UNIT-4
Descriptive analysis

Content

Measures of Central Tendency: Mean, Median, mode, Measures of Dispersion: standard deviation, Co-variance, Coefficient of variation; correlation

What are the measures of 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.

Measures of central tendency help you find the middle, or the average, of a dataset. The 3 most common measures of central tendency are the mode, median, and mean.

- **Mode:** the most frequent value.
- **Median:** the middle number in an ordered dataset.
- **Mean:** the sum of all values divided by the total number of values.

What is the 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

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.

When to use the mode

The mode is most applicable to data from a nominal level of measurement. Nominal data is classified into mutually exclusive categories, so the mode tells you the most popular category. For continuous variables or ratio levels of measurement, the mode may not be a helpful measure of central tendency. That's because there are many more possible values than there are in a nominal or ordinal level of measurement. It's unlikely for a value to repeat in a ratio level of measurement.

What is the 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.

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.
What else do I need to know about the mean?
The population mean is indicated by the Greek symbol μ (pronounced 'mu'). When the mean is calculated on a distribution from a sample it is indicated by the symbol $\bar{x}$ (pronounced X-bar).

What is the 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.

When should you use the mean, median or mode?

The 3 main measures of central tendency are best used in combination with each other because they have complementary strengths and limitations. But sometimes only 1 or 2 of them are applicable to your dataset, depending on the level of measurement of the variable.

- The mode can be used for any level of measurement, but it's most meaningful for nominal and ordinal levels.
- The median can only be used on data that can be ordered – that is, from ordinal, interval and ratio levels of measurement.
- The mean can only be used on interval and ratio levels of measurement because it requires equal spacing between adjacent values or scores in the scale.

Levels of measurement	Examples	Measure of central tendency
Nominal	• Ethnicity • Political ideology	• Mode
Ordinal	• Level of anxiety • Income bracket	• Mode • Median
Interval and ratio	• Reaction time • Test score • Temperature	• Mode • Median • Mean

- The mean is the most frequently used measure of central tendency because it uses all values in the data set to give you an average.
- For data from skewed distributions, the median is better than the mean because it isn't influenced by extremely large values.
- The mode is the only measure you can use for nominal or categorical data that can't be ordered.

How do outliers influence the measures of central tendency?

Outliers are extreme, or atypical data value(s) that are notably different from the rest of the data.

It is important to detect outliers within a distribution, because they can alter the results of the data analysis. The mean is more sensitive to the existence of outliers than the median or mode.

Consider the initial retirement age dataset again, with one difference; the last observation of 60 years has been replaced with a retirement age of 81 years. This value is much higher than the other values, and could be considered an outlier. However, it has not changed the middle of the distribution, and therefore the median value is still 57 years.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 81

As the all values are included in the calculation of the mean, the outlier will influence the mean value.

(54+54+54+55+56+57+57+58+58+60+81 = 644), divided by 11 = 58.5 years

In this distribution the outlier value has increased the mean value.

Despite the existence of outliers in a distribution, the mean can still be an appropriate measure of central tendency, especially if the rest of the data is normally distributed. If the outlier is confirmed as a valid extreme value, it should not be removed from the dataset. Several common regression techniques can help reduce the influence of outliers on the mean value.

Measures of Dispersion

Measures of dispersion are non-negative real numbers that help to gauge the spread of data about a central value. These measures help to determine how stretched or squeezed the given data is. There are five most commonly used measures of dispersion. These are range, variance, standard deviation, mean deviation, and quartile deviation.

The most important use of measures of dispersion is that they help to get an understanding of the distribution of data. As the data becomes more diverse, the value of the measure of dispersion increases. In this article, we will learn about measures of dispersion, their types along with examples as well as various important aspects related to these measures.
What is Measure of Dispersion in Statistics?

Measures of dispersion help to describe the variability in data. Dispersion is a statistical term that can be used to describe the extent to which data is scattered. Thus, measures of dispersion are certain types of measures that are used to quantify the dispersion of data.

