Unit 4: Measures of Center

SUMMARY OF VIDEO

One number most people pay a lot of attention to is the one on their paycheck! Today's workforce is pretty evenly split between men and women, but is the salary distribution for women the same as for men? The histograms in Figure 4.1 show the weekly wages for Americans in 2011, separated by gender.

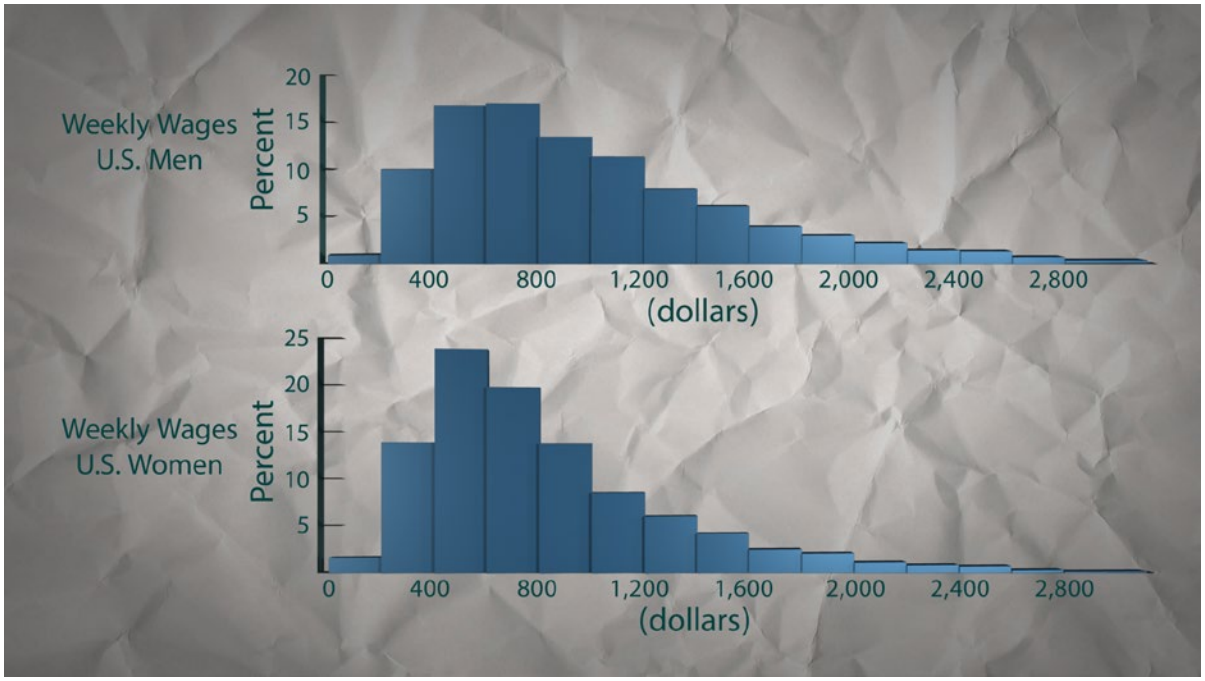

Figure 4.1. Histograms comparing men's and women's wages.

Both histograms are skewed to the right with most people making moderate salaries while a few make much more. For comparison's sake, it would help to numerically describe the centers of these distributions. A statistic called the median splits the distribution in half as shown in Figure 4.2 – half the wages lie above it, and half below. The median wage for men in 2011 was $865. The median wage for women was only $692, about 80% of what men make.

