Continuous Distribution :-

1. Normal Distribution :-

It is also called as Gaussian Distribution.

(i) Definition :-

A random variable x is said to be follow a normal distribution with parameters μ and σ, if it's pdf is given by :-

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

for $-\infty < x < \infty$ and
$-\infty < \mu < \infty$ and
$\sigma > 0$.

(ii) We also write : $X \sim N(\mu, \sigma^2)$.

(iii) $E(x) = \mu$, $Var(x) = \sigma^2$

(iv) If $\mu = 0$, $\sigma^2 = 1$, the x is said to follow a standard Normal Distribution ; $X \sim N(0, 1)$.

(v) The pdf of Standard Normal Distribution is given by :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad ; -\infty < x < \infty.$$

(vi) a] The density of a Normal distribution has it's maximum at $x = \mu$.

b] The density is symmetric around μ.

(vii) Inflexion points of density are $x = (\mu - \sigma)$ and $x = (\mu + \sigma)$.

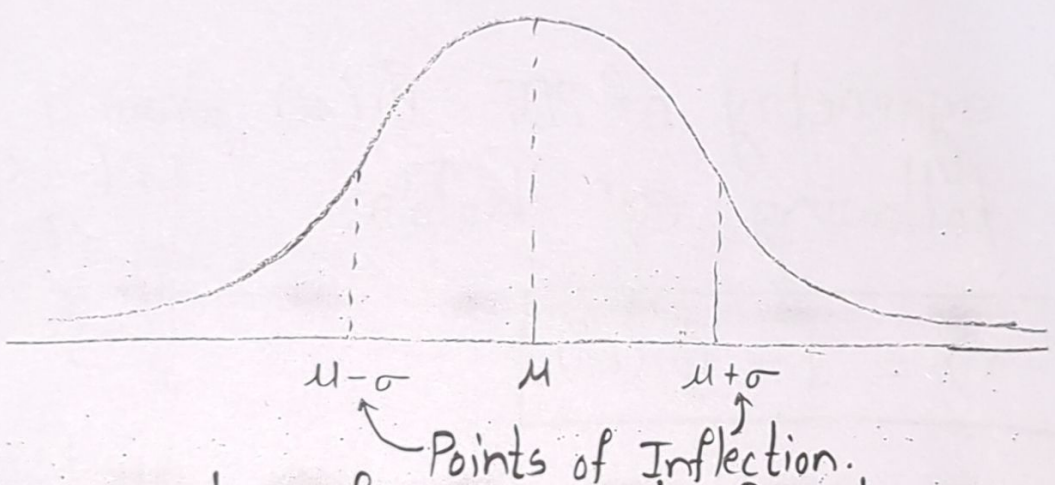
a] A Lower value of σ indicates higher concentration around μ.

b] A high value of σ indicates flatten density.

(viii) The CDF of $X \sim N(\mu, \sigma^2)$ is

$$F(x) = \int_{-\infty}^{x} \phi(t).dt = \Phi(x).$$

(ix) Graph of PDF:-



(x) Calculation rules for Normal Random Variange:

a] Let, $X \sim N(\mu, \sigma^2)$

using the transformation,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Every Normally distributed Random Variable can be transformed into a standard Normal R.V..

we call this transformation as Z transformation.

b] $P(X \leq b) = P\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right)$

$= P\left(Z \leq \frac{b-\mu}{\sigma}\right)$.

$= \Phi\left(\frac{b-\mu}{\sigma}\right)$.

c] $P(X > a) = 1 - P(X \leq a)$

$= 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$.

d] $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$.

e] Because of symmetry of PDF $\Phi(x)$ around it's mean, the following eqⁿ holds.

$$\boxed{\Phi(-a) = 1 - \Phi(a).}$$

$$\boxed{P(-a < Z < a) = 2\Phi(a) - 1.}$$

$P(-a < Z < a) =$
$P(Z \leq a) - P(Z \leq -a)$
$= \Phi(a) - \Phi(-a)$
$= \Phi(a) + \Phi(a) - 1$
$= \underline{2\Phi(a) - 1.}$

- R-Commands :-

$P(X \leq r) = P(X < r) =$ pnorm(r, μ, σ).

For, standard normal = pnorm$(r, 0, 1)$.

Q. Let, x denotes the no. of scores in a test. If x is normally distributed with mean 100 and st. deviation 15, find the prob. that x does not exceed 130.

→ $\mu = 100$. $\sigma = 15$. $\quad Z = \frac{X-\mu}{\sigma}$

$X = 130$.

$\therefore Z = 2$.

$\therefore P(X \leq 130) = \underline{0.9772}$. ← From Table. OR $\int_{-\infty}^{130} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \cdot dx$

Q. The random variable X is normally distributed with $\mu = 9$, and $\sigma = 3$. Find probability when

(i) $X \geq 15$.

(ii) $X \leq 15$.

(iii) $0 < X < 9$

(iv) $X < 15$.

→ $P(X \geq 15) = P(Z \geq 2) = \underline{0.0228}$.

$Z = \frac{15-9}{3} = 2$.

$\therefore P(X \leq 15) = \underline{0.9772}. = 1 - P(X \geq 15)$

$\therefore P(0 < X < 9) = P(X \leq 9) - P(X \leq 0)$

$= P(Z \leq 0) - P(Z \leq -3)$

$= 0.5 - 0.0013$

$= \underline{0.4987}$.

$P(X < 15) = \underline{0.9772}$.

Q. A research scientist reports that a mice will live an average of 40 months, when their diets are sharply restricted and ~~and~~ they enriched with proteins and fats. Assuming that the life time of such mice are normally distributed with $\sigma = 6.3$ months. Find prob. that the given mice will live (i) more than 32 months.
(ii) less than 28 months.
(iii) more than 37 months and/or less than 49 months.

→ $\mu = 40. \quad \sigma = 6.3.$

(i) $x = 32. \quad Z = \frac{X-\mu}{\sigma} = \frac{32-40}{6.3} = -1.27.$

$\therefore P(X > 32) = 1 - P(X \leq 32)$

$= 1 - P(Z \leq -1.27) = 1 - 0.1020$

$= 0.8980$

(ii) $P(X < 28) = P(Z < -1.90)$

$= 0.0287$

(iii) $(37 < X < 49) = P(-0.47 < Z < 1.43) = 0.6044.$

Q. The mean height of 500 students 151 cm. $\sigma = 15$ cm. Assuming that the heights are normally distributed, find how many student's height lies bet ~~150~~ 120 and 155 cm.

