STATISTICS FOR DECISION MAKING

BY

TANUJIT CHAKRABORTY

Indian Statistical Institute

Mail : tanujitisi@gmail.com

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

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

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}.$$

Ex. $U_1 \sim \chi^2_{v_1}$, $U = U_1+U_2 \sim \chi^2_v$,

$$M_X(U) = M(U_1)M(U_2)$$

$$(1-2t)^{-v/2} = (1-2t)^{-v_1/2} \times M(U_2)$$

$$\Rightarrow M(U_2) = (1-2t)^{-(v-v_1)/2}$$

So, $U_2 \sim \chi^2_{v-v_1}$.

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

Let $Y = Z^2$,

$$G(y) = P(Y \le y)$$
$$= P(Z^2 \le y)$$
$$= P(-\sqrt{y} \le Z \le \sqrt{y})$$
$$= 2P(0 \le Z \le \sqrt{y})$$
$$= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \qquad u = z^2.$$
$$= \int_0^{y} \frac{1}{\sqrt{2\pi}} u^{1/2-1} e^{-u/2} du$$

$$\therefore g(y) = G'(y) = \frac{1}{\sqrt{2\pi}} u^{1/2-1} e^{-u/2}$$
$$= \frac{1}{2^{1/2}\Gamma(1/2)} u^{1/2-1} e^{-u/2}$$

So, $Y \sim \chi^2_1$.

So, $Y_1 \sim \chi_1^2, Y_2 \sim \chi_1^2, \dots, Y_n \sim \chi_1^2$.

Let $U = Y_1 + \dots + Y_n$.

$$MGF(U) = E(e^{tU}) = E\left(e^{tY_1 + tY_2 + \dots + tY_n}\right)$$
$$= E(e^{tY_1})E(e^{tY_2})\dots E(e^{tY_n})$$
$$= (1-2t)^{-1/2} \dots (1-2t)^{-1/2}$$
$$= (1-2t)^{-n/2}$$

$\therefore$ So, $U \sim \chi^2_n$.

So, $MGF = \int_0^\infty e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-x/2} x^{-1/2} dx$

$$= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x/2(1-2t)} \cdot x^{-1/2} dx$$

$$= \int_0^\infty \frac{1}{\sqrt{2\pi}(1-2t)} e^{-\frac{u}{2}} \cdot \frac{u^{-1/2}}{(1-2t)^{-1/2}} du \quad \Bigg| \begin{array}{l} x(1-2t) = u \\ dx = \frac{du}{1-2t} \end{array}$$

$$= (1-2t)^{-1/2}.$$

$$\frac{d}{dt}\left[MGF(U)\right] = \frac{d}{dt}\left[(1-2t)^{-n/2}\right]$$

$$= -n/2\,(1-2t)^{-n/2-1}(-2)$$

$$= n(1-2t)^{-n/2-1}$$

$$\left[\frac{d}{dt} M(U)\right]_{t=0} = n = E(\chi_n^2)$$

$$\frac{d^2}{dt^2}\left[MGF(U)\right] = n(n+2) \text{ at } t=0$$

$$\therefore \operatorname{Var}(\chi_n^2) = 2n$$

Q. Show $\Gamma\frac{1}{2} = \sqrt{\pi}$

Sol. $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$

$$\Rightarrow \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1/2$$

$$\Rightarrow \int_0^\infty e^{-x^2/2} dx = \sqrt{\frac{\pi}{2}}.$$

$x^2/2 = u$

$\frac{du}{\sqrt{2u}} = dx$

$$\Rightarrow \int_0^\infty e^{-u/2} \cdot \frac{du}{\sqrt{2u}} = \sqrt{\frac{\pi}{2}}$$

$$\Rightarrow \Gamma\frac{1}{2} = \sqrt{\pi}.$$

Theory & Method of Point Estimation.

Point Estimation:- In statistics, point estimation involves the use of sample data to calculate a single value which is to serve as a "best estimate" of an unknown population parameter.

Statistic:- Statistic is a function of sample values which is itself an observable random variable doesnot contain any parameter. $\bar{X}, s^2, X_{(1)}, X_{(n)}$ are examples of statistic.

Sampling Distribution:- The probability distribution of any statistic is termed as sampling distribution.

Standard Error:- $SE(T) = \sqrt{Var(T)} = \left[E(T-E(T))^2\right]^{1/2}$

The smaller the SE, the better the guess.

- According to Fisher, an estimator is said to be the BEST ESTIMATOR if it is
  - Unbiased
  - consistent
  - Efficient
  - Sufficient

Mean Square Error:- $MSE = E(T-\theta)^2$

$$= E\left[T-E(T)+E(T)-\theta\right]^2$$

$$= E(T-E(T))^2 + \{E(T)-\theta\}^2$$

$$= Var(T) + Bias(\theta, T).$$

1. **Unbiasedness:-** An estimator $\hat{\theta}$ is said to be unbiased for θ if $E(\hat{\theta}) = \theta$.

Ex. 1. $\bar{X}$ is an unbiased estimator for θ where $X_i \overset{iid}{\sim} D(\mu, \sigma^2)$, $i=1(1)n$.

$\rightarrow$ $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\cdot n\mu = \mu.$

2. Show that s^2 is an unbiased estimator of σ^2.

$\rightarrow$ $E(\bar{X}) = \mu$, $V(\bar{X}) = \frac{\sigma^2}{n}$.

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

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

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

$$= \frac{1}{n}\left[\sum_{i=1}^{n}V(x_i) - nV(\bar{x})\right]$$

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

$$= \frac{n-1}{n}\sigma^2$$

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

$\Rightarrow E(s^2) = \sigma^2$.

3. For Binomial distn. fraction defective is an unbiased estimator of p.

Solution:-

$$\hat{p} = \frac{\#\text{ defectives}}{\#\text{ items checked}} = \frac{d}{n}.$$

$$E(\hat{p}) = E\left(\frac{d}{n}\right) = \frac{1}{n}\cdot E(d) = \frac{1}{n}\cdot np = p.$$

2. Consistency:- An estimator $\hat{\theta}_n$ is consistent for θ iff

$$\lim_{n\to\infty} P\left[\,|\hat{\theta}_n - \theta| < \epsilon\right] = 1.$$

Alt. $E(\hat{\theta}_n) = \theta$ and $Var(\hat{\theta}_n) \to 0$ as $n\to\infty$.

Ex. 1. $V(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$ as $n\to\infty$.

So, $\bar{X}$ is consistent for θ.

3. <u>Efficiency</u>:- Consider two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for θ. Then $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if $\frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)} < 1$.

<u>Ex.</u> Let $X_1, \dots, X_n$ be a r.s. from a popln. with mean μ, variance σ^2. Consider three following estimators of μ.

(i) $\hat{\theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$

(ii) $\hat{\theta}_2 = \frac{1}{8}X_1 + \frac{3}{4(n-2)}(X_2 + X_3 + \dots + X_{n-1}) + \frac{1}{8}X_n$

(iii) $\hat{\theta}_3 = \overline{X}$

(a) S.T. each of the three is biased.

(b) Find $e(\hat{\theta}_2, \hat{\theta}_1)$, $e(\hat{\theta}_3, \hat{\theta}_1)$, $e(\hat{\theta}_3, \hat{\theta}_2)$. Which of the three estimators is more efficient?

<u>Solution</u>:- (a)

$$E(\hat{\theta}_1) = \frac{1}{3} \cdot 3\mu = \mu$$

$$E(\hat{\theta}_2) = \frac{1}{8}\mu + \frac{3}{4(n-2)} \cdot (n-2)\mu + \frac{1}{8}\mu = \mu$$

$$E(\hat{\theta}_3) = E(\overline{X}) = \mu.$$

(b)

$$V(\hat{\theta}_1) = \frac{3\sigma^2}{9} = \frac{\sigma^2}{3}$$

$$V(\hat{\theta}_2) = \frac{\sigma^2}{64} + \frac{9}{16(n-2)^2} \cdot (n-2)\sigma^2 + \frac{1}{64}\sigma^2$$

$$= \frac{\sigma^2(n+16)}{32(n-2)}.$$

$$V(\hat{\theta}_3) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

$$e(\hat{\theta}_2, \hat{\theta}_1) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)} = \frac{3(n+16)}{32(n-2)} < 1 \text{ when } n > 3$$

$$e(\hat{\theta}_3, \hat{\theta}_1) = \frac{V(\hat{\theta}_3)}{V(\hat{\theta}_1)} = \frac{3}{n} < 1 \text{ when } n > 3$$

$$e(\hat{\theta}_3, \hat{\theta}_2) = \frac{V(\hat{\theta}_3)}{V(\hat{\theta}_2)} = \frac{32(n-2)}{n(n+16)} < 1 \text{ when } n \geq 15$$

So, for $3 < n \leq 15$, $\hat{\theta}_2$ is more efficient estimator.

