Types of Variables

Qualitative (categorical) ⇒ Grouped data

= express a qualitative atribute such as hairclr, religion, marital status

Nominal
No ordering/ranking is possible

Eg:⇒ hairclr, eyeclr etc

Ordinal
ordering is possible

Eg:⇒ health can be classified as good/Avg/bad

Quantitative (numerical data)

measured in terms of no's such as weight, height

ISM important topics

Discrete
No: without fractions

countable & have finite no. of possibilities such as no. of people

Continuous
No. with fractions

cant be countable & has ∞ no. of s/ns

Eg:⇒ height

Interval
ratio of values of variable do not have any meaning & it doesnot have inherently defined Zero value such as temp.

Ratio
ratio of values of variable have meaning & it have an inherently defined zero val

Eg:⇒ length

① Measure of central tendency

↳ Mean, median, mode

* mean :→ $\bar{y} = \frac{\Sigma fy}{N}$

Mean = $\frac{\text{sum of values}}{\text{tot no. of values}}$

* Mode :- Most repeated no. of times

one mode => unimode

multiple modes => multimode.

mode = 3median − 2mean.

*median :-

↳ segrigrate the given no.'s & find middle value.

↳ if odd no. of values median = middle value

↳ if even no. of values median = Avg of 2 middle values

② Measure of variability

↳ Range, standard deviation, varience

* RANGE :-

distance covered by scores in a distribution.

↳ From smallest to highest value

RANGE = URL x_{max} − LRL x_{min}

Eg :=> [7, 2, 7, 6, 5, 6, 2] ; RANGE hig-low = 7−2 = 5

*Standard deviation:-

↳ Before going to standard deviation lets first understand whats deviation✓

Given data

↳ DAY	1	2	3	4	5	6	7	8	9	10
TIME	39	29	43	52	39	44	40	31	44	35

$\bar{X}$ = mean = $\frac{\text{sum of val}}{\text{tot valy}} = \frac{39+29+43+\cdots+35}{10} = \frac{396}{10} = 39.6$

↳ Now if we substract each of these values (x) with Arthematic mean ($\bar{x}$); that deviation tell about deviation from each of these values from Arthematic mean

① $x-\bar{x} \Rightarrow 39-39.6 = -0.6$

② $x-\bar{x} = 29-39.6 = -10.6$

③ $43-39.6 =$

④ $52-39.6 =$

⑤ $39-39.6 =$

⑥ $44-39.6 =$

⑦ $40-39.6 =$

⑧ $31-39.6 =$

⑨ $44-39.6 =$

⑩ $35-39.6 =$

$\Sigma x-\bar{x} = 0 \Rightarrow$ if we do sum of all deviations $= 0$

*** ↳ It is mandatory that the Algebraic sum of the deviations from Arthematic mean sthd be always "Zero".

Now as sum of deviations is always zero; How to find standard deviation??

Standard deviation:-

3 steps:-

↳ Find the deviation of each score & square each deviation score

↳ Sum of squared deviations.

↳ Avg the squared deviations.

$$\text{Variance} = \frac{\Sigma (x-\mu)^2}{n}$$

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

Eg:- {1, 5, 7, 8, 9}

$$\mu = \text{mean} = \frac{[1+5+7+8+9]}{5} = 6$$

$$\Sigma(x-\mu)^2 = (1-6)^2+(5-6)^2+(7-6)^2+(8-6)^2+(9-6)^2 = 40$$

$$\text{Variance} = \frac{\Sigma(x-\mu)^2}{n} = \frac{40}{5} = 8$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{8}$$ ✓

Standard deviation

- Population standard deviation
- Sample standard deviation.

(i) Population standard deviation:- [entire data]

$$\text{varience} = \frac{\Sigma (x-\mu)^2}{n} \quad [\text{if } n \geq 30]$$

(ii) Sample standard deviation:-

$$\text{varience} = \frac{\Sigma (x-\mu)^2}{n-1} \quad [\text{if } n < 30]$$

↳ More concepts :-

$Q_1 = \left[\frac{n+1}{4}\right]^{th}$ term value — Q_1 = first Quartile => 25%.

$Q_3 = 3\left[\frac{n+1}{4}\right]^{th}$ term value — Q_3 = third Quartile => 75%.

IQR :- [inter Quartile RANGE]:-

↳ measure of varience

↳ Diff b/w 3rd & 1st Quartile

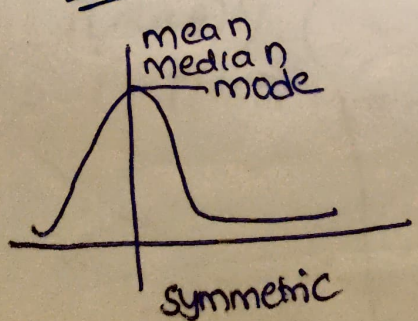
$IQR \Rightarrow Q_3 - Q_1$

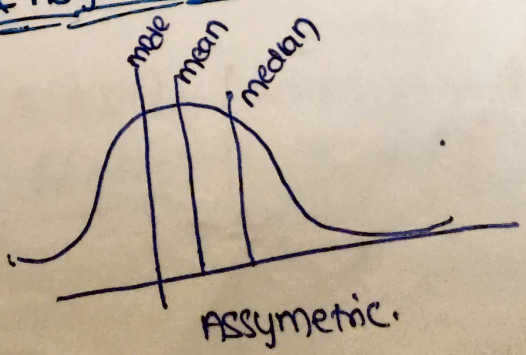
=> Five Point summary :-

To find five point summary; we need to find

i) ~~mean~~ minimum ii) median (iii) ~~mode~~ maximum (iv) Q_1 (v) Q_3

=> DATA :=> Symmetrical & Asymetrical:-





DATA distribution :-[skewed]

