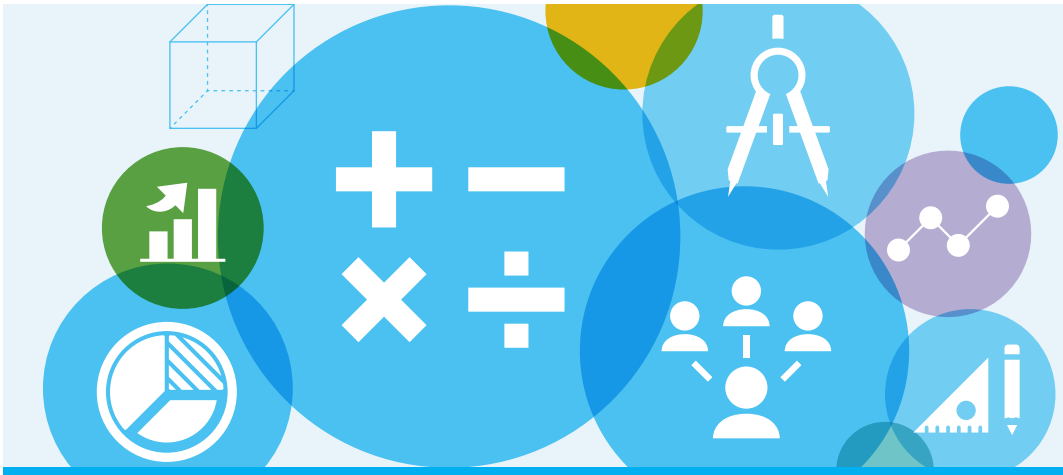


STUDENT VERSION

CENSUS IN COUNTIES — DESCRIBING AND COMPARING HISTOGRAMS TO UNDERSTAND AMERICAN LIFE

Activity Items

There are no separate items for this activity.

Student Learning Objectives

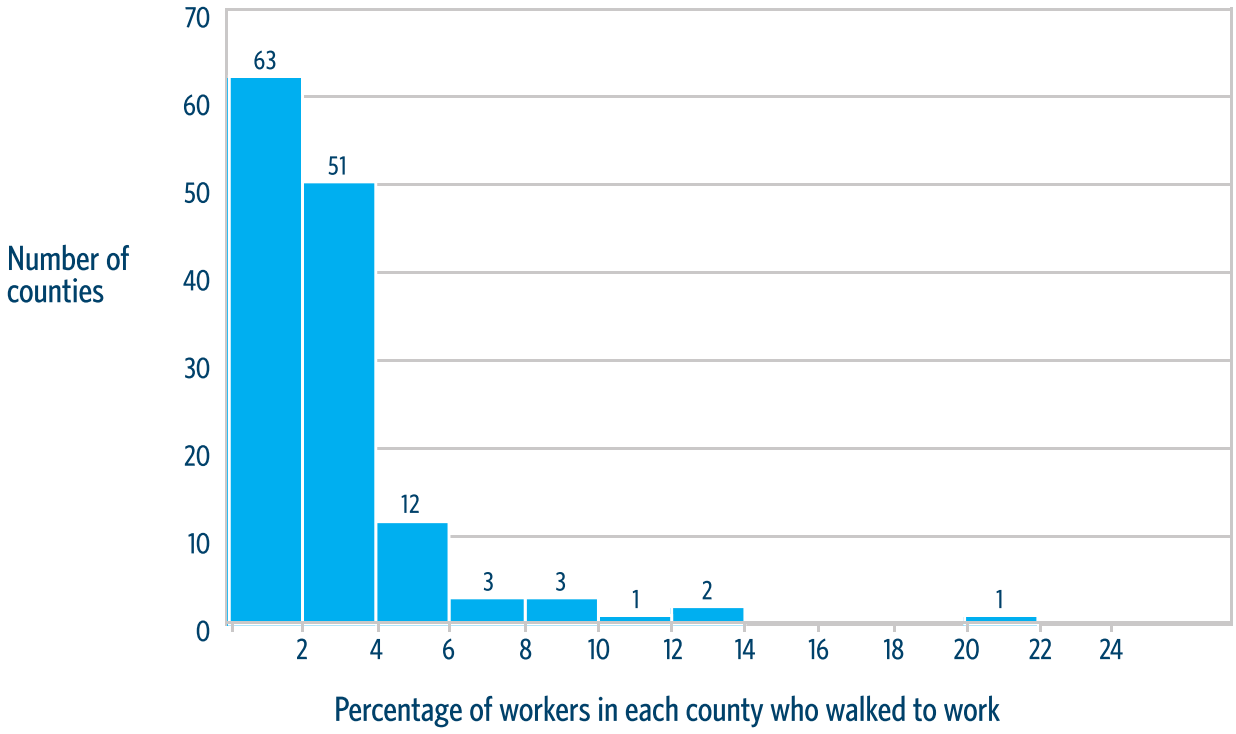
- I will be able to analyze data and describe data distributions.
- I will be able to understand skewness and how it affects the mean and median.
- I will be able to determine the direction of skewness from a distribution's mean and median.
- I will be able to understand the impact of outliers on the standard deviation and interquartile range.

