Chapter 16

Statistics and Probability

STATISTICS

Statistics is a branch of mathematics which deals with collection, grouping, analysis of the data and then to derive the results which are representative of the data collected.

Single value of the variable describing the data is called an average. Average lies generally in the middle of data so such values are called measures of central tendency. Common measures are mean, median and mode.

Mean, Median and Mode

The degree to which the values in a set of data are spread around an average is called dispersion of variation. The different measures of dispersion are range, quartile deviation, mean deviation, standard deviation and variance.

Statistics finds wide application in finance, economics, various surveys, planning, demography and various other analysis. All future planning of a country depends upon statistics of growth or fall in population, industrial production, crop production etc.

Mean

Also called arithmetic mean or simply average.

Unclassified data : The mean of n observations x_1, x_2 .... x_n is defined as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots\dots\dots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Grouped data : Let x_1, x_2 .... x_n be n observation with corresponding frequencies as f_1, f_2, ..., f_n respectively, then

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots\dots\dots + f_nx_n}{f_1 + f_2 + \dots\dots\dots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Shortcut method : Let the assumed mean (usually the middle term of data) be A. The deviation from assumed mean, $d_i = (x_i - A)$ for each term

$$\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i}$$

Step deviation method : When the data is grouped in equal class intervals with class size as h, then to simplify the calculation each $d_i = (x_i - A)$ is divided by h.

$$\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i} \times h \qquad \text{here } A = \text{assumed mean and } d_i = \frac{x_i - A}{h}$$

Weighted mean : If $w_1, w_2, \ldots\ldots, w_n$ are the weights assigned to the values $x_1, x_2, \ldots\ldots, x_n$ respectively then the weighted mean is given by

$$\text{Weighted mean} = \frac{w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n}{w_1 + w_2 + \ldots\ldots + w_n}$$

Combined mean : Let individual means of two sets of data are $\bar{x}_1$ and $\bar{x}_2$ with the size of data as n_1 and n_2 respectively, then

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Properties of Mean

1. If each of the n given observations are increased (decreased) by same quantity a then the new mean of data is also increased (decreased) by a.
2. If each of the n given observation are multiplied (divided) by same quantity $a(a \neq 0)$, the mean of data is also multiplied (divided) by a.
3. The algebraic sum of the deviations of individual values from mean is always zero.

$$\sum_{i=1}^{n} f_i(x_i - \bar{x}) = 0$$

4. The sum of squares of deviation from mean is least.

$$\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 \text{ is least}$$

Median

It is the middle most or central value of the variate in a set of observations, when the observations are arranged either in ascending or in descending order of their magnitudes. Median divides this arranged series in two equal parts.

Median of individual series : If there are total n observations arranged in ascending or descending order, then

(i) If n is odd, median = value of $\left(\frac{n+1}{2}\right)^{th}$ observation.

(ii) If n is even, median = mean of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.

Median of a discrete series : Prepare a cumulative frequency after arranging the data in ascending or descending order.

(i) If n is odd, then median = $\left(\frac{n+1}{2}\right)^{th}$ term of the size arranged either ascending or descending order.

(ii) If n is even, then median = $\dfrac{\left(\frac{n}{2}\right)^{th} \text{ term of the size} + \left(\frac{n}{2}+1\right)^{th} \text{ term of the size}}{2}$.

Median of a continuous series :

Step I Prepare cumulative frequency table

Step II Choose the median class where $\frac{n}{2}^{\text{th}}$ observation lies

Step III Median = $l + \dfrac{\left(\frac{n}{2} - c_f\right)}{f} \times h$

Where l = lower limit of median class

$n = \Sigma f_i$

f = frequency of median class

c_f = cumulative frequency of class preceding to the median class

Mode

Mode = 3 median – 2 mean

It is the value of variate occurring most frequently or the variate having maximum frequency.

Mode of individual series is the value which is repeated maximum number of times.

Mode of discrete series is the value corresponding to maximum frequency.