$n \geq 15$, $\hat{\theta}_3$ is more efficient estimator.

4. **Sufficiency**:- A statistic $\hat{\theta}$ is called sufficient estimator for θ if the conditional distn. of $(X_1, \ldots, X_n)$ given $\hat{\theta} = \theta_0$ is independent of θ.

Example:- 1. $(X_1, \ldots\ldots, X_n)$ be a r.s. from Bin $(1, p)$.
show that $\sum_{i=1}^{n} X_i$ is sufficient for p.

$X_i \sim Bin(1, p)$

$\sum X_i \sim Bin(n, p)$

$\rightarrow$ $P[X_1 = x_1, \ldots, X_n = x_n \mid S = s]$

$$= \frac{P[X_1 = x_1, \ldots\ldots, X_n = x_n]}{P[S = s]} \text{ if } S = \sum_{i=1}^{n} X_i$$

$$= \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{\binom{n}{s} p^s (1-p)^{n-s}} \text{ if } s = \sum x_i \text{, where } x_i = 0 \text{ or } 1 \ \forall i = 1(1)n$$

$$= \frac{1}{\binom{n}{s}} \text{ if } s = \sum x_i \text{, independent of } p.$$

So, $S = \sum_{i=1}^{n} X_i$ is sufficient for p.

Ex. 2. Let (X_1, X_2) be r.s. from Poisson(λ). Show that $X_1 + 2X_2$ is not sufficient for λ.

$\rightarrow$ $$P[X_1 = 0, X_2 = 1 \mid T = 2] = \frac{P[X_1 = 0, X_2 = 1]}{P[X_1 + 2X_2 = 2]}$$

$$= \frac{e^{-\lambda}(\lambda e^{-\lambda})}{P[X_1 = 0, X_2 = 1] + P[X_1 = 2, X_2 = 0]}$$

$$= \frac{\lambda e^{-2\lambda}}{\lambda e^{-2\lambda} + \left(\frac{\lambda^2}{2}\right) e^{-2\lambda}}$$

$$= \frac{1}{\left(1 + \frac{\lambda}{2}\right)} \text{, depend on } \lambda.$$

So, $X_1 + 2X_2$ is not sufficient for λ.

Factorization Theorem:- Let $\underset{\sim}{X} = (X_1, \ldots, X_n)$ be a r.s. from PMF/PDF $f(x;\theta)\ \forall\ \theta \in \Omega$. Then $T(\underset{\sim}{X})$ is sufficient for θ iff

$$\prod_{i=1}^{n} f(x_i;\theta) = g\left(T(\underset{\sim}{x}), \theta\right) \cdot h(\underset{\sim}{x})$$

where, $h(\underset{\sim}{x})$ depends only on $\underset{\sim}{x}$ and $g(T(\underset{\sim}{x}), \theta)$ depends on θ and on $\underset{\sim}{x}$ only through $T(\underset{\sim}{x})$.

For example 1. $\prod_{i=1}^{n} f(x_i, p) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$

$= g\{T(\underset{\sim}{x}), \theta\} \cdot h(\underset{\sim}{x})$, where $h(\underset{\sim}{x}) = 1$ and $T(\underset{\sim}{x}) = \sum_{i=1}^{n} x_i$

So, $T = \sum_{i=1}^{n} x_i$ is sufficient for θ by Factorization theorem.

Ex. If $(X_1, \ldots, X_n)$ be a r.s. from $N(\mu, \sigma^2)$. Then find a two-dimensional sufficient statistic for (μ, σ^2).

Sol.
$$\prod_{i=1}^{n} f(x_i, \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\left(-\frac{\sum x_i^2}{2\sigma^2} + \frac{\mu\sum x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right)}$$

$$= g\left(T(\underset{\sim}{x}), \mu, \sigma\right) \cdot h(\underset{\sim}{x}) \quad ; \quad h(\underset{\sim}{x}) = 1$$

$T(\underset{\sim}{x}) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$ is sufficient for (μ, σ).

Alt.
$$\prod_{i=1}^{n} f(x_i;\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

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

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

$$= g\left(\bar{x}, s^2; \mu, \sigma\right) \cdot h(\underset{\sim}{x}), \quad h(\underset{\sim}{x}) = 1$$

So, $(\bar{x}, s^2)$ is sufficient for (μ, σ).

Ex.3. $f(x;\theta) = \theta x^{\theta-1}, \ 0<x<1$

$$\prod_{i=1}^{n} f(x_i;\theta) = \theta^n \left(\prod x_i\right)^{\theta-1}$$

$$= g_\theta\left\{\prod_{i=1}^{n} x_i\right\} \cdot h(\underset{\sim}{x}) \quad , \quad h(\underset{\sim}{x}) = 1.$$

So, $T = \prod_{i=1}^{n} x_i$ is sufficient for θ.

Method of finding Estimators:-

1. Method of Moments:-

$$E(X^r) = \mu_r' \qquad \mu_2' = E(X^2)$$

$$\mu_1' = E(X). \qquad \mu_2 = \mu_2' - \mu_1'^2.$$

$$\hat{\mu}_1 = \overline{X}. \qquad \mu_2 = E(X^2) - E^2(X)$$

Method of moment doesnot yield the unbiased estimator of population.

Ex.1. $X_i \sim Geo(p)$. Find MME for p.

Sol. $f(x) = p(1-p)^{x-1}$

$$\mu_1' = E(X) = \frac{1}{p} \Rightarrow p = \frac{1}{\mu_1'}$$

so, $\hat{p} = \frac{1}{\overline{X}}$ is MME for p, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Ex. 2. $x_i \sim U(\theta-1, \theta+1)$, Find MME for θ.

Sol. $f(x) = \frac{1}{2}$.

$$E(X) = \mu_1' = \int_{\theta-1}^{\theta+1} \frac{1}{2} x\,dx = \theta.$$

so, $\hat{\theta} = \overline{X} = \frac{\sum x_i}{n}$

Scanned by CamScanner

2. Method of Likelihood Estimation:-

$L(\underset{\sim}{x};\theta)$ is called likelihood function of θ; $L(\underset{\sim}{x},\theta)=\prod_{i=1}^{n} f(x_i,\theta)$.

Ex.1. $X_i \sim Geo(p)$, find MLE for p.

Sol. $L=\prod_{i=1}^{n} p(1-p)^{x_i-1} = \sum_{i=1}^{n}\left(\ln p + x_i \ln(1-p) - \ln(1-p)\right)$

$\ln L = n\ln p + n\bar{x}\ln(1-p) - n\ln(1-p)$

$\frac{\partial}{\partial p}\ln L = \frac{n}{p} - \frac{n\bar{x}}{1-p} + \frac{n}{1-p} = 0 \Rightarrow \hat{p} = \frac{1}{\bar{x}}$

$\frac{\partial^2}{\partial p^2}\ln L = -\frac{n}{p^2} + \frac{n}{(1-p)^2} - \frac{n\bar{x}}{(1-p)^2}$

< 0, since $\bar{x} \geq 1$ for $0 \leq p \leq 1$.

So, MLE of p is $\frac{1}{\bar{x}}$.

Ex.2. $f(x) = \frac{1}{2\theta^3}x^2 e^{-x/\theta}$, $x \in \mathbb{R}^+$, $\theta \in \mathbb{R}^+$

find MLE of θ.

Sol. $L(\underset{\sim}{x}|\theta) = \frac{1}{2\theta^{3n}}\prod_{i=1}^{n}\left(x_i^2 e^{-x/\theta}\right) = \frac{1}{2\theta^{3n}} e^{-\Sigma x_i/\theta}\left(\prod_{i=1}^{n} x_i^2\right)$

$\ln L = -3n\ln(2\theta) + 2\sum_{i=1}^{n}\ln x_i - \frac{\Sigma x_i}{\theta}$

$\frac{\partial}{\partial\theta}\ln L = -\frac{3n\times 2}{2\theta} + \frac{\Sigma x_i}{\theta^2} = 0 \Rightarrow \hat{\theta} = \frac{\bar{x}}{3}$.