Negatively [left] : mean < median

Positively [Right] : mean > median

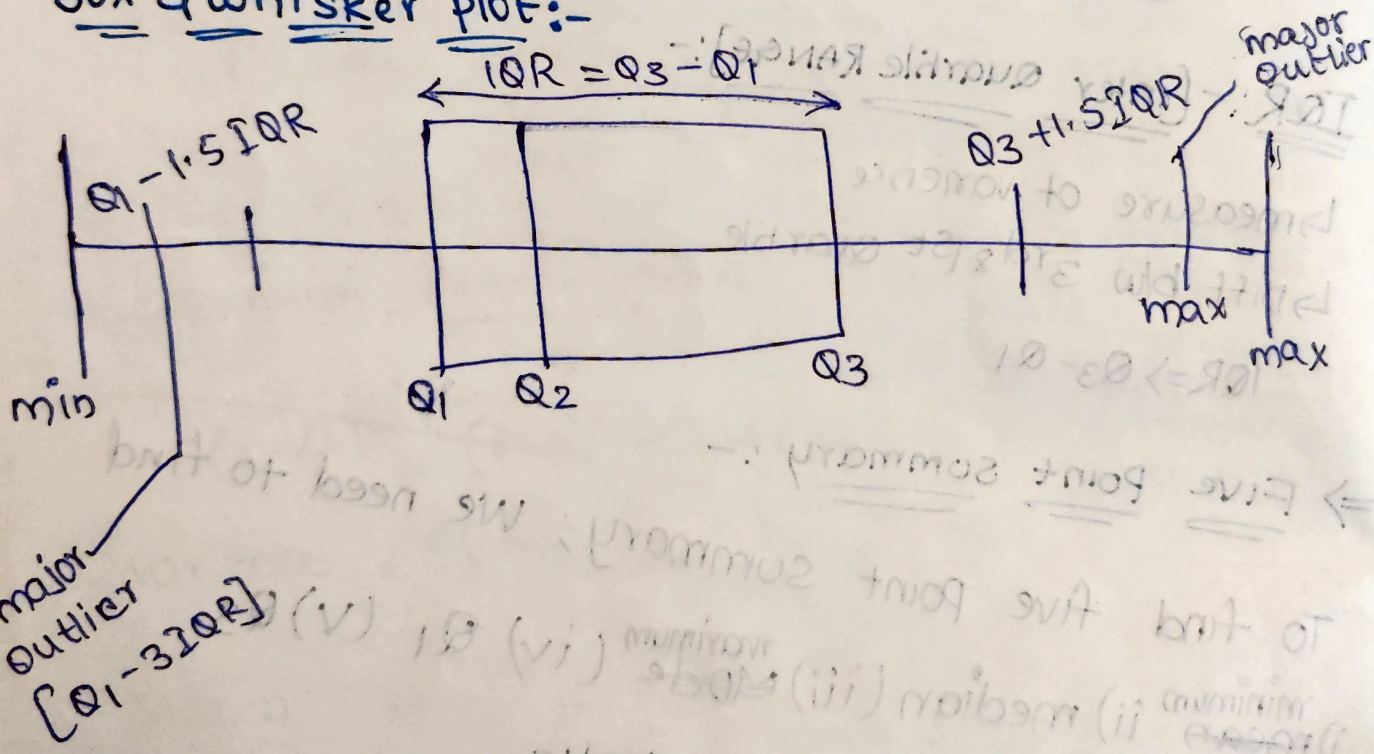
Potential outliners :-

↳ The lower limit & upper limit of data set are given by

lower limit = $Q_1 - 1.5IQR$

upper limit = $Q_3 + 1.5 * IQR$

Box & whisker plot :-



Basic Probability:-

⇒ Random Experiment:-

Random Experiment is used to describe any action whose outcome is not known ~~as~~ in Advance

Eg:- i):- Give a lecture, How many r listening
ii):- Pulling card from deck.

⇒ Sample Space:-

The Sample space of a random experiment is set 'S'. that includes all possible outcomes of experiment.

⇒ Event:-

An Event is a set of outcomes of experiment. this Includes the null (empty) set of Outcomes & set of all Outcomes

Eg:- dice ⇒ Samplespace = {1,2,3,4,5,6}
Event = {even}, {odd}, {prime} ---

A&B are mutually Exclusive Events.

