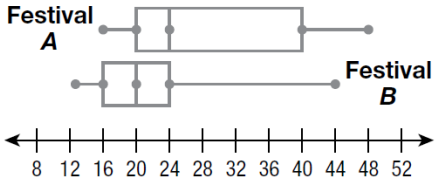
C. Half the skirts have prices that range from $20 to $70.

D. The prices of the skirts are close to the median and not very variable.

Use the box plots and information below for questions 22 and 23.

Music festival *A* and music festival *B* each had 100 volunteers. The box plots show the ages of the volunteers at each festival.

Ages of Music Festival Volunteers



22. Compare the median ages of volunteers at each festival.

_____

23. Which festival has more variability in the ages of its volunteers? Explain your answer.

_____

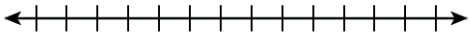
_____

Use the information below for questions 29 and 30.

Mrs. Heath visited her aunt in Nome, Alaska, for the first ten days of January 2012. She recorded the daily low temperature, in degrees Fahrenheit (°F), each day:

−27, −27, −31, −33, −34, −33, −34, −25, −29, −26

29. ORGANIZE Organize these data by displaying them in a box plot. Use the number line provided below.



30. JUSTIFY Mrs. Heath said, "The weather was very, very cold and did not vary much during the trip." Is her statement accurate? Use one or more measures of variability to justify your answer.

_____

_____

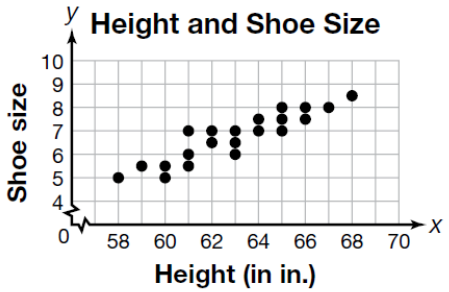
_____

_____

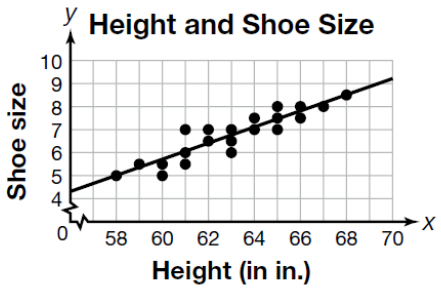
Lesson 3 - Scatter Plots & Linear Regression

Bivariate Data

- Involves the relationship between two variables.
- Can be written as a set (x, y) of ordered pairs and graphed on a coordinate plane.
- This graph is called a **scatter plot**.
- Example: The heights and show sizes of a group of students.



Look at the shape formed by the plotted points. The shape resembles a straight line. This suggests a linear relationship between the variables. You can draw a line to fit, or model, the data. The line you draw represents a linear function. If the line is a good fit, you can use the graph and the equation of the line to interpret and make predictions about the data.



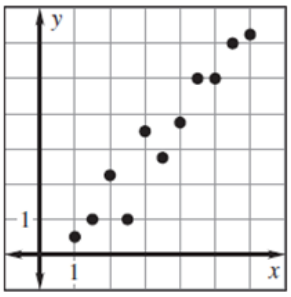
The line appears to be a good fit. The data points slant up from left to right, indicating a positive linear relationship. The line has a positive slope and is close to most data points.

1. What is the equation of the line that fits this data? (HINT: Find the slope & y-intercept)

2. Using that equation find what shoe size you would expect someone to have if she was six feet tall.

Correlation

- Scatterplots are typically used to describe relationships, called **correlations,** between two variables.
- The **correlation coefficient, r** describes how well a line fits the data.
- A **trend line** or **line of best fit** can be drawn to help determine correlation.

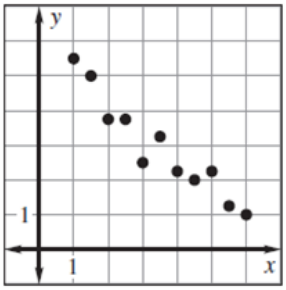


Positive Correlation

As x increases, y increases

r close to 1

Positive Slope

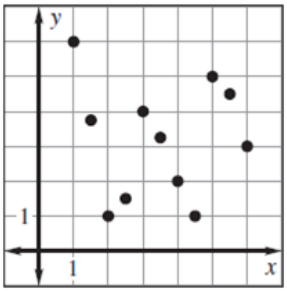


Negative Correlation

As x increase, y decreases

r close to -1

Negative Slope



No Correlation

No relationship between x and y

r close to 0

No Line

Steps for Calculating the Correlation Coefficient & Creating a Model

1. Press [data]
   - Enter 1st row of data into L1
   - Move over one column and enter 2nd row of data into L2

2. Once your data is entered into 2 lists, Press [2nd] [data] to get [stat]
3. Highlight [2: 2-Var Stats] and press [enter]
   - xDATA: L1
   - yDATA: L2
   - Highlight [CALC] and then press [enter]

