19/23

Statistics and Information Theory

=> Gaussian Distribution (μ, σ^2):

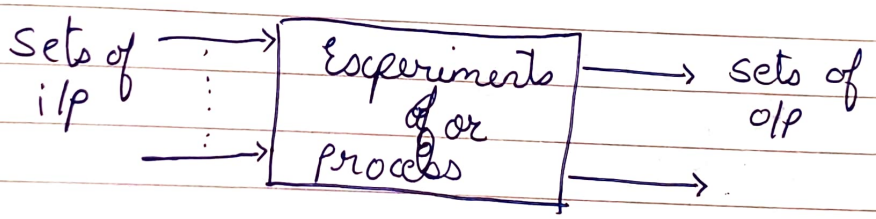
PDF: $f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

=> standard gaussian distribution:

$\mu = 0 \qquad \sigma^2 = 1$

$f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

=> Variance ⟶ Variance of data → spread of data across the mean value.



Agriculture

$y = f(x_1, x_2, x_3)$

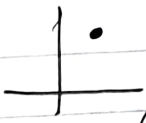
- y: yield of crop (in kg)
- x_1: quantity of fertilizer used.
- x_2, x_3: ...

Model.

$y = f(x_1, x_2)$

- y: Dosage of Meds to be given
- x_1: Age of patient
- x_2: BP level of patient

$y = mx + c$ ~~$\tan\theta = m$~~

Data sheet 1	Data sheet 2
$x = 2, y = 5$ $5 = 2m + c$ $(2, 5)$  Just a point (can't say which line it is on) Not much info	$m = 3, c = 4$ $y = 3x + 4$ Gives multiple points ~~on a~~ on a single line. Much info.

$y = mx + c$ → Variables (x, y)
→ Parameters (m, c)

Model
- → Parameters
- → Variables

- Parameters play a key role here is state
- Parameters : Component of a model which illustrates complete information of model.

Model — Good Model [Many salient features are illustrated]
Bad Model

Process/ Experiment → Collect the data/ ~~observe~~ the data

Process / Experiment → Collect the data / observe the data → Predict the Model / Predict the parameters

Joint PDF: $\frac{1}{(2\pi)^{N/2}} e^{-1/2 \sum_{i=1}^{n} x_i^2}$

$y = mx + c \Rightarrow y = B_0 + B_1 x$

B_0: Intercept parameter, B_1: slope parameter

$$y = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_n x_n$$

Generic Model Regression.

$y = B_0 + B_1 x.$

⇓

Data set N different values of y

$B_0 \Rightarrow \hat{B}_0$

$B_1 \Rightarrow \hat{B}_1$

Model
- Mathematical Model (Theoretical / lack practical Inference)
- Statistical Model

$y = 3x + 2$

- y → yield of crop (in kg)
- x → Quantity of field fertilizer

field of loop (in kg)

$y = 3x + 2$

↓

quantity of fertilizer (in kg)

mathematical model:

(theoretical and lacks practical inferences)

x	y
1	5
2	8
⋮	⋮
100	302

Statistical Model:

$x = 1$ kg $y = 5$ kg → Not exact depend on various factors such as rainfall, draught etc.

$x = 1$ kg $y = 4.7$ kg
$x = 1$ kg $y = 5.1$ kg
} Shows the actual yeild obtained

according to statistical model, on average when the experiment of crop yeild was conducted.

$y = \beta_0 + \beta_1 x$

By statistical model, on an average ~~when the experiment~~ we can get estimated values $\tilde{\beta}_0$ and $\tilde{\beta}_1$ which will be close to β_0 and β_1 respectively.

$\tilde{\beta}_0 = \beta_0 + \epsilon$ $\qquad$ $\hat{\beta}_1 = \beta_1 + \theta$

$\theta, \epsilon \rightarrow$ error

$\therefore y = \beta_0 + \beta_1 x + \epsilon$

B.Tech

↓ 4 Branches

CS, EC, CIVIL, MECH

30 30 30 30

Total students = 400

(chose 30 from each)

Sample: 120 students
Population: 400 students

Sample: subset of population

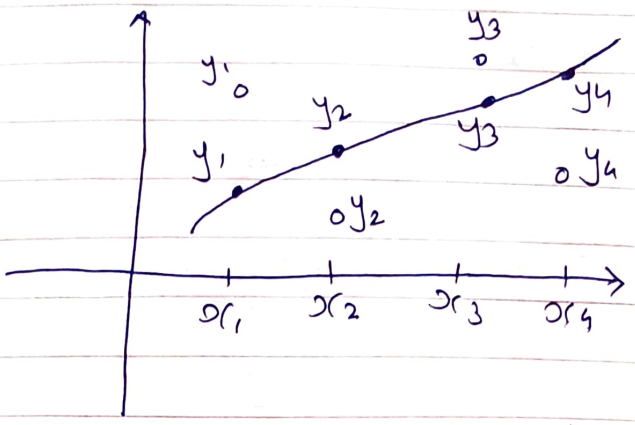
$y = \beta_0 + \beta_1 x$

$\hookrightarrow y = \beta_0 + \beta_1 x + \epsilon$

n observations:

$y(1), y(2), \ldots, y(n) \rightarrow y_1, y_2, \ldots, y_n$

$x(1), x(2), \ldots, x(n) \rightarrow x_1, x_2, \ldots, x_n$.



• → theoretical
o → actual data.

Let individual errors for n observations be $\epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n$

Average Error throughout experiment

$$= \frac{1}{n} \sum_{i=1}^{n} \epsilon_i$$

In some case, $\sum \epsilon_i = 0$ But you can't say that there were no errors in your experiments as there can be +ve and -ve errors at each observation which can add upto zero.

Hence, we use a more accurate metric of ϵ_i^2.

Least Squared Estimation

$\sum_{i=1}^{n} \epsilon_i^2$ ↩ LSE : Least squared estimation

Sum of squared errors across all observations

$\sum_{i=1}^{n} \epsilon_i$

Sum of errors across all observations.

Purpose: To estimate parameters β_0 - β_1 through set of data (x_i, y_i)

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

our task is to min $\left\{ \sum_{i=1}^{n} \varepsilon_i^2 \right\}$

Sample of n values

i/p: $x_1, x_2, \ldots, x_n$

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

o/p: $y_1, y_2, \ldots, y_n$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$\boxed{f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$$

to find minima, we find derivative

$$\frac{\partial f(\cdot)}{\partial \beta_0} = 0 \qquad \frac{\partial f(\cdot)}{\partial \beta_1} = 0$$

$$\sum 2 (y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$-2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$-2 (n\bar{y} - n\beta_0 - n\beta_1 \bar{x}) = 0$$

$$-\bar{y} + \beta_0 + \beta_1 \bar{x} = 0$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$2 (y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i) = 0$$

$$-2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot x_i = 0$$

~~$-2(n\bar{y} - n\beta_0 - n\beta_1 \bar{x})$~~

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \beta_0 n \bar{x}}{\sum_{i=1}^{n} x_i^2}$$

(you need to know value of β_0 to find β_1 and vice

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 n\bar{x}}{\sum_{i=1}^{n} x_i^2} \qquad \hat{\beta}_0 \to (\bar{y} - \hat{\beta}_1 \bar{x})$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \bar{y} n\bar{x} + \hat{\beta}_1 n\bar{x}^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - \hat{\beta}_1 n\bar{x}^2}$$

Denominator term:

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

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

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

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

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Numerator term:

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

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

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

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

LSE → Least Square Estimator

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Hessian Matrix

$$H = \begin{bmatrix} \frac{\partial^2 F(\cdot)}{\partial \beta_0^2} & \frac{\partial^2 F(\cdot)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 F(\cdot)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 F(\cdot)}{\partial \beta_1^2} \end{bmatrix}$$

$$H = \begin{bmatrix} \sum 2 & \sum 2x_i \\ \sum 2x_i & \sum 2x_i^2 \end{bmatrix} = 2\begin{bmatrix} 1 & 1 \\ x_i & x_i \end{bmatrix}$$

$$= 2\begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

$$\det(H) = 4n\sum_{i=1}^{n} x_i^2 - 4n^2\bar{x}^2$$

$$= 4n\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

$$= 4n\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Parameter - fixed unknown value.

