**Variable**: Is a parameter which is changing with outcome for and observed phenomena.

**Data :** The particular value of the variable is data-value.

**Population:** The collection of all data is called population.

**Sample:** A subset of population is called sample.

**e.g. 1;** Suppose a dice is thrown 45 times and outcomes are

6,1,6,2,1,6,6,4,1,5,5,5,1,2,6,6,4,2,3,2,4,1,4,4,6,4,3,6,4,3,2,6,4,3,4,3, 3,5,4,6,3,3,3,6,5. ...**(A)**

It can be arranged as **(A1)**

| $x = number$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| frequency | 5 | 5 | 9 | 10 | 5 | 11 |
| Cumulative freq. (≤) | 5 | 10 | 19 | 29 | 34 | 45 |
| Cumulative freq. (≥) | 45 | 40 | 35 | 26 | 16 | 11 |

(A) Scattered data , & (A1) grouped data yielding the table called frequency distribution

**e.g, 2;** Consider the temperature of a city we have

71.1, 71.4, 71.9, 72.8, 75.9, 76.6, 76.9, 78.6, 80.7, 81.6, 81.8, 83.0, 84.0, 86.2, 87.8, 87.9, 88.8, 88.9, 89.4, 91.9, 92.3, 94.1, 94.4, 94.4, 94.6, 94.7, 95.0, 96.0, 96.8, 99.2, 101.0, 101.7, 103.0, 106.0, 107.5 ...**(B)**

It can be represented in form of class say

70-75, 75-80, 80-85, 85-90, 90-95, 95-100, 100-105, 105-110.

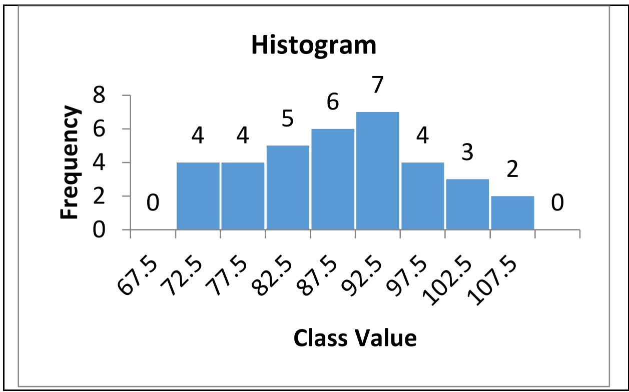
The mid value of an interval is taken as class value.

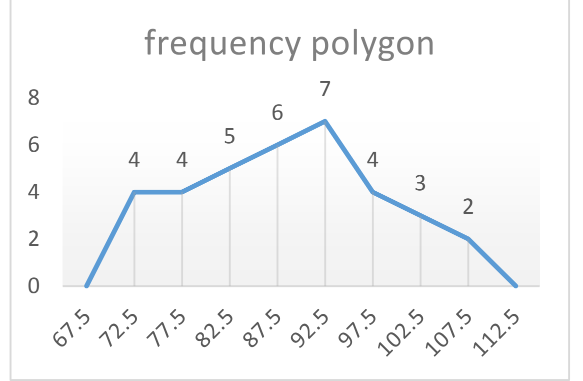
Any data falling on class boundaries to be assigned to higher class

Data can be represented as

| Class bound | 70-75 | 75-80 | 80-85 | 85-90 | 90-95 | 95-100 | 100-105 | 105-110 |
|---|---|---|---|---|---|---|---|---|
| Class value | 72.5 | 77.5 | 82.5 | 87.5 | 92.5 | 97.5 | 102.5 | 107.5 |
| frequency | 4 | 4 | 5 | 6 | 7 | 4 | 3 | 2 |
| Cumulative freq. | 4 | 8 | 13 | 19 | 26 | 30 | 33 | 35 |

Joining midpoint of the tips of the rectangle in the histogram. The polygon is closed on left and right by joining 67.5 112.5 and in this case sum of the area of rectangle equals to area bounded by frequency polygon and $x - axis$.