→ $\mu = 151 \quad \sigma = 15.$

$P(120 < x < 155) = P(-2.06 < Z < 0.27)$

$= 0.5867$

$\therefore$ No. of students $= 0.5867 \times 500$

$= 293$

Q. If x is Normal, with $\mu = 100$ and $\sigma = 5$.

Find: a] $P(95 \leq x < 110)$

b] $P(x < 50)$

c] k, if $P(x > k) = 0.3192$

d] x_1, x_2 if P

→ a] $P(95 < x < 110) = P(-1 < z < 2)$

$= 0.8185.$

b] $P(x < 50) = P(z < 10)$

$= 0.$

c] $P(x > k) = 0.3192.$

$\therefore 1 - P(x \leq k) = 0.6808$

$\therefore z = 0.47 = \frac{x - \mu}{\sigma}$ $\quad \therefore x = 102.35.$

Exponential Distribution:-

It is used when we assume that the future lifetime is independent of the lifetime that has already taken place.

Then the waiting time can be considered to be exponentially Distributed.

(i) Defⁿ:- A R.V. X. is said to follow an exponential distribution with parameters $\lambda > 0$ if it's pdf is given by,

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \text{if } x \geq 0. \\ 0 & , \text{elsewhere.} \end{cases}$$

Also, $X \sim Exp(\lambda)$.

(ii) $E(x) = \frac{1}{\lambda}$ $\quad Var(x) = \frac{1}{\lambda^2}$

(iii) CDF is,

$$F(x) = P(X \leq a) = P(X < a) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0 \\ 0 & , \text{else.} \end{cases}$$

Theorem:- The no. of events Y, occuring within a Continuum of time is Poission Distributed, with Parameter λ, IFF, the time betⁿ two events is exponentially distributed with parameter λ.

- **R-Command:**

$$P(X \leq x) = pexp(x, \lambda).$$

Memorylessness Property:-

If time 't' has already been reached, then the probability of reaching a time greater than $t+\Delta$ does not depend on 't':

i.e. $$\boxed{P(X > t+\Delta \mid X > t) = P(X > \Delta).}$$

$$P(X > t+\Delta \mid X > t) = \frac{P(X > t+\Delta \cap X > t)}{P(X > t)}$$

$$= \frac{P(X > t+\Delta)}{P(X > t)} \quad -(\Delta > 0).$$

$$= \frac{1 - P(X \leq t+\Delta)}{1 - P(X \leq t)}$$

$$= \frac{1 - (1 - e^{-\lambda(t+\Delta)})}{1 - (1 - e^{-\lambda(t)})}$$

$$= e^{-\lambda\Delta} = 1 - (1 - e^{-\lambda\Delta})$$

$$= 1 - P(X \leq \Delta)$$

$$\underline{P(X > t+\Delta / X > t)} = \underline{P(X > \Delta).}$$

Q. The length of telephone conversation is an exponential variate with mean 3; Find the prob., that a call

(i) Ends in less than 3 min.

(ii) Takes betⁿ 3 to 5 min.

(iii)

$\lambda = \frac{1}{3}$.

→ (i) $P(X \leq 3) = 1 - e^{-\lambda \times 3} = 1 - e^{-1} = \underline{\underline{\frac{e-1}{e}}}$.

(ii) $P(3 \leq X \leq 5) = (1 - e^{-5\lambda}) - (1 - e^{-3\lambda})$

$= e^{-1} - e^{-5/3} = \frac{1}{e}\left(1 - \frac{1}{e^{2/3}}\right)$

Distribution of Arithmetic Mean for Normal Distribution :-

Assume that $X \sim N(\mu, \sigma^2)$.

Consider a random sample $X = (x_1, x_2, x_3, \ldots, x_n)$ of n i.i.d. Random Variable with $x_i \sim N(\mu_i, \sigma^2)$, (Independently Identically Distributed),

then $\bar{x}$ = Arithmetic Mean, then

$E(\bar{x}) = \mu_i$ and $Var(\bar{x}) = \frac{\sigma^2}{n}$. ← (Proved Earlier).

$\text{Std. Dev.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$.

Note:- If $X_1, X_2, \ldots, X_n$ are n independent normal random variables with mean $\mu_1, \mu_2, \ldots, \mu_n$ and Variance $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ then for any real no. $a_1, a_2, \ldots, a_n$

$$a_1X_1 + a_2X_2 + \ldots + a_nX_n \sim N(a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n, a_1\sigma_1^2 + a_2\sigma_2^2 + \ldots + a_n\sigma_n^2)$$

holds.

Central Limit Theorem:-

If $\bar{x}$ is the mean of random sample of size n, taken from a population with a mean 'μ', and finite variance σ^2, then the limiting for the distribution is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \text{as } n \to \infty.$$ is Standard Normal Distribution $N(0,1)$.

(i) The sample size $n = 30$, is a guideline to use for the central limit th^m.

(ii) The normal approximation for $\bar{x}$ will be good if $n \geq 30$, provided that the population distribution is not skewed.

(iii) The population is not too different from a normal distribution then the approximation is good for $n < 30$ as well.

Sampling Distribution :-

There are 3 types of Sampling distribution.

1. χ^2 - Distribution.
2. t - Distribution.
3. F - Distribution.

χ^2 - Distribution :-

(i) Defⁿ :- Let, $Z_1, Z_2, \ldots, Z_n$ be i.i.d. Normal Random Variable. Then, the sum of their squares

$$Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$ is χ^2 - Distributed with degree of freedom 'n'. (df = degree of freedom)

(ii) $$f(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} \cdot e^{-x/2}}{\Gamma\frac{n}{2}\, 2^{n/2}}, & x \geq 0. \\ 0, & \text{elsewhere} \end{cases}$$

where Γr is gamma function

i.e.

$$\Gamma r = \begin{cases} (r-1)!, & \text{for } r = +ve \text{ integer} \\ \int_0^\infty t^{r-1} \cdot e^{-t}\, dt, & \text{elsewhere.} \end{cases}$$

(iii) χ^2 - Distribution is <u>not</u> symmetric. <u>(skewed graph)</u>.

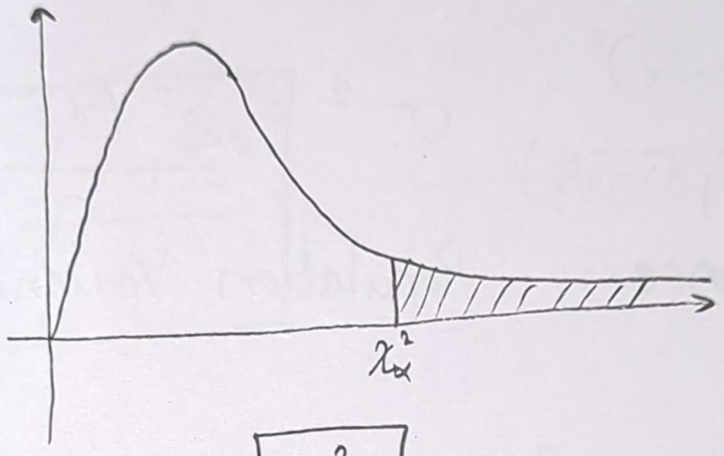
(iv) χ^2 - lies bet^n 0 to ∞.

(v) Mean = $E(X) = df$

$Var(X) = 2.df.$

(vi) Small Value of df gives high value of Tail.

