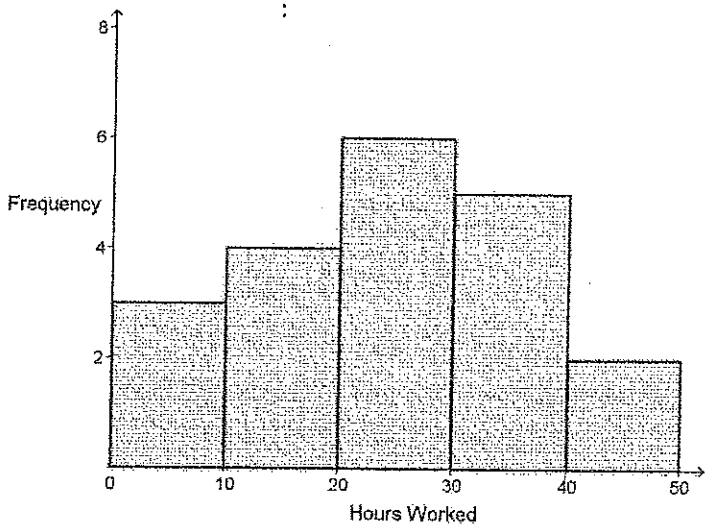
Lesson 1: Distributions and Their Shapes

Create a Histogram to Represent a Distribution

The hours worked for 20 employees are listed below:

5, 5, 8, 10, 12, 12, 15, 20, 20, 20, 22, 24, 25, 30, 30, 31, 35, 39, 40, 41

1. Create a histogram of the hours.



I need to group the data into intervals. The hours range from 5 to 41 hours. I think I will use 10 hour intervals with the first interval being 0 hours up to but not including 10 hours. Then, I need to count the number of data points in each interval to create the height of each bar.

Understand and Answer Questions about the Data

2. Would you describe your graph as symmetrical or skewed? Explain your choice.

   The graph is symmetrical. Most people work between 20 and 30 hours, and there are roughly equal numbers of people who work more or less than that amount.

Symmetrical graphs have the most frequent data points in the middle of the distribution. Skewed graphs have the most frequent data points on the left or right end of the distribution.

3. Identify the typical hours worked by the employees of this business.

Most employees work between 20 and 30 hours.

I need to identify the interval with the most entries.

Then I need to think about the real world. What businesses typically employ most people between 20 and 30 hours per week?

4. What type of business might employ people that work these types of hours? Use the histogram to justify your answer.

This is a small business with most people working less than 40 hours a week. Perhaps it is a restaurant or a coffee shop that employs part-time employees like college students.

Create a Histogram to Represent a Distribution and Answer Questions about the Data

Another company employs 50 people. The hours each employee works are listed below:

8, 8, 8, 16, 16, 16, 16, 16, 16, 16, 16, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 44, 44, 44, 48, 48

5. Create a histogram of the hours.

These hours range from 8 to 48, so I can use the same intervals as before, but there are a lot more employees, so I will need to make the scale on the vertical axis go at least to 25 because it looks like about half of the data points are 40 hours or more.

6. Would you describe your graph as symmetrical or skewed? Explain your choice.

The graph is skewed left. Most of the employees of this business work 40 or more hours per week.

7. Identify the typical hours worked by employees of this business.

Just over half of the employees work 40 or more hours per week.

I need to think of a type of business where most people would work a typical 40 hour work week. When comparing the histograms, I need to think about what is the same and what is different. Since I used the same hours intervals, it is easier to compare the distribution of the two data sets.

8. What type of business might employ people who work these types of hours? Use the histogram to justify your answer.

This is a larger business where more workers appear to work a typical 40 hour work week. Perhaps this is a small manufacturing company, an insurance agency, or a small bank where employees keep full-time, regular hours.

Compare Two Distributions and Their Graphs

9. How would you describe the differences in the two histograms?

The major differences are in the center and distribution of the data. One is symmetric with the hours distributed evenly around the center of the data, and the other is skewed with most employees working 40 or more hours per week. Each hours interval had an entry for the first company but no one at the second company worked from 30 up to 40 hours per week. The second set of data was a much larger set, so the frequencies in each interval are larger since both graphs use the same hours intervals.

Lesson 2: Describing the Center of a Distribution

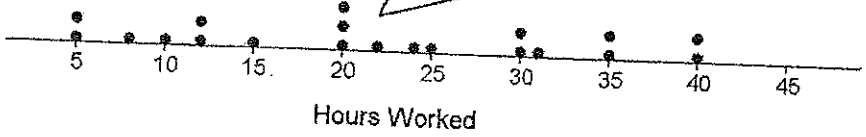
Create a Dot Plot to Represent a Distribution

The hours worked for 20 employees are listed below:

$$5, 5, 8, 10, 12, 12, 15, 20, 20, 20, 22, 24, 25, 30, 30, 31, 35, 35, 40, 40$$

1. Create a dot plot of the hours worked by the 20 employees.

I need to make a number line that includes the highest and lowest values in the data set. I need to scale the number line so I can easily plot the data values, so I will scale it by 5's. Repeated data get stacked on top of one another. Since there are 3 people that worked 20 hours, I need 3 dots at 20. I need to label the graph.



Hours Worked

Calculate the Mean and Median of a Data Set

Add the values in the data set, and divide this sum by the number of values. Since there were 20 employees, I need to divide the sum total of hours by 20.

2. What is the mean of this data set?

$$\frac{5+5+8+10+12+12+15+20+20+20+22+24+25+30+30+31+35+35+40+40}{20} = 21.95$$

The mean hours worked is 21.95.

EUREKA MATH

To find the median, the data set needs to be in order from least to greatest. Then, I need to find the middle number. Since there are 20 entries, the median will be the mean of the 10^{th} and 11^{th} value of the ordered data set.

3. What is the median of the data set?

Since there is an even number of elements in this data set, I must find the mean of the two middle numbers, 20 and 22.

$$\frac{20+22}{2}=21$$

The median is 21 hours.

4. Which numerical summary, the mean or median, is most appropriate for this data set?

Since the distribution is fairly symmetrical, either the mean or the median would be appropriate. Since the mean is slightly higher than the median, the distribution is skewed slightly to the left.