⇒ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Mutually Exclusive:-

$A \cap B = \emptyset$ ⇒ nullset; $P(A \cap B) = 0$

$P(A \cup B) = P(A) + P(B)$

$P(A) = 1 - P(\bar{A})$

⇒ Independent Events:-

We say both A & B are Independent if:

$$P(A \cap B) = P(A) \cdot P(B)$$

⇒ Conditional Probability

The prob of event happening given that another event has already happened

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

⇒ Multiplication Rule:-

let A, B, C are 3 events in Sample Space 'S' then:

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$$

$$P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

⇒ Total Probability :-

The law of tot. Prob.

$$P(B) = \sum_{i=1}^{K} P(B|A_i) \cdot P(A_i)$$

=> Bayer's therom :-

It will help us to find the Prob of something, given that we know something else is happened.

$$P(E_i/A) = \frac{P(E_i) \cdot P(A/E_i)}{\sum_{i=1}^{n} P(E_i) \cdot P(A/E_i)}$$

tot prob $P(A)$

Naive Bayer's therom :-

$$P(x_1, \dots x_n / y) = \prod_{i=1}^{n} P(x_i/y)$$

It works Based on Bayers therom & assumes all features are Independent to each other.

$$P(x_1, x_2/y) = P(x_1/y) \cdot P(x_2/y)$$

=> It is Basically used for Prblems like

S_1/S_2 ; Yes/NO ; Spam/Not Spam; Disease/NO disease

$$P(C/x_1, x_2 \dots x_n) \propto P(C) \cdot P(x_1/C)\, P(x_2/C) \dots P(x_n/C)$$

Eg:-

NO	Clr	legs	Hgt	smelly	species
1	white	3	short	Yes	S_1
2	Green	2	tall	NO	S_1
3	Green	3	short	yes	S_1
4	white	3	short	yes	S_1
5	Green	2	short	NO	S_2
6	white	2	tall	NO	S_2
7	white	2	tall	NO	S_2
8	white	2	short	yes.	S_2

i) find conditional Prob of each attribute {clr, leg, hgt, smelly} for species $\{S_1, S_2\}$

$P(S_1) = 4 \quad P(S_2) = 4$

$P(\text{color} = \text{Green} / S_1) = \frac{2}{4}$

$P(\text{white} / S_1) = 2/4$

$P(\text{legs} = 2 / S_1) = 1/4$

$P(\text{legs} = 3 / S_1) = 3/4$

$P(\text{Hgt} = \text{tall} / S_1) = 1/4$

$P(\text{Hgt} = \text{Short} / S_1) = 3/4$

$P(\text{smelly} \rightarrow \text{yes} / S_1) = 3/4$

$P(\text{Smelly} \rightarrow \text{NO} / S_1) = 1/4$

$P(\text{colour} = \text{Green} / S_2) = \frac{1}{4}$

$P(\text{white} / S_2) = 3/4$

$P(\text{legs} = 2 / S_2) = 4/4$

$P(\text{legs} = 3 / S_2) = 0/4$

$P(\text{Hgt} = \text{short} / S_2) = 2/4$

$P(\text{Hgt} = \text{tall} / S_2) = 2/4$

$P(\text{Smelly} - \text{NO} / S_2) - 3/4$

$P(\text{Smelly} - \text{yes} / S_2) - 1/4$

ii) using Prob Estimate which Instance does this belong to:?

$\{C = \text{Green}, \text{legs} = 2, \text{Hegt} = \text{tall}, \text{smelly} = \text{NO}\}$

$$P(S_1/x) = \frac{P(x/S_1) \cdot P(S_1)}{P(x)} \; ; \; P(S_2/x) = \frac{P(x/S_2) \cdot P(S_2)}{P(x)}$$

denominators r = so compare numera

$$P\left(\frac{S_1}{x}\right) = P\left(\frac{\text{Green}}{S_1}\right) \cdot P\left(\frac{2}{S_1}\right) \cdot P\left(\frac{\text{tall}}{S_1}\right) \cdot P\left(\frac{\text{NO}}{S_1}\right) \cdot P(S_1)$$

$$P\left(\frac{S_2}{x}\right) = P\left(\frac{\text{Green}}{S_2}\right) \cdot P\left(\frac{2}{S_2}\right) \cdot P\left(\frac{\text{tall}}{S_2}\right) \cdot P\left(\frac{\text{NO}}{S_2}\right) \cdot P(S_2)$$

$$P\left(\frac{S_1}{x}\right) > P\left(\frac{S_2}{x}\right)$$

So S_1 is the Instance it belongs to.

Laplace Smoothing:-

As we saw in naive Bayes theorem $P(A/x)$ is calculated as $P(A_1, A_2, \ldots An/x)$ so even if 1 Prob is 0 entire thing comes out as Zero.