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

$S_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 are linear w.r.t. output values / output observations (y_i)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_{xx}}$$

$\sum x_i\bar{y} - \sum\bar{x}\bar{y}$

$i \quad n\bar{x}\bar{y} - \bar{x}\bar{y}n$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{S_{xx}} - \underbrace{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\bar{y}}{S_{xx}}}_{0}$$

x_i are known as they are input values

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{S_{xx}}$$

$$\hat{\beta}_i = \sum_{i=1}^{n} c_i y_i$$

Let $c_i = \frac{x_i - \bar{x}}{S_{xx}}$

$$c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

$$\hat{\beta}_1 = c_1 y_1 + c_2 y_2 + \dots + c_n y_n$$

The estimator $\hat{\beta}_1$ is linear w.r.t. y_i.

$$E[\hat{\beta}_1] = \beta_1$$

↑ average sense.

Unbiased Estimator and Bias of Estimator

- The estimator on an average sense should yield the value of the parameter. (It is estimating).

$$E[\hat{\beta}_1] = \beta_1$$

$$E[\hat{\beta}_0] = \beta_0$$

$$\text{Bias}(\hat{\beta}_1 ; \beta_1) = E(\hat{\beta}_1) - \beta_1$$

$$\text{Bias}(\hat{\beta}_0 ; \beta_0) = E(\hat{\beta}_0) - \beta_0$$

$$E\{\hat{\beta}_1\} = \beta_1 \qquad \hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

$$E\{\hat{\beta}_1\} = E\left\{\sum_{i=1}^{n} c_i y_i\right\} \qquad c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

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

$$= \sum_{i=1}^{n} c_i E\{y_i\}$$

$$= \sum_{i=1}^{n} c_i E\{\beta_0 + \beta_1 x_i + \varepsilon_i\}$$

$\varepsilon_i \sim N(0, \sigma^2)$

$$= \sum_{i=1}^{n} c_i \left[\underbrace{E(\beta_0)}_{\beta_0} + \underbrace{E(\beta_1 x_i)}_{\beta_1 x_i} + \underbrace{E(\varepsilon_i)}_{0}\right]$$

$$E\{\hat{\beta}_1\} = \sum_{i=1}^{n} c_i [\beta_0 + \beta_1 x_i]$$

$$E\{\hat{\beta}_1\} = \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{T_1} + \beta_1 \underbrace{\sum_{i=1}^{n} c_i x_i}_{T_2}$$

Coming to T_1:

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}}$$

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

$$= \frac{1}{S_{xx}} \left[\underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} - \underbrace{\sum_{i=1}^{n} \bar{x}}_{n\bar{x}}\right]$$

$$\boxed{\sum_{i=1}^{n} c_i = 0}$$

$$T_1 = 0$$

$$T_2 = \sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \right) x_i$$

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

$$= \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right]$$

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

$$= 1$$

$\therefore$ $E\{\hat{\beta}_1\} = \beta_1 (1)$

$= \beta_1$

$Bias(\hat{\beta}_1, \beta_1) = E(\hat{\beta}_1) - \beta_1$

$= \beta_1 - \beta_1$

$= 0$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$E\{\hat{\beta}_0\} = E\{\bar{y} - \hat{\beta}_1 \bar{x}\}$

$= E\{\bar{y}\} - E\{\hat{\beta}_1 \bar{x}\}$

$= E(\bar{y}) - E\{\bar{x} \sum_{i=1}^{n} c_i y_i\}$.

$= E(\bar{y}) - E\{\bar{x} \beta_1\}$

$= E(\bar{y}) - \bar{x} \beta_1$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$E(\bar{y}) = E\left\{\frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \varepsilon_i)\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\{\beta_0 + \beta_1 x_i + \varepsilon_i\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}[\beta_0 + \beta_1 x_i + 0]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\beta_1 x_i$$

$$= \frac{1}{n}(n\beta_0) + \frac{1}{n}(\beta_1 \times n\bar{x})$$

$$= \beta_0 + \beta_1 \bar{x}$$

$$E(\hat{\beta}_0) = E(\bar{y}) - \beta_1 \bar{x}$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0$$

$$(x_i, y_i) \Rightarrow (x_1, y_1) \ldots\ldots (x_n, y_n)$$

β_0	β_1
$\downarrow$	$\downarrow$
α	y
$\hat{\beta}_0$	β_1

$$Var(X-Y) = Var(X) + Var(-Y)$$
$$= Var(X) + Var(Y) - 2Cov(X,Y)$$

Variance and Covariance

X, Y

~~Cov(x,y)~~ (

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

$$= E[(X-\bar{X})(Y-\bar{Y})]$$

$E[X]$ → μ_x ; $E(Y)$ → μ_Y

$$Var(X+Y) = Var(x) + Var(Y) + 2Cov(X,Y)$$

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Non-random terms}} + \underbrace{\varepsilon_i}_{\text{Normal distribution}}, \quad \varepsilon_i \overset{IID}{\sim} N(0, \sigma^2)$$

y_i will also follow normal distribution

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i)$$
$$= E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)$$
$$= \beta_0 + \beta_1 x_i$$

$E(y_1) = E(\beta_0 + \beta_1 x + \varepsilon_1) = \beta_0 + \beta_1 x_1$

$E(y_2) = E(\beta_0 + \beta_1 x_2 + \varepsilon_2) = \beta_0 + \beta_1 x_2$

$\Rightarrow E(y_i) = \beta_0 + \beta_1 x_i$

$$\text{Var}(y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i)$$
$$= \text{Var}(\beta_0) + \text{Var}(\beta_1 x_i) + \text{Var}(\varepsilon_i)$$
$$= 0 + 0 + \sigma^2$$
$$= \sigma^2$$

$\text{Var}(\hat{\beta}_0)$ $\text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \text{Var}(\bar{y}) + \text{Var}(\hat{\beta}_1 \bar{x}) - 2\,\text{Cov}(\bar{y}, \hat{\beta}_1 \bar{x})$

$$= \text{Var}(\bar{y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) - 2\,\text{Cov}(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$\text{Cov}(\bar{y}, \hat{\beta}_1 \bar{x}) = \bar{x}\,\text{Cov}(\bar{y}, \hat{\beta}_1)$$

$$= \bar{x}\, E[(\bar{y} - E(\bar{y}))(\hat{\beta}_1 - E(\hat{\beta}_1))]$$

$$= \bar{x}\, E[\underbrace{(\bar{y} - E(\bar{y}))}_{T_1} \underbrace{(\hat{\beta}_1 - \beta_1)}_{T_2}]$$

$$T_1 = \bar{y} - E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i - E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \varepsilon_i) - \frac{1}{n}\sum_{i=1}^{n} E(\beta_0 + \beta_1 x_i + \varepsilon_i)$$
$$= \frac{1}{n}[n\beta_0 + n\beta_1 \bar{x} + n\bar{\varepsilon}] - \frac{1}{n}[n\beta_0 + n\beta_1 \bar{x} + 0]$$
$$= \bar{\varepsilon}$$

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i = \frac{1}{n}[\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n]$$

$$E(\bar{\varepsilon}) = \frac{1}{n}[E(\varepsilon_1) + E(\varepsilon_2) + \cdots + E(\varepsilon_n)]$$
$$= \frac{1}{n}(0 + 0 + \cdots + 0)$$
$$= 0 \qquad (\because \varepsilon_i \sim N(0, \sigma^2))$$

$$E(\bar{\varepsilon}) = 0$$

$\text{Var}(\bar{y}) = \text{Var}\left(\frac{1}{n}\sum y_i\right)$

$\frac{1}{n^2}\sum \text{Var}(y_i)$

$\frac{1}{n^2}\sum \sigma^2 = \sigma^2/n$

$T_2 = \hat{\beta}_1 - \beta_1$

$= \sum_{i=1}^{n} c_i y_i - \beta_1$

$= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i + \varepsilon_i) - \beta_1$

$= \beta_0 \sum_{i=1}^{n} c_i \;(\to 0) + \beta_1 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i \varepsilon_i - \beta_1$

$= \beta_1 (1) + \sum_{i=1}^{n} c_i \varepsilon_i - \beta_1$

$= \sum_{i=1}^{n} c_i \varepsilon_i$

$\text{Cov}(\bar{y}, \hat{\beta}_1 \bar{x}) = \bar{x}\, E\left[\bar{\varepsilon}\left(\sum_{i=1}^{n} c_i \varepsilon_i\right)\right]$

$= \bar{x}\, \bar{\varepsilon} \sum_{i=1}^{n} c_i E[\varepsilon_i]$

$= 0.$

$\text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \text{Var}(\bar{y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}$

$\text{Var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \text{var}(y_i) + 2\sum_{i=1}^{n}\sum_{k=1}^{n} c_i c_k \text{Cov}(y_i, y_k)$

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

$y_1, y_2 \to$ are independent.