<u>Graph:</u>



(vii) <u>Theorem</u> :- $\boxed{\chi^2_{df}}$

Consider two independent random variables which are χ^2_m and χ^2_n distributed, then the sum of these two R.V. is χ^2_{m+n} distributed.

Q. Using stastical tables find

(i) $\chi^2_{0.025}$ when $df = 15$. 27.488

(ii) $\chi^2_{0.05}$ when $df = 25$. 37.652.

Q. Find the prob. that a random sample of 25 observations from a normal population with Var. 6 will have a variance (i) s^2 greater than 9.1

(ii) s^2 lies bet^n 3.462 and 10.745.

P.T.O. ⟶

Theorem:- A χ^2-distributed random variable, having sample variance S^2 in an i.i.d. sample of size n from a normally distributed population is given by

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Formulas:-

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

Sample Variance. Population Variance.

- **Proof:-**

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

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

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

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

$$= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2+(\bar{x}-\mu)^2\times n+2(\bar{x}-\mu)(0)\right\}$$

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

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

$$\sum_{i=1}^{n} z_i^2 = \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$$

$$= \frac{(n-1)s^2}{\sigma^2} + Z^2 \quad \text{—(By Central Limit Th}^m\text{)}.$$

$$\therefore \sum_{i=1}^{n} z_i^2 - Z^2 = \frac{(n-1)s^2}{\sigma^2}.$$

$$\Rightarrow \boxed{\chi^2_{n-1} \sim \frac{(n-1)s^2}{\sigma^2}} \quad \text{—(Hence, Proved).}$$

Q. → Continued. $n = 25$. $\sigma^2 = 6$.

3. $P(s^2 > 9.1) = P\left(\frac{(n-1)s^2}{\sigma^2} > \frac{9.1 \times (n-1)}{\sigma^2}\right)$

$$= P\left(\chi^2_{n-1} > \frac{9.1 \times 24}{6}\right) = P\left(\chi^2_{n-1} > 36.4\right)$$

$$= 0.05.$$

5] $P(9.462 < s^2 < 10.745) = P\left(13.848 < \chi^2_{24} < 42.980\right)$

$$= P\left(\chi^2_{24} > 13.848\right) - P\left(\chi^2_{24} > 42.980\right)$$

$$= 0.95 - 0.01$$

$$= 0.94.$$

Q. An electrical firm manufactures light bulbs that have a length of light that is approximately normally distributed with mean = 800 hrs. and st. dev. = 40 hrs. Find the prob. that a random sample of 16 bulbs will have an average life of less than 775 hrs.

→ $\mu = 800. \quad \sigma = 40. \quad P(\bar{X} < 775) = ?$

$$P(\bar{X} < 775) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{775 - 800}{40/4}\right) = P(Z < -2.5)$$

$$= \underline{0.0062.}$$

□ t- Distribution :-

(i) Let, X and Y be two independent random variables such that $X \sim N(0,1)$ and Y is χ_n^2- distributed, then the ratio $\frac{X}{\sqrt{Y/n}} \sim t_n$.

(ii) pdf is given by :-

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\frac{n}{2}} \cdot \sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \quad -\infty < x < \infty.$$

(iii) **Student's Theorem:-**

Let, $X = (X_1, X_2, \ldots, X_n)$ with $X_i \overset{i.i.d}{\sim} N(\mu, \sigma^2)$

then the ratio $\left(\frac{\bar{x}-\mu}{s}\right)\sqrt{n} = \frac{(\bar{x}-\mu)\sqrt{n}}{\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t_n$, i.e.

is t-distributed with (n-1) degree of freedom.

(iv) If Sample Size, $n < 30$ then we use t-Distribution and for $n \geq 30$, we use Normal Distribution.

(v) $t_\alpha = -t_{1-\alpha}$

Q. Find t-value with dof 14 leaving an area of 0.025 to the left.

→ $\mu = 14$. $\alpha = 1 - 0.025 = 0.975$.

~~$\therefore t_\alpha = 5.629$.~~

$t_\alpha = -t_{(1-\alpha)} = -t_{(0.025)}$

$t_\alpha = -2.145$

Q. $P(-t_{0.025} < T < t_{0.05}) = P(t_{0.975} < T < t_{0.05})$

$= P(-t_{0.025}) - P(t_{0.05}) = P(t_{0.975}) - P(t_{0.05})$

$= P(T > -2.145) - P(T <) = 0.975 - 0.05$

VIMP

$= 0.925$

Q. Find K s.t. $P(K < T < -1.761) = 0.045$ for a random sample of size 15 from a normal population.

→ $\mu = (15-1) \times 30$. Using t-distribution, $\mu = 15 - 1 = 14$.

$$P(T > K) - P(T < -1.761) = 0.045$$

$$P(T > K) - P(T < -t_{(0.05)}) = 0.045$$ VIMP

$$P(T > K) - (1 - 0.05) = 0.045.$$

$$P(T > K) = 1 + 0.045 - 0.05$$

$$= 1 - 0.005$$

$$P(T > K) = 0.995.$$

μ = No. of samples $- 1$

$\mu = n - 1$

~~$P(T > t_K = 0.995 = t_{(1-K)})$.~~

~~$1 - K =$~~ $K = -2.977$

More Easy Way ~

$$P(K < T < t_{0.95}) = 0.045.$$

$\downarrow$

t_α.

$$\alpha - \beta = 0.045.$$

$$\alpha = 0.995.$$

$$K = t_\alpha = -t_{1-\alpha}.$$

$$K = -t_{(0.005)} = -2.977.$$

F-Distribution:-

(i) Let, X and Y be χ^2_m and χ^2_n - distributed R.V.s, then the distribution ratio $\frac{X/m}{Y/n} \sim F_{m,n}$ distributed, with (m,n) dof's.

(ii) pdf is given by:-

$$f(x) = \frac{\Gamma\left(\frac{(n+m)}{2}\right)\left(\frac{n}{m}\right)^{n/2} x^{\left(\frac{n}{2}-1\right)}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)\left(1+\frac{nx}{m}\right)^{\left(\frac{n+m}{2}\right)}}, \quad x>0.$$

(iii)

Q. Find $f_{0.05}$ when dof are 6 and 10.
(ν_1 = 6, ν_2 = 10)

→ $f_{0.05}(6,10) = \underline{3.22}$.

- R-Commands for Quantile:-
  - qchisq(p, df)
  - qt(p, df)
  - qf(p, df_1, df_2)

Inferences :-

- **Simple Random Sample :-**

It is a sample in which, each voter has an equal probability of being selected in the sample, and is independently chosen from sample population.

- **Parameters of Population :-**

It is a numerical value that gives a characteristics of entire population. Denoted by $\bar{P}$.

- **Sample Estimates :-**

Numerical Values calculated from a sample that provides estimates or approximations of population parameters.

