QTMD-2

Percapita income of employed women in Jamshedpur

$PI = \frac{\text{Total Income}}{\text{Tot. No. of employed women}}$

1) Household survey (census)

we are applying simple mathematical tool ratio. But before doing ratio we need to do lot of work to get data.

So, to get a simple value like this we can't afford this tool based on resources

* Statistic is the measure of a sample.
* Parameter " " " " population.

Descriptive statistics

Inferential statistics ← statistical estimate (point estimate, interval estimate) / Testing Hypothesis

Sample → A small representative unit of population.

μ = pop. mean

$\bar{x}$ = Sample mean

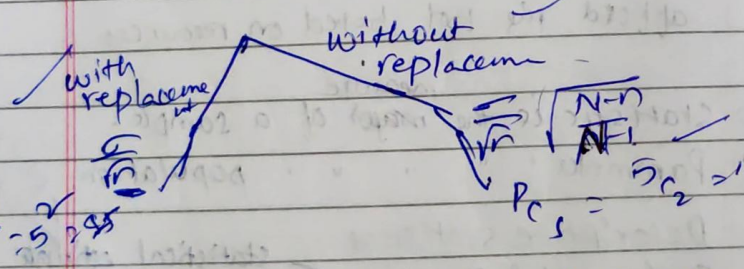
$\sigma \rightarrow$ pop. std deviation

$s \Rightarrow$ sample " "

$$\sigma^2 = \frac{1}{N}\sum (x_i - M)^2 \Rightarrow \sigma = \sqrt{\frac{1}{N}\sum (x_i - M)^2}$$

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

Standard Error $E(\bar{x}) = M$



Standard error is standard deviation also.

Eg:- pop(N) = 3000 $M = 68$, $\sigma^2 = 3$

50 samples of size = 25

a) with replacement.

$\bar{x} =$ . . $\sigma = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$

Mean will be same with and without replacement.

Std. deviation of sampling with replacement

$= \frac{\sigma}{\sqrt{n}} = \frac{3}{5} = 0.6$

without replacement.

$= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{5}\sqrt{\frac{3000-25}{3000-1}}$

$\simeq 1$

$= 0.6$

$\frac{n}{N} < 0.05$

then we can tell size of sample is small compared to population.

Standard Error, $z = \frac{\bar{x} - M}{SE}$

$P(66.8 \leq \bar{x} \leq 68.3) = ? \times 50 =$

$\bar{x}_1 = 66.8$, $z_1 = \frac{66.8 - 68}{0.6} = -2$

$\bar{x}_n = 68.3$; $Z_{\bar{x}} = \frac{68.3 - 68}{0.6} \approx 0.5$

$P(-2 \le Z \le 0.5) = ?$ $0.6687 \times 80 \simeq 53$ samples

$P(\bar{x} \le 66.4)$; $Z = \frac{66.4 - 68}{0.6} = -2.66$

$P(Z \le -2.66) = P(-2.66 \le Z \le 0) = 0.0038$

$\times 80 = 0.304$

Eg2

100 stds → sample

pop → 1546, mean $\mu = 67.45$, $\sigma = 2.93$.

Find 95% confidence interval.

(0) — $\bar{x}$ → $\hat{\mu}$ → point estimate value

1000 5000

10/11/21

Mean, $\hat{\mu} = \left[\bar{x} \pm Z * SE_{\bar{x}}\right]$

confidence → 95%.

95% sure

we can create 100 such intervals. In [ ] such intervals population mean will not be there.

Only 95 intervals are considered as valid.

Just like mean we can have any parameter we can estimate based on the sample statistic.

Eg1 Random sample of 50 QT grades out of total 200 stds mean = 75, $\sigma = 10$. What are 95% confidence limits for estimates of the mean of 200 grades.

$\mu = 75 \pm$. $n(s) = 50$, $n(p) = 200$.

