PROBABILITY AND STATISTICS

UNIT I: Discrete Random variables and Distributions:

Introduction-Random variables- Discrete Random variable-Distribution function-Expectation-Moment Generating function-Moments and properties. Discrete distributions: Binomial, Poisson and Geometric distributions and their fitting to data.

UNIT II: Continuous Random variable and distributions:

Introduction-Continuous Random variable-Distribution function- Expectation-Moment Generating function-Moments and properties.
Continuous distribution: Uniform, Exponential and Normal distributions, Normal approximation to Binomial distribution -Weibull, Gamma distribution.

UNIT III: Sampling Theory:

Introduction - Population and samples- Sampling distribution of means (s known)-Central limit theorem- t-distribution- Sampling distribution of means (s unknown)- Sampling distribution of variances - χ^2 and F-distributions- Point estimation- Maximum error of estimate - Interval estimation.

UNIT IV: Tests of Hypothesis:

Introduction –Hypothesis-Null and Alternative Hypothesis- Type I and Type II errors –Level of significance - One tail and two-tail tests- Tests concerning one mean and proportion, two means- Proportions and their differences- ANOVA for one-way and two-way classified data.

UNIT V: Curve fitting and Correlation:

Introduction - Fitting a straight line –Second degree curve-exponential curve-power curve by method of least squares-Goodness of fit. Correlation and Regression – Properties.

UNIT VI: Statistical Quality Control Methods:

Introduction - Methods for preparing control charts – Problems using x-bar, p, R charts and attribute charts.

Unit-I

Random Variables

A *random* variable, usually written *X*, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables

1. Discrete Random variable
2. Continuous Random variable

Discrete Random Variables

Discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,........ Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

Probability Mass function:

If X is a one-dimensional discrete random variable taking at most a countably infinite number of values $x_1, x_2, ...$then its probabilistic behavior at each real point is described by a function called Probability mass function (or Discrete density function) which is defined below

$P(x) = P(X = x_i) = p_i$ is called a probability mass function

Suppose a random variable X may take k different values, with the probability that X = xi defined to be P(X = xi) = pi. The probabilities pi must satisfy the following:

1: 0 < pi < 1 for each i
2: p1 + p2 + ... + pk = 1.

Example

Suppose a variable X can take the values 1, 2, 3, or 4.
The probabilities associated with each outcome are described by the following table:

Outcome	1	2	3	4
Probability	0.1	0.3	0.4	0.2

The probability that X is equal to 2 or 3 is the sum of the two probabilities:

P(X = 2 or X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7.

Similarly, P(X >1) = 1 - P(X = 1) = 1 - 0.1 = 0.9,

Distribution function:

The distribution function also called the Cumulative distribution function(CDF) or Cumulative frequency function, describes the probability that variable X takes on a value less than or equal to a number x. The distribution function is

Sometimes also denoted by F(x)

Discrete Distribution function is

$$P(X \le x) = \sum_{X \le x} P(x)$$

Properties of Distribution function:

Property 1: If F is the distribution function of the random variable X and if $a < b$, then

$$P(a < X \le b) = F(b) - F(a)$$

Proof: The events $a < X \le b$ and $X \le a$ are disjoint and their union is the event $X \le b$, Hence by addition theorem of probability.

$$P(a < X \le b) + P(X \le a) = P(X \le b)$$
$$P(a < X \le b) = P(X \le b) - P(X \le a)$$
$$= F(b) - F(a)$$

b)

$$P(a < X < b) = F(b) - F(a) - P(X = b)$$

c)

$$P(a \le X < b) = F(b) - F(a) - P(X = b) + P(X = a)$$

d)

$$P(a \le X \le b) = F(b) - F(a) + P(X = a)$$

Property 2:

If F is distribution function of one dimensional random variable X, then

$i)\ 0 \le F(x) \le 1$

$ii)\ \mathrm{F}(x) \le F(y)$ if x < y

Property 3:

If F is the Distribution function of one dimension random variable X, then

$i)$

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

$ii)$

$$F(\infty) = \lim_{x \to \infty} F(x) = 1$$

x	1	2	4	6
$P(x)$	0.2	0.5	0.1	0.2

$$F(x) = P(X \le x) = \sum_{X \le x} P(x)$$

$$F(1) = P(X \le 1) = \sum_{X \le 1} P(x)$$

$$= P(X = 1)$$

$$= 0.1$$

$$F(2) = P(X \le 2) = \sum_{X \le 2} P(x)$$

$$= P(X = 1) + P(X = 2)$$

$$= 0.2 + 0.5$$

$$= 0.7$$

$$F(4) = P(X \le 4) = \sum_{X \le 4} P(x)$$

$$= P(X = 1) + P(X = 2) + P(X = 4)$$

$$= 0.2 + 0.5 + 0.1$$

$$= 0.8$$

$$F(6) = P(X \le 6) = \sum_{X \le 6} P(x)$$

$$= P(X = 1) + P(X = 2) + P(X = 4) + P(X = 6)$$

$$= 0.2 + 0.5 + 0.1 + 0.2$$

$$= 1.0$$

Mathematical Expectation:

Once we have constructed the probability distribution for a random variable, we often want to compute the mean or Expected value of the random variable. The expected value of a discrete random variable is a weighted average of all possible values of the random variable, where the weights are the probabilities associated with the corresponding values. The mathematical expression for computing the expected value of a discrete random variable X with probability mass function P(x) is given below

$$E(x) = \sum_{x} xP(X = x)$$

Properties of Expectation

Addition Theorem of Expectation

If X and Y are random variable then

$$E(X+Y) = E(X) + E(Y)$$

Proof:

$$\begin{aligned}
E(X+Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_j)p_{ij}^{X,Y} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i p_{ij}^{X,Y} + y_j p_{ij}^{X,Y}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij}^{X,Y} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij}^{X,Y} \\
&= \sum_{i=1}^{n} x_i \cdot \left(\sum_{j=1}^{m} p_{ij}^{X,Y}\right) + \sum_{j=1}^{m} y_j \cdot \left(\sum_{i=1}^{n} p_{ij}^{X,Y}\right) \\
&\qquad \text{because we can take } x_i \text{ out of } \textstyle\sum_{j=1}^{m} \text{ because } x_i \text{ does not depend on } j\text{'s} \\
&= \sum_{i=1}^{n} x_i \cdot p_i^X + \sum_{j=1}^{m} y_j \cdot p_j^Y \\
&\qquad \text{because } p_i^X = \textstyle\sum_{j=1}^{m} p_{ij}^{X,Y} \text{ and } p_j^Y = \sum_{i=1}^{n} p_{ij}^{X,Y} \\
&= E(X) + E(Y)
\end{aligned}$$

The mathematical Expectation of the sum of n random variables is equal to the sum of their expectation, provided all the expectations exist.

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Property 2: Multiplication Theorem of Expectation

If the X and Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

The mathematical expectation of the product of a number of independent random variables is equal to the product of their expectations. Symbolically if

$$E(X_1 \times X_2 \times \ldots\ldots \times X_n) = E(X_1) \times E(X_2) \times \ldots \times E(X_n)$$

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$$

Property 3:

If X is random variable and 'a' is constant, then

$$i)\,\mathrm{E}\left(a\psi(X)\right) = aE\left(\psi(X)\right)$$

$$ii)\,\mathrm{E}\left(\psi(X) + a\right) = E\left(\psi(X)\right) + a$$

Property 4:

If X is random variable and 'a' and 'b' are constant, then

$$E(aX + b) = aE(X) + b$$

Property 5:

Expectation of Linear Combination of Random variables:

Let $X_1, X_2, X_3, \ldots\ldots X_n$ be any n random variables and if $a_1, a_2, a_3, \ldots\ldots a_n$ are any n constants, then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

Provide all the expectation exist

Property 6:

If X ≥ 0 then E(X) ≥ 0

Property 7:

If X and Y are two random variables such that Y ≤ X, then

E(Y) ≤ E(X)

Property 8:

$$|E(x)| = E|x|$$

Example:

Repair Costs for a Particular Machine are represented by the following Probability function

x	\$50	\$200	\$350
$P(x)$	0.3	0.2	0.5

What is Expected value of X

And also find the Variance of x

$$E(x) = \sum_x xP(x)$$

$$= (50 \times 0.3) + (200 \times 0.2) + (350 \times 0.5)$$

$$= 15 + 40 + 175$$

$$= \$230$$

$$Variance:$$

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

$$E(x^2)=\sum_x x^2P(x)$$

$$=(50^2\times 0.3)+(200^2\times 0.2)+(350^2\times 0.5)$$

$$=750+8000+61250$$

$$=70000$$

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

$$=70000-(230)^2$$

$$=70000-52900$$

$$=\$17100$$

Moment Generating Function:

The moment generating function (M.G.F) of a random variable X (about origin having the probability function $f(x)$ is given by

$$M_x(t)=E(e^{tx})$$

In Discrete probability function

$$M_x(t)=\sum_x e^{tx}f(x)$$

Moments of MGF

$$\mu_r'=E(x^r)=\left[\frac{d^r}{dt^r}M_x(t)\right]_{t=0}$$

Properties of Moment Generating Funciton

Property 1: $\mathrm{M}_{cx}(t)=M_x(ct)$, where c is a contant

Property 2: The moment generating function of the sum of a number of independent random variables is equal to the product of their respective Moment generating functions

$$M_{X_1+X_2+X_3+....+X_n}(t) = M_{X_1}(t)M_{X_2}(t)M_{X_3}(t).....M_{X_n}(t)$$

$$Mean(\mu_1) = \mu_1^{'} = \left[\frac{d}{dt}M_X(t)\right]_{t=0}$$

$$Variance(\mu_2) = \mu_2^{'} - \left(\mu_1^{'}\right)^2$$

Binomial Distribution

The Probability distribution of the number of success, so obtained is called the Binomial Probability distribution, for the obvious reason that the probabilities of 0, 1, 2,……., n success are the successive terms of the binomial expansion $(q+p)^n$.

Definition: A random variable X is said to follow binomial distribution if it assumes only non-negative and its probability mass function is given by

Let X be a discrete random variable it assume only Non – negative values. Then the probability mass function is defined by

$$P(X = x) = {}^nC_r p^r q^{n-r} \qquad X = 1,2,3......n$$

Where n = No. of trials

x = No. of success

p = Probability of Success

q = Probability of failure

Conditions/Assumptions/ Rules:

1. The no. of trials must be fixed that "n" is finite.
2. The trails are independent of each other.
3. Probability of success (p) value must be constant.
4. The no. of trials must be independent to each other.

Mean of Binomial Distribution

$E(x) = np$

Variance of Binomial Distribution

$V(x) = npq$

Standard deviation=$\sqrt{npq}$

Moment Generation Function of Binomial distribution:

$$M_X(t) = E\left(e^{tx}\right)$$
$$= \sum_{x=0}^{n} \binom{n}{x} \left(pe^t\right)^x q^{n-x}$$
$$= \left(q + pe^t\right)^n$$

Characteristic of the Binomial Distribution:

$$\varphi_X(t) = E\left(e^{itx}\right)$$
$$= \sum_{x=0}^{n} e^{itx} \binom{n}{x} p^x q^{n-x}$$
$$= \sum_{x=0}^{n} \binom{n}{x} \left(pe^t\right)^x q^{n-x}$$

Cumulate generation function of the Binomial distribution is given by

$$K_X(t) = \log\left(M_X(t)\right)$$
$$= n\log\left(q + pe^t\right)$$

Example :

If a coin is tossed 10 times what is the probability that getting head is 2 times

$$P(X=x)={}^{n}C_{x}p^{x}(1-p)^{n-x}$$

$$n=10$$

$$x=2$$

$$p=0.5$$

$$P(X=2)={}^{10}c_{2}(0.5)^{2}(1-0.5)^{10-2}$$

$$=\frac{10!}{2!(10-2)!}(0.5)^{10}$$

$$=45(0.0009765625)$$

$$=0.0439$$

Poisson distribution:

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

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

$$x=0,1,2,......,\infty$$

Moments of Poisson distribution:

$$\mu_r{}'=\mathrm{E}\left(x^{r}\right)$$

$$Mean=\mu_1{}'$$

$$=E(x)$$

$$=\lambda$$

$$\text{Variance}=\mu_2{}'-(\mu_1{}')^{2}$$

$$=\lambda$$

Moment Generating function Poisson distribution:

$$M_x(t) = e^{-\lambda\left(e^t - 1\right)}$$

Characteristic Function of Poisson distribution:

$$\phi_X(t) = e^{-\lambda\left(e^{-it} - 1\right)}$$

Example :
Average number of phone calls on the switch board of the company is 2.5 per minute between 10 am to 1pm
What is the probability that the number of phone calls is 1 per minute
Given that

$\lambda = 2.5$ *per minute*

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

$$P(X = 1) = \frac{e^{-2.5}[2.5]^1}{1!}$$

$= 0.082085 \times 2.5$

$= 0.2052$

Geometric Distribution

This distribution represents the number of failures before you get a success in a series of Bernoulli trials. This discrete probability distribution is represented by the probability density function:

Probability mass funciton of Geometric distribution is

$$P(X = x) = (1 - p)^{x-1} p$$

Where p is probability of success

Example:

Sample question: If your probability of success is 0.2, what is the probability you meet an independent voter on your third try?
Inserting 0.2 as p and with X = 3, the probability density function becomes:

$f(x) = (1 - p)x - 1*p$
$P(X=3) = (1 - 0.2)3 - 1(0.2)$
$P(X=3) = (0.8)2*0.2 = 0.128.$

Unit-II

Continuous Random variable

A continuous random variable is a random variable where the data can take infinitely many values. For example, a random variable measuring the time taken for something to be done is continuous since there are an infinite number of possible times that can be taken.

