Location and dispersion are two important quantitative concepts. But these two measures fail to describe the shape such as the shape of histogram or frequency polygon of the data set. Two data sets may have identical means and identical variance, but their graphical shapes may be different. So, the graphical shape of a distribution frequently plays an important role in determining the appropriate method of statistical analysis.

1. Moments: Moments are constant which used to determine some characteristics/properties of frequency distribution. These characteristics are-

- Skewness
- Kurtosis

These characteristics/properties are known as shape characteristics of the distribution.

r^{th} moments can be written as,

$$\mu_r = \frac{\sum(X_i - \bar{X})^r}{N} \quad ; r = 1,2,3,4, \ldots$$

$$If\ r = 1, \mu_1 = First\ moment = \frac{\sum(X_i - \bar{X})^1}{N}$$

$$If\ r = 2, \mu_2 = Second\ moment = \frac{\sum(X_i - \bar{X})^2}{N}$$

$$If\ r = 3, \mu_3 = Third\ moment = \frac{\sum(X_i - \bar{X})^3}{N}$$

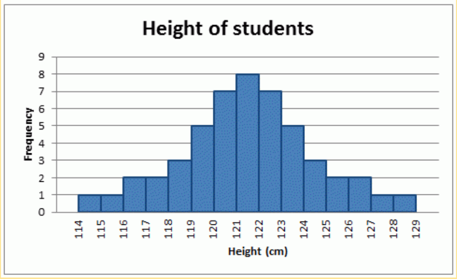
$$If\ r = 4, \mu_4 = Fourth\ moment = \frac{\sum(X_i - \bar{X})^4}{N}$$

2. Skewness: Skewness is the measure of asymmetry or distortion to the symmetric bell-shaped graph in a set of data. Other word, Skewness is a measurement which refers to the lack of symmetry of a distribution. A skewed distribution satisfies the following properties:

a) $Mean \neq Median \neq Mode$

b) The first quartile and third quartiles are not equal distance from the median.

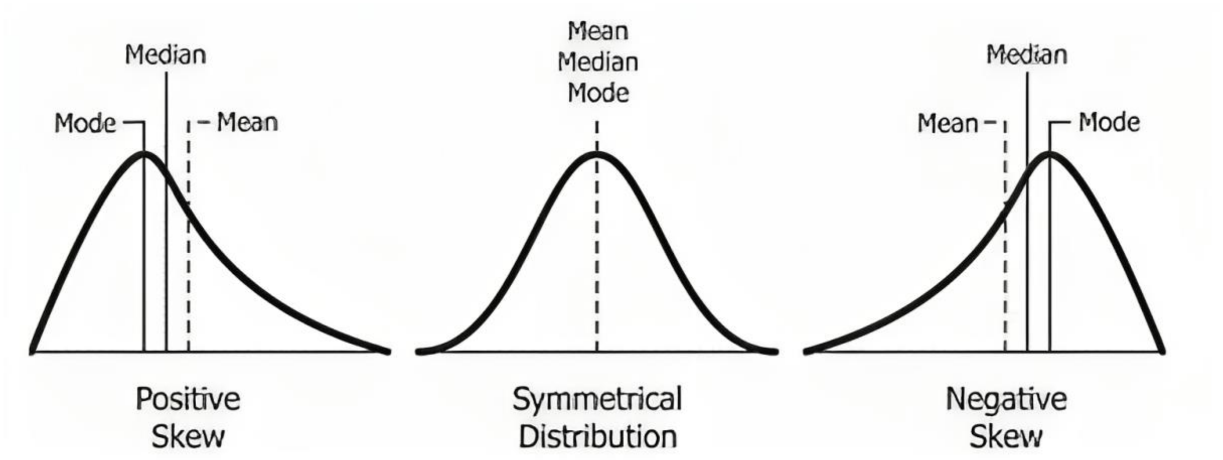
Symmetric Distribution: If the values of a variable in a frequency distribution are equally distributed at the both ends about the mean, the distribution is said to be symmetrical distribution. In this case $\boldsymbol{Mean = Median = Mode}$



3. Types of Skewness: There are two types of skewness or lack of symmetry occurs:

a) Positive skewness: When $Mean > Median > Mode$.

