RANDOM Variable:

Let S be a sample space connected with a random experiment, then mathematical function $X: S \to R$ is called Random Variable, provided Probability of $X \in S \implies P(X \in S) = 1$.

Eg: Tossing of 2 coins.

$S = \{(HH), (HT), (TH), (TT)\}$

X : No. of heads.

X	0	1	2
P(X)	1/4	2/4	1/4

$P(x) = P(x=0) + P(x=1) + P(x=2)$.

$= 1/4 + 2/4 + 1/4$

$= 1$.

Eg: Tossing of 3 coins.

$S = \{(HHH)(HHT)(HTH)(THH)(TTH)(THT)(HTT)(TTT)\}$.

X : No. of tails.

X	0	1	2	3
P(X)	1/8	3/8	3/8	1/8

$P(x) = 1$

Eg:- Rolling 2 dies

X : Sum on faces.

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$P(x) = 1$.

* There are 2 types of random variables. They are

1. Discrete Random Variable
2. Continuous Random Variable.

→ A Random variable X is said to be "Discrete" if it takes finite no. of values in the given interval.

→ If it takes Continuous values then it is said to be "Continuous Random Variable."

Probability Mass function:

Let X be a discrete random variable with distinct values $x_1, x_2, x_3 \dots x_n$. Then, the Probability mass function is denoted by $P(x)$ and defined by

(or)
Probability function: $P(x) = P(X = x_i) = \begin{cases} p_i & x = x_i \\ 0 & x \neq x_i \end{cases}$

Probability Distribution function:

Let X be a discrete Random variable then, Probability distribution function (cummulative function) is denoted by $F_x(x)$ and defined by

$$F_x(x) = P(X \leq x).$$

Eg: Tossing 3 coins

X : No. of heads.

	x_1	x_2	x_3	x_4
X	0	1	2	3
P(x)	1/8	3/8	3/8	1/8
F(x)	1/8	4/8	7/8	1

$F(0) = P(x \leq 0) = P(x=0)$

$F(1) = P(x \leq 1) = P(x=0) + P(x=1)$

$F(2) = P(x \leq 2) = P(x \leq 1) + P(x=2)$

$F(3) = P(x \leq 3) = P(x \leq 2) + P(x=3)$

NOTE:

If $P(x)$ is probability mass function, then it satisfies the following conditions.

1. $P(x_i) \geq 0 \quad \forall i$
2. $\sum_{i=1}^{n} P(x_i) = 1 = F(x)$

(*) A random variable X has the following probability functions.

x	0	1	2	3	4	5	6	7
$P(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

Then.

1. Find k.
2. Compute $P(x<6)$; $P(x \geq 6)$
3. $P(0<x<5)$
4. Find minimum value of 'a' such that $P(X \leq a) > 1/2$
5. Find Probability distribution function.

Sol:-

$$\sum_{i=1}^{n} P(X = x_i) = 1; \quad P(x_i) \geq 0.$$

$$\Rightarrow \quad 9k + 10k^2 = 1 \quad \Rightarrow \quad 10k^2 + 9k - 1 = 0.$$

$$10k^2 + 10k - k - 1 = 0$$

$$10k(k+1) - (k+1) = 0$$

$$k = -1, \ 1/10.$$

$k \neq -1$

if $k = -1$ then $P(1) < 0$. but $P(x_i) \geq 0$

$\therefore k = 1/10.$

x	0	1	2	3	4	5	6	7
$P(x)$	0	0.1	0.2	0.2	0.3	0.01	0.02	0.17

$$P(X<6) = P(X=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4) + P(x=5)$$

$$= 0 + 0.1 + 0.2 + 0.2 + 0.3 + 0.01$$

$$= 0.81$$

$$P(X \geq 6) = P(x=6) + P(x=7) = 0.19$$

(27)

$$P(X \geq 6) = 1 - P(X < 6)$$
$$= 1 - 0.81 = 0.19$$

$$P(0 < X < 5) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$$
$$= 0.1 + 0.2 + 0.2 + 0.3$$
$$= 0.8$$

$P(X \leq a) > \frac{1}{2}$

$F(a) = P(X \leq a) > \frac{1}{2} \Rightarrow$ $F(0) = P(X=0) = 0 \not> 1/2$

$F(1) = P(X \leq 1) = 0.1 \not> 1/2$

$F(4) = P(X \leq 4) = 0.8 > 1/2$

$\therefore a = 4$

(*) A discrete random Variable satisfies the following Conditions. The X assumes the values $-3, -2, -1, 0, 1, 2, 3$. and

$$P(X > 0) = P(X = 0) = P(X < 0)$$
$$P(X = -3) = P(X = -2) = P(X = -1)$$
$$P(X = 1) = P(X = 2) = P(X = 3)$$

Then construct probability mass function and determine the distribution function.

Sol:-

X	-3	-2	-1	0	1	2	3
P(x)	1/9	1/9	1/9	1/3	1/9	1/9	1/9

$$P(X < 0) + P(X = 0) + P(X > 0) = 1$$
$$\Rightarrow 3 \cdot P(X = 0) = 1$$
$$P(X = 0) = 1/3.$$
$$\Rightarrow P(X < 0) = 1/3$$
$$\Rightarrow P(X=-3) + P(X=-2) + P(X=-1) = 1/3 \Rightarrow P(X=-1) = 1/9$$

(28) $P(x>0) = 1/3$

$P(x=1) + P(x=2) + P(x=3) = 1/3$

$P(x=1) = P(x=2) = P(x=3) = 1/9$.

x	-3	-2	-1	0	1	2	3
P(x)	1/9	1/9	1/9	1/3	1/9	1/9	1/9
F(x)	1/9	2/9	3/9	6/9	7/9	8/9	9/9 = 1

*

	a	b	c	d
x	0	1	2	3
P(x)	1/8	3/8	3/8	1/8
F(x)	1/8	4/8	7/8	1

$P(a \le x \le d) = 1$.

$P(a \le x \le d) = P(x=a) + [F(d) - F(a)]$

$P(a \le x \le d) = P(x=a) + P(x=b) + P(x=c) + P(x=d)$

$= 1$

$P(a \le x \le d) = P(x=a) + [F(d) - F(a)]$

$P(a < x \le d) = F(d) - F(a)$

$P(a < x < d) = F(d) - F(a) - P(x=d)$

Probability Density function:

Let X be a continuous random variable defined in the interval $[a,b]$ then, the probability density function is denoted by $f(x)$ and defined by

$f(x) = \frac{d}{dx} F(x)$ ↳ prob. distribution fn.

(or)

$$F(x) = \int_a^b f(x) \cdot dx$$

* If $F(x)$ is a prob. distribution function then it satisfies the following conditions.

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

* In general

$$P(a \leq x \leq b) = \int_{a}^{b} f(x)\,dx$$

(*) A continuous Random Variable X has probability density function $f(x) = 3x^2$; $0 \leq x \leq 1$. Find a, b such that

1) $P(x \leq a) = P(x > a)$

2) $P(x > b) = 0.05$

Sol:-

1) $P(x \leq a) = P(x > a)$

$$\int_{-\infty}^{a} f(x)\cdot dx = \int_{a}^{\infty} f(x)\cdot dx$$

$$\int_{0}^{a} f(x)\cdot dx = \int_{a}^{1} f(x)\cdot dx \implies \int_{0}^{a} 3x^2\cdot dx = \int_{a}^{1} 3x^2\cdot dx$$

$$\Rightarrow \left(x^3\right)_0^a = \left(x^3\right)_a^1$$

$$a^3 = 1 - a^3$$

$$a^3 = 1/2$$

$$\Rightarrow a = \left(\tfrac{1}{2}\right)^{1/3}$$

$$\therefore a = 0.7937$$

2) $P(x > b) = 0.05$

$$\int_{b}^{\infty} f(x)\cdot dx = 0.05$$

$$\Rightarrow \int_{b}^{1} f(x)\cdot dx = 0.05 \implies 1 - b^3 = \frac{1}{20}$$

$$b^3 = \frac{19}{20} \Rightarrow b = \left(\frac{19}{20}\right)^{1/3}$$

$$\therefore b = 0.991$$

(*) Let X be a continuous Random Variable with Probability density function $f(x) = \begin{cases} ax & 0 \le x \le 1 \\ a & 1 \le x \le 2 \\ 3a - ax & 2 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$ then determine the constant 'a',

~~(2 ≤ x ≤ 5)~~ $P(x \le 1.5)$

Sol:- we have $\int_{-\infty}^{\infty} f(x) \cdot dx = 1$

$$\Rightarrow \int_{-\infty}^{0} f(x) \cdot dx + \int_{0}^{1} f(x) \cdot dx + \int_{1}^{2} f(x) \cdot dx + \int_{2}^{3} f(x) \cdot dx + \int_{3}^{\infty} f(x) \cdot dx = 1$$

$$\Rightarrow 0 + \int_{0}^{1} ax \cdot dx + \int_{1}^{2} a \cdot dx + \int_{2}^{3} (3a - ax)\, dx + 0 = 1$$

$$\left(a\frac{x^2}{2}\right)_0^1 + (ax)_1^2 + \left(3ax - \frac{ax^2}{2}\right)_2^3 = 1$$

$$\frac{a}{2} + a + \left(9a - \frac{9a}{2}\right) - \left(6a - 2a\right) = 1$$

$$\frac{3a}{2} + \frac{9a}{2} - 4a = 1$$

$$2a = 1 \Rightarrow a = \frac{1}{2}$$

$$P(x \le 1.5) = \int_{-\infty}^{1.5} f(x) \cdot dx$$

$$= \int_{-\infty}^{0} f(x) \cdot dx + \int_{0}^{1} f(x) \cdot dx + \int_{1}^{1.5} f(x) \cdot dx$$

$$= 0 + \int_{0}^{1} ax \cdot dx + \int_{1}^{1.5} a \cdot dx$$

$$= a\left(\frac{x^2}{2}\right)_0^1 + a(x)_1^{1.5}$$

$$= a\left(\frac{1}{2}\right) + a(1.5 - 1) = \frac{a}{2} + \frac{a}{2}$$

$$= a$$

$$= 1/2$$

(21)

(*) The amount of bread X (in hundreds of pounds) that a certain bakery is able to sell in a day is found to be a numerical value random phenomena with the probability function specified by the Probability density function is given by

$$f(x) = \begin{cases} kx & ; \ 0 \le x \le 5 \\ k(10-x) & ; \ 5 \le x \le 10 \\ 0 & ; \text{otherwise} \end{cases}$$

1) find the value of k such that $f(x)$ is probability density function.

2) what is the probability that no. of pounds of bread that will be sold next day is

a) more than 500 pounds.

b) less than 500 pounds

c) b/w 250 and 750 pounds

Sol:-

1) $\int_{-\infty}^{\infty} f(x) \cdot dx = 1$

$$\int_{-\infty}^{0} f(x) \cdot dx + \int_{0}^{5} f(x) \cdot dx + \int_{5}^{10} f(x) \cdot dx + \int_{10}^{\infty} f(x) \cdot dx = 1$$

$$0 + \int_{0}^{5} kx \cdot dx + \int_{5}^{10} k(10-x)\, dx = 1$$

$$\left(\frac{kx^2}{2}\right)_0^5 + k\left(10x - \frac{x^2}{2}\right)_5^{10} = 1$$

$$k\left(\frac{25}{2}\right) + k\left(100 - \frac{100}{2} - 50 + \frac{25}{2}\right) = 1$$

$$\frac{25k}{2} + k\left(\frac{25}{2}\right) = 1$$

$$k = \frac{1}{25}$$

1 = 100 pounds

2a) $P(X \geq 500)$

$= P(X \geq 5) = \int_5^{\infty} f(x) \cdot dx$

$= \int_5^{10} f(x) \cdot dx = \int_5^{10} k(10 - x) \cdot dx$

$= k\left(10x - \frac{x^2}{2}\right)_5^{10}$

$= \frac{1}{25}\left(100 - 50 - 50 + \frac{25}{2}\right)$

$= \frac{1}{2}$.

2b) $P(X \leq 500)$

$= P(X \leq 5)$

$= \int_0^5 kx \cdot dx = \left(\frac{kx^2}{2}\right)_0^5 = \frac{1}{25}\left(\frac{25}{2}\right) = \frac{1}{2}$

2c) $P(250 \leq X \leq 750)$

$= P(2.5 \leq X \leq 7.5) = \int_{2.5}^{7.5} f(x) \cdot dx$

$= \int_{2.5}^{5} f(x) \cdot dx + \int_5^{7.5} f(x) \cdot dx$

$= \left(\frac{kx^2}{2}\right)_{2.5}^{5} + k\left(10x - \frac{x^2}{2}\right)_5^{7.5}$

$= \frac{1}{25}\left[\frac{25}{2} - \frac{6.25}{2} + 75 - \frac{(7.5)(7.5)}{2} - 50 + \frac{25}{2}\right]$

$= \frac{1}{25}\left[50 - \frac{6.25 + (2.5)(2.5) \cdot 3.3}{2}\right]$

$= \frac{1}{25}\left[50 - \frac{6.25(10)}{2}\right]$

$= \frac{1}{25}\left[\frac{100 - 62.5}{2}\right] = \frac{37.5}{50}$

$= \frac{3.75}{5}$

$= 0.75$

Mathematical Expectation of Random Variable:

* Let X be a random variable then the mathematical Expectation of X denoted by $E(x)$ (or) μ (mean) and defined by

$$E(x) = \sum_{i=1}^{n} x_i P_i \quad \text{if } X \text{ is discrete Random variable.}$$

$$E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx \quad \text{if } X \text{ is Continuous Random Variable.}$$

Eg:

X : No. of heads if 3 coins are tossed.

X	0	1	2	3
P(x)	1/8	3/8	3/8	1/8

Expectation of X

$$E(x) = \mu = \sum_{i=1}^{n} x_i P_i$$

$$= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8}$$

$$= \frac{3}{8} + \frac{6}{8} + \frac{3}{8}$$

$$E(x) = 3/2.$$

Properties:

1. If C is a Constant then

   a) $E(c) = c$

   b) $E(cx) = c \cdot E(x)$

   c) $E(ax \pm b) = a\,E(x) \pm b$

2. If x, y are 2 random variables then

$$E(x + y) = E(x) + E(y)$$

(*) Find the mean of the probability distribution when 2 dice's are thrown and x is the random variable: Sum on the faces.

Sol:-

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$$E(x) = \mu = \sum_{i=1}^{n} x_i P_i$$

$$= \frac{2 + 6 + 12 + 20 + 30 + 42 + 40 + 36 + 30 + 22 + 12}{36}$$

$$= \frac{252}{36} = 7.$$

Variance: The average of sum of squares of deviations of the actual values of the observations taken from its mean value is known as variance of Random Varibale 'x'. It is denoted by $var(x)$ and defined by

$$var(x) = E\left((x-\mu)^2\right)$$

Properties:

1. If 'a' is a constant, then

$$var(a) = 0.$$

$$var(ax) = a^2 \cdot variance(x)$$

2. $var(x) = E\left[(x-\mu)^2\right] = E\left(x^2 + \mu^2 - 2\mu x\right)$

$$= E(x^2) + E(\mu^2) - 2E[x\mu]$$

$$= E(x^2) + \mu^2 - 2\mu E(x)$$

$$= E(x^2) + \mu^2 - 2\mu^2$$

$$= E(x^2) - \mu^2$$

$$\therefore \quad Var(x) = E(x^2) - \mu^2$$

$$= E(x^2) - [E(x)]^2$$

Standard deviation:

The square root value of variance of a random variable x is called Standard deviation.

$$SD = \sqrt{Var(x)}$$

$$SD = \sqrt{\sigma^2} \quad \Rightarrow \quad SD = \sigma$$

* $SD = \sigma$, $Var = \sigma^2$

(*) For the discrete probability distribution

x	-3	-2	-1	0	1	2	3
$P(x)$	k	0.1	k	0.2	$2k$	0.4	$2k$

find k, mean, variance.

Sol:-

$$\sum_{i=1}^{n} P(x_i) = 1 \Rightarrow 6k + 0.7 = 1$$

$$6k = 0.3 \Rightarrow k = 0.05$$

$$\text{mean} = \sum_{i=1}^{n} x_i P_i$$

$$= -3(0.05) - 2(0.1) - 1(0.05) + 0(0.2) + 1(0.1) + 2(0.4) + 3(0.1)$$

$$= -0.15 - 0.2 - 0.05 + 0 + 0.1 + 0.8 + 0.3$$

$$= -0.4 + 1.2$$

$$= 0.8$$

(36)

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$

$$E(x^2) = \sum_{i=1}^{n} x_i^2 P_i$$

$$= 9(0.05) + 4(0.1) + 0.05 + 0.1 + 4(0.4) + 9(0.1)$$

$$= 0.45 + 0.4 + 0.25 + 1.6 + 0.9$$

$$= 0.85 + 1.75 + 0.9$$

$$= 1.75 + 1.75 = 3.5$$

$$\therefore \text{Var}(x) = 3.5 - (0.8)^2$$

$$= 3.5 - 0.64 = 2.86$$

(*) Let X denotes the minimum of 2 no's that appear when a pair of fair dice is thrown once. Determine

1. Discrete probability distribution.
2. Expectation
3. variance
4. SD

Sol:-

X	1	2	3	4	5	6
P(x)	11/36	9/36	7/36	5/36	3/36	1/36

$$E(x^2) = \sum_{i=1} x_i^2 P(x_i) = \frac{11}{36} + 4\left(\frac{9}{36}\right) + 9\left(\frac{7}{36}\right) + 16\left(\frac{5}{36}\right) + 25\left(\frac{3}{36}\right) + 36\left(\frac{1}{36}\right)$$

$$= \frac{11 + 36 + 63 + 80 + 75 + 36}{36}$$

$$= \frac{301}{36} = 8.36$$

$$\text{Mean} = \sum_{i=1}^{n} P_i \cdot X_i$$

$$= \frac{11}{36} + \frac{18}{36} + \frac{21}{36} + \frac{20}{36} + \frac{15}{36} + \frac{6}{36} = \frac{91}{36}$$

$$= 2.52$$

(*) If a random variable x has probability density function $f(x) = \begin{cases} 2e^{-2x} & x>0 \\ 0 & x\le 0 \end{cases}$. Then, find

Probability that it takes

1. between 1 and 3
2. greater than 0.5

Sol:-

1. $P(1 \le x \le 3)$

$$= \int_1^3 f(x)\cdot dx = 2\int_1^3 e^{-2x}\cdot dx = 2\left(\frac{e^{-2x}}{-2}\right)_1^3$$

$$= -e^{-6} + e^{-2}$$

$$= e^{-2} - e^{-6}$$

2. $P(x \ge 0.5)$

$$= \int_{0.5}^{\infty} f(x)\cdot dx = 2\int_{0.5}^{\infty} e^{-2x}\cdot dx$$

$$= 2\left[\frac{e^{-2x}}{-2}\right]_{0.5}^{\infty} = -e^{-\infty} + e^{-2(0.5)}$$

$$= e^{-1} = \frac{1}{e}.$$

(*) The probability density function $f(x)$ of a continuous random variable is given by $f(x) = c\,e^{-|x|}$; $-\infty < x < \infty$

Then, show that

1. $c = 1/2$
2. Find mean, variance.
3. Also find $P(0 \le x \le 4)$

Sol:-

$$\int_{-\infty}^{\infty} c\,e^{-|x|}\cdot dx = 1$$

$$\Downarrow$$

Even fn. $\Rightarrow 2\int_0^{\infty} c\,e^{-|x|}\,dx = 1$

$$2c\int_0^{\infty} e^{-|x|}\cdot dx = 1$$

$$\Rightarrow \quad 2c \int_0^\infty e^{-x}\,dx = 1$$

$$2c \left[-e^{-x} \right]_0^\infty = 1$$

$$2c \left[-e^{-\infty} + e^{-0} \right] = 1$$

$$2c(1) = 1 \Rightarrow c = 1/2.$$

$$\text{Mean} = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$$

$$\text{Mean} = \int_{-\infty}^{\infty} \frac{1}{2} x e^{-|x|} \cdot dx \Rightarrow \text{Mean} = \frac{1}{2} \int_{-\infty}^{\infty} x e^{-|x|} \cdot dx$$

$\Downarrow$ odd function,

$\therefore$ Mean = 0.

$$E(x^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) \cdot dx$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} x^2 e^{-|x|} \cdot dx = \frac{1}{2} 2 \int_0^\infty x^2 e^{-x} \cdot dx$$

$$= \int_0^\infty x^2 e^{-x} \cdot dx$$

$$= \left(x^2 \frac{e^{-x}}{-1} - 2x e^{-x} + 2 \frac{e^{-x}}{-1} \right)_0^\infty$$

$$= 2.$$

$\therefore$ Variance = 2.

$$P(0 \le x \le 4) = \int_0^4 f(x) \cdot dx$$

$$= \int_0^4 \frac{1}{2} e^{-|x|}\,dx = \frac{1}{2} \int_0^4 e^{-x} \cdot dx$$

$$= \frac{1}{2} \left[\frac{e^{-x}}{-1} \right]_0^4$$

$$= \frac{1}{2} \left[1 - \frac{1}{e^4} \right]$$

(*) check whether, the following function

$$f(x) = \begin{cases} \frac{1}{16}(3+x)^2 & -3 \le x \le -1 \\ \frac{1}{16}(6-2x^2) & -1 \le x \le 1 \\ \frac{1}{16}(3-x)^2 & 1 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

is probability density fn. If so find the mean of Random variable.

Sol:-

$$\int_{-\infty}^{\infty} f(x)\cdot dx = \int_{-\infty}^{-3} f(x)\cdot dx + \int_{-3}^{-1} f(x)\cdot dx + \int_{-1}^{1} f(x)\cdot dx + \int_{1}^{3} f(x)\cdot dx + \int_{3}^{\infty} f(x)\cdot dx$$

$$= 0 + \int_{-3}^{-1} \frac{1}{16}(3+x)^2 dx + \int_{-1}^{1} \frac{1}{16}(6-2x^2)\,dx + \int_{1}^{3} \frac{1}{16}(3-x)^2 dx + 0$$

$$= \frac{1}{16}\left[\left[\frac{(3+x)^3}{3}\right]_{-3}^{-1} + \left[6x - \frac{2}{3}x^3\right]_{-1}^{1} + \left[\frac{-(3-x)^3}{3}\right]_{1}^{3} \right]$$

$$= \frac{1}{16}\left[\frac{8}{3} + \frac{16}{3} - \left(-\frac{16}{3}\right) + \frac{8}{3} \right]$$

$$= \frac{1}{16}\left[\frac{16}{3} + \frac{16}{3} + \frac{16}{3} \right] = 1.$$

$\Downarrow$

$f(x)$ is probability density function.

$$\text{Mean} = \int_{-\infty}^{\infty} x f(x)\cdot dx$$

$$= \int_{-3}^{-1} \frac{x}{16}(3+x)^2 dx + \int_{-1}^{1} \frac{x}{16}(6-2x^2)\,dx + \int_{1}^{3} \frac{x}{16}(3-x)^2 dx$$

$$= \frac{1}{16}\left[\left(x\cdot\frac{(3+x)^3}{3} - 1\cdot\frac{(3+x)^4}{12}\right)_{-3}^{-1} + \left(3x^2 - \frac{2}{4}x^4\right)_{-1}^{1} + \left(\frac{-x(3-x)^3}{3} - 1\cdot\frac{(3-x)^4}{12}\right)_{1}^{3} \right]$$

$$= \frac{1}{16}\left[-\frac{8}{3} - \frac{16}{12_3} + 0 - \left(\left(-\frac{8}{3}\right) - \frac{16}{12_3}\right) \right]$$

$$= \frac{1}{16}\left[-4 + 4 \right] = 0.$$

* Let, $f(x)$ be a probability density function. Then

$$\text{Mean} = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$$

→ The value of x, such that $f(x)$ is maximum, is called "Mode."

→ Median is the value of x such that

$$\int_{-\infty}^{M} f(x)\, dx = \int_{M}^{\infty} f(x)\, dx$$

i.e.

$$P(X \leq M) = P(x \geq M)$$

(*) For a probability density function $f(x) = \begin{cases} \frac{1}{2}\sin x & 0 \leq x \leq \pi \\ 0 & \text{otherwise} \end{cases}$

Correlation:

If we want to study 2 variables then the statistical analysis is called "bi-varient" analysis and there exists some relation b/w 2 variables is known as "correlation."

* correlation is a relationship b/w 2 variables. also correlation expresses the relation b/w 2 sets of variables upon each other.
* one variable is called subject (independent) and other variable is called relative (dependent).

Types of variables:

1. Positive and Negative
2. Simple and multiple
3. Partial and total
4. Linear and non-linear

1. positive & Negative correlation:

* positive & Negative correlation depend upon the direction of change of variables.
* If 2 variables tend to move together in the same direction i.e. increase in the value of 1 variable ⇒ increase in the value of other variable. Then the correlation is said to be +ve correlation.

* otherwise i.e. increase in one variable is dependent by decrease in other variable is called -ve correlation.

Eg:
Height & weight are +ve correlated.
Supply & demand are -ve correlated.

2. Simple & Multiple correlation:

when we study only 2 variables then the correlation b/w them is described as Simple correlation.

* If the relation b/w multiple variables is described as Multiple correlation.

3. Partial and total correlation:

The study of 2 variables by excluding some other variables is called partial correlation.

* In total correlation all facts are taken into account.

4. Linear and Non-linear correlation:

If the ratio of change b/w two variables is uniform then there will be a linear correlation. But in a non-linear correlation the ratio of change b/w 2 variables is not uniform.

NOTE:

In a linear correlation if we plot all the points on a graph then we get a straight line.

Method of studying correlation:

There are 2 different methods for finding the relationship b/w 2 variables. They are:

1. Graphical method
2. Mathematical method

Mathematical methods:

1. Karl Pearson coefficient of correlation
2. Spearmans rank coefficient of correlation
3. Method of least squares.

1. Coefficent of correlation:

correlation is a statistical technique used for analysing the behaviour of 2 (or) more variables.

* It analysis deals with the association between 2 (or) more variables.
* The correlation b/w 2 variables is relates to conversation b/w the series of variables but not the function.

Karl Pearson coefficient of correlation:

Measuring the magnitude of linear relationship b/w 2 variables is known as Pearson coefficient of correlation, it is denoted by r, defined by

$$r = \frac{\text{co-varience of } xy}{\sigma_x \cdot \sigma_y}$$

$$r = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2 \frac{1}{n}\Sigma(y_i - \bar{y})^2}}$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} \qquad X = x_i - \bar{x}, \quad Y = y_i - \bar{y}$$

NOTE:

1. Limits of correlation coefficient

$$-1 \le r \le 1$$

2. If $r = 1$ then perfect +ve correlation
   If $r = -1$ then perfect -ve correlation
   If $r = 0$ then there is no correlation b/w them.
3. If X and Y are random variables a, b, c, d are any numbers then

$$r(aX + b, cY + d) = \frac{bd}{|ac|} r(X, Y)$$

4. Two independent variables X and Y are uncorrelated ie.

$$r(X, Y) = 0$$

(*) calculate the coefficient of correlation from the following data.

Sol:

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	y^2	XY
12	14	2	5	4	25	10
9	8	-1	-1	1	1	1
8	6	-2	-3	4	9	6
($\bar{x}$) 10	9($\bar{y}$)	0	0	0	0	0
11	11	1	2	1	4	2
13	12	3	3	9	9	9
7	3	-3	6	9	36	18
		0	0	28	84	46

$\bar{x} = \frac{\Sigma x}{n}$ $\quad \bar{y} = \frac{\Sigma y}{n}$ $\quad r = \frac{\Sigma XY}{\sqrt{\Sigma x^2 \Sigma y^2}}$

$\bar{x} = 10$ $\quad \bar{y} = 9$

$r = \frac{46}{\sqrt{28 \cdot 84}} = 0.9485$

(*) calculate the Karl-Pearson coefficient of correlation blw age of cars and annual maintainence.

Sol:

Age of Cars (x)	Annual (y)	X	Y	X^2	Y^2	XY
2	1600	-5	-200	25	4×10^4	1000
4	1500	-3	-300	9	9×10^4	900
6	($\bar{y}$) 1800	-1	0	1	0	0
($\bar{x}$) 7	1900	0	100	0	10^4	0
8	1700	1	-100	1	10^4	-100
10	2100	3	300	9	9×10^4	900
12	2000	5	200	25	4×10^4	1000
		0	0	70	28×10^4	3700

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

$$r = \frac{3700}{\sqrt{70}\ \sqrt{28\times 10^4}}$$

$$r = \frac{37}{\sqrt{70\times 28}} \Rightarrow r = 0.8357$$

NOTE:

When deviations are taken from assumed mean (when actual mean is not a whole number but a fraction then the direct method will involve a lot of time to calculate correlation coefficient. In that case, we take any one of the entry as assumed mean) then,

$$\bar{r} = \frac{\Sigma XY - \frac{\Sigma X\ \Sigma Y}{n}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}}$$

Where,

$X = x - A$

$Y = y - B$

Where,

A, B are assumed means of x, y respectively.