$\therefore \text{Var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \text{var}(y_i).$

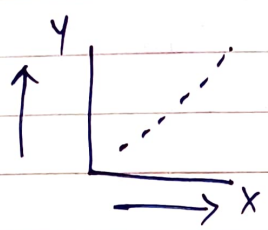
$\sum c_i^2 = \frac{\sum (x_i - \bar{x})^2}{S_{xx}^2} = \sum c_i^2 \sigma^2$

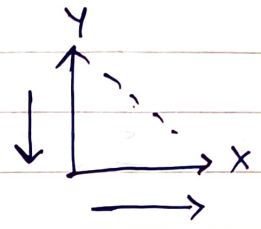
$= \frac{1}{S_{xx}^2} \sum (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \sigma^2 \sum c_i^2 = \frac{\sigma^2}{S_{xx}}.$

$= \frac{1}{S_{xx}^2}\left(\sum x_i^2 - n\bar{x}^2\right) = 1/S_{xx}$

- $Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1)$
$= Cov(\bar{y}, \hat{\beta}_1) - Cov(\hat{\beta}_1 \bar{x}, \hat{\beta}_1)$
$= 0 - \bar{x}\, Cov(\hat{\beta}_1, \hat{\beta}_1)$
$= -\bar{x}\, Var(\hat{\beta}_1)$

$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ $\qquad$ $Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$





- Covariance measures the joint variability of two variables

if $X\uparrow$ with $Y\uparrow$ then $Cov(X,Y)$ is +ve.
if $X\downarrow$ with $Y\uparrow$ then $Cov(X,Y)$ is -ve.
if $X\downarrow$ with $Y\downarrow$ then $Cov(X,Y)$ is +ve.

$Cov(X,Y)$ gives a <u>linear Relationship</u>

X:	-2	-1	0	1	2	} $Cov(X,Y)=0$
Y:	4	1	0	1	4	

$Y = X^2$ $\Rightarrow$ relationship does exist b/w X and Y but it is not linear.

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\qquad$ $\varepsilon_i \sim N(0, \sigma^2)$

<u>IID</u>

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ $\Rightarrow$ InID

InID $\rightarrow$ Independent but not identical distribution

nInID $\rightarrow$ not Independent and not identical distribution

nIID $\rightarrow$ not independent but Identical distribution.

$i = 1 \Rightarrow y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$

$i = 2 \Rightarrow y_2 = \underline{\beta_0 + \beta_1 x_2} + \varepsilon_2$

↳ these are a constant independent value.

as y_i has different mean, it is not identically distributed.

Actual Response : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Fitted Model / Fitted Response : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\boxed{r_i = y_i - \hat{y}_i}$ $\rightarrow$ residual.

Squared Sum of Residual :

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

we know, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$= \sum_{i=1}^{n} [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$SS_{res} = S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}$$

we know that, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\frac{S_{xy}}{S_{xx}} S_{xy}$$

$$= S_{yy} + \frac{S_{xy}^2}{S_{xx}^2} S_{xx} - 2\frac{S_{xy}^2}{S_{xx}}$$

$$= S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy}$$

$$\boxed{SS_{res} = S_{yy} - \hat{\beta}_1 S_{xy}}$$

$S_{xy} = \hat{\beta}_1 S_{xx}$

$$\boxed{SS_{res} = S_{yy} - \hat{\beta}_1^2 S_{xx}}$$

$E\{SS_{res}\} = (n-2)\sigma^2$ // em nam thoki didi

$$\boxed{E\left\{\frac{SS_{res}}{n-2}\right\} = \sigma^2} \Rightarrow SS_{res} \text{ is unbiased estimator of } \sigma^2$$

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-2}$$

Maximum likelihood estimator (MLE)

n R.V.S => $X_1, X_2, \ldots, X_n$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\quad \varepsilon_i \sim N(0, \sigma^2)$ → IID

$y_1, y_2, \ldots, y_n$ are independent.

Joint pdf: $f(y_1, y_2, \ldots y_n) = f(y_1) f(y_2) \ldots f(y_n)$

→ is product of marginal pdf.

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1 - \beta_1 x_1 - \beta_0)^2}{2\sigma^2}}\right) \ldots \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n - \beta_0 - \beta_1 x_n)^2}{2\sigma^2}}\right)$$

$$l(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}$$

⇓

likelihood function

log likelihood function:

$$L(\beta_0, \beta_1, \sigma^2) = \log_e l(\beta_0, \beta_1, \sigma^2)$$

$$L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial L(\cdot)}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial L(\cdot)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i$$

$$\frac{\partial L(\cdot)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial L(\cdot)}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = 0$$

$$n\bar{y} - n\beta_0 - n\beta_1 \bar{x} = 0$$

$$\therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{---} (1)$$

$$\frac{\partial L(\cdot)}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) n\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} = \beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{---} (2)$$

$$\frac{\partial L(\cdot)}{\partial \sigma^2} = 0 \Rightarrow \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i^2 + \beta_0^2 + \beta_1^2 x_i^2 - 2\beta_0 y_i - 2\beta_1 x_i y_i + 2\beta_0 \beta_1 x_i)$$

$$n\sigma^2 = \sum_{i=1}^{n} y_i^2 + n\beta_0^2 + \beta_1^2 \sum x_i^2 - 2n\beta_0 \bar{y} - 2\beta_1 \sum_{i=1}^{n} x_i y_i + 2n\beta_0 \beta_1 \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

. Hessian Matrix

$$H = \begin{bmatrix} \frac{\partial^2 L}{\partial \beta_0^2} & \frac{\partial^2 L}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 L}{\partial \beta_0 \partial \sigma^2} \\ \frac{\partial^2 L}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 L}{\partial \beta_1^2} & \frac{\partial^2 L}{\partial \beta_1 \partial \sigma^2} \\ \frac{\partial^2 L}{\partial \sigma^2 \partial \beta_0} & \frac{\partial^2 L}{\partial \sigma^2 \partial \beta_1} & \frac{\partial^2 L}{\partial \sigma^{2^2}} \end{bmatrix}$$

$\det(H) < 0$

LSE / MLE $\Rightarrow y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\varepsilon_i \overset{IID}{\sim} N(0, \sigma^2)$

LSE:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$\hat{\sigma}_0^2 = \frac{1}{(n-2)} SS_{residual}$

MLE:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$\hat{\sigma}^2 = \frac{1}{n} SS_{residual}$

$X \sim N(E(x_i), Var(x_i))$ $\quad X_i \overset{IID}{\sim} N(\mu, \sigma^2)$

MLE: $\hat{\mu}, \hat{\sigma}^2$

If these both parameter are unknown

Joint PDF: $f(x_1, x_2, \ldots, x_n)$

$= f(x_1)\, f(x_2), \ldots f(x_n)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right) \cdots$$

$$\cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

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

$$L = \log l$$

$$L = \frac{-n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \quad \Rightarrow \frac{\partial L}{\partial \mu} = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

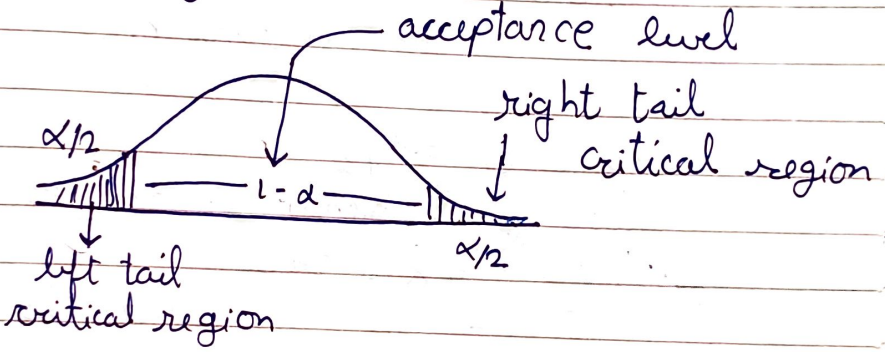
$$\frac{\partial L}{\partial \sigma^2} = 0 \quad \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Hypothesis Testing / Confidence Level / Significance Level

Point estimator → A particular single estimated value of unknown parameter

Interval Estimator → The estimated value of unknown parameter lies in an interval.

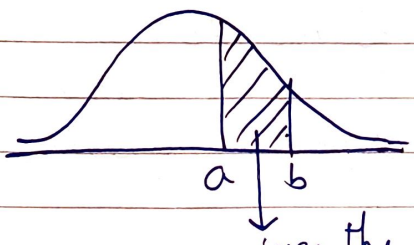
Confidence Interval:
PDF is divided into
(a) acceptance region
(b) critical region.