$\frac{\partial^2}{\partial\theta^2}\ln L = \frac{3n}{\theta^2} - \frac{2\Sigma x_i}{\theta^3} = -\frac{3n}{\theta^2} < 0$

MLE for θ is $\frac{\bar{x}}{3}$.

Ex. 3. $f(x) = \theta x^{\theta-1}$, $\theta \in \mathbb{R}^+$, $0 < x < 1$. Find MLE of θ.

Sol. $L = \theta^n\left(\prod_{i=1}^{n} x_i^{\theta-1}\right)$

$\ln L = n\ln\theta + \left(\Sigma \ln x_i\right)(\theta-1)$

$\frac{\partial}{\partial\theta}\ln L = \frac{n}{\theta} + \Sigma\ln x_i = 0 \Rightarrow \hat{\theta} = -\frac{n}{\Sigma\ln x_i}$

$\frac{\partial^2}{\partial\theta^2}\ln L = -\frac{n}{\theta^2} < 0$. So, MLE of θ is $-\frac{n}{\Sigma\ln x_i}$.

Confidence Interval for μ

$x_i \sim N(\mu, \sigma^2)$ (Here σ^2 is known)

$\bar{x} \sim N(\mu, \sigma^2/n)$.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$P\left[-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[-Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu \le Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$\Rightarrow P\left[\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{x} \le \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Two sided $100(1-\alpha)\%$ C.I. for μ, when σ is known is

$$\left(\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

Sampling error $= \bar{x} - \mu$, when μ increases, C.I. reduces.

Ex. 1. A random sample of size 50 from a particular brand of tea packets produced a mean weight of 15.65 ounces. Weights of brands of tea packets are normally distributed with Normal (s.d. = 0.59 ounce). Find 95% C.I. for μ.

Sol.

$$\mu \in \left(\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$Z_{\alpha/2} = 1.96$ at $\alpha = 0.975$.

$$\in \left(15.65 - \frac{1.96 \times 0.59}{\sqrt{50}},\ 15.65 + \frac{1.96 \times 0.59}{\sqrt{50}}\right)$$

$$= (15.3295, 15.9705)$$

$\bar{x} = 15.65$

$\sigma = 0.59$

$n = 50$

Ex. 2. Mobile phones coming off an assembly line are automatically checked to make sure they are not defective. The manufacturer wants an interval estimate of the percentage of Mobile phones that fail the testing procedure. Compute 90% C.I. based on a r.s. of size 105 in which 17 mobiles failed the testing procedure.

Sol. $n = 105$, $d = 17$, $\hat{p} = \frac{17}{105}$

$\alpha = 0.1$ $X \sim Bin(n, p)$, $X \sim N(np, \sqrt{np(1-p)})$ for large n.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

$$P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = 0.9$$

$$p \in \left(\hat{p} - Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$= \left(0.16 - 1.64 \times \sqrt{\frac{0.16 \times 0.84}{105}}, 0.16 + 1.64\sqrt{\frac{0.16 \times 0.84}{105}}\right)$$

Case II :- (σ is unknown)

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

$$P\left[-t_{\alpha/2, n-1} \le t \le t_{\alpha/2, n-1}\right] = 1 - \alpha$$

$$\Rightarrow P\left[\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}\right] = 1 - \alpha.$$

C.I. for σ :- (μ is known) $\frac{ns^2}{\sigma^2} \sim \chi^2_n$.

$$P\left[\chi^2_{1-\alpha/2, n} \le \chi^2 \le \chi^2_{\alpha/2, n}\right] = 1 - \alpha$$

$$\Rightarrow P\left[\frac{ns^2}{\chi^2_{\alpha/2, n}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{1-\alpha/2, n}}\right] = 1 - \alpha.$$

(μ is unknown) :- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

$$P\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}\right] = 1 - \alpha.$$

C.I. for mean difference:-

$$X_{11}, X_{12}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$$
$$X_{21}, X_{22}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

$$P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = 1-\alpha.$$

$$\Rightarrow P\left((\bar{X}_1 - X_2) - Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le (\mu_1 - \mu_2) \le (\bar{X}_1 - \bar{X}_2) + Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$

then, $v(\bar{X}_1 - \bar{X}_2) = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$t = \frac{(\bar{x}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

C.I. for $\frac{\sigma_1^2}{\sigma_2^2}$:-

$$F = \left(\frac{s_1^2}{s_2^2}\right) \Big/ \left(\frac{\sigma_1^2}{\sigma_2^2}\right) \sim F_{n_1-1, n_2-1}.$$

$$P\left[F_{1-\alpha/2} \le F \le F_{\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[\frac{s_1^2/s_2^2}{F_{\alpha/2, n_1-1, n_2-1}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2/s_2^2}{F_{1-\alpha/2, n_1-1, n_2-1}}\right] = 1-\alpha.$$

TESTING

- Hypothesis: statement $\to$ Null hypothesis / Alternative hypothesis

Type I error: Reject Ho when it is true

Type II error:- Accept Ho when it is false

$$\beta = P(\text{Accept } H_0 \mid H_0 \text{ is false})$$

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

$$\text{power} = 1 - P(\text{Type II error}) = 1 - \beta$$

Test of Mean/Variance

- Mean (Z, t)
  - σ known $\to$ Z-statistic
  - σ unknown $\to$ t-statistic
- Variance (χ^2, F)
  - one-sample $\to$ χ^2
  - two samples $\to$ F

Ex.1 The following sample data (of size $n = 15$) show the stabilized viscosity of Rubberized Asphalt.

3193	3124	3153	3145	3093	3466	3355	2979
3182	3227	3256	3332	3204	3282	3170	

To be suitable for the intended pavement application, the mean stabilized viscosity should be equal to 3200. Test this hypothesis using $\alpha = 0.05$ with assumption that viscosity is normally distd.

Sol. $H_0: \mu = 3200$ Vs $H_1: \mu \neq 3200$

$X = SV \sim N(\mu, \sigma^2)$, σ^2 unknown.

$\bar{X} = 3210.733$, $s = 117.507$, $n = 15$

Test statistic, $T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} = 0.353$

$$T \in \left(-t_{0.025,14}, t_{0.025,14}\right) = (-2.145, 2.145)$$

$\therefore$ We fail to reject Ho.

Ex. 2. The mean contents of coffee cans filled on a particular production line are being studied. Standards specify that the mean contents must be 16.0 oz, and from past experience it is known that the sd of the can contents is 0.1 oz. Find the prob. of type II error and power of the test if the true mean content $\mu_1 = 16.1$ oz, $n = 20$, $\alpha = 0.05$.

Sol. $H_0: \mu_0 = 16$ Vs $H_1: \mu_0 \neq 16$

$\sigma = 0.10$ $\qquad \mu_1 = 16.10, \mu_0 = 16$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \qquad \delta = \mu_1 - \mu_0 = 0.1$$

We fail to reject H_0 when $Z \in (-Z_{0.025}, Z_{0.025}) = (-1.96, 1.96)$

$$\beta = P(\text{accept } H_0 \mid H_0 \text{ is false})$$

$$= \Phi\left(Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

$$= \Phi\left(1.96 - \frac{0.1 \times \sqrt{20}}{0.1}\right) - \Phi\left(-1.96 - \frac{0.1 \times \sqrt{20}}{0.1}\right)$$

$$= \Phi(-2.5121) - \Phi(-6.4321) = 0.006$$

Power of the test $= 1 - \beta = 0.994$.

Ex. 3. The life in hours of a battery is known to be approximately normally distributed with SD $\sigma = 1.25$ hrs. A random sample of 10 batteries has a mean life of $\bar{X} = 40$ hrs.

(a) Is there evidence to support the claim that battery life exceeds 40 hours?

(b) What is the p-value for the test in (a)?

(c) What is the β-error for the test in (a) if true mean is 42 hrs?

solution:- (a) X: Life of the battery

$X \sim N(\mu, \sigma^2)$ with $\sigma = 1.25$

$H_0: \mu = 40$ Vs $H_1: \mu > 40$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{40 - 40}{1.25/\sqrt{10}} = 0 \in (-\infty, 1.64)$$

We have no evidence to reject H_0.

(b) p-value $= P(Z > 0) = 0.05$

(c) β-error $= P(\text{accept } H_0 \mid H_0 \text{ is false})$

$$= \Phi\left(Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) = \Phi\left(1.64 - \frac{2\sqrt{20}}{1.25}\right)$$

$= 0$.