$SE = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{10}{\sqrt{50}} \sqrt{\frac{200-50}{200-1}} = 1.227$

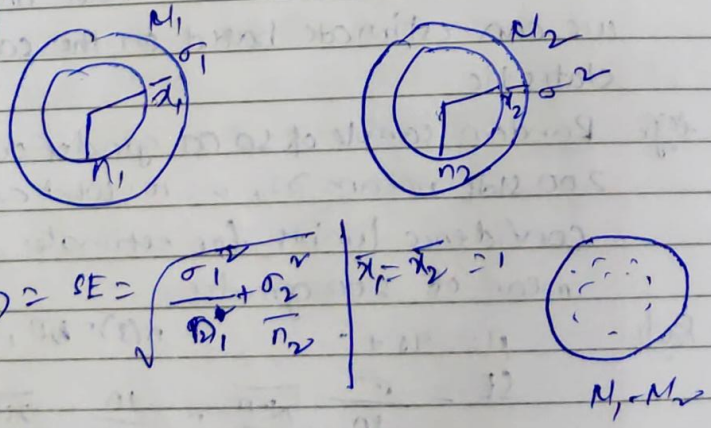
$x_i - \bar{x}$

Estimated interval $= \bar{x} \pm Z_\alpha \times SE_{\bar{x}}$

$= [75 \pm 1.96 \times 1.227]$

$= [72.6, 77.4]$

This interval is based on your sample. This is one of such as valid intervals. Based on this interval we can estimate population.



$SD = SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ | $\bar{x}_1 = \bar{x}_2 = 1$

$M_1 = M_2$

$P_1 : q_1$ n_1 n_2 $P_2 : q_2$ $\sqrt{\frac{PQ}{n}}$

$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$; $Q = 1 - P$

$SE_{P_1 \sim P_2} = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Eg:- A sample of 150 brand A bulbs mean = 1400 hrs. σ = 120 hrs. A sample of 200 brand B lights μ = 1200 h, σ = 80 h. Find 95% confidence limits for the diff of mean life time of the population of brands A & B

Sol:-

$\bar{x}_1 - \bar{x}_2 \pm CD = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(120)^2}{150} + \frac{(80)^2}{200}}$

$= \sqrt{6000} = 11.31$

$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm Z * SE$

$= (1400 - 1200) \pm 1.96 * 11.31$

$= 200 \pm 22.1676$

$= [177.8324, 222.1676]$

(9)(10) random sample of 400 adults and 600 teenagers, 100 adults, 300 teenagers liked it. Construct a 95% confidence limits for the diff in proportion of all adults and teenagers who wa

Sol:

$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{400 \times 100 + 600 \times 300}{1000}$

$= 400 + 1800 = 220.$

$Q = 1 - P =$

$SE = \sqrt{220 \left(\frac{1}{600} + \frac{1}{400}\right)}$

$P_1 = 0.25 \left(\frac{100}{400}\right) \quad P_2 = 0.5 \left(\frac{300}{600}\right)$

$n_1 = 400 \qquad n_2 = 600$

$P = n_1 P_1 = \frac{400 \times 0.25 + 600 \times 0.5}{1000}$

$= \frac{100 + 300}{1000} = 0.4$

$Q = 1 - P = 0.6$

$SE = \sqrt{(0.4)(0.6)\left(\frac{1}{600} + \frac{1}{400}\right)}$

$= \sqrt{0.24 * \left(\frac{1000}{240000}\right)} = 0.031$

$P_1 - P_2 = [0.25 \pm 1.96^{Z} \times 0.031^{SE}]$

$= [0.19, 0.31]$

13) axial length of pistons produced is normally distributed with variance of 64mm. Determ sample size such that mean length of pistons can be obtained within a variation of 2mm with a confidence level of 95%.

Sol) variance = 64 ⇒ SD $\sigma = 8$ $n = ?$

$\mu = \bar{x} \pm Z_{95} \times SE_{\bar{x}}$

$\mu - \bar{x} = \pm Z_{95} \times SE = 2mm$

$\Rightarrow \frac{1.96 \times \sigma}{\sqrt{n}} = 2$

$\frac{1.96 \times 8}{2} = \sqrt{n} \Rightarrow n = 61.54$

→ If the sample mean of 10 elements is 50.