Confidence level $(1-\alpha)$
Significance level (α)

Significance level = 1 - Confidence level

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



gives the probability that r.v X lies b/w a and b

Q. Marks of an exam of 400 students are normally distributed with a population standard deviation of 5.6. A random sample with marks of 40 students is collected which has mean of 32

Estimate the population mean with a confidence of

(a) 80% (b) 90% (c) 98%

N = 400

n = 40.

$\bar{x} = 32$

$\sigma = 5.6$

$1 - \alpha = 0.8$

$\alpha = 0.2$

$\alpha/2 = 0.1$

population $\longrightarrow \mu, \sigma^2$

sample $\longrightarrow \bar{x}, \sigma^2/n$

mean variance

$\frac{32 - \mu}{5.6/\sqrt{40}}$

Population = 400

Choose 4 students from each class $\Rightarrow$

Sample = $4 \times 10 = 40$

$\bar{x} = 32$ for 40 students.

$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

According to point estimator, mean for 400 students will also be 32.

(a) Confidence level 80%

$1 - \alpha = 0.8 \Rightarrow \alpha = 0.2 \Rightarrow \alpha/2 = 0.1$

Interval estimator:

(LCL) Lower Confidence limit: $\bar{x} - E$

(UCL) Upper Confidence limit: $\bar{x} + E$

$E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$n \rightarrow$ sample mean

$$E = z_{0.1} \frac{(5.6)}{\sqrt{40}} = 1.13$$

$$\downarrow$$

$$1.28$$

$$LCL = 32 - 1.13 = 30.87$$
$$UCL = 32 + 1.13 = 33.13$$

$$\therefore \quad \mu \in [30.87, 33.13]$$

(b) 90% Confidence

$$1 - \alpha = 0.9 \Rightarrow \alpha = 0.1 \Rightarrow \alpha/2 = 0.05$$
$$E = z_{0.05} \left(\frac{5.6}{\sqrt{40}}\right) = 1.46$$

$$LCL = 32 - 1.46 = 30.54$$
$$UCL = 32 + 1.46 = 33.46$$

$$\mu \in [30.54, 33.46]$$

(c) 98% confidence

$$1 - \alpha = 0.98 \Rightarrow \alpha = 0.02 \Rightarrow \alpha/2 = 0.01$$
$$E = z_{0.01} \left(\frac{5.6}{\sqrt{40}}\right) = 2.06$$

$$\mu \in [29.94, 32.06]$$

$$X_i \sim N(\mu, \sigma^2)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

$$E(\bar{x}) = \mu \qquad Var(\bar{x}) = \sigma^2/n$$

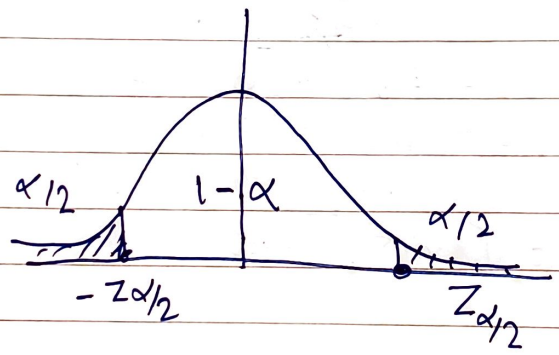
Test statistic $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$$E(Z) = \frac{E(\bar{x}) - E(\mu)}{\sigma/\sqrt{n}} = \frac{\mu - \mu}{\sigma/\sqrt{n}} = 0$$

$$Var(Z) = \frac{1}{(\sigma/\sqrt{n})^2}\left[(Var(\bar{x})) + (Var(\mu))\right]$$

$$= \frac{1}{(\sigma/\sqrt{n})^2}\left((\sigma^2/n) + 0\right)$$

$$= 1$$



$$P[-Z_{\alpha/2} < Z < Z_{\alpha/2}] = 1 - \alpha$$

$$-Z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}$$

$$\underbrace{\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{LCL} < \mu < \underbrace{\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{UCL}$$

Linear Regression Model:

LRM: $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i = \frac{S_{xy}}{S_{xx}}$

$$\hat{\beta}_1 \sim N(E(\hat{\beta}_1), Var(\hat{\beta}_1))$$

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Test statistic : $Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$

$$E(Z) = \frac{E(\hat{\beta}_1) - E(\beta_1)}{\sigma / \sqrt{S_{xx}}} = \frac{\beta_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} = 0$$

$$P[-Z_{\alpha/2} < Z_1 < Z_{\alpha/2}] = 1 - \alpha$$

$$P\left[-Z_{\alpha/2} < \frac{\beta_1 - \hat{\beta}_1}{\sigma / \sqrt{S_{xx}}} < Z_{\alpha/2}\right] = 1 - \alpha.$$

$$P\left[\hat{\beta}_1 - Z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + Z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}}\right] = 1 - \alpha.$$

$\hat{\beta}_1 \rightarrow$ Point Estimator for β_1

$$P\left[\hat{\beta}_1 - Z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + Z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}}\right] = 1 - \alpha.$$

$\rightarrow$ Internal estimators

$$\left[\hat{\beta}_1 - E, \hat{\beta}_1 + E\right]$$

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}}.$$

$\beta_0 \xrightarrow{PE} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\beta_0 \xrightarrow{IE} [\hat{\beta}_0 - E, \hat{\beta}_0 + E]$

$\hat{\beta}_0 \sim N(E(\hat{\beta}_0), Var(\hat{\beta}_0))$

$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right])$

Finding the test statistics:

→ should follow standard normal distribution

$$Z_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$$

$$= \frac{\hat{\beta}_0 - \beta_0}{\theta}$$

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

$$E[Z_0] = E\left[\frac{\hat{\beta}_0 - \beta_0}{\theta}\right]$$

$$= \frac{1}{\theta}\left(E[\hat{\beta}_0] - \beta_0\right]$$

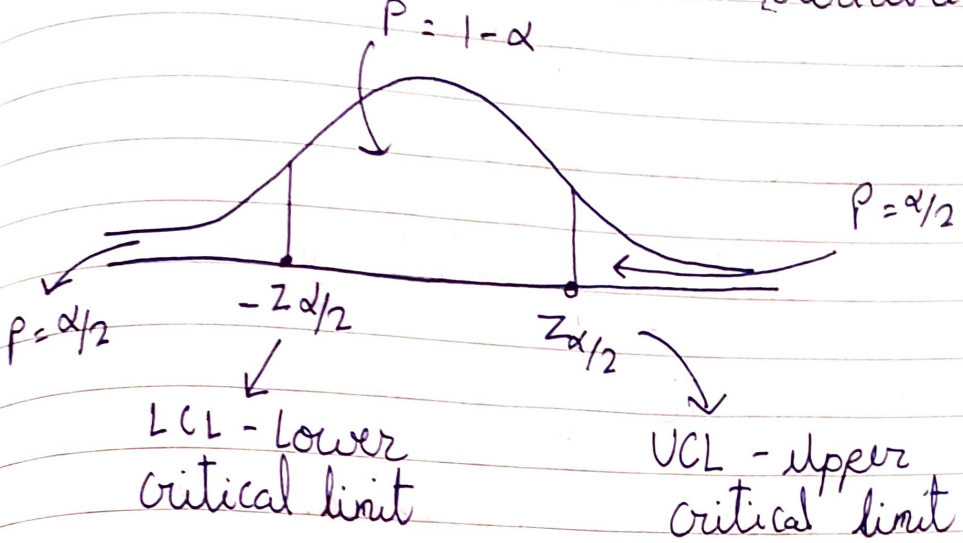
$$= 0$$

$$Var(Z_0) = \frac{1}{\theta^2}\left[Var(\hat{\beta}_0) + Var(\beta_0)\right]$$

$$= \frac{1}{\theta^2}(\theta^2 + 0)$$

$$= 1$$

Thus, Z_0 is standard normal variable.



$$P\left[-Z_{\alpha/2} < Z_0 < Z_{\alpha/2}\right]$$

$$= 1 - \alpha$$

$$= P\left[-Z_{\alpha/2} < \frac{\hat{\beta}_0 - \beta_0}{\hat{\theta}} < Z_{\alpha/2}\right]$$

$$= P\left[-Z_{\alpha/2}\cdot\hat{\theta} < \hat{\beta}_0 - \beta_0 < Z_{\alpha/2}\cdot\hat{\theta}\right]$$

$$= P\left[\hat{\beta}_0 - Z_{\alpha/2}\cdot\hat{\theta} < \beta_0 < \hat{\beta}_0 + Z_{\alpha/2}\hat{\theta}\right]$$

$\hat{\beta}_0 - Z_{\alpha/2}\cdot\hat{\theta}$ → Lower Confidence limit (LCL)

$\hat{\beta}_0 + Z_{\alpha/2}\hat{\theta}$ → Upper Confidence Limit (UCL)

Confidence Interval.