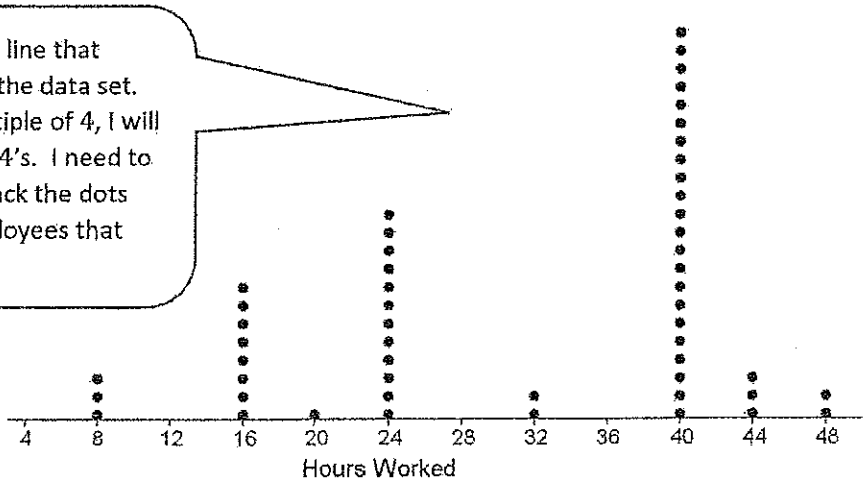
Comparing Distributions

Another company employs 53 people. The hours each employee works are listed below:

8, 8, 8, 16, 16, 16, 16, 16, 16, 16, 16, 20, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 32, 32, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 44, 44, 44, 48, 48

5. Create a dot plot of the hours worked by employees of this company.

I need to make a number line that includes all the values in the data set. Since each entry is a multiple of 4, I will scale the number line by 4's. I need to leave enough room to stack the dots since there were 22 employees that worked 40 hours.



I need to decide whether to use the mean or the median. This distribution is skewed to the left, so the median would be a better choice. This data set is already listed in order from least to greatest, so I do not need to worry about rewriting it in numerical order to find the median. Out of 53 elements, the middle one is the 27th.

6. How many hours per week is typical for employees of this company? Explain how you determined your answer.

Since this distribution is skewed to the left, the median is a better choice to describe the center of the data set. The median hours is 40.

7. Why would it be difficult to report a typical number of hours if we combined the hours employees worked at both companies?

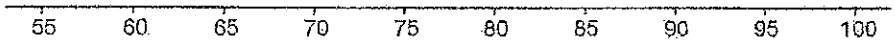
The distribution would have two points that appeared to be the center of the distribution, one around 20 to 24 hours and another around 40 hours. The two companies have very different patterns of hours worked, so combining the two may not accurately represent the trends at either company.

Lesson 3: Estimating Centers and Interpreting the Mean as a Balance Point

Create a Dot Plot to Represent a Distribution and Estimate a Balance Point

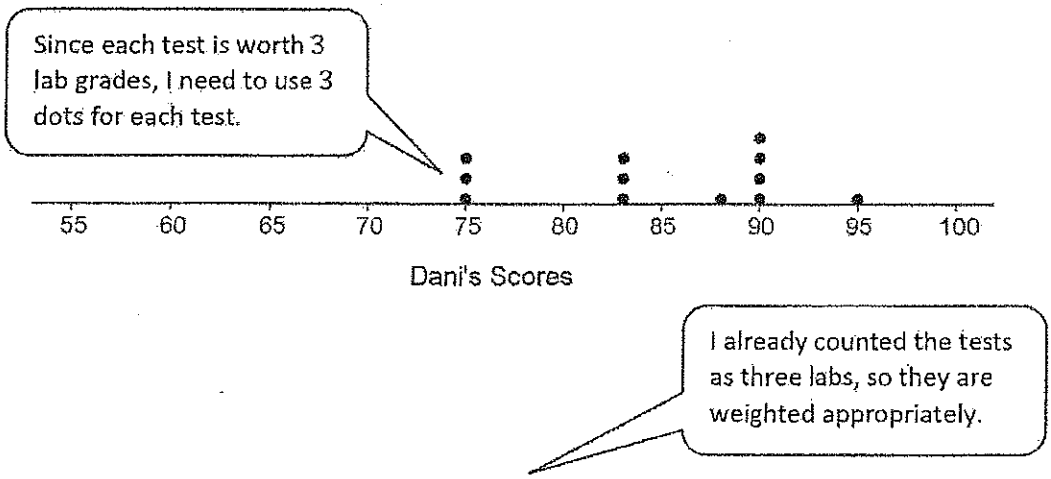
In Mr. Moreno's science class, each test is worth the same amount as three lab reports. Dani earned 75%, 83%, and 90% on her tests, and she earned 90%, 88%, and 95% on her lab reports.

Here is a number line that can be used to plot Dani's grades in science class.



Dani's Scores

1. On the number line, create a dot plot of Dani's science grades. Let one "•" symbol represent one lab report score.

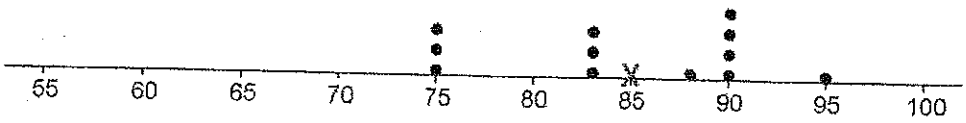


2. To be eligible for the Honors Award, Dani's weighted average in the class must be 85% or higher. Do you think Dani will get the award? Explain your answer.

She has the same number of scores above 85% as she does below 85%. She might be just underneath that average because three of her scores were 75%.

3. Place an X on the number line at a position that you think indicates the balance point of all the "•" symbols.

The X should be located so that the sums of the distances above and below the X and each data value are equal.



Dani's Scores

4. Determine the sum of the distances from the X to each "•" to the left. Determine the sum of the distance from the X to each "•" to the right. Then, explain whether or not you need to adjust your balance point based on these sums.

The distance from 85 to 75 is 10.

$$85 - 75 = 10$$

The distance from 85 to 83 is 2.

$$85 - 83 = 2$$

To find the distance, I can subtract the smaller number from the larger one or use absolute value to get a positive number.

The sum of the distances to the left of 85 is $3 \cdot 10 + 3 \cdot 2$ which is equal to 36.

The distance from 88 to 85 is 3.

$$|85 - 88| = 3$$

The distance from 90 to 85 is 5.

$$|85 - 90| = 5$$

The distance from 95 to 85 is 10.

$$|85 - 95| = 10$$

Since there were four 5's, I can multiply each 5 by 4 and then add to the other distances.

The sum of the distances to the right of 85 is $3 + 4 \cdot 5 + 10$ which is equal to 33.

The balance point should be slightly lower than 85 since the sum of the distances to the left was greater than the sum of the distances to the right.

The balance point is where the sums are equal. I have too large a sum to the left, so the point needs to be lower.

Calculate a Weighted Average and Compare Means and Balance Points

If some data points in a distribution are worth more than others, then a weighted average can be used to describe the center of the distribution. Each data point is counted according to its weight. In this case, tests are worth three lab report grades, so we can multiply each test score by three. When calculating a weighted average, you need to account for the different "weights" of each score, which is why the example below divides the total by 12 instead of 6.

5. Based on these test and lab report grades, what is Dani's weighted average?

Each test is worth 3 labs reports, so multiply those scores by 3. Then, count a total of 12 grades when calculating the mean.

$$\frac{(3 \cdot 75) + (3 \cdot 83) + (3 \cdot 90) + 88 + 90 + 95}{12} = 84.75$$

6. How does the calculated mean compare with your estimated balance point.

They are very close with the mean slightly below the estimated balance point.

Lesson 4: Summarizing Deviations from the Mean

Calculate Deviations from the Mean

The vertical jump in inches of 26 players in an NBA draft is given in the table below. (Data set from Core Math Tools, www.nctm.org)

32	33	33	34	36	36	37	37	37	38	38	38	38
38	38	38	38	38	38	39	39	39	39	40	41	43

It's quicker to write the sum using multiplication when values are repeated.