Statistical measures such as median, mean, range, standard deviation (SD), and interquartile range (IQR) provide useful and interesting snapshots of data sets. In this activity, you will analyze county-level data from the U.S. Census Bureau.

NAME:_____ DATE:_____

Part 1 – Examine Histograms

1. The following graph shows the percentage of workers (16 and older) who walked to work in 2014 in each of the 136 U.S. counties surveyed. The first bar of the histogram, for example, shows that, in 63 of the counties, fewer than 2 percent of workers walked to work. Refer to this graph to answer the questions that follow.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

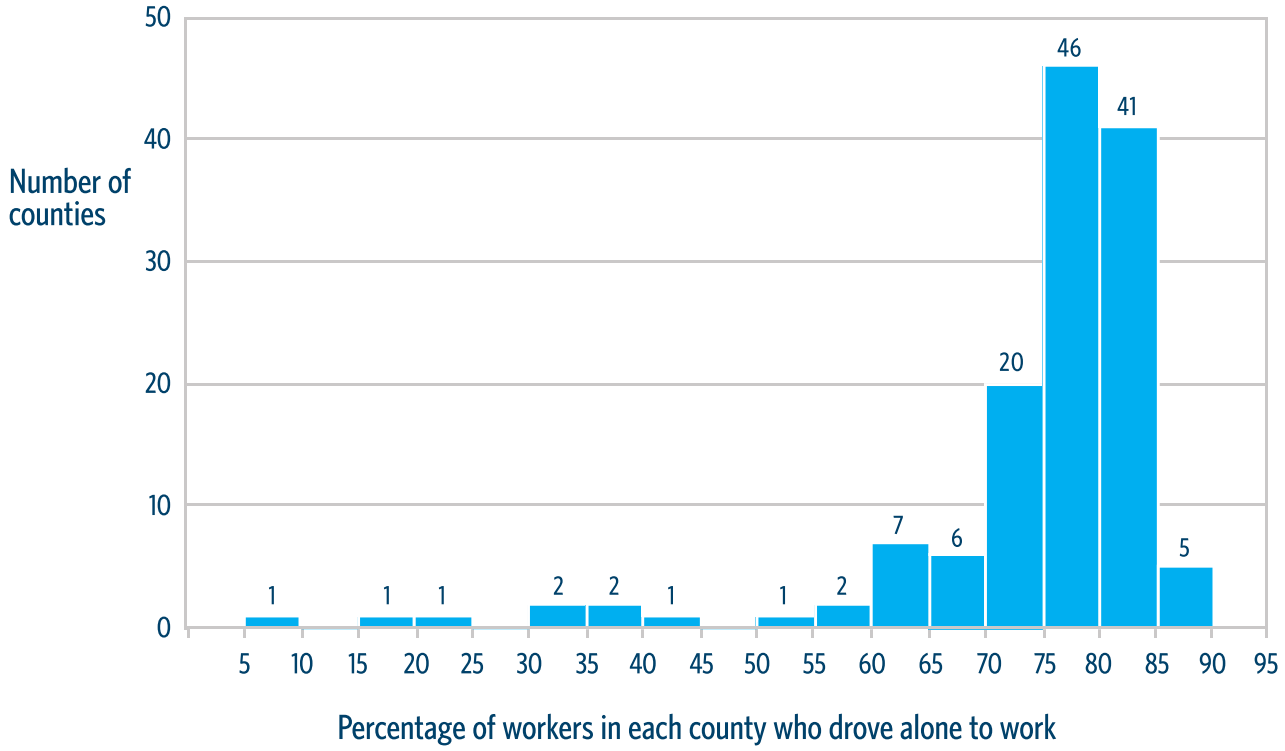
data.census.gov/cedsci/table?q=commute%20to%20work&tid=ACSST1Y2014.S0801&hidePreview=true