Probability Density function:

If X is a one-dimensional Continuous random variable,which is defined below

$f(x)$ is called a probability density function. If it satisfies these conditions

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

$$0 \leq f(x) \leq 1$$

Distribution function:

The distribution function also called the Cumulative distribution function(CDF) or Cumulative frequency function, describes the probability that variable X takes on a value less than or equal to a number x. The distribution function is

Sometimes also denoted by F(x)

Continuous Distribution function is

$$F(x) = P(X \leq x) = \int_{X \leq x} f(x)$$

Mathematical Expectation:

Once we have constructed the probability distribution for a random variable, we often want to compute the mean or Expected value of the random variable. The expected value of a discrete random variable is a weighted average of all possible values of the random variable, where the weights are the probabilities associated with the corresponding values. The mathematical expression for computing the expected value of a discrete random variable X with probability mass function P(x) is given below

$$E(x)=\int xf(x)dx$$

Properties of Expectation

Addition Theorem of Expectation

If X and Y are random variable then

$$E(X+Y)=E(X)+E(Y)$$

Proof:

$$E(X+Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(xy)dxdy$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x)f(xy)dxdy+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y)f(xy)dxdy$$

$$=\int_{-\infty}^{\infty}x\left[\int_{-\infty}^{\infty}f(xy)dy\right]dy+\int_{-\infty}^{\infty}y\left[\int_{-\infty}^{\infty}f(xy)dx\right]dy$$

$$\left(\begin{array}{l}\because f(x)=\int_{-\infty}^{\infty}f(xy)dy \\ f(y)=\int_{-\infty}^{\infty}f(xy)dx\end{array}\right)$$

$$=\int_{-\infty}^{\infty}xf(x)dy+\int_{-\infty}^{\infty}yf(y)dy$$

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

The mathematical Expectation of the sum of n random variables is equal to the sum of their expectation, provided all the expectations exist.

$$E\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}E(X_i)$$

Property 2: Multiplication Theorem of Expectation

If the X and Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

Proof:

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)\, f(xy) dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)\, f(x)\, f(y) dxdy \quad \left(\because f(xy) = f(x)\, f(y)\right)$$

$$= \left[\int_{-\infty}^{\infty} xf(x)dx\right]\left[\int_{-\infty}^{\infty} yf(y)dy\right]$$

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

The mathematical expectation of the product of a number of independent random variables is equal to the product of their expectations. Symbolically if

$$E(X_1 \times X_2 \times ...... \times X_n) = E(X_1) \times E(X_2) \times ... \times E(X_n)$$

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$$

Property 3:

If X is random variable and 'a' is constant, then

$$i)\,\mathrm{E}\left(a\psi(X)\right) = aE\left(\psi(X)\right)$$

$$ii)\,\mathrm{E}\left(\psi(X) + a\right) = E\left(\psi(X)\right) + a$$

Property 4:

If X is random variable and 'a' and 'b' are constant, then

$$E(aX+b)=aE(X)+b$$

Property 5:

Expectation of Linear Combination of Random variables:

Let $X_1, X_2, X_3, \ldots\ldots X_n$ be any n random variables and if $a_1, a_2, a_3, \ldots\ldots a_n$ are any n constants, then

$$E\left(\sum_{i=1}^{n} a_i X_i\right)=\sum_{i=1}^{n} a_i E(X_i)$$

Provide all the expectation exist

Property 6:

If X ≥ 0 then E(X) ≥ 0

Property 7:

If X and Y are two random variables such that Y ≤ X, then

E(Y) ≤ E(X)

Property 8:

$$|E(x)|=E|x|$$

Moment Generating Function:

The moment generating function (M.G.F) of a random variable X (about origin having the probability function $f(x)$ is given by

$$M_x(t) = E\left(e^{tx}\right)$$

In Discrete probability function

$$M_x(t) = \int e^{tx} f(x)\,dx$$

Moments of MGF

$$\mu_r^{'} = E\left(x^r\right) = \left[\frac{d^r}{dt^r} M_x(t)\right]_{t=0}$$

Properties of Moment Generating Funciton

Property 1: $\mathrm{M}_{cx}(t) = M_x(ct)$, where c is a contant

Property 2: The moment generating function of the sum of a number of independent random variables is equal to the product of their respective Moment generating functions

$$M_{X_1+X_2+X_3+....+X_n}(t) = M_{X_1}(t) M_{X_2}(t) M_{X_3}(t) .....M_{X_n}(t)$$

$$Mean(\mu_1) = \mu_1^{'} = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$$

$$Variance(\mu_2) = \mu_2^{'} - \left(\mu_1^{'}\right)^2$$

Continuous Distribution

Uniform distribution

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability.

This distribution is defined by two parameters, a and b where a is the minimum and b is the maximum.The distribution is written as U(a,b).
The general formula for the probability density function (pdf) for the uniform distribution is:

$$f(x) = \frac{1}{b-a}, \ a \le x \le b$$

Mean and Variance of Uniform distribution:

Mean(E(x)):

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \int_a^b x \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \int_a^b x dx \\
&= \frac{1}{b-a} \frac{b^2-a^2}{2} \\
&= \frac{1}{b-a} \frac{(b+a)(b-a)}{2} \\
&= \frac{(b+a)}{2}
\end{aligned}
$$

Variance (V(x)):

$$\begin{aligned} V(X) &= E[(X - E(X))^2] \\ &= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a} dx \end{aligned}$$

Now put $z = \frac{x - \frac{a+b}{2}}{b-a}$ hence $dx = (b-a)dz$,

$$\begin{aligned} V(X) &= (b-a)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} z^2 dz \\ &= (b-a)^2 \left[\frac{z^3}{3}\right]_{-\frac{1}{2}}^{\frac{1}{2}} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Moment Generation for Uniform distribution

$$\begin{aligned} M(t) &= E\left[e^{tX}\right] \\ &= \int_a^b e^{tx} \frac{1}{b-a} dx \\ &= \frac{1}{b-a}\left[\frac{1}{t}e^{tx}\right]_a^b \\ &= \frac{e^{bt} - e^{at}}{t(b-a)} \end{aligned}$$

Example

Suppose the random variable x has a uniform distribution on the interval $[-2, 4]$. Compute the following probability:

Given that a =-2 and b =4

$$f(x) = \frac{1}{b-a}$$

$$= \frac{1}{4-(-2)}$$

$$= \frac{1}{6}$$

$$P(X>2) = \int_2^4 f(x)dx$$

$$= \int_2^4 \left(\frac{1}{6}\right)dx$$

$$= \frac{1}{6}\int_2^4 (1)dx$$

$$= \frac{1}{6}[x]_2^4$$

$$= \frac{4-2}{6}$$

$$= \frac{2}{6}$$

$$= 0.33$$

Normal Distribution

Normal Distribution:

If X is continuous random variable and is said to follow normal distribution then the probability density function is given by

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

$$-\infty < x < \infty$$

$$-\infty < \mu < \infty$$

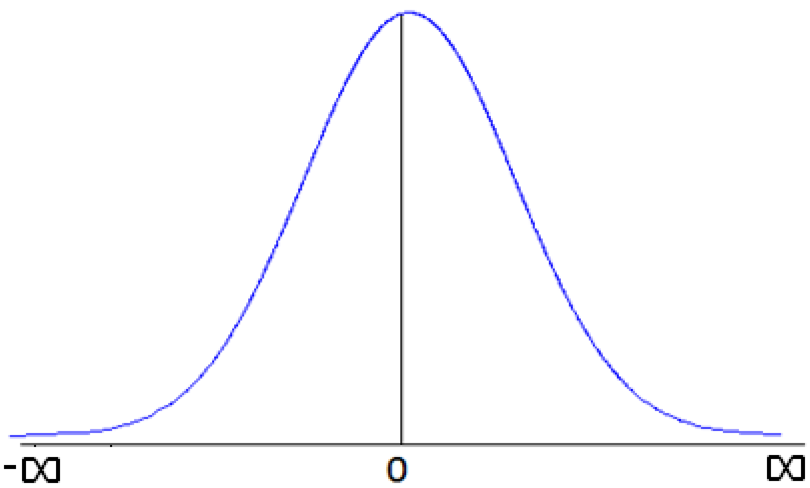
$$0 < \sigma < \infty$$

Properties/Assumptions:

1. The random variable is must be continuous
2. Normal distribution mean is
3. Normal distribution variance
4. Normal distribution standard deviation
5. Mean and variance of the normal distribution are called parameters.
6. If X follows standard normal distribution then probability dencity function is given by Where 'Z' is the standard normal variation

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

Normal curve is symmetrical curve.



Importance:

Normal distribution plays a very important role in statistical theory.

1. Most of the distributions (Binomial, Poisson etc) can be approximated by normal distribution.
2. Many of the sampling distributions (chi-square, t, F distributions) tends to normal distributions for large sample theory.
3. Normal distribution finds the large applications in statistical quality control in industry for setting control limits.

Moment generating function of Normal distribution:

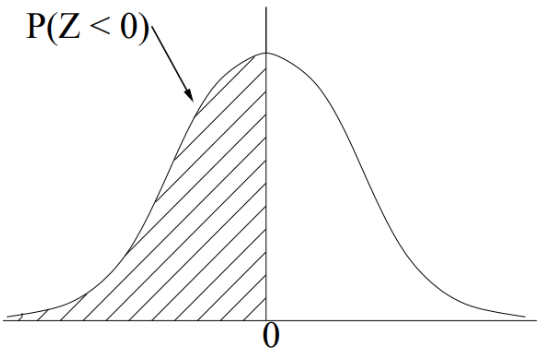
$$\begin{aligned}
M(t) &= E\left[e^{tX}\right] \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}\left(-2tx\sigma^2+(x-\mu)^2\right)} dx \\
&= e^{\mu t+\frac{1}{2}t^2\sigma^2}\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2} dx\right] \\
&= e^{\mu t+\frac{1}{2}\sigma^2 t^2} \qquad -\infty < t < \infty
\end{aligned}$$

Calculating probabilities from the Normal distribution

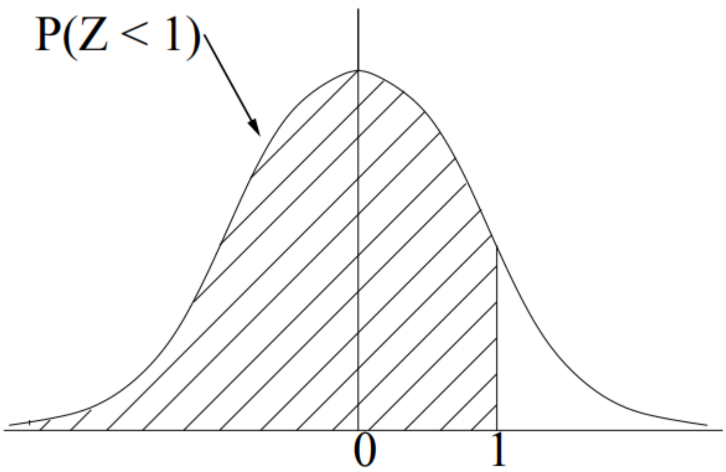
For a discrete probability distribution we calculate the probability of being less than some value x, i.e. $P(X < x)$, by simply summing up the probabilities of the values less than x.

For a continuous probability distribution we calculate the probability of being less than some value x, i.e. P(X < x), by calculating the area under the curve to the left of x.

Suppose we find $P(Z < 0)$

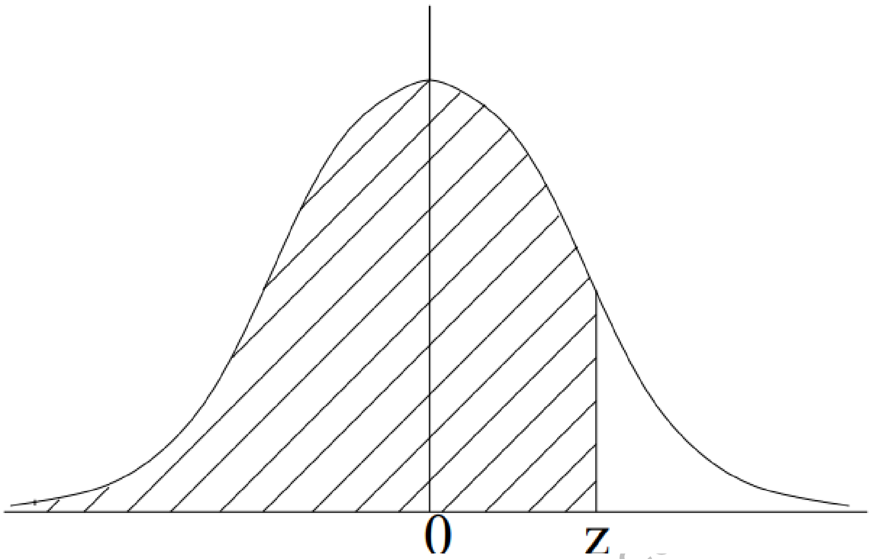


What about $P(Z < 1)$



Calculating this area is not easy and so we use probability tables. Probability tables are tables of probabilities that have been calculated on a computer. All we have to do is identify the right probability in the table and copy it down! Only one special Normal distribution, N(0, 1), has been tabulated. The N(0, 1) distribution is called the standard Normal distribution

The tables allow us to read off probabilities of the form $P(Z < z)$



The Z- table is:

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	0.5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	0.5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	0.6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	0.6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	0.7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	0.7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	0.7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	0.8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	$\boxed{0.8413}$	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	0.8643	8665	8686	8708	8729	8749	8770	8790	8810	8830

Example:

Suppose we know that the birth weight of babies is Normally distributed with mean 3500g and standard deviation 500g. What is the probability that a baby is born that weighs less than 3100g?

Given that

$$Mean(\mu) = 3500$$

$$Standard\ Deviation(\sigma) = 500$$

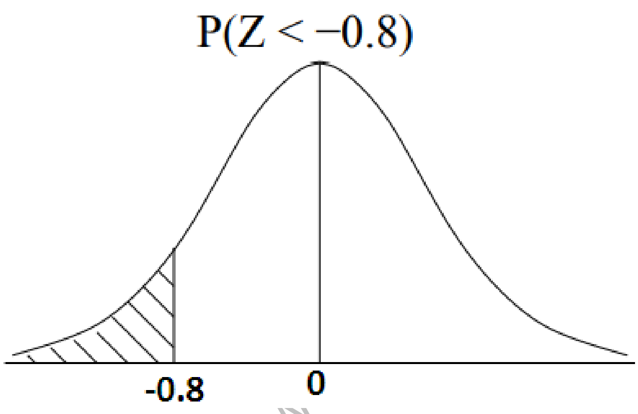
$$P(X < 3100) = P\left(\frac{x-\mu}{\sigma} < \frac{3100-3500}{500}\right)$$

$$= P(Z < -0.8)$$

$$= 1 - P(Z < 0.8)$$

$$= 1 - 0.7881$$

$$= 0.2119$$



Normal Approximation to a Binomial

The normal distribution is used to approximate the binomial distribution when it would be impractical to use binomial distribution to find probability.

If mean $np \geq 5$ *and* $nq \geq 5$, Then the binomial distribution of x is approximately normally distributed with

$$Mean(\mu) = np$$

$$Standard\ Deviaiton(\sigma) = \sqrt{npq}$$

Example:

Previous research shows that 65% of murders are committed with a firearm. if 150 murders are randomly selected, use the normal approximation to the binomial to determine what is the probability that 100 or more murder are committed with a firearm?

Here n=150

$$p = 0.65$$

$$\Rightarrow q = 0.35$$

$$E(X) = np = 97.5$$

$$Var(X) = npq = 52.5$$

Since np > 10,we can use normal approximation for obtaining probabilities.

$$P(X \geq 100) = P\left(\frac{X - np}{\sqrt{npq}} \geq \frac{(100 - 97.5)}{\sqrt{52.5}}\right)$$

$$= P(z \geq 0.345)$$

$$= 1 - P(z \leq 0.345)$$

$$= 1 - 0.63495$$

$$= 0.36505$$

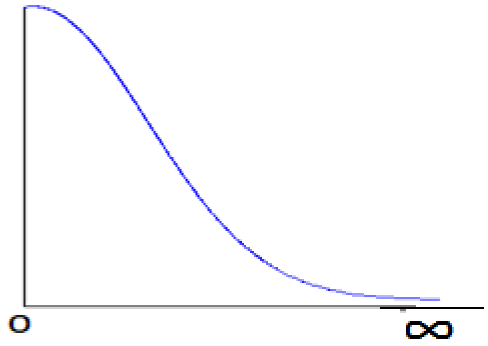
Exponential Distribution

A continuous distribution variable 'X' is said to follow exponential distribution then the probability distribution then the probability density function is given by

$$f(x) = \lambda e^{-\lambda x}$$
$$0 \le x \le \infty$$

Assumptions:

1. Exponential distribution Mean is $\frac{1}{\lambda}$ then exponential distribution variance is $\frac{1}{\lambda^2}$. Then exponential distribution standard deviation is $\frac{1}{\lambda}$
2. Mean and standard deviation of the exponential distribution are the same.
3. It assume only non – negative or positive values
4. Exponential distribution curve is right side curve or positively skewed.