(*) Find the coefficient of correlation for the following data

Sol:

	x	y	X	Y	X^2	Y^2	XY
	28	23	-7	-7	49	49	49
	41	34	6	4	36	16	24
	40	33	5	3	25	9	15
	38	34	3	4	9	16	12
(A)	35	30	0	0	0	0	0
	33	28	-2	-2	4	4	4
			5	2	123	94	104

$$\bar{x} = \frac{\Sigma x}{n} \qquad \bar{y} = \frac{\Sigma y}{n}$$

$$\bar{x} = 215/6 \qquad \bar{y} = 182/6$$

$$\bar{x} = 35.83 \qquad \bar{y} = 30.33$$

Hence, take A, B assumed means.

$$r = \frac{104 - \frac{5 \times 2}{6}}{\sqrt{123 - \frac{25}{6}}\sqrt{94 - \frac{4}{6}}}$$

$$r = \frac{102.33}{\sqrt{118.833}\sqrt{93.333}}$$

$$r = \frac{102.33}{(10.901)(9.6609)}$$

$$r = 0.9717.$$

Rank correlation coefficient:

This method is based on Rank and it is useful in dealing with qualitative characteristics such as Morality, character and intelligence etc.

* It can't be measured quantitatively as in the case of pearsons coefficient of correlation.
* The formula for spearman's rank correlation is given by

$$\rho = 1 - \frac{6 \Sigma D^2}{n(n^2-1)}$$

Where,

ρ - rank correlation coefficient

D^2 - Sum of squares differences of ranks

n - no. of observations

(*) obtain the rank correlation coefficient for the following data.

Sol:

x	y	r_1	r_2	$D = r_1 - r_2$	D^2	
75	85	2	2	0	0	
30	46	7	6	1	1	$n = 8$.
60	54	3	5	-2	4	
80	91	1	1	0	0	
53	58	4	4	0	0	
35	63	6	3	3	9	
15	35	8	8	0	0	
40	43	5	7	-2	4	
		36	36	0	18	

$r_1 \Rightarrow$ rank of x

$r_2 \Rightarrow$ rank of y

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(18)}{8(63)}$$

$$\rho = 1 - 0.214$$

$$\boxed{\rho = 0.7857}$$

(x) A random sample of 5 college students is selected and their grades in maths & statistics are found to be

	1	2	3	4	5
Maths	85	60	73	40	90
Statistics	93	75	65	50	80

Find the rank correlation coefficient.

Sol:

M	S	R_M	R_S	$D = R_M - R_S$	D^2
85	93	2	1	1	1
60	75	4	3	1	1
73	65	3	4	-1	1
40	50	5	5	0	0
90	80	1	2	-1	1
				0	4

$n = 5$.

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(4)}{5(24)}$$

$$\rho = 1 - 0.2$$

$$\boxed{\rho = 0.8}$$

Repeated ranks / Tied ranks:

When observations are repeated then the rank correlation coefficient b/w x and y is given by

$$\rho = 1 - \frac{6\left[\Sigma D^2 + T_x + T_y\right]}{n(n^2-1)}$$

where,

T_x and T_y are sum of correction factors in the series respectively.

$$T_x = T_y = \Sigma \frac{m^3 - m}{12}$$

where,

m - no. of items rank is repeated.

(*) Find the rank correlation coefficient.

x	y	r_1	r_2	$D = r_1 - r_2$	D^2
68	62	4	4	0	0
64	58	5.5	6	-0.5	0.25
75	68	2.5	2.5	0	0
50	45	7	7	0	0
64	81	5.5	1	4.5	20.25
80	60	1	5	-4	16
75	68	2.5	2.5	0	0
				0	36.5

T_x:

75 repeated 2 times

$m = 2$

$T_x = \frac{2^3 - 2}{12} = 0.5$

64 repeated 2 times

$m = 2$

$T_x = 0.5$

T_y:

68 repeated 2 times

$m = 2$

$T_y = 0.5$

$$\rho = 1 - \frac{6[36.5 + (0.5+0.5) + 0.5]}{7(48)}$$

$$\rho = 1 - \frac{6[38]}{7(48)}$$

$$\rho = \frac{28-19}{28} \Rightarrow \rho = 9/28$$

$$\rho = 0.32142$$

(*) Find the rank correlation coefficient for the following data.

Sol:

x	y	r_1	r_2	$D = r_1 - r_2$	D^2
65	68	7	4	3	9
63	66	9	7.5	1.5	2.25
67	68	4.5	4	0.5	0.25
64	65	8	9.5	-1.5	2.25
68	69	2.5	2	0.5	0.25
62	66	10	7.5	2.5	6.25
70	68	1	4	-3	9
66	65	6	9.5	-3.5	12.25
68	71	2.5	1	1.5	2.25
67	67	4.5	6	-1.5	2.25
				0	46

$D = 10$.

T_x:

68 repeated 2 times $\Rightarrow m=2 \Rightarrow T_x = 0.5$

67 repeated 2 times $\Rightarrow m=2 \Rightarrow T_x = 0.5$

T_y:

68 repeated 3 times $\Rightarrow m=3 \Rightarrow T_y = \frac{27-3}{12}$

$T_y = \frac{24}{12}$

$T_y = 2$

66 repeated 2 times $\Rightarrow m=2$

$\Downarrow$

$T_y = 0.5$

65 repeated 2 times $\Rightarrow m=2$

$\Downarrow$

$T_y = 0.5$

$$\rho = 1 - 6\left[\frac{\Sigma D^2 + T_x + T_y}{n(n^2-1)}\right]$$

$$\rho = 1 - 6\left[\frac{46 + (0.5+0.5) + (2+0.5+0.5)}{10(99)}\right]$$

$$\rho = 1 - 6\left[\frac{46+4}{990}\right]$$

$$\rho = 1 - 6\left[\frac{50}{990}\right]$$

$$\rho = 1 - \frac{30}{99} \Rightarrow \rho = 1 - \frac{10}{33}$$

$$\rho = 22/33$$

$$\boxed{\rho = 0.6969}$$

Regression Analysis:

Regression Analysis is a mathematical measure of average relationship between 2 (or) more variables interms of original units of the data.

* In Regression Analysis, there are 2 types of variables.

→ The variable whose value is influenced is called dependent variable and which influences is called independent variable.

* In regression analysis, there are 2 types of lines.

1. Regression line of y on x and it is defined as

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \Rightarrow \text{called as regression coefficient of } y \text{ on } x.$$

$$\Rightarrow y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

2. similarly, regression line of x on y and it is defined as

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where, $b_{xy} = r\left(\frac{\sigma_x}{\sigma_y}\right) \Rightarrow$ called as regression coefficient of x on y.

$$x - \bar{x} = r\left(\frac{\sigma_x}{\sigma_y}\right)(y - \bar{y})$$

$$\Downarrow$$

$$x - \bar{x} = r\left(\frac{\sigma_x}{\sigma_y}\right)(y - \bar{y})$$

NOTE:

To find the regression coefficient of y on x

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

and $$b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

where,

$$x = x - \bar{x} \quad \text{and} \quad y = y - \bar{y}$$

where,

$\bar{x}, \bar{y}$ are original mean's of x and y series respectively.

* If the mean's are Assumed mean's then

$$b_{yx} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$b_{xy} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$$

NOTE:

We have $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ then

$$b_{yx} \cdot b_{xy} = r \cdot \frac{\sigma_y}{\sigma_x} \times r \cdot \frac{\sigma_x}{\sigma_y}$$

$$r^2 = b_{xy} \cdot b_{yx}$$

$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

Alternate method for regression line:

The standard form of regression line of y on x is Equal to

$$y = ax + b.$$

Then by least square method, the normal Equations are:

$$\Sigma y = a\Sigma x + bn$$

$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

* similarly, the standard form of regression line of x on y is

$$x = cy + d$$

The Normal Equations are:

$$\Sigma x = c\Sigma y + dn$$

$$\Sigma xy = c\Sigma y^2 + d\Sigma y$$

(*) Find the regression line of y on x for the following data. Hence find the value of y when $x = 50$.

Sol:

x	y	$x - \bar{x} = X$	$y - \bar{y} = Y$	X^2	XY
0	54	-40	-21	1600	840
20	65	-20	-10	400	240
($\bar{x}$) 40	($\bar{y}$) 75	0	0	0	0
60	85	20	10	400	240
80	96	40	21	1600	840
		0	0	4000	2080

$$\bar{x} = \frac{\Sigma x}{n}$$

$$\bar{y} = \frac{\Sigma y}{n}$$

R.L of y on x: $y - \bar{y} = b_{yx}(x - \bar{x})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

$$b_{yx} = \frac{2080}{4000} = 0.52$$

R.L of y on x: $y - 75 = 0.52(x - 40)$

$$y - 75 = 0.52x - 20.8$$

$$\boxed{y = 0.52x + 54.2}$$

(or)

Standard form for R.L of y on x: $y = ax + b$

x	y	x^2	xy
0	54	0	0
20	65	400	1300
40	75	1600	3000
60	85	3600	5100
80	96	6400	7680
200	375	12000	17080

$200a + 5b = 375$

$12000a + 200b = 17080$

$a = 0.52$

$b = 54.2$

$\boxed{y = 0.52x + 54.2}$

(*) Find the most likely production corresponding to a rain fall 40 from the following data.

	(x) Rainfall	(y) Production
Avg	30	500 kgs
S.D	5	100 kgs
r	0.8	

⇓

corelation coefficient

(20)

y depends on x

R.L of y on x:

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Given,

$$\bar{x} = 30 \quad \bar{y} = 500$$

$$\sigma_x = 5 \quad \sigma_y = 100$$

$$r = 0.8$$

$$y - 500 = 0.8 \left(\frac{20}{5}\right)(x - 30)$$

$$y - 500 = 16x - 480$$

$$\boxed{y = 16x + 20}$$

$x = 40 \Rightarrow y = 16(40) + 20$

$y = 640 + 20$

$y = 660.$

$\mu = 49.97$.

Movements:

* Movements is defined as the Airthmetic mean of various powers of the deviations of items from their mean (Assumed or actual) will give the required Power of movement of the distribution.

* If the deviations of the items are taken from the arithmetic mean of the distribution is known as "Central movement."

* When the actual mean of a distribution is a fraction then it is difficult to calculate the movements. In this case, we select an item as assumed mean and we find the deviations then the movements are called "Raw movement."

1. Central Movements:

⇓ also called as "Movements about original mean."

a) Individual Series:

* Let $\bar{x}$ be the mean of the individual series.
Let d be the deviation of x from its mean $\bar{x}$
i.e., $d = x - \bar{x}$.

* Let N be the total no.of items/observations of the given series, then

1^{st} Central movement $= \mu_1 = \frac{\Sigma di}{N}$

2^{nd} Central movement $= \mu_2 = \frac{\Sigma di^2}{N}$

r^{th} Central movement $= \mu_r = \frac{\Sigma di^r}{N} = \frac{\Sigma (x-\bar{x})^r}{N}$

b) Frequency distribution:

* If 'n' observations $x_1, x_2, x_3 \ldots\ldots x_n$ and frequencies $f_1, f_2, f_3 \ldots\ldots f_n$ respectively then the mean

$$\bar{x} = \frac{\Sigma x_i f_i}{N}$$ where,

$$N = \Sigma f_i$$

Then,

$$\mu_1 = \frac{\Sigma f_i (x_i - \bar{x})}{N} = \frac{\Sigma f_i d_i}{N}$$

$$\mu_2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{N} = \frac{\Sigma f_i d_i^2}{N}$$

$$\mu_r = \frac{\Sigma f_i (x_i - \bar{x})^r}{N} = \frac{\Sigma f_i d_i^r}{N}$$

Properties of Central movements:

1. The first movement about mean is zero

$$\Rightarrow \mu_1 = 0.$$

2. The second central movement gives the variance of the distribution.

$$\Rightarrow \mu_2 = \sigma^2$$

$$\sigma = \pm\sqrt{\mu_2}$$

3. The third movement μ_3 is useful to measure the skewness of the given distribution, ~~if $\mu_3 = 0$~~

a) if $\mu_3 > 0$ $\Rightarrow$ distribution is +vely skewed

b) if $\mu_3 < 0$ $\Rightarrow$ distribution is -ve'ly skewed.

c) if $\mu_3 = 0$ $\Rightarrow$ distribution is Symmetrical

4. The fourth movement μ_4 is useful to measure kurtosis.

Skewness $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

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

Kurtosis.

$\gamma_2 = \beta_2 - 3$

2. Raw movements:

⇓

also called as "Movements about Assumed mean"

* for the distribution, if actual mean is a fraction then it is difficult to calculate a central movement. Then in this case, we select an item as assumed mean (A) then the raw movements

$$\mu_1' = \frac{\Sigma d_i}{N} \qquad \mu_1' = \frac{\Sigma f_i d_i}{N}$$

$$\mu_2' = \frac{\Sigma d_i^2}{N} \qquad \mu_2' = \frac{\Sigma f_i d_i^2}{N}$$

$$\mu_n' = \frac{\Sigma d_i^n}{N} \qquad \mu_n' = \frac{\Sigma f_i d_i^n}{N}$$

$$d_i = x_i - A.$$

(*) Find the 4 movements, skewness, kurtosis for the set of numbers 2, 4, 6, 8.

Sol:- Individual series

⇓

$$\bar{x} = \frac{2+4+6+8}{4} = \frac{20}{4} = 5.$$

x	$d_i = x - \bar{x}$	d_i^2	d_i^3	d_i^4
2	-3	9	-27	81
4	-1	1	-1	1
6	1	1	+1	1
8	3	9	+27	81
	$\Sigma d_i = 0$	$\Sigma d_i^2 = 20$	$\Sigma d_i^3 = 0$	$\Sigma d_i^4 = 164$

$$\mu_1 = \frac{\Sigma d_i}{N} = 0$$

$$\mu_2 = \frac{\Sigma d_i^2}{N} = 5$$

$$\mu_3 = \frac{\Sigma d_i^3}{N} = 0$$

$$\mu_4 = \frac{\Sigma d_i^4}{N} = 41.$$

$$\text{Skewness} = \frac{\mu_3^2}{\mu_2^3} = 0$$

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2} = \frac{41}{25} = 1.64$$

NOTE:

Relation between Central and Raw movements.

$$\mu_1 = \mu_1' - \mu_1' = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4.$$

(*) Find first 4 Central movements for the following data.

x	1	2	3	4	5
f	2	3	5	4	1

Sol:-

$N = \Sigma f_i$

$N = 15$

$$\bar{x} = \frac{\Sigma x_i f_i}{N} = \frac{2+6+15+16+5}{15}$$

$$= \frac{44}{15} = 2.93$$

∴ Assumed mean

$A = 3.$

Raw moments about Assumed mean $A = 3$.

f	x	$d_i = x_i - A$	d_i^2	d_i^3	d_i^4
2	1	-2	4	-8	16
3	2	-1	1	-1	1
5	3	0	0	0	0
4	4	1	1	1	1
1	5	2	4	8	16

(5)

$fi di$	$fi di^2$	$fi di^3$	$fi di^4$
-4	8	-16	32
-3	3	-3	3
0	0	0	0
4	4	4	4
2	4	8	16
$\Sigma fi di = -1$	$\Sigma fi di^2 = 19$	$\Sigma fi di^3 = -7$	$\Sigma fi di^4 = 55$

$$\mu_1' = \frac{\Sigma fi di}{N} = \frac{-1}{15} = -0.06$$

$$\mu_2' = \frac{\Sigma fi di^2}{N} = \frac{19}{15} = 1.266$$

$$\mu_3' = \frac{\Sigma fi di^3}{N} = \frac{-7}{15} = -0.466$$

$$\mu_4' = \frac{\Sigma fi di^4}{N} = \frac{55}{15} = 3.66$$

$\mu_1 = 0.$

$$\mu_2 = 1.266 - 0.0036$$
$$= 1.2624$$

$$\mu_3 = -0.466 - 3(1.26)(-0.06) + 2(-0.06)^3$$
$$= -0.466 + 0.2268 - 0.000432$$
$$= -0.466432 + 0.2268$$
$$= -0.2396.$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$
$$= 3.66 - 4(-0.46)(-0.06) + 6(1.266)(-0.06)^2 - 3(-0.06)^4$$
$$= 3.66 - 0.1104 + 0.027 - 0.00133$$
$$= 3.795$$

(*) calculate first 4 raw movements for the following data.

Class Interval	60-62	63-65	66-68	69-71	72-74
frequency	5	18	42	27	8

Sol:

Assumed mean (A) = 67

X	f	d_i	d_i^2	d_i^3	d_i^4
61	5	-6	36	-216	1296
64	18	-3	9	-27	81
67	42	0	0	0	0
70	27	3	9	27	81
73	8	6	36	216	1296

	$f_i d_i$	$f_i d_i^2$	$f_i d_i^3$	$f_i d_i^4$
	-30	180	-1080	6480
	-54	162	-486	1458
	0	0	0	0
	81	243	729	2187
	48	288	1728	10368
N = 100	45	873	891	20493

$\mu_1 = 0.45$ $\mu_3 = 8.91$

$\mu_2 = 8.73$ $\mu_4 = 204.93$

Moments

Moments are the constants of a Data which help in deciding the characteristics of the population.

Moments help in finding AM, standard deviation and variance of the population directly, and they help in knowing the graphic shapes of the population(Data).

We can call moments as the constants used in finding the graphic shape which helps a lot in characterizing a population(Data).

Thus, Moments are a set of statistical parameters(Mean,Variance,...) to measure a distribution.

Moment can be defined as the average of deviations of observations taken from a point that are raised to a certain power. Broadly they are classified as follows:

1.Central Moments (where deviations will be taken from mean)

2. Raw Moments (where deviations will be taken from an arbitrary point)

1. $\mu_r = r^{th} Central\ moment = \frac{1}{n}\sum(x-\bar{x})^r$ (individual data), where $\bar{x}$ is the Mean of the given data

$$= \frac{1}{N}\sum f * (x-\bar{x})^r \text{ (grouped data)}$$

2. $\mu_r' = r^{th} raw\ moment\ about\ a\ point\ "a" = \frac{1}{n}\sum(x-a)^r$ (individual data)

$$= \frac{1}{N}\sum f * (x-a)^r \text{ (grouped data)}$$

Note:

If we take a=0,the raw moments are called as Moments about Origin, which are given as

$\mu_r' = r^{th} raw\ moment\ about\ origin = \frac{1}{n}\sum x^r$ (individual data)

$$= \frac{1}{N}\sum f * x^r \text{ (grouped data)}$$

Generally we calculate first four moments to describe the data. Some of the important moments are

Central Moments	Raw Moments
$\mu_1 = 1^{st} Central\ moment$ $= \frac{1}{n}\sum(x-\bar{x})^1$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^1$ (grouped data)	$\mu_1' = 1^{st}$ *raw moment about a point a* $= \frac{1}{n}\sum(x-a)^1$ (individual data) $= \frac{1}{N}\sum f*(x-a)^1$ (grouped data) Note: If a=0, μ_1'=Mean
$\mu_2 = 2^{nd} Central\ moment$=Variance= σ^2 $= \frac{1}{n}\sum(x-\bar{x})^2$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^2$ (grouped data) Standard deviation=$\sqrt{Varaince}=\sigma$	$\mu_2' = 2^{nd}$ *raw moment about a point a* $= \frac{1}{n}\sum(x-a)^2$ (individual data) $= \frac{1}{N}\sum f*(x-a)^2$ (grouped data)
$\mu_3 = 3^{rd} Central\ moment$ $= \frac{1}{n}\sum(x-\bar{x})^3$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^3$ (grouped data)	$\mu_3' = 3^{rd}$ *raw moment about a point a* $= \frac{1}{n}\sum(x-a)^3$ (individual data) $= \frac{1}{N}\sum f*(x-a)^3$ (grouped data)
$\mu_4 = 4^{th} Central\ moment$ $= \frac{1}{n}\sum(x-\bar{x})^4$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^4$ (grouped data)	$\mu_4' = 4^{th}$ *raw moment about a point a* $= \frac{1}{n}\sum(x-a)^4$ (individual data) $= \frac{1}{N}\sum f*(x-a)^4$ (grouped data)

Generally we calculate first four moments to describe the data. Some of the important moments are

Central Moments	Raw Moments
$\mu_1 = 1^{st} Central\ moment$ $= \frac{1}{n}\sum(x-\bar{x})^1$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^1$ (grouped data)	$\mu_1' = 1^{st}$ *raw moment about a point* a $= \frac{1}{n}\sum(x-a)^1$ (individual data) $= \frac{1}{N}\sum f*(x-a)^1$ (grouped data) Note: If a=0, μ_1'=Mean
$\mu_2 = 2^{nd} Central\ moment$=Variance= σ^2 $= \frac{1}{n}\sum(x-\bar{x})^2$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^2$ (grouped data) Standard deviation=$\sqrt{Varaince}=\sigma$	$\mu_2' = 2^{nd}$ *raw moment about a point a* $= \frac{1}{n}\sum(x-a)^2$ (individual data) $= \frac{1}{N}\sum f*(x-a)^2$ (grouped data)
$\mu_3 = 3^{rd} Central\ moment$ $= \frac{1}{n}\sum(x-\bar{x})^3$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^3$ (grouped data)	$\mu_3' = 3^{rd}$ *raw moment about a point a* $= \frac{1}{n}\sum(x-a)^3$ (individual data) $= \frac{1}{N}\sum f*(x-a)^3$ (grouped data)
$\mu_4 = 4^{th} Central\ moment$ $= \frac{1}{n}\sum(x-\bar{x})^4$ (individual data) $= \frac{1}{N}\sum f*(x-\bar{x})^4$ (grouped data)	$\mu_4' = 4^{th}$ *raw moment about a point* a $= \frac{1}{n}\sum(x-a)^4$ (individual data) $= \frac{1}{N}\sum f*(x-a)^4$ (grouped data)

Measures of Central tendency:

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within the given data set. As such, they are sometimes called measures of central location. They are also classed as summary statistics.

Each **measure** describes a different indication of the typical or ***central value*** in the distribution.

Three most common measures of central tendency:

- Mean
- Median
- Mode

ARITHMETIC MEAN

The arithmetic mean is the most common measure of central tendency.

It is simply the sum of the observations(numbers) divided by the total number of observations(numbers).

The symbol "μ" is used for the mean.

It is calculated as follows:

- $Mean = \frac{Sum\ of\ observations}{Total\ number\ of\ observations} = \frac{\sum x}{n}$; (in individual data)
- $Mean = \frac{\sum fx}{N}$; (in Grouped data)

Problems:

Problem 1: The heights of five runners are 140 cm, 157 cm, 146 cm, 151 cm and 167 cm respectively. Find the mean height per runner.

Solution: $Mean\ height = \frac{Sum\ of\ heights\ of\ the\ runners}{number\ of\ runners}$

$= \frac{(140+157+146+151+167)}{5} = 152.2$

Problem2: If the mean of 9, 8, 10, x, 12 is 15, find the value of x.

Solution: Mean of the given numbers $= \frac{(9+8+10+x+12)}{5} = \frac{(39+x)}{5}$

According to the problem, mean = 15 (given).

Therefore, $(39 + x)/5 = 15$

$\Rightarrow 39 + x = 15 \times 5$

$\Rightarrow 39 + x = 75$

$\Rightarrow 39 - 39 + x = 75 - 39$

$\Rightarrow x = 36$

Hence, $x = 36$.

Problem3: The mean of 40 numbers was found to be 38. Later on, it was detected that a number 56 was misread as 36. Find the correct mean of given numbers.

Solution: Calculated mean of 40 numbers = 38.

Therefore, calculated sum of these numbers = $(38 \times 40) = 1520$.

Correct sum of these numbers

= [1520 - (wrong item) + (correct item)]

= (1520 - 36 + 56)

= 1540.

Therefore, the correct mean = 1540/40 = 38.5.

Problem4: Calculate Mean for the following data:

Age groups	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	40	53	58	64	72	49	36	25

Solution: To find mean, we calculate the following table:

Age (C.I)	f	Mid Points(x)	f*x
0-10	40	5	200
10-20	53	15	795
20-30	58	25	1450
30-40	64	35	2240
40-50	72	45	3240
50-60	49	55	2695
60-70	36	65	2340
	N= $\sum f$ =397		$\sum$ fx = 14835

$$Mean = \frac{\sum f\,x}{N}$$
$$= \frac{14835}{397}$$
$$= 37.36\ years$$

Problem5: If the mean of 34 individuals age is 14.76 which are distributed in the following data, calculate x and y values

C.I	10-12	12-14	14-16	16-18	18-20
f	4	12	x	14	y

Solution: To find x, we calculate the following table

Age (C.I)	f	Mid points (x)	fx
10-12	4	11	44
12-14	12	13	156
14-16	x	15	15x
16-18	14	17	238
18-20	y	19	19y
	N=30+x+y		438+15x+19y

Given that N=34

$$\Rightarrow 30 + x + y = 34$$

$$\Rightarrow x + y = 4 \ldots\ldots (1)$$

We know that

$$Mean = \frac{\sum f x}{N}$$

$$14.76 = \frac{438+15x+19y}{34}$$

$$\Rightarrow 438+15x+19y=738$$

$$\Rightarrow 15x+19y=300 \ldots\ldots (2)$$

Solving (1) &(2),we get

$x = 3.04 \cong 3 \ and \ y \cong 1$

MEDIAN: It is the middle most value of the data which divide the data into exactly two equal halves.

Individual series:

1. Arrange the data either in ascending or in descending order.
2. Count the number of observations(n).
3. A) If n is odd, Median $=\left(\frac{n+1}{2}\right)^{th}\ term$

   B) If n is even, Median is the average of middle most two values.

   i.e., $Median = \frac{\left(\frac{n}{2}\right)^{th}\ term + \left(\frac{n+2}{2}\right)^{th}\ term}{2}$

Grouped Data: $Median = l + \frac{\frac{N}{2} - cf}{f} * c$

Where l =lower limit of Median class

N=Total frequency

f =frequency of Median class

cf=cumulative frequency of preceeding class to the Median class

c=width of the class interval

Problem4: Find Median for the following data:

CI	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
f	15	10	5	15	20	25	10	15	25

Solution: To find Median, we calculate the following table :

Class interval	frequency	C.f
5-10	15	15
10-15	10	25
15-20	5	30
20-25	15	45
25-30	20	65^{cf}
30-35	25^{f}	90
35-40	10	100
40-45	15	115
45-50	25	140
	N=140	

$$\frac{N}{2} = \frac{140}{2} = 70,$$

$$Median = l + \frac{\frac{N}{2} - cf}{f} * c$$

$$= 30 + \frac{70-65}{25} * 5 = 31$$

MODE: It is the value of data that most often(frequently) occurs.

Individual data, find the value of data that has maximum frequency(highest repetitions)

Grouped data: Mode can be calculated by using the following formula:

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} * c$$

Where

$l = lower\ limit\ of\ Model\ class$(Class that has highest frequency)

$f_1 = frequency\ of\ Model\ class$

$f_0 = frequency\ of\ class\ that\ is\ preceeding\ to\ Model\ class$

$f_2 = frequency\ of class\ that\ is\ suceeding\ to Model\ class$

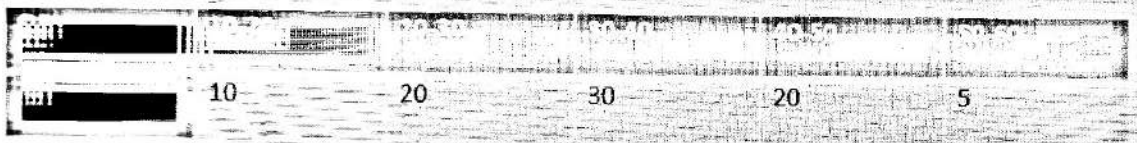
$c = width\ of\ the\ class\ Interval$

Problem1: Calculate Mode for the following:

12,23,34,12,23,45,34,56,34,34,34,67,12,34,15

Solution: As the item '34' has maximum repetitions, Mode for the given data=**34**

Problem2: Calculate Mode for the following:



Solution: To find Mode, we calculate the following table:

10-20	10
20-30	20^{f_0}
l_{30-40}	30^{f_1}
40-50	20^{f_2}
50-60	5

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} * c$$

$$= 30 + \frac{30-20}{2*30-20-20} * 10$$

$$= 35$$

Note: The empirical relation between Mean, Median and Mode is Mode=3Median-2 Mean

2. Probability Distributions.

* there are 2 types of distributions

1. Discrete theoritical distributions.
   a) Binomal distribution
   b) Poison distribution
2. Continuous theoritical distributions.
   a) Normal distribution
   b) Student's T distribution.
   c) chi- distribution.
   d) F distribution.

Binomial Distribution:

Let a random experiment Performed repeatedly 'n' times. Let, the occurrence of an event be a success and it's not occurrence a failure. Assume that the probability of success is P is constant for Each trial. then, the probability of x successes in 'n' independent trials in a specified order is given by the Compound Probability of (SFSSFSS...SFSSS)

$$P(SFSFSS \text{ ---- } SFSSS)$$

$$= P(S) \cdot P(F) \cdot P(S) \cdot P(F) \text{ ------ } P(S) \cdot P(F) \cdot P(S) \cdot P(S) \cdot P(S).$$

* If q is the probability of failure.

$$q = 1 - P$$

$$P + q = 1.$$

* $P(SFSFSS \ldots SFSSS)$

$$= P \cdot q \cdot P \cdot q \cdot P \cdot P \cdots P \cdot q \cdot P \cdot P \cdot P$$

$$= P^x \cdot q^{n-x} \quad (\because \text{'x' successes})$$

* In 'n' trials, x success can occur in ${}^nC_x$ ways. For each of these ways, the probability of x success is same. Hence, the probability distribution of the number of success so obtained is called "Binomial Probability distribution."