a. Which characteristics might the counties represented by the bars on the left of the distribution have in common?

b. Which characteristics might the counties represented by the bars on the right of the distribution have in common?

c. Is this histogram skewed left, skewed right, or symmetrical? How do you know? Do you see any outliers?

2. The following histogram shows the percentage of workers (16 and older) in each of the 136 U.S. counties surveyed who drove alone to work in 2014. The bar on the left, for example, indicates that only one county saw between 5 and 10 percent of its workers drive to work by themselves. Refer to this graph to answer the questions that follow.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

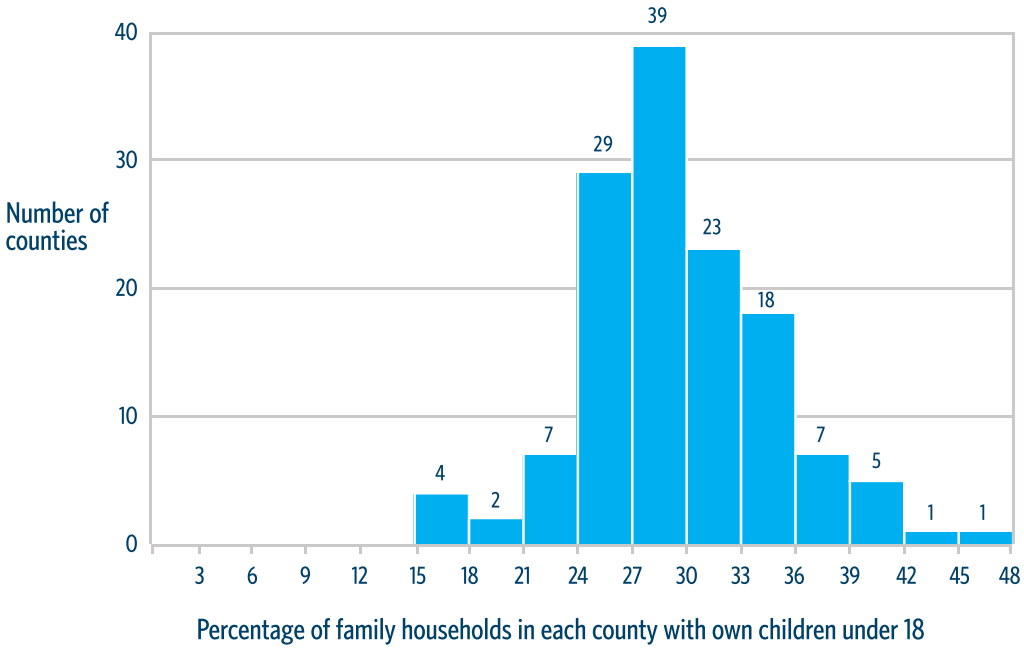
data.census.gov/cedsci/table?q=commute%20to%20work&tid=ACSST1Y2014.S0801&hidePreview=true

a. What characteristics might the counties represented by the bars on the distribution's left have in common?

b. What characteristics might the counties represented by the bars on the distribution's right have in common?

c. Is this histogram skewed left, skewed right, or symmetrical? How do you know? Do you see any outliers?

