Interpreting the Shape of Data Distributions

Differences in Center

You've probably seen data sets in which a few data elements are quite different from the rest. For example:

> Suppose that, in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the more accurate description of the center—the mean or the median?

Solution

$$\text{mean} = \frac{5{,}000{,}000+(49 \cdot 30{,}000)}{50} = 129{,}400$$

median = 30,000

> The median is a more accurate description of the center than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is vastly different from any other piece of data in this set. The 30,000 gives us a better sense of the center of the data.

Note that adding one piece of data to this list has no effect on the median, but an enormous effect on the mean.

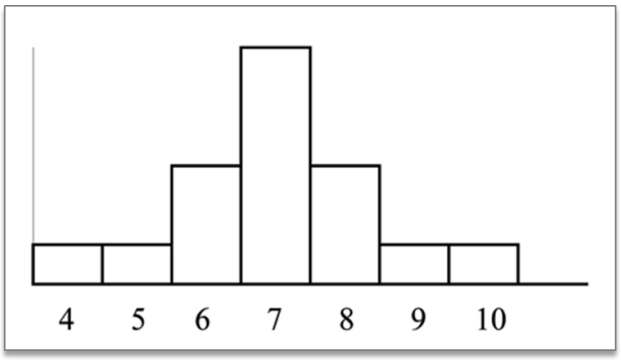
The person making $5,000,000 is an outlier. Outliers are atypical, infrequent observations—that is, values that have an extreme deviation from the center of the distribution. There are no universally agreed-upon criteria to define an outlier, and outliers should be discarded only with extreme caution. However, you should always assess the effects of outliers on the statistical conclusions.

Differences in Spread

Adding an outlier such as in the example above affects spread as well as center. For example, the range, which would be 0 without the outlier (because all the data points had the same value), is now 5,000,000 – 30,000, or 4,970,000. The standard deviation will be affected by the outlier, as well.

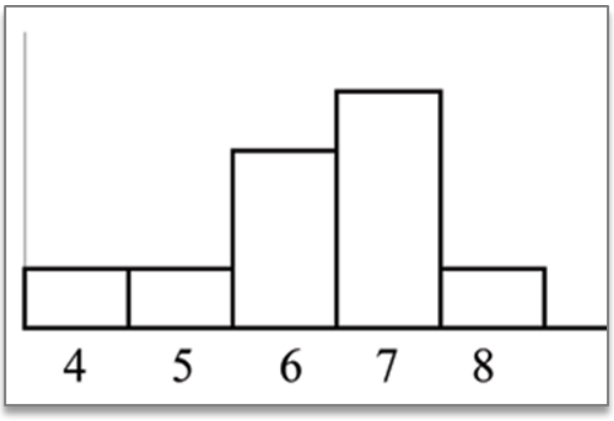
Differences in Shape

Consider this data set: {4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10}. The data produces this histogram. Each interval has a width of one, and each value is located in the middle of an interval. The histogram displays a symmetrical distribution of data.



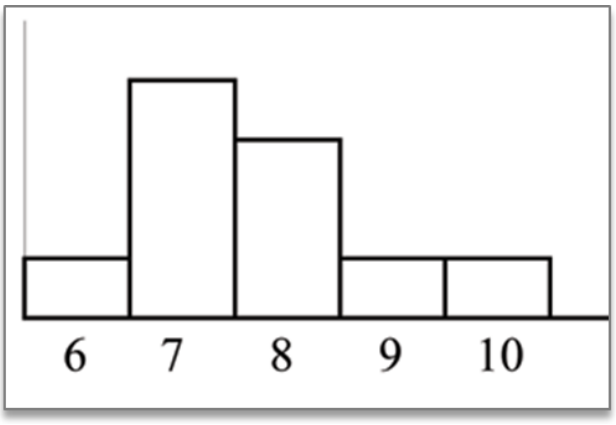
A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shapes to the left and to the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for this data. In a perfectly symmetrical distribution, the mean, the median and the mode are the same.

The histogram shown for the data set {4, 5, 6, 6, 6, 7, 7, 7, 7, 8} is not symmetrical. The right-hand side seems chopped off compared with the left side. We say that the shape of the distribution is skewed to the left because it is pulled out to the left. Another way to say this is that it has negative skew.



The mean is 6.3, the median is 6.5, and the mode is 7. Notice that the mean is less than the median and that they are both less than the mode. The mean and the median both reflect the skewing, but the mean more so.

The histogram for the data set {6, 7, 7, 7, 7, 8, 8, 8, 9, 10} is also not symmetrical. It is skewed to the right. Another way to say this is that it has positive skew.



The mean is 7.7, the median is 7.5, and the mode is 7. Notice that the mean is the largest statistic, while the mode is the smallest. Again, the mean reflects the skewing the most.

To summarize, if the distribution of data is skewed to the left, the mean is less than the median, which is less than the mode. If the distribution of data is skewed to the right, the mode is less than the median, which is less than the mean. (Note that these are just general rules of thumb, and there are some exceptions.)

Skewness and symmetry are important to understand whenever you investigate probability distributions in statistics.

This knowledge article is adapted from the following sources:

Illowsky, Barbara, and Susan Dean. "Collaborative Statistics." *Connexions*. June 22, 2011.

Lane, David M., Project Leader, Rice University. *Online Statistics Education: A Multimedia Course of Study*.