* A random variable X is said to be follow the binomial distribution if it assumes only non-negative values and its probability Mass function is given by

$$P(X=x) = P(x) = \begin{cases} {}^nC_x\, p^x\, q^{n-x} & x = 0, 1, 2, \cdots \\ 0 & \text{otherwise.} \end{cases}$$

NOTE:

The 2 Constants n, p are known as "Parameters of the Binomial distribution."

Mean of Binomial distribution:

$$E(P(x)) = \sum_{x=0}^{n} x \cdot P(x)$$

$$= \sum_{x=0}^{n} x \cdot {}^nC_x \cdot p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \cdot \frac{n!}{(n-x)!\, x!}\, p^x q^{n-x}$$

$$= np \sum_{x=0}^{n} \frac{(n-1)!}{(n-x)!\,(x-1)!}\, p^{x-1} q^{n-x}$$

$$= np\,(p+q)^{n-1}$$

$$= np \cdot 1$$

$$= np$$

$q^{(n-1)-(x-1)} = q^{n-x}$

Variance of Binomial Distribution:

* we have $Var(x) = E(x^2) - [E(x)]^2$

$$E(x^2) = \sum_{x=0}^{n} x^2 P(x)$$

$$= \sum_{x=0}^{n} [x(x-1)+x] P(x)$$

$$= \sum_{x=0}^{n} x \cdot P(x) + \sum_{x=0}^{n} x(x-1) \cdot P(x)$$

$$= np + \sum_{x=0}^{n} x(x-1) \; {}^{n}C_x \; p^x q^{n-x}$$

$$= np + \sum_{x=0}^{n} x(x-1) \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

$$= np + n(n-1)p^2 \sum_{x=0}^{n} \frac{(n-2)!}{(n-x)!\,(x-2)!} p^{x-2} q^{n-x}$$

$$= np + n(n-1)p^2 (p+q)^{n-2}$$

$$= np + n(n-1)p^2$$

$$= np[1 + (n-1)p]$$

$$= n^2p^2 - np^2 + np$$

$$Var(x) = E(x^2) - [E(x)]^2$$

$$= n^2p^2 - np^2 + np - (np)^2$$

$$= np(1-p)$$

$$= npq.$$

Mean B.D	$= np$
Var B.D	$= npq$

Recurrance Relation for the Probabilites of BD:

$$P(x+1) = {}^{n}C_{x+1}\, p^{x+1}\, q^{n-(x+1)}$$

$$P(x) = {}^{n}C_{x}\, p^{x}\, q^{n-x}$$

Now,

$$\frac{P(x+1)}{P(x)} = \frac{{}^{n}C_{x+1}\, p^{x+1}\, q^{n-(x+1)}}{{}^{n}C_{x}\, p^{x}\, q^{n-x}}$$

$$\frac{P(x+1)}{P(x)} = \frac{\frac{n!}{(n-(x+1))!\,(x+1)!}\, p^{x+1}\, q^{n-(x+1)}}{\frac{n!}{(n-x)!\,x!}\, p^{x}\, q^{n-x}}$$

$$\frac{P(x+1)}{P(x)} = \frac{n-x}{x+1}\,\frac{p}{q}$$

$$\therefore \boxed{P(x+1) = \left(\frac{n-x}{x+1}\,\frac{p}{q}\right) P(x)}$$

NOTE:

The recurrance relation is useful only the initial Probability $P(0)$ is known.

(*) 10 coins are thrown simultaneously. Find the Probability of getting atleast 7 heads.

Sol:-

$n = 10$

$x >= 7$

p - getting head (success) $\Rightarrow p = 1/2$

q - getting tail (Failure) $\Rightarrow q = 1/2$

$$P(x \geq 7) = ?$$

$$P(x \geq 7) = P(x=7) + P(x=8) + P(x=9) + P(x=10)$$

(11)

$$= {}^{10}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 + {}^{10}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 + {}^{10}C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 + {}^{10}C_{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0$$

$$= \left({}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\right)\left(\frac{1}{2}\right)^{10}$$

$$= \left(\frac{1}{2}\right)^{10} \left(\sum_{n=7}^{10} {}^{10}C_n\right)$$

$$\therefore \left(\frac{1}{2}\right)^{10} \left(120 + 45 + 10 + 1\right) = \frac{176}{32 \cdot 32} = \frac{11}{64}$$

(*) Two dices are thrown 5 times. Find the probability of getting 6 as sum.

1) 2 times
2) atleast once.
3) $P(1 < x < 5)$

(3,3) (4,2) (2,4) (5,1) (1,5)

Sol:-

$n = 5$

$p = \frac{5}{36}$

$q = \frac{31}{36}$

1) $x = 2 \Rightarrow P(x) = {}^nC_x \, p^x q^{n-x}$

$$P(x=2) = {}^5C_2 \left(\frac{5}{36}\right)^2 \left(\frac{31}{36}\right)^3$$

$$P(x=2) = \frac{10(25)(29791)}{(36)^5}$$

$$P(x=2) =$$

2) $P(x \geq 1) = 1 - P(x=0)$

$$= 1 - {}^5C_0 \, p^0 q^{n-0}$$

$$= 1 - {}^5C_0 \left(\frac{31}{36}\right)^5$$

$$= 1 - \left(\frac{31}{36}\right)^5$$

3) $P(1 < x < 5) = P(x=2) + P(x=3) + P(x=4)$

$= {}^5C_2 \left(\frac{5}{36}\right)^2 \left(\frac{31}{36}\right)^3 + {}^5C_3 \left(\frac{5}{36}\right)^3 \left(\frac{31}{36}\right)^2 + {}^5C_4 \left(\frac{5}{36}\right)^4 \left(\frac{31}{36}\right)$

$= \dfrac{{}^5C_2 (5)^2 (31)^3 + {}^5C_3 (5)^3 (31)^2 + {}^5C_4 (5)^4 (31)}{(36)^5}$

(*) 20 % Percernt of items Produced from a factory are defective. Find the probability that in a Sample of 5 choosen at random

1. None is defective
2. one is defective
3. $P(1 < x < 4)$

Sol:- $X \Rightarrow$ defective

$P = 20\% = 0.2$

$q = 0.8$

$n = 5$.

1. None is defective.

$P(x=0) = {}^5C_0 (0.2)^0 (0.8)^5$

$= 0.3276$

2. one is defective.

$P(x=1) = {}^5C_1 (0.2) (0.8)^4$

$= 0.4096$

3. $P(1 < x < 4) = P(x=2) + P(x=3)$

$= {}^5C_2 (0.2)^2 (0.8)^3 + {}^5C_3 (0.2)^3 (0.8)^2$

$= {}^5C_2 (0.020 + 0.00512)$

$= 10 (0.025)$

$= 0.256$

(*) In 256 sets of 12 tosses of a fair coin, In how many cases one may expect 8 heads.

Sol:

x - head

$n = 12$

$P(x=8) = {}^{12}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^4$

$= 0.1208.$

No. of times 8 heads occurs $f = N \cdot P(x=8)$

$= 256 \times 0.1208$

$= 30.93$

$\simeq 31.$

(*) 7 coins are tossed and the no. of heads are noted. The experiment is repeated 128 times and the following distribution is obtained.

No. of heads	0	1	2	3	4	5	6	7
Frequency	7	6	19	35	30	23	7	1

Fit a binomial distribution for the given data.

Sol:

x - head

$n = 7$

$p = 1/2$

$q = 1/2$

$P(x=0) = {}^{7}C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^7$

$= \frac{1}{128}$

$P(x=1) = {}^{7}C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^6$

$= 7/128$

$P(x=2) = P(2)$

$= 21/128$

$P(x=3) = {}^{7}C_3 \left(\frac{1}{2}\right)^7$

$= 35/128$

$P(x=4) = {}^{7}C_4 \left(\frac{1}{2}\right)^7$

$= 35/128$

$P(x=5) = {}^{7}C_5 \left(\frac{1}{2}\right)^7 = 21/128$

$P(x=6) = {}^{7}C_6 \left(\frac{1}{2}\right)^7 = 7/128$

$P(x=7) = {}^{7}C_7 \left(\frac{1}{2}\right)^7 = 1/128.$

Frequency:

$fe_0 = N \cdot P(0)$

$= 128 \times \frac{1}{128} = 1$

$fe_1 = 7$

$fe_2 = 21$

$fe_3 = 35$

$fe_4 = 35$

$fe_5 = 21$

$fe_6 = 7$

$fe_7 = 1.$

(*) Fit a binomial distribution for the following data.

x	0	1	2	3	4	5
f	2	14	20	34	22	8

Sol:

$n = 5$

for grouped data,

$$\text{mean} = \frac{\sum x_i f_i}{\sum f_i}$$

$$\text{mean} = \frac{284}{100} = 2.84$$

for Binomial distribution,

$$\text{mean} = np$$

$$2.84 = 5(p)$$

$$p = 0.56 \Rightarrow q = 0.44$$

$P(x=0) = {}^5C_0 (0.56)^0 (0.44)^5 = 0.016 \Rightarrow fe_0 = 1.6$

$P(x=1) = {}^5C_1 (0.56)^1 (0.44)^4 = 0.106 \Rightarrow fe_1 = 10.6$

$P(x=2) = {}^5C_2 (0.56)^2 (0.44)^3 = 0.261 \Rightarrow fe_2 = 26.7$

$P(x=3) = {}^5C_3 (0.56)^3 (0.44)^2 = 0.339 \Rightarrow fe_3 = 33.9$

$P(x=4) = {}^5C_4 (0.56)^4 (0.44)^1 = 0.216 \Rightarrow fe_4 = 21.6$

$P(x=5) = {}^5C_5 (0.56)^5 = 0.055 \Rightarrow fe_5 = 5.5$

$\therefore$ $fe_0 = 2$ $\quad fe_3 = 34$

$fe_1 = 10$ $\quad fe_4 = 22$

$fe_2 = 27$ $\quad fe_5 = 5$

Limitation of Binomial Distribution:

* The Binomial distribution fails when

1. the no. of trials 'n' become indefinitely large i.e. $n \to \infty$.
2. the success probability 'p' is very small i.e. $p \to 0$ such that $np = \lambda$ (finite).

* under the above cases, the binomial distribution takes a new form called "Poisson distribution."

Poisson Distribution:

A random variable X is said to follow Poisson distribution if it assumes non-negative values and its probability mass function is given by

$$P(X=x) = P(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0,1,2,\cdots \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0,1,2,\cdots \\ 0 & \text{otherwise} \end{cases}$$

Mean of Poisson Distribution:

$$P(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0,1,2,\cdots \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Mean} = \mu = E(x)$$

$$= \sum_{x=0}^{\infty} x \cdot P(x)$$

$$= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^{x-1}\cdot\lambda}{x(x-1)!}$$

$$= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda e^{-\lambda}\left[\frac{\lambda^0}{0!} + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} \cdots\right]$$

$$= \lambda e^{-\lambda}(e^{\lambda})$$

$$= \lambda.$$

Variance of Poisson distribution:

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$

$$E(x^2) = \sum_{x=0}^{\infty} x^2 P(x)$$

$$= \sum_{x=0}^{\infty} [x(x-1) + x] P(x)$$

$$= \sum_{x=0}^{\infty} x(x-1) P(x) + \underbrace{\sum_{x=0}^{\infty} x \cdot P(x)}_{\text{mean} = \lambda}$$

$$= \lambda + \sum_{x=0}^{\infty} x(x-1) P(x)$$

$$= \lambda + \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \lambda + \sum_{x=0}^{\infty} \frac{x(x-1) e^{-\lambda} \lambda^2 \lambda^{x-2}}{x(x-1)(x-2)!}$$

$$= \lambda + \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}$$

$$= \lambda + \lambda^2 e^{-\lambda} \left[\frac{\lambda^0}{0!} + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \dots \right]$$

$$= \lambda + \lambda^2 e^{-\lambda} e^{\lambda}$$

$$= \lambda^2 + \lambda$$

$$\text{Var}(x) = \lambda^2 + \lambda - \lambda^2$$

$$= \lambda.$$

* In poisson distribution,

$$\text{mean} = \text{Variance} = \lambda$$

$$= np.$$

Recurrance Relation:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$P(x+1) = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}$$

$$\frac{P(x+1)}{P(x)} = \frac{e^{-\lambda}\lambda^{x}\cdot\lambda}{(x+1)\,x!} \times \frac{x!}{e^{-\lambda}\lambda^{x}}$$

$$\frac{P(x+1)}{P(x)} = \frac{\lambda}{x+1} \Rightarrow P(x+1) = \left\{\frac{\lambda}{x+1}\right\} P(x)$$

(*) If the probability that any individual suffers a bad reaction from a certain injection is 0.001. Determine the Probability that out of 2000 individuals

1. Exactly 3
2. more than 2 individuals
3. none suffers a bad reaction.

sol:

success => bad reaction.

$p = 0.001$

$n = 2000$

$np = 2 \Rightarrow \lambda = 2.$

1. Exactly 3

$$P(x=3) = \frac{e^{-2}(2)^{3}}{3!}$$

$$= \frac{e^{-2}(8)}{6} = \frac{4}{3}e^{-2}$$

2. more than 2

$$P(x>2) = 1 - P(x \le 2)$$

$$= 1 - \left[P(x=0) + P(x=1) + P(x=2) \right]$$

$$= 1 - \left[\frac{e^{-2}(2)^0}{0!} + \frac{e^{-2}(2)^1}{1!} + \frac{e^{-2}(2)^2}{2!} \right]$$

$$= 1 - e^{-2}\left[1 + 2 + 2\right]$$

$$= 1 - 5e^{-2}$$

$$= 0.323.$$

3. NONE.

$$P(x=0) = \frac{e^{-2}(2)^0}{0!}$$

$$= e^{-2}$$

$$= 0.135$$

(*) If a bank received on the average 6 bad cheques Per day. Find the Probability that it will receive 4 bad cheques on any given day.

Sol:

$x \Rightarrow$ Bad cheque.

$$\text{Avg} = 6$$

$$\mu = 6 \Rightarrow \lambda = 6$$

$$P(x=4) = \frac{e^{-6}(6)^4}{4!}$$

$$= 0.134$$

(*) A car hire firm has 2 cars which it hires out day by day. The number of demands for 1 car on each day is distributed as a poisson with mean 1.5. Calculate proportion of days

1. on which there is no demand
2. on which demand is refused.

Sol: Given that,

the mean $\lambda = 1.5$ and the demand is in Poisson distribution

$$\therefore P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

1. no demand

$$P(x=0) = \frac{e^{-1.5}(\lambda)^0}{0!}$$

$$= e^{-1.5}$$

$$= 0.223$$

days proportion

$= 365 \times 0.223$

$= 81.395$

$= 81$ days.

2. demand is refused only when the no. of demands is more than 2.

i.e., $P(x>2) = 1 - P(x \leq 2)$

$$= 1 - \left[P(x=0) + P(x=1) + P(x=2)\right]$$

$$= 1 - e^{-1.5}\left[1 + 1.5 + \frac{2.25}{2}\right]$$

$$= 1 - 0.809$$

$$= 0.191.$$

days proportion

$= 365 \times 0.191$

$= 70$ days.

(*) If a random variable has poisson distribution such that $P(1) = P(2)$ then find

1. Mean of the distribution
2. $P(4)$
3. $P(x \geq 1)$

Sol:

$$P(1) = P(2)$$

$$\frac{e^{-\lambda}\lambda^1}{1!} = \frac{e^{-\lambda}\lambda^2}{2!} \Rightarrow \lambda = \frac{\lambda^2}{2}$$

$$\boxed{\lambda = 2}$$

1. Mean $= \mu = \lambda$
   $= 2$

2. $P(4) = \frac{e^{-2}(2)^4}{4!} = 0.090$

3. $P(x \geq 1) = 1 - P(x = 0)$
   $= 1 - \frac{e^{-2}(2)^0}{0!} \Rightarrow 1 - e^{-2}$
   $= 0.865$

(*) A manufacturer of cotter pins knows that 5% of his Product is defective. Pins are sold in boxes of 100. He guarantees that not more than 10 pins will be defective. what is the probability a box will fail to meet the guranteed quality.

(21)

Sol:

$p = 5\% = 0.05$

$n = 100$

$\lambda = np = 100 \times 0.05$

$= 5$

P(fails the Gurantee)

$\Rightarrow P(x > 10) = 1 - P(x \le 10)$

$= 1 - e^{-5}\left[1 + \frac{5}{1!} + \frac{5^2}{2!} + \cdots + \frac{5^{10}}{10!}\right]$

(*) If a Poisson Distribution is such that 3 times of $P(x=4) = \frac{1}{2} \times P(x=2) + P(x=0)$. Then find mean and variance of poison distribution.

Sol:

$$3\ P(x=4) = \frac{1}{2} P(x=2) + P(x=0)$$

$$3\,\frac{e^{-\lambda}\lambda^4}{4!} = \frac{1}{2}\frac{e^{-\lambda}\lambda^2}{2!} + \frac{e^{-\lambda}\lambda^0}{0!}$$

$$\frac{3\lambda^4}{24\,e} = \frac{1}{2}\frac{\lambda^2}{2} + 1$$

$$\lambda^4 - 2\lambda^2 - 8 = 0$$

$\lambda = 2, -2, \sqrt{2}i, -\sqrt{2}i$

$\lambda = 2$ (X), -2 (X, $\downarrow$ $-ve$), $\sqrt{2}i$, $-\sqrt{2}i$ (X, X, $\sqsubset$ imaginary)

$\therefore \lambda = 2$

mean = variance = 2.

(29) (22)

(*) Fit a poisson distribution for the following data.

x	0	1	2	3	4	5	6	7	8
f	56	156	132	92	37	22	4	0	1

$\Sigma f(x) = 500$

Sol:

$$\text{Mean} = \frac{\Sigma x \cdot f(x)}{\Sigma f(x)}$$

$$= \frac{(0\times 56) + (1\times 156) + (2\times 132) + 3(92) + 4(37) + \cdots 8(1)}{500}$$

$$\text{Mean} = \lambda$$

$$= 1.972$$

$$P(x=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$$

$$= e^{-1.972}$$

$$= 0.1391$$

$$P(x=1) = \frac{e^{-1.972} \times (1.972)^1}{1!} = 0.274$$

$$P(x=2) = \frac{e^{-1.972} \times (1.972)^2}{2!} = 0.2704$$

$$P(x=3) = \frac{e^{-1.972} \times (1.972)^3}{3!} = 0.177$$

$$P(x=4) = \frac{e^{-1.972} \times (1.972)^4}{4!} = 0.087$$

$$P(x=5) = \frac{e^{-1.972} (1.972)^5}{5!} = 0.034$$

$$P(x=6) = \frac{e^{-1.972} (1.972)^6}{6!} = 0.011.$$

(23)

$$P(x=7) = \frac{e^{-1.972}(1.972)^7}{7!} = 0.0032$$

$$P(x=8) = \frac{e^{-1.972}(1.972)^8}{8!} = 0.0008.$$

$fe_0 = N \cdot P(0) = 500 \times 0.1391 = 67.55 \simeq 70$

$fe_1 = N \cdot P(1) = 500 \times 0.274 = 137$

$fe_2 = N \cdot P(2) = 500 \times 0.2704 = 135.2 = 135$

$fe_3 = N\ P(3) = 500 \times 0.177 = 88.5 = 89$

$fe_4 = N \cdot P(4) = 500 \times 0.087 = 43.5 = 43$

$fe_5 = N\ P(5) = 500 \times 0.014 = 17.35 = 17$

$fe_6 = N \cdot P(6) = 500 \times 0.011 = 5.7 \simeq 6$

$fe_7 = N \cdot P(7) = 500 \times 0.0032 = 1.6 \simeq 2$

$fe_8 = N \cdot P(8) = 500 \times 0.0008 = 0.4 \simeq 1$

x	0	1	2	3	4	5	6	7	8
f	56	156	132	92	37	22	4	0	1
fe	70	137	135	89	43	17	6	2	1

(*) Fit a Poisson distribution.

x	0	1	2	3	4	5
f(x)	142	156	69	27	5	1

Sol:

$$\text{mean} = \frac{\Sigma x \cdot f(x)}{\Sigma f(x)}$$

$$= \frac{0(142) + 1(156) + 2(69) + 3(27) + 4(5) + 5(1)}{400}$$

$$= 1$$

$$P(x=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-1} = 0.368$$

$$P(x=1) = \frac{e^{-1}(1)^1}{1!} = e^{-1} = 0.368$$

$$P(x=2) = \frac{e^{-1}}{2!} = 0.184 \qquad P(x=4) = \frac{e^{-1}}{4!} = 0.015$$

$$P(x=3) = \frac{e^{-1}}{3!} = 0.0613 \qquad P(x=5) = \frac{e^{-1}}{5!} = 0.0030$$

$fe_0 = N \cdot P(0)$
$= 400 \times 0.368$
$= 147$

$fe_1 = N\ P(1)$
$= 400 \times 0.3678$
$= 1.47$

$fe_2 = N\ P(2)$
$= 400 \times 0.184$
$= 74$

$fe_3 = N\ P(3)$
$= 400 \times 0.0613$
$= 25$

$fe_4 = N\ P(4)$
$= 400 \times (0.015)$
$= 6.0$

$fe_5 = N \cdot P(5)$
$= 400 \times (0.0030$
$= 1.2$
$= 1$

x	0	1	2	3	4	5
f	142	156	69	27	5	1
f_e	147	141	74	25	6	1

Normal Distribution: (1)

(Gaussian Distribution)

* A continuous random variable x is said to be follow Normal distribution if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where,

$$-\infty < x < \infty$$
$$-\infty < \mu < \infty$$
$$\sigma > 0$$

* The constants μ (mean) and σ (S.D) are known as Parameters of the Normal distribution.

* If the random variable X follows normal distribution with mean μ and S.D σ then it is represented as

$$X \sim N(\mu, \sigma)$$
$$X \sim B(x, P)$$
$$X \sim P(\lambda)$$

Standard Normal variable:

* If x is a random variable, follows the Normal distribution then the standard Normal variable is denoted by z and given by

$$z = \frac{x-\mu}{\sigma}$$

Mean of Standard Normal distribution:

$$E(z) = E\left(\frac{x-\mu}{\sigma}\right)$$

$$= \frac{1}{\sigma}\left[E(x) - E(\mu)\right]$$

$$= \frac{1}{\sigma}\left[\mu - \mu\right]$$

$$= 0.$$

$E(\mu)$

$$= \sum_{x=0}^{n} \mu \cdot P(x)$$

$$= \mu \sum_{x=0}^{n} P(x)$$

$$= \mu \cdot 1$$

$$= \mu.$$

Variance of Standard Normal distribution:

$$\text{var}(z) = E(z^2) - \left[E(z)\right]^2$$

$$= E(z^2) - \left[E\left(\frac{x-\mu}{\sigma}\right)\right]^2$$

$$= E\left(\left(\frac{x-\mu}{\sigma}\right)^2\right) - \frac{1}{\sigma^2}\left[E(x-\mu)\right]^2$$

$$= \frac{1}{\sigma^2}\left[E((x-\mu)^2) - \left[E(x-\mu)\right]^2\right]$$

$\text{var}(c) = 0.$

$$= \frac{1}{\sigma^2}\,\text{var}(x-\mu)$$

$$= \frac{1}{\sigma^2}\left[\text{var}(x) - \text{var}(\mu)\right]$$

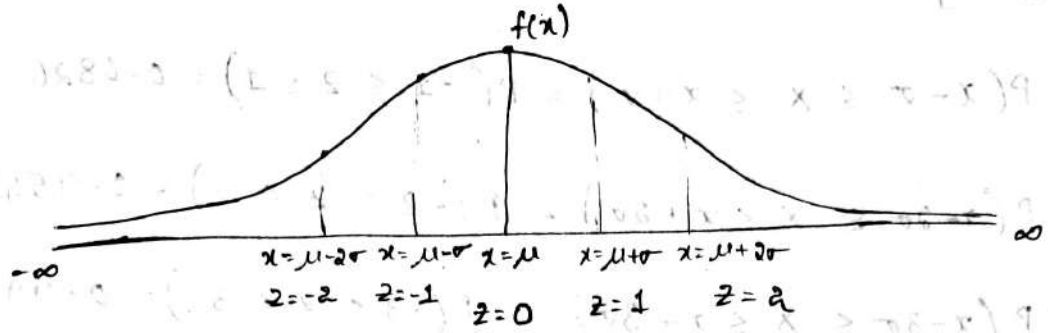
$$= \frac{1}{\sigma^2}\left[\sigma^2 - 0\right]$$

$$= 1.$$

Hence, the standard normal variable (z) follows normal distribution with 0, 1 · where 0 – mean
1 – variance.

$$= N(0, 1).$$

Characteristics of Normal curve: (Gaussian curve)



* The normal probability curve whose density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where,

$$z = \frac{x-\mu}{\sigma}$$

(or)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

has the following properties.

1. It is a bent shaped curve.
2. It is symmetrical about the line $x = \mu$.
3. Mean, mode, median of the normal distribution coincides.
4. The maximum probability occurs when $x = \mu$ and it is given by

$$\left(f(x)\right)_{max} = \frac{1}{\sigma\sqrt{2\pi}}$$

5. x increases numerically, $f(x)$ decreases rapidly.
6. Since, the probability can not be negative, no part of the curve lies below the x-axis.
7. x-axis is a asymptote to the curve.

Area property:

* Probability of $x-\sigma \le X \le x+\sigma$

$$P(x-\sigma \le X \le x+\sigma) = P(-1 \le z \le 1) = 0.6826$$

$$P(x-2\sigma \le X \le x+2\sigma) = P(-2 \le z \le 2) = 0.9544$$

$$P(x-3\sigma \le X \le x+3\sigma) = P(-3 \le z \le 3) = 0.9973$$

(*) X is normally distributed. Mean of X is 12 and Standard deviation is 4. Then, find $P(X \ge 20)$ and $P(0 \le X \le 12)$

Sol:

$\mu = 12$

$\sigma = 4$

$z = \frac{X-\mu}{\sigma}$

⇓

Standard normal variable.

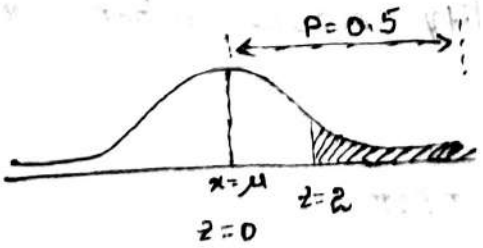
$P(X \ge 20)$

$X = 20$

$z = \frac{20-12}{4}$

$z = 2$

$\Rightarrow P(X \ge 20) = P(z \ge 2)$



$P(X \ge 20) = P(z \ge 2)$

$= 0.5 - P(0 \le z \le 2)$

$= 0.5 - 0.4772$

$= 0.0228.$

$P(0 \le X \le 12)$

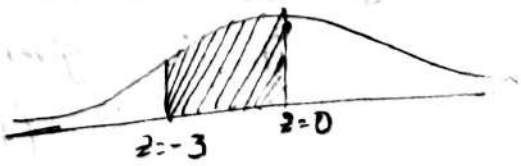
$X = 12$

$z = \frac{12-12}{4} = 0$

$X = 0$

$z = \frac{0-12}{4} = -3$

$\Rightarrow P(0 \le X \le 12) = P(-3 \le z \le 0)$



$P(0 \le X \le 12) = P(-3 \le z \le 0)$

$= P(0 \le z \le 3)$

$= 0.4987.$

Curve is Symmetric about $z=0$ i.e. $x=\mu$.