Measures of Dispersion Definition

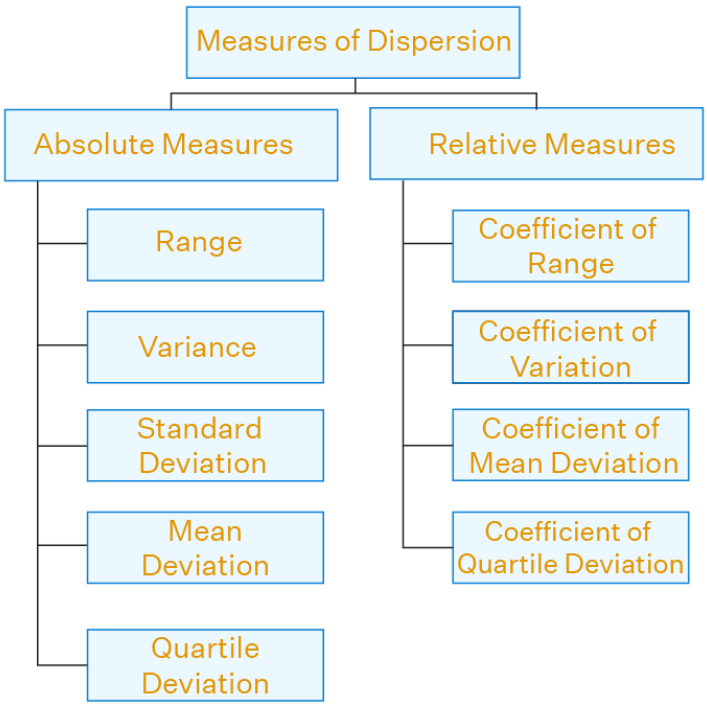
Measures of dispersion can be defined as positive real numbers that measure how homogeneous or heterogeneous the given data is. The value of a measure of dispersion will be 0 if the data points in a data set are the same. However, as the variability of the data increases the value of the measures of dispersion also increases.

Measures of Dispersion Example

Suppose we have two data sets A = {3, 1, 6, 2} and B = {1, 5, 9, 10}. The variance(population) of A is 3.5 and the variance(population) of B is 12.68. This implies that data set B is more variable than data set A. Thus, the variance helps to draw a comparison between the two data sets A and B on the basis of variability.

Types of Measures of Dispersion

The measures of dispersion can be classified into two broad categories. These are absolute measures of dispersion and relative measures of dispersion. Range, variance, standard deviation and mean deviation fall under the category of absolute measures of deviation. These measures have the same unit as the data that is being scrutinized. Coefficients of dispersion are relative measures of deviation. Such dispersion measures are always dimensionless. The upcoming sections will further elaborate on these measures.



Absolute Measures of Dispersion

If the dispersion of data within an experiment has to be determined then absolute measures of dispersion should be used. These measures usually express variations in a data set with respect to the average of the deviations of the observations. The most commonly used absolute measures of deviation are listed below.

Range: Given a data set, the rangecan be defined as the difference between the maximum value and the minimum value.

Range

Range refers to the difference between each series' minimum and maximum values. The range offers us a good indication of how dispersed the data is, but we need other measures of variability to discover the dispersion of data from central tendency measurements. A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

$$\text{Range} = X_{max} - X_{min}$$

Merits of Range

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

Variance: The average squared deviation from the mean of the given data set is known as the variance. This measure of dispersion checks the spread of the data about the mean.

Standard Deviation: The square root of the variance gives the standard deviation. Thus, the standard deviation also measures the variation of the data about the mean.

Standard Deviation

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation.

Merits of Standard Deviation

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations
- Suitable for further mathematical treatment
- Least affected by the fluctuation of the observations
- The standard deviation is zero if all the observations are constant
- Independent of change of origin

Demerits of Standard Deviation

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale

Mean Deviation: The mean deviation gives the average of the data's absolute deviation about the central points. These central points could be the mean, median, or mode.