Mode of continuous series : Modal class is the class having maximum frequency.

$$\text{Mode} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

l = lower limit of modal class

h = width of modal class

f_{m-1} = frequency of class preceding modal class

f_m = frequency of modal class

f_{m+1} = frequency of class succeeding modal class

If mode falls in other class than modal class then, mode = $l + \dfrac{f_{m+1}}{f_{m-1} + f_{m+1}} \times h$

When mean, median and mode coincide, such distribution is called symmetrical distribution. Otherwise

Mode = 3 median – 2 mean, for moderately skew distribution

Mean-deviation : For a frequency distribution, the mean deviation from an average (mean or median) is

$$\text{M.D.(about mean)} = \frac{\sum_{i=1}^{n} f_i \,|\, x_i - \bar{x} \,|}{\sum_{i=1}^{n} f_i} \quad \text{where } \bar{x} \text{ is mean}$$

$$\text{M.D. (about median)} = \frac{\sum_{i=1}^{n} f_i \,|\, x_i - M \,|}{\sum_{i=1}^{n} f_i} \quad \text{where } M \text{ is median}$$

$$\text{Coefficient of mean deviation} = \frac{\text{Mean deviation}}{\text{Corresponding average}}$$

STANDARD DEVIATION (Root mean square deviation)

(a) For ungrouped data

(i) **For Individual series:**

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

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

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

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 + (\bar{x})^2 - 2\bar{x}.\bar{x}$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2$$

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2}}$$

(ii) **For discrete series:**

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

$$= \sqrt{\frac{\sum_{i=1}^{n}f_i x_i^2}{\sum_{i=1}^{n}f_i} - \left(\frac{\sum_{i=1}^{n}f_i x_i}{\sum_{i=1}^{n}f_i}\right)^2}$$

(b) For grouped data

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}f_i(y_i - \bar{y})^2}{\sum_{i=1}^{n}f_i}}$$

where y_i = class marks

$\bar{y}$ = mean

Variance and coefficient of variation

Square fo standard deviation, *i.e.*, σ^2 is called variance, the term coefficient of variation (C.V.) is commonly used for measure of relative variation. A series having lesser C.V. is said to the less varaible as more consistent.

$$\boxed{\text{Coeff. of variance} = \frac{\sigma}{\text{mean}} \times 100\%}$$

Note 1 : *For a symmetrical distribution*

$\bar{x} \pm \sigma$ *covers 68.27% observations*

$\bar{x} \pm 2\sigma$ *covers 95.45% observations*

$\bar{x} \pm 3\sigma$ *covers 99.73% observations*

Note 2 : *Mean deviation* = $\frac{4}{5}\sigma$, *if data is moderately non-symmetrical*

Note 3 :
- *Median can graphically found from ogive.*
- *Mode can be graphically found from histogram*
- *Mean can be graphically found by frequency curve.*

COMBINED STANDARD DEVIATION

Let two data sets having n_1 and n_2 observations have $\bar{x}_1$ and $\bar{x}_2$ as their means and σ_1 and σ_2 their S.D. Then combined S.D. σ_{12}

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where $d_1 = \bar{x}_1 - \bar{x}_{12};\ d_2 = (\bar{x}_2 - \bar{x}_{12})$

and $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

PROBABILITY

Probability or chance is a measure of uncertainty in any event occurring. Life is full of uncertainties. Probability measures the degree of uncertainty between 0 and 1.

Galileo was first to attempt probability while he was dealing with some problem relating to dice in gambling. The foundation of theory of probability was laid by B. Pascal and P. Fermat.

Now probability is widely used in business and economics for making future predictions.

In the game of cards (52 cards) there are four suits (13 each) namely spades (black), clubs (black), hearts (red) and diamonds (red). Aces, Kings, Queens and Jacks (knaves) are called honour cards and Kings, Queens and Jacks are called face cards while others are number cards.

BASIC CONCEPTS

Experiment or trial

An operation which results in some well defined outcomes.

Random Experiment

An experiment whose outcome cannot be predicted with certainty is called a random experiment.

Tossing of a coin is a random experiment.

Throwing a die is a random experiment.