- **Statistics :-**

- A f^n of R.V. is called statistic. It is denoted by $T(x)$.
- A statistic is also a R.V.
- Statistic is used to Estimate a population parameter,
  i.e. $T(x)$ is an estimator of θ.
  i.e. $T(x) = \hat{\theta}$

- **Unbiased Estimator :-**

An Estimator $T(x)$ is unbiased if $\boxed{E_\theta(T(x)) = \theta}$

The Bias of an Estimator is

$$\boxed{\text{Bias}_\theta(T(x)) = E_\theta(T(x)) - \theta.}$$

An Estimator is said to be unbiased, if it's bias is zero. $\underline{Bias_\theta(T(x)) = 0}$.

- **Variance of $T(x)$:-**

$$Var_\theta(T(x)) = E\left\{[T(x) - E(T(x))]^2\right\}$$

- **Mean Square Error (MSE):-**

$$MSE_\theta(T(x)) = E[T(x) - \theta]^2$$

Also:

$$MSE_\theta(T(x)) = Var_\theta(T(x)) + [Bias_\theta(T(x))]^2.$$

□ **Theorem:-**

Let, $X = (x_1, x_2, \ldots, x_n)$ be an i.i.d. sample of sample variable X with population mean μ, and Population var. σ^2, then the A.M. i.e. $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ is an unbiased estimator of μ.

And sample var. $s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$ is an unbiased estimator of σ^2.

- Population Mean :- μ.
- Sample Mean :- $\overline{x}$
- Population Variance :- σ^2
- Sample Variance :- s^2.

• To Prove:- (i) $E(\bar{x}) = \mu$.

(ii) $E(s^2) = \sigma^2$

(i) L.H.S. $= E(\bar{x}) = \frac{1}{n}\sum$

$= E\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n}\left[\frac{E(x_1) + E(x_2) + \ldots + E(x_n)}{}\right]$

$= \frac{n\mu}{n}$

$= \mu$.

L.H.S. = R.H.S.

(ii)

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}\left[(x_i-\bar{x}) + (\bar{x}-\mu)\right]^2$$

$$= \sum_{i=1}^{n}\left[(x_i-\bar{x})^2 + (\bar{x}-\mu)^2 + 2(x_i-\bar{x})(\bar{x}-\mu)\right]$$

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

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + (\bar{x}-\mu)^2 n + (0).$$

Dividing Both sides by $(n-1)$.

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

$\frac{n-1}{\downarrow}$

s^2

$$\therefore S^2 = \frac{n}{n-1}\,\$ - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n-1}$$

Taking Expectation of Both sides,

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

$$E(S^2) = \frac{n}{n-1}\sigma_{\bar{x}}^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sigma_{\bar{x}_i}^2$$

Extreme V.V.IMP.

$$\left[\text{Var}(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \;;\; \frac{\sum \sigma_{\bar{x}_i}^2}{n} = \sigma^2\right].$$

$$E(S^2) = \frac{n}{n-1} \times \frac{\sigma^2}{n} - \frac{n}{n-1}\cdot\sigma^2 = \sigma^2\left[\frac{1}{n-1} - \frac{n}{n-1}\right] = 1\times\sigma^2.$$

$\therefore E(S^2) = \sigma^2$ — (Hence, Proved).

Q. Let, $X_1, X_2, \ldots, X_n$ be i.i.d. ~~of~~ sample of size n, with population mean μ and population variance σ^2.

Prove :- (i) $\tilde{\bar{x}} = \bar{x} + 1 = \frac{\sum(x_i+1)}{n}$ is Biased Estimator of μ.

(ii) $\tilde{S}^2 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}$ is Biased Estimator of σ^2.

→ (i) $E(\tilde{\bar{x}}) = E(\bar{x}+1) = E(\bar{x}) + E(1) = \mu + 1 \neq \mu$.

$\therefore$ Biased Estimator.

(ii) $E(\tilde{S}^2) = E\left(\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n} \times \frac{n-1}{n-1}\right) = E\left(\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n-1} \times \frac{n-1}{n}\right) = E\left(S^2 \times \frac{(n-1)}{n}\right)$

$= \frac{n-1}{n}E(S^2) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2$. $\therefore$ Biased Estimator.

Hypothesis

One sample Test:-

Data is assumed to arise as one sample from defined population.

Two sample Test:-

Data originates in the form of two samples from two different Population.

→ Two Independent Sample Problem.

→ Two Dependent Sample Problem.

Hypothesis:- A claim that we want to test.

i) **Null Hypothesis:-**

Currently accepted value for a parameter. It is opposite of Alternative Hypothesis; denoted by H_0.

i) **Alternative Hypothesis:-**

Involves the claim to be tested. Denoted by H_1 / H_a. It is formulated as deviation from the target value.

Ex: A company has stated that their straws machine makes a straws that are 4 mm diameter. A worker believes the machine no longer makes straws of the size and samples 100 straws to perform a Hypothesis test with 95% confidence.

→ H_0 : $\mu = 4$ mm. = straw diameter.

H_1 : $\mu \neq 4$ mm.

Ex: Doctors believe that the average teen sleeps on an average less than 10 hrs. per day. A researcher believes that teens on average sleep longer. Write H_0 and H_1.

→ H_0 : $\mu \leq 10$ hrs.

H_1 : $\mu > 10$ hrs.

- One and Two Sided Tests:-

For an unknown population Parameter (θ) and a fixed value θ_0, we have:-

Case	H_0	H_1	
A	$\theta = \theta_0$	$\theta \neq \theta_0$	Two sided Test
B	$\theta \geq \theta_0$	$\theta < \theta_0$	One sided test.
C	$\theta \leq \theta_0$	$\theta > \theta_0$	One sided test.

- Type I and Type II errors:-

Decision / Fact	H_0 is True.	H_0 is not True.
H_0 is not Rejected.	Correct Decision.	Type II Error.
H_0 is Rejected.	Type I Error.	Correct Decision.

(i) Type I Error :-

The Hypothesis H_0 is True, but is Rejected.
i.e. H_1 is accepted.

(ii) Type II Error :-

The hypothesis is not rejected, although it is not True.

(a) Significance Level) :-

Probability of Type I Error, i.e.

$$P(H_1 | H_0) = \alpha$$

i.e. Probability of accepting H_1 if H_0 is True.

(b) Level of Confidence :-

Probability of Type II Error.

$$P(H_0 | H_1) = \beta.$$

- We try to fix α and then minimize β.

- Power of Test :-

$$1 - \beta = P(H_1 | H_1)$$

i.e. Prob. of making a decision in favor of the research hypothesis H_1, if it is true. i.e.
the Prob. of detecting a correct research hypothesis.

Steps to Conduct a stastistical Test:-

Step 1:

(i) Define the Assumptions for random variables of interest and specify them in terms of population Parameters. (i.e. θ or μ or σ).

(ii) Formulate Null Hypothesis and Alternative Hypothesis.

(iii) Fix a significance value / level i.e. α.

Step 2:- Consider a stastic, $T(x) = T(x_1, x_2, \ldots, x_n)$ The distribution has to be known under Null hypothesis.

