2. Numerical Descriptive Measures of Central Tendency and Variability

Measures of central tendency

Usually, we focus our attention on two aspects of measures of central location:

- Measure of the central data point (mean, median)
- Measure of variability of the data (range, variance, standard deviation, IQR) about the central point

The central data point reflects the locations of all the actual data points.

Arithmetic mean

This is the most popular and useful measure of central location

Sample Mean **= Sum of Measurements/Number of measurements**

By tradition it is denoted by $\bar{X}$, that is,

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Here n is the sample size, and $X_1, \ldots, X_n$ are observations.

Arithmetic mean

Example 1. The mean of the *sample* of six measurements

$$\mathbf{7, 3, 9, -2, 4, 6}$$

is given by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{7 + 3 + 9 + (-2) + 4 + 6}{6} = 4.5$$

Median

The *median* of a set of measurements is the value that falls in the middle when the measurements are arranged in order of magnitude.

More specifically, median is a such value that splits a data set into halves: *at least half* of the data is *at or below* the median, and *at least half* of the data is *at or above* the median.

Odd number of observations

Example 3. Seven employee salaries (in 1000s) were recorded:

$$28, 60, 26, 32, 30, 26, 29.$$

Find the median salary.

First, sort the salaries. Then, locate the value in the middle:

$$26, 26, 28, \mathbf{29}, 30, 32, 60$$

The median is 29.

Even number of observations

Suppose one employee's salary of $31,000 was added to the group recorded before. Find the median salary.

First, sort the salaries. Then, locate the *two* values in the middle:

$$\mathbf{26, 26, 28, 29, 30, 31, 32, 60}$$

In this case median is $\mathbf{(29+30)/2 = 29.5}$.

Mode

- The mode of a set of measurements is the value that occurs most frequently.
- Set of data may have one mode (or modal class), or two or more modes.

Example 4. The manager of a men's store observes the waist size (in inches) of trousers sold yesterday: $\mathbf{31, 34, 36, 33, 28, 34, 30, 34, 32, 40}$.

The mode of this data set is $\mathbf{34}$ in.

Relationship among mean and median

- If a distribution is symmetrical, then mean, median and mode coincide
- If a distribution is non-symmetrical, and skewed to the left or to the right, the mean and median differ.

A positively skewed distribution ("skewed to the right") typically gives

$$median < mean$$

A negatively skewed distribution ("skewed to the left") typically gives

$$mean < median$$

Geometric mean

This is a measure of the average growth rate.

Let $\boldsymbol{R_i}$ denote the rate of return in period $\boldsymbol{i = 1, \dots, 2}$. The *geometric mean* of the returns $\boldsymbol{R_1, \dots, R_n}$ is the constant that produces the same terminal wealth at the end of period $\boldsymbol{n}$ as do the actual returns for the $\boldsymbol{n}$ periods, i.e.

$$\boldsymbol{R_g = \sqrt[n]{(1+R_1)(1+R_2)\dots(1+R_n)} - 1}$$

Geometric mean - example

Example 5. A firm's sales were \$1,000,000 three years ago. Sales have grown annually by 100%, 100%, -80%. Find the geometric mean rate of growth in sales.

Solution.

$$\boldsymbol{R_g = \sqrt[3]{(1+1)(1+1)(1-.8)} - 1 = -.07 = -7\%}$$

Note that the mean is $\mathbf{40\%}$, and the median is $\mathbf{100\%}$.

Measures of variability

Measures of central location fail to tell the whole story about the distribution. A question of interest still remains unanswered:

- How much *spread* out are the measurements about the central point?

Range

- The range of a set of measurements is the difference between the largest and smallest measurements.
- Its major advantage is the ease with which it can be computed.
- Its major shortcoming is its failure to provide information on the dispersion of the values between the two end points.

Sample variance

This measure of variability reflects the values of *all* the measurements.

The *sample variance* of a sample of n measurements with mean $\bar{X}$ is defined as

$$s^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

Standard deviation

The *sample standard deviation* of a set of measurements is the square root of the sample variance of the measurements.

Sample standard deviation: $s = \sqrt{s^2}$

The standard deviation can be used to

- compare the variability of several distributions
- make a statement about the general shape of a distribution

Empirical rule

If a sample of measurements has a bell-shaped distribution, the interval

- $[\bar{X} - s, \bar{X} + s\,]$ contains approximately 68% of the measurements,
- $[\bar{X} - 2s, \bar{X} + 2s\,]$ contains approximately 95% of the measurements,
- $[\bar{X} - 3s, \bar{X} + 3s\,]$ contains practically all the measurements: 99.7%

Range approximation

By the empirical rule we have:

$$4s < Range < 6s$$

Therefore, we get the following (conservative or from above) estimate of the standard deviation:

$$s \approx \frac{Range}{4}$$

Outliers and z-scores

- z-score of an observation X_i is given by

$$z_{X_i} = \frac{X_i - \bar{X}}{s}$$

- If an observation's z-score is greater than 3 in absolute value, that is, $|z_{X_i}| > 3$, then we call the observation an *outlier.*

Chebyshev theorem

Given any set of observation and a number $k > 1$, the fraction of these observations that lie within k standard deviations of their mean is at least

$$1 - \frac{1}{k^2}$$