3. The following histogram shows the percentage of family households with own children under 18 in each of the 136 U.S. counties surveyed. Refer to this graph to answer the questions that follow.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Living%20arrangements%20&tid=ACSDP1Y2014.DP02&hidePreview=true&y=2014

a. What characteristics might the counties represented by the bars on the far left have in common?

b. What characteristics might the counties represented by the bars on the far right have in common?

c. Is this histogram skewed left, skewed right, or symmetrical? How do you know? Do you see any outliers?

Part 2 – Compare Measures of Spread to Predict a Distribution's Shape

In symmetrical distributions, the mean and the median are close together. In skewed distributions, the mean is drawn toward the longer "tail": Skewed-right distributions typically have a mean that is greater than the median, and skewed-left distributions typically have a mean that is less than the median.

Based on this information, answer the following questions, then check your answers by looking back at the histograms in part 1.

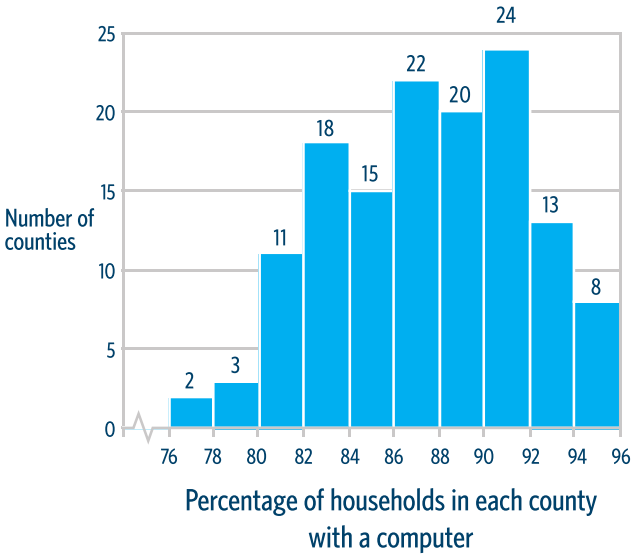
1. Question 1 in part 1 focused on the percentage of workers in each county who walked to work in 2014. The mean of those data is 2.86 percent, and the median is 2.00 percent. Based just on the relationship between these two values, would you expect the graph to be symmetrical, skewed right, or skewed left? Why?

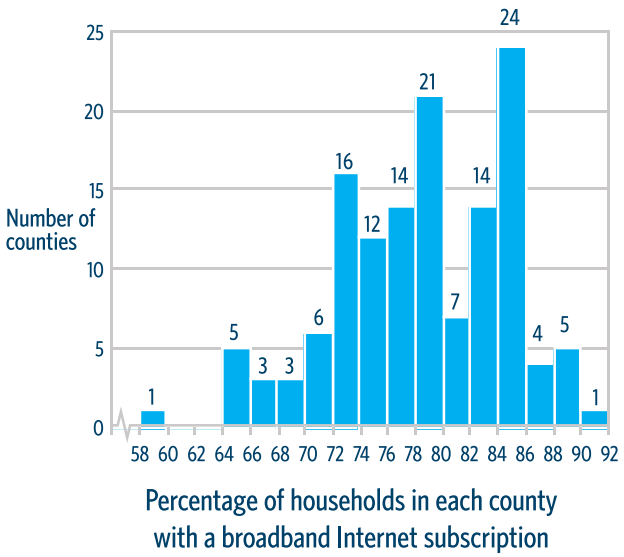
2. Question 2 in part 1 focused on the percentage of workers in each county who drove alone to work in 2014. The mean of those data is 74.03 percent, and the median is 78.20 percent. Based just on the relationship between these two values, would you expect the graph to be symmetrical, skewed right, or skewed left? Why?