Moment Generation for Exponential distribution:

$$\begin{aligned} M(t) &= E\left[e^{tX}\right] \\ &= \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,dx \\ &= \lambda\int_0^\infty e^{(t-\lambda)x}\,dx \\ &= \frac{\lambda}{t-\lambda}\left[e^{(t-\lambda)x}\right]_0^\infty \\ &= \frac{\lambda}{\lambda-t} \end{aligned}$$

Weibull distribution

The Weibull distribution is a continuous probability distribution named after Swedish mathematician Waloddi Weibull. He originally proposed the distribution as a model for material breaking strength, but recognized the potential of the distribution in his 1951 paper A Statistical Distribution Function of Wide Applicability. Today, it's commonly used **to** assess product reliability**,** analyze life data and model failure times**.** The Weibull can also fit a wide range of data from many other fields, including: biology, economics, engineering sciences, and hydrology

Probability density function Weibull distribution:

$$f(x) = \left\{ \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} e^{-\left(x/\lambda \right)^{k}}, X \geq 0 \right.$$

Distribution function of Weibull distribution

$$F(x) = 1 - e^{-\left(x/\lambda \right)^{k}}$$

Mean of Weibul Distribution:

$$\mathrm{E}(X) = \lambda \Gamma\left(1 + \frac{1}{k}\right)$$

Variance of Weibul Distribution:

$$\operatorname{var}(X) = \lambda^2 \left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right) \right)^2 \right].$$

Gamma Distribution

The gamma distribution is another widely used distribution. Its importance is largely due to its relation to exponential and normal distributions. Here, we will provide an introduction to the gamma distribution. In Chapters 6 and 11, we will discuss more properties of the gamma random variables. Before introducing the gamma random variable, we need to introduce the gamma function.

Gamma function: The gamma function is shown by $\Gamma(x)$ is an extension of the factorial function to real (and complex) numbers. Specifically, if n∈ {1,2,3,...} then

$$\Gamma(n) = (n-1)!$$

Probability density functions of Gama distribution:

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\, x^{\alpha-1} e^{-\frac{x}{\beta}} & for \quad x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

Unit-III

Sampling Theory

Population: The aggregate of all units pertaining to a study is called population or universe.

Sample: Set of data drawn from the population is called sample. The process of selection a sample from the population is called sampling.

Example: Suppose there are 3000 students in a college and 250 students are selected in order to estimate the average height of students. This number of 250 students constitutes a sample and the total number of 3000 students is population.

Population size (N) is finite or sometimes infinite and sample size (n) is always finite

Parameter: Population constants [Population Mean μ and population variance "σ^2"] are called parameters.

Statistic: Sample constant [sample Mean $\bar{x}$ and sample variance s^2] are called statistic

In practice parameter values are not known and the estimates based on the sample values are generally used. Thus, statistic which may be recorded as an estimate of parameter obtained from the sample.

Sampling distribution of a statistics

If we draw sample of size 'n' from a given finite population of size 'N' then the total no. of possible samples is called sampling distribution.

Example:

$$^{N}c_n = \frac{N!}{(\text{N-n})!\text{n}!} = k$$

$$^{4}c_2 = \frac{4!}{(4\text{-}2)!2!}$$
$$= \frac{4\times3\times2\times1}{2\times1\times2\times1}$$
$$^{4}c_2 = 6$$

$$(1,2,3,4) = (1,2),(1,3),(1,4),(2,3),(2,4),(3,4)$$

Sampling Distribution of mean$_{(\sigma - Known)}$:

Suppose we draw all possible samples of size n from a population of size N. Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

We know the following about the sampling distribution of the mean. The mean of the sampling distribution ($\mu_{\bar{x}}$) is equal to the mean of the population (μ). And the standard error of the sampling distribution ($\sigma_{\bar{x}}$) is determined by the standard deviation of the population (σ), the population size (N), and the sample size (n). These relationships are shown in the equations below:

$$Mean\left(\mu_{\bar{x}}\right) = \mu$$

$$Standard\ deviation\left(\sigma_{\bar{x}}\right) = \frac{\sigma}{\sqrt{n}}$$

The Central Limit Theorem:

For samples of size 30 or more, the sample mean is approximately normally distributed, with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, where n is the sample size. The standard deviation of the sampling distribution of the mean $\sigma_{\bar{x}}$ is called the standard error of the mean. It is designated by the symbol: The larger the sample size, the better the approximation. Given a population with a finite mean μ and a finite non-zero variance σ2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ2/N as N, the sample size, increases.

Example: The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

Given that

n=100

The sample mean $\bar{X}$ *has mean* $\mu_{\bar{x}} = \mu = 2.61$

and Standard deviation $\left(\sigma_{\bar{x}}\right) = \sigma / \sqrt{n}$

$$= \frac{0.5}{\sqrt{100}}$$

$$= 0.05$$

$$P\left(2.51 \le \bar{X} \le 2.71\right) = P\left(\frac{2.51-2.61}{0.05} \le \frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} \le \frac{2.71-2.61}{0.05}\right)$$

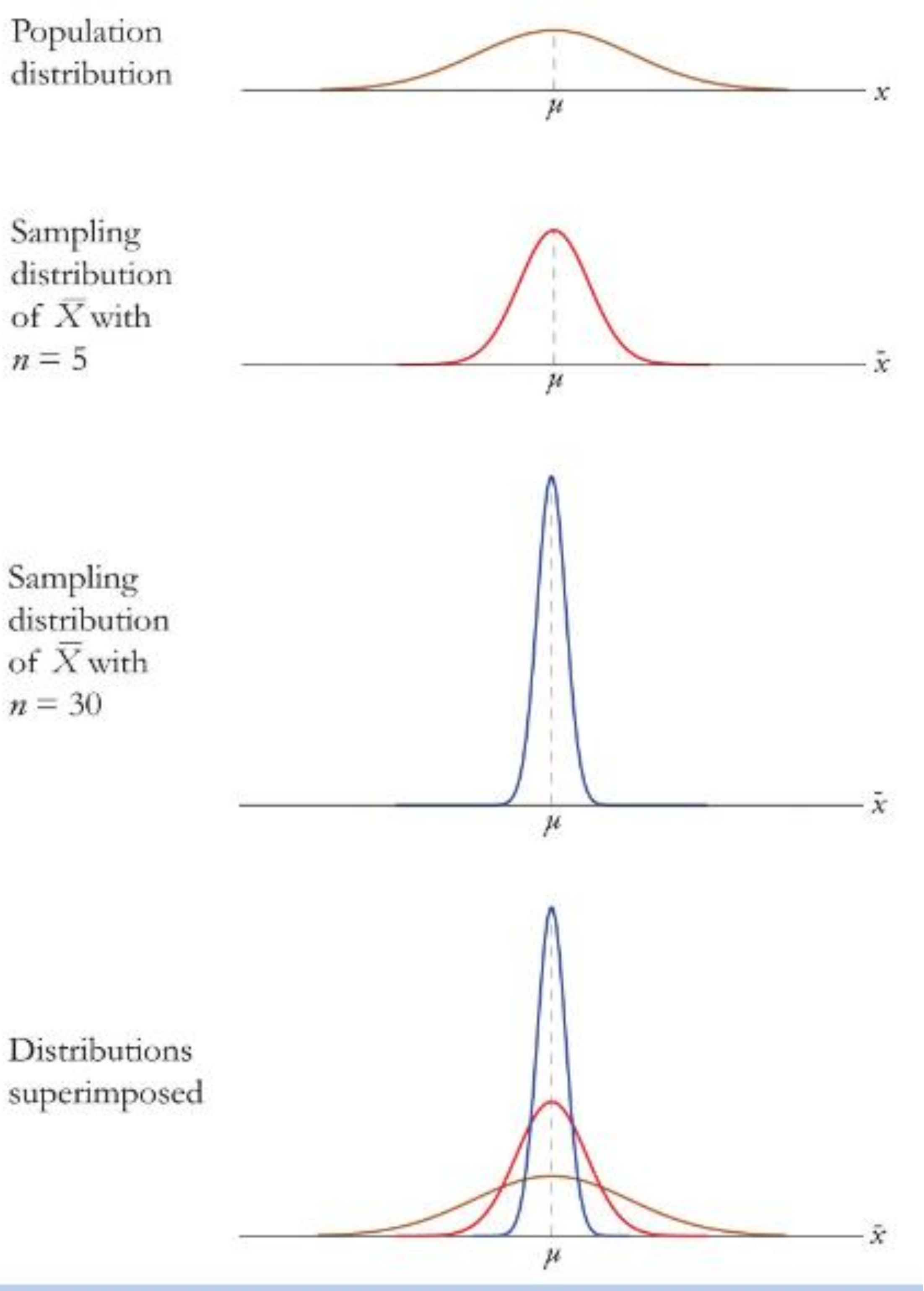
$$= P\left(-2 \le Z \le 2\right)$$

$$= P\left(Z \le 2\right) - P\left(Z \le -2\right) \quad \begin{pmatrix} \because \textit{Using Z table} \\ P\left(Z \le 2\right) = 0.9772 \\ P\left(Z \le -2\right) = 0.0228 \end{pmatrix}$$

$$= 0.9772 - 0.0228$$

$$= \boxed{0.9544}$$

The Central Limit Theorem is illustrated for several common population distributions.



T Distribution (Sampling distribution of Mean $(\sigma - Unknown)$)

Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. It was developed by William Sealy Gosset under the pseudonym Student.

The probability dencity function of t distribution:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Assumption of t distribution:

1. T-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.
2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The fourth assumption is a reasonably large sample size is used. A larger sample size means the distribution of results should approach a normal bell-shaped curve.
5. The final assumption is homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

Sampling Distribution of mean$_{(\sigma - UnKnown)}$:

Suppose we draw all possible samples of size n from a population of size N. Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

$$Mean\left(\mu_{\bar{x}}\right) = \bar{x}$$

$$Standard\ deviation\ of\ sample\ mean\left(\sigma_{\bar{x}}\right) = \frac{s}{\sqrt{n}}$$

Where n is sample size a

S is the Sample standard deviation

Example:

Find the t score for a sample size of 16 taken from a population with mean 10 when the sample mean is 12 and sample standard deviation is 1.5

$$Given\ that$$
$$n = 16$$
$$\mu = 10$$
$$\bar{x} = 12$$
$$s = 1.5$$
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$
$$= \frac{12 - 10}{1.5/\sqrt{16}}$$
$$= \frac{2}{0.375}$$
$$= 5.33$$

Properties of t distribution:

1. The distribution has mean 0
2. The distribution is symmetric about the mean
3. The variance is equal to $\frac{df}{df-2}$
4. The variance always greater than 1 but approaches 1 when df gets bigger

F Distribution

It is defined in terms of ratio of the variances of two normally distributed populations. So, it sometimes also called variance ratio. It is used for comparing the variances of two populations

Probability density function of F-distribution:

$$f(F) = Y_o \frac{F_{\upsilon_1 2-1}}{\left[1+\frac{\upsilon_1}{\upsilon_2}\right](\upsilon_1+\upsilon_2)/2}$$

Where Y0 is a constant depending on the values of v1 and v2 such that the area under the curve is unity.

Properties of F-distribution

1. It is positively skewed and its skewness decreases with increase in v1 and v2 .

2. Value of F must always be positive or zero,ν since variances are squares. So its value lies between 0 and ∞.

3. Mean and variance of F-distribution:

$$Mean = \frac{v_2}{v_2 - 2},\ for\ v_2 > 2$$

$$Variance = \frac{2v_2^{\ 2}\left(v_1 + v_2 - 2\right)}{v_1\left(v_2 - 2\right)^2\left(v_2 - 4\right)},\ for\ v_2 > 4$$

4. Shape of F-distribution depends upon the number of degrees of freedom.

Testing of hypothesis for equality of two variances

It is based on the variances in two independently selected random samples drawn from two normal populations.

Null hypothesis

$$H_0 : \sigma_1^{\ 2} = \sigma_2^{\ 2}$$

$$F = \frac{s_1^{\ 2} / \sigma_1^{\ 2}}{s_2^{\ 2} / \sigma_2^{\ 2}},\ which\ reduces\ to\ F = \frac{s_1^{\ 2}}{s_2^{\ 2}}$$

Degrees of freedom v1 and v2.

Find table value using v1 and v2ν .

If calculated F value exceeds table F value, null hypothesis is rejected.

Example:

A sample of 10 lots from source A yields a variance of 225 and a sample of 11 lots from source B yields a variance of 200. Is it likely that the variance of source A is significantly greater than the variance of source B at α = 0.01?

Solution

$H_0 : \sigma_1^{\ 2} = \sigma_2^{\ 2}$ i.e. the variances of source A and that of source B are the same. The F statistic to be used here is

$$F = \frac{s_1^2}{s_2^2}$$

Where s_1^2 =225 and s_2^2 =200

$$F = \frac{225}{200}$$

$$= 1.1$$

Value of Degrees of freedom (DF)

$$v_1 = n_1 - 1 = 10 - 1$$

$$= 9$$

$$and\ v_2 = n_2 - 1$$

$$= 11 - 1$$

$$= 10$$

at 1% level of significance is 4.49.

Conclusion:

Since computed value of F is smaller than the table value of F, the null hypothesis is accepted. Hence the variances of two populations are same.

Chi-Square$\left(\chi^2\right)$-Distribution

The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution, such as Friedman's analysis of variance by ranks.

The probability density function (pdf) of the chi-square distribution is

$$f(x;k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

The Chi- square (χ^2) test is one of the simplest and most widely used non parametric tests in statistical work. The χ^2 test was first used by Karl Pearson in the year 1900. The quantity χ^2 describes the magnitude of the discrepancy between theory and observation. It is defined as:

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

Properties of χ^2-Distribution:

1. The Mean of χ^2 distribution is equal to the number of degrees of freedom (n)

2. The variance is equal to two times the number of degrees of freedom. i.e The variance of χ^2 distribution is equal to 2n

3. The median of χ^2 distribution divides, the area of the curve into two equal parts, each part being 0.5

4. The mode of When Two Chi- squares □ 2 1 and □ 2 2 are independent □ 2 distribution with □1and □2 degrees of freedom and their sum □ 2 1 + □ 2 2 will follow □ 2 distribution with (□1 + □2) degrees of freedom.distribution is equal to (n-2)

5. Since Chi-square values always positive, the Chi-square curve is always positively skewed.

6. Since Chi-square values increase with the increase in the degrees of freedom, there is a new Chi-square distribution with every increase in the number of degrees of freedom.

7. The lowest value of Chi-square is zero and the highest value is infinity ie $\chi^2 \geq 2$

8.When Two Chi- squares χ_1^2 and χ_2^2 are independent χ^2 distribution with □1and □2 degrees of freedom and their sum $\chi_1^2 + \chi_2^2$ will follow χ^2 distribution with (□1 + □2) degrees of freedom.

