Population

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a **parameter**. For example, All people living in India indicates the population of India.

There are different types of population. They are:

- Finite Population
- Infinite Population
- Existent Population
- Hypothetical Population

Finite Population

The finite population is also known as a countable population in which the population can be counted. Examples of finite populations are employees of a company, potential consumer in a market.

Infinite Population

The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

Existent Population

The existing population is defined as the population of concrete individuals. In other words, the population whose unit is available in solid form is known as existent population. Examples are books, students etc.

Hypothetical Population

The population in which whose unit is not available in solid form is known as the hypothetical population. A population consists of sets of observations, objects etc that are all something in common. Examples are an outcome of rolling the dice, the outcome of tossing a coin.

Sample

A sample is defined as a smaller and more manageable representation of a larger group. A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test. All the students in the class are population whereas the top 10 students in the class are the sample.

Population and Sample Formulas

$$\text{Population MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\text{Population Variance} = (\sigma x)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Sample MAD} = \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\text{Sample Variance} = (Sx)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Population Standard Deviation} = \sigma x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{Sample Standard Deviation} = Sx = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Central Tendency

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

Mode

The mode is the *most commonly occurring value* **in a distribution.**

Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This table shows a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Advantage of the mode:

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Limitations of the mode:

The are some limitations to using the mode. In some distributions, the mode may not reflect the centre of the distribution very well. When the distribution of retirement age is ordered from lowest to highest value, it is easy to see that the centre of the distribution is 57 years, but the mode is lower, at 54 years.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.

In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

Median

The median is the *middle value* **in distribution when the values are arranged in ascending or descending order.**

The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore

the median equals 56.5 years:

52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Advantage of the median:

The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Limitation of the median:

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

Mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

Looking at the retirement age distribution again:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

Advantage of the mean:

The mean can be used for both continuous and discrete numeric data.

Limitations of the mean:

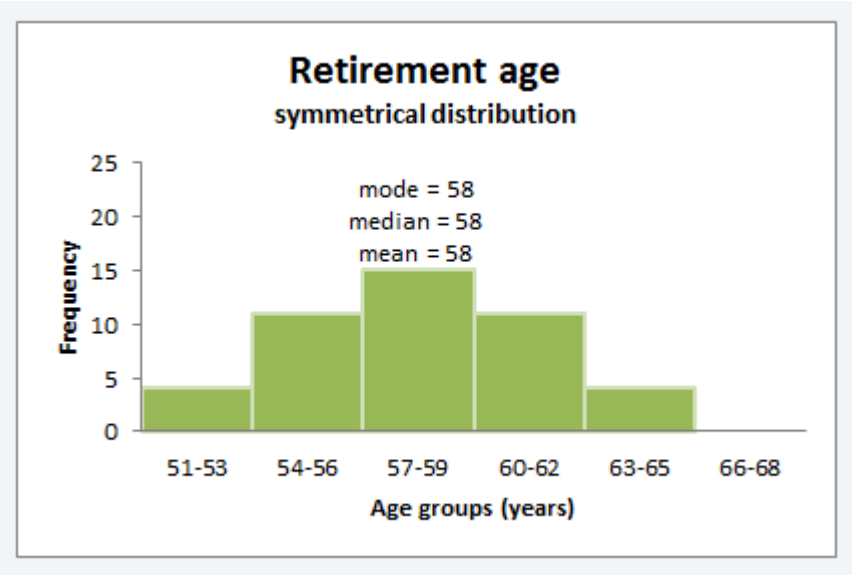
The mean cannot be calculated for categorical data, as the values cannot be summed.

As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

Skewed Distributions and the Mean and Median

Symmetrical distribution:

When a distribution is symmetrical, the mode, median and mean are all in the middle of the distribution. The following graph shows a larger retirement age dataset with a distribution which is symmetrical. The mode, median and mean all equal 58 years.