3. Question 3 in part 1 focused on the percentage of family households in each county with own children younger than 18 in 2014. The mean of those data is 29.42 percent, and the median is 28.95 percent. Based just on the relationship between these two values, would you expect the graph to be symmetrical, skewed right, or skewed left? Why?

Part 3 – Compare Data Distributions

Consider the two histograms below to answer the questions that follow, noting when you make comparisons that these graphs have different data ranges. The histogram on the left shows the percentage of households in each county surveyed that had a computer, and the histogram on the right shows the percentage that had a broadband Internet subscription, both in 2014.





Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Living%20arrangements%20&tid=ACSDP1Y2014.DP02&hidePreview=true&y=2014

1. Which histogram has the greater mean?

2. Based on the nature of these variables, could you have predicted which data set would have a greater center? Explain your logic.

3. Which distribution is more symmetrical? How do you know?

4. Which distribution has the larger spread? How do you know?

Part 4 – Evaluate Spread

Answer the following questions using what you know about the three common measures of spread: the range, the IQR, and the SD (formulas of each shown below).

1. Consider again the two histograms about computer and Internet access — with the means and SDs now highlighted in red — and the table below to answer the questions that follow.

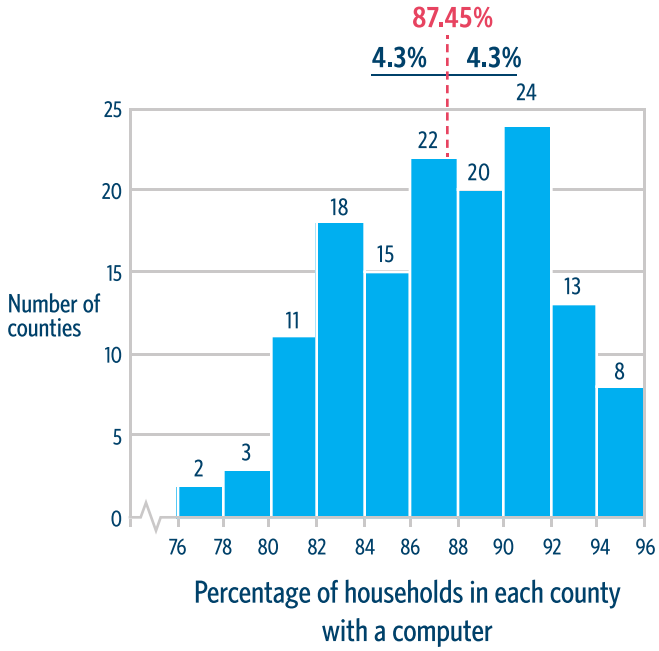
   Range = Maximum - minimum

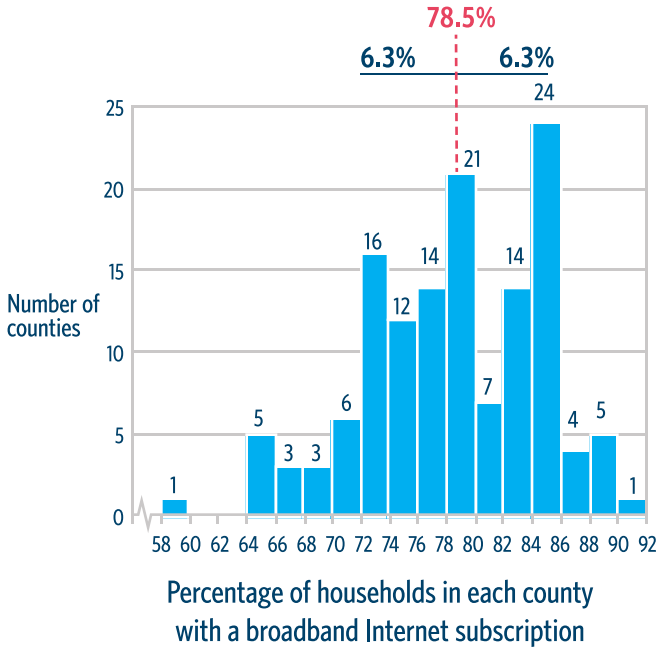
   IQR = Q_3 - Q_1

   Standard Deviation =

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

   Remember: The IQR for these distributions indicates that the widths of the middle 50 percent of the data are 6.7 and 10.1, respectively. The SD says that while some county percentages are closer to and farther from the means, they are within 4.3 and 6.3 percentage points on average, respectively, on both sides of the mean.





Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

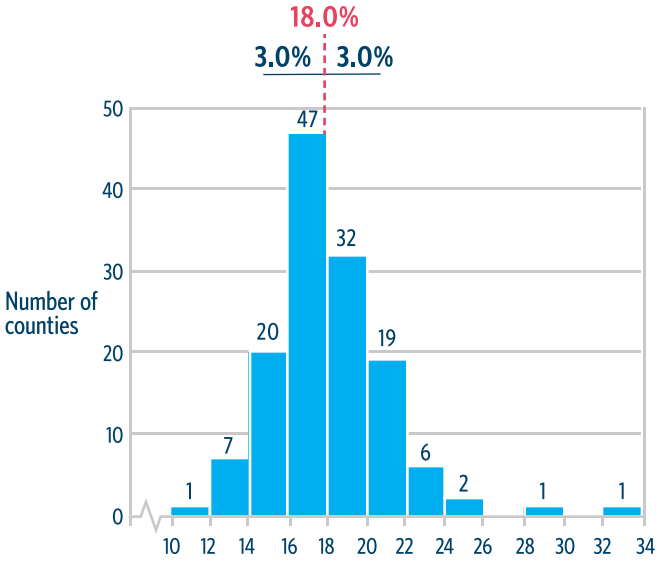
data.census.gov/cedsci/table?q=Living%20arrangements%20&tid=ACSDP1Y2014.DP02&hidePreview=true&y=2014

Values for the "Percentage of households with a computer" graph (in percentage points)			Values for the "Percentage of households with Internet" graph (in percentage points)		
Range	IQR	SD	Range	IQR	SD
18.5	6.7	4.3	32.8	10.1	6.3

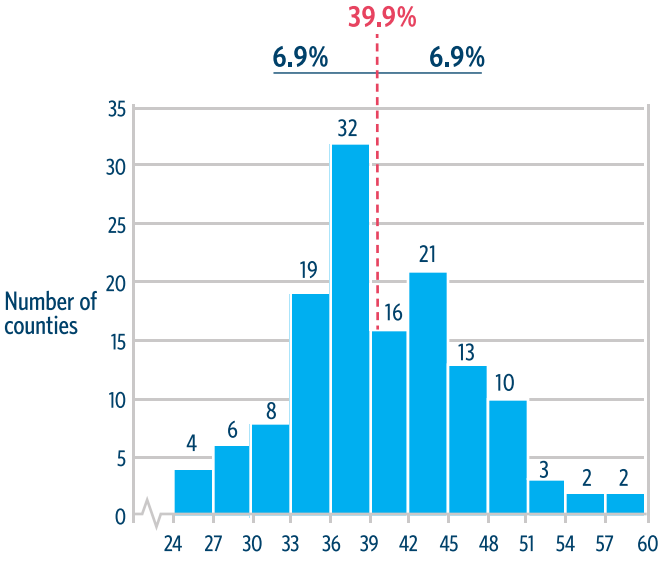
a. What do the values in the table reveal about the spread and variability of the data in both graphs?

b. Mathematically, why do you think that the values for the Internet access distribution are larger than those for the computer ownership distribution?

2. Consider the two histograms and table below to answer the questions that follow. The graph on the left represents the distribution of 136 U.S. counties with percentages of civilians over 16 in service occupations, and the graph on the right shows the percentages in management, business, science, and arts occupations — both in 2014. For example, the tallest bar in the graph on the left indicates that 47 counties had between 16 and 18 percent of their civilians over 16 employed in service occupations.