## Mean

The mean (Arithmetic mean) of sample values $x_1, x_2, \dots x_n$ is given as

$$sample\ mean\ \ (\bar{x}) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or

$$= \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Mean is called average value.

**e.g. 3;** the mean of e.g. 1 is $\frac{5*1+5*2+9*3+10*4+5*5+11*6}{45} = \frac{173}{45} = 3.84$

## Mid range

The mid range of a sorted sample $x_1, x_2, \dots x_n$ is the average of the smallest and the largest value ,

$$mid = \frac{x_1 + x_n}{2}$$

In case (A) $mid = \frac{1+6}{2} = 3.5$ & in case (B) $mid = \frac{72.5+107.5}{2} = 90.0$

## Median $(\tilde{x})$

If the sample data $x_1, x_2, \dots x_n$ is sorted in increasing order the median of the data is given as

$$\tilde{x} = \begin{cases} x_{k+1} & if\ n = 2k+1 \\ \frac{x_k + x_{k+1}}{2} & if\ \ n = 2k \end{cases}$$

**e.g. 4**; for 3,3,5,7,8 $\tilde{x} = 5$ & for 1,2,5,5,7,8,8,9 $\tilde{x} = \frac{5+7}{2} = 6$

In case of (A1) $\tilde{x} = 3$ (23rd term), and in case of (B) since data is grouped into the class, we can find median in two ways.

Since the number of data is 35 hence median is 18th term. Using the cumulative frequency it is the second term of class 85-90, thus

i. Simply $\tilde{x} = 87.5$ that is the class value.

ii. Linearly interpolate in the class $\tilde{x} = 85 + \frac{5}{6} * 5 = 89.166$

Clearly (ii) will be good approximation to the median.

In case of grouped frequency distribution median is obtained by the formula

$$\tilde{x} = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

***N***: is population (or sample) size.

***l***: is lower limit of median class, i.e.; c.f. is just greater than N/2

***f*:** is frequency of the median class,

***h***: width of the median class,

***C***: c.f. of the pre-median class.

## Mode

It is the value which occurs with the greatest frequency (maybe more than one value). In case of grouped frequency it is given as

$$mode = l + \frac{h(f_m - f_1)}{2f_m - (f_1 + f_2)}$$

***l***: lower limit of modal class, i.e.; clas with max. frequency

$\boldsymbol{f_m}$**:** frequency of modal class

$\boldsymbol{f_1}$ & $\boldsymbol{f_2}$: Frequency of pre-modal class and post-model class

***h*:** width of modal class

In case of (B) $mode = 90 + \frac{5(7-6)}{2*7-(6+4)} = 91.25$

## Variance and standard deviation

Consider two samples

S1 : 7,9,9,10,10,11,14 & S2 : 7,7,8,10,11,13,14

For both the samples $\bar{x} = 10$ both has first and last term same, but in S1 values are closely clustered about the mean $\bar{x}$ then the values in S2.

Consider a sample of values $x_1, x_2, \dots x_n$ and suppose $\bar{x}$ is the mean of the sample. The difference $(x_i - \bar{x})$ is called the deviation of the data value $x_i$ from the mean $\bar{x}$. It is positive or negative accordingly as $x_i$ is greater or less than $\bar{x}$.

The variance $\sigma^2$ is given as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 \ , \ where \quad N = \sum_{i=1}^{n} f_i$$

$$= \frac{1}{N}\left\{\sum_{i=1}^{n} f_i(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i^2 - \frac{2\bar{x}}{N}\sum_{i=1}^{n} f_i x_i + \frac{\bar{x}^2}{N}\sum_{i=1}^{n} f_i$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i^2 - \bar{x}^2 \qquad or \qquad = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i^2 - \left(\frac{\sum_{i=1}^{n} f_i x_i}{N}\right)^2$$

And *standard deviation (S.D.)* is denoted by $\sigma$ (non-negative square root of variance).