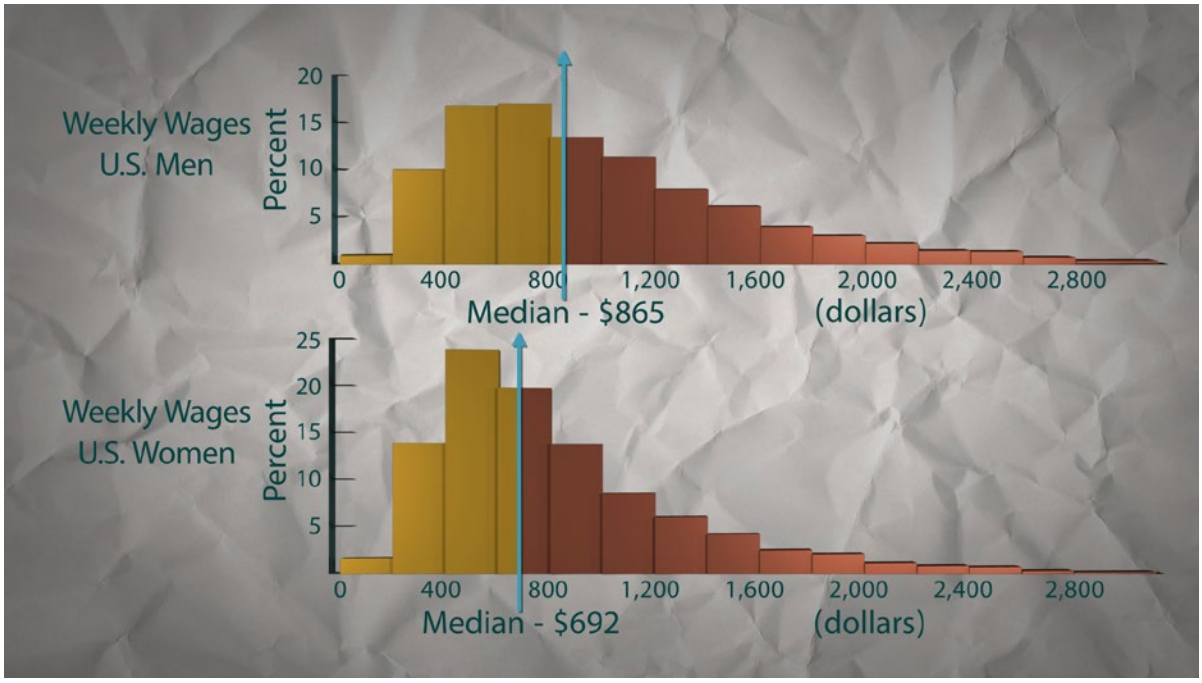

Figure 4.2. Locating the median on histograms of wages.

Simply using the median, we have identified a real disparity in wages, but it is much harder to figure out why it exists. Some of the difference can be accounted for by differences in education, age, and years in the workforce. Another reason for the earnings gap is that women tend to be concentrated in lower-paying jobs – but that begs the question: Are these jobs worth less in some sense? Or are these jobs lower paid because they are primarily held by women? This is the central issue in the debate over comparable worth – the idea that men and women should be paid equally, not only for the same job but for different jobs of equal worth.

Back in 1988 the city of Colorado Springs, Colorado, was at the forefront of this debate. As part of its normal operation, the city government evaluated all municipal jobs with criteria like working conditions, necessary skills, and accountability required. Each job got a numerical rank. It turned out that many clerical jobs, which are mostly filled by women, scored the same number of points as operations and maintenance jobs, which are mostly filled by men. However, the median wage for men's jobs was always higher than the corresponding median wage for the women even though these jobs were judged to be exactly equal in requirements and responsibility. A group of clerical workers used this evidence to pressure the city for a more equitable pay structure. The numbers were hard to argue with and the clerical workers won. The city agreed to equalize the median salaries for jobs of comparable worth. And the plan had a benefit for the city as well – the relatively high turnover rate for jobs held by women decreased.

Colorado Springs relied on the median statistic to identify the inequality in men's and women's salaries. Next, we take a look at how to calculate this measure of center. Below are the weekly salaries from a small hypothetical company that has 19 employees. The salaries have been arranged in order from the lowest, $290 for an entry-level receptionist, up to the highest, $2,000 for the president.

290	350	400	400	450	450	450	500	500	500
550	550	650	750	800	1200	1300	1500	2000	

The median represents a typical wage. To calculate the median, determine the number of observations, *n*. In this case, we have 19 salaries and so, *n* = 19. The location of the median is at (*n* + 1)/2, or (19 + 1)/2 = 10. Count up 10 spots from the bottom (or down 10 spots from the top) and read off the median: $500. Since we had an odd number of paychecks, it is easy to count up 10 places to the middle number. But what if we had an even number of observations to deal with? Suppose we add a paycheck of $550 as shown below.

290	350	400	400	450	450	450	500	500	500
550	550	550	650	750	800	1200	1300	1500	2000

With the additional paycheck, *n* = 20. Now, we count up (20 + 1)/2, or 10.5 spaces. That puts us right in between the two middle values of 500 and 550. So, the median is actually halfway between those two salaries, $525.