CL: $1-\alpha \longrightarrow (1-\alpha)100$ %

SL: $1-CL \longrightarrow (100-CL)$ %.

θ —PE→ $\hat{\theta}$

θ —IE→ $\hat{\theta}-E$ to $\hat{\theta}+E$

Test of Hypothesis:

- Null Hypothesis:

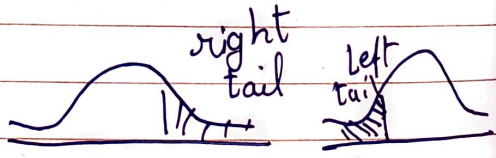
  Something presented by H_0

- Alternate Hypothesis:

→ Something represented by H_1.
→ H_0 and H_1 are complementary to each other
→ Alternate hypothesis always contradicts null hypothesis.

Verdict RT-PCR \ truth	Covid -ve Innocent	Covid +ve Guilty
Covid -ve Set Free	✓	Type 2 Error Misdetection (False -ve)
Covid +ve Jailed.	Type 1 error False Alarm x False +ve	✓

① One tailed test
② Two tailed test

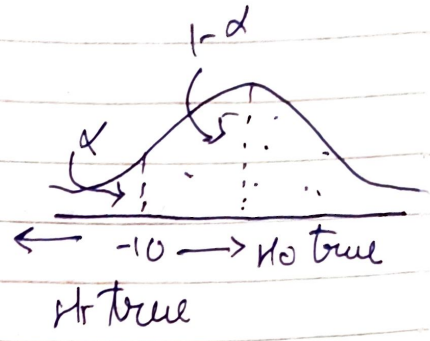


By comparing test statistics value with the critical value. Cobtained from table) we come to the conclusion whether to reject the Null Hypothesis. H_0 or we fail to reject the Null Hypothesis.

Left Tail test:

$H_0 : \theta \geq -10$

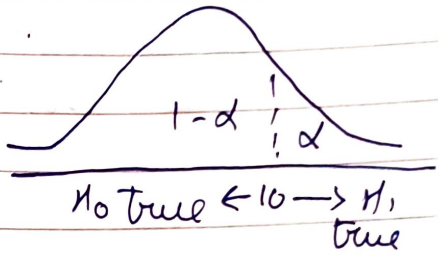
$H_1 : \theta < -10$



Right Tail test:

$H_0 : \theta \leq 10$

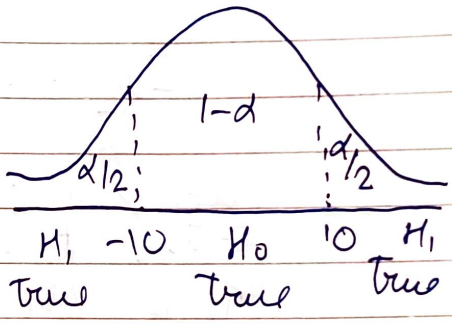
$H_1 : \theta > 10$



Two tailed test:

$H_0 : |\theta| \leq 10$

$H_1 : |\theta| > 10$



Manufacturer is manufacturing 1200 boxes and claims that average weight of a box is 1.84 kg.
Customer randomly chooses 64 boxes and find sample average weight as 1.88 kg. Considering that population standard deviation is 0.3 kg, use 95% confidence level and test for hypothesis that true average weight of shipment is 1.84 kg.

$N = 1200$ $\quad \mu = 1.84$ $\quad 1-\alpha = 0.95$

$n = 64$ $\quad \bar{x} = 1.88$ $\quad \sigma = 0.3$ $\quad \alpha = 0.05$

$$-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96$$

-1.69

$H_0: \mu = 1.84$ $\quad$ 0.0735

$H_1: \mu \neq 1.84$ $\quad$ $1.8065 < \mu < 1.9535$

$N = 1200$ $\quad$ $n = 64$

$\sigma = 0.3$ $\quad$ $\bar{x} = 1.88\ kg$

$\sigma^2 = 0.09$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$100(1-\alpha)\% = 95\%$

$\alpha = 0.05$ $\quad$ $\alpha/2 = 0.025$

$$Z = \frac{1.88 - 1.84}{0.3/\sqrt{64}} = 1.07$$

$Z_{\alpha/2} = 1.96$

We can see that Z lies between $-Z_{\alpha/2}$ to $Z_{\alpha/2}$

Thus, Null Hypothesis is true.

Decision failed to reject Null Hypothesis

Inconclusive evidence to suggest that avg. weight.

Q. Estimating the Difference in Mean of two Normal Distribution.

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a Normal Population having mean μ_1 and variance σ_1^2.

Let $Y_1, Y_2, \ldots, Y_m$ be a sample of size m from a Normal population having mean μ_2 and variance σ_2^2.

Point of concern: $\mu_1 - \mu_2$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad x_i \sim N(\mu_1, \sigma_1^2)$$

$$E(\bar{x}) = \mu_1 \qquad Var(\bar{x}) = \frac{\sigma_1^2}{n} \qquad \bar{x} \sim N(\mu_1, \sigma_1^2/n)$$

$$\bar{y} = \frac{Y_1 + Y_2 + \ldots + Y_m}{m} \qquad Y_i \sim N(\mu_2, \sigma_2^2)$$

$$E[\bar{y}] = \mu_2 \qquad Var(\bar{y}) = \frac{\sigma_2^2}{m} \qquad \bar{y} \sim N(\mu_2, \sigma_2^2/m)$$

$\mu_1 - \mu_2 \longrightarrow \bar{x} - \bar{y}$ (Point estimator)

→ Interval estimator?
= Point estimator $\pm E$
$= (\bar{x} - \bar{y}) \pm E$

$$\bar{x} - \bar{y} \sim N(E(\bar{x} - \bar{y}), Var(\bar{x} - \bar{y}))$$

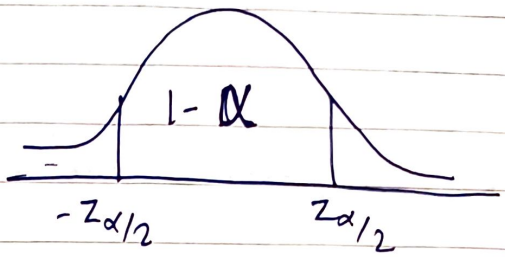
$$\underbrace{\bar{x} - \bar{y}}_{\theta} \sim N\left(\underbrace{\mu_1 - \mu_2}_{E(\theta)}, \underbrace{\sigma_1^2/n + \sigma_2^2/m}_{Var(\theta)}\right)$$

Test statistics:

$$Z_{test} = \frac{\theta - E(\theta)}{\sqrt{Var(\theta)}} = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$$

In exam, value of Z multiplier means the value of $Z_{\alpha/2}$.

$$P(-Z_{\alpha/2} < Z_{test} < Z_{\alpha/2}) = 1-\alpha.$$



$$P\left[-Z_{\alpha/2} < \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} < Z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[(\bar{x}-\bar{y}) - Z_{\alpha/2}\sqrt{\sigma_1^2/n + \sigma_2^2/m} < \mu_1-\mu_2 < (\bar{x}-\bar{y}) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right] = 1-\alpha$$

$100(1-\alpha)\%$ two sided confidence interval for the estimator of difference of mean of two distribution.

Trying to find interval estimator

LRM → Linear Regression Model

β_1 : Slope parameter $\quad y = \beta_0 + \beta_1 x$

$H_0 : \beta_1 = \beta_1^K$ [β_1^K : known constant value]

$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xxc})$

Test statistics:

$$z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}}$$

$$E(z_1) = \frac{1}{\sqrt{\sigma^2 / S_{xx}}} \left[E(\hat{\beta}_1) - E(\beta_1) \right] = 0$$

$$\text{Var}(z_1) = \frac{1}{\sqrt{\sigma^2 / S_{xx}}}$$

$$\text{Var}(z_1) = \frac{1}{\sigma^2 / (S_{xx})} \left(\text{Var}(\hat{\beta}_1) - \text{Var}(\beta_1) \right)$$

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

$$= 1$$

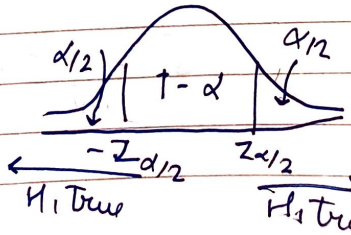
z_1 (test statistic) can be used to test hypothesis, significance of slope parameter.

