

NUMERICAL DESCRIPTIONS OF QUANTITATIVE DATA: MEASURES OF CENTER

Numerical descriptions of quantitative data allow us to be more precise in describing various characteristics of a data set. Numerical descriptions of data do not replace a good picture (i.e., graphical display) of the data though. Numerical descriptions and graphical displays should be used together.

MEASURES OF CENTER

The center of a distribution is usually the most important aspect to describe. Measures of center are often used to describe or represent what the "typical" value is.

We will talk about three measures of center: mean, median, and mode.

Mean

The **mean** or **average** of a set of numbers is their sum divided by how many numbers there are. The mean $\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$. The notation n is used for the number of observations in a data set.

The notation \bar{x} is reserved for the *sample mean*. The Greek letter μ is used to denote the *population mean*. Most often, a data set contains *sample* data instead of data on the entire population of interest.

For example, suppose the American Medical Association wishes to find the average weight for adult women with height 5 feet, 1 inch, in the age group 35-50 years. The AMA would like to know μ = average weight, but it will be impossible to observe weights for every individual in the population of interest. The AMA settles for estimating μ based on a sample. That is, they estimate the population mean μ by the sample mean \bar{x} .

Median

Another useful measure of center is the median. The median is the middle value of a set of numbers when the numbers are arranged in order. When there are an even number of observations, the median is the average of the *two* middle observations.

The median is such that half of the observations are less than it and half of the observations are larger than it. This property of median does **not** hold for the mean.

Example

A sample of 10 adults was asked to report the number of hours they spent on the internet the previous month. Find the mean and median.

0, 7, 12, 5, 33, 14, 8, 0, 9, 22

Mean = $110/10 = 11.0$

For median, order the observations first.

0, 0, 5, 7, 8, 9, 12, 14, 22, 33

Median = $(8 + 9)/2 = 8.5$

BEHAVIOR OF MEAN AND MEDIAN

Suppose the respondent who reported 33 hours on the internet actually reported 133 hours.

The high outlier pulls the mean internet usage from 11.0 to 21.0, **but the median stays the same**. The mean is pulled in the direction of extreme observations while the median stays the same.

We say the median is a **resistant** measure of center, because it can resist the influence of outliers. The mean, on the other hand, cannot resist the influence of outliers. The mean is highly influenced by extreme values.

When a data set contains extreme values, the median will usually be more representative of what is "typical."

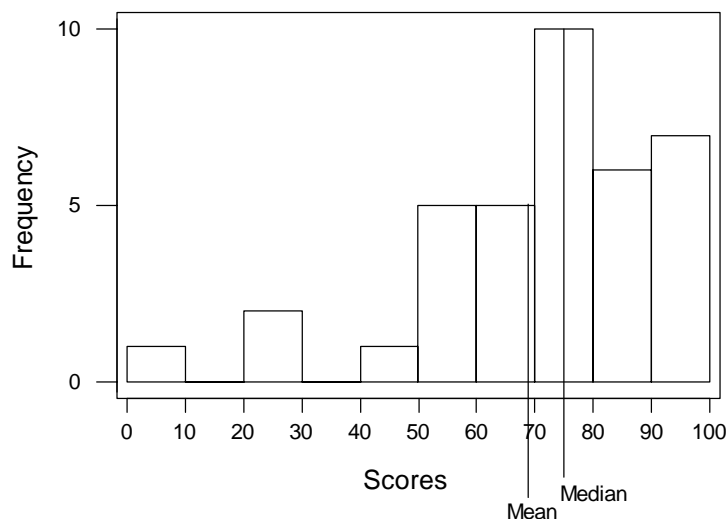
MEAN, MEDIAN, SYMMETRY, AND SKEWNESS

- ◆ The mean and median of a symmetric distribution are close together.
- ◆ In a right-skewed distribution, large (high) observations "pull" the mean right of the median. The mean is pulled toward the long tail.
- ◆ In a left-skewed distribution, small (low) observations "pull" the mean left of the median and toward the long tail.

Example

90	95	94	93	96	90	92	87	82	86	81	86	86	71	71
75	75	77	70	75	75	77	75	62	69	61	62	61	56	56
58	53	51	40	20	21	3								

n	Mean	Median
37	69.51	75.00



Mode

Mode is another measure of center, but can be misleading. The mode of a data set is the value of the observation that occurs most frequently.

For the internet usage data,

0, 0, 5, 7, 8, 9, 12, 14, 22, 33

the mode is 0, and is not a good representation of what is "typical."

There can be more than one mode or no mode. How?

For categorical data where the ordering of the categories is not relevant, mean and median are not appropriate measures of center, but mode can be used.

Example: Suppose a sample of 100 individuals contains 15 left-handers, 80 right-handers, and 5 ambidextrous individuals. The data entry is coded as follows: 1 = right-handed, 2 = left-handed, 3 = ambidextrous.

The sample mean is 1.25 - meaningless!
The mode is 1 = right-handed.

Comment: When the graphical display of a distribution shows two peaks, it is often described as *bimodal*. A bimodal distribution might be indicative of two natural groupings in the data. In a bimodal distribution, numerical measures of center often do not describe the "typical" value.

4. If you had data for all students in your school on the amount of money spent in the previous year on overnight stays in a hospital, probably the median and mode would be 0 but the mean would be positive. Explain why. Make reference to the definitions of median, mode, and mean in your explanation. (You can use an example to explain, if you want.)

5. A sample of 99 distances has a mean of 24 feet and a median of 24.5 feet. Unfortunately, it has just been discovered that an observation which was erroneously recorded as "30" actually had a value of "35." If we make this correction to the data, then:

- A) the mean remains the same, but the median is increased.
- B) the mean and median remain the same
- C) the median remains the same, but the mean is increased
- D) the mean and median are both increased.
- E) we do not know how the mean and median are affected without further calculations; but the standard deviation is increased.

6. When testing water for chemical impurities, results are often reported as "bdl," which means "below detection limit." The following are nine measurements of the amount of lead in a series of water samples taken from inner city households (measured in parts per million, ppm).

5 7 12 bdl 10 8 bdl 20 6

Which of the following is correct?

- A) The mean lead level in the water is about 10 ppm.
- B) The mean lead level in the water is about 8 ppm.
- C) The median lead level in the water is 7 ppm.
- D) The median lead level in the water is 8 ppm.
- E) Neither the mean nor the median can be computed because some values are unknown.

7. For the following histogram, what is the proper ordering of the mean, median, and mode?

- A) I = mean, II = median, III = mode
- B) I = mode, II = median, III = mean
- C) I = median, II = mean, III = mode
- D) I = mode, II = mean, III = median
- E) I = mean, II = mode, III = median