The median is not the only measure for center. Another way to measure the center of a distribution of values is by taking the average. Statisticians call this number the mean, which is denoted by $\bar{x}$. It is calculated by adding up all the values and dividing by the number of values:

$$\bar{x} = \frac{\sum x}{n}$$

If we return to our original 19 paychecks, we find the mean as follows:

$$\bar{x} = \frac{\$290 + \$350 + \ldots + \$1500 + \$2000}{19} = \frac{13{,}590}{19} \approx \$715.26$$

Notice that the mean, which is about $715, is higher than the median of $500. You can think of the mean as the balancing point of all the values. It is the value of the pivot point shown in Figure 4.3 that balances all the observations.

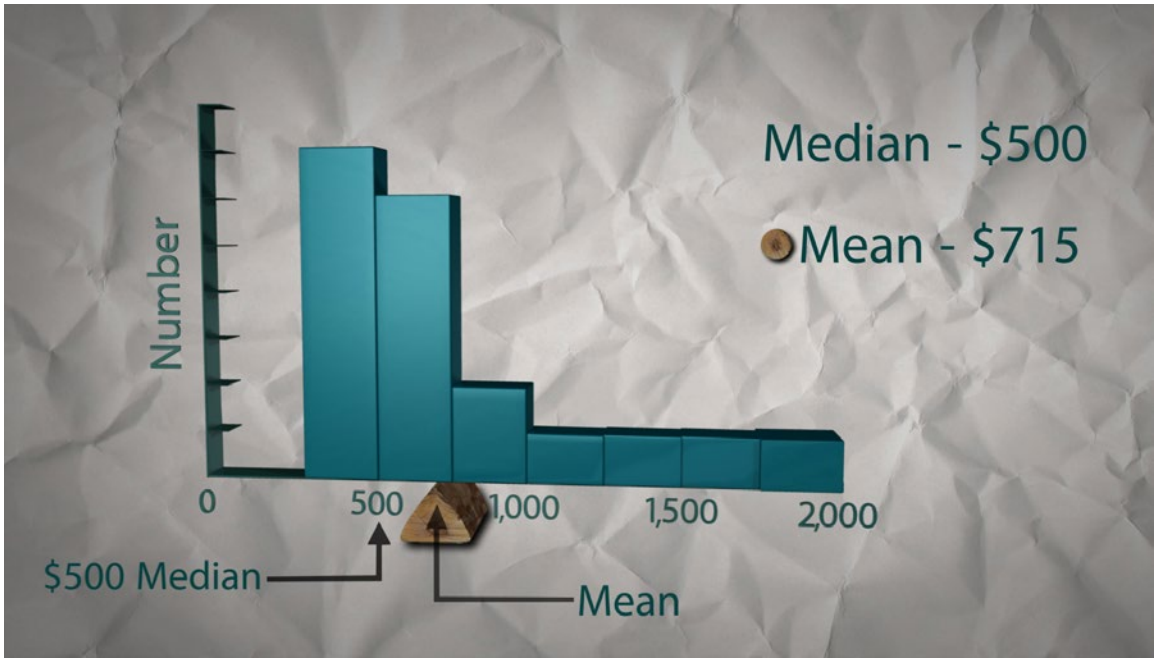

Figure 4.3. The mean as the balancing point.

The mean is influenced much more by the one high salary going to the president of the company. The median, on the other hand, is what statisticians call resistant. The median doesn't depend on what the values are out there at the extremes of our distribution. If the president doubled his salary while everyone else stayed at the same wage, the mean would bump up to $820.53, or around $821. But our median would stay at $500.

The shape of a distribution can give you some hints about the relationship between the mean and median. For a fairly symmetric distribution, such as the one shown in Figure 4.4, the mean and median are roughly the same.