1. Calculate the mean vertical jump for these players in the NBA draft.

$$\frac{32 + 2 \cdot 33 + 34 + 2 \cdot 36 + 3 \cdot 37 + 10 \cdot 38 + 4 \cdot 39 + 40 + 41 + 43}{26} = 37.5$$

2. Calculate the deviations from the mean for these vertical jumps, and write your answers in the table below.

Vertical Jump	32	33	33	34	36	36	37	37	37	38	38	38	38
Deviation from the Mean	−5.5	−4.5	−4.5	−3.5	−1.5	−1.5	−0.5	−0.5	−0.5	0.5	0.5	0.5	0.5
Vertical Jump	38	38	38	38	38	38	39	39	39	39	40	41	43
Deviation from the Mean	0.5	0.5	0.5	0.5	0.5	0.5	1.5	1.5	1.5	1.5	2.5	3.5	5.5

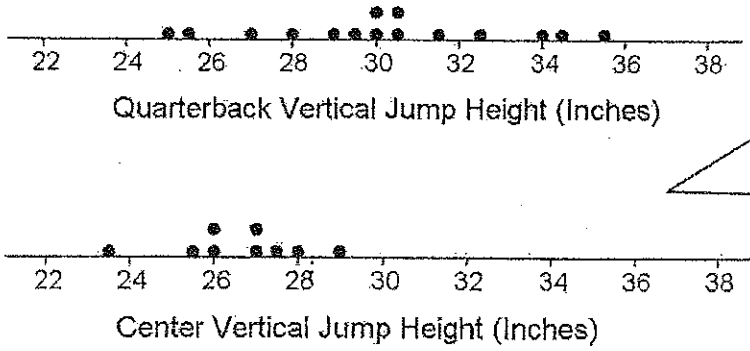
3. Write an expression for the deviation from the mean for a jump height of 32 inches.

$32 - 37.5$

We don't use absolute value for deviation from the mean, so the deviations will be either positive, negative, or zero depending on whether the value is above, below, or at the mean.

Compare the Variability of Two Distributions Considering Deviations from the Mean

The vertical jumps of the quarterbacks and the centers from a recent NFL combine are shown on the dot plots below.



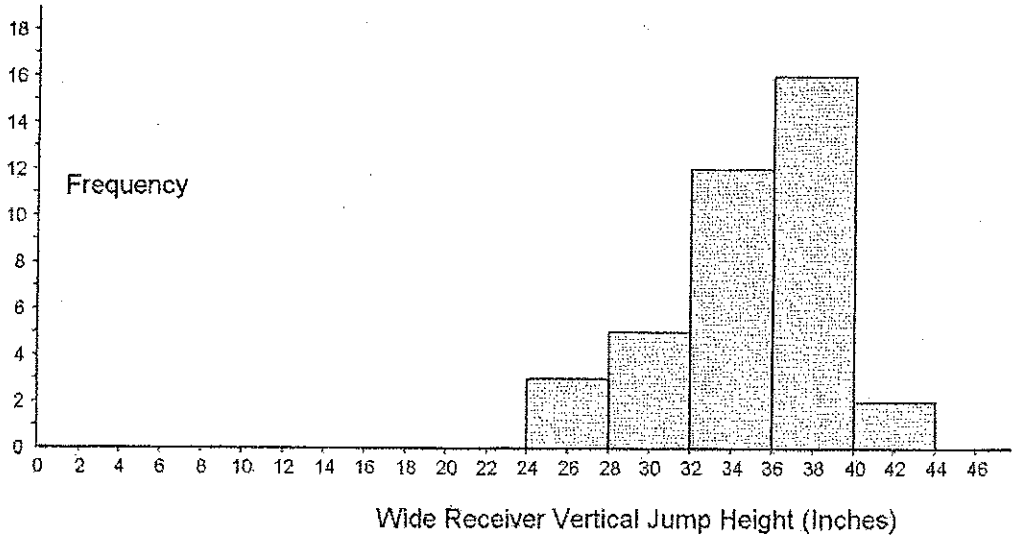
I need to look at how far away each data point would be from the mean. Since the distributions are fairly symmetrical, the mean will be near the middle.

4. Based on the data, which position, quarterback or center, has the greatest deviation from the mean?

The quarterback position has the greater deviation from the mean because the jump heights range from 25 to nearly 36 inches. For the centers, the jump heights range from around 23 to 29 inches.

Estimate Mean and Deviation from the Mean Given a Histogram

The vertical jumps of the wide receivers from a recent NFL combine are shown on the histogram below.



5. How many wide receivers jumped around 34 inches?

Twelve wide receivers jumped around 34 inches.

In a histogram, the height of the bar represents the number of values in each interval.

6. How many wide receivers participated in the combine in total?

The height of the first bar is 3, which represents 3 players in that interval. The sum of the frequencies in each interval is $3 + 5 + 12 + 16 + 2$ which is equal to 38. There were 38 receivers in the combine.

7. Suppose the three players represented by the bar centered at 26 inches each jumped exactly 26 inches and the players in the next bar each jumped exactly 30 inches, and so on. If you were to add up all the jump heights, what result would you get?

$3 \cdot 26 + 5 \cdot 30 + 12 \cdot 34 + 16 \cdot 38 + 2 \cdot 42 = 1328$

I can multiply the middle value in each bar by the frequency and add.

8. What is the mean jump height for the wide receivers?

$\frac{1328}{38} \approx 34.9$

The mean jump height is approximately 34.9 inches.

9. What is a typical deviation from the mean for this data set? Explain your reasoning.

A typical deviation from the mean would be 4 to 6 inches. Each bar represents a 4-inch interval of jumps. Most jumps are within 4 inches of the mean, but a few are outside of that range, so the typical deviation from the mean should be a little bit greater than 4 inches.

I need to think about how far each value of the data set would be from 34.9 inches. Based on the bar heights (frequency), I can see that almost all the heights were within 4 to 6 inches of the mean.

Lesson 5: Measuring Variability for Symmetrical Distributions

Calculate the Standard Deviation of a Data Set

1. Ten of the members of a high school boys' basketball team were asked how many hours they studied in a typical week. Their responses (in hours) are shown in the table below.

Number of Hours Studied	20	12	9	6	13	10	14	11	11	12
Deviation from the Mean	8.2	0.2	−2.8	−5.8	1.2	−1.8	2.2	−0.8	−0.8	0.2
Squared Deviation from the Mean	67.24	0.04	7.84	33.64	1.44	3.24	4.84	0.64	0.64	0.04

a. Calculate the mean study time for data set.

$$\frac{20+12+9+6+13+10+14+11+11+12}{10} = 11.8$$

Find the sum of the data values, and divide by the total number of players, which is 10.

b. Calculate the deviations from the mean, and write your answers in the second row of the table.

First entry: $20 - 11.8 = 8.2$

Second entry: $12 - 11.8 = 0.2$

Third entry: $9 - 11.8 = -2.8$

I need to subtract the mean from each value in the data set.