=> So to Avoid this we add 1 to every count, including unseen event

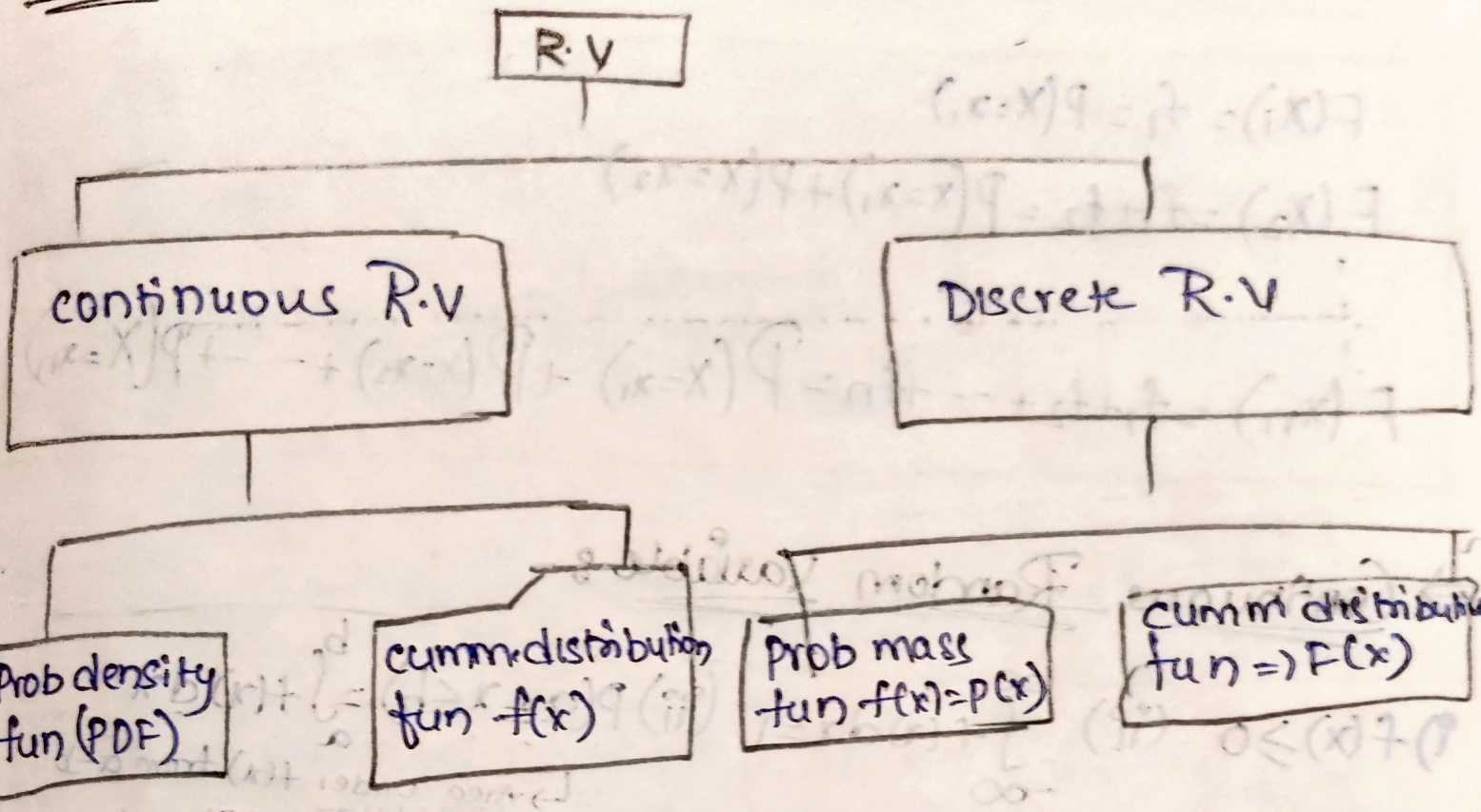
$$P(w/c) = \frac{count(w,c)+1}{\sum count(w_i,c)+y}$$

$y \rightarrow$ no. of unique words

Eg:- $P(A\ very\ close\ Game/sports) = ?$

text	TAG
"A Great Game"	Sport
the election was over	NSport
"A Clean but unforgettable Game"	Sport
"It was close election	NSport

Random variable :-



1) Discrete R.V

$$\Sigma f(x) = 1 \; ; \; f(x) \geq 0$$

⇒ Mean / Expected value of Discrete R.V :-

$$E(x) = \mu = \Sigma x \cdot f(x)$$

⇒ Variance of Discrete R.V is :-

$$Var(x) = \sigma^2 = E(x-\mu)^2 = \Sigma (x-\mu)^2 \cdot f(x)$$

(or)

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

$$E(x)^2 = \Sigma x^2 \cdot f(x)$$

⇒ Cummilative prob distribution funi (CDF) of Discrete R.V

$$F(x_1) = f_1 = P(X=x_1)$$

$$F(x_2) = f_1 + f_2 = P(X=x_1) + P(X=x_2)$$

$$\vdots$$

$$F(x_n) = f_1 + f_2 + \cdots + f_n = P(X=x_1) + P(X=x_2) + \cdots + P(X=x_n)$$

$P(X=x_2) \hookrightarrow = 1$

⇒ Continuous Random Variable :-

(i) $f(x) \geq 0$ (ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (iii) $P(a \leq x \leq b) = \int_a^b f(x)\,dx$

↳ Area under $f(x)$ from a–b for any a & b

↳ Basic formulaes of ∫ & derivatives by sir:-

Integrations :-

i) $\int_a^b x^n\,dx = \left|\frac{x^{n+1}}{n+1}\right|_a^b = \frac{1}{n+1}\left[b^{n+1} - a^{n+1}\right]$

ii) $\int_\alpha^\beta e^{ax}\,dx = \left|\frac{e^{ax}}{a}\right|_\alpha^\beta$

(iii) $\int a\,dx = ax + c$ ↳ const.

derivatives

i) $\frac{d}{dx}(x^n) = n \cdot x^{n-1}$

ii) $\frac{d}{dx}(e^{ax}) = a \cdot e^{ax}$.

