Notation

We use an upper-case letter to denote a variable, and the corresponding lower-case letter to denote a general value of the variable. For example, when X is used to denote a variable, x is used to denote its particular value.

Suppose, our sample consists of n values of a variable X. We use $\sum x$ to denote the sum of all these values.

Describing the center of the data

There are three common measures to describe the central tendency of the sample data. These are:

1. Mean
2. Median
3. Mode

Mean

Let a sample of n values of variable X be taken. Data: $x_1, x_2, \cdots, x_n$. The sample mean is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example

Data: 4, 8, 5, 9, 15

$$\bar{x} = \frac{1}{n}\sum x = \frac{1}{5}(4 + 8 + 5 + 9 + 15) = 8.2$$

- Keep the result as fraction or decimal even if the variable is discrete.
- Mean cannot be calculated for categorical variables.

Median

It is the middlemost value in the sorted data. If n is an odd number, median is the middle value, i.e., $\left(\frac{n+1}{2}\right)$th value of the sorted data. If n is an even number, median is the average of the two middle values, i.e.,

$$\text{Median} = \frac{\frac{n}{2}\text{th value} + \left(\frac{n}{2}+1\right)\text{st value}}{2}$$

When sample size is large, approximately 50% values are less (more) than the median.

Example

Data: 4, 8, 5, 9, 15

Sorted data: 4, 5, 8, 9, 15

Median = 8

Example:

Data: 4, 8, 5, 9, 15, 13

Sorted data: 4, 5, 8, 9, 13, 15

Median $= \frac{1}{2}(8+9) = 8.5$

Mode

Mode is the value that occurs most frequently.
Sometimes two or more values occur with highest frequency.

- If there are two modes, the data is bimodal.
- If there are more than two modes, the data is multimodal.

If all values occur with equal frequency, there is no mode.

Example:

20 people were asked to give satisfaction rating after a restaurant meal on a scale of 1 (not satisfied) to 10 (extremely satisfied).
Data: 9, 3, 7, 5, 5, 10, 8, 9, 9, 10, 9, 8, 9, 6, 9, 8, 7, 7, 10, 6.
Mode = 9 (occurred 6 times in the data)

Which measure to use when

For categorical data, mode can be used.

For numerical (discrete or continuous) data, any of the three measures can be used. However, for mathematical reasons, mean or median is preferred.

Center for numerical data: mean or median?

Data: 2, 3, 4, 5, 7

Here, mean = 4.2, median = 4.

(Results are close. Mean is preferred because it is easy to calculate and mathematically solid.)

Data: 2, 3, 4, 5, 507 (the last value is an ‘outlier’)

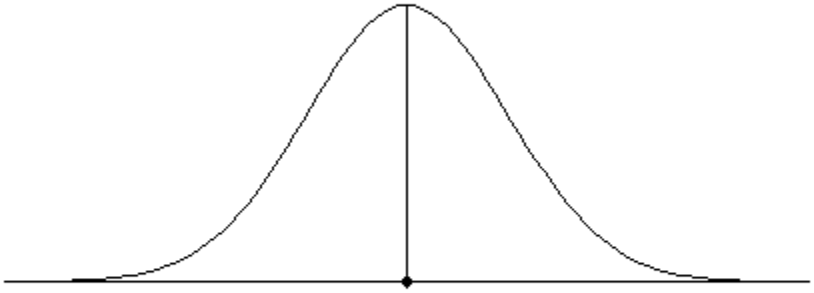
Here, mean = 104.2, median = 4.

Median represents the majority of the data. Mean represents neither the majority, nor the outlier. Median is preferred because it gives reasonable result.

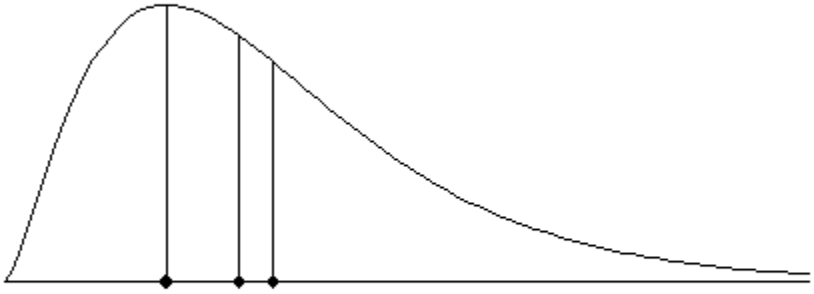
- When data have outliers, median is preferred.