$\frac{x_1 + x_2 + \dots x_{10}}{10} = 50$.

Out of these 10 elements can change independently arbitarily so that the mean still remains at 50

Max 9 elements randomly can be changed, 10th will be depends on sum of 9 elements and 500.

No. of independent components sample $= n-1$

No. of " components is called as "Degrees of freedom".

$E(\bar{x}) = \mu$ $s^2_{sam} \cdot \sigma^2_{pop}$

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

expected value $E(s^2) \neq \sigma^2$ $s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$

If this is replaced by $n-1$ then

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

Bessel's correction → n is changed to n-1.

$P = \{2, 3, 4, 5, 6\}$

$n(s) = 2$ $\bar{x} \Rightarrow 2, 2.5, 3, 3.5, 4, 4.5$

possibles 1, 2, 3, 4

Using Basil's correction $E(s^2) = \sigma^2$.

$n-1$ is considered (interpreted) as degree of freedom.

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

↳ sample size − 1

called as degrees of freedom.

practice

Descriptive Statistics

A set of data, and how do we describe data.

→ How we describe a set of data is called descriptive statistics.

Measure of Central Tendancy → describes about central/focal point of dataset.

" " desperation

→ A random variable is a variable which is associated with a random experiment. The variable may be discrete or continuous.

→ Frequency distribution

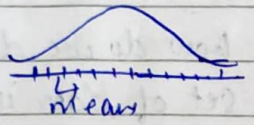
Marks → variable

How datapoints occurs frequently in a distribution

→ Statistic → It is a characteristic of a sample.

Parameter → " " " " population

→ Central limit theorem → When the size of the sample increases the sampling distribution of mean is same as normal distribution.



$E(\bar{x}) = \mu$.

Sampling Error with replacement

" " Without "

No. of samples drawn:

with replacement $= {}^{ps}5^2 = 25$ $\left| \frac{\sigma}{\sqrt{n}} = SE \right.$ (pop size, sample size)

without " ${}^{ps}5_{C_2} = 10$ $\rightarrow \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = SE$

Point Estimation

Eg: pop = 3000, $\mu = 68$. $\sigma = 3$,

sample size = 25 $\bar{x} = ?$ $s = ?$, no. of samples = 80

Mean will be same

with replacement = 68, | without = 68

$\sigma = \frac{3}{\sqrt{25}} = 3/5 = 0.6$ | $\frac{3}{\sqrt{25}} \sqrt{\frac{3000-25}{3000-1}} = 0.6 \times 1$

$= 0.6$

This happens b/c size of sample is less when compared to size of population.

If size of the sample by size of pop $\frac{n}{N} < 0.05$ then we can say size of the sample is small compared to size of population.

Then with replacement and without replacement are same.

→ In how many samples of the above problem would you expect to find the mean

a) b/w 66.8 and 68.3 inches.

Total 80 samples 68

$$z = \frac{\bar{x} - \mu}{SE}$$

$P(66.8 \le \bar{x} \le 68.3) = ? \times 80 = ?$

$\bar{x}_1 = 66.8$, $z_1 = \frac{66.8 - 68}{0.6} = -2$

$\bar{x}_2 = 68.3$, $z_2 = \frac{68.3 - 68}{0.6} = 0.5$ → pnorm(-2) pnorm(0.5)

$-2 \le z \le 0.5 \rightarrow P(-2 \le z \le 0.5) = 0.6687 \times 80$

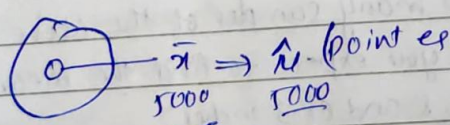
$= 53$

b) < 66.4 $P(\bar{x} \le 66.4)$; $z = \frac{66.4 - 68}{0.6} = -2.66$.