Applications of Chi square test

1) To test of Goodness of Fit
2) To test Independence of Attributes
3) To test of Homogeneity

Example of Chi- square Distribution Problem

Out of 8,000 graduates in a town 800 female out of 1600 graduate employees 120 are female.

Gender	Employed	Not Employed	Total
Male	1480	5720	7200

Female	120	680	800
Total	1600	6400	8000

Use χ^2 to determine if any distinction in made in appointment on the basis of gender. value of χ^2 at 5% level for one degrees of freedom is 3.84.

Null Hypothesis:

The appointment does not based on the gender

$H_0 : \sum O_i = \sum E_i$

Alternative Hypothesis:

The appointment is based on the gender

$H_1 : \sum O_i \neq \sum E_i$

Level of significance$(\alpha) = 5\%$

Under H_0, The test statistic value can be defined as

$$\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

Solution:

We find Expected frequencies

$$\mathrm{E}(O_i) = \frac{(Row\ Total) \times (\text{Coloum}\ Total)}{Grand\ \ Total\ \ of\ \ all\ \ cells}$$

$$\mathrm{E}(1480) = \frac{7200 \times 1600}{8000}$$

$$= 1440$$

$$\mathrm{E}(5720) = \frac{7200 \times 6400}{8000}$$

$$= 5760$$

$$\mathrm{E}(120) = \frac{800 \times 1600}{8000}$$

$$= 160$$

$$\mathrm{E}(680) = \frac{800 \times 6400}{8000}$$

$$= 640$$

Observed Frequency(Oi)	Expected Frequency(Ei)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1480	1440	40	1600	1.11
5720	5760	-40	1600	0.28
120	160	-40	1600	10
680	640	40	1600	2.5
8000	8000			$\sum\left(\frac{(O_i - E_i)^2}{E_i}\right) = 13.89$

$$\chi^2 = \sum\left(\frac{(O_i - E_i)^2}{E_i}\right) = 13.89$$

$$Df = (r-1)(c-1)$$

$$= (2-1)(2-1)$$

$$= 1$$

$$\chi^2_{Critical} = 3.84$$

Conclusion:

Here we observe that the test statistic value of $\chi^2\,(13.89)$ *is* greater than Critical value(3.84).So, we reject the null hypothesis

Therefore we conclude that there is a sufficient evidence to support that the The appointment is based on the gender

ESTIMATION

Definition: When the data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences or generalization about that population from the information embodied in the sample. Statistical estimation, or briefly estimation is concerned with the methods by which population characteristics are estimated form sample information.

With respect to estimating a parameter, the following two types of estimates are possible:

- Point estimation
- Interval estimation

Point estimation

The point estimation in a single number which is used as an estimate of the unknown population parameter. The procedure in point estimation is to select a random sample of 'n' observations X_1, X_2, X_3 ….. X_n from a population f (X, θ) and then to use some preconceived method to arrive from these observations at a number say $\hat{\theta}$ (read theta hat)which we accept as an estimator of θ. the estimator θ is a single point on the real number scale on thus the name point estimation, $\hat{\theta}$ depends on the random variables that generate the sample and hence, it too is a random variable with its own sampling distribution.

{Notes: The symbol θ is generally used to denote a parameter that could be a mean, median or some measure of variability, etc.}

Interval estimation or confidence level:

As distinguished form a point estimate which provides one single value of the parameter. An interval estimate of a population parameter is a statement of two values between which is estimated that the parameter lies. An interval estimate would always be specified by two values, i.e., the lower one and the upper one. In more technical terms, interval estimation refers to the estimations of a parameter by a random interval called the confidence interval, whose end points L and U with L < U, are functions of the observed random variables such that the probability that the inequality L < θ <U is satisfied in terms of predetermined number. L and U are called the confidence limits and are the random end points of interval estimate.

If we estimate the average income of the people living in a village as Rs.875 it will be a point estimate the average income of the could lie between Rs.800 and Rs.950, it will be an interval estimate.

On comparing these two methods of estimation we find that point estimation has an advantage as much as it provides and exact value for the parameter under investigation.

Properties of a good estimator:

A distinction is made between an estimate and an estimator. The numerical value of the sample mean is said to an estimate of the population mean figure, for example, the sample mean $\bar{x}$ is an estimator of the population mean.

A good estimator, as common sense dictates, is close to the parameter being estimated. Its quality is to evaluated in terms of the following properties.

1. **Unbiasedness:**

An estimator is said to be unbiased if its expected value is identified with the population parameter being estimated. That is if $\hat{\theta}$ is an unbiased estimate of θ , then we must have

E($\hat{\theta}$) = θ many estimators are "Asymptomatically Unbiased" in the sense of the biases reduce to practically insignificant values zero when 'n' becomes sufficiently large. The estimator s^2 is an example.

2. Consistency:

If an estimator, say $\hat{\theta}$, approaches the parameter θ closer and closer as the sample size 'n' increases, $\hat{\theta}$ is said to be a consistent estimator of θ . Stating somewhat more rigorously. The estimator $\hat{\theta}$ is said to be a consistent estimator of θ if as 'n' approaches infinity, the probability approaches 1 that $\hat{\theta}$ will differ from the parameter θ by not more than an arbitrary small constant.

3. Efficiency:

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for, a given sample size) is said to be relatively more efficient. Stated in a somewhat different language, estimator $\hat{\theta}_1$ is said to be more efficient than another estimator $\hat{\theta}_2$ for θ if the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated.

4. Sufficiency:

An estimator is said to be sufficient if it conveys as much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exits, it is absolutely unnecessary to consider any other estimator: a sufficient estimator ensures that all information a sample can furnish with respect to the estimation of a parameter is being utilized.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties.

Unit-IV

Testing of Hypothesis

Introduction

In many realities, inferences about populations are to be drawn based on the characteristics of samples. As discussed earlier, sampling enables a researcher to draw an inference about a population. The inference may be pertaining to certain hypothesis.

Hypothesis is an assumption about a population. Consider, a study relating to buyers behavior. A few sample hypothesis are presented as follows:

- Mean purchase made by females (**μ_1**) is more than or equal to mean purchases made by males (**μ_2**) in a textile store (**$\mu_1 \geq \mu_2$**).
- Mean age of female shoppers (**μ_1**) is less than or equal to that of male shoppers (**μ_2**) in a book exhibition (**$\mu_1 \leq \mu_2$**).
- Mean monthly income of buyers (**μ**) in shop is more than or equal to Rs.10,000 (**$\mu \geq$** 10,000)

Test of Hypothesis or Test of significance: A very important aspect of the sampling theory is the study of test of significance which enables us to decide on the basis of the sample results. The deviation between the observed sample statistic and hypothetical parameter value.

A test of statistical hypothesis is a two action decision problem after the experimental sample values have been obtained the two actions being acceptance (or) rejection of hypothesis under consideration.
Testing of hypothesis are two types:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)\

Null hypothesis (H_0)

It is usually a hypothesis of no difference is called null hypothesis. It is usually denoted by "H_0". It should be completely impartial and should have no brief for any party or company nor should be allow his personal views to utilize the decision.

Example: Let us consider the light bulbs problem. Suppose that the bulbs manufactured under some standard manufacturing process have an average life of " hours and is proposed to test a new procedure "" for manufacturing light bulbs. Thus, we have two populations of bulbs those manufacture by standard process and those manufacture by new process.

In this problem the following three hypotheses may be set up

1. Standard process is greater than new process.
2. Standard process is less than to new process
3. There is no difference between standard process and new process.

Null hypothesis (H_0): There is no difference between new process and standard process.

Alternative hypothesis (H_1):

Any hypothesis which is complimentary to the null hypothesis is called alternative hypothesis, which is denoted by H_1.

Example: Above light bulbs alternative hypothesis (H_1) is: New process is better than standard process (or) new process is inferior to standard process.

Let us, suppose that the bulbs manufactured under some standard manufacturing process have an average life of 'μ_1' hours. If 'μ_2' is the mean life of the bulbs manufactured by the new process.

Procedure for Testing of Hypothesis:

We know summarize below the various steps in testing of statistical hypothesis in a systematic manner.

1. Null Hypothesis: Set up the Null Hypothesis (H_0)
2. Alternative Hypothesis: Set up the Alternative Hypothesis (H_1). This will enable us to decide whether we have to use a single – tailed (right or left tailed) test or two – test .
3. Level of Significance: Choose the appropriate levee of Significance (α) depending on reliability of the estimates and permissible risk. This is to be decide before sample is drawn, i.e., α is fixed in advance.
4. Test Statistic (or Test criterion): Compute the test statistic

   Under the null hypothesis

$$Z = \frac{t - E(t)}{S.E(t)}$$

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of Z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Type –I and Type-II Errors

	Decision	
	Accept H_0	Reject H_1
H_0 true	Correct decision	Type-I error
H_0 is false	Type-II error	Correct decision

Type-I error: reject H_0 when it is true is called type-I error

P (Type-I error) = α (Level of significance)

Type-II error: accept H_0 when it is false is called type-II error

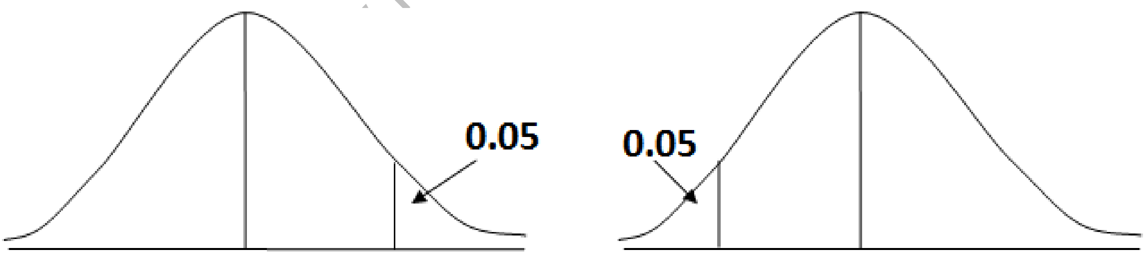
P (Type-II error) **= 1-** α

Level of significance

Having formulated the hypothesis, the next step is its validity at certain level of significance. The confidence with which a null hypothesis is accepted or rejected depends upon the significance level. A significance level of say 5% means that the risk of making a wrong decision is 5%. The researcher is likely to be wrong in accepting false hypothesis or rejecting a true hypothesis by 5 out of 100 occasions. Therefore, a 1% significance level provides greater confidence to the decision than 5% significance level.

One-tailed:

A hypothesis test may be one-tailed or two-tailed. In one tailed test the test-statistic for rejection of null hypothesis falls only in one-tailed of sampling distribution curve.

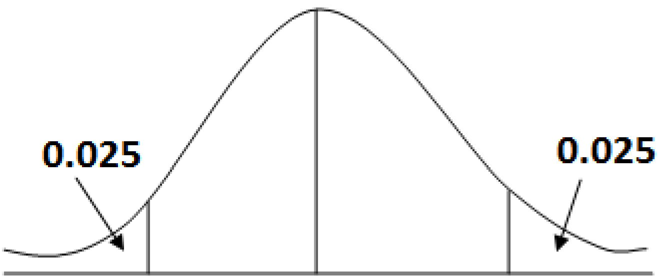


Right tailed tests Left tailed tests

Whether the test is one-sided or two sided-depends on alternate hypothesis.

Two tailed tests

As two tailed test is one in which the test statistics leading to reading to rejection of null hypothesis falls on both tails of the sampling ∞distribution curve shown



Two tailed tests

When we should apply a hypothesis test that is one-tailed or two-tailed depends on the nature of the problem. One-tailed test is used when the research's interest is primarily on one side of the issue
Example: "Is the current advertisement less effective than the proposed new advertisement"?

A two tailed test is appropriate, when the researcher has no reason to focus on one side of the issue. Example "Are the two markets- Mumbai and Delhi different to test market a product?"

Sign of alternate hypothesis	Type of test
$\neq$	Two-sided
$<$	One-sided to left
$>$	One-sided to right

Large Samples

If the sample size in greater than or equal to 30 ($n \geq 30$). Then it is called a "Large Sample".

In this section, we will study the following tests which are based upon on a large sample

- Test for single mean
- Test for two means
- Test for single proportion
- Test for two proportion

Test for Single Mean

Let Xi (i = 1, 2, 3… n) be a random sample size 'n' drawn form a normal population with mean μ

Null hypothesis: There is no significance difference between the sample mean and the population mean.

Now under null hypothesis (H_0), the test statistic is

$$Z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}}$$ (If standard deviation is known)

Where $\bar{x}$ = Sample mean
μ = Population mean
σ = Standard deviation

$$Z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}$$ (If standard deviation is unknown)

Where s = Sample Standard deviation.

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for population mean μ formulae is $\bar{x} \pm (Z_{tab})\frac{\sigma}{\sqrt{n}}$

95% Confidence limits for the population mean μ are $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

99% Confidence limits for the population mean μ are $\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$

90% Confidence limits for the population mean μ are $\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}$

Example:

A sample of 36 house holds in village was taken and the average income was found to be Rs.116.2 per day with a standard deviation of 25.

to test the hypothesis that the average income of households in a village 115 per day.

Solution:

Sample size, $n = 36$

The sample mean, $\bar{x} = 116.2$

Population mean, $\mu = 115$

Population standard deviation, $\sigma = 25$

1)

We test to determine whether the average income significantly different from 115.

Hypothesis :

The null and alternative hypotheses are,

$H_0 : \mu = 115$

vs

$H_a : \mu \neq 115$

Teststatistics :

Under H_0, the test statistic us defined as

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{116.2 - 115}{25/\sqrt{36}}$$

$$= \boxed{0.29}$$

*U*sing Z table to the critical(table value)

$Z_{critical} = 1.96$

Concludion :

Here we observe that the calculated value of z is less than table of Z test

So we accept null hypothesis(H_0)

Therefore we conclude that there is no signficance different from Mean of Households income in sample and Population

Test for two Means or Test of significance for difference of two Means

Let $\bar{x}_1$ be the mean of a sample of size n_1 drawn from a population with mean μ_1 and variance σ_1^2 and $\bar{x}_2$ be the mean of another independent sample of size n_2 drawn from another population with mean μ_2 and variance σ_2^2.

Null hypothesis (H_0): There is no significance difference between two population means.

Now under null hypothesis H_0, the test statistic is

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where n_1 = no. of observations of first sample.
n_2 = no. of observations of second sample.
$\bar{x}_1$ = First sample mean
$\bar{x}_2$ = Second sample mean
σ_1 = Standard deviations of first sample
σ_2 = Standard deviation of second sample

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for $|\mu_1 - \mu_2|$ i.e., for the difference in the two means of populations formulae is $|\bar{x}_1 - \bar{x}_2| \pm (Z_{tab})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

95% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x}_1 - \bar{x}_2| \pm (1.96)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

99% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x}_1 - \bar{x}_2| \pm (2.58)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

90% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x}_1 - \bar{x}_2| \pm (1.645)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Test for Single Proportion

If a random sample size of n. let X is a no. of persons possessing the given attribute. Then the sample proportion of success is $p = \frac{X}{n}$, we have proved that $E(p) = P$.

Null hypothesis (H_0): There is no significance difference between sample proportion and population proportion.