Step 3: Consider a Critical region, K for the statistics T, i.e. a region where if T falls in this region - H_0 is rejected, such that

$$P_{H_0}(T(x) \in K) \leq \alpha$$

Step 4:-

Find $t(x) = T(x_1, x_2, \ldots, x_n)$ based on sample value $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

Step 5:- Decision Rule:-

If $t(x)$ falls into critical region K, then null Hypothesis H_0 is rejected.

ie. $t(x) \in K$: H_0 is rejected.

$\Rightarrow H_1$ is significant.

If $t(x)$ falls outside K, null hypothesis is not rejected.

ie. $t(x) \notin K$; H_0 is not rejected.

$\Rightarrow H_0$ is accepted and is significant.

□ Tests:-

[I] One-Sample Gauss Test (z-test):-

This test is used ~~when~~ to test whether unknown mean differs from a specific value of mean of sample.

- We assume $\sigma^2 = \sigma_0^2$ is given.

◇ Steps:-

1. (i) The random variable X follows a $N(\mu, \sigma^2)$, distribution with known variance σ_0^2.

   (ii) formulating H_0 and H_1.

   $H_0: \mu = \mu_0$; $H_1: \mu \neq \mu_0$ (Two sided test).

   $H_0: \mu \leq \mu_0$; $H_1: \mu > \mu_0$ (One sided test).

   $H_0: \mu \geq \mu_0$; $H_1: \mu < \mu_0$ (one sided test).

   (iii) $\alpha = 0.05$ (used as default if not given in question).

2. Constructing a test statistics:-

   We know that if X_i's are i.i.d. then Sample Mean is Normally Distributed.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right)$$

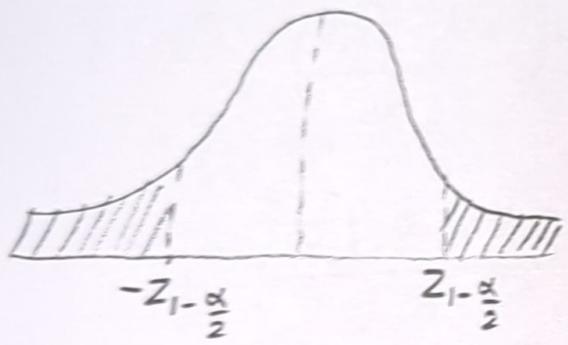
$$\therefore T(x) = \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma_0} \sim N(0, 1).$$

3. Critical Region.

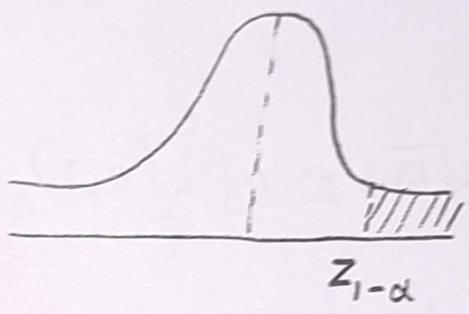
   Consider the following table.

Case	H_0	H_1	Critical Region.
Two sided a.	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$
One sided b.	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (z_{1-\alpha}, \infty)$
One sided c.	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, z_\alpha = -z_{1-\alpha})$.

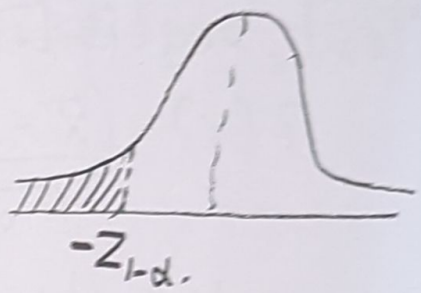
a)



b)



c)



4. Realization of Test statistics:

For observed samples $x_1, x_2, \ldots, x_n$ the arithmetic mean $\bar{x} = \frac{\sum x_i}{n}$ is used to calculate realized test statistic, $t(x)$ ie

$$t(x) = T(x_1, x_2, \ldots, x_n)$$
$$= \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma_0}$$

5. Decision Rule:-

If $t(x) \in K \Rightarrow H_0$ is Rejected.
$\Rightarrow H_1$ is Accepted.

If $t(x) \notin K \Rightarrow H_0$ is Accepted.
$\Rightarrow H_1$ is Rejected.

Q. A sample of 100 tires is taken from a lot. The mean life of tires in the sample was found to be 39350 kms with the population standard deviation of 3260 km. Test the hypothesis, at 1% level of significance that the mean life of tire is 40,000 kms.

→ (1) $n = 100$. $\bar{x} = 39350$ kms $\sigma = 3260$ kms. $\alpha = 0.01$.

$H_0 : \mu = 40000$ kms

$H_1 : \mu \neq 40000$ kms.

2. Test statistic:

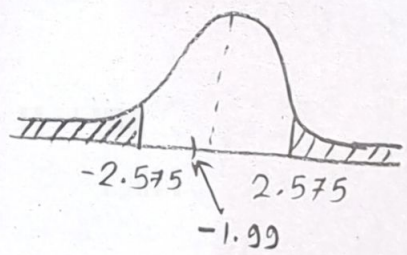
$$T(x) = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma_0} \sim N(0,1).$$

3. We take third critical region.

$$K = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$$

$$= (-\infty, -z_{0.995}) \cup (z_{0.995}, \infty)$$

$$K = (-\infty, -2.575) \cup (2.575, \infty)$$



4. $$t(x) = \left(\frac{\bar{x} - \mu_0}{\sigma_0}\right)\sqrt{n} = \left(\frac{40000 + 39350}{3260}\right) \times 10 = -1.99$$

5. $t(x) \notin K$. ∴ H_0 is accepted. H_1 is rejected.

Q. A random sample of 100 recorded deaths in the US during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years does this seem to indicate that mean life span today is greater that 70 years? Use a 0.05 level of significance.

→ 1. $n = 100$. $\bar{x} = 71.8$ years. $\sigma = 8.9$ years. $\alpha = 0.05$.

$H_0 : \mu \leq 70$ years.

$H_1 : \mu > 70$ years.

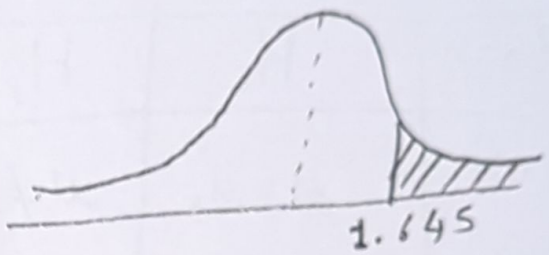
2. Test stastic:-

$$T(x) = \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma_0} \sim N(0,1).$$

3. Critical Region:-

$$K = (Z_{1-\alpha}, \infty).$$

$$K. = (Z_{0.95}, \infty). = (1.645, \infty).$$



4. $t(x) = \left(\frac{\bar{x} - \mu_0}{\sigma_0}\right)\sqrt{n}$

$$= \left(\frac{71.8 - 70}{8.9}\right) \times 10 = 2.02247.$$

5. $t(x) \in K. \Rightarrow H_1$ is <u>accepted</u>. H_0 is <u>Rejected</u>.