Continue these calculations; the remaining values are in the second row of the table above.

c. Square the deviations from the mean, and write them in the third row of the table.

First entry: $8.2^2 = 8.2 \cdot 8.2 = 67.24$

Second entry: $0.2^2 = 0.2 \cdot 0.2 = 0.04$

Squaring a number means to multiply that number by itself.

Continue these calculations; the remaining values are in the third row of the table above.

d. Find the sum of the square deviations.

The sum means to add all of the numbers.

$$67.24 + 0.04 + 7.84 + 33.64 + 1.44 + 3.24 + 4.84 + 0.64 + 0.64 + 0.04 = 119.6$$

e. What is the value of n for this data set?
Divide the sum of the squared deviations by $n-1$.

n stands for the number of values in the data set. We divide by one less than that number.

$n = 10$

$\frac{119.6}{9} = 13.2\overline{8}$

f. Take the square root of your answer to part (e).
Round your answer to the nearest hundredth.

The nearest hundredth means two places after the decimal. Use the "≈" symbol when rounding. I will need to use a calculator to compute the square root.

$\sqrt{13.2\overline{8}} \approx 3.65$

2. Find the standard deviation of the following data set: $3, 5, 10, 23, 23, 30, 34, 40$.

Find the mean.

$$Mean = \frac{3+5+10+23+23+30+34+40}{8} = 21$$

Find the deviations from the mean.

$3 - 21 = -18$

$5 - 21 = -16$

$10 - 21 = -11$

$23 - 21 = 2$

$23 - 21 = 2$

$30 - 21 = 9$

$34 - 21 = 13$

$40 - 21 = 19$

I can follow the steps below to find standard deviation:

- Find the mean.
- Find the deviations from the mean.
- Square the deviations.
- Sum the squares of the deviations.
- Divide by $n-1$.
- Take the square root.

Sum the square of the deviations.

$$(-18)^2 + (-16)^2 + (-11)^2 + 2^2 + 2^2 + 9^2 + 13^2 + 19^2 = 1320$$

The sum of the squares is divided by $n-1$.

$$\frac{1320}{7} \approx 188.57$$

The standard deviation is the square root of the number.

$$\sqrt{188.57} \approx 13.73$$

Standard deviation is greater when a data set has more variability. It is a measure of the spread of the data.

3. Which data set, the one from Exercise 1 or the one from Exercise 2, has the greatest spread (variability)?

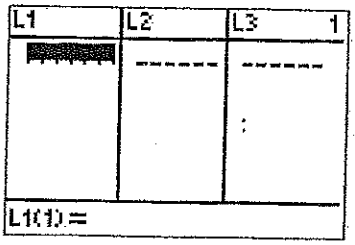
The one from Exercise 2 because it had the larger standard deviation.

Lesson 6: Interpreting Standard Deviation

Calculate the Mean and Standard Deviation Using a TI-83 or TI-84 Calculator

Instructions may vary based on the type of calculator or software used. The instructions below are based on a Texas Instruments TI-83 or TI-84 calculator using data stored in L1.

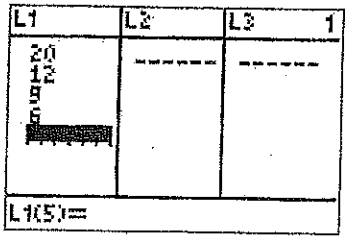
1. From the home screen, press STAT, ENTER to access the stat editor.



After pressing STAT and ENTER, I see this screen, and I can start typing the data values in the L1 column.

2. If there are already numbers in L1, clear the data from L1 by moving the cursor to "L1" and pressing CLEAR, ENTER.

3. Move the cursor to the first entry of L1, type the first data value, and press ENTER. Continue entering the remaining data values to L1 in the same way.

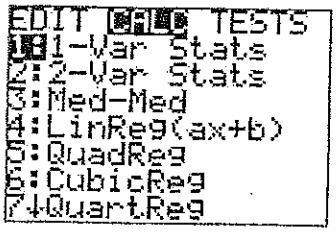


This is what the data set {20,12,9,6} would look like after I enter it in L1.

4. Press 2ND, QUIT to return to the home screen.

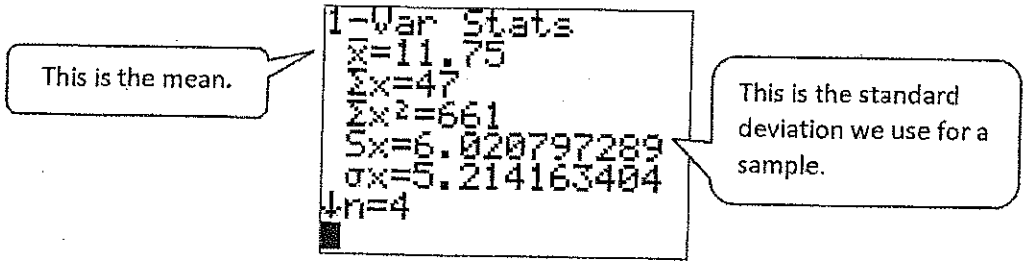
This step is optional.

5. Press STAT, select CALC, select 1-Var Stats, and press ENTER.

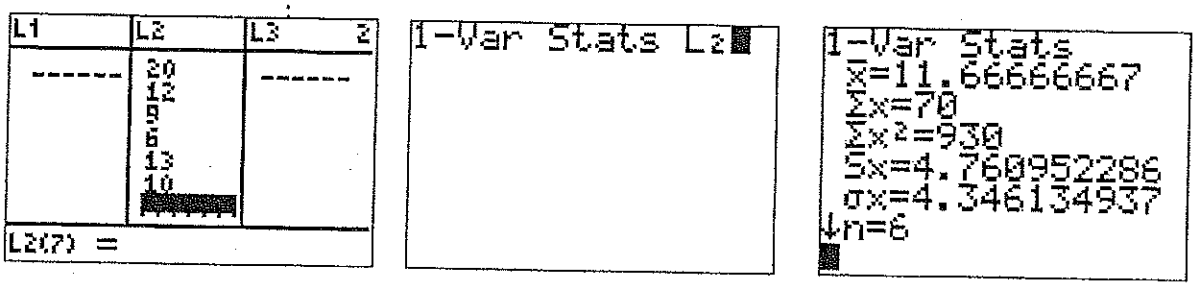


This step assumes I have already entered data specifically into L1.

6. The screen should now show summary statistics for your data set. The mean is the $\bar{x}$ value, and the standard deviation for a sample is the s_x value.