Mean Deviation

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \ldots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is

$$\text{Mean deviation from average } A = 1/n \, [\sum_i |x_i - A|]$$

For a grouped frequency, it is calculated as:

$$\text{Mean deviation from average } A = 1/N \, [\sum_i f_i |x_i - A|], \; N = \sum f_i$$

Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

Merits of Mean Deviation

- Based on all observations
- It provides a minimum value when the deviations are taken from the median
- Independent of change of origin

Demerits of Mean Deviation

- Not easily understandable
- Its calculation is not easy and time-consuming
- Dependent on the change of scale
- Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

Quartile Deviation: Quartile deviation can be defined as half of the difference between the third quartile and the first quartile in a given data set.

Quartile Deviation

The quartiles divide a data set into quarters. The first quartile, (Q_1) is the middle number between the smallest number and the median of the data. The second quartile, (Q_2) is the median of the data set. The third quartile, (Q_3) is the middle number between the median and the largest number.

Quartile deviation or semi-inter-quartile deviation is

$$Q = ½ \times (Q_3 - Q1)$$

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores 50% of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

Relative Measures of Dispersion

If the data of separate data sets have different units and need to be compared then relative measures of dispersion are used. The measures are expressed in the form of ratios and percentages thus, making them unit less. Some of the relative measures of dispersion are given below:

Coefficient of Range: It is the ratio of the difference between the highest and lowest value in a data set to the sum of the highest and lowest value.

Coefficient of Variation: It is the ratio of the standard deviation to the mean of the data set. It is expressed in the form of a percentage.

Coefficient of Mean Deviation: This can be defined as the ratio of the mean deviation to the value of the central point from which it is calculated.

Coefficient of Quartile Deviation: It is the ratio of the difference between the third quartile and the first quartile to the sum of the third and first quartiles.

Measures of Dispersion and Central Tendency

Both measures of dispersion and measures of central tendency are used to describe data. The table given below outlines the difference between the measures of dispersion and central tendency.

Measures of Dispersion	**Central Tendency**
When we want to quantify the variability of data we use measures of dispersion.	Measures of central tendency help to quantify the data's average behaviour.
Measures of dispersion include variance, standard deviation, mean deviation, quartile deviation, etc.	Measures of central tendency are mean, median, and mode.

Difference between Standard Deviation and Coefficient of Variation

[Click Here for Sample Questions]

When measuring the spread of values in a dataset, the coefficient of variation and the standard deviation are both used. The table below summarises the key differences between the two measures.

It is a relative measure of dispersion	It is an absolute measure of dispersion

It measures the ratio of the standard deviation to the mean	It measures how far a data point lies from the mean
Coefficient of variation is usually used to compare the variation of different data sets	Standard deviation is used to measure the dispersion of data in a single data set

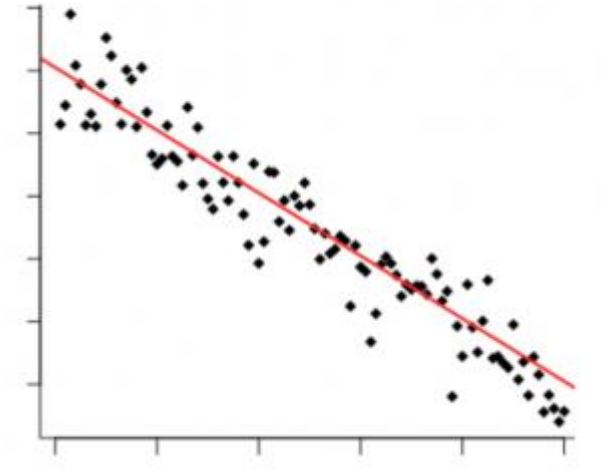
What is a Correlation?

A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. Using a scatterplot, we can generally assess the relationship between the variables and determine whether they are correlated or not.