## Moments

The r<sup>th</sup> moment of a variable $x$ about any point $x = a$ denoted by $\mu_r'$ is given as

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - a)^r$$

In case $a = \bar{x}$ then the moment is about mean, called central moment, denoted as $\mu_r$ is given as

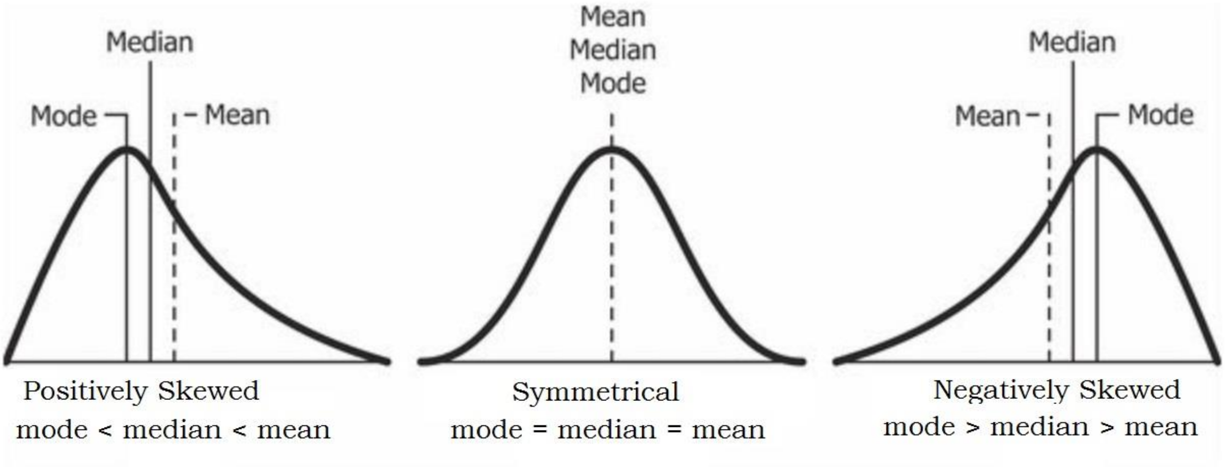
$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^r$$

In particular we have $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = \sigma^2$ & $\mu_0' = 1$, $\mu_1' = \bar{x} - a$

If $a = 0$, $\mu_1' = \bar{x}$, i.e.; first moment about the origin is the mean.

## Skewness and kurtosis

**Skewness** is the measure of lack of symmetry. A distribution is symmetric or 'normal' when frequencies are symmetrically distributed about mean. If the frequency curve of distribution has a long tail to right of central maximum the distribution is said to be skewed right or positively skewed, and if reverse is true then said to be skewed to left or negatively skewed.



The expression

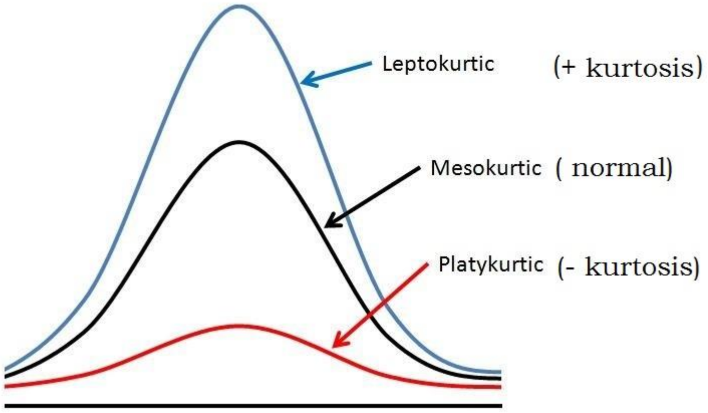
$$\frac{mean - mode}{S.D.}$$

is coefficient of skewness called Karl-Pearson's Coefficient of Skewness. It is a dimensionless number. For symmetric distribution, it is zero.
Thus skewness speaks about the manner in which the data is spread across the mean.

**Kurtosis** is the degree of peakedness or flatness of the frequency curve (usually taken relative to the normal distribution).

The data could be concentrated across the mean or dispersed equally across it which determines the peakedness of the curve.