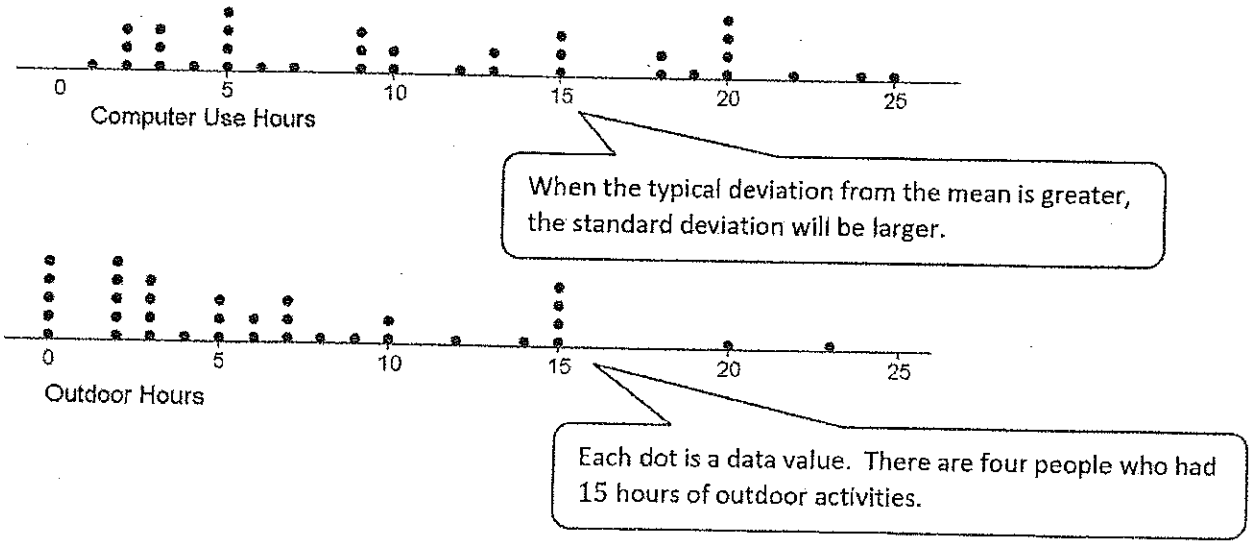
If data is stored in another list, it will need to be referred to after selecting 1-Var Stats in Step 5. For example, if data was entered in L2:

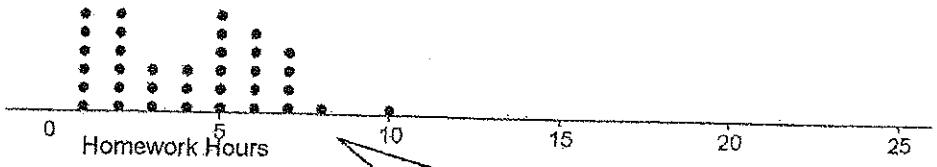


Press STAT, select CALC, select 1-Var Stats, and then refer to L2. This is done by pressing 2ND, L2 (i.e., "2ND" and then the "2" key). The screen will display 1-Var Stats L2. Then, press ENTER.

Calculating the Standard Deviation Given a Box Plot

A random sample of 35 ninth graders reported they spent the following number of hours each week on the following activities as shown in the dot plots below.





When more of the data values are closer to the mean, the standard deviation will be smaller.

1. Which of the three activities has the smallest standard deviation? Which had the largest? Justify your answer.

   Homework has the smallest standard deviation because the times are clustered around the mean (center) of the data. Either Computer Use or Outdoor Activities will have the largest standard deviation because the times are more spread out from the mean (center) of the data.

2. Estimate the mean and standard deviation of computer use hours.

   The data is pretty evenly distributed, so the mean will be near the center. The mean is approximately 11. The standard deviation is a measure of typical deviation from the mean, which appears to be around 6 or 7.

3. Use a calculator to determine the mean and standard deviation of each variable, and record the information in the table below. Round answers to the nearest hundredth.

First, I need to enter each data set into a list on the calculator.

	Homework	Outdoor Activities	Computer Use
Mean	4.14	6.86	11.11
Standard Deviation	2.39	6.03	7.30

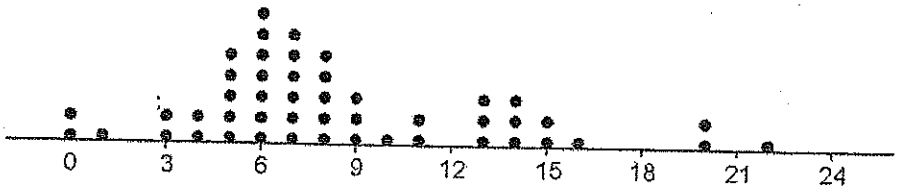
On the calculator, $\bar{x}$ is the mean, and s_x is the standard deviation.

Then, I need to press STAT, CALC, 1-Var Stats, and ENTER for data stored in L1.

For data entered in another list, I need to type that list name after 1-Var Stats before pressing ENTER.

Lesson 7: Measuring Variability for Skewed Distributions (Interquartile Range)

The dot plot displays the number of hours a sample of 48 ninth graders spend playing video games during one week.

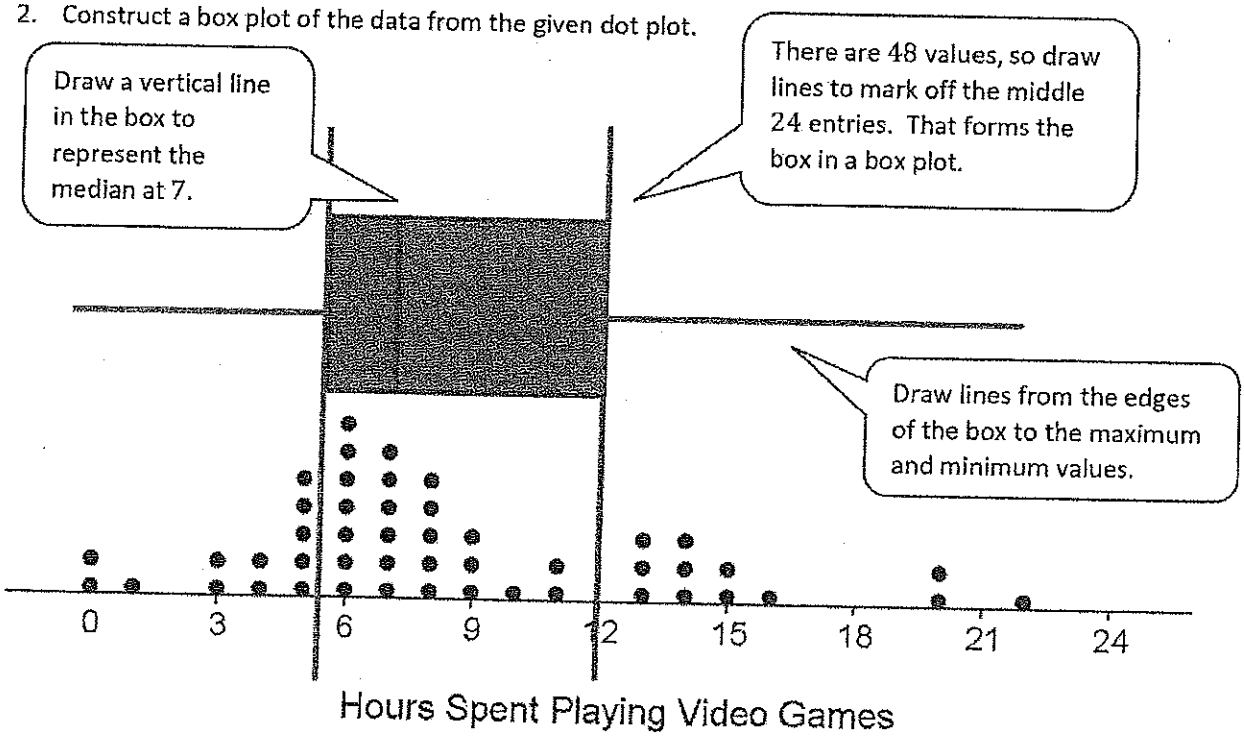


Hours Spent Playing Video Games

1. Identify the data set shown in the dot plot as skewed to the left or skewed to the right.

   The data set is skewed right because it spreads out longer on the right side.