[II] t-test :-

This test is used when σ^2 is unknown.

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

  Population variance, $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$.

- <u>Steps</u> :-

1. (i) (ii) (iii) } → Same as Z-test.

2. Constructing a test statistic:-

$$T(x) = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S_x} \quad \text{with } n-1 \text{ degrees of freedom.}$$

3. Critical Region:- (degree of freedom, $\nu = n-1$)

Case	H_0	H_1	K.
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty, -\lvert t_{1-\frac{\alpha}{2}} \rvert) \cup (\lvert t_{1-\frac{\alpha}{2}} \rvert, \infty)$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$(-\infty, -\lvert t_{1-\alpha} \rvert)$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$(\lvert t_{1-\alpha} \rvert, \infty)$.

4. Realization of test statistics:-

For an observed sample $x_1, x_2, \ldots, x_n$

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{and} \quad t(x) = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S_x}$$

5. Decision Rule.

(Same as z-test).

Q. A manufacturer of a certain brand of energy bar claims that the average saturated fat content in the bar is 0.5 gms. Will you support his claim if the 8 bars that you examined for fat content were found to contain 0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4 and 0.2 gms of saturated fats? Take $\alpha = 0.05$.

→

1. H_0: $\mu = 0.5$ gms. $\quad \alpha = 0.05$

H_1: $\mu \neq 0.5$ gms. $\quad n = 8 \quad \nu = 7.$

$$\bar{x} = \frac{0.6+0.7+0.7+0.3+0.4+0.5+0.4+0.2}{8} = 0.475.$$

2. $T(x) = \left(\frac{\bar{x} - \mu_0}{S_x}\right)\sqrt{n}.$

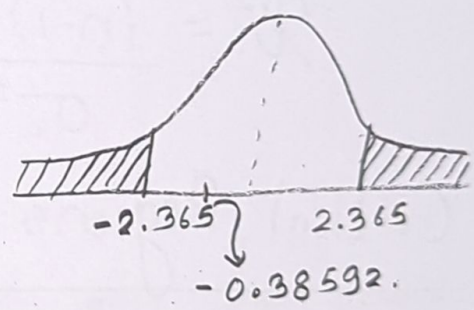
3. Critical Region:

$$K. = (-\infty, -|t_{1-\frac{\alpha}{2}}|) \cup (|t_{1-\frac{\alpha}{2}}|, \infty)$$

$$= (-\infty, -|t_{0.975}|) \cup (|t_{0.975}|, \infty)$$

$$= (-\infty, -|-t_{0.025}|) \cup (|-t_{0.025}|, \infty)$$

$$K. = (-\infty, -(2.365)) \cup (2.365, \infty).$$



4. $t(x) = \left(\frac{\bar{x} - \mu_0}{S_x}\right)\sqrt{n} = \left(\frac{0.475 - 0.5}{0.183225}\right)$ $\quad S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$t(x) = \underline{-0.38592}$

5. Decision :-

$t(x) \notin K.$ $\quad \therefore$ H_0 is Accepted.

H_1 is Rejected.

[III] χ^2-Test:

- This test is used when $\sigma^2 = \sigma_0^2$ is given.
- We assume the Distribution of the Population Sample is Normal.

- **Steps:-**

1. (i) (ii) (iii) } → Same as Previous tests.

2. **Constructing a test statistics:-**

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

3. Critical Regions:-

Case	H_0	H_1	K.
(a)	$\theta = \theta_0$	$\theta \neq \theta_0$	$K = (0, \chi^2_{1-\frac{\alpha}{2}}) \cup (\chi^2_{\frac{\alpha}{2}}, \infty)$
(b)	$\theta \geq \theta_0$	$\theta < \theta_0$	$K = (0, \chi^2_{1-\alpha})$
(c)	$\theta \leq \theta_0$	$\theta > \theta_0$	$K = (\chi^2_{\alpha}, \infty)$

4. Realization of test stastics:-

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

5. Decision:

Same as Previous tests.

Q. A manufacturer of car batteries claims that the life of a company's battries is approx. normally distributed with a standard deviation of 0.9 years. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think $\sigma > 0.9$ years? use $\alpha = 0.05$.

→ 1. H_0 : $\sigma \leq 0.9$ years. $\quad \alpha = 0.05$

$\sigma > 0.9$ years. $\quad s = 1.2$ years. $\quad n = 10$. $\quad \nu = 9$.

2. ~~$T(x) =$~~ $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

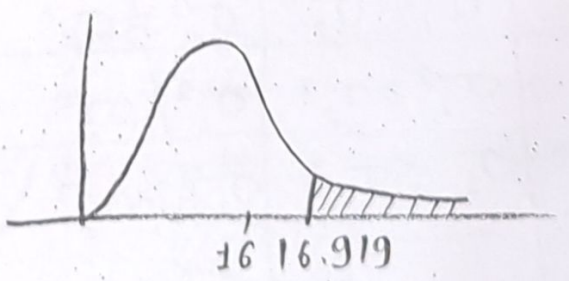
3. Critical Region.

$K = (\chi_\alpha^2, \infty)$

$K = (\chi_{0.05}^2, \infty) = (16.919, \infty)$

4. Realization of test stastics:-

$\chi^2 = \dfrac{(10-1) \times (1.2)^2}{(0.9)^2} = 16.$



5. Decision:-

$\chi^2 \notin K^2$

$\therefore$ H_1 is Rejected $\quad$ H_0 is Accepted

[IV] f-test :-

* We use this test for comparing variance i.e. whether $\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$.
* We assume the population to be normally Distributed.

Steps :-

1. (i) (ii) (iii) } → Same as Previous tests.

2. Constructing a Test statistics :-

$$f = \frac{S_1^2}{S_2^2}$$. Also, we have : $f_\alpha(\nu_1, \nu_2) = \frac{1}{f_{1-\alpha}(\nu_2, \nu_1)}$.

3. Critical Regions :

Case	H_0	H_1	Region (K).
(a)	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$(0, f_{1-\frac{\alpha}{2}}(\nu_1, \nu_2)) \cup (f_{\alpha/2}(\nu_1, \nu_2), \infty)$
(b)	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$(0, f_{1-\alpha}(\nu_1, \nu_2))$
(c)	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$(f_\alpha(\nu_1, \nu_2), \infty)$

4. Realization of test statistics :-

$$f = \frac{S_1^2}{S_2^2}$$

5. Decision :-

Same as Previous tests.

Q. A study is conducted to compare the lengths of time required by men and women to assemble a certain product. Past experience indicates that the distribution of times for both men and women is approx. normal but variance of times for a women is less than that of a man. A random sample of times for 11 men and 14 women gives resp. standard dev. 6.1 and 5.3 Test the hypothesis $\sigma_1^2 = \sigma_2^2$ against $\sigma_1^2 > \sigma_2^2$.

M W.