Ex.4. A soft-drink machine at a steak house is regulated so that the amount of drink dispensed is approximately normally distributed with a mean of 200 mililiters and s.d. of 15mls. The machine is checked periodically by taking a sample of 9 drinks and computing the average content. If x-bar falls in the interval $191 < \bar{X} < 209$, the machine is thought to be operating satisfactorily; otherwise we conclude that $\mu = 200$ ml.

(a) Find the prob. of committing a type I error when $\mu = 200$ ml.

(b) Find the prob. of committing a type II error when $\mu = 215$ ml.

Solution:-

(a)
$$
\begin{aligned}
P(\text{Type I error}) &= P(\text{Reject } H_0 \mid H_0 \text{ is true}) \\
&= 1 - P(\text{accept } H_0 \mid H_0 \text{ is true}) \\
&= 1 - P(191 < \bar{X} < 209 \mid \mu = 200) \\
&= 1 - P\left(\frac{191-200}{15/\sqrt{9}} < Z < \frac{209-200}{15/\sqrt{9}}\right) \\
&= 1 - P(-1.8 < Z < 1.8) \\
&= 0.0718
\end{aligned}
$$

(b)
$$
\begin{aligned}
P(\text{Type II error}) &= P(\text{accept } H_0 \mid H_0 \text{ is false}) \\
&= P(191 < \bar{X} < 209 \mid \mu = 215) \\
&= P\left(\frac{191-215}{15/\sqrt{9}} < Z < \frac{209-215}{15/\sqrt{9}}\right) \\
&= P(-4.8 < Z < -1.2) \\
&= 0.1151
\end{aligned}
$$

Ex.5. The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company:

46.4 46.1 45.8 47.0 46.1 45.9 45.8 46.9 45.2 46.0

Find a 95% C.I. for variance assuming normal distribution.

Solution:- μ is unknown.

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,\,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,\,n-1}}\right]$$

$$= \left[\frac{9 \times 0.2862}{19.02} < \sigma^2 < \frac{9 \times 0.2862}{2.70}\right]$$

$$= \left[0.135 < \sigma^2 < 0.954\right]$$

Ex.6. The Edison Electric Institute has published figures on the number of kilowatt hrs used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hrs per year. If a r.s. of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hrs per year with SD = 11.9 kilowatt hrs, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hrs annually? Assume Normal distn.

Solution:- $H_0: \mu = 46$ $\quad H_1: \mu < 46$

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

$$= \frac{42 - 46}{11.9/\sqrt{12}} = -1.164, \text{ under } H_0.$$

$t_{0.95,11} = -1.796$

So, we fail to reject H_0.

So, vacuum cleaners on average doesn't use less than 46 kwhr annually.

Ex.7. Engineers at a large automobile manufacturing company are trying to decide whether to purchase brand A or brand B tires for the company's new models. To help the arrive at a decision, an experiment is conducted using 12 of each brand. The tires are run until they wear out.

$$\bar{X}_1 = 37900 \text{ Km}, \quad s_1 = 5100 \text{ Kms}; \quad n_1 = 12$$
$$\bar{X}_2 = 39800 \text{ Km}, \quad s_2 = 5900 \text{ Kms}; \quad n_2 = 12$$

Test the hypothesis that there is no difference in the average wear of the two brands of tires. Assume the popl^n. to be approximately normally distributed with equal variances.

Solution:-
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$
$$= \frac{s_1^2 + s_2^2}{2} = 30410000$$
$$s_p = 5514.53$$

$$\therefore t = \frac{37900 - 39800}{5514.53\sqrt{\frac{1}{12}+\frac{1}{12}}} = -0.844, \text{ under } H_0.$$

$$\in \left(-t_{0.025, 22}, t_{0.025, 22}\right) = (-2.074, 2.074)$$

So, we fail to reject H_0.

$\therefore$ There is no difference in the average wear out of the two brands of tires.

Scanned by CamScanner

Two Sample Tests

- **Paired t-test:-**

Sample I	Sample II	d (difference)
$\vdots \Big\} n_1$	$\vdots \Big\} n_2$	$x_i - \bar{x}_d$
		$\vdots$

In Paired t-test, no. of sample sizes should be same for both samples, but for two sample t-test it need not be.

Sample units are same in both the situation.
No. of sample sizes should be same for both samples.

Test hypothesis is $H_0: \mu_d = 0$ $\quad$ $d = X_2 - X_1$

$H_1: \mu_d > 0$ $\quad$ $V(d) = Var(X_2) + Var(X_1) - 2Cov(X_2, X_1)$

So, test ~~statistic~~ is $t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim t_{n-1}$.

Two Sample Tests

Test for mean → σ - known ; σ - unknown → $\sigma_1 = \sigma_2$; $\sigma_1 \neq \sigma_2$

1. Check for Normality for both sample data.
2. Check for equality of variances.
3. Test the equality of means.

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

$H_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,\, n_2-1}$$

Q. The concentration of active ingredients in a liquid loundry detergent is thought to be effected by the type of catalyst used in the process. 10 observations on concentration are taken with each catalyst and the data following:

Catalyst-1 : 57.9, 66.2 65.4 65.4 65.2 62.6 67.6 63.7 67.2 71.0

Catalyst-2: 66.4 71.7 70.3 69.3 64.8 69.6 68.6 69.4 65.3

Is there any evidence to indicate that the mean active concentration depends on the choice of catalyst.

[check for equality of variances]

Sol. $s_1^2 = (3.444)^2 = 11.46$

$s_2^2 = (2.355)^2 = 5.546$

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad F = \frac{s_1^2}{s_2^2} = \frac{11.86}{5.546} = 2.138$

$H_1: \sigma_1^2 \neq \sigma_2^2$

Df. (9,8).

$F_{0.05,9,8} = 3.39$

$F_{0.95,9,8} = \frac{1}{F_{0.05,9,8}} = 0.3095$

So, there is no evidence to reject H_0.

We can assume variance of populations are same.

(Test for equality of means)

$\overline{X}_1 = 65.22 \quad ; \quad s_1^2 = (3.444)^2$

$\overline{X}_2 = 68.38 \quad ; \quad s_2^2 = (2.355)^2$

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Now, $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

Now, $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$

$= \sqrt{\dfrac{9 \times (3.444)^2 + 8 \times (2.355)^2}{10+9-2}}$

$= 2.98$

C.I. is $\quad (\overline{X}_1 - \overline{X}_2) \pm s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{0.025, n_1+n_2-2}$

$= -3.16 \pm 2.98 \sqrt{\frac{1}{10} + \frac{1}{9}} \cdot 2.110$

$= (-6.049, -0.2709)$

$t_c = \dfrac{-3.16 - 0}{2.98 \sqrt{\frac{1}{10} + \frac{1}{9}}} = -2.307.$

C.R. is $(t > 2.110, t < -2.110)$; so reject H_0.

$\therefore$ p-value $= P(t > 2.307, t < -2.307)$

$= 2P(t > 2.307)$

$= 2 \times 0.01750$

$= 0.03516.$

<u>Note:-</u> When $\sigma_1^2 \neq \sigma_2^2$; $T = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$$DF = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Q.

AC City	AC Village
3	48
7	44
25	40
10	38
15	33
6	21
12	20
25	12
15	1
7	18

Sol. (Check for variance)

$\bar{x}_1 = 12.5$; $s_1^2 = (7.634)^2$

$\bar{x}_2 = 27.5$; $s_2^2 = (15.35)^2$

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

$H_1: \sigma_1^2 \neq \sigma_2^2$

Now, $F = \dfrac{s_1^2}{s_2^2} = 0.2473$

Now, $F_{0.05, 9, 9} = 3.18$; $F_{0.95, 9, 9} = 0.3144$.

$\therefore$ Reject the null hypothesis. (Check for equality of means)

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Now, $$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$= -2.766$

$df = \dfrac{863.77}{65.46} = 13$. Now, $CR = (t_{13} < -1.771,\ t_{13} > 1.771)$.

Falls in CR. So reject null hypothesis.

i.e. $\mu_1 \neq \mu_2$.

p-value $= P(t_{13} < -2.766) + P(t_{13} > 2.766)$

$= 2 . P(t_{13} > 2.766)$

$= 2 \times 0.00857$

$= 0.01714$.