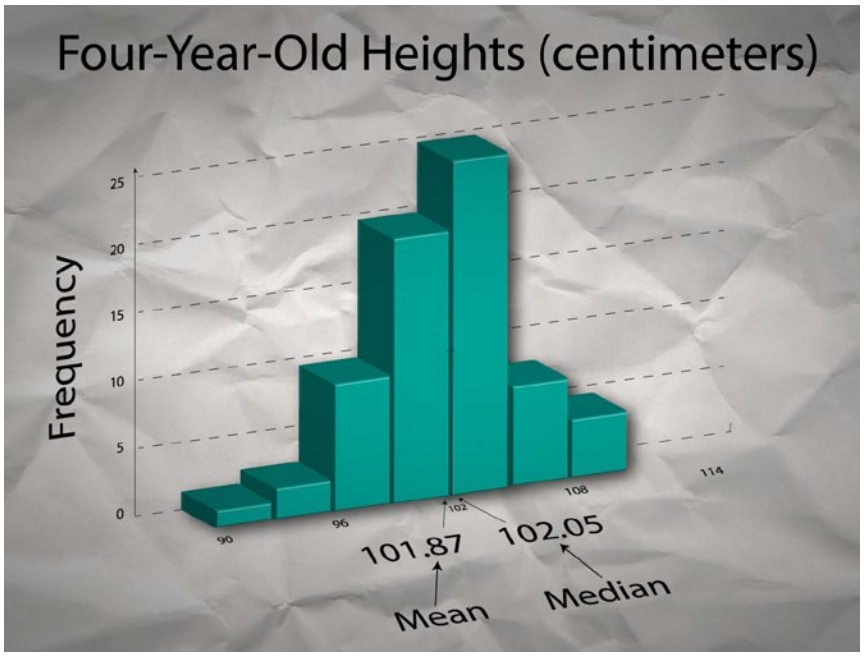

Figure 4.4. Mean and median are close.

If the distribution is skewed to the right, like the scores on the difficult exam pictured in Figure 4.5, the mean is larger than the median.

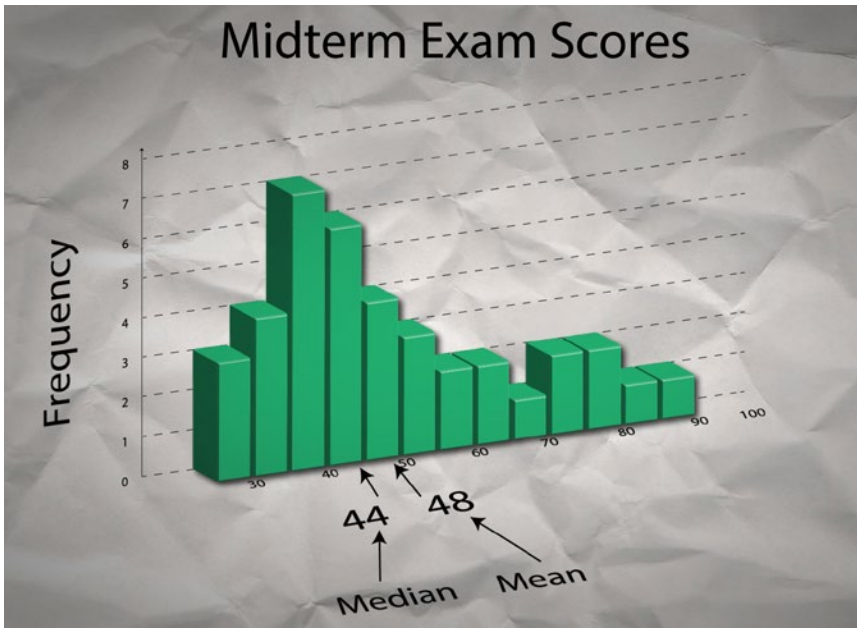

Figure 4.5. Mean larger than median.

Likewise, if the distribution is skewed to the left, like the scores on one easy exam shown in Figure 4.6, the mean is smaller than the median. Remember the mean is influenced by values at the extremes and the median is not.