Sample Space

The set of all possible outcomes of an experiment is called the sample space of that experiment.

It is denoted by S.

When a coin is tossed the sample space $S = \{H, T\}$

When a die is thrown, the sample space $S = \{1, 2, 3, 4, 5, 6\}$

EVENT

A subset of the sample space is called an event.

Types of events

Simple event or elementary event :

An event is called a simple event, if it is a singleton subset of the sample space S.

For example, getting of three heads in a simultaneous throw of 3 coins is a simple event.

Mixed event or compound event :

A subset of the sample space S which contains more than one element is called a mixed event.

For example, getting the same number on both die, when cast together is a compound event.

Impossible event :

Impossible event is represented by empty set ϕ.

For example, getting a number 7 when a die is thrown is an impossible event.

Sure event or certain event :

The event represented by sample space S itself is a sure event or certain event.

For example, the event of getting one red or black card, when a card is drawn from a well shuffled pack of cards is a sure event.

Equally likely events :

Events are said to be equally likely when one event does not happen more often than the other event.

For example, the event of drawing a black card and the event of drawing a red card are equally likely, when a card is drawn from a pack of 52 cards.

Mutually exclusive or disjoint events :

A set of events is said to be mutually exclusive, if the happening of one excludes the happening of the others.

Thus events $A_1, A_2, ..., A_n$ are mutually exclusive if and only if $A_i \cap A_j = \phi$ for all $i \neq j$.

For example, when husband and wife both appear for an interview for one post, the selection of husband and the selection of wife are mutually exclusive events.

Exhaustive events (cases) :

A set of events is said to be exhaustive if the performance of the experiment always results in the occurrence of at least one of them.

If $A_1, A_2, \ldots\ldots, A_n$ are exhaustive events then, $A_1 \cup A_2 \cup \ldots\ldots \cup A_n = S$

For example, the events {1, 2}, {3, 4, 5} and {6} are exhaustive event for the sample space {1, 2, 3, 4, 5, 6}.

ALGEBRA OF EVENTS

Let E and F be two events. Then

(i) E' or $\overline{E}$ or E^c stands for the non-occurrence or negation of E.

(ii) $E \cup F$ stands for the occurrence of at least one of E and F. (Also denoted by $E + F$, E or F)

(iii) $E \cap F$ stands for the simultaneous occurrence of E and F. (Also denoted by EF or E and F)

(iv) $E' \cap F'$ stands for the non-occurrence of both E and F

(v) $E \subseteq F$ stands for the occurrence of E implies occurrence of F.

If $E \cap F = \phi \Rightarrow E$ and F are mutually exclusive events.

(vi) $E–F$ denotes the occurrence of event E but not F

$E–F = E \cap F'$

(vii) $E \cup F = F \cup E$ and $E \cap F = F \cap E$

(viii) $(E \cup F)' = E' \cap F'$ and $(E \cap F)' = E' \cup F'$

PROBABILITY

Probability of occurrence of an event is a number lying between 0 and 1 (inclusive of 0 and 1), i.e. $0 \le P(E) \le 1$

Let S be the sample space, then the probability of occurrence of an event E is denoted by $P(E)$ and is defined as

$$P(E) = \frac{n(E)}{n(S)} = \frac{\text{number of elements in } E}{\text{number of elements in } S}$$

$$= \frac{\text{number of cases favourable to event } E}{\text{total number of cases}}$$

If $P(E) = 0 \quad \Rightarrow \quad$ Event is impossible

$P(E) = 1 \quad \Rightarrow \quad$ Event is sure

$P(E) \ngtr 1$ and $P(E) \nless 0$

Remarks :

1. More is the probability of an event, more are chances of its happening.
2. $P(\phi) = 0$ & $P(S) = 1$

i.e. nothing outside the sample space can occur.