b) Negative skewness: When $Mean < Median < Mode$



4. Coefficient of Skewness: Coefficient of skewness can be written as,

$$S_k = \frac{3(Mean - Median)}{Standard\ Deviation}$$

$S_k > 0$: ***Positively skewed***
$S_k < 0$: ***Negatively skewed***
$S_k = 0$: ***Symmetric***

Math: Calculate the coefficient of skewness: $15, 18, 2, 6, 4$

Solution: Organize data into ascending order: $2, 4, 6, 15, 18$

$$Mean, \bar{X} = \frac{\sum X_i}{n} = 9$$

$$Median = \left(\frac{n+1}{2}\right)^{th} observation = \left(\frac{6}{2}\right)^{th} obs. = 3rd\ obs. = 6$$

$$SD = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}} = \sqrt{\frac{(15-9)^2 + (18-9)^2 + (2-9)^2 + (6-9)^2 + (4-9)^2}{4}} = 7.07$$

$$\therefore S_k = \frac{3(Mean - Median)}{SD} = 1.27$$

Since $S_k > 0$, thus the distribution is positively skewed distribution.

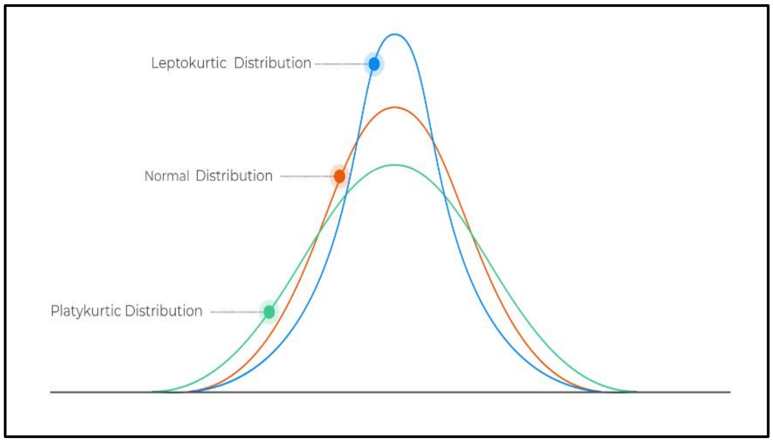
5. Kurtosis: Kurtosis refers to the degree or peaked or flatness of a distribution.

Kurtosis can be divided into three types:

a) Leptokurtic distribution

b) Mesokurtic/Normal/Standard distribution

c) Platykurtic distribution



6. Coefficient of kurtosis: Kurtosis can be calculated as,

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$\boldsymbol{If\ \beta_2 > 3: Leptokurtic}$
$\boldsymbol{If\ \beta_2 = 3: Mesokurtic}$
$\boldsymbol{If\ \beta_2 < 3: Platykurtic}$

Math: Calculate the coefficient of kurtosis: $3, 4, 2, 4, 6, 2, 5$ [Ans: Platykurtic]

Extra:

a) Marks of 10 students in a class: $15, 25, 48, 50, 65, 95, 18, 85, 75, 55$. Calculate coefficient of skewness and interpret the result.

b) Time of reading newspaper of 6 people (in hour) $3, 4, 2, 4, 6, 2, 5$. Calculate coefficient of kurtosis.

1. Box-Whisker Plot: A box-whisker plot, also known as a box plot, is a graphical representation of the distribution of a dataset. It displays the five-number summary of the data, which includes the minimum, first quartile (Q_1), median (second quartile or Q_2), third quartile (Q_3), and maximum. The plot consists of a rectangular "box" that represents the interquartile range (IQR) between Q_1 and Q_3, and "$whiskers$" that extend from the box to the minimum and maximum values, indicating the spread of the data. Outliers may also be depicted as individual points or dots outside the whiskers. Box plots provide insights into the central tendency, spread, and potential outliers of the data distribution.

- 1st Quartile (Q_1)
- 2nd Quartile (Q_2)
- 3rd Quartile $Q_3)$
- Smallest/minimum value
- Highest/Maximum value
- Inter Quartile Range $(IQR = Q_3 - Q_1)$

2. Outlier: An outlier is an observation or data point that significantly differs from the rest of the data in a dataset. Outliers are values that are unusually high or low compared to the other values in the dataset and can skew the overall analysis or interpretation of the data.

