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.

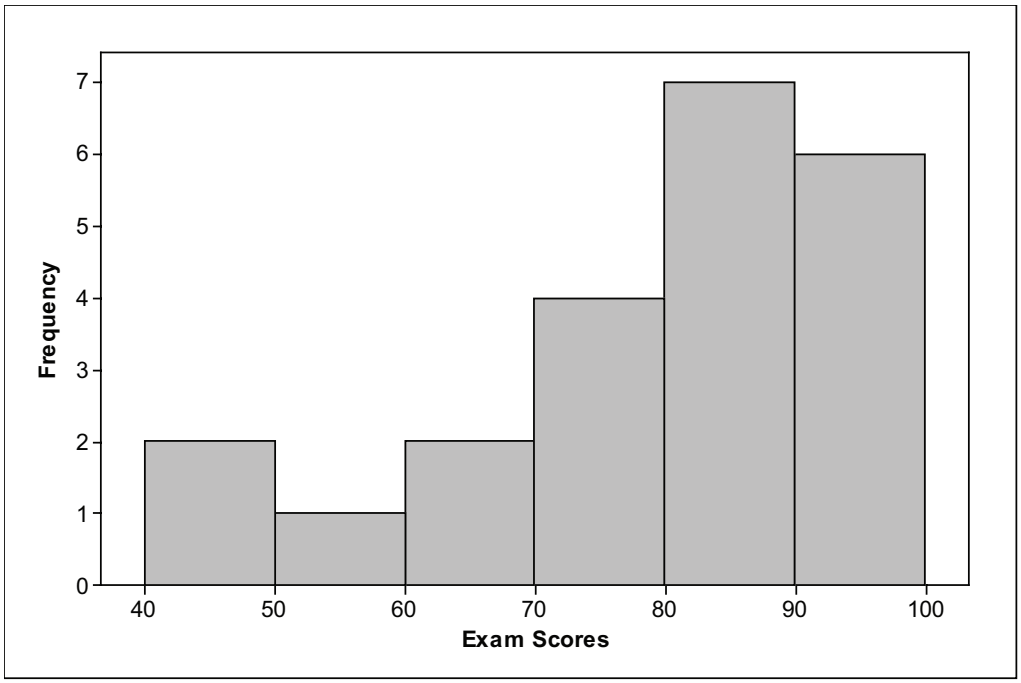

Figure 4.7. Histogram of exam scores.

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.