ODDS IN FAVOUR OR AGAINST

If an event E can happen in a ways and fails to happen in b ways then;

(i) $P(E) = \dfrac{a}{a+b}$

(ii) $P(E^c) = \dfrac{b}{a+b}$

(iii) The odds in favour of event $E = \frac{a}{b}$

(iv) The odds against event $E = \frac{b}{a}$

If P is the probability of an event E, then

(i) Probability of E not happening = $1 - P$

(ii) Odds in favour of $E = \frac{P}{(1-P)}$

(iii) Odds against $E = \frac{(1-P)}{P}$

ADDITION THEOREM

1. If $p_1, p_2, \ldots\ldots\ldots\ldots p_n$ be the probabilities of n mutually exclusive events $E_1, E_2, \ldots\ldots\ldots E_n$, then the probability p, that any one of these events will happen is given by

   $p = p_1 + p_2 + \ldots\ldots\ldots\ldots + p_n$

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

2. If E and F are any two events (not mutually exclusive), then

   $P(E \cup F) = P(E + F) = P(E) + P(F) - P(E \cap F) = P(E) + P(F) - P(EF)$

   In general, If $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ are n events, then

$$P(E_1 \cup E_2 \cup \ldots\ldots\ldots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{1 \le i < j \le n} P(E_i \cap E_j) + \sum_{1 \le i < j < k \le n} P(E_i \cap E_j \cap E_k) - \ldots\ldots$$

$$\ldots + (-1)^{n-1} P(E_1 \cap E_2 \cap \ldots\ldots\ldots\ldots \cap E_n)$$

IMPORTANT RESULTS

(i) If E and F are two events, then

P (exactly one of E, F occurs) $= P(E \cup F) - P(E \cap F) = P(E' \cup F') - P(E' \cap F')$

(ii) $P(E' \cup F') = 1 - P(E \cap F)$, $P(E' \cap F') = 1 - P(E \cup F)$

(iii) If $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ are n events then

$P(E_1 \cup E_2 \cup \ldots\ldots\ldots\ldots \cup E_n) \le P(E_1) + P(E_2) + \ldots\ldots\ldots\ldots + P(E_n)$

(iv) $P(E_1 \cap E_2 \cap \ldots\ldots\ldots\ldots \cap E_n) \ge 1 - P(E'_1) - P(E'_2) - \ldots\ldots\ldots\ldots - P(E'_n)$

(v) $P(E_1 \cap E_2 \cap \ldots\ldots\ldots\ldots \cap E_n) \ge P(E_1) + P(E_2) + \ldots\ldots\ldots\ldots + P(E_n) - (n-1)$

(vi) If $E \subseteq F \Rightarrow P(E) \le P(F)$

Conditional Probability :

The probability of occurrence of an event E, given that an event F has already occurred is called the conditional probability of occurrence of E. It is denoted by $P(E/F)$.

$$P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{n(E \cap F)}{n(F)} \qquad (P(F) \ne 0)$$

$$\Rightarrow P(E/F) = \frac{N_{E \cap F}/N}{N_F/N} = \frac{P(E \cap F)}{P(F)} \qquad (\text{N = total no. of outcomes})$$

Hence, $P(E \cap F) = \begin{cases} P(F).P(E/F) & (\text{if } P(F) \ne 0) \\ P(E).P(F/E) & (\text{if } P(E) \ne 0) \end{cases}$

Independent Events :

Two events E and F are independent if the occurrence or non-occurrence of E(or F) does not affect the probability of occurrence or non-occurrence of F(or E), that is

$P(F/E) = P(F)$ provided $P(E) \neq 0$.

Here $P(F/E)$ stands for probability of F if E has already happened. Thus, E and F are independent if and only if $P(E \cap F) = P(E) \cdot P(F)$

Multiplication Laws of Probability :

If $E_1, E_2, \ldots E_{n+1}$ be $(n+1)$ events such that $P(E_1 \cap E_2 \cap \ldots\ldots \cap E_n) > 0$ then.

$$P(E_1 \cap E_2 \cap \ldots\ldots \cap E_{n+1}) = P(E_1).\, P(E_2 / E_1).P(E_3 / E_1 \cap E_2) \ldots\ldots P(E_{n+1} / E_1 \cap E_2 \cap \ldots\ldots \cap E_n)$$