Now under null hypothesis H_0, the test statistic is

$$Z = \frac{|p - P|}{\sqrt{\frac{PQ}{n}}}$$

Where 'p' = Sample proportion

P = Population proportion

Q = 1-P

n = no. of observations or samples

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for population proportion P is $|p - P| \pm (Z_{tab})\sqrt{\frac{PQ}{n}}$

95% Confidence interval for population proportion P are $|p - P| \pm (1.96)\sqrt{\frac{PQ}{n}}$

99% Confidence interval for population proportion P are $|p - P| \pm (2.58)\sqrt{\frac{PQ}{n}}$

90% Confidence interval for population proportion P are $|p-P| \pm (1.645)\sqrt{\frac{PQ}{n}}$

Test for two proportions

Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 form the two populations respectively. Then sample proportions are

$$p_1 = \frac{X_1}{n_1} \quad \text{And} \quad p_2 = \frac{X_2}{n_2}$$

Null hypothesis (H_0): There is no significance difference between two population Proportions.

Now under null hypothesis H_0, the test statistic is

$$Z = \frac{|p_1 - p_2|}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Where n_1 = no. of observations of first sample.
n_2 = no. of observations of second sample.
p_1 = First sample proportion
p_2 = Second sample proportion
P = Population proportion

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2},$$

$$Q = 1 - P$$

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for $|P_1 - P_2|$ i.e., for the difference in the two proportions of populations formulae is $|p_1 - p_2| \pm (Z_{tab})\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

95% Confidence limits for $|P_1 - P_2|$ are $|p_1 - p_2| \pm (1.96)\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

99% Confidence limits for $|P_1 - P_2|$ are $|p_1 - p_2| \pm (2.58)\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

90% Confidence limits for $|P_1 - P_2|$ are $|p_1 - p_2| \pm (1.645)\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Example:

In a large ciry A, 20% of a random sampel of 900 school children had defective eye-sight. In other large city B, 15% of random sample of 1600 children had same defect. Is this difference between the two populaitons significant? at 5% level of significance.

Solution:

Null hypothesis:

There is no significance difference between the proportion of School children who had defective eye-sight in two cities A and B

$H_0 : P_1 = P_2$

Alternative hypothesis:

There is a significance difference between the proportion of School children who had defective eye-sight in two cities A and B

$H_1 : P_1 \neq P_2$

Level of significance value $(\alpha) = 0.05$

Under H_0, The test statistic value can be defined as

$$Z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Proportion of School children who had defective eye-sight in city A $(p_1) = 0.25$

Proportion of School children who had defective eye-sight in city B $(p_2) = 0.15$

$n_1 = 500$

$n_2 = 600$

$$Pooled\ \text{Proportion}(P) = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$= \frac{500(0.20)+600(0.15)}{900+1600}$$

$$= 0.168$$

$$Z = \frac{0.20 - 0.15}{\sqrt{0.168(1-0.168)\left(\frac{1}{900}+\frac{1}{1600}\right)}}$$

$$= \frac{0.05}{0.0156}$$

$$= 3.21$$

Using Z critcal values table

$$Z_{\text{Critical}} = 1.96$$

Conclusion:

Here we observe that the test statistic value (3.21) is greater than the critical value of z (1.96).So we reject the null hypothesis.

Therefore we coclude that there is no significance difference between Proportion of School children who had defective eye-sight in two cities A and B

Analysis of Variance (ANOVA)

Analysis of variance is a powerful statistical tool for significantly the test based on t – test is an adequate procedure only for testing the significance between the sample means.

In a situation when we have two or more samples to consider at a time an alternative procedure is called analysis of variance.

Eg: Suppose five fertilizers are applied at random to four plots each in a field consists of 20 plots of the same shape and same size and the yield of wheat on each to these plots is given we may be interested to finding out whether the affect of these fertilizers the yields is significantly different or in other words.

The answer of this problem is providing by the technique of analysis of variance (ANOVA) is to test the homogeneity of the several means (more than two means).

Assumptions for ANOVA test:

ANOVA test is based on the test statistics F for the validity of the F test in ANOVA the following assumptions are.

- The sample observations are independent
- Various treatments and environment effects or additive in nature.
- The sample have been drawn from the normal population

In the following selections we will discuss the analysis of variance

- One way classification
- Two way classification.

Example:

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full – size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha = 5\%$.

Compact Cars(X1)	Midsize Cars(X2)	Full-Size Cars(X3)
643	469	484
655	427	456
702	525	402

Solution :

We want to test whether the hypothesis there is a significance difference between the mean head pressure of three types of cars

Null Hypothesis:

There is no significance difference between the mean head pressure of three types of cars

$H_0 : \mu_1 = \mu_2 = \mu_3$

Alternative hypothsis:

There is a significance difference between the mean head pressure of three types of cars

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

Level of Significance $(\alpha) = 5\%$

Calcualtions:

Compact Cars(X1)	Midsize Cars(X2)	Full-Size Cars(X3)
643	469	484
655	427	456
702	525	402
$\sum X_1 = 2000$	$\sum X_2 = 1421$	$\sum X_3 = 1342$
$\overline{X}_1 = 666.67$	$\overline{X}_2 = 473.67$	$\overline{X}_3 = 447.33$

We find Grand Total $(G) = 643 + 655 + 702 + 469 + ... + 402$

$= 4763$

Total number of Observation $(N) = 9$

Correction Factor $(C.F) = \dfrac{G^2}{N}$

$$= \frac{(4763)^2}{9}$$

$= 2520685.44$

Sum of Square of Total $(SST) = \left(643^2 + 655^2 + 702^2 + 469^2 + ... + 402^2\right) - C.F$

$= 2616989 - 2520685.44$

$= 96303.56$

Sum of square of Between the groups (SSB)

$$=\left(\frac{\left(\sum X_1\right)^2}{n_1}+\frac{\left(\sum X_2\right)^2}{n_2}+\frac{\left(\sum X_3\right)^2}{n_3}\right)-C.F$$

$$=\left(\frac{(2000)^2}{3}+\frac{(1421)^2}{3}+\frac{(1342)^2}{3}\right)-2520685.44$$

$$=2606735-2520685.44$$

$$=86049.56$$

Sum of Square of Error $(SSE)=SST-SSB$

$$=96303.56-86049.5556$$

$$=10254$$

Degrees of Freedom (Df):

DF of Total =N-1

=9-1

=8

DF of Groups=n-1

=3-1

=2

DF of Error=N-n

=9-3

=6

ANOVA					
Source of Variation	*SS*	*Df*	*MS*	$F = \frac{MSB}{MSE}$	*F critical(From F table)*
Between Groups	86049.56	2	43024.78	25.17541	5.143253
Within Groups(Error)	10254	6	1709		
Total	96303.56	8			

From The ANOVA Table
The test statistic value of F =25.18
Critical value of F =5.14

Decision:
Here we observe that the test statistic value of F(25.18) is greater than the critical value(5.14).So we reject the null hypothesis.

Conclusion:
Therefore we conclude that there is a significance difference between the mean head pressure of three types of Cars.

Unit-V

Curve Fitting

Curve fitting is to develop methods for establishing the relationship between two variables whose values have been obtained by experiment. Before doing this it is necessary to consider the algebraic relationships that give rise to standard geometric shapes.

Standard curves:

Straight line

Second-degree curves

Exponential curves

Power Curve

Straight Line:

The equation of a straight line is a first-degree relationship and can always be expressed in the form:

$$y = ax + b$$

The normal equations are

$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

Example:

Use the least square method to determine the equation of line of best fit for the data. Then plot the line.

X	Y	xy	x^2
56	147	8232	3136
42	125	5250	1764
72	160	11520	5184
36	118	4248	1296
63	149	9387	3969
47	128	6016	2209
55	150	8250	3025
49	145	7105	2401
38	115	4370	1444
42	140	5880	1764
68	152	10336	4624
60	155	9300	3600
$\sum x =$ 628	$\sum y =$ 1684	$\sum xy =$ 89894	$\sum x^2 =$ 34416

$$\sum y = na + b\sum x$$

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

$$1684 = 12a + 628b -----------(1)\times 628$$

$$89894 = 628a + 34416b --------(2)\times 12$$

$$(1)-(2)$$

$$1057552 = 144a + 394384b$$

$$1078728 = 144a + 412992b$$

$$----------------------$$

$$-21176 = 0a - 18608\text{b}$$

$$-------------------$$

$$b = \frac{21176}{18608}$$

$= 1.14$

We substitute the b value in the equation (1)

$1684 = 12a + 628b$

$1684 = 12a + 628(1.14)$

$1684 = 12a + 715.92$

$12a = 1684 - 715.92$

$12a = 968.08$

$$a = \frac{968.08}{12}$$

$= 80.67$

The straight line equation is

Y = a+bX

$\boxed{Y = 80.67 + 1.14X}$

Second-degree curves:

The general second-degree curve is:

$$y = ax^2 + bx + c$$

Normal Equations

$$\sum y = a\sum x^2 + b\sum x + nc$$

$$\sum xy = a\sum x^3 + b\sum x^2 + c\sum x$$

$$\sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2$$

Example: Fit a Second Degree curve to the following Data?

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Solution:

x	X_i	Y_i	X_i^2	X_i^3	X_i^4	X_iY_i	$X_i^2Y_i$
1	-4	2	16	-64	256	-8	32
2	-3	6	9	-27	81	-8	54
3	-2	7	4	-8	16	-14	28
4	-1	8	1	-1	1	-8	8
5	0	10	0	0	0	0	0
6	1	11	1	1	1	11	11
7	2	11	4	8	16	22	44
8	3	10	9	27	81	30	90
9	4	9	16	64	256	36	144
N=9	$\sum X_i = 0$	$\sum Y_i = 74$	$\sum X_i^2 = 60$	$\sum X_i^3 = 0$	$\sum X_i^4 = 708$	$\sum X_iY_i = 51$	$\sum X_i^2Y_i = 411$

{Here, we made the equation } We use the method called "the method of least squares".

The Equation of parabola is $y = a + bx + cx^2$

Hence, the normal equations are

$$\sum Y_i \ = \ Na + b\sum X_i + c\sum X_i^2$$

$$\sum X_iY_i \ = \ a\sum X_i + b\sum X_i^2 + c\sum X_i^3$$

$$\sum X_i^2Y_i \ = \ a\sum X_i^2 + b\sum X_i^3 + c\sum X_i^4$$

$$74 = 9a + b(0) + 60c$$

$$9a + 60c = 74 - - - - (i)$$

$$51 = a(0) + 60b + 0c \text{ -----}(ii)$$

$$b = \frac{51}{60}$$
$$= 0.85$$
$$411 = 60a + 0b + 708c$$
$$411 = 60a + 708c \quad ------(iii)$$
$$Solving\ (i)\ and\ (iii)\ simultaneously,$$
$$we\ get\ a = 10.004\ ,\ c = -0.267$$
$$Second\ Degree\ curve\ is$$
$$y = 10.004 + 0.85X - 0.267X^2$$
$$= 10.004 + 0.85(x-5) - 0.267(x-5)^2$$
$$= 10.004 + 0.85x - 4.25 - 0.267(x^2 - 10x + 25)$$
$$= 10.004 + 0.85x - 4.25 - 0.267x^2 + 2.67x - 6.675$$
$$\boxed{y = -0.921 + 3.52x - 0.267x^2}$$

Exponential Curve:

Exponentials are often used when the rate of change of a quantity is proportional to the initial amount of the quantity. If the coefficient associated with b and/or d is negative, y represents exponential decay. If the coefficient is positive, y represents exponential growth.

$$y = ae^{bx}$$
$$We\ logarthims\ on\ both\ sides$$
$$\log y = \log\left(ae^{bx}\right)$$
$$\log y = \log a + \log e^{bx}$$
$$\log y = \log a + bx$$

Where logy = Y

$\log a = A$

$b = B$

$Y = A + Bx$

The normal equations are

$\sum Y = nA + B\sum X$

$\sum XY = A\sum X + B\sum X^2$

Power Curve:

Suppose we have data that, when plotted, appear to have a power-law character. If we choose a power function to represent the data, we write

$$y = ax^b$$

$y = ax^b$

We apply logarithms on both sides

$\log y = \log\left(ax^b\right)$

$\log y = \log a + \log x^b$

$\log y = \log a + b\log x$

$Y = \log y$

$A = \log a$

$X = \log x$

$B = b$

$Y = A + BX$

The Normal equations are

$\sum Y = nA + B\sum X$

$\sum XY = A\sum X + B\sum X^2$

Correlation

First we know some basic terms:

Uni-variate Distribution: the distribution which involves one variable is called univariate distribution

Bi-variate Distribution: The distribution which involves two or more variables is called Bi-variate distribution.

Correlation:

The relation between two variables is called correlation. It is used to measure the relationship between two variables.

If the change one variable affects a change in other variable, the variables, are correlated. Correlation broadly classified into three ways.

Positive Correlation

If two variables deviated in the same direction. If increase in one variable in a corresponding increase in other variable. (Or)

If decrease in one variable a corresponding decrease in the other variable. This is the same direction. This type of correlation is called positive correlation.

Example:

1. Height and weight of certain group of persons
2. Income and expenditure.
3. Rainfall and agricultural production.

Negative Correlation:

If tow variables deviated in the opposite direction. If the increase in one variable in a corresponding decrease in other variable. (Or)

If the decrease in one variable in a corresponding increase in the other variable. This is the opposite direction this type of correlation is called Negative correlation.

Example: Price and demand of commodity.

Zero Correlation or Independent Correlation:

There is no relation between two variables. This is also known as zero correlaton

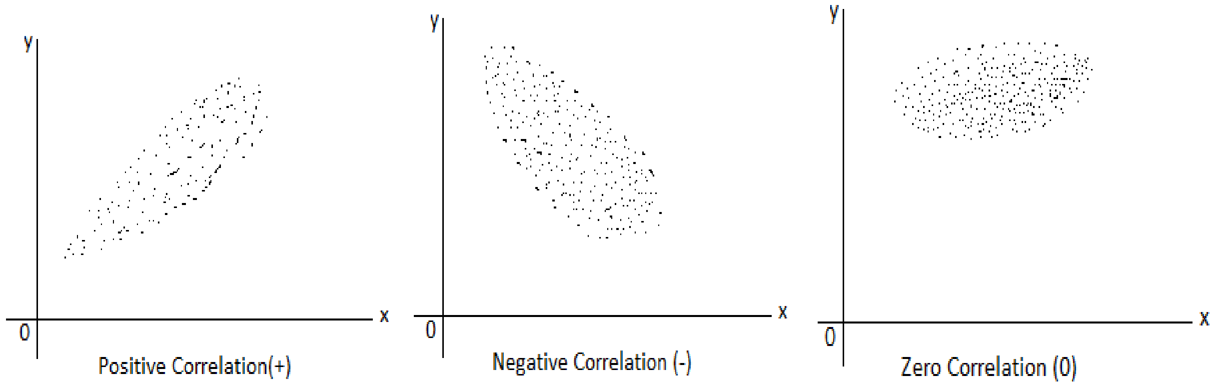
Example: Beautiful and Intelligence.

Scatter Diagram:

It is the simplest way of the diagrammatic representation of the Bi – variate data. For Bi-variate distribution if the values of the variables (x,y) I = 1,2,3….n are plotted along the x-axis and y- axis respectively in the xy-plane.