If the data is more concentrated closer to the mean it is called as leptokurtosis. As the data gets more disperse, the peakedness reduces and the curve becomes Mesokurtic and if the data is much more dispersed, it is termed as platy kurtosis.



# Probability

For any experiment the collection of all possible outcomes (events) is called sample space ($S$) and is exhaustive.
If no preference is given to any outcome, i.e., all may occur with equal opportunity then events are called equally likely.
If occurrence of one event precludes the occurrence of any other event of the sample space then such events are known as mutually exclusive.

**Definition**: (Classical)
If a random trial may result in '$n$' exhaustive, mutually exclusive and equally likely outcomes, out of which '$m$' are favouring to occurrence of any event $E$ then the probability '$p$' of that event $E$ is given as

$$p = P(E) = \frac{no. of\ favourable\ cases}{total\ no. of\ cases} = \frac{m}{n}$$

From the definition it is clear that $0 \le P(E) \le 1$.

$P(E) = 0$ is impossible event.

$P(E) = 1$ is sure event.

Also the probability ($q$) of not occurrence of E , i.e., $\bar{E}\ or\ E^c$ will be,

$$q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(E) = 1 - p$$

Also we have, $p + q = 1$ ie., $P(E) + P(\bar{E}) = 1$

**Axiomatic approach**

The concept of probability can be represented by set algebra. $S$ (sample space) is the universal set. Outcomes are elements of $S$ and any subset of $S$ (proper or improper) is an event $E$ then

$$P(E) = \frac{n(E)}{n(S)} = \frac{\text{no. of element in set } E}{\text{no. of element in set } S}$$

**Definition**

Let $S$ be a sample space and an event $E \in S$ , the probability of event $E$ is $P(E)$ when the following axiom holds;

Axm.1. $0 \le P(E) \le 1 \quad for\ each \quad E \in S$

Axm.2. $P(S) = 1$

Axm.3. For any two disjoint events $E_1\ \&\ E_2$ we have

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Using the Axm.3 we can extended it to any number of disjoint events

$P(E_1 \cup E_2 \cup E_2 \cup \ldots \cup\ E_n) = P(E_1) + P(\ E_2) + \cdots + P(E_n)$

**Theorem**

i. $P(\phi) = 0$

ii. $P(\bar{E}) = 1 - P(E)$

iii. $P(E_1) \leq P(E_2)\ ,\ if\ \ E_1 \subseteq E_2$

iv. $P(E_2 \backslash E_1) = P(E_2) - P(E_1\ \cap\ E_2)$

**Proof**:

i. $S \cup \phi = S \Longrightarrow P(S \cup \phi) = P(S)$

$P(S) + P(\phi) = P(S) \Longrightarrow P(\phi) = 0 \qquad \because\ S\ \&\ \phi\ \ are\ disjoint$

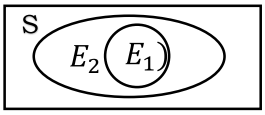
ii. $S$ be sample space then $E \cup \bar{E} = S \Longrightarrow P(E) + P(\bar{E}) = P(S) = 1$

$P(\bar{E}) = 1 - P(E)$

iii. $E_2 = E_1 \cup (E_2 \backslash E_1)$

$or\ \ P(E_2) = P(E_1) + P(E_2 \backslash E_1) \qquad \because\ \ E_1\ \&\ E_1\ are\ disjoint.$

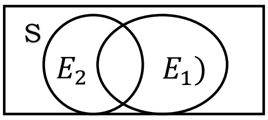
$\Longrightarrow P(E_2) \geq P(E_1) \qquad \because P(E_2 \backslash E_1) \geq 0$



iv. $E_2 = (E_2 \backslash E_1) \cup (E_1 \cap E_2)$

$P(E_2) = P(E_2 \backslash E_1) \cup P(E_1 \cap E_2)$

$P(E_2 \backslash E_1) = P(E_2) - P(E_1 \cap E_2)$

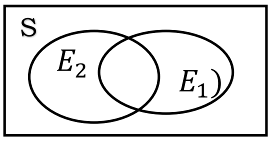


The sample space considered above is finite sample space that is number of outcome of a trial is finite for number of element in set S is finite. If all outcomes possess equal probabilities then it is called equiprobable space.