⇒ Mean & varience of continuous R.V :-

i) $\mu = E(x) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \Rightarrow$ Mean

ii) $V(x) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2 \Rightarrow$ varience

iii) Standard deviation $= \sqrt{\text{varience}}$

$e = 2.718$

Joint Prob distribution :-

Let $x = \{x_1, x_2, \ldots x_n\}$ & $y = \{y_1, y_2, \ldots, y_n\}$ be a discrete R.V

then $P(x,y) = J_{ij}$ is called Joint Prop distribution of x & y

If it satisfies condition

i) $J_{ij} \geq 0$ (ii) $\sum_{i=1}^{m} \sum_{j=1}^{n} J_{ij} = 1$

	y_1	y_2	- - - - - - -	y_n	
x_1	J_{11}	J_{12}	- - - -	J_{1n}	$f(x_1)$
x_2	J_{21}	J_{22}	- - - -	J_{2n}	$f(x_2)$
⋮	⋮	⋮	- - - -	⋮	⋮
x_m	J_{m1}	J_{m2}	- - - -	J_{mn}	$f(x_m)$
Sum	$g(y_1)$	$g(y_2)$	- - - -	$g(y_n)$	sum = 1

marginal Prob.

Note: Marginal distribution of x & y alone are:

* Discrete $\Rightarrow g(x) = \sum_y f(x,y)$ & $h(y) = \sum_x f(x,y)$
* Continuous $\Rightarrow g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ & $h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$

Conditional prob of distribution func:-

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \qquad P(B/A) = \frac{P(B \cap A)}{P(A)}$$

if x,y is Discrete R.V the condition prob distribution fun:-

$$P(X=x, Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

if x,y is continuous R.V then:

$$f_{x|y} = \frac{f_{x,y}(x,y)}{f_Y(y)}$$

$f_{x,y}(x,y) \Rightarrow$ Joint PDF of x,y

$f_y(y) \Rightarrow$ Marginal PDF of y

Bernoulli distribution:-

Success (1) with Prob $\Rightarrow P$

Success (0) " " $\Rightarrow (1-P) = q$

mean $= P$

variance $= pq$

PMF :- $P(x=x) = \begin{cases} p^x q^{1-x} & \forall x = 1,0 \\ 0 & \text{elsewhere} \end{cases}$

Binomial distribution :-

Binomial distribution will be applied from flowing experimental conditions

⇒ No. of trails (n) is ⇒ finite.

⇒ Trails are Independent to each other

⇒ Prob of success P is constant of each trail

⇒ Each trail has 2 mutually exclusive events success/fail

Mean $= E(x) = np$	$(1-p) = q$	④ PMF $= P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$ $= nC_x p^x q^{n-x}$
Varience $= \sigma^2 = npq$		$x \Rightarrow$ No. of success in n trails
SD $= \sigma = \sqrt{npq}$		$n \Rightarrow$ tot trails

Poisson Distribution :-

Poisson distribution is suitable for rare events for which Prob of occurence 'p' is very small & no. of trails n is very large

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

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

$\mu = \lambda$ in Poission

Normal Distribution

$$E(X) = \mu = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Var}(x) = \sigma^2 = \left(\int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx\right) - \mu^2$$

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

$z =$

---

1) neither nor $\Rightarrow (A \cup B)^c$

mean, SD

2) or/and $\Rightarrow A \cup B$ And $\Rightarrow A \cap B$ Or $\Rightarrow A \cup B$

$\left(\frac{n+1}{4}\right)^{th}$ term

$$\text{mr } x = \int_{LL}^{UL} f(x)\,dy; \quad y = \int_{LL}^{UL} f(y)\,dx$$

Σ

Naive Bayers

Y/N $\Rightarrow$ sunny kinda Qsn

$P(Y) =$? $P(N) =$?

$P(Y/S) =$? $P(N/S)$

$$P(Y/S) = \frac{P(S/Y) \cdot P(Y)}{P(S)}$$

In Naive Bayers

$$P(N/S) = \frac{P(S/N) \cdot P(N)}{P(S)}$$

$P(S/Y) = \frac{P(S \cap Y)}{P(Y)}$ = find & substitute in above.

1) Central limit therom (CLT).

(i) $Z\ score = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ $\quad n > 30$

Avg mean $= \mu$
Prob mean $= \bar{x}$
σ = standard dev
n = tot values

(ii) $Z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}\left[\sqrt{\frac{N-n}{N-1}}\right]}$

⇒ we use (ii) when N is given & $n/N \leq > 0.5$

⇒ if $n/N < 0.5$ ⇒ we use (i) only N ⇒ not needed in this case.

2) Sample distribution of Sample Proportion (%):-

⇒ In such questions % will be given (or) directly they ask sample proportion

(i) $P = x/n$

x ⇒ sample workers (category).
n ⇒ tot workers

$$Z = \frac{\hat{P} - P}{\sqrt{\frac{Pq}{n}}}$$

$P \Rightarrow$ Population Prop $= \frac{x}{n}$.