The diagram of dots obtained is known as scattered diagram. From this scattered diagram we can form a fairly good, though vague, idea whether the variables are correlated or not.

Example: if the dots are very dense, that is very close to each other. We should expect a fairly good amount of correlation between the variables. If the dots are widely scattered, we should expect a bad correlation.



Karl Pearson's Correlation Coefficient:

This is used to measure the degree of linear relationship between two variables. If x and y are two variables then "Karl Pearson's correlation coefficient (r)" is

$$r_{xy} \text{ or } r(x,y) = \frac{n\sum xy - \left(\sum x \sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

1. Correlation co-efficient is always lies between -1 and +1. That is, $-1 \le r \le +1$.
   If r = +1, the correlation is perfect and positive.
   If r = 0, the correlation is zero, there is no relation between two variables.
   If r = -1, the correlation is perfect and negative.
2. Correlation coefficient is independent of change of origin and scale that is r(x, y) = r(u, v)
3. Independent variables are un-correlated
4. Karl Pearson's correlation co-efficient deals with the quantitative characteristics only.

Example:

The following table gives information on ages and cholesterol levels for a random sample of 10 men.

Age	58	69	43	39	63	52	47	31	74	36
Cholesterol Level	189	235	193	177	154	191	213	165	198	191

Calculate the correlation Coefficient

$$Correlation\ Coefficient(r) = \frac{n\sum xy - \left(\sum x \sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$= \frac{10(98667) - (512 \times 1906)}{\sqrt{\left[10(28110) - (512)^2\right]\left[10(368000) - (1906)^2\right]}}$$

$$= \frac{10798}{\sqrt{[281100 - 262144][3680000 - 3632836]}}$$

$$= \frac{10798}{\sqrt{(18956)(47164)}}$$

$$= \frac{10798}{\sqrt{894040784}}$$

$$= \frac{10798}{29900.51}$$

$$= 0.3611$$

Comment : There is postive correlation between the variables Age and Cholestrol level

Probable error of Correlation Coefficient:

If r(x, y) is correlation coefficient in a sample of "n" pairs of observations then standard error is given by

$$\text{Standard Error (S.E)}=\frac{1-r^2}{\sqrt{n}}$$

Probable error of correlation coefficient is defined as

$$\text{P.E (r)}=0.675(\text{S.E})$$
$$=0.675\left(\frac{1-r^2}{\sqrt{n}}\right)$$

Where P.E = probable error of correlation coefficient.

Spearmen's Rank Correlation Coefficient

X_i, y_j be the ranks of two characteristics A and B respectively the spearmen's correlation coefficient is denoted by

$$\rho(x,y)=1-\frac{6\sum d^2}{n(n^2-1)}$$

Where 'n' = number of observations

d = the rank of x – rank of y (R(x)-R(y))

Spearmen's Tide Rank correlation co-efficient: Let (x, y) be the same repeated ranks of the two characteristics A and B then spearmen's tide rank correlation coefficient is denoted by

$$\rho=1-\frac{6\left(\sum d^2+T_x+T_y\right)}{n(n^2-1)}$$

Where n = number of observations

T_x = Tide rank in x-series

$$T_x = \frac{\sum_{i=1}^{n} m_i (m_i^2 - 1)}{n(n^2 - 1)}$$

Where m_i = number of repeated times in i^{th} highest value.

T_y= Tide rank in y-series

$$T_y = \frac{\sum_{j=1}^{n} m_j (m_j^2 - 1)}{n(n^2 - 1)}$$

Where m_j= number of repeated times in j^{th} highest value.

Example:

The scores for nine students in physics and math are as follows:
Physics: 56,75,45,71,62,64,58,80,76,61
Mathematics: 66,70,40,60,65,56,59,77,67,63

Compute the student's ranks in the two subjects and compute the Spearman rank correlation.

Solution

Physics(x)	Maths(y)	R(x)	R(y)	$d = R(x) - R(y)$	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

This indicates a strong positive relationship between the ranks individuals obtained in the maths and physics.

Regression

The literal meaning of regression analysis is "stepping back towards the average". The regression analysis was first derived by "Sir. Francis Galton". The regression analysis is the mathematical measure of average relation between two or more variables in terms of the origin unit of the data.

The main aim of regression analysis is to estimate or predict unknown values from the given known values. Example: $y = a + bx$

Simple Regression or Linear Regression:

The average relation between one dependent variable and one independent variable is called regression.

Example: $y = a + bx$

Multiple Regressions:

The average relationship between one dependent variable and two or more independent variables is called multiple regression

Example: $y = a + b_1x_1 + b_2x_2 + b_3x_3 .... + b_nx_n$

Dependent and Independent Variables

In the regression analysis, there are two types of variables. One is dependent variable and other one is independent variable.

The variable whose value influenced or predicted is known as dependent variable or explained variables .Which values influences or prediction by other variables in known as independent variable or explanatory variable.

Regression Lines:

Let (x_i, y_j) I =1, 2, 3, ….n be the bi-variate data, 'y' is dependent variable and 'x' is independent variable then regression equation of 'y' on 'x' is defined as

$$y = a + bx$$

Where a is a constant

b is the regression coefficient

The regression equation of 'x' on 'y' is defined as

$$x = a + by$$

Any line passes through the points $\overline{x}$ and $\overline{y}$ respectively then the regression equation of 'y' on 'x' is defined as

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

Where b_{yx} is the regression coefficient of 'y' on 'x'.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$or$$

$$b_{yx} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\frac{1}{n}\sum x^2 - (\bar{x})^2}$$

The regression equation of 'x' on 'y' is defined as

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

Where b_{xy} is regression co efficient of 'x' on 'y'.

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$or$$

$$b_{xy} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\frac{1}{n}\sum y^2 - (\bar{y})^2}$$

Properties of Regression Co efficient:

- Correlation coefficient is geometric mean of two regression coefficients

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

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

$$= \sqrt{r \times r}$$

$$= \sqrt{r^2}$$

$$= r$$

So the Geometric mean of two regression coefficients is equal to correlation coefficient

2) The arithmetic mean of two regression coefficient is greater than the correlation coefficient.

$$\frac{b_{yx} + b_{xy}}{2} > r$$

3) If one regression coefficient is greater than unity, then the other regression coefficient must be lesser than unity.

$$b_{yx} > 1 \text{ and } b_{xy} < 1$$

4) Regression coefficient is independent of change of origin but not scale.

$$U = \frac{x-a}{h}, V = \frac{y-b}{k}$$

$$x = a + hU$$

$$E(x) = a + hE(U)$$

$$y = b + kv$$

$$E(y) = b + kE(V)$$

$$x - E(x) = h\left(U - E(U)\right)$$

$$y - E(y) = k\left(V - E(V)\right)$$

The Regression coefficient of y on x

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{yx} = \frac{Cov(x, y)}{\sigma_x}$$

$$Cov(x, y) = E\left(\left(x - E(x)\right)\left(y - E(y)\right)\right)$$

$$= E\left[h\left(U - E(U)k\left(V - E(V)\right)\right)\right]$$

$$= hkE\left(U - E(U)\right)\left(V - E(V)\right)$$

$$= hkCov(UV)$$

$$\sigma_x{}^2 = V(x) = E\left(x - E(x)\right)^2$$

$$= E\left(h\left(U - E(U)\right)\right)$$

$$= h^2 E\left(U - E(U)\right)^2$$

$$b_{yx} = \frac{Cov(x,y)}{\sigma_x{}^2}$$

$$= \frac{hkCov(U,V)}{h^2 E(U - E(U))^2}$$

$$= \frac{k}{h} b_{VU}$$

Similarly

$$b_{xy} = \frac{h}{k} b_{UV}$$

So the Regression coefficient is independent of change of origin but not scale

Example:

The following table gives information on ages and cholesterol levels for a random sample of 10 men.

Age	58	69	43	39	63	52	47	31	74	36
Cholesterol Level	189	235	193	177	154	191	213	165	198	191

Taking age as an independent variable and cholesterol level as a dependent variable,
1) Find the regression of cholesterol level on age.
2) Briefly explain the meaning of the values of a and b
3). Predict the cholesterol level of a 60 year old man.

Solution:
1) Objective: We find Regression Model to forecast Cholesterol level From Age. In this problem Cholesterol Level is Dependent variable it is denoted as Y and Age is a Independent variable it is denoted by X

The Regression Equation is

Cholestrol Level(Y) = a+b Age(X(

'a' is a Intercept or Constant

'b' Slope or Regression Coefficient

Solution

Age(X)	Cholestrol Level(Y)	xy	x^2	y^2
58	189	10962	3364	35721
69	235	16215	4761	55225
43	193	8299	1849	37249
39	177	6903	1521	31329
63	154	9702	3969	23716
52	191	9932	2704	36481
47	213	10011	2209	45369
31	165	5115	961	27225
74	198	14652	5476	39204
36	191	6876	1296	36481
$\sum x = 512$	$\sum y = 1906$	$\sum xy = 98667$	$\sum x^2 = 28110$	$\sum y^2 = 368000$

Regression line of y on x

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

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

$$= \frac{512}{10}$$

$$= 51.2$$

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

$$= \frac{1906}{10}$$

$$= 190.6$$

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

$$= \frac{10(98667) - (512 \times 1906)}{10(28110) - (512)^2}$$

$$= \frac{986670 - 975872}{281100 - 262144}$$

$$= \frac{10798}{18956}$$

$$= 0.5693$$

b_{yx} or b =0.5693

Regression line is

$$y - 190.6 = 0.5693(x - 51.2)$$

$$= 0.5693x - (0.5693 \times 51.2)$$

$$= 0.5693x - 29.1482$$

$$y = 0.5693x - 29.1482 + 190.6$$

$$\boxed{y = 0.5693x + 161.45}$$

Thus Regression Equation is

$$Cholestrol(y) = 161.45 + 0.5693 Age(x)$$

2)

Briefly explain the meaning of the values of a and b

Interscept (a): *If the Age is 0 years, then we predict the the Cholestrol level is 161.45*

$$y = 161.45 + 0.5693(x)$$
$$= 161.45 + 0.5693(0)$$
$$= 161.45$$

Slope (b): *If the Age is increased by 1 year then Cholestrol level is increased by 0.5693 units.*

3)

We estimate the Cholestrol level when age is 60-years old man

So $Age(x) = 60$

$$Cholestrol(y) = 161.45 + 0.5693 Age(x)$$
$$y = 161.45 + 0.5693(60)$$
$$= 195.608 \cong 196$$

The Cholestrol level when the age is 60 years old man is 196

Unit-VI
Control Charts

The epoch-making discovery and development of control charts was made by a young physicist Dr. Walter A. Shewart of Bell Telephone laboratories in 1924 and the following years. Based on the theory of probability and sampling Stewarts's Control charts provide a powerful tool of discovering and correcting the assignable causes of variation outside the "stable pattern" of chance causes, thus enabling us to stabilize and control our processes at desired performances and thus bring the process under statistical control

In industry one is faced with two kinds of problems

1. To check whether the process is conforming to standard laid down
2. To improve the level of standard and reduce variability consistent with cost considerations.

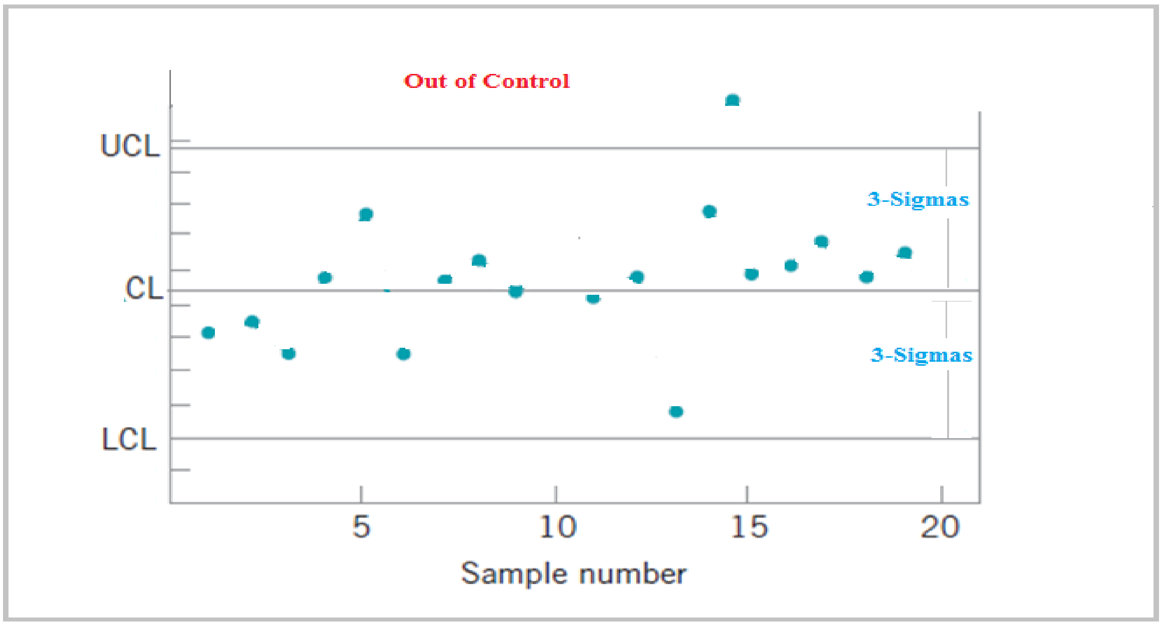
Shewhart's control charts provide an answer to both. Control charts provide criteria for detecting lack of statistical control.

A typical control chart consists of the following three horizontal lines:

1. A central line(CL) to indicate the desired standard of the level of the process
2. Upper control line (UCL)
3. Lower control line (LCL),

Together with a number of sample points as exhibited in the following diagram which depicts the principle of Shewhart's control charts.

Outline of a Control Chart



Objectives of Control Charts:

- Controlling ongoing processes by finding and correcting problems as they occur.
- Predicting the expected range of outcomes from a process.
- Determining whether a process is stable (in statistical control).
- Analyzing patterns of process variation from special causes (non-routine events) or common causes (built into the process).
- Determining whether the quality improvement project should aim to prevent specific problems or to make fundamental changes to the process.

Control Charts for variables: Variable control charts are used when quality is measured as variables (length, weight, tensile strength, etc.). The main purpose of the variable control charts is to monitor the process mean and the standard deviation.

$\bar{x}$ **and R chart:** No production process is perfect enough to produce all the items exactly alike. Some amount of variation, in the produced items, is inherent in any production scheme. This variation is the totally of numerous characteristics of the production process viz., raw material, machine, setting and handling, operators, etc. As pointed out earlier, this variation is the result of (1) chance causes and (2) assignable causes. The control limits in the $\bar{x}$ and R charts are so placed that they reveal the presence or absence of assignable causes of variation in the

a) Average($\bar{x}$)- Mostly related to machine setting,
b) Range (R) –Mostly related to negligence on the part of the operator.