Another formula for df :- $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

Q. Two types of insection modeling machine are used to form plastic parts, a part is considered defective if it has excessive shrink case or is discoloured. Two random sample each of 300 are selected and 15 defective parts are found in the sample from machine 1 while 8 defective parts are found in the sample from machine 2. Is it reasonable to conclude that both machines produce the same fraction of defective parts using $\alpha = 0.05$.

Ans. $H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$$\hat{p}_1 = \frac{d_1}{n_1} = 0.05$$

$$\hat{p}_2 = \frac{d_2}{n_2} = 0.0267$$

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = E\left(\frac{d_1}{n_1}\right) - E\left(\frac{d_2}{n_2}\right) = p_1 - p_2$$

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

under H_0.

Now, $$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim N(0,1)$$

$$\hat{p} = \frac{d_1 + d_2}{n_1 + n_2} = \frac{15+8}{300+300} = 0.0383$$

$$Z = \frac{(0.05 - 0.0267) - 0}{\sqrt{0.0383(1-0.0383)\left(\frac{2}{300}\right)}} = 1.4869$$

$$CR = (Z > 2.578,\ Z < -2.578)$$

$\therefore$ point not fall in the C.R.

So, we can't reject the null hypothesis, i.e., the machine produce the same fraction defective.

Chi-square test is used for:-

1. Goodness of fit
2. Independence of attributes
3. Equality of several proportions.

Q. The no. of cars passing East Bound has been tabulated by a group of students, following data obtained:

Vehicle/min (x_i)	f_i (freq)
40	14
41	24
42	57
43	111
44	194
45	256
46	296
47	378
48	250
49	185
50	171
51	150
52	110
53	102

x_i	f_i
54	96
55	90
56	81
57	73
58	64
59	61
60	59
61	50
62	42
63	29
64	18
65	15
Total	N = 2976

Whether the assumption of a Poisson distn. seem appropriate as a probability model for this process.

Sol. $P(r) = \dfrac{e^{-\lambda}\lambda^{r}}{r!}$; $\hat{\lambda} = \bar{x} = \dfrac{\Sigma x_i f_i}{\Sigma f_i} = 49.67406$

$P(40) = 0.022928$

$N \times P(40) = 2976 \times 0.022928 = 65.472 =$ Expected freq.

Observed frequency $= 14$

So, data does not follow Poisson distn.

Q. Let X denotes the no. of flaws observed on a large coil of galvanized steel. 75 coils are inspected and the following data are observed:

x_i	$O_i = f_i$	$P(n)$	$E_i = N \times P(n)$	$(O_i - E_i)^2/E_i$
0	0	0.0074	0.555	
1	1	0.036	2.72	
2	11	0.089	6.68	
3	8	0.145	10.92	
4	13	0.178	13.40	
5	11	0.175	13.15	
6	12	0.1433	10.75	
7	10	0.1004	7.54	
8	9	0.0616	4.62	
>9	0	0.062	4.66	
	N=75	1		

$\sum x_i f_i = 368$; $\hat{\lambda} = \frac{368}{75} = 4.90$; $P(1) = \frac{e^{-\lambda} \cdot \lambda^1}{1!} =$

$$\chi_c^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1} = \chi^2_4 .$$

$\chi_c^2 = 2.524$; $\chi^2_{0.05,4} = 9.487$

$\therefore$ Distribution fits the data well.

Q. For Normal time interval is given. Test whether Normal distn. will fit the data well or not.

Class interval	$O_i = f_i$	$x_i = \frac{a+b}{2}$	$x_i f_i$	$f_i(x_i - \bar{x})^2$	Pr	$N \times P(n)$	$(O-E)^2/E$
2–6	2	4	8	291.7	.0223	1.1591	
6–10	4	8	32	260.95	.0941	4.8943	0.00047
10–14	10	12	120	166.21	.2241	11.6553	0.235194
14–18	18	16	288	6.11	.3015	15.6762	0.34447
18–22	12	20	240	184.69	.2291	11.9141	0.0006
22–26	5	24	120	184.69	.0983	5.1142	0.019544
26–30	1	28	28	142.16	.0238	1.2382	
Total	N = 52	112	836	1359.69	0.99		0.6002

$\bar{x} = \sum x_i f_i / \sum f_i = \frac{836}{52} = 16.07692$; $N = 52$

$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N-1}} = 5.16$

Prob $= \Phi\left(\frac{6-\bar{x}}{s}\right) - \Phi\left(\frac{2-\bar{x}}{s}\right) =$ ——

$\chi_c^2 = 0.6002 \sim \chi^2_{0.05,3}$ $\boxed{24}$

Linear Regression Model

$$y_i = \alpha + \beta x_i + \epsilon_i \ , \ i = 1(1)n \ , \ \epsilon_i \overset{IID}{\sim} N(0, \sigma^2).$$

OLS Estimates :- $\hat{\beta} = b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}$

$\hat{\alpha} = a = \bar{y} - b\bar{x}$.

$$E(\hat{\beta}) = E(S_{xy}/S_{xx}) = \frac{1}{S_{xx}} E(S_{xy}) = \frac{1}{S_{xx}} E\left[\sum(x_i - \bar{x})(y_i - \bar{y})\right]$$

$$= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i\right), \text{ since } \sum(x_i - \bar{x})\bar{y} = 0$$

$$= \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} E(x_i - \bar{x}) y_i\right]$$

$$= \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} E(x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i)\right]$$

$$= \frac{1}{S_{xx}} E\left[\beta \sum_{i=1}^{n} (x_i - \bar{x}) x_i + \sum(x_i - \bar{x})\epsilon_i\right]$$

$$= \frac{1}{S_{xx}} E\left[\beta \ \sum(x_i - \bar{x})^2 + \sum(x_i - \bar{x})\epsilon_i\right]$$

$$= \frac{1}{S_{xx}} \cdot \beta S_{xx} = \beta.$$

$\therefore \hat{\beta}$ is an unbiased estimator of β.

$$E(\hat{\alpha}) = E(\bar{y} - b\bar{x}) = E(\bar{y}) - E(b\bar{x})$$

$$= \alpha + \beta\bar{x} - \beta\bar{x}$$

$$= \alpha.$$

$$V(\hat{\beta}) = E\left(\hat{\beta} - E(\hat{\beta})\right)^2$$

$$= E\left(\hat{\beta} - \beta\right)^2$$

$$= E\left(\frac{S_{xy}}{S_{xx}} - \beta\right)^2$$

$$= \frac{E\left(S_{xy} - \beta S_{xx}\right)^2}{S_{xx}^2}$$

$=$

$$V(\hat{\beta}) = \frac{1}{S_{xx}^2} V\left[\sum (x_i - \bar{x}) y_i\right]$$

$$= \frac{1}{S_{xx}^2} \cdot V\left[\sum_{i=1}^{n} (x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i)\right]$$

$$= \frac{1}{S_{xx}^2} V\left[\beta \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (x_i - \bar{x}) \epsilon_i\right]$$

$$= \frac{V\left(\sum (x_i - \bar{x}) \epsilon_i\right)}{S_{xx}^2} = \frac{\sigma^2 \sum (x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

$$V(\hat{\alpha}) = V(\bar{y} - \hat{\beta}\bar{x}) = V\left(\alpha + \beta\bar{x} + \bar{\epsilon} - \hat{\beta}\bar{x}\right)$$

$$= V(\bar{\epsilon}) + \bar{x}^2 V(\hat{\beta})$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2}{S_{xx}} \cdot \sigma^2$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

Derivation of OLS estimates:- $y_i = \alpha + \beta x_i + \epsilon_i$

Minimize $S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - \alpha - \beta x_i]^2$

$$\frac{\partial S}{\partial \alpha} = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i)(-1) = 0$$

$\Rightarrow \sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$ ——————— 1^{st} normal equation

$$\frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} 2(y_i - \beta x_i)(-x_i) = 0$$

$\Rightarrow \sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$ ——————— 2^{nd} normal equation

Solving ① and ②, we have

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - b\bar{x}.$$

Q. Show that $E(SSE) = (n-2)\sigma^2$

Ans.

$$E(SSE) = E\left[\sum_{i=1}^{n} (y_i - a - bx_i)^2\right]$$

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

$$= E\left[\left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2\right]$$

$$= \left(\sum_{i=1}^{n} E(y_i^2)\right) - n E(\bar{y}^2) - \sum_{i=1}^{n} (x_i - \bar{x})^2 E(b^2)$$

$$= \left(\sum_{i=1}^{n} Var(y_i) + E(y_i)^2\right) - n\left[Var(\bar{y}) + E^2(\bar{y})\right] - \sum_{i=1}^{n} (x_i - \bar{x})^2 \left[Var(b) + E^2(b)\right]$$

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

$$= n\sigma^2 - \sigma^2 - \sigma^2$$

$$= (n-2)\sigma^2$$

Scanned by CamScanner

(Linear Regression)

Ex. 1.

x	1.6	9.4	15.5	20	22	35.5	43	40.5	33
y	240	181	193	155	172	110	113	75	94

Find simple regression model using L.S. Find an estimate of σ^2.