H_1 : $\beta_1 \neq \beta_1^*$

Decision Rule : When to reject H_0?

ans: $z_1 > z_{\alpha/2}$ or $z_1 < -z_{\alpha/2}$

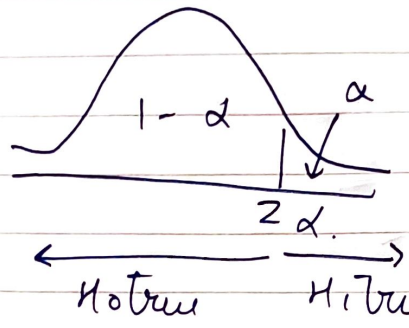
$|z_1| > z_{\alpha/2}$



$H_1 : \beta_1 > \beta_1^K$

Decision Rule:
Reject H_0

$z_1 > z\alpha$



$H_1 : \beta_1 < \beta_1^K$

Decision Rule:
Reject H_0

$z_1 < -z_\alpha$

Normal distribution
(μ, σ^2)
Population parameter

$x_i \sim N(\mu, \sigma^2)$

Sample is considered
$(\bar{x}, s^2)$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Population mean

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

population variance.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

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

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

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

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

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

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

$$\sigma^2 = E\{x_i^2\} - \mu^2$$

$$E\{x_i^2\} = \sigma^2 + \mu^2$$

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

$$\sigma^2/n = E(\bar{x}^2) - \mu^2.$$

$$E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2$$

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

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

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

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

$$\hat{\sigma}^2 = \left(\frac{n}{n-1}\right) s^2$$

If σ^2 is unknown, from above expression what is estimator of σ^2??

$$E\left(\left(\frac{n}{n-1}\right) s^2\right) = \sigma^2$$

Can I say s^2 is an unbias estimator of σ^2.

if s^2 was unbiased estimator for σ^2, we could have

$\hat{\sigma}^2 = s^2$ but here, its not True

So, s^2 is biased estimator.

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

$$E\{s^2\} = \sigma^2$$

(μ, σ^2)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Confidence interval (CI) $100[1-\alpha)\%$ for μ

$$\left[\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

LCL UCL

Interval estimator for population parameter μ when population variance σ^2 was known.

When σ^2 itself is unknown and you need to find Interval estimator for μ.

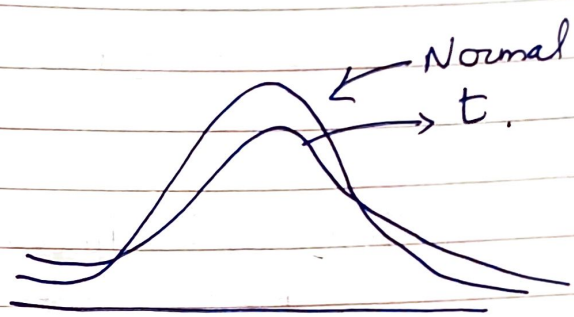
Test statistics:

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

freedom as you used s instead of σ.

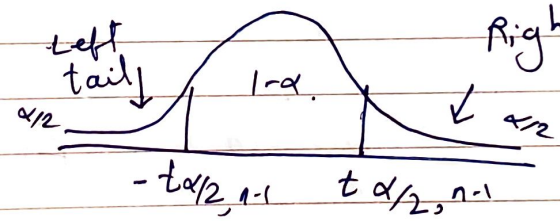
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

t-distribution:



Degree of freedom ↑ ⟶ t will closely match normal. (close margin)

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



Aim: To find interval estimator for unknown population mean parameter μ, when σ is unknown.

$100(1-\alpha)\%$ confidence, formulate the interval estimator for μ.

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

$$P\left[-t_{\alpha/2, n-1} < \frac{\bar{x} - \mu}{S/\sqrt{n}} < t_{\alpha/2, n-1} \right] = 1 - \alpha$$

$$P\left[\bar{x} - (t_{\alpha/2, n-1})\frac{S}{\sqrt{n}} < \mu < \bar{x} + (t_{\alpha/2, n-1})\frac{S}{\sqrt{2}} \right] = 1 - \alpha$$

$$E = \frac{S}{\sqrt{n}} t_{\alpha/2} \qquad S^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2 = \sigma^2$$

Intercept Parameters :

β_0 :

$H_0 : \beta_0 = \beta_0^c$

$H_1 : \beta_0 \neq \beta_0^c$

Interval estimate

$(1-\alpha)\,100\%$ CI

Test statistics $Z_0 = \dfrac{\hat{\beta}_0 - \beta_0^c}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}}$

standard normal.

$$E(Z_0) = \frac{1}{\theta}\left[E(\hat{\beta}_0) - E(\beta_0^c) \right]$$

$$= \frac{1}{\theta}\left[\beta_0^c - \beta_0^c \right]$$

$$= 0$$

$$\text{Var}(Z_0) = \frac{1}{\theta^2}\left(\text{Var}(\hat{\beta}_0) - \text{Var}(\beta_0^c) \right)$$

($\text{Var}(\beta_0^c) = 0$)

$$= \frac{1}{\theta^2} \cdot \theta^2 = 1$$

→ When σ^2 is KNOWN:

. For parameter β_1

$$CI: \left[\hat{\beta}_1 - Z_{\alpha/2}\sqrt{\sigma^2/S_{xxc}},\ \hat{\beta}_1 + Z_{\alpha/2}\sqrt{\sigma^2/S_{xxc}}\right]$$

. for parameter β_0.

$$CI: \left[\hat{\beta}_0 - Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xxc}}\right)},\ \hat{\beta}_0 + Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xxc}}\right)}\right]$$

→ When σ^2 is UNKNOWN (so we use t-distribution)

Test statistics $\quad t_1 = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/S_{xxc}}} \sim t(n-2)$

$H_0: \beta_1 = \beta_1^k$

Reject H_0 when H_1

We know that: $E\left[\dfrac{SS_{res}}{n-2}\right] = \sigma^2$

$\hat{\sigma}^2$

So, $\hat{\sigma}^2 = \dfrac{SS_{res}}{n-2}$

$\therefore \hat{\sigma} = \sqrt{\dfrac{SS_{res}}{n-2}}$

↳ derived earlier

→ Confidence CI: $\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}},\ \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]$

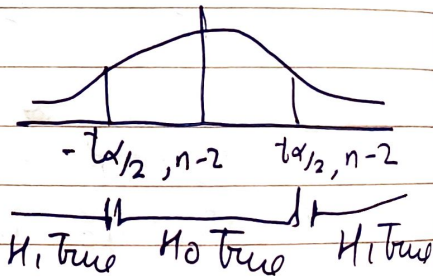
So for β_0:

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)$$

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-2}$$

$$CI: \left[\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)},\ \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right]$$

* $H_0: \beta_0 = \beta_0^c$
$H_1: \beta_0 \neq \beta_0^c$
Reject H_0 when $|t_0| \geq |t_{\alpha/2}|$



test statistic: $t_0 = \frac{\hat{\beta}_0 - \beta_0^c}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$

Chi-Square Distribution:

Consider 'n' dataset:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2$$

$\varepsilon_i \sim N(0, \sigma^2)$

↳ Not in standard Normal form

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

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$= \sum \left[(\beta_0 + \beta_1 x_i + \varepsilon_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

$$SS_{res} = \sum_{i=1}^{n} \left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \varepsilon_i \right]^2$$

$$\frac{SS_{res}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \varepsilon_i \right]^2$$

$$\frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sigma^2} = \frac{\varepsilon_1^2}{\sigma^2} + \frac{\varepsilon_2^2}{\sigma^2} + \dots + \frac{\varepsilon_n^2}{\sigma^2}$$

$\frac{\varepsilon_i}{\sigma} \sim N(0, 1)$ => follows chi sq. distribution with 'n' degree of freedom.

$$\frac{SS_{res}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \varepsilon_i \right]^2$$

2 degrees of freedom are lost as we are using estimated value $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\therefore \frac{SS_{res}}{\sigma^2} \sim \chi^2(n-2)$$

$H_1 : \lambda < \theta$

We reject H_0 if $P(x \le x_c) = \alpha/2$

$$E\left[\frac{SSres}{\sigma^2}\right] = n-2 \Rightarrow E\left[\frac{SSres}{n-2}\right] = \sigma^2$$

$$\therefore \hat{\sigma}^2 = \frac{SSres}{n-2}$$

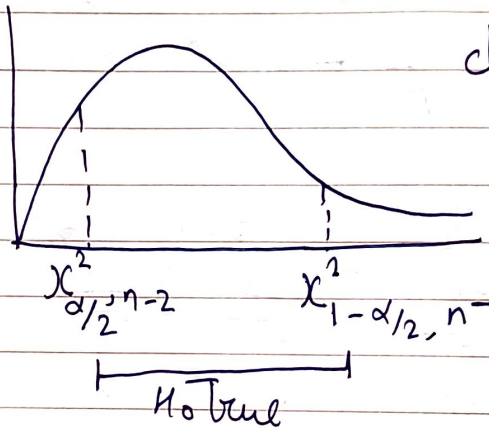
- Test of hypothesis for σ^2.