Example: Construct the control chart for mean and the range for the following data on the basis of fuses, samples of 5 being taken every hour (each set of 5 has been arranged in ascending order of magnitude). Comment on whether the production seems to be under control, assuming that these are the data

42,42,19,36,42,51,60,18,15,69,64,61,
65,45,24,54,51,74,60,20,30,109,90,78,
75,68,80,69,57,75,72,27,39,113,93,94,
78,72,81,77,59,78,95,42,62,118,109,109,
87,90,81,84,78,132,138,60,84,153,112,136

Solution

Sample no.	Sample observations					Total	Sample Mean($\bar{x}$)	Sample Range
1	42	65	75	78	87	347	69.4	45
2	42	45	68	72	90	317	63.4	48
3	19	24	80	81	81	285	57.0	62
4	36	54	69	77	84	320	64.0	48
5	42	51	57	59	78	287	57.4	36
6	51	74	75	78	132	410	82.0	81
7	60	60	72	95	138	425	85.0	78

8	18	20	27	42	60	167	33.4	42
9	15	30	39	62	84	230	46.0	69
10	69	109	113	118	153	562	112.4	84
11	64	90	93	109	112	468	93.6	48
12	61	78	94	109	136	478	95.6	75
						Total	859.2	716

From the above data, we get

$$\bar{\bar{x}} = \frac{\sum \bar{x}}{12} = \frac{859.2}{12} = 71.60$$

$$\bar{R} = \frac{\sum R}{12} = \frac{716}{12} = 59.67$$

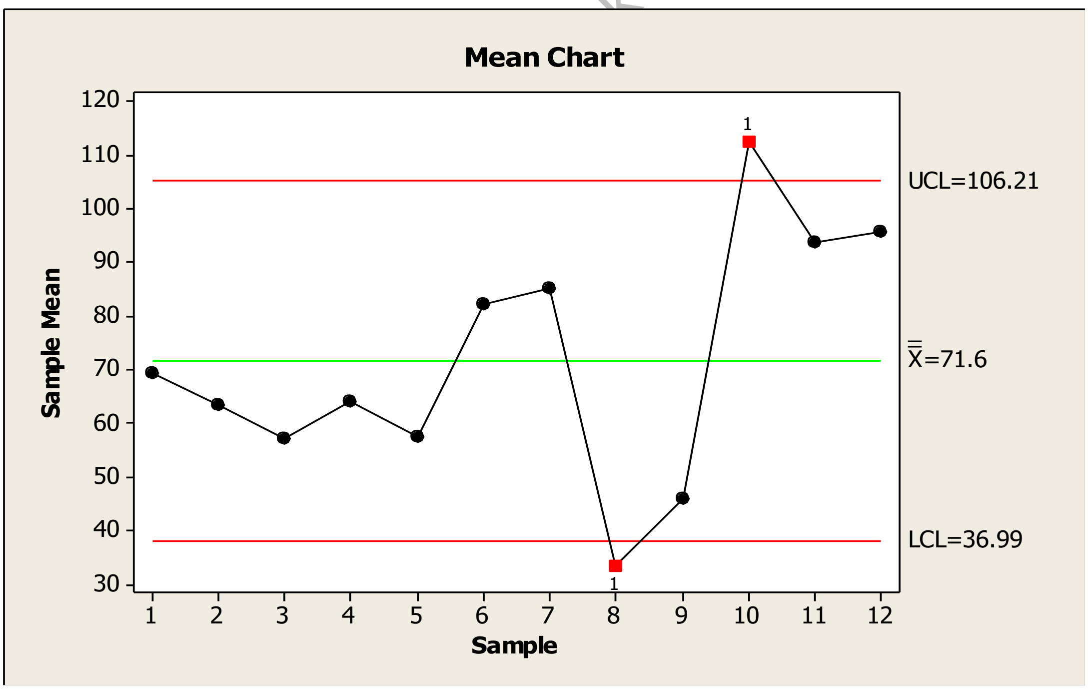
From the Tables, for sub sample size(n)= 5, we have A_2=0.58,D_3 =0 and D_4=2.115

$\bar{x} - Chart$

Central Line (C.L) $= \bar{\bar{x}} = 71.60$

Upper control Line (UCL) = $\bar{\bar{x}} + A_2\bar{R} = 71.60 + (0.58 \times 59.67) = 71.60 + 34.61 = 106.21$

Lower control Line (LCL) = $\bar{\bar{x}} - A_2\bar{R} = 71.60 - (0.58 \times 59.67) = 71.60 - 34.61 = 36.99$



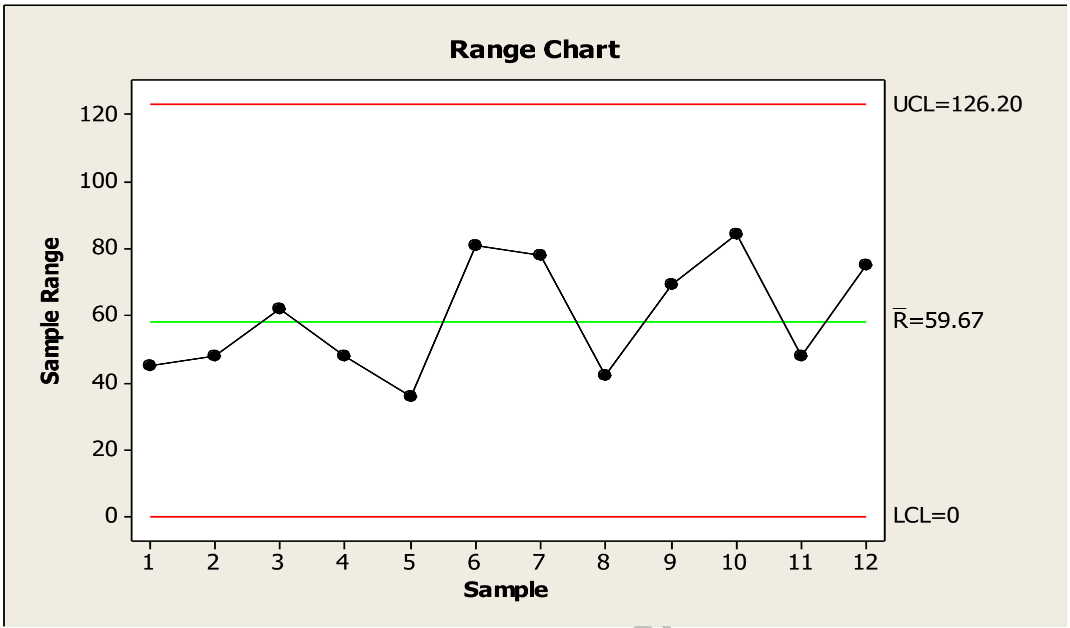
Sample points corresponding to sample number 8 and 10 lie outside the control limits in the Mean chart.

$R - Chart$

Central Line (C.L) = $\overline{R} = 59.67$

Upper control Line (UCL) = $D_4\overline{R} = (2.115 \times 59.67) = 126.20$

Lower control Line (LCL) = $D_3\overline{R} = (0 \times 59.67) = 0$



Since the entire sample points fall within the control limits, the Range chart shows that process is in control

Comment: Although R-chart depicts Control, the process can't be regarded to be in statistical control Mean ($\overline{x}$) Chart shows lack of control.

Interpretation of Mean ($\overline{x}$) and R-chart

In order to judge if a process is in control, $\overline{x}$ and R charts should be examined together and the process should be deemed in statistical control if both the charts show a state of control. Situations exist where R-chart is in a state of control but $\overline{x}$ chart is not. We summarize below, in a tabular form, such different situations and the interpretation to be accorded to each.

	Situation in		Interpretation
S .No	**$\overline{x}$-chart**	R-chart	
1.	In control	Points beyond limits only on one side	Level of process has shifted
2.	In control	Points beyond limits on both sides	Level of process is changing in erratic manner
3.	Out control	Points beyond limits	Variability has increased

		on both sides	
4.	Out of control	Out of control on one side`	Both level and variability have changed
5.	In control	Run of 7 or more points on one side of central line	Shift in process level
6.	In control	Trend of 7 or more points. No points outside control limits	Process level is gradually changed
7.	Runs of 7 or more points above central line	---	Variability has increased
8.	Points too close to the central line	----	Systematic differences within subgroups
9.	-----	Points too close to central line	Systematic differences within subgroups.

Standard Deviation (σ)Chart-: This chart is constructed to get a better picture of the variations in the quality standard in a process than that is obtained from the range chart provided the standard deviation(σ) of the various samples are readily available.

Control Charts for Attributes

In spit of wide applications of $\bar{x}$ and R (orσ) charts as a powerful tool of diagnosis of sources of trouble in a production process, their use is restricted because of the following limitations:

- They are charts for variables only i.e., for quality characteristics which can be measured and expressed in numbers.
- In certain situations they are impracticable and un-economical e.g., if the number of measurable characteristics, each of which could a possible candidate be for $\bar{x}$ and R chart, say 30,000 or so then obliviously there can't be 30,000 control charts.

As an alternative to $\bar{x}$ and R charts, we have the control chart for attributes which can be used for quality characteristics:

1. Which can be observed only as attributes by classifying an item as defective of non-defective i.e., conforming to specifications or not
2. Which are actually observed as attributes even though they could be measured as variables e.g., go and no-go gauge test results.

P-Chart or Control for fraction defective chart: This chart is constructed for controlling the quality standard in the average fraction defective of the products in a process when the observed sample items are classified into defectives & non-defectives.

Example: The following are the figures of defective in 22 lots each containing 2,000 rubber bolts:
425,430,216,341,225,322,280,306,337,305,356,402,216,264,126,409,193,326,280,389,451,420.
Draw control chart for fraction defective and comment on the state of control of the process

Solution: Here we have a fixed sample size n=2,000 for each lot if d_i and p_i are respectively the number of defectives and the sample fraction defective for i^{th} lot then

$$p_i = \frac{d_i}{2000}, (i = 1, 2, ..., 22)$$

Which are given the following table

S. No	d	p	S. No	d	P
1	425	0.2125	12	402	0.2010
2	430	0.2150	13	216	0.1080
3	216	0.1080	14	264	0.1320
4	341	0.1705	15	126	0.0630
5	225	0.1125	16	409	0.2045
6	322	0.1610	17	193	0.0965
7	280	0.1400	18	326	0.1630
8	306	0.1530	19	280	0.1400
9	337	0.1685	20	389	0.1945
10	305	0.1525	21	451	0.2255
11	356	0.1780	22	420	0.2100
					$\sum p_i = 3.5095$

In the usual notations, we have

$$\bar{p} = \frac{\sum p_i}{22} = \frac{305095}{22} = 0.1595$$

$$\bar{q} = 1 - \bar{p} = 1 - 0.1595 = 0.8505$$

Central line (C.L) = $\bar{p}$ = 0.1595

Upper control line (UCL) =

$$= \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.1595 + 3\sqrt{\frac{0.1595 \times 0.8505}{2000}}$$

$$= 0.1595 + 3\sqrt{0.000067}$$

$$= 0.1595 + 3 \times 0.0082$$

=0.1595+0.0246

UCL=0.1841

Lower control line (LCL) =

$$= \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.1595 - 3\sqrt{\frac{0.1595 \times 0.8505}{2000}}$$

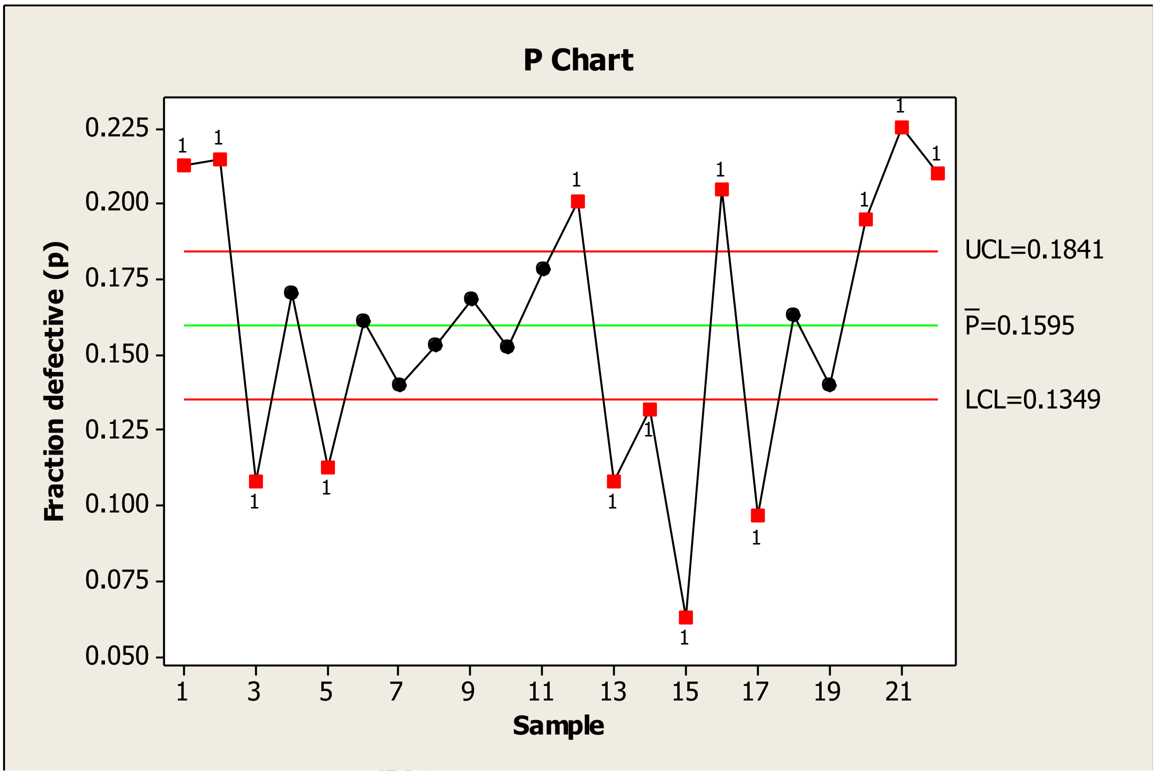
$$= 0.1595 - 3\sqrt{0.000067}$$

$$= 0.1595 - 3 \times 0.0082$$

=0.1595-0.0246

LCL=0.1349

Then we draw the p-chart based on the fraction defective values in above table.



Comment: From the above p-chart, we find that a number of points fall outside the control limits, hence the process cannot be regarded in the statistical control.

np-Chart

This chart is constructed for controlling the quality standard of attributes in a process where the sample size is equal & it is required to plot the no. of defectives (np) in samples instead of fraction defectives (p).