Sol.

$\sum X = 220.5$

$\sum Y = 1333$

$\sum X^2 = 7053.67$

$\sum Y^2 = 220549$

$\sum XY = 26864.4$

$S_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 1651.42$

$S_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = -5794.1$

$S_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 23118.37$

$$b = \frac{S_{XY}}{S_{XX}} = -3.51$$

$$a = \bar{y} - b\bar{x} = 234.105$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = -0.94$$

$$\hat{\sigma}^2 = \frac{S_{YY} - bS_{XY}}{n-2} = \frac{S_{YY}(1-r^2)}{n-2} = 397.297$$

Regression line: $\hat{y} = 234.07 - 3.508x$

x	y	$\hat{y}$	$e_i = y_i - \hat{y}_i$ (Residuals)
22	172	156.87	15.155

$y_i = \alpha + \beta x_i + \epsilon_i$

$H_0: \beta = 0$ Vs. $H_1: \beta \neq 0$

$E(b) = \beta$, $V(b) = \frac{\sigma^2}{S_{XX}}$; $t = \frac{b - \beta}{\frac{\sigma}{\sqrt{S_{XX}}}} = \frac{-3.508 - 0}{0.4911} \sim t_{n-2}$

$t_{0.025,7} = 2.365$

(Multiple Regression)

Ex. 2. A study was performed on wear of a bearing y and its relation to X_1 (oil viscosity) and X_2 (load)

Y	X_1	X_2
293	1.6	851
230	15.5	816
172	22	1058
91	43	1201
113	33	1357
125	40	1115

Fit a multiple linear model to these data.

Sol. The regression equation is

$$y = 384 - 3.64x_1 - 0.112x_2$$

$S = 12.3539$, $R^2 = 98.58$, $R^2_{adj} = 97.58$

ANOVA Table

Source	DF	SS	MS	F	P
Regression	2	29787	14894	98.59	$0.002 < 0.005$
Residual Error	3	458	153		
Total	5	30245			

So, significant.

* calculations are done in MINITAB *

Scanned by CamScanner

Multiple Linear Regression:- $y = a + b_1 x_1 + b_2 x_2$

$$\underset{\sim}{y}_{6\times1} = X_{6\times3}\,\beta_{3\times1} + \epsilon_{6\times1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ 1 & x_{61} & x_{62} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$$SS = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$$

$$= Y'Y - X\beta'Y - Y'X\beta + (X\beta)'(X\beta)$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$$\frac{\partial SS}{\partial \beta} = -2X'Y + 2X'X\beta = 0$$

So, $X'X\beta = X'Y$

$$\underset{\sim}{\beta} = (X'X)^{-1}(X'Y)$$

$E(\hat{\beta}) = \beta$ (Show)

Note:- For linear regression, $R^2 = 1 - \frac{SSE}{TSS}$ = Proportion of variability in Y explained by the model.

$$R^2_{adjusted} = 1 - \frac{SSE/n-p-1}{TSS/n-1}$$

$$= 1 - \frac{SSE}{TSS} \times \frac{n-1}{n-p-1}$$

$$= 1 - \frac{(1-R^2)(n-1)}{(n-p-1)}$$

where p = number of predictors and
n = Total sample size.

Multicollinearity:- Relationship (linear) among predictor variables.

$$E(\hat{\beta}) = E\left[(X'X)^{-1}(X'Y)\right]$$

$$= E\left[(X'X)^{-1} X'(X\beta + \epsilon)\right]$$

$$= E\left[(X'X)^{-1} X'X\beta + (X'X)^{-1}X'\epsilon\right]$$

$= \beta$ since $E(\epsilon) = 0$ and $(X'X)^{-1}X'X = I$.

$\therefore \hat{\beta}$ is an unbiased estimator of β.

Analysis of Variance

One-way ANOVA:- $H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n$

H_1: at least one dquality is violated

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad ; \quad \epsilon_{ij} \sim N(0, \sigma^2)$$
$$i = 1(1)a, \quad j = 1(1)b.$$

$$\bar{y}_{oo} = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} / ab$$

$$\bar{y}_{io} = \sum_{j=1}^{b} y_{ij} / b$$

$$TSS = SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{oo})^2 = TSS$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{b} \left[(y_{ij} - \bar{y}_{io}) + (\bar{y}_{io} - \bar{y}_{oo}) \right]^2$$

Fixed effect Model:- $\alpha_i = \mu_i - \mu$

$$\sum_{i=1}^{a} \alpha_i = \sum (\mu_i - \mu) = 0$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{io})^2 + b \sum_{i=1}^{a} (\bar{y}_{io} - \bar{y}_{oo})^2$$

$$= SS_{within} + SS_{between}$$

Q. Show that $E(SSA) = (a-1)\sigma^2 + b\left(\sum_{i=1}^{a} \alpha_i^2\right)$

Solution:-

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$
$$\bar{y}_{io} = \mu + \alpha_i + \bar{\epsilon}_i$$
$$\bar{y}_{oo} = \mu + \bar{\epsilon}_{oo} \quad \left(\because \sum \alpha_i = 0\right)$$

$$\sum_{i=1}^{a} (\bar{y}_{io} - \bar{y}_{oo})^2 = \sum_{i=1}^{a} \alpha_i^2 + \sum_{i=1}^{a} (\bar{\epsilon}_{io} - \bar{\epsilon}_{oo})^2$$

$$SSA = b\sum_{i=1}^{a} \alpha_i^2 + b\sum_{i=1}^{a} (\bar{\epsilon}_{io} - \bar{\epsilon}_{oo})^2$$

$$E(SSA) = b\sum_{i=1}^{a} \alpha_i^2 + b.(a-1)\cdot\frac{\sigma^2}{b}$$

$$= (a-1)\sigma^2 + b\left(\sum_{i=1}^{a} \alpha_i^2\right)$$

$$E(MSA) = \sigma^2 + \frac{n\left(\sum \alpha_i^2\right)}{a-1}$$

$$E(SSE) = a(n-1)\sigma^2$$

Q. $\boxed{E(MSE) = \sigma^2}$

$\rightarrow$ $$SSE = \sum_{i=1}^{a} \sum_{j=1}^{b} \left(y_{ij} - \bar{y}_{i0}\right)^2$$

$$SSE = \sum\sum\left(\epsilon_{ij} - \bar{\epsilon}_{i0}\right)^2$$

$$E(SSE) = \sum_{i=1}^{a} (n-1)\sigma^2$$

$$= a(n-1)\sigma^2$$

$$E(MSE) = \frac{E(SSE)}{df} = \frac{a(n-1)\sigma^2}{a(n-1)} = \sigma^2.$$

Non-parametric Inference

1. Sign Test:- $H_0: \tilde{\mu} = \tilde{\mu}_0$ Vs. $H_1: \hat{\mu} \neq \hat{\mu}_0$

Data: 7.91 7.85 6.82 8.01 7.46 6.95 7.05 7.35 7.25 7.42

We want to know whether the population median is 7 or not.

Data-median: 0.91 0.85 −0.18 1.01 0.46 −0.05 0.05 0.35 0.25 0.42

\# +ve sign = 8

\# −ve sign = 2

$$\text{p-value} = 2\left[{}^{10}C_8 (.5)^8 (.5)^2 + {}^{10}C_9 (.5)^9 (.5) + {}^{10}C_{10} (.5)^{10}\right]$$

$$= 0.109 > 0.05.$$

So, median can be 7. we accept the null hypothesis.

For $n \geq 10$,
$$Z = \frac{R^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{8-5}{\sqrt{\frac{10}{4}}} = 1.897 \in (-1.96, 1.96)$$

$\therefore$ We accept H_0.