$H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 \neq \sigma_0^2$

$$\frac{SSres}{\sigma^2} \sim \chi^2(n-2)$$

Test Statistic : $C_0 = \frac{SSres}{\sigma^2} \sim \chi^2(n-2)$



chi-sq is Not Symmetric

Decision Rule:

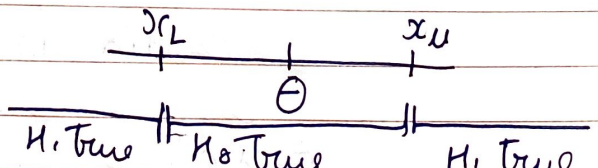
Reject H_0 if $C_0 < \chi^2_{\alpha/2, n-2}$ or $C_0 > \chi^2_{1-\alpha/2, n-2}$

* ~~Poisson~~ Poisson Distribution Hypothesis Testing:

↳ $f_x(x) = e^{-\lambda}\frac{\lambda^x}{x!}$ $\quad E[x] = \lambda$.

$H_0 : \lambda = \theta$

$H_1 : \lambda \neq \theta$



(with some energy margin)

We reject H_0 if $P(x \le x_c) = \alpha/2$ or $P(x \ge x_u) = \alpha/2$

$H_1 : \lambda > 0$ (We reject H_0 if $P(x \ge x_u) = \alpha/2$

Information Theory

- Information ⟶ Engineering concept

(1) T H _?_ _ => {E
(2) C _?_ T => {A, U, O}
(3) _?_ A T => {G, E, H, P, M, F, _ _ .}

(Guess list)

The Guess list keeps on increasing.

CO _?_?_ AL
OL
AT
LD

Randomness in my choosing is MORE.

COA _?_ T
L

Randomness further decreases.

Information is reduction of uncertainty (randomness)

Entropy / Information

CNN ~~dot~~
RNN Neural Network

Randomness ↑
information ↑
probability ↓

Modeling the Uncertainty :

Prime Numbers less than 10 => {2, 3, 5, 7}
Uncertainty = 4

C _?_ T => {A, U, O}
Uncertainty = 3

Measure of Uncertainty ??

↓

Must be modelled mathematically

"$\log_2(\ )$"

$\mathcal{X}$: set of possible answers.

$|\mathcal{X}|$: No. of possible answers

$\mathcal{X} = \{2, 3, 5, 7\}$

$|\mathcal{X}| = 4$

x_i : i^{th} answer

$P(x_i)$: Probability that the x_i^{th} answer is correct.

Information : $\boxed{I(x_i) = \log_2\left(\frac{1}{P(x_i)}\right)}$

$I(x_i) = \log_2\left(\frac{1}{P(x_i)}\right)$

(event) E_1 : The sun rises from East

E_2 : It will be snowfall in Ahmedabad in the month of May/June.

$P(E_1) = 1$

$P(E_2) \approx 0$

$$I(E_1) = \log_2\left(\frac{1}{P(E_1)}\right)$$
$$= \log_2 1$$
$$= 0$$

(No information)

$$I(E_2) = \log_2\left(\frac{1}{P(E_2)}\right)$$
$$= \log_2(\infty)$$
$$= \infty$$

(Highly informative)

Entropy ($\mathcal{X}$)

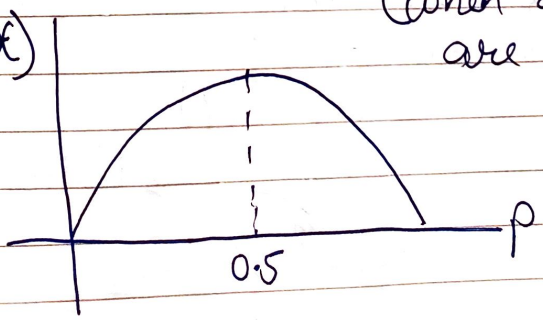
$$H(\mathcal{X}) = \sum_{i=1}^{K \to |\mathcal{X}|} -P(x_i) \log P(x_i)$$

$$H(\mathcal{X}) = \sum_{i=1}^{K} P(x_i) \cdot \log_2\left(\frac{1}{P(x_i)}\right)$$

=> Uncertainty / Guess-list may NOT be equally likely / Probable.

=> Uncertainty can ~~be~~ vary based on knowledge

(When all events are equally likely)



More Entropy

Most uncertainty

2nd years

↓

10 students are equally likely to get gold medal.

Prediction : Full of uncertainty.

2019

↓

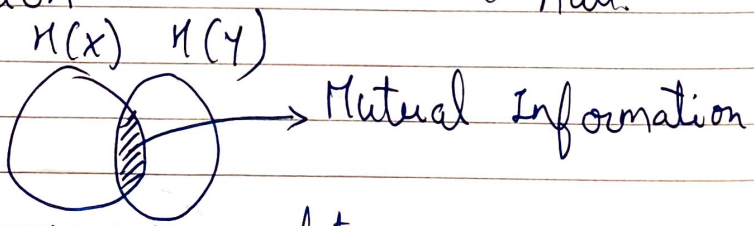
Uncertainty = 0

※ Entropy Venn Diagram

X : Thaltej (traffic) Y : traffic in Marine Drive

X and Y are independent, so H(X) has nothing to do with H(Y).

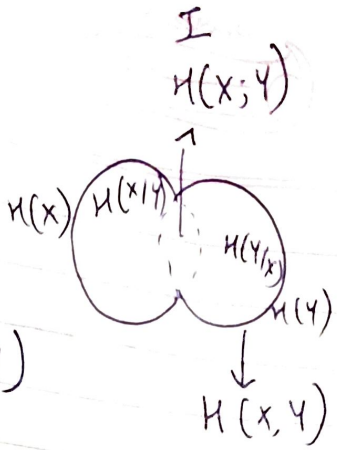
X : traffic in Thaltej Netrostation Y : Traffic in Palladium Mall.



There is some dependency between X and Y.

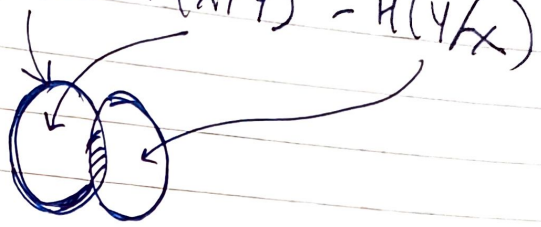
$I(X;Y)$

$$I(X) = H(X) - H(X/Y)$$
$$I(X;Y) = H(Y) - H(Y/X)$$



$$I(X) = H(X) + H(Y) - H(X,Y)$$
$$I(X;Y)$$

$$I(X) = H(X,Y) - H(X/Y) - H(Y/X)$$



- If we have message m_1, and m_2, independent with probability P_1 and P_2

Source → Encoding

msg/sys → BITS

$$P = P_1 \cdot P_2$$

$$I = \log_2\left(\frac{1}{P}\right) = \log_2\left(\frac{1}{P_1 \cdot P_2}\right)$$

$$I = \log_2\left(\frac{1}{P_1}\right) + \log_2\left(\frac{1}{P_2}\right)$$

$$I = I_1 + I_2$$

Decoding → Sink

- If there are 'm' messages from a source and all the messages are equally likely

$$m = 2^N = 16 \Rightarrow N = 4$$

$$p_i = \frac{1}{m} \qquad 1 \le i \le 16$$

$$H = \sum_{i=1}^{m=16} p_i \log_2 \left(\frac{1}{p_i}\right)$$

$$= \sum_{i=1}^{m} \frac{1}{m} \log_2 (m)$$

$$= \frac{m}{m} \log_2 (m)$$

$$= \log_2 (2^N)$$

$$= N$$

$$M = \{m_1, m_2, \ldots, m_{16}\}$$
$$\quad \downarrow 16 \quad \downarrow 16 \qquad\qquad \downarrow 16.$$

$$H = 4 \text{ bits / sym} \xrightarrow{\text{msg}}$$

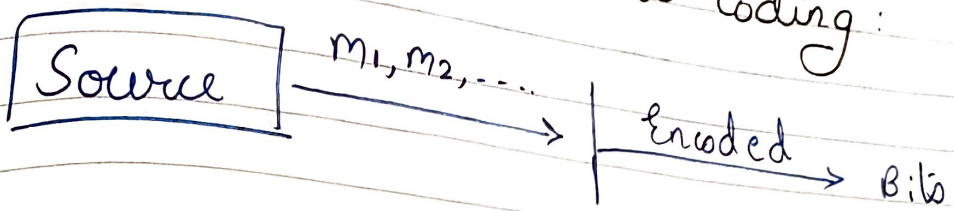
$$I_i = \log_2 \left(\frac{1}{p_i}\right) \qquad i = 1, \ldots, N$$