1. $n_1 = 11$, $n_2 = 14$. $\nu_1 = 10$, $\nu_2 = 13$.
   $S_1 = 6.1$ $S_2 = 5.3$ $\alpha = 0.05$ - (default).

$H_0 : \sigma_1^2 \leq \sigma_2^2$

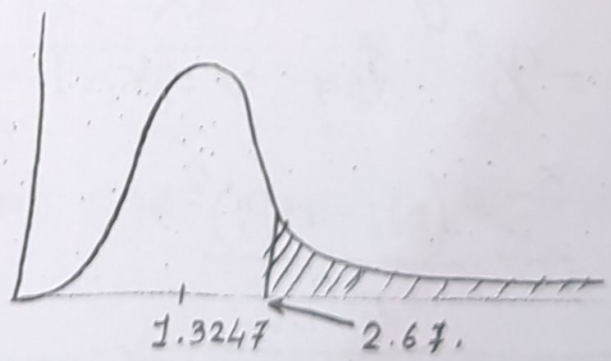
$H_1 : \sigma_1^2 > \sigma_2^2$.

2. $f = \frac{S_1^2}{S_2^2}$.

3. $K = (f_\alpha(\nu_1, \nu_2), \infty) = (f_{0.05}(10,13), \infty)$

$K = (2.67, \infty)$.

4. $f = \frac{(6.1)^2}{(5.3)^2} = 1.3247$.



5. $f \notin K$. $\therefore$ H_1 is Rejected.
   $\therefore$ H_0 is Accepted.

□ Note:- Consider two samples X and Y in i.i.d. of size m and n from normal population, and their sample variance are given by $S_x = \frac{1}{m-1}\sum(X_i - \bar{X})^2$ and $S_y = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$ then $\frac{S_x^2}{\sigma^2} \sim F_{(m-1),(n-1)}$

[V] χ^2 Goodness of fit:-

- used when observed absolute frequencies are compared with the expected absolute frequencies under H_0.

Steps:-

1. (i) (ii) (iii) } → Same as previous tests.

2. Test Statistic:-

$$T(x) = \chi^2 = \sum_{i=1}^{K} \frac{(N_i - np_i)^2}{np_i},$$

where N_i are the absolute frequencies of observations of the sample x_i in ~~class~~ class i, N_i is a random variable with realization n_i in the observed sample.

3. Critical Region:

$K = \chi^2_\alpha$ for $\nu = K-1$ degree of freedom.

4. $\chi^2 = \sum_{i=1}^{K} \frac{(N_i - np_i)^2}{np_i}$

5. If $\chi^2 > K$, then H_0 is rejected.
H_1 is Accepted.

Q. Gregor Mendel conducted crossing experiment with pea plants of different shape and color. Let, us look at the outcome of a pea crossing experiment with the following result:-

Crossing Result	RY	RG	EY	EG.
Observations	315	108	101	32.

$n = 556.$

Mendel has hypothesis that the four different types occur in proportions of 9:3:3:1 i.e.

$$P_1 = \frac{9}{16}, \quad P_2 = \frac{3}{16}, \quad P_3 = \frac{3}{16}, \quad P_4 = \frac{1}{16}.$$

→ 1. $H_0 : P(X=i) = P_i \qquad i = 1,2,3,4 \qquad \alpha = 0.05$

$H_1 : P(X=i) \neq P_i \qquad \nu = 4-1 = 3 = K-1.$

2. $$\chi^2 = \sum_{i=1}^{K} \left(\frac{N_i - nP_i}{nP_i}\right)^2 = \sum_{i=1}^{K} \frac{(N_i - nP_i)^2}{nP_i}$$

3. $K = \chi_\alpha^2 = 7.815$

4. $$\chi^2 = \frac{\left(315 - 556 \times \frac{9}{16}\right)^2}{556 \times \frac{9}{16}} + \frac{\left(108 - 556 \times \frac{3}{16}\right)^2}{556 \times \frac{3}{16}} + \frac{\left(101 - 556 \times \frac{3}{16}\right)^2}{556 \times \frac{3}{16}} + \frac{\left(32 - 556 \times \frac{1}{16}\right)^2}{556 \times \frac{1}{16}}$$

$$= \frac{(2.25)^2}{34.75 \times 9} + \frac{(3.75)^2}{34.75 \times 3} + \frac{(3.25)^2}{34.75 \times 3} + \frac{(2.75)^2}{34.75}$$

$$= \frac{1}{34.75}\left[\frac{(2.25)^2}{9} + \frac{(3.75)^2}{3} + \frac{(3.25)^2}{3} + (2.75)^2\right] = \frac{16.333}{34.75}$$

$\chi^2 = 0.47002$

5. $0.47 < 7.815.$

$\therefore H_0$ is Accepted. H_1 is Rejected.

□ Test Decisions Using Confidence Intervals:-

1. If H_0 is Rejected at the significance level α, then there exist a $100(1-\alpha)\%$ confidence interval which yields the same conclusion as the test.

2. This is called as "Duality".

eg. Consider 2 drugs A and B and we want to find whether average changes due to drug B is greater (higher) than for drug A i.e. $H_1 : S_B > S_A$.

So we want to decide whether H_1 is significant or not.

This is equivalent to constructing $100(1-\alpha)\%$ confidence interval for $S_B - S_A$ and checking whether H_0 is significant or not.

◈ Conditions when Tests are used :-

1. Z-test: (i) When σ is known.
   (ii) $n \geq 30$ or Data is Normally distributed.
   (iii) Comparing the sample mean to a known mean of population.

2. t-Test :- (i) when σ is unknown.
   (ii) $n < 30$ or Data is Approx. Normally Distributed.

3. χ^2-Test:- When Distribution is Normal and fixed variance/ standard deviation is given.

4. f-test:- When variance need to be compared.

5. χ^2 Goodness of fit Test :- Used when observed ~~when observed~~ Absolute frequencies are compared with the expected absolute freq. under H_0.

Linear Regression :- → fitting

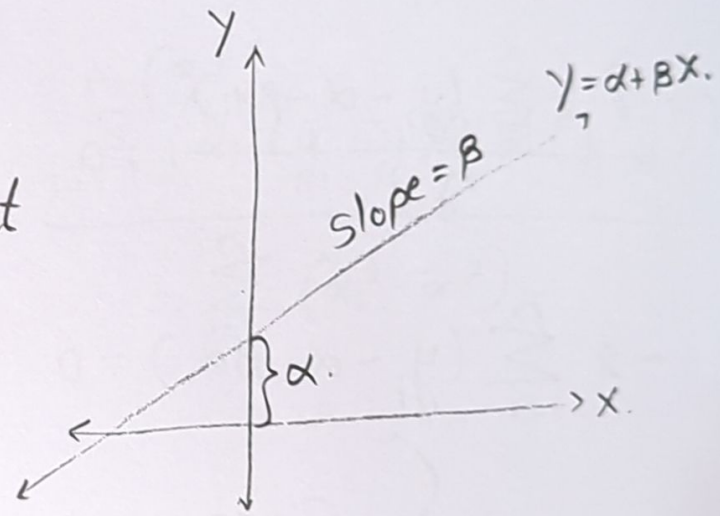
1. Variables (X, Y, ..)