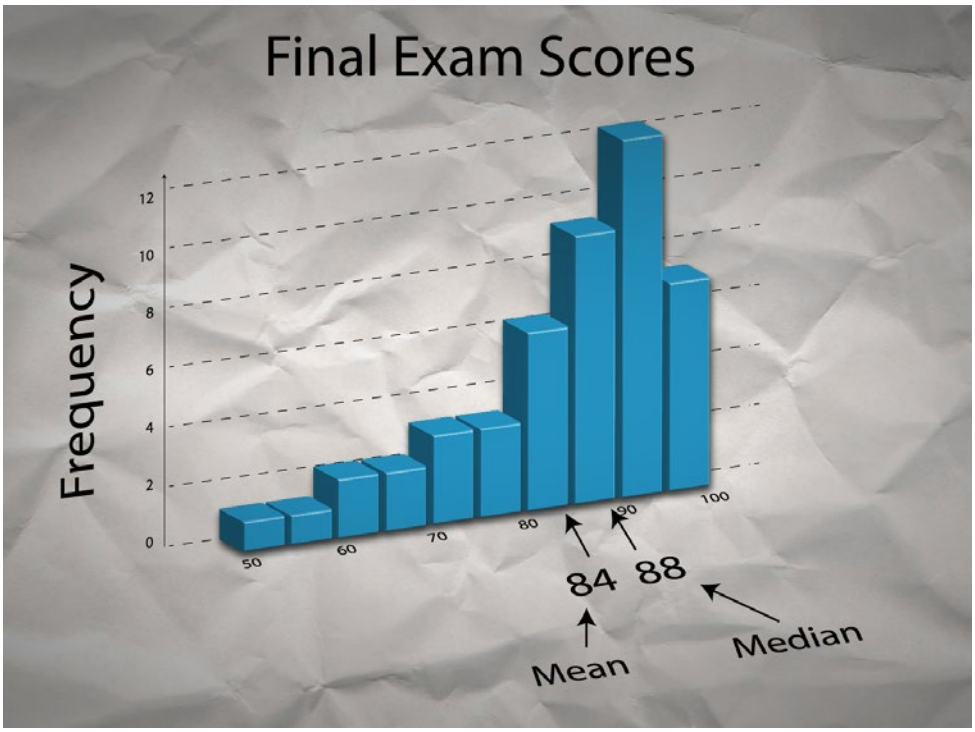

Figure 4.6. Median larger than mean.

STUDENT LEARNING OBJECTIVES

A. Understand that graphical descriptions of data are more meaningful when supplemented with numerical measures of center.

B. Know that the median (midpoint or typical value) and mean (arithmetic average) are common measures of center (or location) for a distribution. Sometimes the mode is also used as a measure of center.

C. Be able to calculate the median, mean, and mode of a small data set.

D. Know that the mean and median should be close in symmetric distributions and that the mean is pulled toward the long tail of a skewed distribution. Know that the mean is a non-resistant measure of center because it is strongly influenced by extreme observations and that the median is a resistant measure of center.

E. Be able to choose an appropriate measure of center in practice.

CONTENT OVERVIEW

Before describing data numerically, we always begin with a graph such as a stemplot or a histogram. The graph shows us the overall pattern of the data and any striking deviations such as outliers. The next step is to give a numerical description of some important aspects of the data. The focus of this unit is on numerical descriptions of the center or location of a distribution. The median, mean, and mode are three numerical measures that use different ideas of "center." We begin with the median.

The **median** is the midpoint of a distribution, the value with half the observations lying below it and half above. Instructions for calculating this midpoint number are given below.

Calculating the Median of *n* Observations

Step 1: Arrange the observations from smallest to largest.

Step 2: Determine the location of the median: (n + 1)/2.

Step 3: Find the median in the ordered list from Step 1:

If *n* is odd, count up (n + 1)/2 spots in the ordered list and select this value. The median will be the middle number in the ordered list.

If *n* is even, count up the number of spots on either side of (n + 1)/2 and average these two values. This median will be the average of the two middle numbers.

The median is easy to calculate once the data are ordered from smallest to largest. However, if the data set is large, use software to sort the data from largest to smallest. Another approach to ordering the data would be to make a stemplot. As an example, consider the 22 exam scores listed below.

40 41 50 68 69 72 76 79 79 80 82 85 86 87 88 88 90 91 92 93 96 98

The exam scores have already been ordered from smallest to largest. Notice that repeat scores are included in the list. For example, two people scored 88 and so, the score of 88 appears twice on this list.

Now, we compute the location of the median:

$$\frac{n+1}{2} = \frac{22+1}{2} = 11.5$$

Start at 40 and count up 11 and 12 positions. Exam scores 82 and 85 are in the 11th and 12th position. The median is the average of these two numbers:

$$\text{median} = \frac{82+85}{2} = 83.5$$

Next, we discuss the mean as a measure of center. The **mean** is the average value. It is the balance point of the distribution (See Figure 4.3.). If the observations are from a sample of *x* values, we often use the notation $\bar{x}$ to represent the mean.

