Variation or dispersion

A measure of central tendency tells us where the centre of data lies but does not tell us how the observations are scattered around this central value. Two sets of data may have the same arithmetic mean, but the items in one set may scatter widely around its average, while in the other case, observations may be close to the arithmetic mean. This implies that a measure of central tendency alone cannot describe a distribution adequately.

Dispersion is the extent to which values of a distribution are spread out or scattered. Dispersion implies that observations in a set of data are not all the same. There may be small differences or large differences between observations in a distribution. If the observations are all the same, then there is no variability. If there are small differences between the observations, then the variability is small. If there are large differences between the observations, then the variability is large.

Measures of variation or dispersion

Measures of dispersion provide a quantitative measure of the degree to which observations in a distribution are spread out or clustered together. The values of a distribution may be widely spread out or may be closely clustered around the arithmetic mean. The dispersion of values is indicated by the extent to which these values tend to be spread out over an interval rather than cluster closely around an average.

A good measure of dispersion serves three purposes.

(1) It tells whether the values of a distribution are clustered close together or are widely spread out over a large distance.

(2) It tells us how much distance to expect between one observation and another or how much distance to expect between an individual data value and the arithmetic mean.

(3) It measures how well an individual value represents the entire distribution.

Classification of measures of Dispersion

Measures of dispersion are classified into two categories:

1. **Distance Methods**

The distance measures describe the spread or dispersion of values in terms of difference among certain values in the data set. Distance measures include.

(i) The range – This is the difference between the lowest and the largest values in the data set.

(ii) Interquartile range – This is the difference between the lower quartile and the upper quartile. It measures the range covered by the middle 50% of the observations. It is also used to identify outliers.

(iii) Semi interquartile range – It measures the range covered by the middle 25% of the data values in the data set.

These measures of variation do not indicate how values in a data set are scattered around a central value.

2. <u>Average deviation measures</u>

The average deviation measures describe the average deviation of observations from the arithmetic mean. They measure the average amount by which values in a data set deviate from a measure of central tendency.

Average deviation measures include:

1. Mean absolute deviation
2. Variance
3. Standard deviation
4. Coefficient of variation.

If we calculate the difference between each data value and the arithmetic mean, we would have both positive and negative differences. The average of these differences will always be zero.

Since average deviation of individual values in the data set from their actual arithmetic mean is always equal to zero, such a measure would not indicate any variation even when there is variation among observations. The problem can be solved in two ways.

(i) Ignore the signs of the deviation by taking absolute values.

For instance, the absolute value of 3 – 5 is written as $|3-5|$ = 2 i.e the positive difference.

(ii) Square the deviations because the square of a negative number is positive

$(3-5)^2 = (-2)^2 = 4$

Mean absolute deviation

The mean absolute deviation of a set of observations is the arithmetic mean of the absolute deviations of the observations from their arithmetic mean. It measures how closely a set of observations cluster around their arithmetic mean.

Variance

It is also called the mean squared average deviation. It is based on the squared deviations of the observed values in the data set from their arithmetic mean.

It is the average of the squared deviations of observation in a data set from their arithmetic mean. If the values are near the arithmetic mean, the variance will be small. In contrast, if the values are far from the arithmetic mean, the variance will be large.

Standard deviation

The variance is difficult to interpret because it is expressed in square units. To get an interpretable measure of variation expressed in the units of the original data, we take a positive square root of the variance to get the standard deviation. Standard deviation is the square root of the average of the squared deviations of observations from the arithmetic mean of a data set.

Finding the square root of the variance puts the standard deviation; in the same units as the original data. A small standard deviation tells us that the observations cluster closely around their arithmetic mean, while a large standard deviation says that the observations are much more scattered.

The larger the variance or standard deviation, the more the data values are dispersed.

Finding the variance and standard deviation of a distribution.

I. Find the arithmetic mean of the data.
II. Subtract the arithmetic mean from each data value.
III. Square the differences.
IV. Find the sum of the squares.
V. Divide the sum by n to get the variance.
VI. Take the square root of the variance to get the standard deviation.

Example I

Determine the mean deviation for the following distribution: 15, 30, 45, 50, 100.

Example II

Determine the mean deviation for the following distribution.

x	15	30	45	50	100
f	2	8	4	4	2

Example 1

Find the variance and the standard deviation of 2, 4, 6 and 8.

Example 2

Find the variance and the standard deviation of the following distribution:

x	2	4	6	8
f	3	2	4	1

Example 3 Find the variance and the standard deviation of the following distribution:

Age group	20-30	30-40	40-50	50-60	60-70	70-80
No. of persons	4	6	5	4	4	2

Coefficient of variation

It is a measure of relative dispersion and is computed by expressing the standard deviation of a distribution as a percentage of the arithmetic mean.

$$CV = \frac{\delta}{x} x100$$

It is used for

(i) Comparing the dispersions of two or more data sets expressed in different units of measurement e.g. Height in meters and weight in Kg.