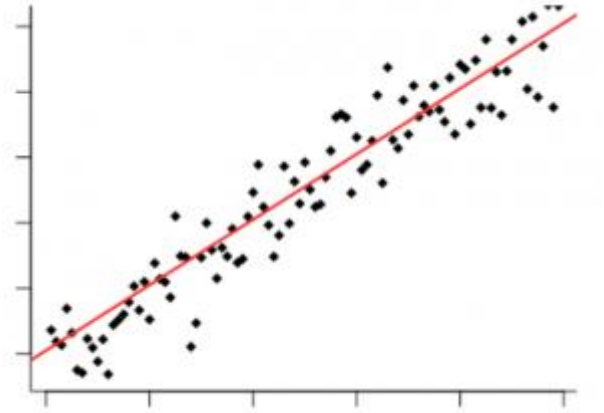
The correlation coefficient is a value that indicates the strength of the relationship between variables. The coefficient can take any values from -1 to 1. The interpretations of the values are:

- **-1:** Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).
- **0:** No correlation. The variables do not have a relationship with each other.
- **1:** Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

One of the primary applications of the concept in finance is portfolio management. A thorough understanding of this statistical concept is essential to successful portfolio optimization.
When a correlation is **positive** there is an increase of value.



If the the correlation is **negative** it will show a decrease of value:



When the variables show minimum difference and the line that goes through all the dots in the chart has almost no difference we are talking about a very strong correlation. The highest correlation number is 1.0 which means all dots are aligned perfectly.

Let's use a sample with this. I own a bicycle shop and I want to see if there is a correlation between temperature during the months and my sales amount.

Interpreting a correlation coefficient

The value of the correlation coefficient always ranges between 1 and -1, and you treat it as a general indicator of the strength of the relationship between variables.

The **sign** of the coefficient reflects whether the variables change in the same or opposite directions: a positive value means the variables change together in the same direction, while a negative value means they change together in opposite directions.

The **absolute value** of a number is equal to the number without its sign. The absolute value of a correlation coefficient tells you the magnitude of the correlation: the greater the absolute value, the stronger the correlation.

There are many different guidelines for interpreting the correlation coefficient because findings can vary a lot between study fields. You can use the table below as a general guideline for interpreting correlation strength from the value of the correlation coefficient. While this guideline is helpful in a pinch, it's much more important to take your research context and purpose into account when forming conclusions. For example, if most studies in your field have correlation coefficients nearing .9, a correlation coefficient of .58 may be low in that context.

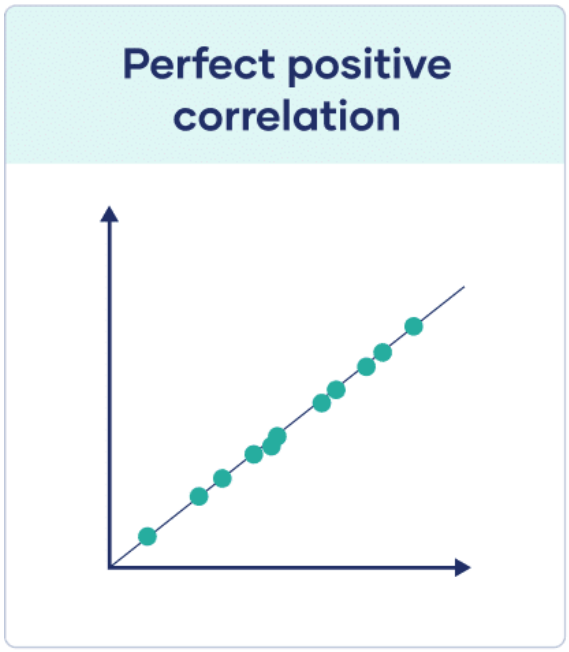
Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