(5)

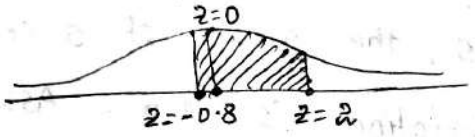
(*) X is a normal variable with mean $\mu = 30$ and standard deviation $\sigma = 5$. Then find $P(26 \le X \le 40)$ and $P(X \ge 45)$ and $P(|X-30| < 5)$

Sol: $\mu = 30$, $\sigma = 5$ $\qquad z = \frac{X-\mu}{\sigma}$

1) $P(26 \le X \le 40)$ $\qquad X = 26$ $\qquad X = 40$

$\Downarrow$ $\qquad z = \frac{26-30}{5}$ $\qquad z = \frac{40-30}{5}$

$P(-0.8 \le z \le 2)$ $\qquad z = -0.8$ $\qquad z = 2$



$P(-0.8 \le z \le 2) = P(-0.8 \le z \le 0) + P(0 \le z \le 2)$

$= P(0 \le z \le 0.8) + P(0 \le z \le 2)$

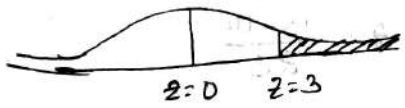
$= 0.2881 + 0.4772$

$= 0.7653$

2) $P(X \ge 45)$ $\qquad X = 45$

$= P(z \ge 3)$ $\qquad z = \frac{45-30}{5}$

$z = 3$



$P(z \ge 3) = 0.5 - P(0 \le z \le 3)$

$= 0.5 - 0.4987$

$= 0.0013$

3) $P(|X-30| < 5)$ $\qquad |X-30| < 5$

$= P(-1 < z < 1)$ $\qquad X - 30 < 5$ and $30 - X < 5$

$= P(-1 < z \le 0) + P(0 \le z < 1)$ $\qquad X < 35$ and $-X < -25$

$= P(0 \le z < 1) + P(0 \le z < 1)$ $\qquad X > 25$

$= 2 \cdot P(0 \le z < 1)$ $\qquad -5 < X - 30 < 5$

$= 2(0.3413)$ $\qquad 25 < X < 35$

$X = 25$ $\qquad X = 35$

$z = -1$ $\qquad z = 1$

(or)

3) $P(|x-30|<5)$ $\quad z = \frac{x-\mu}{\sigma}$

$= P\left(\left|\frac{x-30}{5}\right|<1\right)$ $\quad z = \frac{x-30}{5}$

$= P(|z|<1)$

$= P(-1 \le z \le 1)$

$= 2\,P(0 \le z \le 1) = 2 \times 0.3413$

$= 0.6826.$

(*) In a sample of 1000 cases, the mean of a certain test is 14 and standard deviation is 2.5. Assuming the distribution to be normal then find

1. How many students score between 12 and 15.
2. How many score above 18.

Sol:

$\mu = 14$

$\sigma = 2.5$ $\quad z = \frac{x-\mu}{\sigma} = \frac{x-14}{2.5}$

1. $P(12 \le x \le 15)$

$= P(-0.8 \le z \le 0.4)$

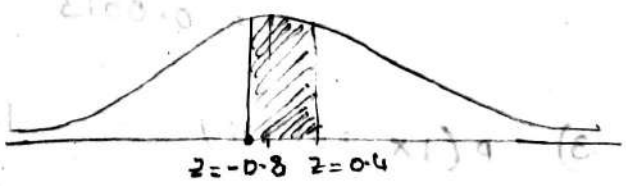
$= P(0 \le z \le 0.8) + P(0 \le z \le 0.4)$

$= 0.2881 + 0.1554$

$= 0.4435$

$x = 12$ $\quad z = \frac{12-14}{2.5}$ $\quad z = \frac{-2}{2.5}$ $\quad z = -0.8$

$x = 15$ $\quad z = \frac{15-14}{2.5}$ $\quad z = \frac{1}{2.5}$ $\quad z = 0.4$



∴ No. of students b/w 12 & 15 = total students × Probability

$= 0.4435 \times 1000$

$= 443.5$

$= 444.$

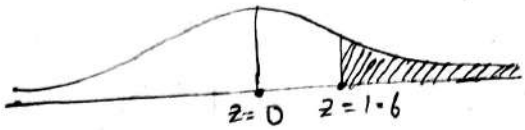
2. $P(X \geq 18)$ $\quad x = 18$

$= P(Z \geq 1.6)$ $\quad z = \frac{18-14}{2.5}$

$= 0.5 - P(0 \leq Z \leq 1.6)$ $\quad z = \frac{4}{2.5} \Rightarrow z = 1.6$

$= 0.5 - 0.4452$

$= 0.0548.$



∴ No. of students above 18. $= 0.0548 \times 1000$

$= 54.8$

$= 55$

(*) Suppose the weights of 800 students are normally distributed with mean $\mu = 140$ and S.D $\sigma = 10$. Find the no. of students whose weights are

1. b/w 142 to 154
2. more than 152.

Sol:-

$z = \frac{x - 140}{10}$

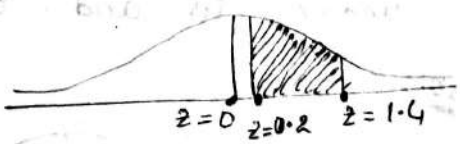
1. $P(142 \leq x \leq 154)$ $\quad x = 142$ $\quad z = \frac{142-140}{10}$ $\quad z = 0.2$ $\quad x = 154$ $\quad z = \frac{154-140}{10}$ $\quad z = 1.4$

$= P(0.2 \leq z \leq 1.4)$

$= P(0 \leq z \leq 1.4) - P(0 \leq z \leq 0.2)$

$= 0.4192 - 0.0793$

$= 0.3399.$



No. of students $= 0.3399 \times 800$

$= 271.92$

$= 272$

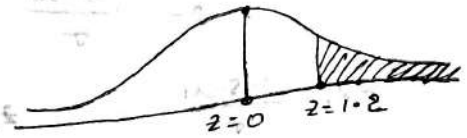
2. $P(x \geq 152)$ $\quad x = 152$ $\quad z = \frac{152-140}{10}$ $\quad z = 1.2$

$= P(z \geq 1.2)$

$= 0.5 - P(0 \leq z \leq 1.2)$

$= 0.5 - 0.3849$

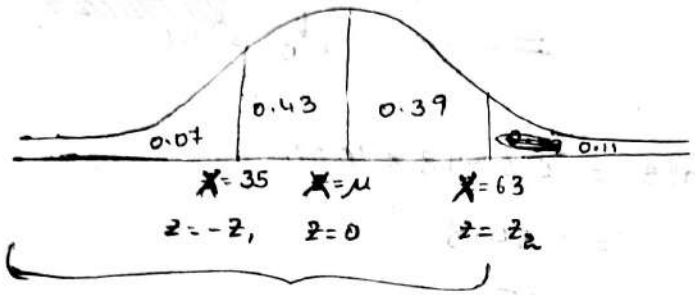
$= 0.1151$



No. of students $= 0.1151 \times 800 = 92.08 = 92$

(*) In a normal distribution, 7% of items are under 35 and a 89% are under 63. Determine mean and variance of the distribution.

Sol:



89% = 0.89

$P(0 \le z \le z_2) = 0.39$

$\Downarrow$

$z_2 = 1.23$

$P(-z_1 \le z \le 0) = P(0 \le z \le z_1)$

$= 0.43$

$\Downarrow$

$z_1 = 1.48$

$z = \frac{x-\mu}{\sigma}$

$-z_1 = \frac{35-\mu}{\sigma}$ $\quad z_2 = \frac{63-\mu}{\sigma}$

$\frac{63-\mu-35+\mu}{\sigma} = z_2 + z_1$

$\frac{28}{\sigma} = 2.71$

$\sigma = 10.33$

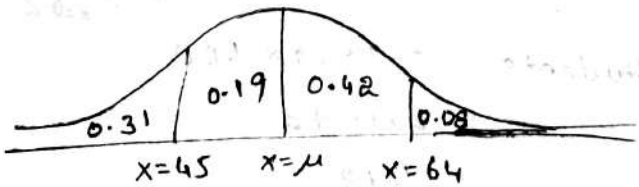
$-1.48 = \frac{35-\mu}{10.33}$

$\mu = 35 + 1.48(10.33)$

$\mu = 50.28$

(*) In a normal distribution, 31% of items are under 45 and 8% are over 64. Find μ, σ.

Sol:



$P(0 \le z \le z_2) = 0.42$

$\Downarrow$

$z_2 = 1.41$

$P(0 \le z \le z_1) = 0.19$

$z_1 = 0.5$

$z = \frac{x-\mu}{\sigma}$

$-z_1 = \frac{45-\mu}{\sigma}$ $\quad z_2 = \frac{64-\mu}{\sigma}$

$\frac{64-\mu-45+\mu}{\sigma} = 1.91$

$\frac{19}{1.91} = \sigma$

$\sigma = 9.94$

$$-0.5 = \frac{45-\mu}{9.94}$$

$$\mu = 45 + 9.94(0.5)$$

$$\mu = 49.97.$$

(1)

Joint Probability Distribution:

* If (x, y) (ordered pair) is a 2-dimensional discrete random variable then the Joint function of x, y is known as Joint Probability Mass function of x, y.

⟶ It is denoted by P_{xy} and defined by

$$P_{x,y}(x_i, y_i) = \begin{cases} P(x = x_i, y = y_i) & x = x_i \\ & y = y_i \\ 0 & \text{otherwise} \end{cases}$$

Probability Distribution function:

(or)

Cummulative Distribution function:

* Let X, Y are two Random variables of the 2-D discrete function, a real valued function F is said to be Joint distribution function if it is defined as for the values

$$F_{x,y}(x_i, y_i) = P(x \leq x_i, y \leq y_i)$$

Eg: Tossing of a ~~die~~ coin and rolling of a die simultaneously with random variables X, Y respectively, Then the Joint Probability distribution is

X \ Y	1	2	3	4	5	6
H	1/12	1/12	1/12	1/12	1/12	1/12
T	1/12	1/12	1/12	1/12	1/12	1/12

* Marginal distribution function of X:

* Marginal distribution of $X = x_i$ is denoted by $P(X = x_i)$ or $P(x_i)$ defined by

$$P(X = x_i) = P(X = x_i, Y = y_1) + P(X = x_i, Y = y_2) + \cdots + P(X = x_i, Y = y_n)$$

$$= \sum_{j=1}^{n} P(x_i, y_j)$$

* Similarly, the marginal distribution function of $Y = y_i$ is denoted by $P(Y = y_i)$ or $P(y_i)$ and defined by

$$P(Y = y_i) = P(X = x_1, Y = y_i) + P(X = x_2, Y = y_i) + \cdots + P(X = x_n, Y = y_i)$$

$$P(Y = y_i) = \sum_{j=1}^{n} P(x_j, y_i)$$

X \ Y	y_1 1	y_2 2	y_3 3	y_4 4	y_5 5	y_6 6	M·D·X
x_1 1	$1/36$ $P(x_1,y_1)$	$1/36$ $P(x_1,y_2)$	$1/36$	$1/36$	$1/36$	$1/36$ $P(x_1,y_6)$	$P(x_1)=\sum_{j=1}^{6}P(x_1,y_j)$
x_2 2	$1/36$ $P(x_2,y_1)$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$P(x_2)=\sum_{j=1}^{6}P(x_2,y_j)$
x_3 3	$1/36$ $P(x_3,y_1)$	.....				$1/36$	
x_4 4	⋮						
x_5 5							
x_6 6	$1/36$ $P(x_6,y_1)$	.....				$1/36$	$P(x_6)=\sum_{j=1}^{6}P(x_6,y_j)$
M·D·Y	$P(y_1)=\sum_{j=1}^{6}P(x_j,y_1)$					$P(y_6)=\sum_{j=1}^{6}P(x_j,y_6)$	1

NOTE:

* for a Joint Probability Mass function $P(x_i, y_i)$ must satisfy

1. $P(x_i, y_i) \geq 0$
2. $\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) = 1$

(*) The Joint Probability distribution of 2 random variables x, y is given by

$$P(x=0, y=1) = k/3$$

$$P(x=1, y=-1) = k/3$$

$$P(X=-1,\ Y=1) = k/3 \quad (12)$$

Then find

1. k
2. Marginal distribution of X, Y.

Sol:

	x_1	x_2	x_3
X / Y	-1	0	1
y_1 -1	0	0	$k/3$
y_2 1	$k/3$	$k/3$	0

$$\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) = 1$$

$$\frac{k}{3} + \frac{k}{3} + \frac{k}{3} = 1$$

$$\boxed{k = 1}$$

y \ X	-1	0	1	
-1	0	0	$1/3$	$P(-1) = 1/3$
1	$1/3$	$1/3$	0	$P(1) = 2/3$
	$P(-1) = 1/3$	$P(0) = 1/3$	$P(1) = 1/3$	1

Marginal probability of X

X	-1	0	1
M.P of X	$1/3$	$1/3$	$1/3$

Marginal probability of Y.

Y	-1	1
M.P of Y	$1/3$	$2/3$

(*) from the following bivarient Probability distribution of x and y

x \ y	1	2	3	4	5	6	
0	0	0	1/32	2/32	2/32	3/32	$P(x=0)=\frac{8}{32}$
1	1/16	1/16	1/8	1/8	1/8	1/8	$P(x=1)=\frac{5}{8}$
2	1/32	1/32	1/64	1/64	0	2/64	$P(x=2)=\frac{4}{32}$
	3/32	3/32	11/64	13/64	6/32	16/64	

Then find

1. Probability of $x \le 1, y = 2$
2. Probabity of $x \le 1$
3. $P(y \le 3)$
4. $P(x \le 2, y \le 4)$

Sol:-

1) $P(x \le 1, y = 2) = P(x=0, y=2) + P(x=1, y=2)$

$= 0 + 1/16$

$= 1/16$

2) $P(x \le 1) = P(x=0) + P(x=1)$

$= \frac{8}{32} + \frac{5}{8} = \frac{28}{32}$

3) $P(y \le 3) = P(1) + P(y=2) + P(y=3)$

$= 3/32 + 3/32 + 11/64 = 23/64$

4) $P(x \le 2, y \le 4) = P(y=1) + P(y=2) + P(y=3) + P(y=4)$

$= 36/64$

A 2 dimensional random variable have a bivarient distribution given by

$$P(X = x, Y = y) = \frac{x^2 + y}{32} \qquad x = 0, 1, 2, 3 \quad y = 0, 1$$

Then Construct Joint probability distribution, Marginal distribution for X and Y.

Sol:

Y \ X	0	1	2	3	
0	0	1/32	4/32	9/32	$P(Y=0) = \frac{14}{32}$
1	1/32	2/32	5/32	10/32	$P(Y=1) = \frac{18}{32}$
	$P(X=0)$ = 1/32	$P(X=1)$ = 3/32	$P(X=2)$ = 9/32	$P(X=3)$ = 19/32	

Marginal P. of X

X	0	1	2	3
M.P of X	1/32	3/32	9/32	19/32

Marginal P. of Y

Y	0	1
M.P of Y	14/32	18/32

JOINT Conditional probability mass function:

* Let (X, Y) is random variable in 2-D space then the conditional probability mass fn. is denoted by $P_{X/Y}(x/y)$ and defined by

$$P_{X/Y}(x/y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

$$= \frac{P(X=x \cap Y=y)}{P(Y=y)}$$

Similarly,

$$P_{Y/X}(y/x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$\Downarrow$

Probability of $Y=y$ for given $X=x$.

NOTE:

$$P(X=x, Y=y) = P(X=x \cap Y=y)$$

(*)

$$P(X=x, Y=y) = \frac{x^2+y}{32} \qquad x=0,1,2,3 \quad y=0,1$$

Find

$P_{X/Y}(2/1)$

$P_{Y/X}(0/3)$

Y \ X	0	1	2	3
0	0	$1/32$	$4/32$	$9/32$
1	$1/32$	$2/32$	$5/32$	$10/32$

$$P_{X/Y}(2/1) = \frac{P(X=2,\ Y=1)}{P(Y=1)}$$

$$= \frac{5/32}{18/32} = \frac{5}{18}$$

$$P_{Y/X}(0/3) = \frac{P(X=3,\ Y=0)}{P(X=3)}$$

$$= \frac{9/32}{19/32} = \frac{9}{19}.$$

JOINT Probability Density function:

* Let (x, y) is a random variable of 2 dimensional then the Joint probability density function is denoted by $f_{xy}(x, y)$ and defined by

$$f_{xy}(x, y) = P(X=x,\ Y=y)$$

* Also, the probability density fn. is obtained by differentiating the probability distribution fn. (or) Probability cummulative fn. i.e.
if $F(x, y)$ is a probability distribution fn. then,

$$f(x, y) = d(F(x, y))$$

$$F(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx$$

NOTE:

for a 2 dimensional random variable (x, y), a function $f_{xy}(x, y)$ is said to be probability density fn. if and only if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{xy}(x, y)\, dy\, dx = 1.$$

Marginal probability distribution fn. of X:

* The marginal distribution of x is denoted by $F_x(x)$ and defined by

$$F_x(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

* similarly, the marginal distribution of y is denoted by $F_y(y)$ and defined by

$$F_y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx.$$

Conditional probability density function:

* The Conditional density fn. of $X=x$ for given $Y=y$ is defined as

$$f_{x/y}(x/y) = \frac{f_{xy}(x,y)}{F_y(y)}$$

similarly,

$$f_{y/x}(y/x) = \frac{f_{xy}(x,y)}{F_x(x)}$$

(*) The Joint probability density fn. of a Dimensional random variable (x,y) is given by

$$f(x,y) = \begin{cases} 2 & 0<x<1 \;;\; 0<y<x \\ 0 & \text{otherwise.} \end{cases}$$

Then find marginal density fn of x and y. Find the Conditional density fn of x for given $Y=y$ and Conditional density fn of y for given $X=x$.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_{x=0}^{1}\int_{y=0}^{x} 2\,dy\,dx$$

$$= 2\int_{x=0}^{1} (y)_0^x\,dx = 2\int_{x=0}^{1} x\,dx$$

$$= 2\left(\frac{x^2}{2}\right)_0^1$$

$$= 1$$

$\therefore f(x,y)$ is a Joint Probability distribution function.

Marginal P. density fn. of $X = f_x(x)$

$$= \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$= \int_{0}^{x} 2\,dy = 2x$$

Marginal p. density fn. of $Y = f_y(y)$

$$= \int_{-\infty}^{\infty} f(x,y)\,dx$$

$$= \int_{0}^{1} 2\,dx = 2.$$

Conditional density function of x given $y = y$

$$f_{x/y}(x/y) = \frac{f_{xy}(x,y)}{f_y(y)}$$

$$= \frac{2}{2} = 1$$

Conditional density function of y given $x = x$

$$f_{y/x}(y/x) = \frac{f_{xy}(x,y)}{f_x(x)}$$

$$= \frac{2}{2x} = 1/x.$$

(*) Suppose the random variable (X, Y) have the joint density fn. defined by

$$f_{XY}(x,y) = \begin{cases} C(2x+y) & 2<x<6,\ 0<y<5 \\ 0 & \text{otherwise.} \end{cases}$$

then find C, $P(x>3, y>2)$, $P(x>3)$, $P(x+y<4)$

Sol:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

$$\int_{x=2}^{6}\int_{y=0}^{5} C(2x+y)\,dy\,dx = 1$$

$$C\int_{x=2}^{6}\left(2xy+\frac{y^2}{2}\right)_0^5 dx = 1$$

$$C\int_{x=2}^{6}\left(10x+\frac{25}{2}\right)dx = 1 \implies C\left[5x^2+\frac{25x}{2}\right]_2^6 = 1$$

$$C\left[(180+75)-(20+25)\right]=1$$

$$C\left[180+30\right]=1$$

$$\therefore \boxed{C=\frac{1}{210}}$$

$$P(x>3, y>2) = \int_{x=3}^{6}\int_{y=2}^{5}\frac{2x+y}{210}\,dy\,dx$$

$$= \frac{1}{210}\int_{x=3}^{6}\left(2xy+\frac{y^2}{2}\right)_2^5 dx$$

$$= \frac{1}{210}\int_{x=3}^{6}\left(10x+\frac{25}{2}-4x-\frac{4}{2}\right)dx$$

$$= \frac{1}{210}\int_{x=3}^{6}\left(6x+\frac{21}{2}\right)dx = \frac{61}{210}\int_{x=3}^{6}\left(6x+\frac{21}{2}\right)dx$$

$$= \frac{1}{210}\left[108+63-27-\frac{63}{2}\right]$$

$$= \frac{1}{210}\left[81+\frac{63}{2}\right] = \frac{162+63}{2(210)}$$

$$= \frac{225}{2(210)} = \frac{15}{28}.$$

$$P(x>3) = \int_{x=3}^{6} f(x,y)\,dx$$

$$= \int_{y=0}^{5}\int_{x=3}^{6} \frac{1}{210}(2x+y)\,dx\,dy = \int_{y=0}^{5} \frac{1}{210}\left(x^2+xy\right)_3^6\,dy$$

$$= \int_{y=0}^{5} \frac{1}{210}\left(36+6y-9-3y\right)dy$$

$$= \int_{y=0}^{5} \frac{1}{210}\left(27+3y\right)dy$$

$$= \int_{y=0}^{5} \frac{1}{70}\left(y+9\right)dy$$

$$= \frac{1}{70}\left(\frac{y^2}{2}+9y\right)_0^5$$

$$= \frac{1}{70}\left(\frac{25}{2}+45\right)$$

$$= \frac{23}{28}.$$

$$P((x+y)<4)$$

$$= \int_{y=0}^{2}\int_{x=2}^{4-y} \frac{1}{210}(2x+y)\,dx\,dy$$

$$= \int_{y=0}^{2} \frac{1}{210}\left(x^2+xy\right)_2^{4-y}\,dy$$

$$= \int_{y=0}^{2} \frac{1}{210}\left((4-y)^2+y(4-y)-4-2y\right)dy$$

$$= \int_{y=0}^{2} \frac{1}{210}\left(12-6y\right)dy$$

$$= \frac{6}{210}\left(2y-\frac{y^2}{2}\right)_0^2 = \frac{6}{210}(4-2)$$

$$= \frac{12}{210} = \frac{2}{35}.$$

NOTE:

The random variables (X, Y) are said to be mutually independent if $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

(*) show that the Joint probability function

$$f_{XY}(x, y) = \begin{cases} 9\, e^{-3x}\, e^{-3y} & x, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

is a valid density fn. Also check whether the random variables are mutually independent or not.

sol:

$f(x, y)$ is Joint probability density fn. if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

$$\int_{x=0}^{\infty} \int_{y=0}^{\infty} 9\, e^{-3x}\, e^{-3y}\, dy\, dx$$

$$= 9 \int_{x=0}^{\infty} e^{-3x} \left[\frac{e^{-3y}}{-3} \right]_0^{\infty} dx = 9 \int_{x=0}^{\infty} e^{-3x} \left[\frac{1}{3} \right] dx$$

$$= 3 \left(\frac{e^{-3x}}{-3} \right)_0^{\infty}$$

$$= 3\left(\frac{1}{3}\right) = 1$$

$\therefore$ Valid density function.

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$= \int_{0}^{\infty} 9\, e^{-3x}\, e^{-3y}\, dy$$

$$= 9\, e^{-3x} \left[\frac{e^{-3y}}{-3} \right]_0^{\infty}$$

$$= 3\, e^{-3x}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

$$= \int_{0}^{\infty} 9\, e^{-3x}\, e^{-3y}\, dx$$

$$= 9\, e^{-3y} \left[\frac{e^{-3x}}{-3} \right]_0^{\infty}$$

$$= 3\, e^{-3x}$$

$$f_X(x) \cdot f_Y(y) = 9 e^{-3x} e^{-3y}$$ (22)

$$= f_{XY}(x, y)$$

$\therefore$ X, Y are mutually independent.

SAMPLING DISTRIBUTION

→ The outcome of a statistical experiment may be recorded as numerical value or discriptive

→ But the statistician is primarily consult with the analysis of numerical data.

→ When a pair of dice are toised and the sum of the numbers on the faces of dices is the outcome and it is recorded as numerical value.

→ If the students of certain school are given to blood test then the representation is like discriptive.

eg:- A+ve, B+ve, B-ve

Therefore the classification of blood types may be convenient to use the numbers like 1 - A+ 2 - A-

→ The no of observations may be finite or infinite.

eg:-1) Blood types of students in a particular school is finite observations.

2) If we toss a pair of dices indefinitely and recorded the sum which is occur then we get a infinite set of values, the each set values represent in the result of a single toss of paired dice.

POPULATION:-

Population consists of the totality of observations with which are concerned

Population is a collection of (group of) objects

→ The no. of observations in a population is defined to be size of the population and it is denoted by N.

→ The population size 'N' may be finiter or infinite

eg:- The students in a college is finite population

2) production of electric bulbs from a factory is infinite population.

SAMPLE:

A sample is a subset of a population, the no. of observations in a sample is called as size of the sample and it is denoted by 'n'.

NOTE:-

1) When information is collected in respect of every individual item, then the enquiry is said to be done by complete enumeration or senses. But this method involves more administration, expenses, time, etc.,

2) The data collected from population and examine then the process is called as sampling survey, the results are generalised and made applicable to whole field this is known as sampling.

3) The population is a universal set for the sample.

4) The statistical constants like mean, standard deviation, correlation coefficient etc., for the population is called as parameter

5) The statistical constants like mean, standard deviation, variance for the sample is called as sample statistics.

DIFFERENT METHODS OF SAMPLING:

There are two types of sampling distribution majorly. That is

1) probability sampling distribution

2) Non - probability sampling distribution.

1) Probability Sampling Distribution:

i) Random sampling (or) probability sampling:

It is a process of drawing a sample from a population in such a way that each member of the population has an equal chance of being included in the sample. This process is known as random sampling.

eg:- 1) selecting 20 words randomly from a

NOTE:

If each element of a population may be selected more than once then it is called as sampling with replacements, if not it is sampling without replacement.

If N is the size of a population and sample is the size of sample then

a) The no. of samples with replacement is equals to N^n

b) The no. of samples without replacements $= {}^{N}C_{n}$

ii) Stratified sampling (or) stratified random sampling:

In this type of sampling the population is divided into several parts (or groups) according to some relevant characteristics and each sub grouped is called as sub-population or strata, then a small sample (sub-sample) is selected from each strata at random, all sub samples combine together to form the stratified sampling which represents the population properly. This process of obtaining or examining is called as stratified sampling.

eg:- Blood In a country, to check which political party is good we use stratified sampling.

iii) Quasi Random Sampling (or) Systematic sampling:

It means forming the samples in some systematic manner by taking items at regular intervals.

→ In this method first all of the units of population arranged in some order then from the first 'k' items selected at random → This unit and every kth unit of serially listed population combine together constitute a systematic sampling.

eg:- Arranging 100 students in a class by taking 10th interval

2) Non-Probability Sampling Distribution:-

i) Purposive sampling (or) judgement sampling:

When the choice of the individual items of a sample entirely depends on the individual judgement of the investegator (sampler). It is called a purposive sampling.

For example, if a sample of 20 students is to be selected from a class of 100 students for the purpose of extra curricular activities.

ii) Sequential Sampling:

It consists of a sequence of a sample drawn one after another from the population depending on the results of previous samples, if the result of 1st sample leads to a decision which is not acceptable, the lot from which the sample was drawn is rejected but if the result of the sample is accepted a new sample is drawn but the 1st sample leads to not a clear decision then the 2nd sample is drawn. If second sample is giving acceptance then the second sample

drawn from the population is accepted.

CLASSIFICATION:

Samples are classified into 2 ways

i) Large sample:- If the size of a sample is $n \geq 30$ then the sample is said to be large sample.

ii) Samp Small sample:- If the size of a sample $n < 30$ then the sample is said to be small small or exact sample.

NOTE:

→ The sampling from a finite population with replacement can be considered theoritically as sampling from infinite population (An element of the population can be choosed at more than once)

→ Sampling from a finite population without replacement can be considered theoritically as sampling from finite population (An element of the population can not be choosen more than once and it is not repeated).

Parameters and statistic:

Parameter is a statistical measure based on the all the units (all observation of a population). for example the statistical constant of the population mean, variance are referred

as parameters. statistic (or sample statistic is the statistical measure based on only, the all the units selected in a sample.

eg: this statistical measures computed from sample observations alone. sample mean ($\bar{x}$), sample variance (s^2) referred as statistics.

sample mean: If $x_1, x_2, \ldots x_n$ represents a random sample of size 'n' then the sample mean is defined as statistic i.e, $\bar{x} = \sum_{i=1}^{n} x_i/n$

Sample variance: If $x_1, x_2 \ldots x_n$ represents a random sample of size 'n' then the sample variance defined by standard the statistic is $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

standard deviation of a sample

$$s = \sqrt{\text{i.e. } \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Sampling Distribution (of a statistic):

The probability distribution of a statistic is called a sampling distribution

If we draw a sample of size 'n' from a given population (finite) of size 'N' then the total no. of possible samples is ${}^{N}C_{n}$. i.e.,

$${}^{N}C_{n} = \frac{N!}{n!(N-n)!}$$

Standard error (S.E) of a statistic:

The standard deviation of a sampling distribution of a statistic is known as standard error of a statistic and it is denoted by S.E

1) The standard error of sample mean is

$$\bar{x} = \frac{\sigma}{\sqrt{n}}$$

→ standard error of sample proportion is

$$p = \sqrt{\frac{pq}{n}}$$

→ standard error of sample standard deviation

$$= \frac{\sigma}{\sqrt{2n}}$$

SAMPLING DISTRIBUTION OF MEAN (σ known):
The probability distribution of $\bar{x}$ is called as the sampling distribution of means. The sampling distribution of a statistic depends on the size of the population, the size of the samples and the method of choosing samples.

Let $x_1, x_2, x_3 \ldots x_n$ be the 'n' random samples drawn from a population of size 'N' with mean 'μ' and variance σ^2 and $\bar{x}$ is the mean of the sample.

Infinite population:
Suppose the samples are drawn from infinite population or sampling is done with replacement then

→ mean $\mu_{\bar{x}} = \mu$

→ variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

→ standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

NOTE:
→ provided sample size is large

→ The no. of samples drawn from the infinite population of size 'N' with replacement then the total no. of possible samples of n is N^n

→ Finite Population:
Consider a finite population of size 'N' with mean 'μ' and standard deviation σ, draw the all possible samples of size 'n' without replacement then

→ The mean of the sampling distribution of means ($N > n$) i.e., $\mu_{\bar{x}} = \mu$

→ The variance is $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

→ standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$

NOTE:
Here $\frac{N-n}{N-1}$ is called as the finite population correction factor.

SAMPLING DISTRIBUTION OF PROPORTION
Let p be the probability of occurrence of an event (success) and $q = 1-p$ is the probability of non-occurrence (failure). Then draw all the possible samples of 'n' from an infinite population, compute the proportion of 'p' of success for each of these samples. Then the mean μ_p and variance σ_p^2 of sampling distribution of proportion are given by

$$\mu_p = p \qquad \sigma_p^2 = \frac{pq}{n} = \frac{p(1-p)}{n}$$

For finite population (without replacement) of size 'N' then we have $\mu_p = p$

$$\sigma_p^2 = \frac{pq}{n}\left[\frac{N-n}{N-1}\right]$$

SAMPLING DISTRIBUTION OF DIFFERENCES AND SUMS :

Let μ_{s_1} and σ_{s_1} be the mean and standard deviation of sampling distribution of statistic s_1 obtained by computing s_1 for all possible statistics of size 'n_1' drawn from population A. Let μ_{s_2} and σ_{s_2} be the mean and standard deviation of sampling distribution of statistic s_2 obtained by computing s_2 for all possible samples of size 'n_2' drawn from the another population B.