$\hat{P} \Rightarrow$ sample proportion

$q \Rightarrow 1 - P$

$n =$ Sample size.

⇒ Statistical Inference

- Point estimation
- interval estimation
- Hypothesis estimation

⇒ Standard deviation of Sample Prop!-

finite population	Infinite Population
$\sigma_{\bar{P}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$	$\sigma_{\bar{P}} = \sqrt{\frac{P(1-P)}{n}}$

Confidence Interval

* if confidence Interval = 95
  then $(1-\alpha) = 95$;
  $\alpha = 5\%$.

This is based on N.

$$\left[\bar{x} \pm z_{\alpha/2}\left[\frac{\sigma}{\sqrt{n}}\right]\right] \Big/ \left[\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\left[\sqrt{\frac{N-n}{N-1}}\right]\right]$$

BITS - WILP

AIML

ISM ASSIGNMENT

∴ **confidence Interval** is

$$\bar{x} - z\alpha/2\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + z\alpha/2\left(\frac{\sigma}{\sqrt{n}}\right)$$

↳ Same way for other too.

⇒ In Qsn if they give conf Interval = 95%.

what is α ⇒ $1 - \alpha = 95\%$; so $\alpha = 5\%$.

If $\alpha = 1\%$.	If $\alpha = 5\%$. ⇒ 0.5	If $\alpha = 10\%$.
$z\alpha/2 = 2.58$	$z\alpha/2 \Rightarrow 1.96$	$z\alpha/2 = 1.645$

$$\bar{x} - z\alpha/2\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$ ⟹ for 2 means & 2 samples

if Sp is given
↳ Pooled Sp

then.

$$\bar{x} - z\alpha/2\left[SE(\bar{x}_1 - \bar{x}_2)\right]$$

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (\text{or}) \quad \sqrt{s_p\left(1/n_1 + 1/n_2\right)}$$

*

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z\alpha/2\left[SE(\bar{x}_1 - \bar{x}_2)\right].$$

Hypothesis

⇒

- **NULL**
  - ⇒ if $\mu_1 = \mu_2$
  - ⇒ No diff
  - ⇒ Unbiased / Neutral
- **Alternative**
  - ⇒ if $\mu_1 \neq \mu_2$ / $\mu_1 > \mu_2$ / μ_1
  - ⇒ diff → yes
  - ⇒ Biased.

Level of Significance (α)

	1%	5%	10%
two tailed test ($H_a \neq$)	$\lvert z_{\alpha/2} \rvert = \pm 2.58$	$\lvert z_{\alpha/2} \rvert = \pm 1.96$	$\lvert z_{\alpha/2} \rvert = \pm 1.645$
Right-tailed test ($H_a >$)	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
left-tailed test ($H_a <$)	$z_\alpha = -2.33$	$z_\alpha = -1.65$	$z_\alpha = -1.28$

⇒ **IN Right tailed**

if $z > z_\alpha$
└ reject

in left tailed

if $z < z_\alpha$
└ reject

2-tailed

if $-z_{\alpha/2} \leq z \leq +z_{\alpha/2}$ ⇒ Accept
└ else reject

Test significance for proportion: -

$$Z_p = \frac{p - P}{\sigma_p}$$

$$\sigma_p = \sqrt{\frac{PQ}{n}} \sqrt{\frac{N-n}{N-1}}$$

Z score
$n > 30$

If N is unknown $\Rightarrow \sigma_p = \sqrt{\frac{PQ}{n}}$

If P is unknown $\Rightarrow \sigma_p = \sqrt{\frac{pq}{n}}$

$p \Rightarrow$ Prop of survival $\Rightarrow x/n$

$q = (1-p)$

$P \Rightarrow$ Given Prob Proportion

$Q \Rightarrow 1 - P$

$n \Rightarrow$ sample size

Difference of Proportion $\Rightarrow$ Test significance

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \Rightarrow \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$n < 30$
$\rightarrow$ T score

T-test ($n < 30$)

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

T-test

diff b/w mean of 2 population :-

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} \quad (3)$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$s_p \Rightarrow$ Pooled SD

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

$n < 30$

μ isn't.

If μ isn't given then

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Steps

1) State H_0 & H_a
2) α
3) t-distribution
4) compute t-score
5) reject/accept H_0

difference b/w mean of Paired observations :-

└ befor & after

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

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

chisquare-test

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$E_{ij} = \frac{r_i c_j}{n}$$

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

In chi-square => If calculated χ^2 > critical χ^2

⟹ Reject NULL Hypothesis

df & α → see table find critical χ^2.