Relation between mean, median and mode

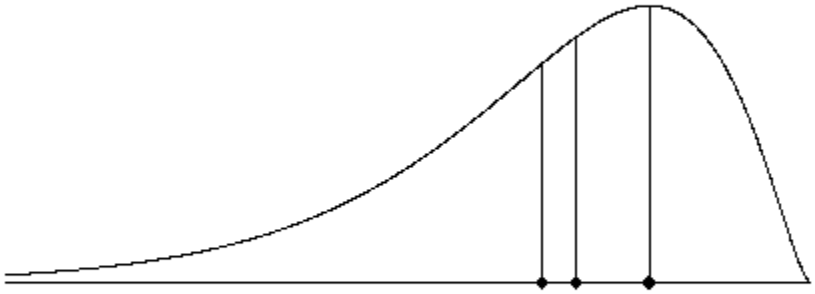
For symmetric bell-shaped distribution:
mean = median = mode (shown with a bullet point in the plot below).



For positively skewed distribution:
mean > median > mode (shown with 3 bullet points in the plot below).



For negatively skewed distribution:
mean < median < mode (shown with 3 bullet points in the plot below).



Exercise

Consider the data: 2, 4, 10, 10, 12, 6, 11, 12, 12, 8. Compute mean, median and mode. Comment on the shape of the distribution.

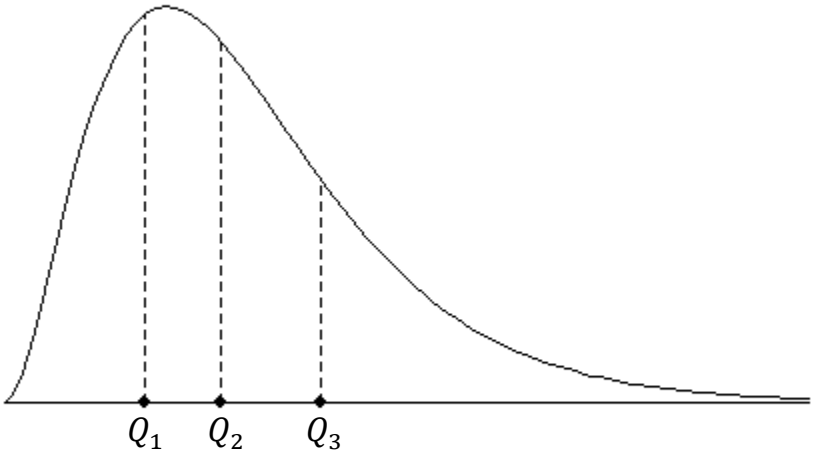
Solution

Sorted data: 2, 4, 6, 8, 10, 10, 11, 12, 12, 12.
Mean = 8.7
Median = (10 + 10)/2 = 10
Mode = 12

Since Mean < Median < Mode, the distribution is negatively skewed (or skewed to the left).

Quartiles

There are three quartiles that divide the total area of the histogram in 4 equal parts. The first quartile is denoted by Q_1. The second quartile Q_2 is the median. The third quartile is denoted by Q_3.



Example

20 customers' satisfaction ratings:

5, 1, 7, 3, 5, 10, 10, 9, 8, 8, 10, 8, 8, 9, 9, 8, 8, 10, 9, 9.

Sorted data:

1, 3, 5, 5, 7, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 10.

Median = (8+8)/2 = 8

Q_1 = (7+8)/2 = 7.5

Q_3 = (9+9)/2 = 9

Five-number summary

We often describe a set of data by using a **five-number summary**. The summary consists of (1) minimum (the smallest value) (2) the first quartile Q_1 (3) the median (4) the third quartile Q_3 and (5) maximum (the largest value).

Example

The five-number summary of the previous data: 1, 7.5, 8, 9, 10.

Percentiles

When data are arranged in increasing order, the p^{th} percentile is a value such that p percent of the values fall at or below the value, and $(100 - p)$ percent of the values fall at or above the value. There are 99 percentiles that divide the total area of the histogram in 100 equal parts.

Example

Let the 83^{rd} percentile = 39.5. This means 83% values in the data are less than 39.5, and (100 – 83) % = 17% values are more than 39.5.