Then now compute the statistic $s_1 - s_2$, the difference of the statistic from all the possible combinations of these samples from the 2 populations A & B.

Then

→ The mean $\mu_{s_1 - s_2} = \mu_{s_1} - \mu_{s_2}$

→ $\sigma_{s_1 \pm s_2} = \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}$

NOTE:

Assuming that samples are independent sampling distribution of sum of statistics as given by $\mu_{s_1 + s_2} = \mu_{s_1} + \mu_{s_2}$

$$\sigma_{s_1 + s_2} = \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}$$

→ For example, for infinite population the sampling distribution of sum of means as $\mu_{\bar{x}_1 + \bar{x}_2}$ and $\sigma_{\bar{x}_1 + \bar{x}_2}$ given by

$$\mu_{\bar{x}_1 + \bar{x}_2} = \mu_{\bar{x}_1} + \mu_{\bar{x}_2} = \mu_1 + \mu_2$$

and $$\sigma_{\bar{x}_1 \pm \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

→ For sample distribution of differences of proportion we have

$$\mu_{p_1 - p_2} = \mu_{p_1} - \mu_{p_2} = p_1 - p_2$$

$$\sigma_{p_1 - p_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

1. Find the value of finite population correction factor for $n = 10$, $N = 100$

Sol:- Correction factor $= \left(\frac{N-n}{N-1}\right)$

The population size $N = 100$ and sample size $n = 10$.

$$= \frac{100-10}{100-1} = \frac{90}{99} = 0.909$$

2. A population consists of 5 numbers 2, 3, 6, 8, 11. Consider all possible samples of size 2 which can be drawn

i) With replacement

ii) Without "

From this population find a) mean of the population b) standard deviation of population c) the mean of sampling distribution of means d) standard deviation of sampling distribution of means

Sol:- The population size $N=5$ sample size $n=2$

i) With replacement:

No. of samples $N^n = 5^2 = 25$

a) Mean of population $(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\mu = \frac{2+3+6+8+11}{5} = 6$$

b) S.D of population $(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$

$$\sigma = \sqrt{\frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}}$$

$$= \sqrt{\frac{16+9+4+25}{5}} = \sqrt{\frac{54}{5}} = \sqrt{10.8}$$

$$= 3.2863$$

c) Sample mean $(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$

Samples = {(2,3) (2,6) (2,8) (2,11) (2,2)
(3,2) (3,6) (3,8) (3,11) (3,3)
(6,2) (6,6) (6,3) (6,8) (6,11)
(8,2) (8,3) (8,6) (8,8) (8,11)
(11,2) (11,3) (11,6) (11,8) (11,11)}

mean = {2.5 2.5 4 5 6.5 2
2.5 3 4.5 5.5 7 8.5 3
4 4.5 6 7 8.5
5 5.5 7 8 9.5
6.5 7 8.5 9.5 11}

= 1 1 2 1 2 2 2 2 1 2 4

2 2.5 3 4 4.5 5 5.5 6 6.5 7

$$= \frac{2+2.5+4+5+6.5+\dots+11}{25}$$

$$= \frac{150}{25} = 6$$

d) Sample S.D $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$

$$= \sqrt{\frac{(2-6)^2+(2.5-6)^2+\dots+(11-6)^2}{25}}$$

$$= \sqrt{\begin{matrix}16+12.25+4+1+0.25+0.25+2.25+0.25\\+1+9+4+0+2.25+1+6.25+1+0.25\\+1+4+12.25+0.25+1+6.25+12.25\end{matrix}}{25}$$

$$= \sqrt{\frac{123}{25}} = \sqrt{4.92} = 2.218$$

ii) Without replacement:

$$N_C{}_n = {}^5C_2 = 10$$

a) Mean of population $(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$$

b) S.D of population $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$

$$= \sqrt{\frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5}}$$

$$\sigma = \sqrt{10.8} = 3.286$$

c) Sample mean $(\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$

Samples = { (2,3) (2,6) (2,8) (2,11)
(3,6) (3,8) (3,11)
(6,8) (6,11) (8,11) }

means of samples = { 2.5 4 5 6.5
4.5 5.5 7
7 8.5 9.5 }

Mean of the sample

$$\bar{x} = \frac{2.5 + 4 + 5 + 6.5 + 4.5 + 5.5 + 7 + 8.5 + 9.5 + 7}{10}$$

$$= 6$$

d) Sample S.D $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$

$$s = \sqrt{\frac{(2.5-6)^2 + (4-6)^2 + \cdots + (9.5-6)^2}{10}}$$

$$s = 2.012$$

3. A population consists of 5, 10, 14, 18, 13, 24. consider all possible samples of size 2 i) with replacement ii) Without replacement from the combination and find a) the mean of population b) the standard deviation of population c) the mean of sampling distribution of means d) the standard deviation of sampling distribution of mean

Sol:- population size $N = 6$ sample size $= 2$

i) with replacement:

$$N^n = 6^2 = 36$$

a) Mean of population $\mu = \sum_{i=1}^{n} \frac{x_i}{n}$

$$\mu = \frac{5 + 10 + 14 + 18 + 13 + 24}{6}$$

$$= \frac{84}{6} = 14$$

b) S.D of population $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$

$$\sigma = \sqrt{\frac{(5-14)^2 + (10-14)^2 + (14-14)^2 + (18-14)^2 + (13-14)^2 + (24-14)^2}{6}}$$

$$\sigma = 5.972$$

c) sample mean $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$

samples = { (5,5) (5,10) (5,14) (5,18) (5,13) (5,24)
(10,5) (10,10) (10,14) (10,18) (10,13) (10,24)
(14,5) (14,10) (14,14) (14,18) (14,13) (14,24)
(18,5) (18,10) (18,14) (18,18) (18,13) (18,24)
(13,5) (13,10) (13,14) (13,18) (13,13) (13,24)
(24,5) (24,10) (24,14) (24,18) (24,13) (24,24) }

Mean = $\begin{Bmatrix} 5 & 7.5 & 9.5 & 11.5 & 9 & 14.5 \\ 7.5 & 10 & 12 & 14 & 4.5 & 17 \\ 9.5 & 12 & 14 & 16 & 13.5 & 19 \\ 11.5 & 14 & 16 & 18 & 15.5 & 21 \\ 9 & 11.5 & 18.5 & 15.5 & 13 & 18.5 \\ 14.5 & 17 & 19 & 21 & 18.5 & 24 \end{Bmatrix}$

$= \frac{504}{36} = 14$

d) Sample $S.D = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

$= \sqrt{\frac{(5-14)^2 + (7.5-14)^2 + (9.5-14)^2 + \cdots (18.5-14)^2 + (24-14)^2}{36}}$

$= 4.2229$

ii) Without replacement

$N_C{}_n = {}_6C_2 = 15$

a) Mean of population $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$

$\mu = \frac{5+10+14+18+13+24}{6}$

$= \frac{84}{6} = 14$

b) S.D of population $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$

$\sigma = \sqrt{\frac{(5-14)^2 + (10-14)^2 + \cdots (13-14)^2 + (24-14)^2}{6}}$

$\sigma = 5.972$

c) Sample mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Samples = $\{(5,10)(5,14)(5,18)(5,13)(5,24)$
$(10,14)(10,18)(10,13)(10,24)$
$(14,18)(14,13)(14,24)$
$(18,13)(18,24)$
$(13,24)$

Mean = $\begin{Bmatrix} 7.5 & 9.5 & 11.5 & 9 & 14.5 \\ 12 & 14 & 11.5 & 17 \\ 16 & 13.5 & 19 \\ 15.5 & 21 \\ 18.5 \end{Bmatrix}$

$= \frac{210}{15} = 14$

d) Sample of $S.D = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

$= \sqrt{\frac{(7.5-14)^2 + (9.5-14)^2 + \cdots (18.5-14)^2}{15}}$

$= 3.7771$

EXPONENTIAL DISTRIBUTION

By,

Vinay Kumar V

Exponential Distributions

A continuous random variable X is said to be follow an Exponential distribution if its probability density function defined by

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x\geq 0,\lambda\geq 0\\ 0 & otherwise\end{cases}$$

For exponential distribution λ is called rate parameter.

$f(x)$ is said to be probability density function if

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\,dx$$

$$= \lambda \int_{0}^{\infty} e^{-\lambda x}\,dx = \lambda \left(\frac{e^{-\lambda x}}{-\lambda} \right)_{0}^{\infty}$$

$$= -1(0-1)$$

$$= 1$$

$$\text{Mean} = \mu = E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$= \int_{0}^{\infty} \lambda x e^{-\lambda x}\, dx = \lambda \int_{0}^{\infty} x e^{-\lambda x}\, dx$$

$$= \lambda \left(x \frac{e^{-\lambda x}}{-\lambda} - 1. \frac{e^{-\lambda x}}{-\lambda} \right)_{0}^{\infty}$$

$$= \left(-\left(-\frac{1}{\lambda^2} \right) \right)$$

$$= \frac{1}{\lambda}$$

$$\therefore \text{mean} = \frac{1}{\lambda}$$

λ is the inverse of the Expected duration (μ)

Variance = $E(x^2) - \left(E(x)\right)^2$

$$E(x^2) = \int_0^\infty \lambda x^2 e^{-\lambda x}\, dx$$

$$= \lambda \int_0^\infty x^2 e^{-\lambda x}\, dx$$

$$= \lambda \left(x^2 \frac{e^{-\lambda x}}{-\lambda} - 2x \frac{e^{-\lambda x}}{\lambda^2} + 2 \frac{e^{-\lambda x}}{-\lambda^3} \right)_0^\infty$$

$$= \lambda \left(-\frac{2}{-\lambda^3} \right) = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2}$$

$\therefore$ Variance = $= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2$

$$= \frac{1}{\lambda^2}$$

- If the expected duration is 5 minutes then the rate parameter value λ is 0.2.
- The cumulative distribution function can be written as the probability of the lifetime being less than some value x.
- $P(X \leq x) = 1 - e^{-\lambda x}$
- $E(X) = 1/\lambda$
- $Var(X) = 1/\lambda^2$

Problem 1

Assume that the length of a phone call in minutes is an exponential random variable X with parameter $\lambda = 1/10$. If someone arrives at a phone booth just before you arrive, find the probability that you will have to wait

a) Less than 5 minutes
b) Greater than 10 minutes
c) Between 5 and 10 munutes
d) Compute the Expected value and Variance.

Sol: a) $$P(X \le 5) = 1 - e^{-5*1/10} = 1 - e^{-\frac{1}{2}}$$

$$= 1 - 0.6066 = 0.3934$$

b) $P(X \geq 10) = 1 - (P(X < 10))$

$= 1 - P\left(1 - e^{-10 * \frac{1}{10}}\right)$

$= e^{-1} = 0.3678$

c) $P(5 \leq X \leq 10) = 1 - (P(X \leq 5) + P(X \geq 10))$

$\therefore P(5 \leq X \leq 10) = 1 - (0.3934 + 0.3678)$

$= 1 - 0.7612 = 0.2388$

$Mean(\mu) = E(X) = 1/\lambda$

$= \frac{1}{1/10} = 10$

$Variance(\sigma^2) = 1/\lambda^2 = 100$

Problem 2:

If X is an exponential variable with mean 5 .Evaluate

1) $P(0 \le X \le 1)$
2) $P(-\infty < X \le 10)$

Sol: We have Exponential distribution is $f(x)=\lambda e^{-\lambda x}; 0 \le x < \infty$

$$Mean(\mu)=E(X)=1/\lambda = 5 \Rightarrow \lambda = 1/5 = 0.2$$

$$\therefore f(x)=\frac{1}{5}e^{-\frac{x}{5}}; 0 < x < \infty$$

1) $P(0 \le X \le 1)=\int_0^1 f(x)dx$

$$=\int_0^1 \frac{1}{5}e^{-\frac{x}{5}}dx$$

$$=\frac{1}{5}\left(\frac{e^{-\frac{x}{5}}}{-\frac{1}{5}}\right)_0^1 = 1-e^{-0.2} = 0.1813$$

2) $P(-\infty < X \le 10)=\int_{-\infty}^0 f(x)dx+\int_0^{10} f(x)dx$

$$=0+\int_0^{10}\frac{1}{5}e^{-\frac{x}{5}}dx = 0.8647$$

Gamma Distribution:-

Add a shape parameter to the Exponential distribution.

(Recall the definition of the gamma function)

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx ; \quad k>0$$

making the substitution $y = x/\lambda$ where λ is a positive real constant the gamma function becomes

$$\Gamma(k) = \int_0^\infty (\lambda y)^{k-1} e^{-\lambda y} \lambda dy , \quad k>0$$

divide both sides by $\Gamma(k)$ yielding

$$1 = \int_0^\infty \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma(k)} dy$$

Def'- A continuous Random Variable X with Probability density function (pdf)

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} ; \quad x>0$$

for some real constants $\lambda>0$ and $k>0$ is a gamma (λ, k) random variable.

The cdf for $X \sim \text{gamma}(\lambda, k)$ is

$$F(x) = \int_0^x \frac{\lambda^k w^{k-1} e^{-\lambda w}}{\Gamma(k)} dw$$

$$= \frac{\lambda^k}{\Gamma(k)} \int_0^x w^{k-1} e^{-\lambda w} dw$$

$$= \frac{\lambda^k}{\Gamma(k)} \int_0^{\lambda x} \left(\frac{y}{\lambda}\right)^{k-1} e^{-y} \cdot \frac{1}{\lambda} dy$$

$$= \frac{1}{\Gamma k} \int_0^{\lambda x} y^{k-1} e^{-y} dy$$

Gamma Distribution:-

The probability density function

PDF : $f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$

where $x = 0, 1, \ldots \infty$
$\alpha > 0$
$\beta > 0$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$

Note:- $\Gamma(\alpha+1) = \alpha \Gamma(\alpha)$

μ = Mean $(E(x))$:- $\int_0^\infty x f(x)\,dx$

$$\int_0^\infty x \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}\,dx$$

$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-x/\beta}\,dx$$

Let $t = x/\beta$; $x = \beta t$; $dx = \beta\,dt$

$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty (\beta t)^\alpha e^{-t}\,\beta\,dt$$

$$= \frac{1}{\Gamma(\alpha)\beta^\alpha}\,\beta \int_0^\infty t^\alpha e^{-t}\,dt$$

Let $\alpha = \delta - 1$, $\delta = \alpha + 1$

$$= \frac{\beta}{\Gamma(\alpha)} \int_0^\infty t^{\delta-1} e^{-t}\,dt$$

$$\text{Mean} = \frac{\beta\,\Gamma(\alpha+1)}{\Gamma(\alpha)}$$

$$\text{Mean} = \beta\alpha$$

$$\boxed{\text{Mean} = \alpha\beta}$$

Variance :-

$$\sigma^2 = E(x^2) - (E(x))^2$$

$$E(x^2) = \int_0^\infty x^2 \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx$$

$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha+1} e^{-x/\beta} dx$$

Let $t = x/\beta$; $x = \beta t$; $dx = \beta dt$

$$E(x^2) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int (\beta t)^{\alpha+1} e^{-t} \beta dt$$

$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \beta^{\alpha+1} \beta \int_0^\infty t^{\alpha+1} e^{-t} dt$$

$$= \frac{\beta^2}{\Gamma(\alpha)} \int t^{\alpha+1} e^{-t} dt$$

$$= \frac{\beta^2}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}$$

$$= \frac{\beta^2}{\Gamma(\alpha)} = \beta^2 \frac{(\alpha+1)\alpha\Gamma(\alpha)}{\Gamma(\alpha)}$$

$$= \beta^2(\alpha+1)\alpha$$

$$Var(x) = E(x^2) - (E(x))^2$$

$$= \beta^2\alpha(\alpha+1) - \alpha^2\beta^2 = \boxed{Var(x) = \alpha\beta^2}$$

we know that,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

$$\therefore \Gamma(\alpha+2) = \int_0^\infty t^{\alpha+2-1} e^{-t} dt$$

$$\Gamma(\alpha+2) = \frac{= \int_0^\infty t^{\alpha+1} e^{-t} dt}{\text{eq}}$$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

$$\Gamma(\alpha+2) = (\alpha+1)\Gamma(\alpha+1)$$

$$\boxed{\Gamma(\alpha+2) = (\alpha+1)\alpha\Gamma(\alpha)}$$

Example:-

The daily consumption of milk in a city in excess of 30,000 gallons is approximately distributed as gamma variate with parameters $\alpha = 2$ & $\lambda = \frac{1}{20,000}$. The city has daily stock of 40,000 gallons. what is the probability of stock is not sufficient on a particular day?

Sol:-

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$$

$$= \frac{\left(\frac{1}{20000}\right)^2 x^1 e^{\frac{-x}{20,000}}}{\Gamma(2)}$$

$$= \left(\frac{1}{20000}\right)^2 x\, e^{-x/20,000}$$

P(Stock is insufficient) =

P(daily consumption > daily stock)

$= P(x > 30,000)$

$= P(x + 30,000 > 40,000)$

$= P(x > 40000 - 30000)$

$$= P(x > 10,000)$$

$$= \int_{10,000}^{\infty} \left(\frac{1}{20000}\right)^2 x e^{-x/20000}$$

$$= \left(\frac{1}{20,000}\right)^2 \left[x \left(\frac{e^{-x/20,000}}{-1/20,000} \right) - 1 \frac{e^{-\frac{x}{20000}}}{\left(-\frac{1}{20,000}\right)^2} \right]_{10000}^{\infty}$$

$$= \left(\frac{1}{20,000}\right)^2 \left[0 - \frac{e^{-\frac{10,000}{20000}}}{\left(-\frac{1}{20000}\right)^2} \right]$$

$$= e^{-1/2} = \underline{0.6065}$$

Ex-2

Engineers designing the next generation of space shuttles plan to include two fuel pumps - one active, the other in reserve. If the primary pump malfunctions, the second is automatically brought on line. Suppose a typical mission is expected to require that fuel be pumped for at most 50 hrs. According to the manufacturer's specifications, pumps are expected

to fail once every 100 hours. what are the chances that such a fuel pump system would not remain functioning for the full 50 hours?

Sol:-

We are given that λ, the average number of failures in 100 hours interval is 1

i.e. θ the mean waiting time until the first failure is $1/\lambda$ or 100 hours

Let x denote the time elapsed until the α=2nd pump breaks down.

Assuming the failures follow a poisson process, the probability density function of x is

$$f_x(x) = \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$$

(Alternative definition taking α instead of 'k')

given, $\alpha = 2$

$\lambda = 1/100$

$$f(x) = \frac{1}{100^2 \Gamma(2)} \cdot e^{-x/100} x^{2-1}$$

$$= \frac{1}{10000} x e^{-x/100}$$

for $y>0$, Therefore the probability that the system fails to last 50 hours is

$$P(x<50) = \int_0^{50} \frac{1}{10000} x e^{-x/100}\, dx$$

Let $u = x$ and $dv = e^{-x/100}$

then $du = dx$ and $v = -100 e^{-x/100}$

$$P(x<50) = \frac{1}{10000}[uv - \int v\,du]$$

$$= \frac{1}{10000}\left\{\left[-100x e^{-x/100}\right]_0^{50} - \int (-100 e^{-x/100} dx)\right\}$$

$$= \frac{1}{10000}\left\{(-5000 e^{-1/2}) + 100\left[(-100) e^{-x/100}\right]_0^{50}\right\}$$

$$= \frac{1}{10000}\left\{-5000 e^{-1/2} - 10000(e^{-1/2} - 1)\right\}$$

$$= -\frac{1}{2} e^{-1/2} - e^{-1/2} + 1$$

$$= 1 - 3/2\, e^{-1/2} = 0.09$$

TESTING OF HYPOTHESIS

Introduction:

The main problems in statistical problems can be classified into 2 areas

1) The area of estimation of population parameter and
2) setting up of confidence interval for them i.e; the area of point estimation and interval estimation
2) Test of statistical hypothesis

In the previous chapter (estimation) we have seen how a parameter can be estimated from sample data. To decide whether a statement of parameter is true or false of estimated value of the parameter we test by using test of hypothesis.

The hypothesis being tested as 'H' there are two possibilites H is true or false. If

i) H is true but it is rejected then it is type I error (α error)

ii) If H is false but accepted, it is type II error (β error).

There are two types of hypothesis.

i) Null hypothesis ii) Alternative hypothesis

Null Hypothesis:

For applying the test of significance we first setup a hypothesis which is a statement about population parameter such a hypothesis is usually a hypothesis of no differences is called null hypothesis.

In otherwords a null hypothesis is a hypothesis which is tested for possible rejection under the assumption i.e., true, and it is denoted by H_0.

Alternative Hypothesis:

Any hypothesis which is complementary to the null hypothesis is called as an alternative hypothesis and denoted by H_1.

For example, if we want to test a null hypothesis that the population has a specified mean μ_0 (say) i.e., null hypothesis $H_0 : \mu = \mu_0$

Then the alternative hypothesis be

i) $H_1 : \mu \neq \mu_0$ [i.e.,

ii) $H_1 : \mu > \mu_0$

iii) $H_1 : \mu < \mu_0$

The alternative hypothesis (i) is known as two tailed test, the alternative hypothesis (ii) is known as right tailed test & the alternative hypothesis (iii) is said to be left tailed test

NOTE:

→ The setting of an alternative hypothesis is very important to decide whether we have to use single tailed test (left or right) or two tailed test.

→ The sizes of type I error also called as producers risk, & the size of type II error is called as consumer risk.

Critical region:-

A region corresponding to a statistic in the sample space 's' which leads to the rejection of H_0 is called as the critical region or rejection region, and those region which leads to acceptance of H_0 is called acceptance region.

Level of Significance:-

The probability α that a random value of the statistic 't' belongings to the critical region is known as the level of significance. In otherwords level of significance is the size of type I error.

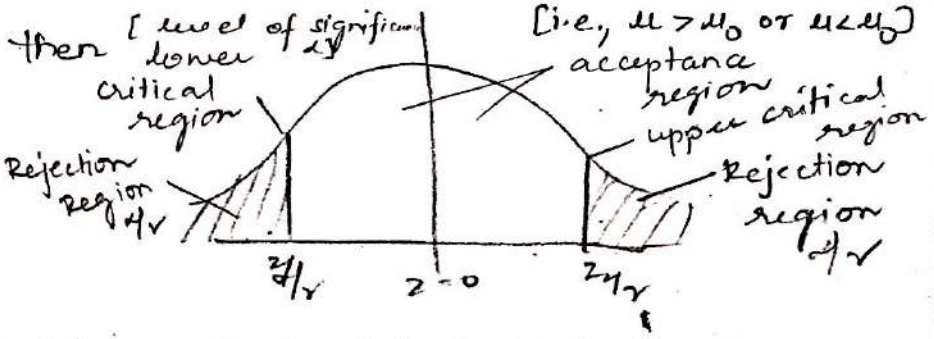
In any test the critical region represented by a portion or partion of a area under the normal curve (probability curve) of the sampling distribution of the test statistic.

In two tailed test, consider the area

of both tails of the curve represented by sampling distribution

Null Hypothesis $H_0 : \mu = \mu_0$

Alternative " $H_1 : \mu \neq \mu_0$ [i.e., $\mu > \mu_0$ or $\mu < \mu_0$]

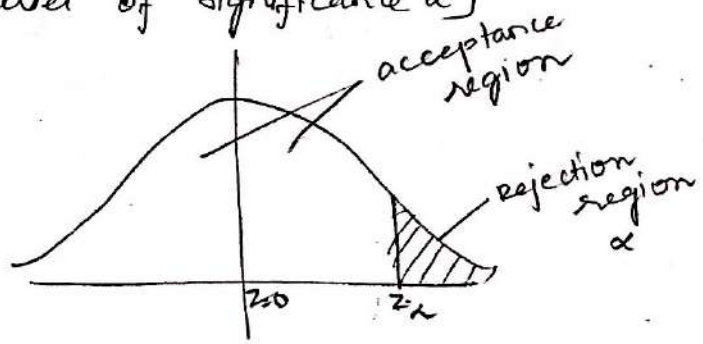


In a single tailed test, the area on the right hand side of an ordinate (say $x = x_0$) is taken into account then

i) Null hypothesis $H_0 : \mu = \mu_0$

ii) Alternative " $H_1 : \mu > \mu_0$

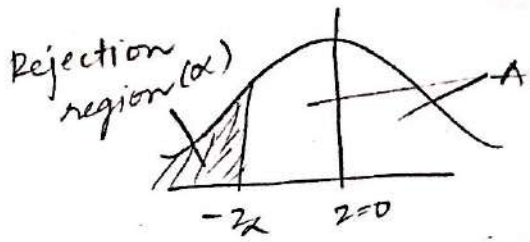
[level of significance α]



ii) Null hypothesis $H_0 : \mu = \mu_0$

Alternative " $H_1 : \mu < \mu_0$ [left tailed test]

Level of significance α



Critical values of z (z_α):

Critical values of z_α	level of significance		
	1% 0.01	5% 0.05	10% 0.1
Two tailed test	$z_{\alpha/2} = 2.58$	$z_{\alpha/2} = 1.96$	$z_{\alpha/2} = 1.645$
Right tailed test	$z_\alpha = 2.33$	$z_\alpha = 1.6455$	$z_\alpha = 1.28$
left tailed test	$z_\alpha = -2.33$	$-z_\alpha = -1.645$	$-z_\alpha = -1.28$

NOTE:

If 'n' is small ($n \leq 30$) then the sampling distribution of the test statistic 'z' will not be normal. That time we use t-test, F-test, χ^2-test

PROCEDURE FOR TESTING OF HYPOTHESIS:

Step 1: Null Hypothesis: set up a null hypothesis H_0

Step 2: Alternative Hypothesis: set up the alternative " H_1 this will help us to decide whether we have to use single tailed test (left or right) or two tailed ".

step 3: level of significance : choose the appropriate level of significance 'α'.

step 4: Calculate the test statistic z under the null hypothesis i.e., $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

step 5: Conclusion:
compare the calculated 'z' value with the tabulated value at the given level of significance α.

→ If $|z| < z_\alpha$ if we accept null hypothesis H_0

→ If $|z| > z_\alpha$ we reject null hypothesis H_0

1. A dice is tossed 960 times and it falls with 5 upwards 184 times, is the die unbiased at a level of significance α = 0.01

Sol:- n = 960

p = probability of getting 5 = 1/6

q = 1 − p = 5/6

$\mu = np = 960 \times \frac{1}{6} = 160$

$\sigma = \sqrt{npq} = \sqrt{960 \times \frac{1}{6} \times \frac{5}{6}} = 11.54$

x is no. of success = 184

Null Hypothesis H_0: die is unbiased ($\mu = \mu_0$)

Alternative " H_1: die is biased ($\mu \neq \mu_0$)

level of significance α = 0.01

the test $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{184-160}{11.54} = 2.079$

$|z| = 2.07$

tabulated value z = 2.58

$|z| < z$

∴ the null hypothesis is accepted i.e., die is biased

2. A die is tossed 256 times and it turns up with even digit 150 times, is the dice biased, at 1% level

Sol:- n = 256

p = even digit (2, 4, 6) = 3/6 = 1/2

q = 1 − p = 1/2

$\mu = np = 256 \times \frac{1}{2} = 128$

$\sigma = \sqrt{npq} = \sqrt{144 \times \frac{1}{2}} = 8$

x = 150 times

H_0: die is biased

H_1: die is unbiased

level of significance α = 1.96

the test statistic $z = \frac{x-\mu}{\sigma} = \frac{150-128}{8} = 2.75$

$|z| = 2.75$

tabulated value z = 1.96

$|z| > z$

$\therefore$ the null hypothesis is rejected, i.e, die is unbiased.

(iii) Test of significance for large samples

under the last large sample test we will see four important tests to test the significance.

i) Test of significance for single mean
ii) " " " " " differences of means
iii) " " " single proportion
iv) " " " differences of proportions

i) Test of significance for single mean

Suppose we want to test whether the given sample of size 'n' has been drawn from a population with mean μ. We will setup null hypothesis that there is no difference b/w $\bar{x}$ & μ. when $\bar{x}$ is the mean of the sample, then the test statistic $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.

If σ is unknown then we use $z = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

NOTE
The values $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ are called the 95% confidence limits for the mean of the population corresponding to the given sample

1. According to the norms established for a mechanical aptitude test 18th persons who are 18 years old have an average height 73.2 with a standard deviation of 8.6. If 4 randomly selected persons of the average 76.7, test the hypothesis $\mu = 73.2$ against the alternative hypothesis $\mu > 73.2$ at the 0.01 level of significance.

Sol: Null hypothesis H_0 : $\mu = 73.2$
Alternative " H_1 : $\mu > 73.2$ (Right tailed)

$\sigma = 8.6$ $\bar{x} = 76.7$ $n = 4$
level of significance $\alpha = 0.01$
The test statistic $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{76.7-73.2}{8.6/\sqrt{4}}$

$= 0.813$

$z_{cal} = 0.813$ $z_{tab} = z_{0.01} = 2.33$

$\therefore |z_{cal}| < z_{tab}$

$\therefore$ Null hypothesis is accepted i.e; $\mu = 73.2$

2. An Ambulance service claims that it takes an average time less than 10 mins to

reach the destination in emergency calls. A sample of 36 calls as a mean of 11 mins and the variance of 16 mins test the claim at 0.05 level of significance

Sol:-

$H_0 : \mu = 10$

$H_1 : \mu < 10$ (left tailed)

$n = 36 \quad \sigma = 4 \quad \bar{x} = 11 \quad \mu = 10$

level of significance $\alpha = 0.05$

test statistic $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{11-10}{4/\sqrt{36}} = 1.5$

$z_{cal} = 1.5$

$z_{tab} = z_{at\,0.05} = 1.645$

$\because |z_{cal}| > z_{tab}$

$1.5 > -1.645$

$\therefore$ Null hypothesis is rejected.