3. What are some factors or circumstances that can lead to the presence of outliers in a dataset?

Outlier may result from:

— errors in data collection,

— measurement variability,

— genuine extreme observations

— Sampling Bias

— Data Entry Mistakes

4. Why is the detection of outliers crucial?

Outliers can impact statistical measures such as the mean, potentially pulling it towards their direction. They can also affect the distribution of the data, leading to misleading conclusions if not properly identified and handled. Outlier detection is an important step in data analysis to ensure accurate and reliable insights.

5. Inner fences:

- Use: Inner fence is statistical boundaries used in outlier detection methods, particularly in the context of box plots or other data distributions. They help identify potential outliers within a dataset.
- Inner Fence: The inner fence is a range around the interquartile range (IQR) that defines the limits within which data points are considered typical or "normal". It is typically set at 1.5 times the IQR above the third quartile (Q_3) and below the first quartile (Q_1). Data points outside the inner fence are considered potential mild outliers/suspected outliers. Formula: $Inner\ fences\ = (Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$

6. Use of box-whisker plot:

a) To get idea of the shape of the distribution

b) To detect outliers from the data.

c) To get idea about the spread ness of the data set.

Math: The monthly starting salaries in dollar for a random sample 12-business school graduates are as follows: $2900, 2765, 2960, 2890, 2880, 2720, 2930, 2950, 2860, 3060, 3260, 3525$ Compute the summary of the Box-Whisker plot and identify any outliers (if present).

Solution:

First, we arrange the data set in order from smallest to largest:

$$2720, 2765, 2860, 2880, 2890, 2900, 2930, 2950, 2960, 3060, 3260, 3525$$

Now, we calculate the box-whisker plot summary.

$\boldsymbol{a}) \; Position\ of\ Q_1 = \frac{(i \times n)}{4} = \frac{(1 \times 12)}{4} = 3$

Since the position value is an integer,

$\Rightarrow Q_1 = \frac{3^{rd} Observation + 4^{th} Observation}{2} = \frac{(2860 + 2880)}{2} = 2870$

$\boldsymbol{b}) \; Position\ of\ Q_2 = \frac{(i \times n)}{4} = \frac{(2 \times 12)}{4} = 6$

Since the position value is and integer,

$\Rightarrow Q_2 = \frac{(6^{th} Observation + 7^{th} Observation)}{2} = \frac{(2900 + 2930)}{2} = 2915$

$\boldsymbol{c}) \; Position\ of\ Q_3 = \frac{(i \times n)}{4} = \frac{(3 \times 12)}{4} = 9$

Since the position value is an integer,

$\Rightarrow Q_3 = \frac{(9^{th} Observation + 10^{th} Observation)}{2} = \frac{(2960 + 3060)}{2} = 3010$

$\boldsymbol{d}) \; Smallest\ value = 2720$

$\boldsymbol{e}) \; Largest\ value = 3535$

$\boldsymbol{f}) \; IQR = Q_3 - Q_1 = 3010 - 2870 = 140$

Interpretation of IQR:

To detect outlier, we need to calculate Inner fences range and outer fences range.

$Inner\ fences = (Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$

$$= \big((2870 - (1.5 \times 140)), (3010 - (1.5 \times 140))\big) = (2660, 3220)$$

By considering the inner fences range, we determine that two values, namely 3260 and 3525, fall outside this range, suggesting they might be outliers (suspected outliers).

Math: Construct a Box-and- Whisker Plot for these data and identify if any outliers:

$$3, 9, 10, 2, 6, 7, 5, 8, 6, 6, 4, 9, 22$$

Math: Construct box plot

53, 55, 57, 58, 58, 59, 59, 60, 61, 62, 65, 65, 66, 68, 68, 68, 69, 73, 76, 77, 78, 80, 82, 90

Math: Construct box plot

68, 70, 74, 76, 78, 79, 79, 80, 82, 83, 86, 87, 87, 87, 88, 88, 90, 90, 93, 96, 96, 97, 98, 98