(ii) Comparing the dispersions of data sets that are in the same unit of measurement but the arithmetic means of the data sets are not the same e.g. comparing the variability in sales of large firms and small firms.

Find the coefficient of variation for the following:

Subject	**Arithmetic Mean**	**Standard deviation**
Statistics	45	5.5
Accounting	18.5	2.6

Which subject has greater variability?

Outliers

An outlier is an extremely high or extremely low data value when compared with the rest of the data values in the data set.

These are several reasons why outliers may occur:

(i) The data value may have resulted from a measurement or observational error i.e. the researcher measured the variable incorrectly.

(ii) The data value may have resulted from a recording error, that is, it may have been written or typed incorrectly.

(iii) The value may have been obtained from a subject that is not in the defined population.

(iv) The data value might be a legitimate value that occurred by chance.

Procedure for Identifying Outliers

Step 1 Arrange the data in order and find Q_1 and Q_3.

Step 2 Find the interquartile range.

Step 3 Multiply the interquartile range by 1.5.

Step 4 Subtract the value obtained in step 3 from Q_1 and add the value to Q_3. The values obtained are called outlier boundaries.

Step 5 Check the data set for any data value that is smaller than $Q_1 - 1.5\,IQR$ or larger than $Q_3 + IQR$

Determination of Quartiles for Individual Series

Step 1 Arrange the data in ascending order i.e. from the lowest to the highest.

Step 2 Find the median of the distribution. This is the value of Q_2.

Step 3 Find the median of the data values that fall below Q_2. This is the value of Q_1.

Step 4 Find the median of the data values that fall above Q_2. This is the value of Q_3.

Box Plot

It is also called box and whisker plot. It is a graphical summary of a data set which is based on five numbers. These numbers are the lowest number that is not an outlier or the boundary for outliers, the lower quartile, the median, the upper quartile, and the highest value that is not an outlier or the boundary for outliers.

The first step in drawing a box plot is to create an appropriate scale along the horizontal axis. Next, we draw a box that starts at Q1 and ends at Q3. Inside the box we place a vertical line to represent the median. Finally, we extend horizontal lines from the box out to the minimum value and the maximum value.

A distribution is skewed if:

1) Median is not located at the centre of the box
2) The dashed lines to the right and to the left of the box are unequal.

ILLUSTRATION

Check the following data set for outliers and draw a Box Plot.
5, 6, 18, 13, 15, 50, 12, 22.

SYMMETRY AND SKEWNESS

Symmetry and skewness describe the shape of a distribution i.e how data values are distributed between the extreme values. Data values can be evenly distributed along the horizontal axis or concentrated at either the lower end or the high end of the scale along the horizontal axis.
The shape of a distribution can be examined with the help of a frequency curve. A frequency curve is a graph of frequencies against class midpoints.
Consider the following distribution.

x	25	35	45	55	65
f	4	7	8	7	4

Required:

i) Represent the distribution using a frequency curve.
ii) Determine the Arithmetic Mean, Median, and Mode.

The vertical line divides the distribution into two equal parts; Part A and Part B. The vertical line is a line of symmetry and therefore the frequency curve is symmetrical in nature. Part A is a minor image of Part B.

Observations

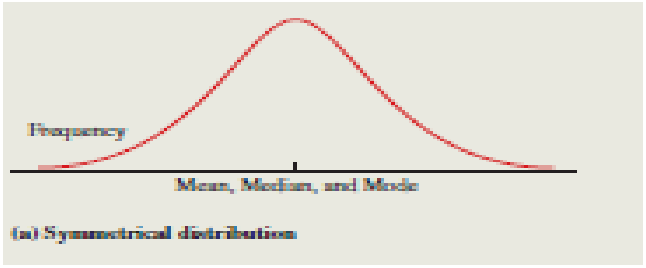
1) Mean = Mode = Median. The median, mode and the mean appear at the centre of the distribution.
2) The data values are evenly spread around the measures of central tendency. This implies that the data values below the mean, mode and median are a mirror image of those above.
3) The frequency curve is symmetrical about the arithmetic mean, that is, observations equidistant from the mean to the left and right have equal frequencies.

Conclusion

The distribution is symmetrical. A distribution that deviates from symmetry is a skewed distribution.