3. A sample of 400 items is taken from a population whose S.D is 10 the mean of the sample is 40. Test whether the sample is came for the population with mean 38 also calculate 95% confidence interval for the population.

$H_0 = \mu = 38$

$H_1 : \mu \neq 38$

$\alpha = 0.05$

$n = 400 \quad \sigma = \sqrt{100} = 10 \quad \bar{x} = 40 \quad \mu = 38$

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{40-38}{10/\sqrt{400}} = 4$

$z_{cal} = 4 \qquad z_{tab}$ at 1.96

19/05/15

In a random sample of 60 workers, the average time taken by them to get to the work is 33.8 mins with a S.D of 6.1 mins. Can we reject the null hypothesis $\mu = 32.6$ mins in favour of alternative hypothesis $\mu > 32.6$ at $\alpha = 0.025$ level of significance (Note: the tabulated value of z at 0.025 level of significance is 2.58)

Sol:- $n = 60 \quad \bar{x} = 33.8 \quad \sigma = 6.1 \quad \mu = 32.6$

$H_0 : \mu = 32.6$

$H_1 : \mu > 32.6$

Test statistic $z = \frac{33.8 - 32.6}{\frac{6.1}{\sqrt{60}}} = 1.52$

$z_{tab} = z_{0.025} = 2.58$

$z < z_{tab}$

$\therefore$ null hypothesis is accepted.

It is claimed that a random sample of 49 tyres has a mean life of 15200 kms and the sample was drawn from population whose mean is 15,150 kms & S.D of 1200 kms test the significance at 0.05 level.

Sol:- $n = 49$

$\bar{x}$ $\mu = 15200$ $\bar{x} = 15,150$ $\sigma = 1200$

H_0: $\mu = 15,150$ H_1: $\mu \neq 15,150$

Test statistic $z = \frac{15,200 - 15150}{\frac{1200}{\sqrt{49}}} = 0.291$

$z_{tab} = 1.96$

$z < z_{tab}$

null hypothesis is accepted.

Test of significance for differences of means

Let $\bar{x}_1$ be the mean of a sample size n_1 from a population with mean μ_1 and σ_1^2 (variance). Let $\bar{x}_2$ be the mean of a sample of size n_2 from a population with mean μ_2 and variance σ_2^2.

To test whether there is any significant difference between $\bar{x}_1$ & $\bar{x}_2$ we have the test statistic $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

NOTE

① If the samples have been drawn from same population then $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then the test statistic is $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$

② If σ is unknown we can estimate of

$\sigma^2 = \frac{n_1 s_1^2 - n_2 s_2^2}{n_1 + n_2}$

① The mean of 2 large sample of sizes 1000 and 2000 members are 67.5 inches and 68 inches respectively. can the samples be regarded as from the same population of S.D 2.5 inches

Sol:- $n_1 = 1000$ $n_2 = 2000$ $\bar{x}_1 = 67.5$ $\bar{x}_2 = 68$ $\sigma = 2.5$

H_0: Two samples are taken from same population ($\mu_1 = \mu_2$)

H_1: 2 samples are not taken from same population ($\mu_1 \neq \mu_2$)

$\alpha = 0.05$

Test statistic $z = \frac{67.5 - 68}{\sqrt{\frac{2.5^2}{1000} + \frac{2.5^2}{2000}}} = -5.16$

$z_{tab} = z_{\alpha = 0.05} = 1.96$

$|z| < z_{tab} \Rightarrow |-5.16| < 1.96$

$z > z_{tab}$

∴ null hypothesis is not accepted.

In a survey of buying habits 400 women shoppers are choosen at random in a super market A located in a certain section of city. Their average weekly food expenditure is Rs.259 with S.D Rs.40/-. For 400 women shoppers choosen at random in a super market B in another section of city, their average weekly food expenditure is Rs.220/- with a S.D of Rs.55/-. Test at 1% of level of Significance whether the average weekly food expenditure of shoppers are equal.

Sol:- $n_1 = 400$ $n_2 = 400$ $\bar{x}_1 = 250$ $\bar{x}_2 = 220$ $\sigma_1 = 40$ $\sigma_2 = 55$

H_0: ($\mu_1 = \mu_2$) Weekly food expenditure are equal

H_1: $\mu_1 \neq \mu_2$ Weekly food expenditure are not equal

level of significance $\alpha = 0.01$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{250 - 220}{\sqrt{\frac{40^2}{400} + \frac{55^2}{100}}}$$

$Z = 8.85$

$Z_{tab} = Z_{\alpha = 0.01} = 2.58$

$|Z| > Z_{tab}$

Null hypothesis is rejected

20/08/15

Samples of students were drawn from a two universities and from their weights in kgs, mean and s.D are calculated and shown below. Make a large sample test to test the significance of difference between the means.

	Mean	S.D	size of sample
Uni A (Varsity)	55	10	400
Uni B (Varsity)	57	15	100

Sol:- $H_0 : \bar{x}_1 = \bar{x}_2$ $\alpha = 0.05$

$H_1 : \bar{x}_1 \neq \bar{x}_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{55 - 57}{\sqrt{\frac{10^2}{400} + \frac{15^2}{100}}} = -1.26$$

$Z_{tab} = 1.96$

$|Z| < Z_{tab} \Rightarrow |-1.26| < 1.96$

$1.26 < 1.96$

∴ Null hypothesis is accepted

i.e., Avg weight of universities A and Avg weight of " B are equal.

The research investigator is interested in studying whether there is a significance difference in the salaries of MBA graduate in 2 Metropolitan cities. A random sample of size 100 from mumbai and their average income of 20,150 another random sample of size 60 from chennai and their average income 20,250. If the variances of both populations are given as $\sigma_1^2 = 40,000$ & $\sigma_2^2 = 32,400$ respectively.

Sol:- $n_1 = 100$ $n_2 = 60$ $\bar{x}_1 = 20,150$ $\bar{x}_2 = 20,250$

$\sigma_1^2 = 40,000$ $\sigma_2^2 = 32,400$

Null hypothesis $H_0 : \bar{x}_1 = \bar{x}_2$

Alternative " $H_1 : \bar{x}_1 \neq \bar{x}_2$

level of significance $\alpha = 0.05$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{20,150 - 20,250}{\sqrt{\frac{40,000}{100} + \frac{32,400}{60}}} = -3.26$$

$z_{tab} = 1.96$

$|z| < z_{tab} \Rightarrow |-3.26| < 1.96$

$z > z_{tab}$

$\therefore$ null hypothesis is not accepted.

A company claims its bulbs are superior to those of its main competitor. If a study showed that a sample 40 of its bulbs have a mean life time of 647 hrs of continuous use with a S.D of 27 hrs. While a sample of 40 bulbs made by its main competetor has a mean of 638 hrs continuously with S.D of 31 hrs. Test the significance b/w the differences of a means with 5% of level of significance.

Sol:-

$n_1 = 40 \quad \bar{x}_1 = 647 \quad \sigma_1 = 27$

$n_2 = 40 \quad \bar{x}_2 = 638 \quad \sigma_2 = 31$

$H_0: \bar{x}_1 = \bar{x}_2$

$H_1: \bar{x}_1 \neq \bar{x}_2$

$\alpha = 0.05$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{647-638}{\sqrt{\frac{27^2}{40} + \frac{31^2}{40}}} = 1.38$$

$z_{tab} = 1.96$

$z < z_{tab}$

$\therefore$ null hypothesis is accepted.

The average marks scored by 32 boys is 72 with a S.D of 8. While that for 36 girls is 70 with S.D of 6. Does this indicate that the boys performed better than girls at the level of significance 0.05

Sol:- $n_1 = 32 \quad n_2 = 36 \quad \bar{x}_1 = 72 \quad \bar{x}_2 = 70 \quad \sigma_1 = 8$

$\sigma_2 = 6$.

$H_0: \bar{x}_1 = \bar{x}_2$

$H_1: \bar{x}_1 > \bar{x}_2$

$\alpha = 0.05$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{72-70}{\sqrt{\frac{64}{32} + \frac{36}{36}}} = 1.15$$

$z_{tab} = 2.33$

$z < z_{tab}$

$\therefore$ null hypothesis is accepted

The nicotent in mg of 2 samples of tobacco were found to be as follows. Find the S.E and confidence limits for the difference b/w the means at 0.05 level.

Sample A	24	27	26	23	25	
Sample B	29	30	30	31	24	36

Sol: sample A mean $\bar{x}_1 = 25$ $s_x^2 = 1.99^2$

B " $\bar{x}_2 = 30$ $s_y^2 = 12.32$

$$S.E = \sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}$$

$$= \sqrt{\frac{1.99^2}{5} + \frac{12.36}{6}} = 1.568$$

Confidence limits $= \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$= (\bar{x}_1 - \bar{x}_2) \pm 3 S.E (\bar{x}_1 - \bar{x}_2)$

$= (25-30) \pm 1.96 (25-30)$

$= (-20.288, 10.288)$

26/09/15

TESTING OF HYPOTHESIS FOR SINGLE PROPORTIONS:

Suppose a sample of size 'n' is taken from a normal population and sample is large then to test the significance difference between the sample proportions p and population proportion P we use the statistic $z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

NOTE:-

Limits for population propotion 'P' are given by

$p \pm 3\sqrt{\frac{PQ}{n}}$; $Q = 1 - P$

1. A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory confirmed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 are faulty. Test his claim at 5% level of significance

Sol: sample $n = 200$

No. of pieces confirming to the specifications

$= 200 - 18$

$= 182$

$p = \frac{182}{200} = 0.91$

$P = \frac{95}{100} = 0.95$

$Q = 1 - P = 1 - 0.95 = 0.05$

Null hypothesis $H_0 : P = 0.95$

$H_1 : P \geq 0.95$

level of significance $\alpha = 0.05$

test of statistic $z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$= \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.05}{200}}}$

$= -2.595$

$z_{cal} = -2.595$

$z_{tab} = z_{\alpha 0.05} = 1.645$

$\therefore |Z_{cal}| > Z_{tab}$

i.e., $2.595 > 1.645$

$\therefore$ Null hypothesis is rejected

i.e., the manufacturer claimed is false.

2. In a sample of 1000 people in karnataka 540 are rice eaters and the rest of them are wheat eaters. Can we assume that both rice & wheat eaters are equally popular at 1% level of significance.

Sol:- $n = 1000$

Rice eaters are 540 $P = \frac{540}{1000} = 0.54$

$P = \frac{1}{2} = 0.5 \quad Q = \frac{1}{2} = 0.5$

Null hypothesis: $H_0: P = 0.5$

Alternative " $H_1: P \neq 0.5$

level of significance $\alpha = 0.01$

test statistic $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$= \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{1000}}} = 2.52$

$Z_{cal} = 2.52$

$Z_{tab} = 2.58$

$Z_{cal} < Z_{tab}$ i.e., $2.52 < 2.58$

$\therefore$ Null hypothesis is accepted.

3. In a big city 325 men out of 600 men is found to be a smokers. Does this information support the conclusion that the majority of men in the city are smokers

Sol:- $n = 600$

$p = \frac{325}{600} = 0.541$

$P = 0.5 \quad Q = 0.5$

Null hypothesis: $H_0: P = 0.5$

$H_1: P \neq 0.5$

$\alpha = 005$

$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.541 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{600}}}$

$Z_{cal} = 1.97$

$Z_{tab} = 1.645$

$Z_{cal} > Z_{tab}$ i.e., $1.97 > 1.645$

Null hypothesis is not accepted

4. A dice was thrown 9000 times and in this 3220 times it is 3 or 4. Is this consistent with the hypothesis that a dice is unbiased

Sol:- $n = 9000$

$p = \frac{3220}{9000} = 0.3577$

$P = 2/6 = 1/3 \quad Q = 1 - \frac{1}{3} = 2/3$

$H_0: p = 1/3$

$H_1: p \neq \frac{1}{3}$

$\alpha = 0.05$

$$z = \frac{0.357 - 0.3333}{\sqrt{\frac{0.3333 \times 2/3}{2000}}} = 4.91$$

$z_{cal} = 4.91$

$z_{tab} = 1.96$

$z_{cal} \geq z_{tab}$

Null hypothesis is ~~accepted~~ rejected

5. A random sample of 500 pineapples was taken from a large consequent & 65 were found to be bad. Find the percentage of bad pineapples in the consequent. A social worker believes that

6. A manufacturer claims that only 4% of his products are defected. A random sample of 500 are taken among which 100 were defective. Test the hypothesis at 0.05 level.

Sol: $n = 500$

No. of pineapples were good $= 500 - 65$

$= 435$

$p = \frac{435}{500} = 0.87$ $\quad q = 1 - p$

$= 1 - 0.13 = 0.87$

$P = 0.87$ $\quad Q = 0.13$

$$Z = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.87 \times 0.87}{500}} = 0.0745$$

limits for population proportion

$= [0.87 \pm 3(0.022)]$

$= [0.804, 0.936)$

Sol: $n = 500$

No. of pieces which are not defective $= 500 - 100$

$= 400$

$P = \frac{100}{500} = 1/5 = 0.2$

$q = 0.8$ $\quad H_0: P = 0.04$ $\quad H_1: P > 0.04$

$P = \frac{4}{100} = 0.04$; $Q = 0.96$

$$Z = \frac{0.2 - 0.04}{\sqrt{\frac{0.04(0.96)}{500}}} = 18.25$$

$|Z|_{cal} = 18.25$

$|Z|_{tab} = 1.96$

$|Z|_{cal} > Z_{tab}$

H_0 is rejected

In a random sample of 160 workers exposed to a certain amount of radiation, 24 experienced some ill effects. Construct 99% confidence interval for the corresponding true percentage.

Sol: $n = 160$

$p = \frac{24}{160} = 0.15 \quad q = 0.85$

Confidence limit

$= \left(p \pm 3\sqrt{\frac{pq}{n}}\right)$

$= \left(0.15 \pm 3\sqrt{\frac{0.15 \times 0.85}{160}}\right)$

$= (0.065, 0.234)$

Test of significance of difference between two sample proportions:

Let p_1 & p_2 will be the sample proportions in two large random samples of sizes n_1 & n_2 drawn from a two populations P_1, P_2. Then test whether the 2 samples have been drawn from the same population

i) Null hypothesis $H_0: P_1 = P_2$

ii) Alternative hypothesis $H_1: P_1 \neq P_2$

iii) Test statistic:

There are 2 ways to compute test static

a) When the population proportions P_1 & P_2 are known in this case $Q_1 = 1 - P_1$ $Q_2 = 1 - P_2$ then S.E of difference i.e.,

$$S.E(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

Hence the test statistic $z = \frac{p_1 - p_2}{S.E(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$

b) When the population proportions P_1 & P_2 are unknown but the sample proportions p_1 & p_2 are unknown. In this case we have 2 methods to find P_1 & P_2

* Method of substitution: In this method sample proportions p_1 & p_2 are substituded for P_1 & P_2 then S.E of $(p_1 - p_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

Hence the test statistic $z = \frac{p_1 - p_2}{S.E(p_1 - p_2)}$

$$= \frac{p_1 - p_2}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

* Method of pooling: In this method the estimated value for the 2 population proportions is obtained by pooling. The 2 sample proportions p_1 & p_2 into a single proportions p by the faumula given below.

sample proportions of two samples are estimated value of $p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$; $q = 1 - p$

$$S.E(p_1 - p_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

$$z = \frac{p_1 - p_2}{S.E(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$$

A random sample of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were infavour of the proposal. Test the hypothesis that the proportions of men and women infavour of the proposal are same at level of.

Sol:- $n_1 = 400 \quad n_2 = 600$

$p_1 = \frac{200}{400} = 0.5 \quad q_1 = 1 - 0.5 = 0.5$

$p_2 = \frac{325}{600} = 0.541 \quad q_2 = 1 - 0.541 = 0.459$

Null hypothesis : $H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

level of significance $\alpha = 0.05$

Test statistic $z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}}}$

$= \frac{0.5 - 0.541}{\sqrt{\frac{0.5 \times 0.5}{400} + \frac{0.541 \times 0.459}{600}}}$

$z = -1.272$

$|z_{cal}| = 1.27.$

$z_{tab} = 1.96$

$\therefore |z_{cal}| < z_{tab}$ i.e, $1.28 < 1.96$

$\therefore$ We Accept the null hypothesis H_0.

A manufacture of electronic equipment subjects samples of 2 completing brands of transistors to an accelerated performance test. 45 of 180 transistors of the 1st kind and 34 of 120 transistors of the 2nd kind failed the test. what we can conclude at the level of significance 0.05 about the difference corresponding sample proportions.

Sol:- $n_1 = 180 \quad n_2 = 120$

$p_1 = \frac{45}{180} = 0.25 \quad q_1 = 1 - 0.25 = 0.75$

$p_2 = \frac{34}{120} = 0.283 \quad q_2 = 1 - 0.283 = 0.717$

Null hypothesis : $H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

$\alpha = 0.05$

$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}}}$

$= \frac{0.25 - 0.283}{\sqrt{\frac{0.25 \times 0.75}{180} + \frac{0.283 \times 0.717}{120}}}$

$z = -0.631$

$|z_{cal}| = |-0.631| = 0.631$

$z_{tab} = 1.96$

$\therefore |z_{cal}| < z_{tab}$ i.e, $0.631 < 1.96$

$\therefore$ Null Hypothesis is accepted.

In 2 large populations there are 30% & 25% respectively fair haired people. Is the difference likely to be hidden in samples of 1200 & 900 respectively from the 2 populations.

Sol:- $n_1 = 1200$ $n_2 = 900$

$P_1 = \frac{30}{100} = 0.3$ $Q_1 = 1 - P_1 = 1 - 0.3 = 0.7$

$P_2 = \frac{25}{100} = 0.25$ $Q_2 = 1 - 0.25 = 0.75$

Null hypothesis $H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

$\alpha = 0.05$

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

$$= \frac{0.3 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}}}$$

$$= 2.553$$

$Z_{cal} > Z_{tab}$

∴ Null hypothesis is rejected.

In a city A 20% of a random sample of 900 school boys has a certain slide physical defect. In another city B 18.5% of a random sample of 1600 school boys has the same defect. Is the difference b/w the proportions is significant at 0.05 level of significance.

Sol:- Defected persons in A $= \frac{20}{100} \times 900 = 180$

" " B $= \frac{18.5}{100} \times 1600 = 296$

$P_1 = \frac{180}{900} = 0.2$ $P_2 = \frac{296}{1600} = 0.185$

$q_1 = 1 - P_1 = 0.8$ $q_2 = 1 - P_2 = 0.815$

$H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

$\alpha = 0.05$

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}}}$$

$$= \frac{0.2 - 0.185}{\sqrt{\frac{0.2 \times 0.8}{900} + \frac{0.185 \times 0.815}{1600}}}$$

$$= 0.909$$

$Z_{cal} < Z_{tab}$

∴ Null hypothesis is accepted.

18-05-2021

Analysis of Variance (ANOVA):

ANOVA is a statistical technique to make inference about multiple parameters relating to population mean. Inferences conserning multiple population mean will be considering in ANOVA.

In this analysis, one can suppose that the experiment has available results of k independent random samples from k different populations & he or she is concerned

with testing the hypothesis that means of these k populations are all equal.

One way classification

Suppose that the mean of R.V. depends only on a single factor, namely the sample, the variable is from, then this scenario is said to be constituting a one way analysis of variance.

consider k independent random samples each of size n_i, where the members of the i^{th} sample $y_{i1}, y_{i2}, \ldots, y_{in_i}$ are normal random variables with unknown mean and unknown variance

Observations

sample-1 :	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1j}	y_{1n_1}
sample-2 :	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2j}	y_{2n_2}
sample-i :	y_{i1}	y_{i2}	y_{i3}	$\cdots$	y_{ij}	y_{in_i}
sample-k :	y_{k1}	y_{k2}	y_{k3}		y_{kj}	y_{kn_k}

mean	sum of squares
$\bar{y}_1$	$\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2$
$\bar{y}_2$	$\sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2$
$\bar{y}_i$	$\vdots$
$\bar{y}_k$	$\vdots$

To simplify the calculations, let

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \qquad N = \sum_{i=1}^{k} n_i$$

then, the overall mean / grand mean

$$\bar{\bar{y}} = \frac{T}{N}$$

this analysis leads to a comparison of k different population means consist essentially of splitting the sum of squares about the grand mean $\bar{\bar{y}}$ into a component due a variation between the samples and a component due to variation within the sample. i.e., each observation y_{ij} will be decomposed as

$$y_{ij} = \bar{\bar{y}} + (\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i)$$

$$\Rightarrow (y_{ij} - \bar{\bar{y}}) = (\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i) \quad \text{---} *)$$

variance due to grand mean
= variance due to between sample
+ variance due to within the sample squaring on both sides and taking Σ on both sides.

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$+ 2 \sum_{j=1}^{k} \sum_{i=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})(y_{ij} - \bar{y}_i)$$

$$= \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)$$

since $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 + \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$SST = SS_b + SSW.$$

Total sum of squares = sum of squares b/w samples + Error sum of squares

Similarly, the decomposition of degrees of freedom associated with the above sum of squares, for the general one way analysis of variance is,

Total degrees of freedom = dof between samples + dof within the sample

$$N - 1 = (k-1) + (N-k).$$

further each sum of squares is converted a mean square to test for the equality of means & it is given by,

Mean square (MS) = $\frac{\text{sum of squares}}{\text{corresponding dof}}$

$$M_{ib} = \frac{SS_b}{k-1} \quad ; \quad M_{sc} = \frac{SS_c}{N-k}.$$

now consider the ratio between the 2 mean squares.

$$F = \frac{SS_b}{MS_c} M_{ib} = \frac{SS_b/k-1}{SS_c/N-k}$$

here the variable has an F distribution with $k-1$ & $N-k$ degrees of freedom. To test for the equality of means we compare the ANOVA. by comparing these P value with F-table value for the given level of significance, i.e., if $F < F_{k-1, N-k, X}$

★ $$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_j} (y_{ij} - \bar{y})^2$$

$$SS_b = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

$$SSE = \sum_{j=1}^{n_j}\sum_{i=1}^{k} (y_{ij} - \bar{y}_i)^2$$

Mean square = $\frac{\text{sum of square}}{\text{dof}}$

$$MS_b = \frac{SS_b}{k-1}$$

$$SSE = \frac{SSE}{N-k}$$

$$F = \frac{MS_b}{MSE} = \frac{SSb/k-1}{SSE/N-k}$$

source of variation	degrees of freedom	sum of squares	Mean squares	F
rows	$r-1$	SS_r	$MS_r = \frac{SS_r}{(r-1)}$	$F_r = \frac{MS_r}{MSE}$
columns	$c-1$	SS_c	$MS_c = \frac{SS_c}{(c-1)}$	$F_c = \frac{MS_c}{MSE}$
error	$(r-1)(c-1)$	SSE	$MSE = \frac{SSE}{(r-1)(c-1)}$	

Problems

1. An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following cleanness readings were obtained.

	inj 1	inj 2	inj 3
Det·A	45	43	51
Det·B	47	46	52
Det·C	48	50	55
Det·D	42	37	49

Obtain the appropriate ANOVA.

Sol: $r=4, c=3 \quad T_{..} = 565$

~~$T_{..} = 8302$~~

$$C = \frac{T_{..}^2}{rc} = \frac{(565)^2}{4(3)} = 26602.08$$

$$SS_R = \frac{1}{c}\sum_{i=1}^{r} T_{i.}^2 - C$$

$$= \frac{1}{c}\left[T_{1.}^2 + T_{2.}^2 + T_{3.}^2 + T_{4.}^2\right] - C$$

$$= \frac{1}{3}\left[139^2 + 145^2 + 153^2 + 128^2\right] - C$$

$$= \frac{80139}{3} - 26602.08$$

$$= 26713 - 26602.08$$

$$= 110.92$$

$$SS_C = \frac{1}{r}\sum_{i=1}^{c} T_{.i}^2 - C$$

$$= \frac{1}{4}\left[182^2 + 176^2 + 207^2\right] - C$$

$$= \frac{106949}{4} - 26602.08$$

$$= 26737.25 - 26602.08$$

$$= 135.17$$

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}^2 - C$$

$$=$$

$SST = 1262 - 960 = 302$

$\Rightarrow \boxed{SST = 302}$

$SSE = SST - SSb$

$= 302 - 270 = 32$

$\boxed{SSE = 32}$

Mean squares:

$MSb = \frac{SSb}{K-1} = \frac{270}{3-1} = 135$

$\boxed{MSb = 135}$

$MSE = \frac{SSE}{N-K} = \frac{132}{15-3} = \frac{32}{12} = 2.667$

Now, $F = \frac{MSB}{MSE} = \frac{135}{2.667} = 50.618$

Source of sample	dof	sum of squares of squ.	Mean squ.	F.
Between sample	2	270	135	50.618
error	12	32	2.667	

2. The following are the number of mistakes made in 5 successive days for 4 techniques technicians working for photographic laboratory. Construct one way ANOVA table

Tech1	6	14	10	8	11
Tech2	14	9	12	10	14
Tech3	10	12	7	15	11
Tech4	9	12	8	10	11

$K = 4, n_1 = 5, n_2 = 5, n_3 = 5, n_4 = 5, N = 20$

$T_1 = 49, T_2 = 59, T_3 = 55, T_4 = 50$

$T_. = 213$

$C = \frac{T_.^2}{N} = \frac{(213)^2}{20} = 2268.45$

$SSb = \sum_{i=1}^{K} \frac{T_i^2}{n_i} - C$

$= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} - C$

$= \frac{(49)^2}{5} + \frac{(59)^2}{5} + \frac{(55)^2}{5} + \frac{(50)^2}{5} - 2268.45$

$= 2281.4 - 2268.45$

$= 12.95$

$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}^2 - C$

$= 517 + 717 + 639 + 510 - 2268.45$

$= 2383 - 2268.45$

$= 114.55$

$SSE = SST - SSb$

$= 114.55 - 12.95 = 101.6$

Mean squares:

$$MSb = \frac{SSb}{K-1} = \frac{12.95}{3} = 4.317$$

$$MSE = \frac{SSE}{N-K} = \frac{101.6}{16} = 6.35$$

$$F = \frac{MSb}{MSE} = \frac{4.317}{6.35} = 0.677$$

20-05-24

3. The effectiveness of three methods of teaching programming of a certain language to be compared. Random samples of size 4 are taken from large group of persons taught by these various methods. The following are scores which they obtained in an appropriate achievement test.

Method A : 73 77 67 71
Method B : 91 81 87 85
Method C : 72 77 76 79

Perform one way ANOVA

Sol $k=3, n_1=n_2=n_3=4 \quad N=12$

$T_1 = 288, \; T_2 = 344, \; T_3 = 304$

$T_. = 936 \qquad C = \frac{T_.^2}{N} = \frac{(936)^2}{12}$

$C = 73008$

$$SSb = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C.$$

$$= \frac{(288)^2}{4} + \frac{(344)^2}{4} + \frac{(304)}{4} - 73008$$

$$= 73424 - 73008$$

$$= 416$$

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - C.$$

$$= 20788 + 29636 + 23130 - 73008$$

$$= 546$$

$$SSE = SST - SSb$$

$$= 546 - 416 = 130.$$

$$MSb = \frac{SSb}{K-1} = \frac{416}{2} = 208$$

$$MSE = \frac{SSE}{N-K} = \frac{130}{9} = 14.44$$

$$F = \frac{MSb}{MSE} = 14.4044$$

S.S	DOF	S.S	M.S	F
b/w	2	416	208	14.4
error	9	130	14.44	
Total	11	546		

Two way Classification

Here we consider models that assume there are two factors that determine the mean value of a variable in these models, the variables can be thought of being arranged in a rectangular array with the mean value of a specified variable depending on both row and column in which it is located.

Let y_{ij} denote the observation pertaining to the i^{th} sample and j^{th} method where, $i = 1, 2, 3, \ldots, r$;
$j = 1, 2, 3, \ldots, c$

$\bar{y}_{i.}$ = the mean of the c observations for i^{th} row.

$\bar{y}_{.j}$ = mean of r observations for the j^{th} column.