↓ Information

$$H = \underbrace{p_1 . I_1 + p_2 . I_2 + \ldots + p_N I_N}_{\text{Average Information}}$$

$$H = \sum_i p_i \log_2 \left(\frac{1}{p_i}\right) = -\sum_i p_i \log_2 p_i$$

Encoding Techniques / Source Coding:



Shannon - Fano Encoding:

- It is an entropy encoding technique for data compression.
- It assigns a code to each symbol/message based on their probability.
- Variable length coding scheme.
- The source symbols are split into two subsets "0" and "1" with probability as close to one-half. Then each of the sub-sets is split into two and so on until there is only one symbol / message left in the set.

Example:

Find the codewords

Example:

Source $S = \{S_1, S_2, S_3, S_4\}$

Probability of occurence of each source message $P = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$

msg

Message	Prob.	1st stage	2nd stage	3rd stage	Codeword
S_1	1/2	0			0
S_2	1/4	1	0		10
S_3	1/8	1	1	0	110
S_4	1/8	1	1	1	111

$$H = \sum_i P_i \log_2 \frac{1}{P_i}$$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8}$$

$$= 1/2 + 1/2 + 3/4$$

$$= 7/4$$

Avg. Codeword length: $= \sum_{i=1}^{4} P_i l_i$

$\hat{H} \; |L| \bar{L}$

$$= P_1 l_1 + P_2 l_2 + P_3 l_3 + P_4 l_4$$

$$= \left(\frac{1}{2} \times 1\right) + \left(\frac{1}{4} \times 2\right) + \left(\frac{1}{8} \times 3\right)2$$

$$= 7/4$$

$$\eta\,[\text{Efficiency}] = \frac{\text{Entropy}}{\text{Avg. Codeword length}}$$

$$= \frac{7/4}{7/4}$$

$$= 1$$

Redundancy [Re] = $1 - \eta = 0$.

Efficiency = 100 % when both the subsets are equally probable in the 1st, 2nd, 3rd stage.

$\eta < 1$ when subset probability are not equal.

$\eta_{SES} < \eta_{HES}$

η_{SES} ↓ shannon encoding scheme

η_{HES} ↓ Hoffmann encoding scheme

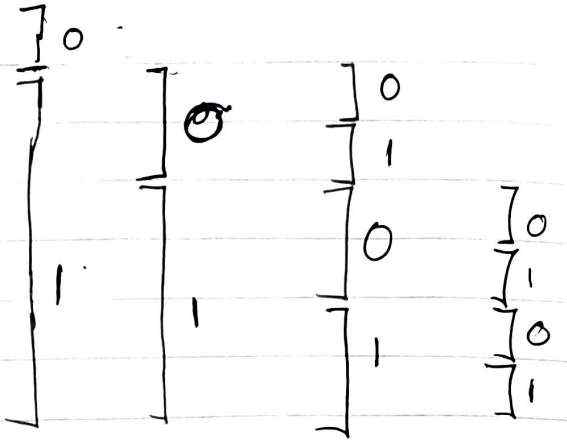
Example:

$M = \{m_1, m_2, \ldots, m_7\}$

$P = \{0.4, 0.2, 0.12, 0.08, 0.08, 0.08, 0.04\}$

By shannon's encoding scheme,

message	Prob	Codeword
m_1	0.4	0
m_2	0.2	100
m_3	0.12	101
m_4	0.08	1100
m_5	0.08	1101
m_6	0.08	1110
m_7	0.04	1111



$$H = 0.4 \log_2\left(\frac{1}{0.4}\right) + \ldots\ldots + 0.04 \log_2\left(\frac{1}{0.04}\right)$$

$$\bar{L} = 0.4(1) + \cancel{0.6}\, 0.2 \times 3 + \ldots$$

$$\eta = \frac{H}{\bar{L}} < 1$$

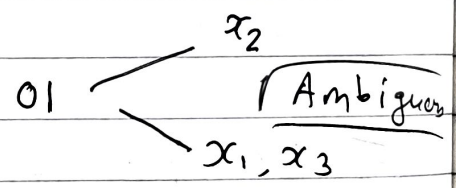
↓

There will be redundancy in SEC.

$X = \{x_1, x_2, x_3, x_4\}$

↓ ↓ ↓ ↓

0 01 1 10



→ Shannon Encoding scheme is prefix-free code

Kraft's Inequality / Theorem: $\sum_{i=1}^{n} 2^{-l_i} \leq 1$

If this condition is satisfied, we obtain codes which are prefix-free code.

Any encoding scheme that does not have prefix subset, is called prefix free code.

Q. Refer shannon example 1st question:

$\{S_1, S_2, S_3, S_4\}$.

$l_1 = 1 \quad l_2 = 2 \quad l_3 = 3 \quad l_4 = 3$

$$\sum_{i=1}^{4} 2^{-l_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3}$$
$$= 2^{-1} + 2^{-2} + 2^{-2}$$
$$= 2^{-1} + 2^{-1}$$
$$= 1$$

Huffman Encoding:

(Shannon - Fano)

↓

Robert Fano
Prof. at MIT.

Amnon Shashua
Israeli

↓

Prof. of
Hebrew University
Jerusalem.

Intel

Huffmann Coding Scheme:

① The source symbols are arranged in order of decreasing probability. Then the two of lowest probabilities are assigned bit 0 and 1.

② Group the last 2 symbols/messages and encode them and move the grouped symbol as high as possible.

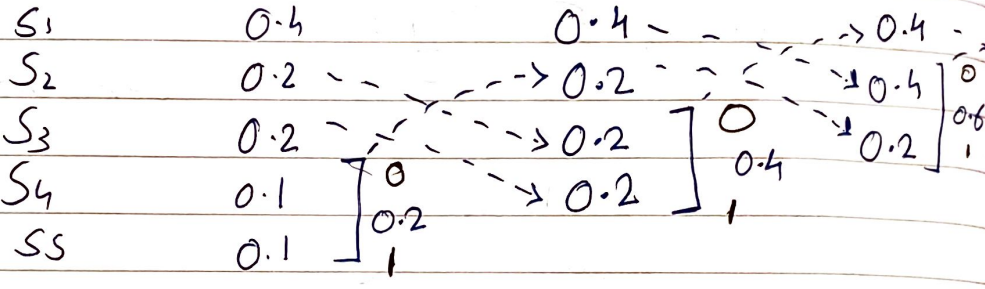
③ Repeat the step (2) until only 2 symbols are left.

④ Codeword for each symbol is found by moving in reverse direction.

Q. $S = \{S_1, S_2, S_3, S_4, S_5\}$

$P = \{0.4, 0.2, 0.2, 0.1, 0.1\}$

Symbol	Prob.
S_1	0.4
S_2	0.2
S_3	0.2
S_4	0.1
S_5	0.1



S_1	00	$l_1 = 2$
S_2	10	$l_2 = 2$
S_3	11	$l_3 = 2$
S_4	010	$l_4 = 3$
S_5	011	$l_5 = 3$

$3 \times 2^{-2} + 2 \times 2^{-3}$

$\frac{3}{4} + \frac{1}{4}$

1

prefix free code.

$$H = \sum_{i=1}^{5} P_i \log_2 \left(\frac{1}{P_i}\right)$$

$$\bar{L} = \sum_{i=1}^{5} P_i l_i$$

$= 0$

$$\eta = \frac{H}{\bar{L}} \qquad Re = 1 - \eta$$

efficiency (Hoffmann) better than efficiency (Shannon)

$$\eta = \frac{H}{\bar{L} \log_2 M}$$

$M = 2$

$\log_2 2 = 1$ (Binary)