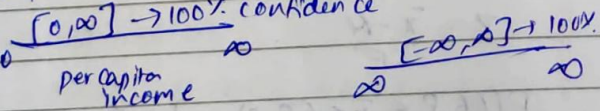
$P(z \le -2.66) = P(-2.66 \le z \le 0) = 0.0038 \times 80$

$= 0.304 = 0$

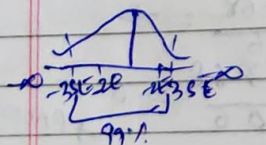
② heights of 100 male stds
Pop = 1546, $\mu = 67.45$, $\sigma = 2.93$
find 95% confidence interval for estimating the mean height of the stds.
point estimate



Sample is representative of population.



This interval is meaningless



$$\mu \pm Z = \frac{\bar{x} - \mu}{SE}$$

$$\hat{\mu} = [\bar{x} \pm Z \cdot SE]$$

95% 99% 90%

Based on the confidence level we can find the interval.

If the confidence is 99% then 1% is incorrect

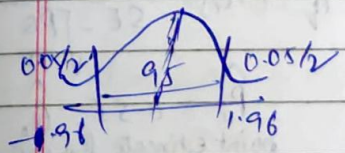
$$\hat{\mu} = [\bar{x} \pm Z \times SE] \quad \sigma = 2.93$$

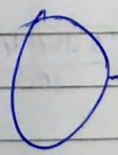
$$= [67.45 \pm 0.995 \times 2.93] \quad Z = \frac{\bar{x} - \mu}{\sigma} = \frac{}{2.93}$$

$$= \quad Z(0.995)$$

$$SE = \frac{2.93}{\sqrt{100}} \sqrt{\frac{1546-100}{1546-1}} = \frac{2.93}{10} \times \sqrt{\frac{1446}{1545}} = 0.2834$$

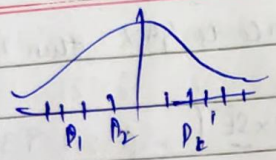
$$\hat{\mu} = [67.4 \pm 0.2834 \times 1.96]$$ → Confidence of 95%
Z = 1.96





$P(S) = \frac{200}{1000} = 1/5$ p & q represent proportions

$\bar{q} = 1 - P(S) = 4/5$



$SE_p = E(p_i) = p$

$SE_p = \sqrt{\frac{PQ}{n}}$

$SE_p = \sqrt{\frac{p_1 q_1}{100}} =$

Pb3 Sample of 100 voters chosen, 55% of them are in favor of a particular candidate. (all voters)

$n = 100 \quad p = 0.55 \quad q = 0.45 \quad SE = \sqrt{\frac{PQ}{n}} =$

(estimate) $\hat{p} = p \pm Z_{95\%} \, SE$

$= 0.55 \pm$

$[0.55 \pm 1.96 \times 0.04974]$

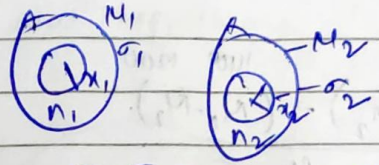
[ ]

$\hat{P}_0 = 0.55, \quad \hat{Q}_0 = 0.45$

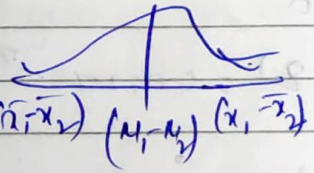
point estimate for pop. proportion

$SE_p = \sqrt{\frac{0.55 \times 0.45}{100}} = 0.04974$

2 population's



$\bar{x}_1 - \bar{x}_2$ is a new statistic and this also follow a distribution



Std. dev of pop, & pop,

$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Size of sample of pop, and pop,

$E(\bar{x}_1 - \bar{x}_2) = M_1 - M_2$

Pb4 Manfa A → mean = 1400, $\sigma = 200$ (pop)

B → " = 1200, $\sigma = 100$

Random sample of 125 (A) 125 (B)

what is prob that brand A bulbs will have a mean life time that is at least

a) 100 hrs more than the brand B bulbs.