**Addition law of probability**

If $E_1$ & $E_2$ are any two events of the sample space $S$ , then

$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$



From the figure we have

$E_1 \cup E_2 = (E_1 \backslash E_2) \cup E_2$ , where $E_1 \backslash E_2$ & $E_2$ are disjoint sets.

$P(E_1 \cup E_2) = P(E_1 \backslash E_2) + P(E_2) = P(E_1) - P(E_1 \cap E_2) + P(E_2)$

$\qquad = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

For the events $E_1 , E_2 , E_3$

$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_2)$

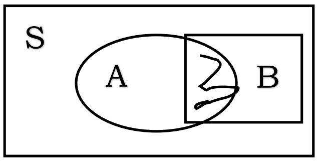
And similarly can be extended to any number of events.

**Conditional probability**

The probability for the event $A$ to occur while it is known that the event $B$ has already occurred is called conditional Probability and defined as following

$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \ldots(1)$

$P(A \cap B) = \dfrac{n(A \cap B)}{n(S)}\ , \qquad P(B) = \dfrac{n(B)}{n(S)} \quad \Longrightarrow P(A|B) = \dfrac{n(A \cap B)}{n(B)}$



**Multiplication law of probability**

From equation (1) we have

$P(A \cap B) = P(B)P(A|B) \quad or \quad P(A)P(B|A)$

In case of two Independents events $A$ & $B$ with $P(A) \neq 0$ & $P(B) \neq 0$ it becomes

$$P(A \cap B) = P(A)P(B) \qquad ...(2)$$

(as occurrence of one does not influence occurrence of other)

*The relation given in equation (2) is the condition of being two events independent.*

Its extension for three events $A, B$ *and* $C$ will be,

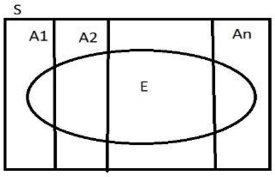
i. $P(A \cap B \cap C) = P(A)P(B)P(C)$
ii. $A, B, \& C$ are pair wise independent

Both the condition holds simultaneously.

**Partition , Total probability, and Bay's formula**

Let set $S$ is the union of mutually disjoint subsets $A_1, A_2, \ldots, A_n$ then $A_1, A_2, \ldots, A_n$ forms the partition of the set $S$. Let $E$ be any event of $S$ then

$E = E \cap S = E \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$
$\quad = (E \cap A_1) \cup (E \cap A_2) \cup\ldots\cup (E \cap A_n)$



Moreover the $'n'$ subsets in the right side are mutually disjoint and form partition of $E$ .

$P(E) = P(E \cap S) = P(E \cap A_1) + P(E \cap A_2) + \ldots + P(E \cap A_n)$ (∵ mutually disjoint sets)

$$= P(A_1)P(E|A_1) + P(A_2)P(E|A_2) + +P(A_n)P(E|A_n) \qquad ...(3)$$

Relation (3) is called total probability.

**Bay's formula**

Let $E$ be an event in the sample space $S$ and let $A_1, A_2, \ldots, A_n$ are disjoint event whose union is $S$ , i.e.;

$S = \cup_{i=1}^{n} A_i$ , then for $i = 1,2, \ldots n$

$$P(A_i|E) = \frac{P(A_i \cap E)}{P(E)} = \frac{P(A_i \cap E)}{P(A_1)P(E|A_1) + P(A_2)P(E|A_2) + +P(A_n)P(E|A_n)}$$

**Independent events**

If the occurrence of an event does not get influenced by occurrence of other events is called independent.

If the occurrence of $A$ does not affect the occurrence of $B$ , then $A$ & $B$ are called independent events, i.e.; $P(A|B)$ *or* $P(B|A)$ is same as $P(A)$ *or* $P(B)$ ,

Hence, $P(A \cap B) = P(B)P(A|B)$ becomes $P(A \cap B) = P(A)P(B)$ .

*(Mutually exclusive are different than independent)*