2. Wilcoxon Signed-Rank Test:- Assumption:- symmetric, continuous.

Rank: 9 8 −3 10 7 −1.5 1.5 5 4 6

Sum of +ve ranks = 50.5

Sum of −ve ranks = 4.5

$W = \min\{50.5, 4.5\} = 4.5 < 8$; we reject H_0.

$n \geq 20$,

$$Z = \frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

3. Mann-Whitney Test:- Assumption:- same shape, spread, but location differs.

$H_0: \mu_1 = \mu_2$ Vs. $H_1: \mu_1 \neq \mu_2$.

Sample 1: 25 27 29 31 30 26 24 32 33 38

Sample 2: 31 33 32 35 34 29 38 35 37 30

Arrangement:- $\overset{(1)}{24}$ $\overset{(1)}{25}$ $\overset{(1)}{26}$ $\overset{(1)}{27}$ $\overset{(1)}{29}$ $\overset{(2)}{29}$
$\overset{(1)}{30}$ $\overset{(2)}{30}$ $\overset{(1)}{31}$ $\overset{(2)}{31}$ $\overset{(1)}{32}$
$\overset{(2)}{32}$ $\overset{(1)}{33}$ $\overset{(2)}{33}$ $\overset{(2)}{34}$ $\overset{(2)}{35}$
$\overset{(2)}{35}$ $\overset{(2)}{37}$ $\overset{(1)}{38}$ $\overset{(2)}{38}$

R_1 = Sum of the ranks from sample 1 = 77

R_2 = " " " " " 2 = 133

$$R_1 + R_2 = 210 = \frac{n(n+1)}{2} = \frac{20 \times 21}{2}$$

$n_1 \geq 8$, $n_2 \geq 8$,

$$Z = \frac{R_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} = \frac{-4.23}{2} = -2.12 \notin (-1.96, 1.96)$$

$\therefore$ We reject H_0.

4. Kruskal-Wallis Test:- $H_0: \mu_1 = \mu_2 = \mu_3$
H_1: at least one inequality holds

More than two samples.

Method 1:- 553 550 568 541 537

Method 2: 553 599 579 545 540

Method 3: 492 530 528 510 571

Arrangement:- $\overset{(3)}{492}$ $\overset{(3)}{510}$ $\overset{(3)}{528}$ $\overset{(3)}{530}$ $\overset{(1)}{537}$
$\overset{(2)}{540}$ $\overset{(1)}{541}$ $\overset{(2)}{545}$ $\overset{(1)}{550}$ $\overset{(2)}{553}$
$\overset{(1)}{553}$ $\overset{(1)}{568}$ $\overset{(3)}{571}$ $\overset{(2)}{579}$ $\overset{(2)}{599}$

R_i = Sum of i's Rank
N = Total no. of readings
a = 3

$R_1 = 43.5$
$R_2 = 53.5$
$R_3 = 23$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_i^2}{n_i} - 3(N+1) \sim \chi^2_{\alpha, a-1}$$

$= 4.835$

$$p \text{ value} = \int_{4.835}^{\infty} f(\chi^2_2)\, d\chi^2 = 0.0898 > 0.05$$

We accept H_0.

Contingency Table

$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$

$H_0: p_1 = p_2 = \dots = p_K$
H_1: at least one equality is violated

V_i: Vendor i

	V_1	V_2	V_3	V_4	Total	
Good	182 $\boxed{188.02}$	147 $\boxed{141.02}$	89	163	581	R_1
Bad	18	3	5	11	37	R_2
Total	200	150	94	174	618 = G	
	C_1	C_2	C_3	C_4		

$$E_{11} = \frac{R_1 \times C_1}{G} = \frac{581 \times 200}{618} = 188.02$$

$$E_{12} = \frac{R_1 \times C_2}{G} = \frac{581 \times 150}{618} = 141.02$$

$E_{13} =$

$\vdots$

$$\chi^2_{calculate} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 7.577 < \chi^2_{0.05, 3}$$

Tabulated χ^2:-

Calculate $\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,3} = 7.815$

So, we fail to reject H_0.

$\therefore$ all vendors provide more or less same proportion of defective.

Working formula:- $H_0: p_{ij} = p_{i0} \times p_{0j}$

V_1	V_2	V_3	V_4
$d_1 = 18$	$d_2 = 3$	$d_3 = 5$	$d_4 = 11$

$$\hat{p} = \frac{18+3+5+11}{618} = \frac{37}{618} = 0.06$$

$E_1 = n_1\hat{p} = 200 \times 0.06 = 12$
$E_2 = n_2\hat{p} = 150 \times 0.06 = 9$
$E_3 = n_3\hat{p} = 94 \times 0.06 = 5.64$
$E_4 = n_4\hat{p} = 174 \times 0.06 = 10.44$

$$\chi^2_{calculated} = \sum_{i=1}^{4} \frac{(d_i - n_i\hat{p})^2}{\sqrt{n_i\hat{p}(1-\hat{p})}}$$

$= 7.577$

$$\boxed{\text{Show that } E(SSE) = (n-2)\sigma^2}$$

$\rightarrow$

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$= S_{yy} - bS_{xy}$$

$$E(S_{yy}) = E\left[\sum_i (y_i - \bar{y})^2\right]$$

$$= E\left[\sum_{i=1}^{n} (\alpha + \beta x_i + \epsilon_i - \alpha - \beta\bar{x} - \bar{\epsilon})^2\right]$$

$$= E\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \beta^2 + \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2\right]$$ [Product term vanishes]

$$= \beta^2 S_{xx} + (n-1)\sigma^2$$

$$E(bS_{xy}) = E\left(\frac{S_{xy}}{S_{xx}} \cdot S_{xy}\right) = \frac{1}{S_{xx}} E(S_{xy}^2)$$

$$E(S_{xy}) = E\left[\sum_i (x_i - \bar{x}) y_i\right]$$

$$= E\left[\sum_{i=1}^{n} (x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i)\right]$$

$$= \beta S_{xx} \quad ; \quad E(\epsilon_i) = 0$$

$$V(S_{xy}) = V\left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i\right) = V\left[\sum_{i=1}^{n} (x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i)\right]$$

$$= 0 + 0 + V\left(\sum_{i=1}^{n} (x_i - \bar{x})\epsilon_i\right)$$

$$= \sigma^2 S_{xx}$$

$$E(S_{xy}^2) = (\sigma^2 + \beta^2 S_{xx}) S_{xx}$$

So, $E(SSE) = E(S_{yy}) - E(bS_{xy})$

$$= (n-1)\sigma^2 - \sigma^2$$

$$= (n-2)\sigma^2$$

Solved Examples:-

1. A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline assumption in km/litre are given below:

Car :	1	2	3	4	5	6	7	8	9	10	11	12
Radial Tires:	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
Belted Tires:	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9

Can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires? Assume the population to be normal.

Solution:- $d_i = x_i - y_i$: 0.1   −0.2   0.4   0.1   −0.1   0.1
0   0.2   0.5   0.2   0.1   0.3

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

$H_0: \mu_d = 0$
$H_1: \mu_d > 0$

$$= \frac{0.14 - 0}{0.197/\sqrt{12}}, \text{ under } H_0$$

$$= 2.46 > t_{0.05,11} = 1.796$$

We reject H_0.

We conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires.

2. Ten engineering colleges in India were surveyed. The sample contained 250 electrical engineers, 80 being women; 175 chemical engineers, 40 being women. Compute a 90% C.I. for the difference between the proportions of women in these two fields of engineering. Is there a significant difference between the two proportions?

<u>Solution:-</u> $n_1 = 250$, $n_2 = 175$

$$\hat{p}_1 = \frac{80}{250} = 0.32, \quad \hat{p}_2 = \frac{40}{175} = 0.228$$

90% CI for $(p_1 - p_2)$:-

$$\left[\hat{p}_1 - \hat{p}_2 - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ \hat{p}_1 - \hat{p}_2 + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

$$= (0.018, 0.168)$$

$H_0: p_1 = p_2$

$H_1: p_1 \neq p_2$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad p = \frac{d_1 + d_2}{n_1 + n_2} = 0.282$$

$$= \frac{0.32 - 0.228}{\sqrt{0.282 \times (1 - 0.282\left(\frac{1}{250} + \frac{1}{250}\right)}}$$

$$= 2.04 \notin (-1.64, 1.64)$$

We reject H_0.