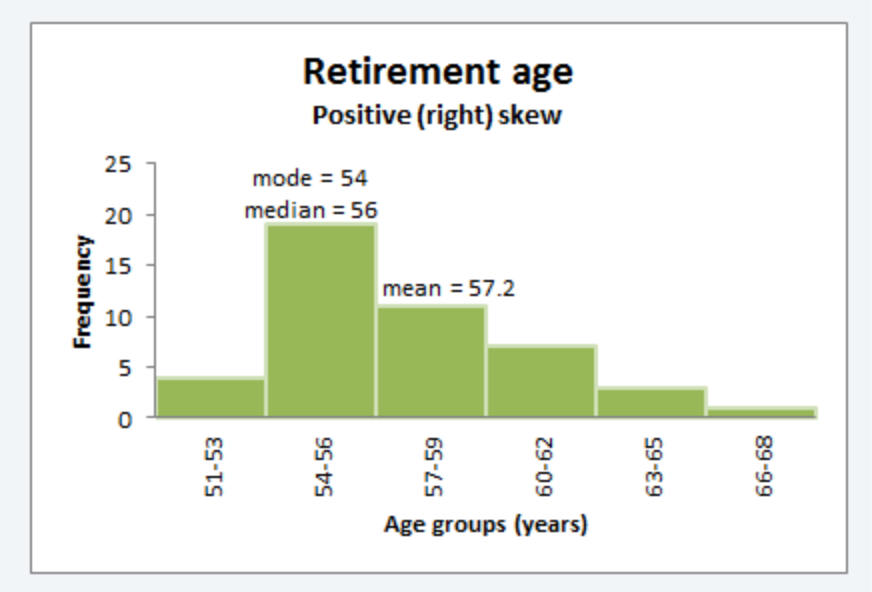


Skewed distributions:

When a distribution is skewed the mode remains the most commonly occurring value, the median remains the middle value in the distribution, but the mean is generally 'pulled' in the direction of the tails. In a skewed distribution, the median is often a preferred measure of central tendency, as the mean is not usually in the middle of the distribution.

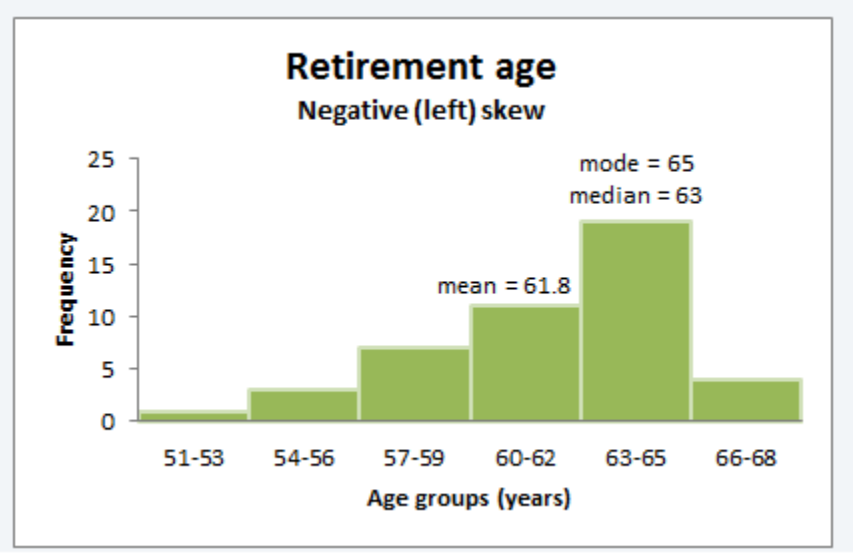
A distribution is said to be **positively or right skewed** when the tail on the right side of the distribution is longer than the left side. In a positively skewed distribution it is common for the mean to be 'pulled' toward the right tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be less than the mean value.

The following graph shows a larger retirement age data set with a distribution which is right skewed. The data has been grouped into classes, as the variable being measured (retirement age) is continuous. The mode is 54 years, the modal class is 54-56 years, the median is 56 years and the mean is 57.2 years.



A distribution is said to be **negatively or left skewed** when the tail on the left side of the distribution is longer than the right side. In a negatively skewed distribution, it is common for the mean to be 'pulled' toward the left tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be greater than the mean value.

The following graph shows a larger retirement age dataset with a distribution which left skewed. The mode is 65 years, the modal class is 63-65 years, the median is 63 years and the mean is 61.8 years.



Probability Distribution

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

The **central limit theorem** states that if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed. In a normal distribution, data is symmetrically distributed with no skew.