Symmetrical distribution – Zero Skewness

This is a distribution of data in which the data values are evenly distributed around the measures of central tendency. In such a case the mean, mode and median are always equal and are located at the centre of the distribution. The distribution has the same shape on either side of the centre i.e the right half of the distribution is a mirror image of the left half of the distribution: A symmetric distribution is bell shaped as shown below:



A symmetrical distribution is represented by a curve that can be divided by a vertical line into two parts that are mirror images.

Skewed distribution – Non symmetrical distribution

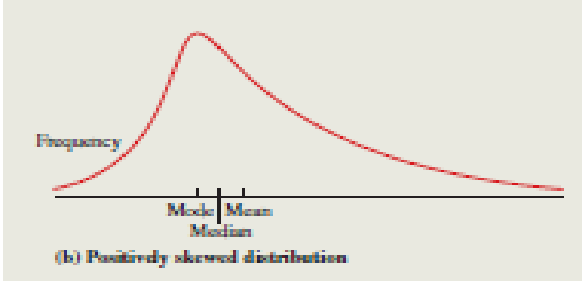
Skewness exists if there is lack of symmetry. The effect of this is that the mode, mean, and median are different. Values in a skewed distribution are concentrated at either the low end or the high end of the scale along the horizontal axis. A skewed distribution is therefore represented by a frequency curve that lacks symmetry.

Skewed distribution denotes that data are sparsely distributed at one end of the distribution and piled up at the other end. The mode of the distribution is at the apex (high point) of the frequency curve. The skewed portion is the long thin part of the curve where data are sparsely distributed. The arithmetic mean tends to be located toward the longer tail of the distribution because the mean is particularly affected by extreme values. The median always occurs between the mode and the arithmetic mean. A distribution may be positively skewed or negatively skewed.

Positively skewed distribution

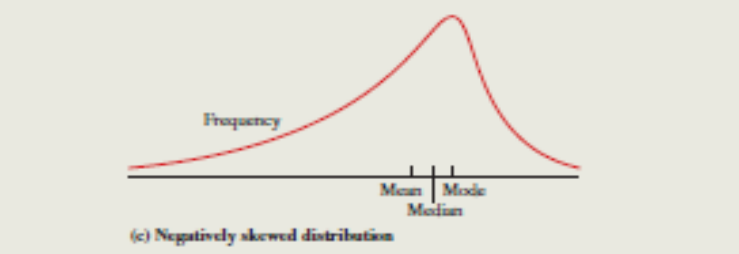
This is a distribution which is skewed towards the right. In a positively skewed distribution, the right tail of the frequency curve is more pronounced than the left tail i.e the tail extends more to the right than to the left. The mean is usually greater than the mode and median. There are more observations below the arithmetic mean than above the arithmetic mean. The arithmetic mean is the largest of the three measures of central tendency because the arithmetic mean is influenced more than the median or mode by a few extremely high values. In a positively skewed distribution, the majority of the data values cluster at the lower end of the distribution.

A positively skewed distribution is represented by the following diagram.



Negatively skewed distribution

In a negatively skewed distribution, extreme values in the data set move towards the left tail. Since arithmetic mean of values in any data set gets affected due to the presence of outliers in one tail of a distribution, the arithmetic mean shifts substantially in the direction of these values. If a distribution is negatively skewed the arithmetic mean is the lowest of the three measures. The median is greater than the arithmetic mean and the mode is the largest of the three measures. The mean is influenced by a few extremely low observations. A negatively skewed distribution is represented by the following diagram. In a negatively skewed distribution the majority of the data values cluster at the upper end of the distribution.



Measures of Skewness

Measures of skewness describe the degree to which values in a data set deviate from symmetry on both the sides of the central value. In a symmetric distribution, the data values are evenly distributed on both sides of the mean. The mean, median ad mode are the same and are at the centre of the distribution. When a distribution is extremely skewed, the value of the arithmetic mean will be pulled toward the tail.

The degree of skewness in a distribution can be measured both in the absolute and relative sense.

Absolute Skewness

For a skewed distribution, mean moves away from the mode, so the distance between mean and mode is used to measure the degree of skewness. The larger the difference between mean and mode, the more is the skewness.

Thus, absolute skewness = mean – mode. For a positively skewed distribution SK is a positive value, otherwise a negative value.

Limitations

- The absolute skewness cannot be used for comparing skewness of two or more distributions having different units of measurement.
- It cannot be used to compare distributions of unequal size.

Relative Measures of Skewness

It is used to compare skewness of two or more distributions.

Karl Pearson coefficient of skewness

$$S_{kp}=\frac{Mean-Mode}{Standard\ deviation}$$

Income distribution tend to be positively skewed, since there is a lower limit of O, but practically no upper limit on how much a select few might earn.