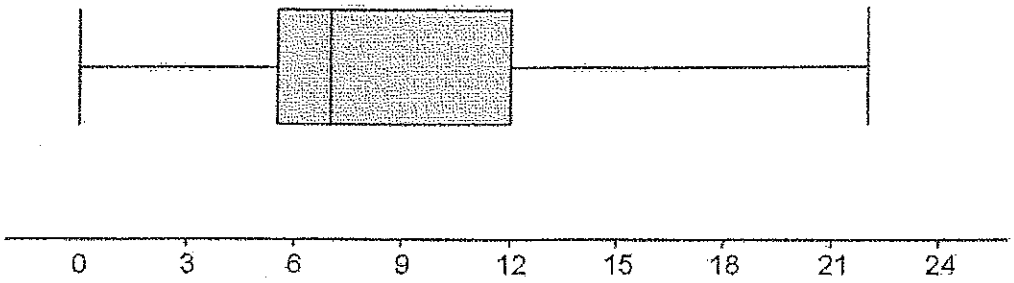
2. Construct a box plot of the data from the given dot plot.



Hours Spent Playing Video Games

EUREKA MATH

The final box plot should look like this.



Hours Spent Playing Video Games

Calculate the Five-Number Summary for a Data Set and the Interquartile Range

3. What is the five-number summary for the data set?

$$0, 1, 2, 3, 3, 6, 6, 6, 7, 7, 9, 12, 20, 22$$

Minimum Value: 0

Lower Quartile or Q1: 3

Median: 6

Upper Quartile or Q3: 9

Maximum Value: 22

There are 14 data values. Once the set is in order, I need to cut the data set into quarters. The middle is between the 7th and 8th values, and the quartiles are the 4th and 11th values.

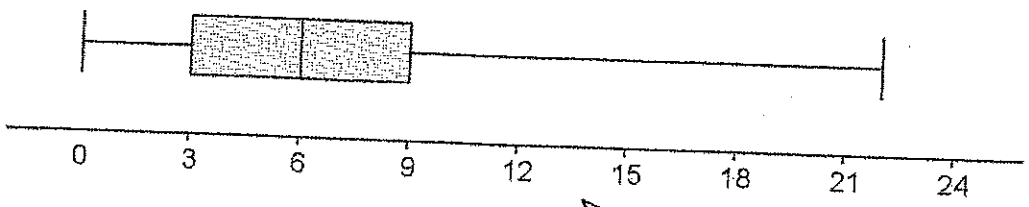
The interquartile range (IQR) is the difference between the third and first quartiles.

4. What is the interquartile range?

The lower quartile is 3. *The upper quartile is* 9. *The difference is* $9 - 3 = 6$. *The interquartile range is* 6.

Construct a Box Plot for a Data Set Using the Five-Number Summary

5. Construct a box plot of the data set from Problem 3.



The data ranges from 0 to 22, so I will make the number line go slightly above and slightly below these numbers.

The lower quartile is 3, and the upper quartile is 9. These values mark the sides of the box. Then, draw a line at the median, 6.

Finally, extend a horizontal line from the sides of the box to the minimum value 0 and the maximum value 22.

6. Identify any outliers in the data set from Problem 3.

The interquartile range is 6. Next, multiply this number by 1.5.

$$1.5 \cdot 6 = 9$$

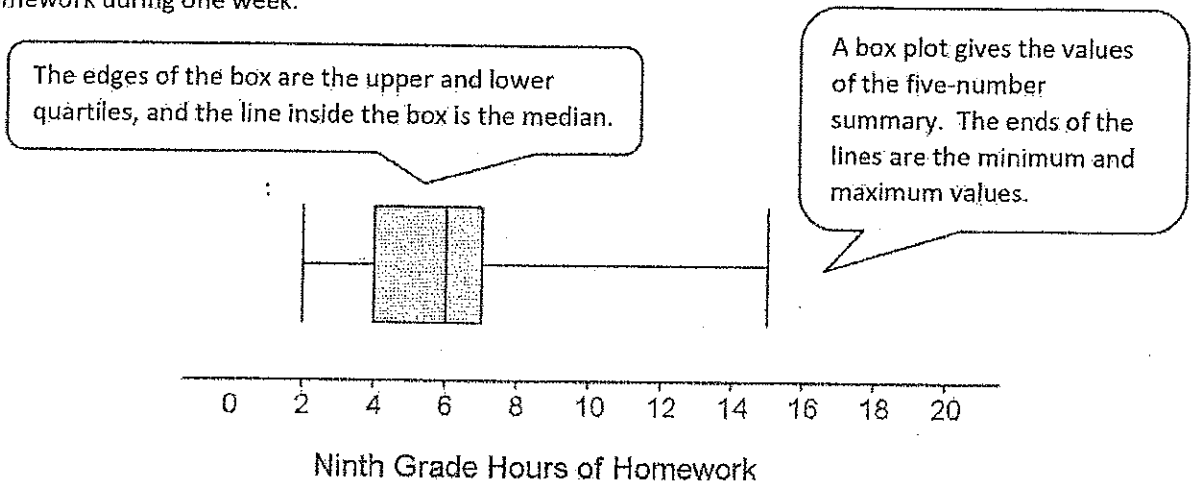
Any data value that is more than 9 away from the upper or lower quartile is considered an outlier. Thus, both 20 and 22 are outliers in this data set since $9 + 9 = 18$.

EUREKA MATH

Lesson 8: Comparing Distributions

Construct a Possible Dot Plot from a Box Plot

The box plot displays the number of hours a random sample of 24 ninth graders at a high school spent doing homework during one week.



1. Construct a possible dot plot of the sample of 24 ninth graders.
   a. What is the five-number summary for this data set?

      Minimum value: 2

      Lower Quartile (Q1): 4

      Median: 6

      Upper Quartile (Q3): 7

      Maximum Value: 15

   b. How many data values are located between the upper and lower quartiles?

      Half of the data set lies between the upper and lower quartiles. There must be 12 data values.

   c. How many data values will be below the lower quartile, and how many will be above the upper quartile?