$SE = \sqrt{\frac{(200)^2}{125} + \frac{(100)^2}{125}} = \sqrt{\frac{}{125}}$

$= 20$

$$Z = \frac{\bar{x} - \mu}{SE}$$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}}$$ (1400, 1200)

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - 200}{20} \qquad \frac{\bar{x}_2 + 160 - \bar{x}_2 - 200}{20}$$

$$Z = -2$$

$$P(Z) = P(-2) = 0.0227$$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}} = \frac{7 - 6}{1.38} = \frac{1}{1.38} = 0.72$$

$$SE = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = 1.38$$

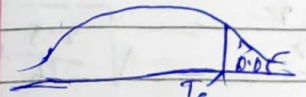
Sample doesn't give enough evidance to reject null hypothesis

⇒ Mean weight = >100 gms sample of 25

$\mu = 97$, variance = 64 claim of manager at 0.05

Null hypothesis $H_0: \mu \leq 100$

$H_a: \mu > 100$



Variance $s^2 = 64$

sample size < 30

So, we have to take T-distribution

Degrees of freedom = $n - 1 = 25 - 1 = 24$.

$T_{critical} = T_c = 1.711$

$$t = \frac{\bar{x} - \mu}{SE} \qquad SE = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{25}} = \frac{8}{5} = 1.6$$

$$= \frac{97 - 100}{1.6} = \frac{-3}{1.6} = -1.875 < t_{critical}.$$

So, it doesn't fall in rejection.

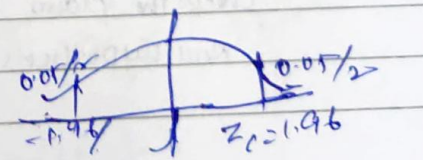
$T < T_c$ failed to reject. His claim is false

diameter (bt) = b_2 → sample 20 verify claim of producer

Sample b → Mean = 39.5, Var = 2.5 mean = 40, var = 4

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$



$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{SE}$ | $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$SE = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ $= \sqrt{\frac{2.5}{16} + \frac{4}{20s}}$

$= \sqrt{\frac{12.5+16}{80}} \Rightarrow$

$SE = SP\sqrt{\frac{1}{16} + \frac{1}{20}}$

$= 0.59$

$\frac{2.5-4}{0.59} = \frac{1.5}{0.59} = 2.517$

If falls in the rejection area of null hypo. so, reject the null hypothesis

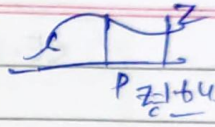
Managers claim is wrong.

$\Rightarrow$ >90%. Sample = 80, it cured 68% pati. check the claim at 0.05

Null hypothesis $\rightarrow h_0 : P \leq 90\%$

↓ population proportion

Alt hypothesis $h_1 : P_1 > 90\%$



Standard error.

$SE = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.9 \times 0.1}{80}}$

$= \sqrt{\frac{0.09}{80}} = 0.034$

$z = \frac{P-P}{SE} = \frac{0.68 - 0.9}{0.034} = -6.47$

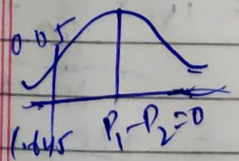
Not in the rejection area then failed to reject null hypothesis.

$\rightarrow$. $RT_2 > RT_1$,

Sample 1 = 110 days on 90 days target > | $S_2 = 130$ days 112 days target >

check at 0.05 level

Null $H_0 : P_2 \leq P_1$ $H_1 : P_2 > P_1$



$SE_{P1-P2} = \frac{P_{n_1}P_1 + n_2P_2}{n_1+n_2}$ $P_1 = \frac{90}{110}$ $P_2 = \frac{112}{130}$

$Q = 1 - P$

$PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

SE = 0.042 $Z = \frac{(P_1 - P_2) - (P_1 - P_2)}{SE}$

$= \frac{\left(\frac{110}{110} - \frac{112}{130}\right)}{0.042} = -0.922$

It is not in rejection area.

So, we can't reject null hypothesis.

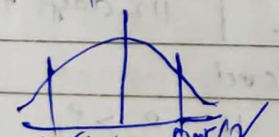
→ Sample 1 City X → 135 out of 150 Sig. level = 0.05

Sample 2 City Y → 133 out of 160.