Percentage of civilians over 16 in each county in service occupations



Percentage of civilians over 16 in each county in management, business, science and arts

Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Occupation%20of%20civilian%2016%20and%20over%20&tid=ACSST1Y2018.S2401&t=Occupation&hidePreview=true

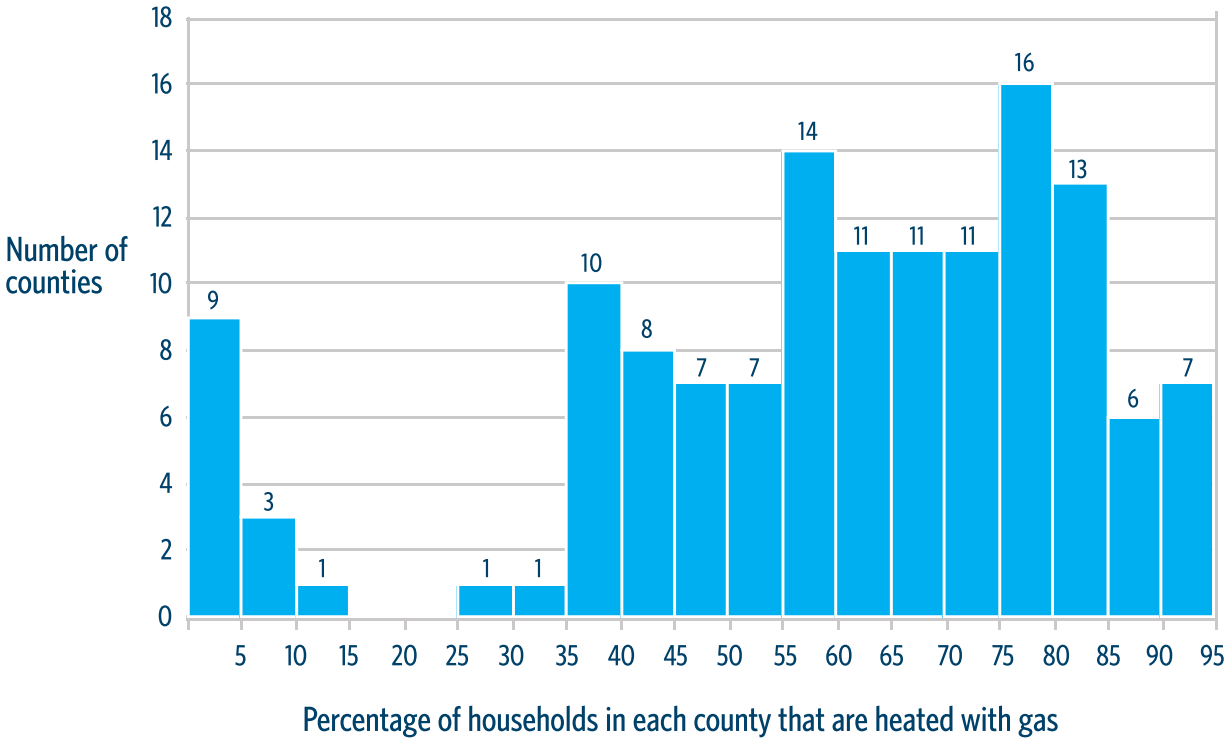
Values for the "Percentage of civilians over 16 in service occupations" graph (in percentage points)					Values for the "Percentage of civilians over 16 in management, business, science, and arts occupations" graph (in percentage points)				
Mean	Median	Range	IQR	SD	Mean	Median	Range	IQR	SD
18.0	17.7	21.3	3.2	3.0	39.9	38.6	35.2	8.5	6.9

a. Describe the shape, center, and spread of each histogram.

b. Now compare your descriptions for these graphs, thinking about which measures are best for comparing distributions with different scales.

c. Summarize what your comparisons mean in context.

3. Consider the following histogram, which shows the percentage of households in each county that used gas heat in 2014.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

Note: ACS 1-year data include data for all counties in the United States. This activity only focuses on 120 counties to illustrate the statistical concept.