Here's how to calculate the mean:

Calculating the Mean

For *n* observations of *x* values: $\bar{x} = \frac{\text{sum of the observations}}{\text{number of observations}} = \frac{\sum x}{n}$

Returning to our example of 22 exam scores, we calculate the mean as follows:

$$\bar{x} = \frac{40+41+\ \ldots\ +96+98}{22} = \frac{1730}{22} \approx 78.6$$

If we had started with a graphic display of the exam scores as shown in Figure 4.7, we should have expected a mean that was less than the median. The histogram is skewed to the left, with a few exam scores in the left tail of the distribution. The median is unaffected by these scores, but they drag the mean down.

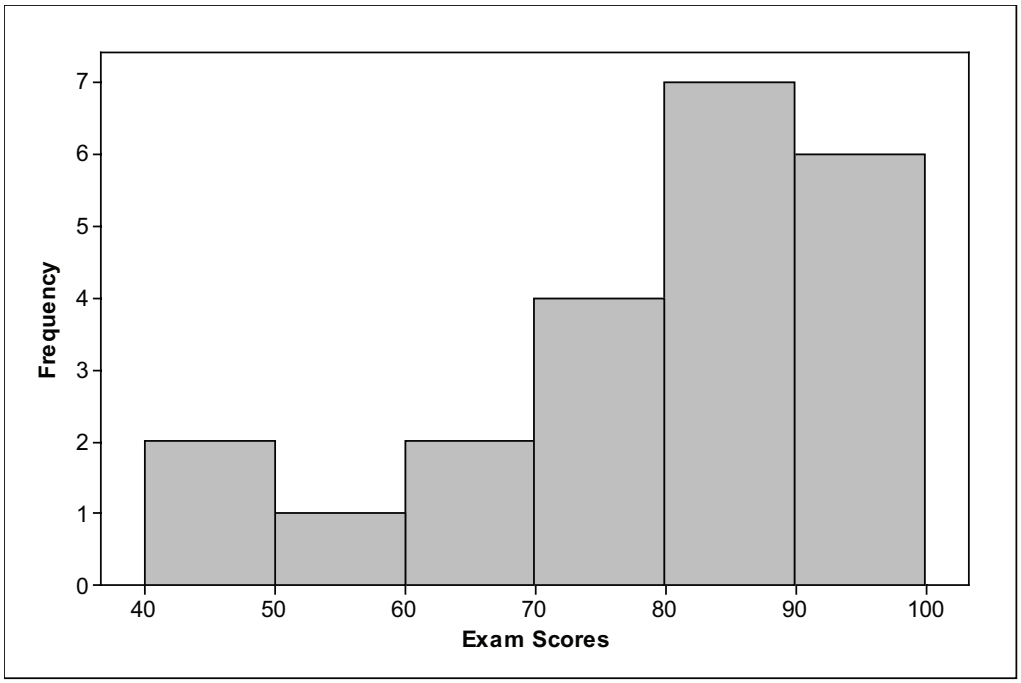

Figure 4.7. Histogram of exam scores.

Lastly, there is one other measure that is sometimes used as a measure of center, and that is the mode. The **mode** is the most frequent observation. In our list of exam scores, there are two scores that appear twice in the list, 79 and 88. Since both of these scores are tied for occurring most frequently, the mode is not unique – instead there are two modes.

We have discussed three measures of center or location, the median, mean, and mode. How do you decide which is best for a given situation? In choosing an appropriate measure of center, start with a graphic display of the data. Consider the overall shape of the data and deviations from that shape before deciding whether to use the mean or median to summarize the location of the data. Keep in mind that the median is a **resistant** measure of center, which is not influenced by a few extreme data values whereas a few extreme outliers can pull the mean in the direction of the extreme values.

For roughly symmetric distributions the mean and median will be close in value. For highly skewed data, or data with extreme outliers, the median is generally the better choice for a measure of the center or location of the data. For data sets with multiple peaks, the modes may give a better indication of location.

KEY TERMS

The **median** gives the midpoint of a set of data – it separates the upper half of the data from the lower half. To calculate the median, order the data from smallest to largest and count up $(n + 1)/2$ places in the ordered list.

The **mean** is the arithmetic average or balance point of a set of data. To calculate the mean, sum the data and divide by the number of data:

$$\overline{x} = \frac{\sum x}{n}$$

The **mode** is the data value that occurs most frequently.

A **resistant measure** of some aspect of a distribution (such as its center) is relatively unaffected by a small subset of extreme data values.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What *variable* is examined in comparing men and women workers at the beginning of the video?

2. Would you describe the shape of the distribution of men's weekly wages as symmetric, skewed to the left or skewed to the right?