Special Cases :

(i) If $E_1, E_2 \ldots E_n$ are independent events. Then $P(E_1 \cap E_2 \ldots\ldots \cap E_n) = P(E_1)\, P(E_2) \ldots\ldots P(E_n)$

(ii) The probability that none of $E_1, \ldots E_n$ happen is $= (1 - P(E_1)) \cdot (1 - P(E_2)) \ldots\ldots (1 - P(E_n))$

(iii) If p is the probability that an independent event happens in one trial, then the probability that it will happen in a succession of k trials $= p^k$

Total Probability Rule :

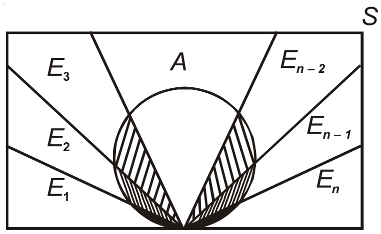
Let $\{E_i\}$, $i = 1, 2 \ldots n$ be n events such that $E_i \cap E_j = 0$ for $j \neq i$ and $\bigcup_{i=1}^{n} E_i = S$ (they are mutually exclusive and exhaustive). Suppose $P(E_i) > 0$ $(i = 1, 2 \ldots n)$. Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(E_i)\, P(A/E_i) = P(E_1)\, P(A/E_1) + \ldots + P(E_n)\, P(A/E_n).$$

Baye's Rule :

Let $\{E_i\}$ be mutually exclusive events such that $P(E_i) > 0$ for $i = 1, \ldots n$ and $S = \bigcup_{i=1}^{n} E_i$. Let A be any event with $P(A) > 0$. Then for $i = 1, 2, \ldots n$

$$P(E_i/A) = \frac{P(E_i).P(A/E_i)}{\sum_{i=1}^{n} P(E_i).P(A/E_i)}$$



Facts to Remember :

Let E and F are mutually exclusive and exhaustive events and G be any other event. Then

(i) $G = (G \cap E) \cup (G \cap F)$

(ii) $P(G) = P(E)\, P(G/E) + P(F)\, P(G/F)$

(iii) $P(E/G) = \dfrac{P(E) \cdot P(G/E)}{P(E) \cdot P(G/E) + P(F) \cdot P(G/F)}$

PROBABILITY DISTRIBUTION OF A RANDOM VARIATE

A random variable which can take only finite or countably infinite number of values is called discrete random variable otherwise continuous random variable.

If the values of a random variable together with the corresponding probabilities are given, it is called the probability distribution of random variable.

e.g., the probability distribution of number of heads when two coins are tossed together is

X	:	0	1	2
$P(X)$	:	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Binomial Distribution

(i) If n independent trials are performed and if p be the probability of success of each trial and q of its failures, then the probabilities of getting 0, 1, 2, 3,........n successes are given by the respective terms of the binomial expansion

$$(q+p)^n = q^n + {}^nC_1\, q^{n-1} p + {}^nC_2\, q^{n-2} p^2 + \ldots\ldots\ldots + p^n$$

$$= P(0) + P(1) + P(2) + \ldots\ldots\ldots + P(n)$$

The values of p and q remain the same for all trials in an experiment

(ii) The probability of exactly r successes in n trials is given by

$P(r) = {}^nC_r \,.\, q^{n-r} \,.\, p^r$

(iii) The probability of at least one success

$= P(1) + P(2) + \ldots\ldots + P(n) = 1 - P(0)$

The probability of at least two successes

$= 1 - [P(0) + P(1)]$ and so on

Note **:** *The mean, variance and standard deviation of a binomial distribution are np, npq and* $\sqrt{npq}$ *respectively.*

Use of Multinomial Theorem :

Let a die having m faces (with numbers 1, 2, 3.......m) be tossed n times. Then the probability that the sum of the numbers obtained on the die is p is given by the coefficient of x^p in the expansion of

$$\frac{(x + x^2 + x^3 + \ldots\ldots + x^m)^n}{m^n}$$

❑ ❑ ❑