If α isn't given => $\alpha = 0.05$

ANOVA :-

1) one-way ANOVA

&

2) two-way ANOVA

ONE WAY ANOVA

(Parametric test)

Step1

Sum of all values $\Rightarrow G$

$CF = G^2/n$

Step2

Sum of sq of each element $\Rightarrow \Sigma\Sigma x_{ij}^2$

Step3 $\Sigma\Sigma x_i^2 - CF = SST$

Step4 :- $SSTR = \frac{\Sigma C_i^2}{n} \Rightarrow$ each clmn sq, tot

Step5 $SSTR = SST - SSTR$

$\alpha = 0.05$ If not given

Eg:-

	Grp_1	Grp_2	Grp_3	Grp_4
1	a	b	c	d
2	e	f	G	H
3	I	J	K	L
4	m	N	O	P

Step1 :-

$a+b+c+d - - - + P = G$

$CF = \frac{G^2}{n}$

step2

$\Sigma x_{ij}^2 = a^2+b^2+c^2+d^2+ - - - +P^2$

step3

$SST = \Sigma x_{ij}^2 - CF$

step4

$\frac{\Sigma C_i^2}{n} = \frac{(a^2+e+i+m)^2}{4} + - - + \frac{(d+H+L+P)^2}{4}$

Step 5 -

$$SSTR = \frac{\Sigma Ci^2}{n} - CF.$$

$K-1 \Rightarrow df1$
$n-K \Rightarrow df2$
} in Anova table

& $SSE = SST - SSTR$

Anova table [Step 6]

Source of variance	df	sum of squares	mean square	F ratio
B/w groups	$K-1$	SSTR	$\frac{SSTR}{K-1} = MSTR$	
within groups	$n-K$	SSE	$\frac{SSE}{n-K} = MSE$	$\frac{MSTR}{MSE} = F$
tot.	$n-1$	(SSTR+SSE)		$F(\alpha, K-1, n-K)$

$n \Rightarrow$ total values from a to p

$K \Rightarrow$ No. of groups.

Conclusion if calculated (F) > table value → H_0: Reject

2-Way ANOVA:-

	Grp1	Grp2	G3	G4
1	a	b	c	d
2	e	f	g	h
3	i	J	K	L

Step1

$$G = a+b+c+d+\cdots+L$$

$$\frac{G^2}{n} = CF$$

Step2

$$\Sigma\Sigma x_{ij} = a^2+b^2+c^2+d^2+\cdots+L^2$$

$$SST = \Sigma\Sigma x_{ij}^2 - CF$$

Step3:-

$$\frac{\Sigma C_i^2}{n_i} = \frac{(a+e+i)^2}{n} + \frac{(b+f+J)^2}{n} + \cdots + \frac{(d+h+L)^2}{n}$$

Here $n=3$

$$SSTR = \frac{\Sigma C_i^2}{n_i} - CF$$

Step4:-

SSBL for Row

$$\hookrightarrow \frac{\Sigma R_i^2}{n} - CF = SSBL$$

$$\Rightarrow \frac{(a+b+c+d)^2}{n} + \frac{(e+f+g+h)^2}{n} + \frac{(i+J+K+L)^2}{n} = \frac{\Sigma R_i^2}{n}$$

$n=4$ here.

$SSBL = \frac{1}{k} \sum \frac{R_i^2}{n} - CF$

Step5:-

$SSE = SST - SSTR - SSBL$

Step6 : table

Source of variance	df	sum of squares	mean square	F ratio
b/w treatments	$k-1$	SSTR	$MSTR = \frac{SSTR}{(k-1)}$	for treatment $F = \frac{MSTR}{MSE}$ $F(\alpha, \text{cal value})$
b/w blocks	$m-1$	SSBL	$MSBL = \frac{SSBL}{m-1}$	for BLOCK $F = \frac{MSBL}{MSE}$ $F(\alpha, \text{cal value})$
within Groups	$(k-1)*(m-1)$	SSE	$MSE = \frac{SSE}{(k-1)(m-1)}$	

$K=4 \Rightarrow$ 4 treatments $\Rightarrow G_1, G_2, G_3, G_4$ } in this example.

$m=3 \Rightarrow$ 3 Blocks $\Rightarrow 1, 2, 3$ }

If $F_{cal} < F_{table}$ Accept H_0 Reject H_1

$F_{treatment} \Rightarrow F(\alpha, (k-1), (k-1)(m-1))$

$F_{Block} \Rightarrow F(\alpha, (m-1), (k-1)(m-1))$

Maximum likelihood Estimation :-

Normal

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

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

$$\frac{dl}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sigma^2 = \frac{\Sigma(x_i-\bar{x})^2}{n}$$

Bernoulli

$$L_p = p^x(1-p)^{n-x}$$

log

$$l_p = x\log p + n-x\log(1-p)$$

$$\frac{dl}{dp} = \frac{x}{p} - \frac{n-x}{1-p} = 0$$

$$\theta = x_{ln}$$

Binomial

Poisson

$$\frac{\partial l}{\partial \mu} = -n + \frac{n\bar{x}}{\mu}$$

$$\hat{\mu} = \bar{x}$$

$$\lambda = \frac{\Sigma x_i}{n}$$

$$\frac{dl}{d\mu} = -n + \frac{\Sigma x_i}{\mu}$$

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

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

Least square

$$d(b_0, b_1) = \sum_{i=1}^{n}\left[y_i - (b_0 + b_1 x_i)\right]^2$$

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

no calculator so $x!=0$

Line of Regression / Linear regression

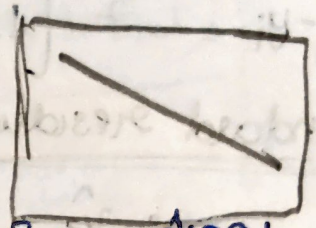
x	y	x^2	y^2	xy
1	5	1	25	5
2	6	4	36	12
3	7	9	49	21
4	8	16	64	32
Σx	Σy	Σx^2	Σy^2	Σxy