3. What is the most important difference between the distributions of weekly wages for men and for women?

4. Would a few very large incomes pull the mean of a group of incomes up, down, or leave the mean unaffected?

5. Would a few very large incomes pull the median of a group of incomes up, down, or leave the median unaffected?

UNIT ACTIVITY:
MEAN, MEDIAN AND DISTRIBUTION SHAPE

This activity will provide an opportunity to practice computing the mean and the median. In addition, the activity will emphasize the relevance of a distribution's shape to the relationship between the mean and median. You will need to access Stemplots from the Interactive Tools menu.

1. Work in Quiz mode. Set the number of observations to 10 and the maximum to 100. Assume that the Stemplots tool is generating 10 hypothetical test scores.

a. Make a copy of the stemplot. Based only on the shape of the stemplot, which do you think is larger, the mean or the median? Justify your choice.

b. Calculate the mean and median. Show your calculations. Submit your answers to make sure they are correct. (If not, revise your answers and re-submit.) Now that you have done the calculations, was your answer to (a) correct?

2. Repeat question 1 with a new sample of 10 test scores.

3. Use the Stemplots tool to generate 17 final exam scores. The final exam is worth 150 points; so, set the Maximum Observation Value to 150.

a. Make a copy of the stemplot.

b. Based only on the shape of the stemplot, which do you think is larger, the mean or the median? Justify your choice.

c. Calculate the mean and median. Use the Stemplots tool to check that your calculations are correct. Was your answer to (b) correct?

4. Repeat question 3 with a new sample of 17 final exam scores.

5. Below are 70 exam scores from a very difficult exam given to a large class.

64	78	67	35	74	73	69	66	36	69
74	38	72	79	36	46	77	69	39	38
63	32	36	80	35	35	36	39	35	35
67	73	58	43	64	64	69	69	69	37
50	63	36	39	74	36	35	60	62	65
69	69	35	34	49	67	65	61	33	36
36	37	36	36	65	69	40	72	69	66

a. Work in calculation mode. Enter the exam scores into the Stemplots tool. Use the interactive tool to make the stemplot. Describe the shape of the plot.

b. Determine the median, mean, and mode(s) for the exam scores.

c. Based on the plot, which gives a better description of the location of these data, the median, mean, or mode(s)? Explain.

6. Below are 30 exam scores from a statistics exam.

90	76	78	76	75	74	85	74	65	78
75	60	75	76	75	78	70	75	65	85
72	74	70	76	72	80	80	72	78	74

a. Work in calculation mode. Enter the exam scores into the Stemplots tool. Use the interactive tool to make the stemplot. Describe the shape of the plot.

b. Determine the median, mean, and mode(s) for the exam scores.

c. Based on the plot, which gives a better description of the location of these data, the median, mean, or mode(s)? Explain.

EXERCISES

1. Here are the starting salaries, in thousands of dollars, offered to the 20 students who earned degrees in computer science in 2011 at a university.

63	56	66	77	50	53	78	55	90	65
64	69	59	76	48	54	49	68	51	50

a. Make a graph to describe the distribution and write a brief description of its important features.

b. Find the median salary.

c. Find the mean salary.

d. Find the mode of the salaries.

e. Is the mean about the same as the median or not? What feature of the distribution explains the difference between the mean and the median? Is the mode a good measure of the center for these data?

2. Each month, the Commerce Department reports the “average” price of new single-family homes. For August 2012, the two “averages” reported were $256,900 and $295,300. Which of these numbers was the mean price and which was the median price? Explain your answer.

3. In 1961 New York Yankee outfielder Roger Maris held the major league record for home runs in a single season, with 61 home runs. That record held for 37 years. Here are Maris’s home run totals for his 10 years in the American League.

13, 23, 26, 16, 33, 61, 28, 39, 14, 8

a. Find the mean number of home runs that Maris hit in a year, both with and without his record 61. How does removing the record number of home runs affect his mean number of runs?

b. Find the median number of home runs that Maris hit in a year, both with and without his record 61. How does removing the record number of home runs affect his median number of runs?

c. If you had to choose between the mean and median to describe Maris's home run hitting pattern, which would you use?