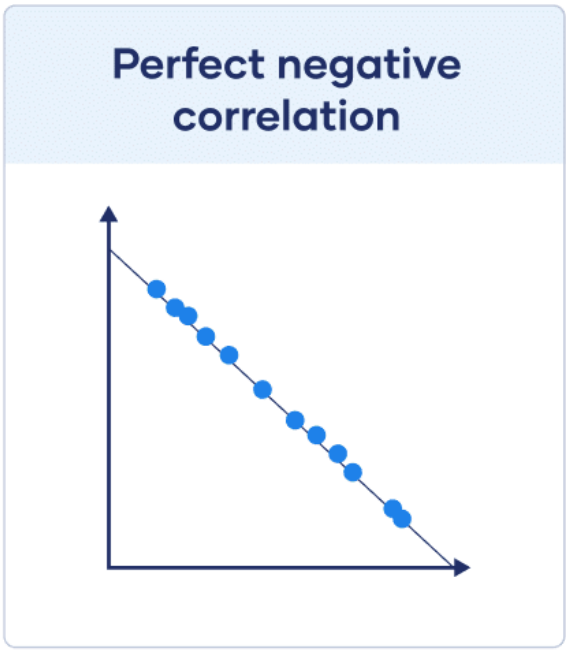
Visualizing linear correlations

The correlation coefficient tells you how closely your data fit on a line. If you have a linear relationship, you'll draw a straight line of best fit that takes all of your data points into account on a scatter plot.

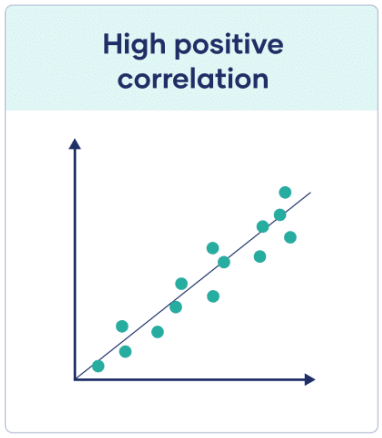
The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

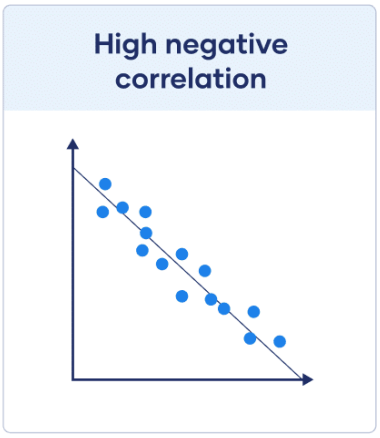
If all points are perfectly on this line, you have a **perfect** correlation.



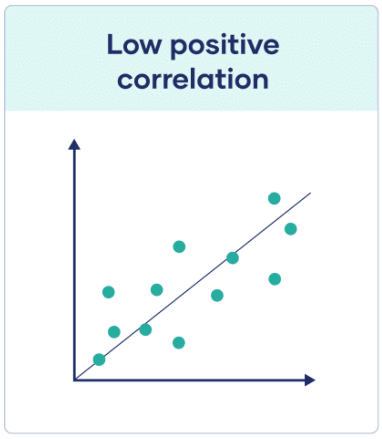


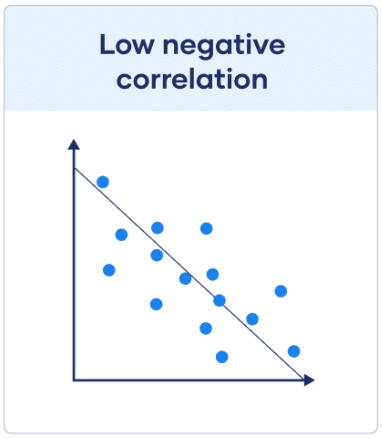
If all points are close to this line, the absolute value of your correlation coefficient is **high**.





If these points are spread far from this line, the absolute value of your correlation coefficient is **low**.





Note that the steepness or slope of the line isn't related to the correlation coefficient value. The correlation coefficient doesn't help you predict how much one variable will change based on a given change in the other, because two datasets with the same correlation coefficient value can have lines with very different slopes.