2. Regression Parameters. (a, b, α, β, ..)

Linear Model :-

$$Y = \alpha + \beta X.$$

- Y : Dependent Variable (Response)
- α : y-Intercept
- β : Slope
- X : Independent Variable (covariate).



Method of Least Square :-

Consider n sets of observation, given by $P_i = (x_i, y_i)$, $i = 1, 2, \ldots n$ obtained on two variables $P = (X, Y)$.

The method of least squares says that, a line can be fitted to the given data set such that the errors are minimized. we need to determine α and β, denoted by $\hat{\alpha}$ and $\hat{\beta}$ s.t. the sum of squared distances bet^n data points and the line $Y = \alpha + \beta X$ is minimized.

$$y_1 = \alpha + \beta x_1 + e_1$$
$$y_2 = \alpha + \beta x_2 + e_2 \quad \Rightarrow \boxed{e_i = y_i - \alpha - \beta x_i}$$
$$\vdots$$
$$y_i = \alpha + \beta x_i + e_i \qquad \text{we have to minimize, } \boxed{\hat{\sum_{i=1}} e_i^2}$$

$$\therefore \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

To minimize, using principle of Maxima and Minima.

Partially Diff. w.r.t. α and β, and equating with '0'.

$$\frac{\partial \left(\sum_{i=1}^{n} e_i^2\right)}{\partial \alpha} = 0 \quad ; \quad \frac{\partial \left(\sum_{i=1}^{n} e_i^2\right)}{\partial \beta} = 0.$$

$$\frac{\partial \left(\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\right)}{\partial \alpha} = 0 \quad ; \quad \frac{\partial \left(\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\right)}{\partial \beta} = 0.$$

$$-2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \quad ; \quad -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0.$$

Since, we are estimating $\hat{\alpha}$ and $\hat{\beta}$, replace α, β with $\hat{\alpha}, \hat{\beta}$.

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\alpha} - \sum_{i=1}^{n} \hat{\beta} x_i = 0.$$

$$n\bar{y} - n\hat{\alpha} - \hat{\beta} n\bar{x} = 0.$$

$$\therefore n\hat{\alpha} = n\bar{y} - \hat{\beta} n\bar{x}$$

$$\boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$

$$\sum_{i=1}^{n} (x_i y_i - \hat{\alpha} x_i - \hat{\beta} x_i^2) = 0.$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\alpha} n\bar{x} - \sum_{i=1}^{n} \hat{\beta} x_i^2 = 0.$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 - n\bar{x}(\bar{y} - \hat{\beta}\bar{x}) = 0.$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \hat{\beta}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 0.$$

$$\therefore \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

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

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

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

$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i^2 - \bar{x}^2)}}$$

□ R-Command :-

Tut-11. Q.5.

```
x ← c(77, 50, ..., 67)
y ← c(82, 66, ...)
lm(y ~ x)  ⟶ This will directly give, α̂ and β̂.
```

Properties of Linear Regression:-

1. We should interpret $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ only in the interval $[x_1, x_n]$.

2. The point $(\bar{x}, \bar{y})$ always lies on the regression line.
   $(\bar{x}, \bar{y})$ satisfies $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, because of $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$.

3. The sum of Residuals is zero.

Residuals: It is the difference bet^n y_i and $\hat{y}_i$.
Denoted by $\hat{e}_i$. $\Rightarrow \hat{e}_i = y_i - \hat{y}_i$

Proof:

To Prove: $\sum_{i=1}^{n} \hat{e}_i = 0$.

$$L.H.S. = \sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))$$

$$= \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta} x_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} + \sum_{i=1}^{n} \hat{\beta}\bar{x} - \sum_{i=1}^{n} \hat{\beta} x_i$$

$$= n\bar{y} - n\bar{y} + \hat{\beta} n\bar{x} - \hat{\beta} n\bar{x} = 0.$$

L.H.S. = R.H.S.

4. The Arithmetic mean of $\hat{y}$ is equal to the arithmetic mean of y.

To prove: $\bar{\hat{y}} = \bar{y}$

Proof:

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \frac{1}{n} \sum_{i=1}^{n} (\hat{\alpha} + \hat{\beta} x_i) = \frac{1}{n} \sum_{i=1}^{n} (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i)$$

$$= \frac{n\bar{y} - n\hat{\beta}\bar{x} + n\hat{\beta}\bar{x}}{n} = \bar{y}$$

L.H.S. = R.H.S.

$\therefore \bar{\hat{y}} = \bar{y}$

Multilinear Regression :-

When we have 2 Co-variates, then the regression line becomes:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 .$$

$\hat{y} \equiv \hat{y}_i \quad \forall i = 1, 2, \ldots, n.$

$\hat{x}_1 \equiv \hat{x}_{1i} \quad \forall i = 1, 2, \ldots, n. \qquad \hat{x}_2 \equiv \hat{x}_{2i} \quad \forall i = 1, 2, \ldots, n.$

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

$$e_i = y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i}$$

We need to minimize $\sum_{i=1}^{n} e_i^2$.

By Using Principle of Maxima and Minima,

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \alpha} = 0 \; ; \quad \frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_1} = 0 \; ; \quad \frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_2} = 0 .$$

$$-2 \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0 .$$

$$-2 \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i}) x_{1i} = 0 .$$

$$-2 \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i}) x_{2i} = 0 .$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n\alpha - \beta_1 \sum_{i=1}^{n} x_{1i} - \beta_2 \sum_{i=1}^{n} x_{2i} = 0 . \quad \text{—(I)}$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_{1i} = \alpha \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i} x_{2i} \quad \text{—(II)}$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_{2i} = \alpha \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 . \quad \text{—(III)}.$$

Solving 3 simultaneous eq^{ns}, I, II and III, we will get values of $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$.

Then Regression line will be

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

When we have p covariates i.e. $x_1, x_2, \ldots, x_p$.
then regression line will be

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p.$$

Further,

$$y_1 = \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \ldots + \hat{\beta}_p x_{1p} + e_1$$

$$y_2 = \hat{\beta}_0 + \hat{\beta}_1 x_{22} + \ldots + \hat{\beta}_p x_{2p} + e_2$$

$$y_n = \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \ldots + \hat{\beta}_p x_{np} + e_n.$$

It can be written as:

$$Y = x\beta + e$$

where,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}_{n\times(p+1)}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

then,

$$\boxed{\hat{\beta} = (X'X)^{-1} X'Y.}$$

where, X' is Transpose of X.

R - Commands:-

1. T - test:

$x_1, x_2, \ldots, x_n$ are sample elements.

$X \longleftarrow c(x_1, x_2, \ldots, x_n)$.

(i) If μ_0 is given.

t.test (X, mu = μ_0)

(ii) Two sided test.

t.test (X, alternative = 'two-sided')

(iii) One sided.