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

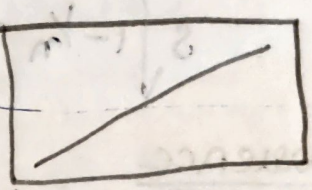
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \cdot \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$\hat{\beta}_0 = \frac{\Sigma y_i}{n} - \hat{\beta}_1 \cdot \frac{\Sigma x_i}{n}$$

If $\beta_1 < 0$ diag:



If $\beta_1 > 0$ diag:



Sum of sq of error / Residuals

$$SSE = \Sigma y_i^2 - \hat{\beta}_0 \Sigma y_i - \hat{\beta}_1 \Sigma x_i y_i$$

$$\sigma^2 = s^2 = \frac{SSE}{n-2}$$

Total sum of squares

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$SST = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

Co-efficient of determination

$$r^2 = 1 - \frac{SSE}{SST}$$ [blw 0 & 1]

(or)

$$r^2 = \frac{SSRg}{SST}$$

atlest ≥

$$y_i - \hat{y}_i = \sigma^2 \cdot \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{Sxx}\right]$$

Standard residual:-

$$e_i = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{Sxx}}}$$

covarience

$$var_x = \frac{\Sigma (x - \bar{x})^2}{N-1} = \frac{\Sigma (x - \bar{x}) \cdot (x - \bar{x})}{N-1}$$

$$cov_{xy} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N-1}$$

formulae for finding corelation

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{(N\Sigma x^2 - (\Sigma x)^2)(N\Sigma y^2 - (\Sigma y)^2}}$$

fitting a trend to timeseries:- (tita linear trend)

for odd num:

$$x = \frac{t - \text{middle value of } t}{\text{Interval}(h)}$$

for even no's

$$x = \frac{t - (\text{mean of middle 2 values of } t)}{\frac{1}{2}\,\text{Interval}(h)}$$

$$\hat{b} = \frac{\Sigma x y_t}{\Sigma x^2} \qquad \hat{a} = \frac{\Sigma y_t}{n} \qquad \hat{y}_t = \hat{a} + \hat{b}x$$

year t	Production y_t	$x = \frac{t - \text{value}}{\text{Interval}}$	x^2	xy_t
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
	Σy_t	Σx	Σx^2	Σxy_t

Exponential smoothing technique:-

L compute moving Avg

simple exponential smoothing with α

$$F_1 = F_2 = y$$

$$F_n = \alpha \cdot y_{n-1} + (1-\alpha) F_{n-1}$$

L Performance measures

$$MSE = \frac{\Sigma(\Delta t)^2}{n} \qquad MAPE = \frac{\Sigma \frac{\Delta t}{y_t}}{n} * 100$$

$$MAD = \frac{\Sigma |\Delta t|}{n} \qquad LAD = MAX\,|\Delta t|$$

$$\Delta t = y_t - P_t$$

Gmm

	RB	NT
Literacy		5
ID 1	20	
		3
ID 2	30	
		7
ID 3	50	

cluster : [(20,1)(10,1), 0.5
cluster [(50,7)(10,1), 0.5

in order of μ σ π

Ans

cluster 1

$\mu_1 = 20$ $\sigma_1 = 10$ $\pi_1 = 0.5$

$\mu_2 = 1$ $\sigma_2 = 1$

cluster 2

$\mu_1 = 50$ $\sigma_1 = 10$ $\pi_2 = 0.5$

$\mu_2 = 7$ $\sigma_2 = 1$

$$N(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[- \frac{(x-\mu)^2}{2\sigma^2} \right]$$

cluster 1 & 2

for ID 1 = (70, 5) for cluster 1

feature 1 ($N(x_1, \mu_1, \sigma_1)$

$= N(70, 20, 10)$

$$g_{12} = \frac{1}{2\pi(10)} \exp\left[- \frac{(70-20)^2}{2(10)^2} \right) = 1.4867198 \times 10^{-4}$$

feature 2 for 201 (70,5)

$\sim N(x_2, \mu_2, \sigma_2)$

$= N(5, 1, 1)$

$g_2 = \frac{1}{2\pi(1)} \cdot \exp\left[-\frac{(5-1)^2}{2(1)^2}\right] = 1.33830226 \times 10^{-4}$

Product → gaussian 1 * gaussian 2

$= 1.4687 \times 10^{-7} * 1.338 \times 10^{-4} = 1.9648 \times 10^{-4}$

Weight $= \pi \cdot p$

$= 0.5$ (product)

$= 0.5 (1.9648 \times 10^{-11}) = 9.9512 \times 10^{-12}$

$\gamma_{i1} = \frac{\text{Weight}_1}{\text{Weight}_1 + \text{weight}_2}$

$\gamma_2 = \frac{\text{Weight}_2}{\text{Weight}_2 + \text{weight}_2}$

Same way

cluster 2 for feature 1 & feature 2

then next point [30,3]

do cluster 1 → feature 1
→ feature 2

do cluster 2 → feature 1
→ feature 2

Spearman

Rank of correlation coefficient

$$\rho_{\text{Spearman}} R = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$