4. You should see a list of calculations. Here are the ones that we will use:

   1: n = (number of values you entered in each list)

   2: $\overline{x}$ = (mean of L1)

   5: $\overline{y}$ = (mean of L2)

   (Skip down to the letters)

   D: a = (**slope** of the line of best fit)

   E: b = (**y-intercept** of the line of best fit)

   F: r = (correlation coefficient)

Equation of Line of Best Fit from Calculator

y = **a**x + **b**

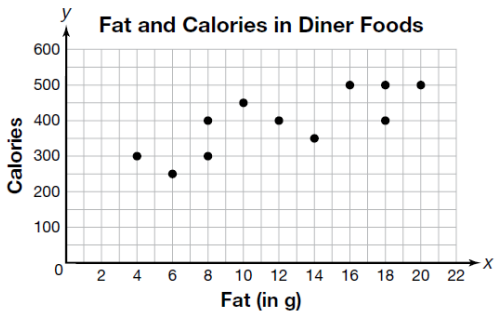
slope, m **y-intercept**

EXAMPLE For a health project, Dylan recorded the number of grams of fat and the number of calories in lunch entrees sold at his favorite diner.

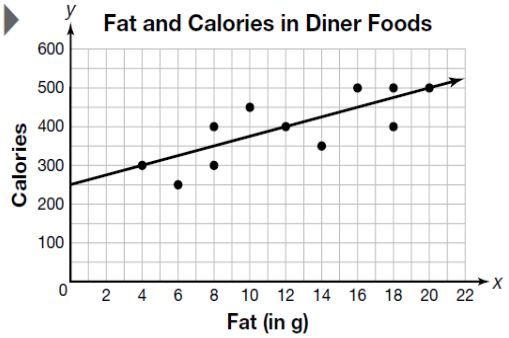
Fat (in grams)	4	6	8	8	10	12	14	16	18	18	20
Calories	300	250	300	400	450	400	350	500	400	500	500

Create a scatter plot for the data. Draw a line to fit the data. Find the equation of the line.

1 Use the ordered pairs of data items to create a scatter plot.



2 Draw a line to fit the data.



3 Write an equation for the line of fit.

The points (4, 300) and (12, 400) are on the line. Use those points to find the slope.

▶ The equation of the line is

DISCUSS

Explain what the slope of the line tells you in this context. Do the data show a positive linear relationship or a negative linear relationship?

3. Using your equation, estimate how many calories an entrée would have if it had 30 grams of fat.

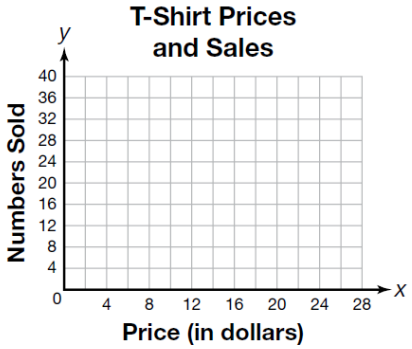
4. Using technology, find the line of BEST fit. (Use calculator steps on previous page)

5. Find the correlation coefficient, r.

The table below shows T-shirt sales data for a store one weekend.

Price, *x* (in dollars)	4	8	8	12	12	16	20	20	24	24
Number Sold, *y*	32	26	30	22	26	20	12	20	14	10

6. Create a scatter plot for the data. Then draw a line of fit for the data.

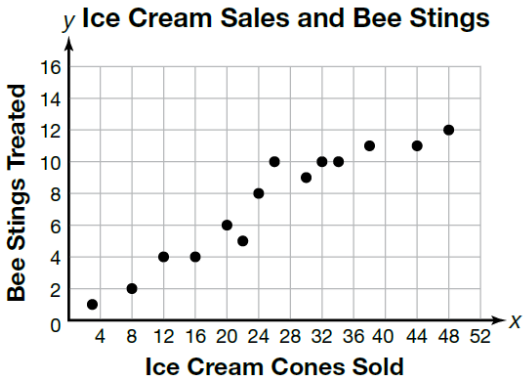


7. Find the slope of the line of fit. What does it represent in the context of this problem?

8. Using technology, find the equation for the line of fit. (Use calculator instructions)

9. Using your equation, predict how many $18 t-shirts the store could sell in a weekend.

CONCLUDE The scatter plot below shows data for the number of ice cream cones sold and the number of bee stings treated at a lake resort. Based on the data, can you conclude that eating ice cream causes bee stings? If not, what can you conclude?