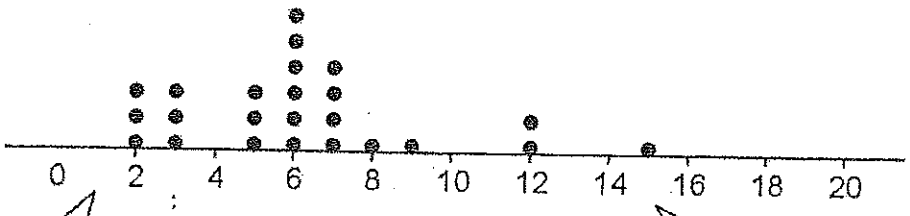
      There must be 6 values below the lower quartile and 6 values above the upper quartile.

d. List your possible data set below.

2, 2, 2, 3, 3, 3, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 9, 12, 12, 15

The median must be 6. I need to have the minimum value, 2, and the maximum value, 15, in my data set.

e. Create a dot plot using the sample you created.

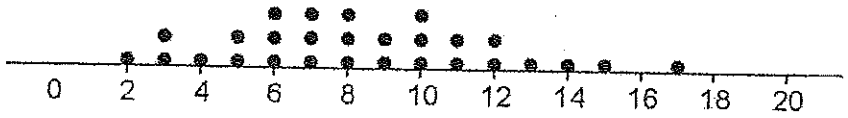


My possible data set needs to have $Q1 = 3$ and $Q3 = 9$.

This is not the only possible data set that could be represented by the given box plot.

Construct a Box Plot from a Dot Plot

The dot plot below shows the hours of homework for a random sample of 28 juniors from the same high school during one week.



Eleventh Grade Hours of Homework

2. Construct a box plot from this dot plot.

   a. What is the five-number summary for this data set?

   Minimum value: 2

   Lower Quartile: 6

   Median: 8

   Upper Quartile: 11

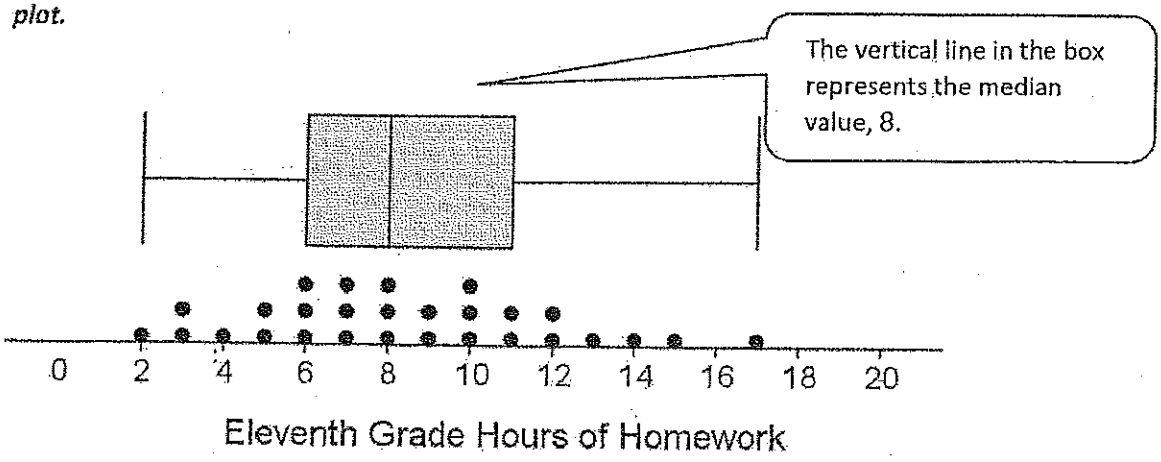
   Maximum Value: 17

The lower quartile is between the 7th and 8th data values. Both of these values are 6, so the lower quartile

EUREKA MATH

b. Create a box plot using the information in the five-number summary.

The box plot is superimposed on the dot plot to illustrate how to construct the box plot from a dot plot.



Compare Two Distributions

The median describes a typical value for a data set. The interquartile range (IQR) describes the variability of a data set.

3. What is a typical amount of homework for ninth graders and eleventh graders?

The median for ninth graders is 6 hours and for eleventh graders it is 8 hours.

4. Do the two data sets have a similar variability? Use the inner quartile range to support your answer.

The IQR is the difference between the upper and lower quartiles. The ninth grade IQR is 3 hours, and the eleventh grade IQR is 5 hours. The data sets have different variability. Eleventh graders have more variety in the number of hours spent doing homework than ninth graders do because they have the greater IQR.

5. Why might eleventh graders typically spend more time on homework than ninth graders?

Perhaps they are taking more challenging courses or have more academic courses that require homework.

6. Why might the hours that eleventh graders spend on homework have more variability than the hours spent by ninth graders?

Most ninth graders could be taking similar classes. By the time you are in eleventh grade, you have more options for courses or career paths such as advanced placement, vocational, or regular courses, so homework hours could vary more.

Lesson 9: Summarizing Bivariate Categorical Data

Survey Design

A random sample of 50 ninth graders were surveyed regarding their favorite type of music. Twenty-nine of the students in the sample were females.

- Eight students liked country, and 6 of those were females.
- Eight students liked rock, and 5 of those were males.
- Only 4 females and no males liked pop music.
- Ten females and 6 males liked rap/hip hop.
- Five students liked techno/electronica, and 2 of those were females
- Four females and 5 males preferred other types of music.

Some of the numbers here are represented with symbols and some with words.

1. What questions might have been asked to gather this information?

   What is your gender? What is your favorite type of music?

There are two types of data in the bulleted list: gender and types of music. So my questions need to ask about that.

How could you best randomly survey ninth graders at your high school about their favorite type of music?

Answers will vary. Sample response: Randomly select five students from each ninth grade English class since all ninth graders take an English class.

I need to think of a way that doesn't bias the results but provides a fairly easy way to get answers to my questions.

2. Would the results of a random survey of ninth graders at your high school be representative of all ninth graders in the United States? Explain your reasoning.

   No because this would not account for differences in geographic locations.

This answer should be no but student reasons will vary. Reasons should point out differences due to demographics and geographic location not necessarily representing the total U.S. population of ninth graders.

EUREKA MATH

Summarize Bivariate Categorical Data in a Two-Way Frequency Table

3. Complete a two-way frequency table using the survey results from the 50 ninth graders.

The values of the favorite type of music categorical variable are the different types of music in the top row including the other value.

Since there were 29 females out of 50 total, there must be 21 males. The entry in the lower right cell is always the total number surveyed.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Female	6	3	4	10	2	4	29
Male	2	5	0	6	3	5	21
Total	8	8	4	16	5	9	50

If 6 out of the 8 people that liked country were girls, then the other two have to be boys. The 6 and 2 are called joint frequencies.

The marginal frequencies in the bottom row and right-most column are the total number of responses for each value of the categorical variable.

The marginal frequencies should always add up to the total surveyed.

4. Do you think there is a difference in the responses of males and females? Explain your answer.

Answers will vary.

Making comparisons of the joint frequencies is tricky because the numbers of males and females are not equal.