3. The grades in statistics course for a particular semester were as follows:

Grade	A	B	C	D	E
f	14	18	32	20	16

Test the hypothesis at 0.05 level of significance that the distribution of grades is uniform.

Sol. $H_0: X \sim \text{Uni}$
$H_1: X \nsim \text{Uni}$

$N = \sum_{i=1}^{n} f_i = 100.$ Expected freq. $= N \times f(x) = 100 \times \frac{1}{5} = 20 = E_i$

$$\chi_c^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 10 > \chi^2_{0.05,4} = 9.49$$

So, we reject H_0, distn. is not uniform.

4. In an experiment to study the dependence of hypertension on smoking habits, the following data were taken on 180 individuals:

	Non-smoker	Moderate	Heavy	Row sum
Hypertension	21	36	30	87
No hypertension	48	26	19	93
column sum	69	62	49	180

Test the hypothesis that the presence or absence of hypertension is independent of smoking habits using $\alpha = 0.05$.

Sol. H_0: smoking has no effect
H_1: smoking has no effect

$E_{11} = \frac{R_1 \times C_1}{G} = \frac{69 \times 87}{180} = 33.35$; $E_{12} = 29.97$

$E_{13} = \frac{R_1 \times C_3}{G} = \frac{87 \times 49}{180} = 23.68$; $E_{21} = 35.65$

$E_{22} = \frac{R_2 \times C_2}{G} = \frac{93 \times 62}{180} = 32.03$; $E_{23} = 25.32$

$$\chi_c^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 14.46 > \chi^2_{0.05,2} = \chi^2_{\alpha,(r-1)(c-1)} = 5.99$$

we reject H_0.

5. A study was made by a retail merchant to determine the relation between weekly advertising expenditures & sales.

Advt. Costs (X)	Sales (Y)	X^2	Y^2	XY
40	385	1600	148225	15400
20	400	400	160000	8000
25	395	625	156025	9875
20	365	400	133225	7300
30	475	900	225625	14250
50	440	2500	240100	22000
40	490	1600	176400	19600
20	420	400	313600	8400
50	560	2500	275625	28000
40	525	1600	230400	4000
25	480	625	260100	25500
50	510	2500		
Total: 410	5445	15650	2512925	191325

(a) $y = a + bx$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} = 3.22$$

$$\hat{a} = \bar{y} - b\bar{x} = \frac{5445}{12} - 3.22 \times \frac{410}{12} = 343.73$$

Regression equation is given by $y = 343.73 + 3.22x$.

(b) Suppose $x = 35$.

then estimate of weekly sales $= y = 343.73 + 3.22 \times 35$

$= 456.43$.

One-sample Testing Procedure

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region	Two sided $(1-\alpha)100\%$ C.I.
$H_0: \mu = \mu_0$ σ^2 known	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$H_1: \mu \neq \mu_0$ $H_1: \mu > \mu_0$ $H_1: \mu < \mu_0$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 > Z_\alpha$ $Z_0 < -Z_\alpha$	$\bar{x} - Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ β-error $= \Phi\left(Z_{\alpha/2} - \dfrac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \dfrac{\delta\sqrt{n}}{\sigma}\right)$ $\delta = \mu_1 - \mu_0$
$H_0: \mu = \mu_0$ σ^2 unknown	$t_0 = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$	$H_1: \mu \neq \mu_0$ $H_1: \mu > \mu_0$ $H_1: \mu < \mu_0$	$\lvert t_0 \rvert > t_{\alpha/2, n-1}$ $t_0 > t_{\alpha, n-1}$ $t_0 < -t_{\alpha, n-1}$	$\bar{x} - t_{\alpha/2, n-1}\dfrac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2, n-1}\dfrac{s}{\sqrt{n}}$
$H_0: \sigma^2 = \sigma_0^2$	$\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$H_1: \sigma^2 \neq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha/2, n-1}$ $\chi_0^2 > \chi^2_{\alpha, n-1}$ $\chi_0^2 < \chi^2_{1-\alpha, n-1}$	$\dfrac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \le \sigma^2 \le \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$
$H_0: p = p_0$ (proportion in Binomial)	$Z_0 = \dfrac{x - np_0}{\sqrt{np_0(1-p_0)}}$ $Z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$	$H_1: p \neq p_0$ $H_1: p > p_0$ $H_1: p < p_0$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 > Z_\alpha$ $Z_0 < -Z_\alpha$	$\hat{p} - Z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + Z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ $\hat{p} = \dfrac{d}{n}$. $95\% \rightarrow$, $Z_{\alpha/2} = 1.96$ $90\% \rightarrow$, $Z_{\alpha/2} = 1.64$

Two sample Testing Procedures

Null hypothesis	Test statistic	Alt. hypothesis	Rejection criteria	Two sided $100(1-\alpha)\%$ CI
$H_0: \mu_1-\mu_2=\Delta_0$ σ_1^2, σ_2^2 known	$z_0 = \dfrac{\bar{x}_1-\bar{x}_2-\Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$H_1: \mu_1-\mu_2 \neq \Delta_0$ $H_1: \mu_1-\mu_2 > \Delta_0$ $H_1: \mu_1-\mu_2 < \Delta_0$	$\lvert z_0\rvert > z_{\alpha/2}$ $z_0 > z_\alpha$ $z_0 < -z_\alpha$	$\bar{x}_1-\bar{x}_2-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le$
$H_0: \mu_1-\mu_2=\Delta_0$ $\sigma_1^2=\sigma_2^2=$ unknown	$t_0 = \dfrac{\bar{x}_1-\bar{x}_2-\Delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $s_p^2 = \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$H_1: \mu_1-\mu_2 \neq \Delta_0$ $H_1: \mu_1-\mu_2 > \Delta_0$ $H_1: \mu_1-\mu_2 < \Delta_0$	$\lvert t_0\rvert > t_{\frac{\alpha}{2}, n_1+n_2-2}$ $t_0 > t_{\alpha, n_1+n_2-2}$ $t_0 < -t_{\alpha, n_1+n_2-2}$	$\bar{x}_1-\bar{x}_2-t_{\alpha/2, n_1+n_2-2}\cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le$
$H_0: \mu_1-\mu_2=\Delta_0$ $\sigma_1^2 \neq \sigma_2^2$ unknown	$t_0 = \dfrac{\bar{x}_1-\bar{x}_2-\Delta_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ $df = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$	$H_1: \mu_1-\mu_2 \neq \Delta_0$ $H_1: \mu_1-\mu_2 > \Delta_0$ $H_1: \mu_1-\mu_2 < \Delta_0$	$\lvert t_0\rvert > t_{\alpha/2, df}$ $t_0 > t_{\alpha, df}$ $t_0 < -t_{\alpha, df}$	$\bar{x}_1-\bar{x}_2-t_{\alpha/2, df}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1-\mu_2$
$H_0: \mu_D=0$ Paired data	$t_0 = \dfrac{\bar{d}-\mu_{D(0)}}{s_d/\sqrt{n}}$	$H_1: \mu_d \neq 0$ $H_1: \mu_d > 0$ $H_1: \mu_d < 0$	$\lvert t_0\rvert > t_{\alpha/2, n-1}$ $t_0 > t_{\alpha, n-1}$ $t_0 < -t_{\alpha, n-1}$	$\bar{d}-t_{\alpha/2, n-1}\cdot\frac{s_d}{\sqrt{n}} \le \mu_D \le \bar{d}+t_{\alpha/2, n-1}\cdot\frac{s_d}{\sqrt{n}}$
$H_0: \sigma_1^2=\sigma_2^2$	$f_0 = \dfrac{s_1^2}{s_2^2}$	$H_1: \sigma_1^2 \neq \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$f_0 > f_{\alpha/2, n_1-1, n_2-1}$ $f_0 > f_{\alpha, n_1-1, n_2-1}$	$\frac{s_1^2}{s_2^2}\cdot f_{1-\alpha/2, n_1-1, n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot f_{\alpha/2, n_1-1, n_2-1}$
$H_0: p_1=p_2$ Binomial	$z_0 = \dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $\hat{p} = \dfrac{d_1+d_2}{n_1+n_2}$	$H_1: p_1 \neq p_2$ $H_1: p_1 > p_2$ $H_1: p_1 < p_2$	$\lvert z_0\rvert > z_{\alpha/2}$ $z_0 > z_\alpha$ $z_0 < -z_\alpha$	$\hat{p}_1-\hat{p}_2-z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1-p_2 \le$