Correlation vs Causation

Correlation: implies a mutual relationship between two or more things. A strong relationship between two variables could be a coincidence or caused by additional factors. Typically, correlations use the words noticed and showed.

Correlations only show relationships…they cannot be used to make conclusions!!

Causation: implies a relationship in which one action or event is the direct consequence of another (cause and effect).

Correlation	Causation
• Smoking is correlated with alcoholism *(but it doesn't cause it).* • The more ice cream consumed on a beach, the increased number of people who go in the water *(eating ice cream doesn't cause you to go in the water more).*	• The more you smoke, the chances of developing lung cancer increase. (*Does smoking cause lung cancer?)* • The less calories you eat, the more weight you lose (*Does eating less cause you to lose weight?)*

Example: Determine if the following relationships show a correlation or causation:

a. A recent study showed that college students were more likely to vote than their peers who were not in school.

b. Dr. Shaw noticed that there was more trash in the hallways after 2nd period than 1st period.

c. You hit your little sister and she cries.

d. The number of miles driven and the amount of gas used on your trip to Disneyworld.

e. The age of a child and his/her shoe size.

Lesson 4 - Two-Way (Frequency) Tables

UNDERSTAND Data can be classified as being either quantitative data or categorical data. **Quantitative data** involve numbers that usually result from measurement. Temperature, height, cost, and population are examples of quantitative data. **Categorical data** take on values that are names or labels. Gender, profession, and nationality are examples of categorical data.

When researchers collect data, they often ask more than one question. Comparing the results of those questions can reveal relationships among the data. To compare two categorical variables, you can enter the frequencies for each category into a **two-way frequency table**.

The two-way frequency table below displays the results of a survey that examined the relationship between gender and video game play. The table shows **joint frequencies** and **marginal frequencies**.

	Play Daily	Play Occasionally	Total
Boys	16	8	24
Girls	4	12	16
Total	20	20	40

Joint frequencies are in the body of the table.

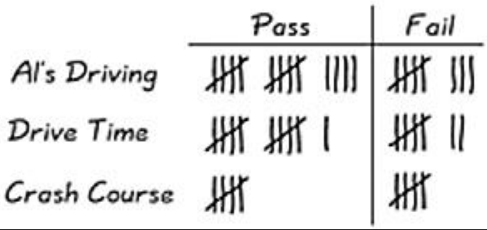
Marginal frequencies are in the "Total" row and "Total" column.

Sometimes you are less interested in the actual frequency count than in the percentage of data values that fall into each category. These percentages are the **relative frequencies**. When

Suppose the table above represents a middle school art class' responses to the question "How often do you play video games on WiiU, PS4, or the Xbox One?"

1. How many people are in the class?
2. How many girls are in the class?
3. How many boys played video games daily?
4. What percent of girls played video games daily?
5. What percent of daily players were girls?
6. Do you see an association between gender and video game use in this survey?

Francine is evaluating three driving schools. She asked 50 people who attended the schools whether they passed their driving tests on the first try.



	Pass	Fail
Al's Driving	𝍸 𝍸 𝍸 ||||	𝍸 𝍸 |||
Drive Time	𝍸 𝍸 |	𝍸 𝍸 ||
Crash Course	𝍸	𝍸

7. Create a frequency table comparing results and schools.

8. Of all students who passed, what portion went to Drive Time?

9. Which school had the most failures?

10. At which school are students least likely to fail?

A random group of high school students was surveyed. Each student was asked whether it should be mandatory for all high school students to participate in a sport. The results are partially summarized in the two-way table.

	Agree	Disagree	No Opinion	Total
Freshman	53	12	7	
Sophomore	65		2	104
Junior	18	42	12	
Senior	56			
Total		158		375

11. How many seniors were surveyed?

12. What percent of students surveyed were seniors?

13. What percent of students disagree with the mandate?

14. What percent of students disagreeing were seniors?

15. What percent of seniors disagreed with the mandate?

16. Do you see an association between grade level and opinion in this survey?