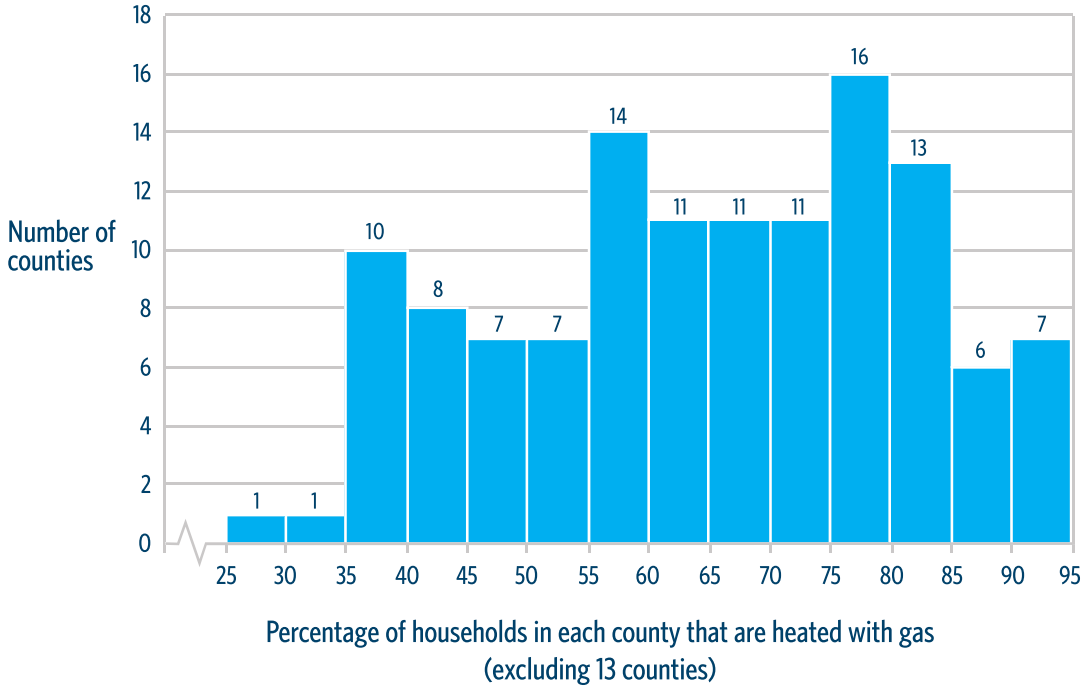
data.census.gov/cedsci/table?q=DP04&tid=ACSDP1Y2014.DP04&hidePreview=true&y=2014

The data are somewhat bimodal, with a group of 13 counties on the left that have low percentages of households heated with gas and a group of more than 120 counties on the right that have over one-third of their households heated with gas. A single measure of center would not be appropriate here.

The 5-number summary, mean, and IQR for the data in this histogram are as follows:

5-Number Summary						
Minimum	Q1	Median	Q3	Maximum	Mean	IQR
2.4%	45.5%	63.5%	77.5%	91.8%	59.4%	32.0%

If the data for the 13 counties with very low percentages (whose mean and median are 4.7 percent and 3.8 percent, respectively) were removed, the graph would appear as follows:



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

Note: ACS 1-year data include data for all counties in the United States. This activity focuses on only 120 counties to illustrate the statistical concept.

data.census.gov/cedsci/table?q=DP04&tid=ACSDP1Y2014.DP04&hidePreview=true&y=2014

The 5-number summary, mean, and IQR for this revised histogram would then be:

5-Number Summary						
Minimum	Q1	Median	Q3	Maximum	Mean	IQR
27.6%	52.1%	67.2%	78.5%	91.8%	65.2%	26.4%

a. Which measure of center was more affected by the omission of the 13 counties: the mean or the median? Explain your reasoning, using what you know about the calculations of these two measures.

Part 5 – Draw Conclusions About Data

1. Reflect on what you have learned: Write about when it would be best to use the median or the mean to describe the center of a distribution and when it would be best to use the range, the SD, and/or IQR to describe the spread in different kinds of distributions.