$\bar{y}_{..}$ = grand mean of r×c observations methods.

	m-1	m-2	m-j	m-c	
sample1	y_{11}	y_{12} ..	y_{1j} ..	y_{1c}	$\bar{y}_{1.}$
sample 2	y_{21}	y_{22} ..	y_{2j} ..	y_{2c}	$\bar{y}_{2.}$
samplei	...	...	...	..	
sampler	y_{r1}	y_{r2}	y_{rj} ..	y_{rc}	$\bar{y}_{r.}$
	$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\bar{y}_{.j}$	$\bar{y}_{.c}$	

$$T_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij}$$

$$\bar{y}_{..} = \frac{T_{..}}{rc}$$

sum of squares due to rows

$$SS_r = \frac{1}{c} \sum_{i=1}^{r} T_{i.}^2 - C$$

sum of squares due to columns

$$SS_c = \frac{1}{r} \sum_{j=1}^{c} T_{.j}^2 - C$$

where, $C = \frac{T_{..}^2}{rc}$

Also, Total sum of squares

$$SST = \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij}^2 - C$$

$\therefore SSE = SST - SS_r - SS_c$

Mean square values are

Mean square due to rows:

$$MS_r = \frac{SS_r}{r-1}$$

Mean square due to columns:

$$MS_c = \frac{SS_c}{c-1}$$

$$MSE = \frac{SSE}{(r-1)(c-1)}$$

The ratios among the mean squares:

about rows: $F_r = \frac{MS_r}{MSE}$

about columns: $F_c = \frac{MS_c}{MSE}$

19-05-2021

One Way ANOVA Table

source of sample	degrees of freedom	Sum of squares	mean squares	F
Between sample	$k-1$	$SS_b = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$MS_b = \frac{SS_b}{k-1}$	$F = \frac{MS_b}{MSE}$
Error	$N-k$	$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$MSE = \frac{SSE}{N-k}$	

Note: Let $C = \frac{T_.^2}{N}$

$$SS_b = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$SS_E = SS_T - SS_b$$

here $T_. = \sum_{i=1}^{k} T_i$; $T_i = \sum_{j=1}^{n_i} y_{ij}$

Problems

1. Suppose 3 drying formulae for curing a glue are studied and the following curing times observed. Construct the one way ANOVA table for the data.

Formula A	13	10	8	11	8	
Formula B	13	11	14	14		
Formula C	4	1	8	4	2	4

Sol $n_1 = 5$; $n_2 = 4$; $n_3 = 6$, $k = 3$

$T_1 = 50$; $T_2 = 52$; $T_3 = 18$

$T_. = T_1 + T_2 + T_3 = 120$.

$$C = \frac{T_.^2}{N} = \frac{(120)^2}{15} = 960$$

$$SS_b = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

$$= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - C$$

$$= \frac{(50)^2}{5} + \frac{(52)^2}{4} + \frac{(18)^2}{6} - 960$$

$$= 500 + 52(13) + 18(3) - 960$$

$$\boxed{SS_b = 270}$$

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$= (y_{11}^2 + y_{12}^2 + y_{13}^2 + y_{14}^2 + y_{15}^2) + (y_{21}^2 + y_{22}^2 + y_{23}^2 + y_{24}^2) + (y_{31}^2 + y_{32}^2 + y_{33}^2 + y_{34}^2 + y_{35}^2 + y_{36}^2) - C$$

Small Sample Tests

if sample size (n) < 30 then we say it is small sample.

We have small sample Test's (Topics):-

I. Student T-Test (About Means)
   (i) T-Test for single Mean
   (ii) T-Test for difference of Means
   (iii) T-Test for paired data

II. F-Test (About variances)

III. χ^2-Test (chi-square Test) ⇒ (large sample also)
   (i) goodness of fit
   (ii) Independent attributes.

Student's T-Test

(i) Student's T-Test for single Mean:-

(1) Null Hypothesis (H_0): μ = given value (or) $\mu = \bar{x}$

(2) Alternative Hypothesis (H_1): $\mu \neq$ given (or) $\mu >$ given (or) $\mu <$ given

(3) Level of significance $(1-\alpha)$.

degrees of freedom $\nu = n-1$

(4) Test statistic

$$t_{cal} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \quad (or) \quad t_{cal} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n-1}}$$

(5) Conclusion

$|t_{cal}| < t_{tab} \Rightarrow$ Accepted H_0

(or) $|t_{cal}| \geq t_{tab} \Rightarrow$ reject H_0

Example 1:-

A random sample of 10 boys had following IQ's: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100

(i) Do these data support assumption of a population mean IQ of 100?

(ii) Find a reasonable range of a most of the mean IQ's values of sample of 10 boy's? [Confidence Interval for True Mean]

Solution:-

10 boy's IQ's given

70, 120, 110, 101, 88, 83, 95, 98, 107, 100

Single Mean ($\bar{x}$) = $\frac{70+120+110+101+88+83+95+98+107+100}{10}$

$\therefore \bar{x} = 97.2$

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

x	$x-\bar{x}$	$(x-\bar{x})^2$
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	+0.8	0.64
107	9.8	96.04
100	2.8	7.84

$\Sigma(x-\bar{x})^2 = 1833.6$

$\therefore s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

$s^2 = \frac{1833.6}{10-1}$

$s^2 = 203.73$

$s = 14.273$

(i) given $\mu = 100$

(ii) Null Hypothesis (H_0): $\mu = 100$

(iii) Alternative Hypothesis (H_1): $\mu \neq 100$

(iv) Level of significance (α) = 0.05

degrees of freedom (ν) = $n-1 = 10-1 = 9$

(v) Test statistic $t_{cal} = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} = \frac{97.2-100}{14.273/\sqrt{10-1}}$

$\therefore t_{cal} = -0.588$

$\therefore |t_{cal}| = 0.588$

(vi) $|t_{cal}| = 0.588 < t_{tab} = 1.83$

$\therefore$ we accept null hyp H_0 (t_{tab} at 0.05 LOS & 9 dof)

(ii) Confidence Interval True Mean
($\because t_\alpha = t_{tab}$ value)

$$\left[\bar{x} - t_\alpha\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{x} + t_\alpha\left(\frac{s}{\sqrt{n}}\right)\right]$$

$$\left[97.2 - 1.83\left(\frac{14.273}{\sqrt{10}}\right) \le \mu \le 97.2 + 1.83\left(\frac{14.273}{\sqrt{10}}\right)\right]$$

$$C.I = [88.94 \le \mu \le 105.459]$$

H.W

Example (2):-

A random sample of 8 ~~envelops~~ envelops is taken from letter box of a post office and their weights in grams are found to be 12.1, 11.9, 12.4, 12.3, 11.9, 12.1, 12.4, 12.1

(i) Does this sample indicate at 1% level that the Average weight of envelops received at their post office is 12.35 gram

(ii) Find 99% Confidence interval for the Mean cost of the envelops received at that post office.

Example (3):-

A random sample of 10 bags of pesticides are taken who weights are 50, 49, 52, 44, 45, 48, 46, 45, 49, 45 (in kgs) Test whether the average packing can be taken to be 50 kgs?

(ii) Student T-Test for difference of Means

(i) Null Hypothesis (H_0): $\mu_1 = \mu_2$

(ii) Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$ (or)

$\mu_1 > \mu_2$ (or)

$\mu_1 < \mu_2$

(iii) Level of significance (LOS) $= 1-\alpha$

degrees of freedom (dof) $(v) = n_1 + n_2 - 2$

(iv) Test statistic

$$t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $S = \sqrt{\frac{\Sigma(x_i - \bar{x}_1)^2 + \Sigma(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}}$

(v) conclusion

if $|t_{cal}| < t_{tab} \Rightarrow$ Accepted H_0

If $|t_{cal}| \geq t_{tab} \Rightarrow$ rejected H_0

Note:-

Confidence Interval

$$\left[(\bar{x}_1 - \bar{x}_2) - t_\alpha S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu \leq (\bar{x}_1 - \bar{x}_2) + t_\alpha S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

Example 1:-

Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results.

Horse	A	28	30	32	33	33	29	34
Horse	B	29	30	30	24	27	29	-

Test whether the two horses have the same running capacity.

Sol:-

(i) Null Hypothesis (H_0): $\mu_1 = \mu_2$

i.e horse A & B having same running capacity

(ii) Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$

(not same running capacity

(iii) Level of Significance $(1-\alpha) = 0.05$

degrees of freedom $(\nu) = n_1 + n_2 - 2$

$\nu = 7 + 6 - 2 = 11$

(iv) Test Statistic

$$t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

∴ Horse A:

sample size $(n_1) = 7$

Sample Mean $(\bar{x}_1) = \frac{28+30+32+33+33+29+34}{7} = 31.3$

~~Sample Variance Standard Deviation~~

Horse B:

sample size $(n_2) = 6$

Sample mean $(\bar{x}_2) = \frac{29+30+30+24+27+29}{6} = 28.17$

x_1	$x_1 - \bar{x}_1$	x_2	~~$x_2 - \bar{x}_2$~~	$x_2 - \bar{x}_2$	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)^2$
28	−3.3	29	~~0.83~~	0.83	10.89	0.689
30	−1.3	30	~~1.83~~	1.83	1.69	3.35
32	0.7	30		1.83	0.49	3.35
33	1.7	24		−4.17	2.89	17.39
33	1.7	27		−1.17	2.89	1.145
29	−2.3	29		0.83	5.29	0.689
34	2.7				7.29	
				$\Sigma(x_1 - \bar{x}_1)^2 = 31.43$		26.613 $\Sigma(x_2 - \bar{x}_2)$

$$\therefore S^2 = \frac{\Sigma(x_1 - \bar{x})^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$S^2 = \frac{31.43 + 26.613}{7+6-2} \Rightarrow S^2 = 5.276$$

$$\therefore S = \sqrt{5.276} = 2.29$$

$$\therefore t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{31.3 - 28.17}{2.29\sqrt{\frac{1}{7} + \frac{1}{6}}}$$

$$\therefore t_{cal} = 2.5$$

(v) $|t_{cal}| = 2.5 > t_{tab} = 1.8$

∴ we reject null hyp H_0

i.e Horses A & B not having same running capacity.

H.W.

Example (2) :-

Two independent samples of 8 and 7 items respectively had the following values

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	

Is the difference between the means of samples significant?

(iii) Student T-Test for paired data (or) Paired T-Test

(i) Null Hypothesis (H_0): $M_1 = M_2$

(ii) Alternative Hypothesis (H_1): $M_1 \neq M_2$ (or) $M_1 > M_2$ (or) $M_1 < M_2$

(iii) Level of significance $(1-\alpha)$

Degrees of freedom (v) = no. of pairs $- 1$

$v = n - 1$

(iv) Test statistic

$$t_{cal} = \frac{\bar{d}}{s/\sqrt{n}} \quad \text{where} \quad \bar{d} = \frac{\Sigma d}{n}$$

$$s^2 = \frac{\Sigma (d_i - \bar{d})^2}{n-1}$$

(v) Conclusion

if $|t_{cal}| < t_{tab} \Rightarrow$ Accepted H_0

if $|t_{cal}| \geq t_{tab} \Rightarrow$ rejected H_0

Example 1:

Blood pressure of 5 women before and after intake of certain drug is

Before	110	120	125	132	125
After	120	118	125	136	121

Test at 1% level of significance?

Sol:-

(i) Null Hypothesis (H_0): no significance difference in blood pressure before and after ($M_1 = M_2$)

(ii) Alternative Hypothesis (H_1): $M_1 < M_2$

(iii) level of significance = 0.01

degrees of freedom

v = no. of paired data -1

$v = 5-1 = 4$

(iv) Test statistic

$$t_{cal} = \frac{\bar{d}}{s/\sqrt{n}}$$

Before x	after y	$d = y - x$	d^2
110	120	10	100
120	118	−2	4
123	125	2	4
132	136	4	16
125	121	−4	16
		$\Sigma d = 10$	$\Sigma d^2 = 140$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{10}{5} = 2 \Rightarrow \boxed{\bar{d} = 2}$$

$$s^2 = \frac{\Sigma(d-\bar{d})^2}{n-1} = \frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1} = \frac{140 - 2^2 \times 5}{5-1}$$

$$s^2 = 30 \Rightarrow \boxed{s = 5.477}$$

$$\therefore t_{cal} = \frac{2}{5.477/\sqrt{5}} = 0.81$$

(v) $|t_{cal}| = 0.81 < t_{tab} = 3.75$

(t_{tab} at 0.01 LOS & 4 d.o.f)

$\therefore$ we Accept null Hypp (H_0)

(H.W)

Example 2 :-

In a study of the effectiveness of Physical Exercise in weight reduction a group of 16 persons engaged in a prescribed program of physical exercise for months showed the following results

weight before	209	178	169	212	180	192	159	180	170	153
weight After (Exercise)	196	171	160	207	171	190	158	180	164	152

183	165	201	179	243	144
179	162	199	173	231	140

Use 0.1 level of significance to test whether the prescribed program of Exercise is effective?

F-Test

(for variances)

(i) Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$

(ii) Alternative Hypothesis (H_1): $\sigma_1^2 \neq \sigma_2^2$

(iii) Level of significance ($1-\alpha$)

degrees of freedom

$V_1 = n_1 - 1$ (which has large variance)

$V_2 = n_2 - 1$ (which has small variance)

(iv) $$f_{cal} = \frac{\text{greater variance}}{\text{smaller variance}} = \frac{S_1^2}{S_2^2} \quad (\text{if } S_1^2 > S_2^2)$$

(or) $$f_{cal} = \frac{S_2^2}{S_1^2} \quad (\text{if } S_2^2 > S_1^2)$$

(v) if $|f_{cal}| < f_{tab} \Rightarrow$ Accepted H_0

if $|f_{cal}| \geq f_{tab} \Rightarrow$ rejected H_0

Note:

$$\text{variance} = \frac{\text{Sum of Squares of deviations}}{\text{degrees of freedom}}$$

i.e $$S_1^2 = \frac{\Sigma(x_i - \bar{x})^2}{n_1 - 1}$$

$$S_2^2 = \frac{\Sigma(x_i - \bar{y})^2}{n_2 - 1}$$

Notation:
∴ Sum of Squares of deviation of x variable
$S = \Sigma(x_i - \bar{x})^2$

Example (1) :-

Two independent samples of 8 and 7 items respectively had the following values of the variables

Sample 1	9	11	13	11	16	10	12	14
Sample 2	11	13	11	14	10	8	10	

Do the this estimates of population variances differ significantly?

Sol:- (i) Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$ (or) $S_1^2 = S_2^2$

(ii) Alternative Hypothesis (H_1): $S_1^2 \neq S_2^2$

(iii) Level of significance $(1-\alpha) = 0.05$

degrees of freedom $v_1 = n_1 - 1 = 8-1 = 7$

$v_2 = n_2 - 1 = 7-1 = 6$

(iv) $f_{cal} = \dfrac{\text{greater variance}}{\text{smaller variance}}$

x Sample 1	y Sample 2	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$
9	11	-3	9	0	0
11	13	-1	1	2	4
13	11	1	1	0	0
11	14	-1	1	3	9
16	10	4	16	-1	1
10	8	-2	4	-3	9
12		0	0		
14	10	2	4	-1	1
			$\Sigma(x-\bar{x})^2$		$\Sigma(y-\bar{y})^2$
			$= 36$		$= 24$

$\bar{x} = 12$ $\quad$ $\bar{y} = 11$

$$\therefore S_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1-1} = \frac{36}{8-1} = \frac{36}{7} = 5.14$$

$$S_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2-1} = \frac{24}{7-1} = \frac{24}{6} = 4$$

$$\therefore F_{cal} = \frac{\text{greatvariance } (S_1^2)}{\text{smaller variance } (S_2^2)} = \frac{5.14}{4}$$

$$\therefore F_{cal} = 1.285$$

$$\therefore |F_{cal}| = 1.285 < F_{tab} = 4.21$$

$\therefore$ Accepted H_0

ie $S_1^2 = S_2^2$

$F_{(7,6)}$ at 0.05 los $= 4.21$

from Table

$v_1 = 7$

$v_2 = 6$

$\alpha = 0.05$

(4.10)

Example (2):-

Five measures of the output of two units have given the following results. Assume that both samples have been obtained from normal populations, to test at 10% significance level if two populations have the same variance

Unit A	14.1	10.1	14.7	13.7	14.0
Unit B	14.0	14.5	13.7	12.7	14.1

χ^2 - Test

(Chi - Square Test)

(i) χ^2 - Test for Goodness of fit

(i) null Hypothesis (H_0): $O_i = E_i$

i.e (observed = Expected)

(ii) Alternative Hypothesis (H_1): $O_i \neq E_i$

(iii) level of significance $(1-\alpha)$

(iv) Test Statistic

$$\chi^2_{cal} = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

(v) if $|\chi^2_{cal}| < \chi^2_{tab} \Rightarrow$ Accepted H_0

if $|\chi^2_{cal}| \geq \chi^2_{tab} \Rightarrow$ Rejected H_0

Note:

O_i = observed frequency

E_i = Expected frequency

χ^2 - Test for Independent Attributes

Let us consider the given contigency table

A	a	b
B	c	d

⇒ observed values

			row sum
	a	b	a+b
	c	d	c+d
column sum	a+c	b+d	N = a+b+c+d total

Expected values:-

$$E(a) = \frac{(\text{respective row sum}) \times (\text{respective column sum})}{\text{Total}}$$

$$\text{i.e } E(a) = \frac{(a+b)(a+c)}{N}$$

$$\text{Ily } E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(c+d)(a+c)}{N}$$

$$E(d) = \frac{(c+d)(b+d)}{N}$$

then $\chi^2_{cal} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$

Example 1:

200 digits were choosen at random from a set of tables. The frequencies of the digits were

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use χ^2 Test to assess the correctness of Hypothesis that the digits were distributed in equal number in the table.

Sol:

(i) null Hypothesis (H_0):

The frequence of the digits occurance is equal

ie the Expected frequencies are $= \frac{200}{10} = 20$ for each digit

(ii) Alternative Hypothesis (H_1):

frequency is not equal

(iii) level of significance (α) = 0.05

degrees of freedom = no. of observations (n) -1 ($\nu = n-1$)

(iv) Test statistic

$$\chi^2_{cal} = \sum \left(\frac{O_i - E_i)^2}{E_i} \right)$$

digits	Observed frequency (O_i)	expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	18	20	-2	0.2
1	19	20	-1	0.05
2	23	20	3	0.45
3	21	20	1	0.05
4	16	20	-4	0.8
5	25	20	5	1.25
6	22	20	2	0.2
7	20	20	0	0
8	21	20	1	0.05
9	15	20	-5	1.25
				$\Sigma = 4.3$

$$\therefore \chi^2_{cal} = \Sigma \left(\frac{(O_i - E_i)^2}{E_i} \right) = 4.3$$

$\chi^2_{tab} = 16.919$ at $0.05 = LOS$ & $9 = dof$

$$\therefore \left| \chi^2_{cal} \right| = 4.3 < \chi^2_{tab} = 16.919$$

$\therefore$ we Accepted H_0

i.e frequency of distribution is equal

Example ②:-

From the following table test whether the colour of the son's eyes is associated with that of the fathers.

eye colour of sons → / eye colour of fathers ↓	Brown	Black
Brown	417	151
Black	148	230

Sol:-

i) Null Hypothesis (H_0):

Son's eye colour independent on fathers eye colour.

ii) Alternative Hypothesis (H_1):

Son's eyes colour depends on fathers eye colour.

(iii) Level of significance $= \alpha = 0.05$

degrees of freedom = (no. of rows − 1) × (no. of columns)

i.e $= (2-1) \times (2-1) = 1$

(iv) Test statistic

$$\chi^2_{cal} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Son's → Father's	Brown	Black	row sum
Brown	471	151	622
Black	148	230	278
column sum	619	381	N = 1000

$$E(471) = \frac{622 \times 619}{1000} = 385$$

$$E(151) = \frac{622 \times 381}{1000} = 236.98 = 237$$

$$E(148) = \frac{278 \times 619}{1000} = 172.08 = 172$$

$$E(230) = \frac{278 \times 381}{1000} = 105.91 = 106$$

O_i	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
471	385	86	19.21
151	237	-86	31.20
148	172	-24	3.34
230	106	124	145.05

$$\Sigma = 198.8 \simeq 199$$

$$\therefore |\chi^2_{cal}| = 199 > \chi^2_{tab} = 3.841$$

∴ we reject null Hyp H_0

i.e son's eye colour depends on father's eye colour.

Example (3):-

Test whether the intelligence of sons depends on the intelligence fathers from the following data, using χ^2 test at 0.05 level of significance

Father's / Son's ↓	Intelligent	dull
intelligent	200	110
dull	50	600

Example (4):-

The following data gives the classification of 100 workers according to gender and nature of work. Using χ^2 Test Examine whether the nature of work is independent of the gender of the worker

Gender	Nature of work	
	Skilled	Unskilled
Male	40	20
female	10	30

Example (5):-

A die is thrown 60 times with the following results

face	1	2	3	4	5	6
frequency	8	7	12	8	14	11

Test at 5% level of significance.

Example (6):-

Among 64 off springs of certain cross between guinea pigs 34 were red, 10 were black and 20 were white. According to the genetic model these numbers should be in the ratio 9:3:4. Are the data consistent with the model at 5% level.

TIME SERIES

1.1. Meaning

An arrangement of statistical data in accordance with time of occurrence or in a chronological order is called a time series. The numerical data which we get at different points of time-the set of observations-is known as time series.

In time series analysis, current data in a series may be compared with past data in the same series. We may also compare the development of two or more series over time.These comparisons may afford important guide lines for the individual firm. In Economics, statistics and commerce it plays an important role.

1.2 Definition and Examples

A time series is a set of observations made at specified times and arranged in a chronological order.

For example, if we observe agricultural production, sales, National Income etc., over a period of time, say over the last 3 or 5 years, the set of observations is called time series. Thus a time series is a set of time, quantitative readings of some various recorded at equal intervals of time. The interval may be an hour, a day, a week, a month, or a calendar year. Hourly temperature reading, daily sales in a shop, weekly sales in a shop, weekly sales in a market, monthly production in an industry, yearly agricultural production, population growth in ten years, are examples of time series.

From the comparison of past data with current data, we may seek to establish what development may be expected in future. The analysis of time series is done mainly for the purpose of forecasts and for evaluating the past performances. The chronological variations will be object of our study in time series analysis.

The essential requirements of a time series are:

- The time gap, between various values must be as far as possible, equal.

- It must consist of a homogeneous set of values.
- Data must be available for a long period.

symbolically if „t" stands for time and „y_t" represents the value at time t then the paired values (t, yt) represents a time series data.

Ex 1: Production of rice in Tamilnadu for the period from2010-11 to 2016-17.

Table 1.1. Production of rice in Tamilnadu (in '000 metric tons)

Year	Production
2010-11	400
2011-12	450
2012-13	440
2013-14	420
2014-15	460
2016-17	520

1.3 Uses of Time Series

The analysis of time series is of great significance not only to the economists and business man but also to the scientists, astronomists, geologists, sociologists, biologists, research worker etc. In the view of following reasons

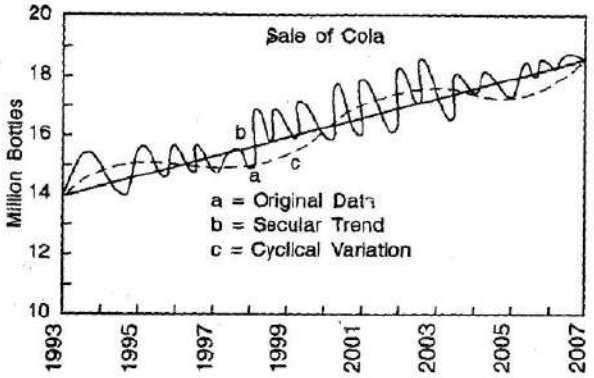
- It helps in understanding past behavior.
- It helps in planning future operations.
- It helps in evaluating current accomplishments.
- It facilitates comparison.

1.4 Components of Time Series

The values of a time series may be affected by the number of movements or fluctuations, which are its characteristics. The types of movements characterizing a time series are called components of time series or elements of a time series.

These are four types

- Secular Trend
- Seasonal Variations
- Cyclical Variations
- Irregular Variations



Secular Trend

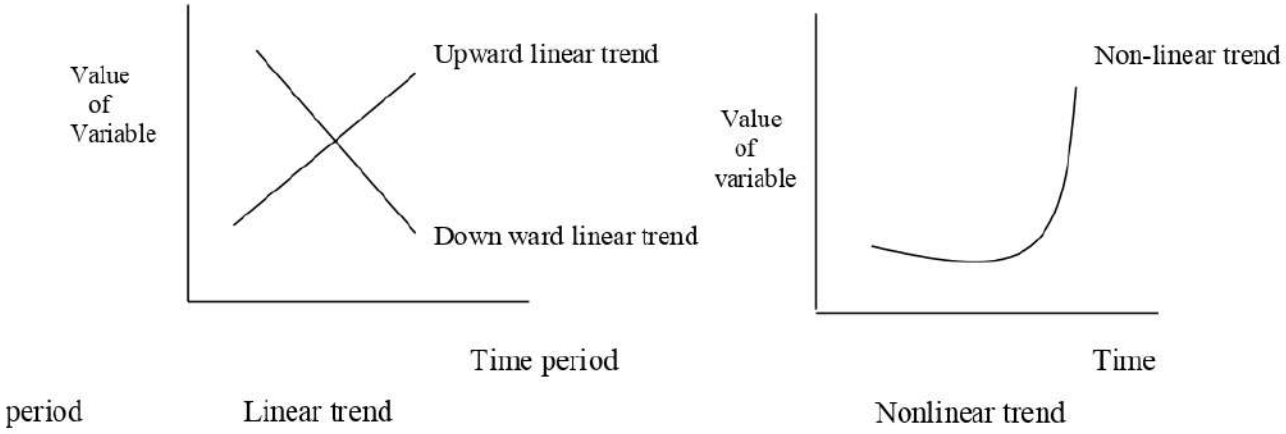
Secular Trend is also called long term trend or simply trend. The secular trend refers to the general tendency of data to grow or decline over a long period of time. For example the population of India over years shows a definite rising tendency. The death rate in the country after independence shows a falling tendency because of advancement of literacy and medical facilities. Here long period of time does not mean as several years. Whether a particular period can be regarded as long period or not in the study of secular trend depends upon the nature of data. For example if we are studying the figures of sales of cloth store for 1996- 1997 and we find that in 1997 the sales have gone up, this increase cannot be called as secular trend because it is too short period of time to conclude that the sales are showing the increasing tendency.

On the other hand, if we put strong germicide into a bacterial culture, and count the number of organisms still alive after each 10 seconds for 5 minutes, those 30 observations showing a general pattern would be called secular movement.

Mathematically the secular trend may be classified into two types

1. Linear Trend
2. Curvi-Linear Trend or Non-Linear Trend.

If one plots the trend values for the time series on a graph paper and if it gives a straight line then it is called a linear trend i.e. in linear trend the rate of change is constant where as in non-linear trend there is varying rate of change.



period

Figure 1.1. Linear Trend and Nonlinear Trend

Seasonal Variations

Seasonal variations occur in the time series due to the rhythmic forces which occurs in a regular and a periodic manner with in a period of less than one year. Seasonal variations occur during a period of one year and have the same pattern year after year. Here the periodof time may be monthly, weekly or hourly. But if the figure is given in yearly terms then seasonal fluctuations does not exist. There occur seasonal fluctuations in a time series due to two factors.

- Due to natural forces
- Manmade convention.

The most important factor causing seasonal variations is the climate changes in the climate and weather conditions such as rain fall, humidity, heat etc. act on different products and industries differently. For example during winter there is greater demand for woolen clothes, hot drinks etc. Where as in summer cotton clothes, cold drinks have a greater sale and in rainy season umbrellas and rain coats have greater demand.

Though nature is primarily responsible for seasonal variation in time series, customs, traditions and habits also have their impact. For example on occasions like dipawali, dusserah, Christmas etc. there is a big demand for sweets and clothes etc., there is a large demand for books and stationary in the first few months of the opening of schools and colleges.

Cyclical Variations or Oscillatory Variation

This is a short term variation occurs for a period of more than one year. The rhythmic movements in a time series with a period of oscillation(repeated again and again in same manner) more than one year is called a cyclical variation and the period is called a cycle. The time series related to business and economics show some kind of cyclical variations.

One of the best examples for cyclical variations is „Business Cycle". In this cycle there are four well defined periods or phases.

- Boom
- Decline
- Depression
- Improvement

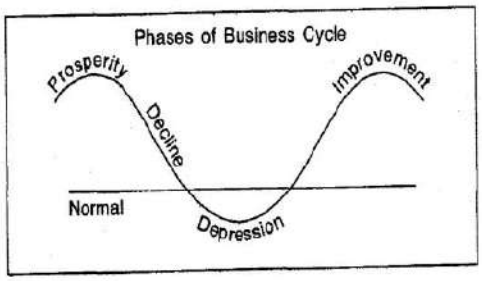


Figure 2. Phases of Business Cycle

Irregular Variation

It is also called Erratic, Accidental or Random Variations. The three variations trend, seasonal and cyclical variations are called as regular variations, but almost all the time series including the regular variation contain another variation called as random variation. This type of fluctuations occurs in random way or irregular ways which are unforeseen, unpredictable and due to some irregular circumstances which are beyond the control of human being such

as earth quakes, wars, floods, famines, lockouts, etc. These factors affect the time series in the irregular ways. These irregular variations are not so significant like other fluctuations.

1.5 Mathematical Model

In classical analysis, it is assumed that some type of relationship exists among the four components of time series. Analysis of time series requires decomposition of a series, to decompose a series we must assume that some type of relationship exists among the four components contained in it.

The value Y_t of a time series at any time t can be expressed as the combinations of factors that can be attributed to the various components. These combinationsare called as models and these are two types.

- Additive model
- Multiplicative model

Additive model

In additive model $Y_t = T_t + S_t + C_t + R_t$

Where T_t = Trend value at time t

S_t = Seasonal component

C_t = Cyclical component

R_t = Irregular component

But if the data is in the yearly form then seasonal variation does not exist, so in that situation $Y_t = T_t + C_t + R_t$

Generally the cyclical fluctuations have positive or negative value according to whether it is in above or below the normal phase of cycle.

Multiplicative model:

In multiplicative model $Y_t = T_t . S_t . C_t . R_t$.