4. Refer to Table 3.3 (Unit 3). This table gives the number and percentage of residents 65 and older in each state and the District of Columbia.

a. Unit 3, Exercise 1(a) asked you to draw a histogram of the numbers of residents 65 and older. (If you haven't already done so, draw the histogram.) Compute the mean and median of these data. Which measure of location, the mean or the median, better describes the location of the numbers of residents 65 and older? Justify your choice based on a histogram of these data.

b. Unit 3, Exercise 2(a) asked you to draw a histogram of the percentage of residents 65 and older. (If you haven't already done so, draw the histogram.) Compute the mean and median of these data. Which measure of location, the mean or the median, better describes the location of the percentage of residents 65 and older? Justify your choice based on a histogram of the percentages.

REVIEW QUESTIONS

1. A man in a nursing home has his pulse taken every day. His pulse readings (beats per minute) over a one-month period appear below.

72 56 56 68 78 72 70 70 60 72 68 74
76 64 70 62 74 70 72 74 72 78 76 74
72 68 70 72 68 74 70

a. Make a stemplot of the pulse data. Break the stem into 5 (for digits 01, 23, 45, 67, and 89).

b. Determine the mean, median and mode for these data. Be sure to include units in your answers.

c. Based on these data, which measure (or measures) from (b) do you think best describes the man's typical pulse rate? Explain your reasoning.

2. Eating fish contaminated with mercury can cause serious health problems. Mercury contamination from historic gold mining operations is fairly common in sediments of rivers, lakes, and reservoirs today. A study was conducted on Lake Natoma in California to determine if the mercury concentration in fish in the lake exceeded guidelines for safe human consumption. A sample of 83 largemouth bass was collected and the concentration of mercury from sample tissue was measured. Mercury concentration is measured in micrograms of mercury per gram or μg/g. The histogram in Figure 4.8 presents results from this study.

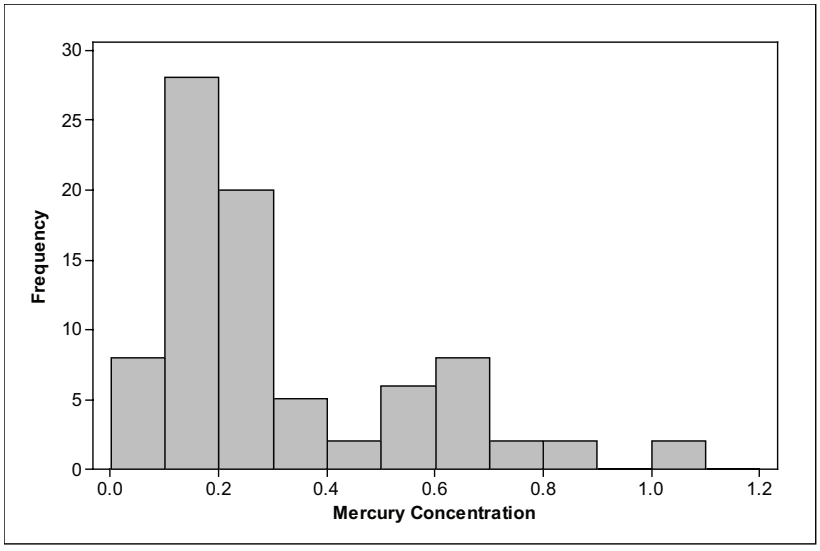

Figure 4.8. Histogram of mercury concentration in fish.

a. The primary objective of the study was to determine if mercury concentrations in fish tissue exceeded safety guidelines for human consumption. The U.S. Environmental Protection Agency (USEPA) human health criterion for methylmercury in fish is 0.30 μg/g. Approximately how many of the fish in the sample had mercury concentrations below the level set by the EPA (and hence were considered safe for human consumption)?

b. Approximately what percentage of the sample had mercury concentrations higher than the level set by the EPA? Show how you arrived at your answer.

c. Would the mean mercury concentration be larger, smaller, or about the same as the median mercury concentration? Explain.

3. A student often orders french fries at a local fast-food place. She keeps track of the number of french fries in each small bag she buys. Here are her counts:

42, 47, 49, 58, 43, 47, 44, 38, 38, 28, 55, 40, 46
54, 45, 45, 51, 35, 46, 37, 46, 40, 43, 49, 37

a. Calculate the mean and median for these data. Show how you computed these values.

b. Make a stemplot of the distribution. Describe the overall shape of the distribution. Are there any outliers?

c. Do you prefer the mean or the median as a brief description of the center of this distribution? Why?