Example:An inspection of 10 samples of size 400 each from 10 lots reveals the following number of defectives:17, 15, 14, 26, 9, 4, 19, 12, 9, and 15

Calculate control limits for the number of defective units.
Solution: n = 400, k(no. of sample) = 10,
Total no. of defectives ($\sum d$)

$$\sum d = 17+15+14+26+9+4+19+12+9+15$$

$$\sum d = 140$$

$$n\overline{p} = \frac{\sum d}{k}$$

$$= \frac{140}{10}$$

$$n\overline{p} = 14$$

Now, $\overline{p}$ is

$$\overline{p} = \frac{n\overline{p}}{n}$$

$$= \frac{14}{400}$$

$$\overline{p} = 0.035$$

$$\overline{q} = 1-\overline{p} = 1-0.035 = 0.965$$

Central line (C.L) = $n\overline{p} = 14$
Upper control line (UCL) =

$$= n\overline{p} + 3\sqrt{n\overline{p}\overline{q}} = 14 + 3\sqrt{400\times0.035\times0.965}$$

$$= 14 + 3\sqrt{13.51}$$

$$= 14 + 3\times3.675$$

$$=14+11.025$$

UCL=25.025
Lower control line (LCL) =

$= n\bar{p} - 3\sqrt{n\bar{p}\bar{q}} = 14 - 3\sqrt{400 \times 0.035 \times 0.965}$

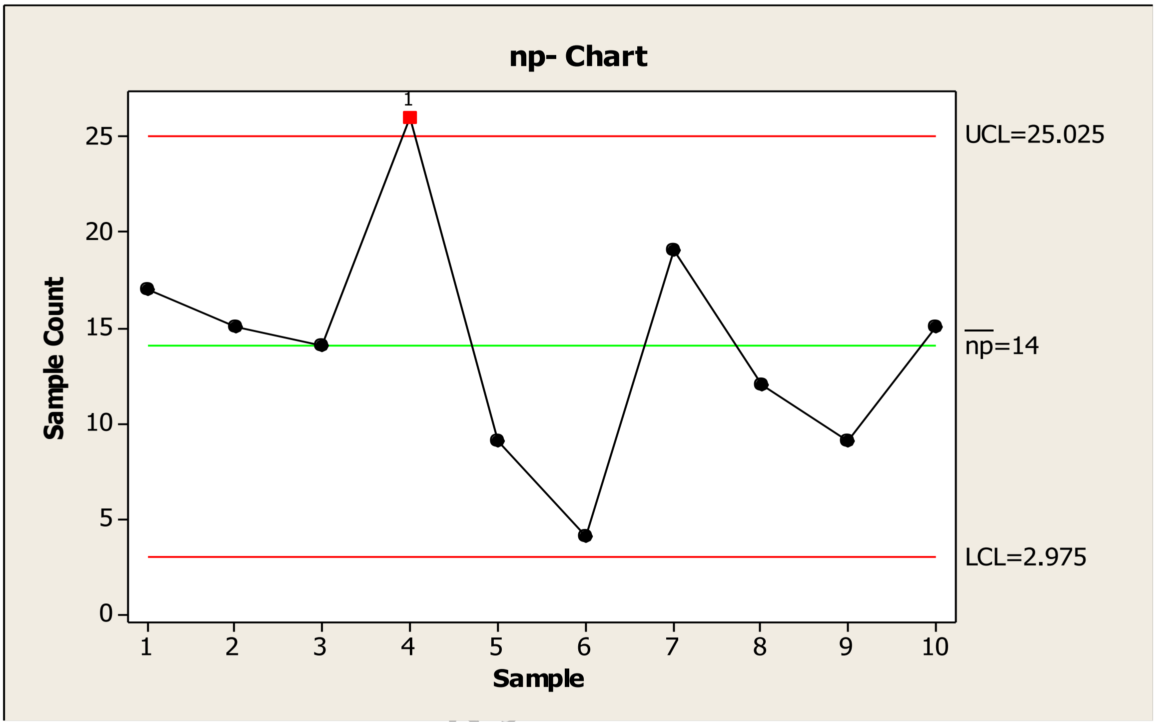
$= 14 - 3\sqrt{13.51}$

$= 14 - 3 \times 3.675$

=14-11.025

LCL=2.975

Then we draw the np-chart based on the fraction defective values in above table.



Comment: From the above np-chart, a sample point corresponding to sample number 4 lie outside the control limits, So the process is out of statistical control.

C-Chart

This chart is used for the control of no. of defects per unit say a piece of cloth/glass/paper/bottle which may contain more than one defect. The inspection unit in this chart will be a single unit of product. The probability of occurrence of each defect tends to remain very small.

Advantages of C-chart:

The following are the field of application of C-Chart

- Number of defects of all kinds of aircraft final assembly.
- Number of defects counted in a roll of coated paper, sheet of photographic film, bale of cloth etc.
- C –chart is the number of break downs at weak spots in insulation in a given length of insulated wire subject to a specified test voltage
- C-chart technique can be used with advantage in various fields other than industrial quality control, e.g., it has been applied (i) to accident statistics (both of industrial and highway accidents). (ii) in chemical laboratories, and (iii) in epidemiology

Example: In welding of seams, defects included pinholes, cracks, cold laps, etc. A record was made of the number of defects found in one seam each hour and is given below

Date	Time	No.of defect (d)	Date	Time	No.of defects (d)
1-12-83	8 am	2		12 am	6
	9 am	4		1 pm	4
	10 am	7		2 pm	9
	11 am	3		3 pm	9
	12 am	1	3-12-83	8 am	6
	1 pm	4		9 am	4
	2 pm	8		10 am	3
	3 pm	9		11 am	9
2-12-83	8 am	5		12 am	7
	9 am	3		1 pm	4
	10 am	7		2 pm	7
	11 am	11		3 pm	12
				Total	$\sum d = 144$

Draw the control chart for number of defects and give your comment.

Solution: Average number of defects per sample is

$$\bar{c} = \frac{\sum d}{n}$$

$$= \frac{144}{24}$$

$$\bar{c} = 6$$

Central line (C.L) = $\bar{c} = 6$

Upper control line (UCL) =

$$= \bar{c} + 3\sqrt{\bar{c}} = 6 + 3\sqrt{6}$$

$$= 6 + 3 \times 2.45$$

$$=6+7.35$$

UCL=13.35

Lower control line (LCL) =

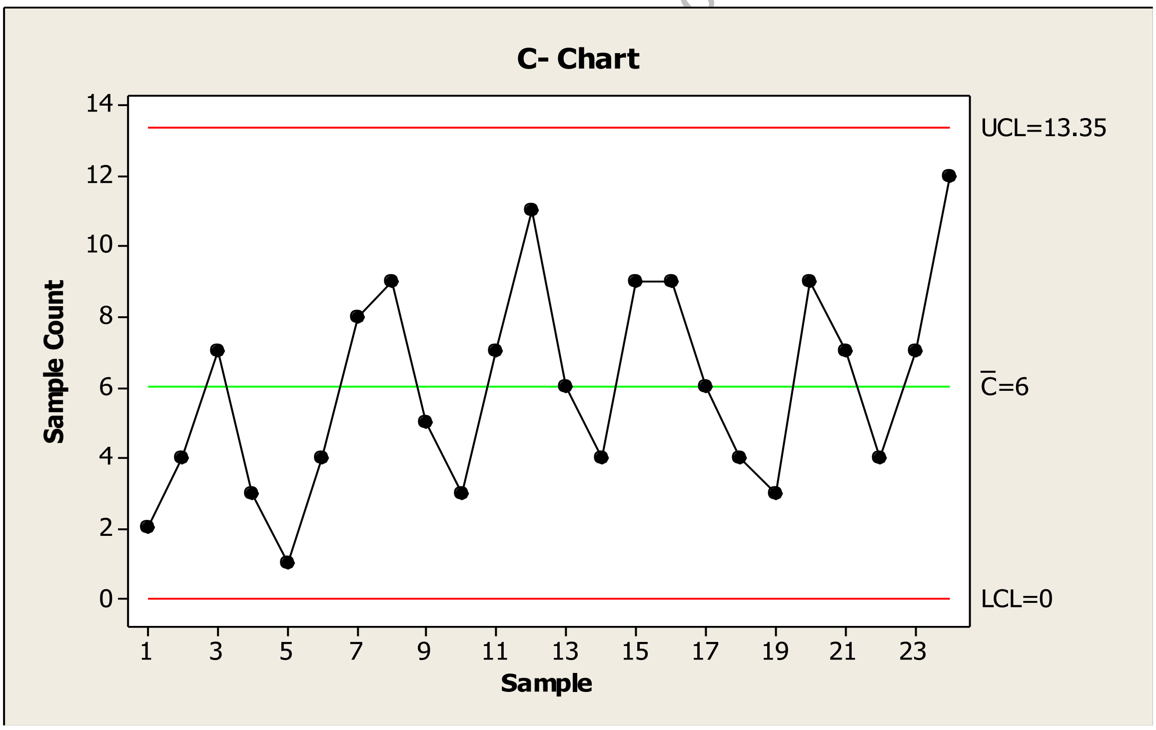
$$= \bar{c} - 3\sqrt{\bar{c}} = 6 - 3\sqrt{6}$$

$$= 6 - 3 \times 2.45$$

$$=6-7.35$$

LCL=-1.35

Then we draw the c-chart based on the fraction defective values in above table.



Comment: Since none of the 24 points falls outside the control limits, process average may be regarded in the state of statistical control.

U-Chart:u-chart measuresthe number of events defects, or non-conformities per unit or time period, and the' sample' size can be allowed to vary. In the case of inspection of cloth or other surfaces, the area examined may be allowed to vary and the *u*-chart will show the number of defects per unit area, e.g. per square meter. The design of the *u*-chart is similar to the design of the *p*-chart for proportiondefective. As in the *p*-chart, it is necessary to calculate the process average defect rate. In this case we introduce the symbol *u*.

Example: A supply chain engineering group monitors shipments of materials through the company distribution network. Errors on either the delivered material or the accompanying documentation are tracked on a weekly basis. Fifty randomly selected shipments are examined and the errors recorded. Data for twenty weeks are shown in table. Set up a *u* control chart to monitor this process.

Sample number	Sample Size	Total number of Errors	Average number of Errors
1	50	2	0.04
2	50	3	0.06
3	50	8	0.16
4	50	1	0.02
5	50	1	0.02
6	50	4	0.08
7	50	1	0.02
8	50	4	0.08
9	50	5	0.10
10	50	1	0.02
11	50	8	0.16
12	50	2	0.04
13	50	4	0.08
14	50	3	0.06
15	50	4	0.08
16	50	1	0.02
17	50	8	0.16
18	50	3	0.06
19	50	7	0.14
20	50	4	0.08
	Total	74	1.48

Solution:

$$\bar{u} = \frac{\sum u}{20}$$

$$= \frac{1.48}{20}$$

$$\bar{u} = 0.0740$$

Central line (C.L) = $\bar{u}$ = 0.074

Upper control line (UCL) =

$$= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} = 0.074 + 3\sqrt{\frac{0.074}{50}}$$

$$= 0.074 + 3 \times 0.038$$

$$= 0.074 + 0.114$$

UCL=0.19

Lower control line (LCL) =

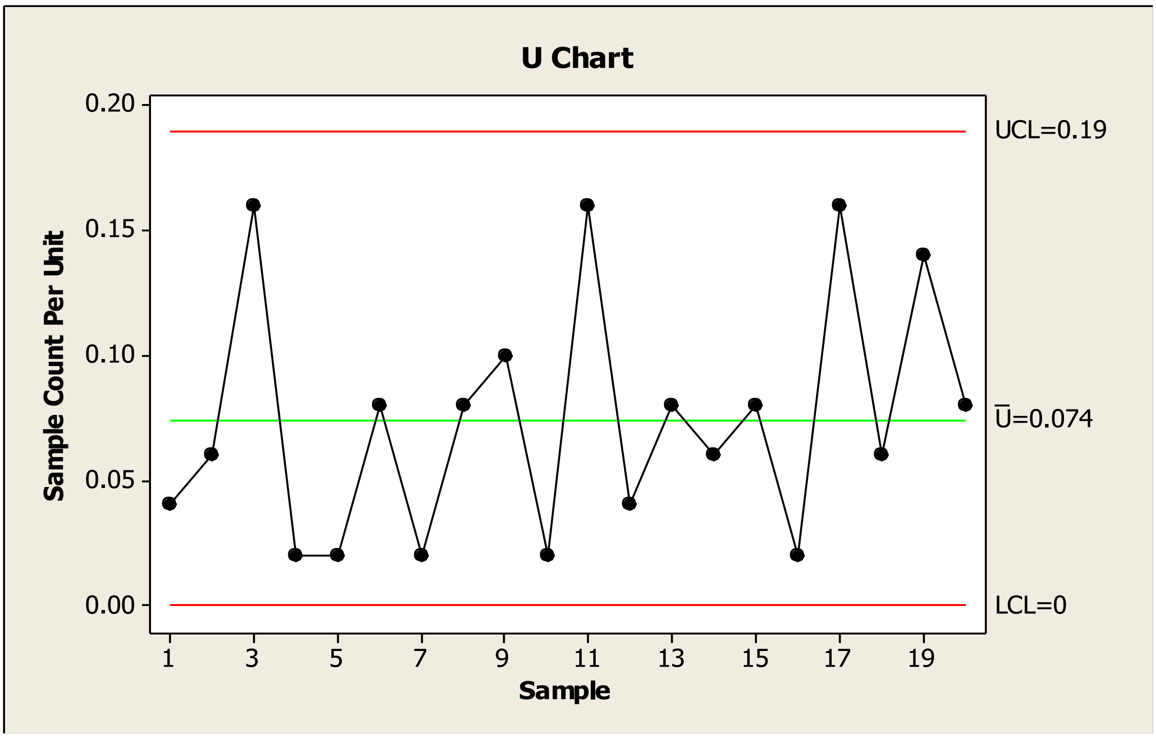
$$= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} = 0.074 - 3\sqrt{\frac{0.074}{50}}$$

$$= 0.074 - 3 \times 0.038$$

$$= 0.074 - 0.114$$

LCL=-0.04

Then we draw the u-chart based on the fraction defective values in above table.



Comment: Since none of the 24 points falls outside the control limits, process average may be regarded in the state of statistical control.

Attribute data in Non-conforming:

What is measured	*Chart name*	*Attribute charted*	*Centre-line*	*Warning lines*	*Action or control lines*	*Comments*
Number of defectives in sample of constant size n	'np' chart or 'pn' chart	np – number of defectives in sample of size n	$n\bar{p}$	$n\bar{p} \pm 2\sqrt{n\bar{p}(1-\bar{p})}$	$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})}$	n = sample size p = proportion defective $\bar{p}$ = average of p
Proportion defective in a sample of variable size	'p' chart	p – the ratio of defectives to sample size	$\bar{p}$	$\bar{p} \pm 2\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}^{*}$	$\bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}^{*}$	$\bar{n}$ = average sample size $\bar{p}$ = average value of p
Number of defects/flaws in sample of constant size	'c' chart	c – number of defects/flaws in sample of constant size	$\bar{c}$	$\bar{c} \pm 2\sqrt{\bar{c}}$	$\bar{c} \pm 3\sqrt{\bar{c}}$	$\bar{c}$ = average number of defects/flaws in sample of constant size
Average number of flaws/defects in sample of variable size	'u' chart	u – the ratio of defects to sample size	$\bar{u}$	$\bar{u} \pm 2\sqrt{\frac{\bar{u}}{\bar{n}}}^{*}$	$\bar{u} \pm 3\sqrt{\frac{\bar{u}}{\bar{n}}}^{*}$	u = defects/flaws per sample $\bar{u}$ = average value of u n = sample size $\bar{n}$ = average value of n

Uses of Control Charts

1. Helps in determining the quality standard of the products.
2. Helps in detecting the chance & assignable variations in the quality standards by setting two control limits
3. Reveals variations in the quality standards of the products from the desired level
4. Indicates whether the production process is in control or not
5. Ensures less inspection cost & time in the process control.
6. Determining whether the quality improvement project should aim to prevent specific problems or to make fundamental changes to the process
7. Predicting the expected range of outcomes from a process
8. Controlling ongoing processes by finding and correcting problems as they occur