Lesson 10: Summarizing Bivariate Categorical Data with Relative Frequencies

Construct a Relative Frequency Table and Interpret the Results

Consider the two-way frequency table for a random sample of 50 ninth graders surveyed regarding their favorite type of music.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	6	3	4	10	2	4	29
Males	2	5	0	6	3	5	21
Total	8	8	4	16	5	9	50

1. Calculate the relative frequencies for each of the cells to the nearest thousandth.

Three places after the decimal point

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	$\frac{6}{50} = 0.120$	$\frac{3}{50} = 0.060$	$\frac{4}{50} = 0.080$	$\frac{10}{50} = 0.200$	$\frac{2}{50} = 0.040$	$\frac{4}{50} = 0.080$	$\frac{29}{50} = 0.580$
Males	$\frac{2}{50} = 0.040$	$\frac{5}{50} = 0.100$	$\frac{0}{50} = 0.000$	$\frac{6}{50} = 0.120$	$\frac{3}{50} = 0.060$	$\frac{5}{50} = 0.100$	$\frac{21}{50} = 0.420$
Total	$\frac{8}{50} = 0.160$	$\frac{8}{50} = 0.160$	$\frac{4}{50} = 0.080$	$\frac{16}{50} = 0.320$	$\frac{5}{50} = 0.100$	$\frac{9}{50} = 0.180$	$\frac{50}{50} = 1.000$

I need to divide each count by the total surveyed and write my answer as a decimal.

This means that 12% of the people surveyed were boys that liked rap. It doesn't mean that 12% of boys liked rap. To convert a decimal to a percent, I need to think of it using hundredths.

$$0.120 = \frac{120}{1000} = \frac{12}{100} = 12\%$$

2. What is the relative frequency of students whose favorite music is rap/hip hop?

   The relative frequency is 0.320 or 32%. This is the relative frequency for the cell that corresponds to the total number of students whose favorite music was rap/hip hop.

3. What is the relative frequency of males whose favorite music is rap/hip hop?

   The relative frequency is 0.120 or 12%. This is the relative frequency for the cell that corresponds to the total number of males whose favorite music was rap/hip hop.

4. Why might someone question whether or not the students who completed the survey were selected at random? Explain your answer.

   You would expect to see equal numbers of males and females. Nearly 60% of those surveyed were females.

   If the survey values are very different from the population, then the survey might not be random.

5. If another student was selected at random from this school, do you think their favorite type of music would be pop? Explain your answer.

   No, looking at the relative frequencies in the last row, we can see that only 8% of students reported pop as their favorite type of music.

   Survey results can be used to make predictions about a population.

Lesson 11: Conditional Relative Frequencies and Association

Construct a Row Conditional Relative Frequency Table and Interpret the Results

Consider the two-way frequency table for a random sample of 50 ninth graders surveyed regarding their favorite type of music.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	6	3	4	10	2	4	29
Males	2	5	0	6	3	5	21
Total	8	8	4	16	5	9	50

The word *row* indicates that I need to divide each frequency count in a given row by the row total.

1. Construct a row conditional relative frequency table for this data. Give answers to the nearest thousandth.

The first row total is 29. The frequency count in the first cell is 6. The row relative frequency for females whose favorite music is country rounded to the nearest thousandth would be

$$\frac{6}{29} \approx 0.207.$$

The row relative frequency for males whose favorite music is country would be $\frac{2}{21} \approx 0.095$.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	$\frac{6}{29} \approx 0.207$	$\frac{3}{29} \approx 0.103$	$\frac{4}{29} \approx 0.138$	$\frac{10}{29} \approx 0.345$	$\frac{2}{29} \approx 0.069$	$\frac{4}{29} \approx 0.138$	$\frac{29}{29} = 1.000$
Males	$\frac{2}{21} \approx 0.095$	$\frac{5}{21} \approx 0.238$	$\frac{0}{21} = 0.000$	$\frac{6}{21} \approx 0.286$	$\frac{3}{21} \approx 0.143$	$\frac{5}{21} \approx 0.238$	$\frac{21}{21} = 1.000$
Total	$\frac{8}{50} = 0.160$	$\frac{8}{50} = 0.160$	$\frac{4}{50} = 0.080$	$\frac{16}{50} = 0.320$	$\frac{5}{50} = 0.100$	$\frac{9}{50} = 0.180$	$\frac{50}{50} = 1.000$

This means that approximately 28.6% of the boys surveyed indicated that their favorite music was rap/hip hop. It does not mean that approximately 28.6% of those that liked rap/hip hop were boys.

2. For what types of music are the row conditional relative frequencies for females and males very different?

   They were fairly different for all categories. The most similar was rap/hip hop, which was the most popular type of music for males and females.

3. If Pedro, a ninth grade male at this school, completed the favorite type of music survey, what would you predict was his response?

   He would probably like either rap/hip hop, rock, or other.

   I need to think about which entries were greatest in each row.

4. If Ali, a ninth grade female at this school, completed the favorite type of music survey, what would you predict was her response?

   She would most likely indicate her favorite type of music was rap/hip hop or maybe country.

5. Is it fair to say that males and females equally prefer *other* types of music since they had nearly equal frequency counts?

   No. The row conditional relative frequencies are different.

6. Do you think there is an association between gender and favorite type of music for ninth graders at this school? Explain.

   While the survey revealed differences between the genders, the differences were not that large, and the survey only included 50 students. We cannot say there is strong evidence for an association.

The word *column* indicates that I need to divide each frequency count in a given column by the column total.

Construct a Column Conditional Relative Frequency Table and Interpret the Results

7. Construct a column conditional relative frequency table for this data. Give answers to the nearest thousandth.

The first column total is 8. The frequency count in the first cell is 6. The column relative frequency for females whose favorite music is country rounded to the nearest thousandth is

$$\frac{6}{8} = 0.750.$$

The column relative frequency for males whose favorite music is country would be $\frac{2}{8} = 0.250$.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	$\frac{6}{8} = 0.750$	$\frac{3}{8} = 0.375$	$\frac{4}{4} = 1.000$	$\frac{10}{16} = 0.625$	$\frac{2}{5} = 0.400$	$\frac{4}{9} \approx 0.444$	$\frac{29}{50} = 0.580$
Males	$\frac{2}{8} = 0.250$	$\frac{5}{8} = 0.625$	$\frac{0}{4} = 0.000$	$\frac{6}{16} = 0.375$	$\frac{3}{5} = 0.600$	$\frac{5}{9} \approx 0.556$	$\frac{21}{50} = 0.420$
Total	$\frac{8}{8} = 1.000$	$\frac{8}{8} = 1.000$	$\frac{4}{4} = 1.000$	$\frac{16}{16} = 1.000$	$\frac{5}{5} = 1.000$	$\frac{9}{9} = 1.000$	$\frac{50}{50} = 1.000$

This means that 37.5% of those surveyed that liked rap/hip hop were boys. It does not mean that 37.5% of boys liked rap/hip hop.

8. If you wanted to know the relative frequency of females surveyed whose favorite music was country, would you use a row conditional relative frequency or a column conditional relative frequency?

I would use a row conditional relative frequency.

The category is females, and the condition is country music. That category is in a row.

9. If you wanted to know the relative frequency of students who liked rap/hip hop that were males, would you use a row conditional relative frequency or a column conditional relative frequency?

I would use a column conditional relative frequency.

The category is rap/hip hop and the condition is male. That category is in a column.

EUREKA MATH