The multiplicative model can be put in additive model by taking log both sides. However most business analysis uses the multiplicative model and finds it more appropriate to analyze business situations.

According to this model, the simple

One of the most important tasks before economists and businessmen these days is to make estimates for the future. For example, a businessman is interested in finding out his likely sales in the year 2016 or as a long-term planning in 2025 or the year 2030 so that he could adjust his production accordingly and avoid the possibility of either unsold stocks or inadequate production to meet the demand. Similarly, an economist is interested in estimatingthe likely population in the coming year so that proper planning can be carried out with regard to food supply, jobs for the people, etc. However, the first step in making estimates forthe future consists of gathering information from the past. In this connection one usuallydeals with statistical data which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as „time series". Thus when we observe numerical data at different points of time the set of observations is known as time series. For example if we observe production, sales, population, imports, exports, etc. at different points of time, say, over the last 5 or 10 years, the set of observations formed shall constitute time series. Hence, in the analysis of time series, time is the most important factor because the variable is related to time which may be either year, month, week, day and hour or even- minutes or seconds.

1.6 Measurement of Secular trend:

Secular trend is a long term movement in a time series. This component represents basic tendency of the series. The following methods are generally used to determine trend in any given time series. The following methods are generally used to determine trend in any given time series.

- Graphic method or eye inspection method
- Semi average method
- Method of moving average
- Method of least squares

Method of Moving Average:

It is a method for computing trend values in a time series which eliminates the short term and random fluctuations from the time series by means of moving average. Moving average of a period m is a series of successive arithmetic means of m terms at a time starting with 1^{st}, 2^{nd}, 3^{rd} and so on. The first average is the mean of first m terms; the second averageis the mean of 2^{nd} term to $(m+1)^{th}$ term and 3^{rd} average is the mean of 3^{rd} term to $(m+2)^{th}$ term and so on.

If m is odd then the moving average is placed against the mid value of the time interval it covers. But if m is even then the moving average lies between the two middle periods which does not correspond to any time period. So further steps has to be taken to place the moving average to a particular period of time. For that we take 2-yearly moving average of the moving averages which correspond to a particular time period. The resultant moving averages are the trend values.

Ex:1 Calculate 3-yearly moving average for the following data.

Years	Production	3-yearly moving avg (trend values)
2001-02	40	
2002-03	45	(40+45+40)/3 = 41.67
2003-04	40	(45+40+42)/3 = 42.33
2004-05	42	(40+42+46)/3 = 42.67
2005-06	46	(42+46+52)/3 = 46.67
2006-07	52	(46+52+56)/3 = 51.33
2007-08	56	(52+56+61)/3 = 56.33
2008-09	61	

Ex :2 Calculate 4-yearly moving average for the following data.

Years	Production	4-yearly moving avg	2-yealry moving avg (trend values)
2001-02	40		
2002-03	45		
		→ (40+45+40+42)/3 = 41.75	
2003-04	40		→ 42.5
		→ (45+40+42+46)/3 = 43.15	
2004-05	42		→ 44.12
		→ (40+42+46+52)/3 = 45	
2005-06	46		→ 47
		→ (42+46+52+56)/3 = 49	
2006-07	52		→ 51.38
		→ (46+52+56+61)/3 = 53.75	
2007-08	56		
2008-09	61		

Advantages:

- This method is simple to under stand and easy to execute.
- It has the flexibility in application in the sense that if we add data for a few more time periods to the original data, the previous calculations are not affected and we get a few more trend values.
- It gives a correct picture of the long term trend if the trend is linear.
- If the period of moving average coincides with the period of oscillation (cycle), the periodic fluctuations are eliminated.
- The moving average has the advantage that it follows the general movements of the data and that its shape is determined by the data rather than the statistician"s choice of mathematical function.

Disadvantages:

- For a moving average of 2m+1, one does not get trend values for first m and last m periods.

- As the trend path does not correspond to any mathematical; function, it cannot be used for forecasting or predicting values for future periods.
- If the trend is not linear, the trend values calculated through moving averages may not show the true tendency of data.
- The choice of the period is sometimes left to the human judgment and hence may carry the effect of human bias.

Method of Least Squares:

This method is most widely used in practice. It is mathematical method and with its help a trend line is fitted to the data in such a manner that the following two conditions are satisfied.

1. $\sum(Y - Y_c) = 0$ i.e. the sum of the deviations of the actual values of Y and the computed values of Y is zero.
2. $\sum(Y - Y_c)^2$ is least, i.e. the sum of the squares of the deviations of the actual values and the computed values is least.

The line obtained by this method is called as the "line of best fit".

This method of least squares may be used either to fit a straight line trend or a parabolic trend.

Fitting of a straight line trend by the method of least squares:

Let Y_t be the value of the time series at time t. Thus Y_t is the independent variable depending on t.

Assume a straight line trend to be of the form $Y_{tc} = a + bt$ ..... (1)

Where Y_{tc} is used to designate the trend values to distinguish from the actual Y_t values, a is the Y-intercept and b is the slope of the trend line.

Now the values of a and b to be estimated from the given time series data by the method of least squares.

In this method we have to find out a and b values such that the sum of the squares of

the deviations of the actual values Y_t and the computed values Y_{tc} is least.

i.e. $S = \sum (Y_t - Y_{tc})_2$ should be least

i.e. $S = \sum (Y_t - a - bt)^2$ ..... (2) Should be least

Now differentiating partially (2) w.r.to a and equating to zero we get

$$\frac{\partial S}{\partial a} = 2\sum (Y_t - a - bt)(-1) = 0$$

$$\Rightarrow \sum (Y_t - a - bt) = 0$$
$$\Rightarrow \sum Y_t = \sum a + b\sum t$$

$$\Rightarrow \sum Y_t = na + b\sum t \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Now differentiating partially (2) w.r.to b and equating to zero we get

$$\frac{\partial S}{\partial b} = 2\sum (Y_t - a - bt)(-t) = 0$$

$$\Rightarrow \sum t(Y_t - a - bt) = 0$$
$$\Rightarrow \sum tY_t = a\sum t + b\sum t^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

The equations (3) and (4) are called „normal equations"

Solving these two equations we get the values of a and b say $\hat{a}$ and $\hat{b}$.

Now putting these two values in the equation (1) we get

$$Y_{tc} = \hat{a} + \hat{b}t$$

which is the required straight line trend equation.

Note: The method for assessing the appropriateness of the straight line modal is the method of first differences. If the differences between successive observations of a series are constant

(nearly constant) the straight line should be taken to be an appropriate representation of the trend component.

Illustration 10. Below are given the figures of production (in thousand quintals) of a sugar factory :

Year	2001	2002	2003	2004	2005	2006	2007
Production (in '000 qtls.)	80	90	92	83	94	99	92

(*i*) Fit a straight line trend to these figures.
(*ii*) Plot these figures on a graph and show the trend line.
(M. Com., Jiwaji Univ.; M. Com., Ajmer Univ.; B. Com., HPU; B.Com., Bangalore Univ.)

Solution. (*i*) FITTING THE STRAIGHT LINE TREND

Year	*Production ('000 qtls.)* Y	X	XY	X^2	*Trend values* Y_c
2001	80	−3	−240	9	84
2002	90	−2	−180	4	86
2003	92	−1	−92	1	88
2004	83	0	0	0	90
2005	94	+1	+94	1	92
2006	99	+2	+198	4	94
2007	92	+3	+276	9	96
$N = 7$	$\Sigma Y = 630$	$\Sigma X = 0$	$\Sigma XY = 56$	$\Sigma X^2 = 28$	$\Sigma Y_c = 630$

The equation of the straight line is $Y_c = a + bX$.

To find a and b we have two normal equations

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Since $\Sigma X = 0$; $a = \frac{\Sigma Y}{N}$, $b = \frac{\Sigma XY}{\Sigma X^2}$

$\Sigma Y = 630$, $N = 7$, $\Sigma XY = 56$, $\Sigma X^2 = 28$,

$\therefore$ $a = \frac{630}{7} = 90$; and $b = \frac{56}{28} = 2$

Hence the equation of the straight line trend is $Y_c = 90 + 2X$.

Origin, 2004 : X units, one year; Y units, production in thousand quintals.

For $X = -3$, $Y_c = 90 + 2(-3) = 84$

For $X = -2$, $y_c = 90 + 2(-2) = 86$

For $X = -1$, $y_c = 90 + 2(-1) = 88$.

Similarly, by putting $X = 0, 1, 2, 3$, we can obtain other trend values. However, since the value of b is constant, first trend value need be obtained and then if the value b is positive we may continue adding the value of b to every preceding value. For 2002 it will be 84 + 2 = 86, for 2003 it will be 86 + 2 = 88, and so on. If b is negative, then instead of adding we will deduct.

(*ii*) The graph of the above data is given below :

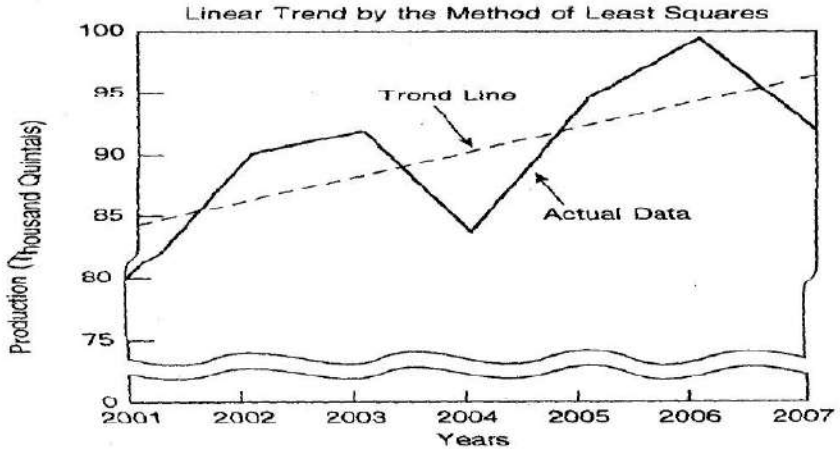


Illustration 32. Calculate trend values by the method of least-squares from the data given below :

Year	2000	2001	2002	2003	2004	2005	2006	2007
Sales	80	90	92	83	94	99	92	104

Plot the data showing also the trend line. *(B.Com., HPU; M.A. Econ., GNDU)*

Solution. FITTING STRAIGHT LINE TREND BY METHOD OF LEAST SQUARES

Years	*Sales* Y	*Deviations from 2003.5*	*Deviations multiplied by 2* X	XY	X^2	Y_c
2000	80	− 3.5	− 7	− 560	49	83.0
2001	90	− 2.5	− 5	− 450	25	85.5
2002	92	− 1.5	− 3	− 276	9	88.0
2003	83	− .5	− 1	− 83	1	90.5
2004	94	+ .5	+ 1	+ 94	1	93.0
2005	99	+ 1.5	+ 3	+ 297	9	95.5
2006	92	+ 2.5	+ 5	+ 460	25	98.0
2007	104	+ 3.5	+ 7	+ 728	49	100.5
$N = 8$	$\Sigma y = 734$			$\Sigma XY = 210$	$\Sigma X^2 = 168$	$\Sigma Y_c = 734$

The equation of the straight line is $Y_c = a + bX$.

To find a and b we have two normal equations

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Since $\Sigma X = 0$, $a = \frac{\Sigma Y}{N} = \frac{734}{8} = 91.75$, $b = \frac{\Sigma XY}{\Sigma X^2} = \frac{210}{168} = 1.25$

The required line equation is $Y = 91.75 + 1.25X$

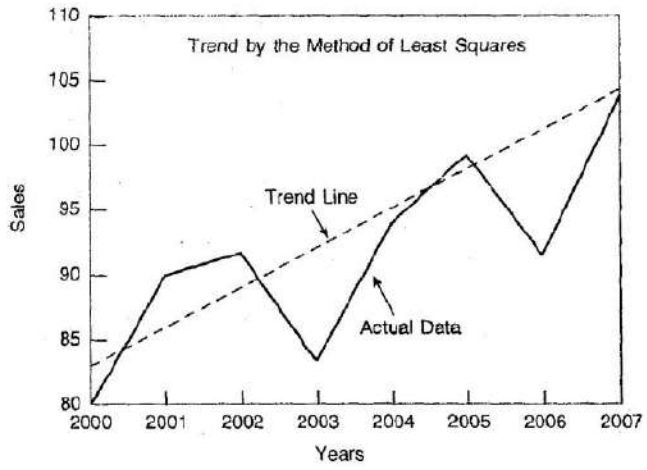
The trend values for various years are

$Y_{2000} = 91.75 + 1.25(-7) = 91.75 - 8.75 = 83$

For finding these trend values, double the value of b, *i.e.*, $1.25 \times 2 = 2.5$ and add to the preceding value :

$Y_{2001} = 83 + 2.5 = 85.5$

and so on



Fitting of a parabolic trend by the method of least squares

Let Y_t be the value of the time series at time t. Thus Y_t is the independent variable depending on t.

Assume a parabolic trend to be of the form $Y_{tc} = a + bt + ct_2$..... (1)

Now the values of a, b and c to be estimated from the given time series data by the method of least squares.

In this method we have to find out a, b and c values such that the sum of the squares of the deviations of the actual values Y_t and the computed values Y_{tc} is least.

i.e. $S = \sum (Y_t - Y_{tc})^2$ should be least

i.e. $S = \sum (Y_t - a - bt)^2$ ..... (2) Should be least

Now differentiating partially (2) w.r.to a and equating to zero we get

$$\frac{\partial S}{\partial a} = 2\sum (Y_t - a - bt - ct^2)(-1) = 0$$

$$\Rightarrow \sum (Y_t - a - bt - ct^2) = 0$$

$$\Rightarrow \sum Y_t = \sum a + b\sum t + c\sum t^2$$

$$\Rightarrow \sum Y = na + b\sum t + c\sum t^2$$

$$t \qquad \text{.....(3)}$$

Now differentiating partially (2) w.r.to b and equating to zero we get

$$\frac{\partial S}{\partial b} = 2\sum (Y_t - a - bt - ct^2)(-t) = 0$$

$$\Rightarrow \sum t(Y_t - a - bt - ct^2) = 0$$

$$\Rightarrow \sum tY_t = a\sum t + b\sum t_2 + c\sum t_3 \qquad \text{..... (4)}$$

Now differentiating partially (2) w.r.to c and equating to zero we get

$$\frac{\partial S}{\partial c} = 2\sum (Y_t - a - bt - ct^2)(-t^2) = 0$$

$$\Rightarrow \sum t^2 (Y_t - a - bt - ct^2) = 0$$

$$\Rightarrow \sum t^2 Y_t = a\sum t^2 + b\sum t^3 + c\sum t^4 \qquad \text{..... (5)}$$

The equations (3), (4) and (5) are called „normal equations"

Solving these three equations we get the values of a, b and c say $\hat{a}, \hat{b}$ and $\hat{c}$.

Now putting these three values in the equation (1) we get

$$Y_{tc} = \hat{a} + \hat{b}t + \hat{c}t^2$$

Which is the required parabolic trend equation

Note: The method for assessing the appropriateness of the second degree equation is the method of second differences. If the differences are taken of the first differences and the results are constant (nearly constant) the second degree equation be taken to be an appropriate

representation of the trend component.

Illustration 14. The prices of a commodity during 2002-2007 are given below. Fit a parabola $Y = a + bX + cX^2$ to these data. Estimate the price of the commodity for the year 2008 :

Year	Prices	Year	Prices
2002	100	2005	140
2003	107	2006	181
2004	128	2007	192

Also plot the actual and trend values on the graph. *(B.Com. (H), DU; M. Com., M.D. Univ.)*

Solution : To determine the values of a, b and c, we solve the following normal equations :

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2 \quad ...(i)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 \quad ...(ii)$$

$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 \quad ...(iii)$$

Year	Prices (Rs.) Y	X	X^2	X^3	X^4	XY	X^2Y	Trend Values (Y_c)
2002	100	− 2	4	− 8	16	− 200	400	97.717
2003	107	− 1	1	− 1	1	− 107	107	110.401
2004	128	0	0	0	0	0	0	126.657
2005	140	+ 1	1	+ 1	1	+ 140	140	146.485
2006	181	+ 2	4	+ 8	16	+ 362	724	169.885
2007	192	+ 3	9	+ 27	81	+ 576	1728	196.857
$N = 6$	$\Sigma Y = 848$	$\Sigma X = 3$	$\Sigma X^2 = 19$	$\Sigma X^3 = 27$	$\Sigma X^4 = 115$	$\Sigma XY = 771$	$\Sigma X^2 Y = 3,099$	$\Sigma Y_c = 848.002$

$$848 = 6a + 3b + 19c \quad ...(i)$$

$$771 = 3a + 19b + 27c \quad ...(ii)$$

$$3,099 = 19a + 27b + 115c \quad ...(iii)$$

Multiplying the second equation by 2 and keeping the first as it is, we get

$$848 = 6a + 3b + 19c$$

$$1,542 = 6a + 38b + 54c$$

$$-694 = -35b - 35c \quad ...(iv)$$

or $35b + 35c = 694$

Multiplying Eqn. (*ii*) by 19 and Eqn. (*iii*) by 3, we get

$$14,649 = 57a + 361b + 513c$$

$$9,297 = 57a + 81b + 345c$$

$$5,352 = 280b + 168c \quad ...(v)$$

Multiplying equation (iv) by 8, we have

$$280b + 280c = 5,552$$

Solving equations (*iv*) and (*v*)

$$280b + 280c = 5,552$$

$$280b + 168c = 5,352$$

$$112c = 200 \quad \text{or} \quad c = 1.786$$

Substituting the value of c in Eqn. (*iv*),

$$35b + (35 \times 1.786) = 694$$

$$35b = 694 - 62.5 = 631.5 \text{ or } b = 18.042$$

$$848 = 6a + 3(18.042) + 19(1.786) = 6a + 54.126 + 33.934$$

$$6a = 759.94 \quad \text{or} \quad a = 126.657$$

Thus $a = 126.657$, $b = 18.042$ and $c = 1.786$

Substituting these values in the equation,

$$Y = 126.657 + 18.042X + 1.786X^2$$

when $X = -2$

$$Y = 126.657 + 18.042(-2) + 1.786(-2)^2$$
$$= 126.657 - 36.084 + 7.144 = 97.717$$

when $X = -1$

$$Y = 126.657 + 18.042(-1) + 1.786(-1)^2$$
$$= 126.657 - 18.042 + 1.786 = 110.401$$

when $X = 1$,

$$Y = 126.657 + 18.042 + 1.786 = 146.485$$

when $X = 2$,

$$Y = 126.657 + 18.042\,(2) + 1.786\,(2)^2 = 169.885$$

when $X = 3$,

$$Y = 126.657 + 18.042\,(3) + 1.786\,(3)^2 = 196.857$$

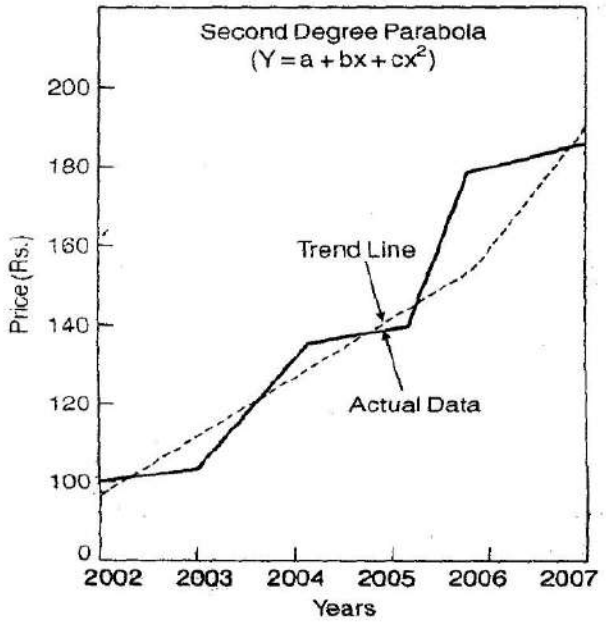
Price for the year 2008

For 2008 X would be equal to 4. Putting $X = 4$ in the equation,

$$Y = 126.657 + 18.042\,(4) + 1.786\,(4)^2$$
$$= 126.657 + 72.168 + 28.576 = 227.401.$$

Thus the likely price of the commodity for the year 2008 is Rs. 227.41 approx.

The graph of the actual and trend values is given below:



Advantages

- This is a mathematical method of measuring trend and as such there is no possibility of subjectiveness i.e. everyone who uses this method will get same trend line.
- The line obtained by this method is called the line of best fit.
- Trend values can be obtained for all the given time periods in the series.

Disadvantages

- Great care should be exercised in selecting the type of trend curve to be fitted i.e. linear, parabolic or some other type. Carelessness in this respect may lead to wrong results.
- The method is more tedious and time consuming.

- Predictions are based on only long term variations i.e trend and the impact of cyclical, seasonal and irregular variations is ignored.
- This method can not be used to fit the growth curves like Gompertz curve $\left(Y = Ka^{b^X}\right)$, logistic curve $\left(Y = \frac{1}{K + \quad}\right)$ etc.

14) Fit a staright line trend by the method of least square to the following data. Also find an estimate for the year 2000;

Year	: 1990	1991	1992	1993	1994	1995	1996	1997
Production (tonnes)	: 38	40	65	72	69	67	95	104

$(Y = 68.75 + 4.404X;)$

15) From the following data, calculate trend by 4 yearly moving average and find short-term oscillations:

Year	**Production(tonnes)**	**Year**	**Production(tonnes)**
1984	5	1990	9
1985	6	1991	10
1986	7	1992	9
1987	7	1993	10
1988	6	1994	11
1989	8	1995	11

UNIT-2

SEASONAL VARIATIONS

2.1 Introduction

Seasonal variations are regular and periodic variations having a period of one year duration. Some of the examples which show seasonal variations are production of colddrinks, which are high during summer months and low during winter season. Sales of sarees in a cloth store which are high during festival season and low during other periods.

The reason for determining seasonal variations in a time series is to isolate it and to study its effect on the size of the variable in the index form which is usually referred as seasonal index.

2.2 Measurement of seasonal variations:

The study of seasonal variation has great importance for business enterprises to plan the production schedule in an efficient way so as to enable them to supply to the public demands according to seasons.

There are different devices to measure the seasonal variations. These are

- Method of simple averages.
- Ratio to trend method
- Ratio to moving average method
- Link relative method.

Ratio to trend method:

This method is an improvement over the simple averages method and this method assumes a multiplicative model i.e

$$Y_t = T_t S_t C_t R_t$$

The measurement of seasonal indices by this method consists of the following steps.

- Obtain the trend values by the least square method by fitting a mathematical curve, either a straight line or second degree polynomial.
- Express the original data as the percentage of the trend values. Assuming the multiplicative model these percentages will contain the seasonal, cyclical and irregular components.
- The cyclical and irregular components are eliminated by averaging the percentages for different months (quarters) if the data are In monthly (quarterly), thus leaving us with indices of seasonal variations.
- Finally these indices obtained in step(3) are adjusted to a total of 1200 for monthly and 400 for quarterly data by multiplying them through out by a constant K which is

given by $K = \frac{1200}{Total\ of\ the indices}$ for monthly,

$K = \frac{400}{Total\ of\ the indices}$ for quarterly.

Advantages:

- It is easy to compute and easy to understand.
- Compared with the method of monthly averages this method is certainly a more logical procedure for measuring seasonal variations.
- It has an advantage over the ratio to moving average method that in this method we obtain ratio to trend values for each period for which data are available where as it is not possible in ratio to moving average method.

Disadvantages:

- The main defect of the ratio to trend method is that if there are cyclical swings in the series, the trend whether a straight line or a curve can never follow the actual data as closely as a 12- monthly moving average does. So a seasonal index computed by the ratio to moving average method may be less biased than the one calculated by the ratio to trend method.

Example 1:

Calculate seasonal indices by Ratio to moving average method from the following data.

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	30	40	36	34
2004	34	52	50	44
2005	40	58	54	48
2006	54	76	68	62
2007	80	92	86	82

Solution. For determining seasonal variation by ratio-to-trend method, first we will determine the trend for yearly data and then convert it to quarterly data.

CALCULATING TREND BY METHOD OF LEAST SQUARES

Year	*Yearly totals*	*Yearly average* Y	*Deviations from mid-year* X	XY	X^2	*Trend values*
2003	140	35	− 2	− 70	4	32
2004	180	45	− 1	− 45	1	44
2005	200	50	0	0	0	56
2006	260	65	+ 1	+ 65	1	68
2007	340	85	+ 2	+ 170	4	80
$N = 5$		$\Sigma Y = 280$		$\Sigma XY = 120$	$\Sigma X^2 = 10$	

The equation of the straight line trend is $Y = a + bX$.

$$a = \frac{\Sigma Y}{N} = \frac{280}{5} = 56 \qquad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{120}{10} = 12$$

Quarterly increment $= \frac{12}{4} = 3$.

Calculation of Quarterly Trend Values. Consider 2003, trend value for the middle quarter, *i.e.*, half of 2nd and half of 3rd is 32. Quarterly increment is 3. So the trend value of 2nd quarter is $32 - \frac{3}{2}$, *i.e.*, 30.5 and for 3rd quarter is $32 + \frac{3}{2}$, *i.e.*, 33.5. Trend value for the 1st quarter is 30.5 – 3, *i.e.*, 27.5 and of 4th quarter is 33.5 + 3, *i.e.*, 36.5. We thus get quarterly trend values as shown below :

TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	27.5	30.5	33.5	36.5
2004	39.5	42.5	45.5	48.5
2005	51.5	54.5	57.5	60.5
2006	63.5	66.5	69.5	72.5
2007	75.5	78.5	81.5	84.5

The given values are expressed as percentage of the corresponding trend values.
Thus for 1st Qtr. of 2003, the percentage shall be (30/27.5) × 100 = 109.09, for 2nd Qtr. (40/30.5) × 100 = 131.15, etc.

GIVEN QUARTERLY VALUES AS % OF TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	109.09	131.15	107.46	93.15
2004	86.08	122.35	109.89	90.72
2005	77.67	106.42	93.91	79.34
2006	85.04	114.29	97.84	85.52
2007	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
S.I. Adjusted	92.05	117.36	102.12	88.46

Total of averages = 92.77 + 118.28 + 102.92 + 89.15 = 403.12.
Since the total is more than 400 an adjustment is made by multiplying each average by $\frac{400}{403.12}$ and final indices are obtained.

Ratio to moving average method:

The ratio to moving average method is also known as percentage of moving average method and is the most widely used method of measuring seasonal variations. The steps necessary for determining seasonal variations by this method are

- Calculate the centered 12-monthly moving average (or 4-quarterly moving average) of the given data. These moving averages values will eliminate S and I leaving us T and C components.
- Express the original data as percentages of the centered moving average values.
- The seasonal indices are now obtained by eliminating the irregular or random components by averaging these percentages using A.M or median.
- The sum of these indices will not in general be equal to 1200 (for monthly) or 400 (for quarterly). Finally the adjustment is done to make the sum of the indices to a total of 1200 for monthly and 400 for quarterly data by

multiplying them through out by a constant K which is given by

$$K = \frac{1200}{Total\ of\ the\ indices}$$ for monthly

$$K = \frac{400}{Total\ of\ the\ indices}$$ for quarterly

Advantages:

- Of all the methods of measuring seasonal variations, the ratio to moving average method is the most satisfactory, flexible and widely used method.
- The fluctuations of indices based on ratio to moving average method is less than based on other methods.

Disadvantages:

- This method does not completely utilize the data. For example in case of 12-monthly moving average seasonal indices cannot be obtained for the first 6 months and last 6 months.

Illustration 24. Calculate seasonal indices by the ratio to moving average method, from the following data :

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2005	68	62	61	63
2006	65	58	66	61
2007	68	63	63	67

Solution. CALCULATION OF SEASONAL INDICES BY 'RATIO TO MOVING AVERAGE' METHOD

Year	Quarter	Given figures	4-figure moving totals	2-figure moving totals	4-figure moving average	Given figure as % of moving average
2005	I	68				
	II	62	→ 254			
	III	61	→ 251 → 505		63.186	96.54
	IV	63	→ 247 → 498		62.260	101.19
2006	I	65	→ 252 → 499		62.375	104.21
	II	58	→ 250 → 502		62.750	92.43
	III	66	→ 253 → 503		62.875	104.97
	IV	61	→ 258 → 511		63.875	95.50
2007	I	68	→ 255 → 513		64.125	106.04
	II	63	→ 261 → 516		64.500	97.67
	III	63				
	IV	67				

CALCULATION OF SEASONAL INDEX

Year	*Percentage to Moving Average*			
	1st Quarter	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2005	—	—	96.63	101.20
2006	104.21	92.43	104.97	95.50
2007	106.04	97.67	—	—
Total	210.25	190.10	201.60	196.70
Average	105.125	95.05	100.80	98.35
Seasonal Index	105.30	95.21	100.97	98.52

Arithmetic average of averages $= \frac{399.32}{4} = 99.83$

By expressing each quarterly average as percentage of 99.83, we will obtain seasonal indices.

Seasonal index of 1st Quarter $= \frac{105.125}{99.83} \times 100 = 105.30$

Seasonal index of 2nd Quarter $= \frac{95.05}{99.83} \times 100 = 95.21$

Seasonal index of 3rd Quarter $= \frac{100.80}{99.83} \times 100 = 100.97$

Seasonal index of 4th Quarter $= \frac{98.35}{99.83} \times 100 = 98.52$