Let $H_0 : P_1 = P_2$ proportion of DC holders from X, Y

$H_a : P_1 \neq P_2$

Sampling distribution of proportion



Z_c

$Z = \frac{\hat{p} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}}{}$

$= \frac{150 \times \frac{135}{150} + 160 \times \frac{133}{160}}{150 + 160}$

$= \frac{268}{310} = 0.86$

$Q = 0.14$

$SE = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.86 \times 0.14\left(\frac{1}{160} + \frac{1}{150}\right)}$

$= 0.04$

$Z = \frac{P_1 - P_2}{SE} = \frac{\left(\frac{135}{150} - \frac{133}{160}\right)}{0.04} = 1.$

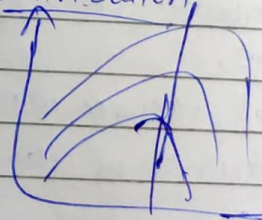
$= 1.77$

Z lies within range.

Null hypothesis is not rejected

Failed to reject null hypothesis

t-distribution



pop. when standard dev is not given and pop. size ≤ 30

If n > 30 then it is similar to Z-dist.

+ve skewed range $[0, \infty]$

26/10/21

Chi-square and F-test

F-test we use for 2 populations.

F-calc: $F = \frac{S_1^2}{S_2^2}$

Compare variance of 2 pop by taking 2 samples

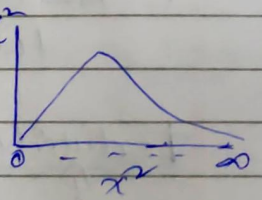
In F-table: Col level → Num of F; Row level → Denom of F.

Chi-square is also testing variance w.r.t. one population.

Inferential, hypothesis testing either with 1 population or 2 populations.

S.D of sample

$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ ← SD of pop



+ve skewed distribution starts from 0 to ∞.

We will tell sample variance w.r.t. pop. variance.

Other way of representation of chi-square.

$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ → observed frequency; E_i → Expected frequency

Will comp with Table F-val

This way of Pearson chi-square is used for goodness of fit.

Eg:- No. of car accidents per month in a certain city were as follows:

12, 18, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident conditions were same during this 10 month period.

Sol:- Null hypothesis H_0: Accident conditions were same

H_1: conditions were not same.

Total No. of accidents = 110.

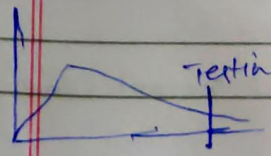
If conditions same no. of accidents/month = 11 ∴

Based on this,

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

$O_i - E_i > 0$ or < 0

If the diff is significantly high then χ^2 value will be high



If χ^2 is beyond certain limit we can tell the alt. hypothesis is true or if it is within the limit then we can tell that we can't reject the null hypothesis.

$$\chi^2 = \frac{(8-11)^2 + (18-11)^2 + (20-11)^2 + (2-11)^2 + (14-11)^2 + (10-11)^2 + (15-11)^2 + (6-11)^2 + (9-11)^2 + (4-11)^2}{11}$$

$$= \frac{1+49+81+81+9+1+16+25+4+49}{11}$$

$$= \frac{2\times 81 + 2\times 49 + 316}{11} = \frac{316}{11} \quad 28.7$$

10 observations

Degree of freedom = $n-1 \rightarrow 9$

At 5% level of significance of DOF $\rightarrow 9$ then $\chi^2 = 16.19$

we can infer reject null hypothesis

$(28.7 > 16.19)$

So, the accident conditions are not same.

Based on the data we are checking wether our assumption is correct or not. This is good fit.

Pb 2:) Theory predicts proportion of item in 4 grps A, B, CD → 9:3:3:1. In an exp among 1600 items, the no's in grp were 882, 313, 287 and 118. Does this exp support the thy.

Sol:) Null hypothesis H_0:

H_1:

Expected = $\frac{\text{Total}}{4} = \frac{1600}{4} = 400$

$$\chi^2 = \frac{(882-400)^2 + (313-400)^2 + (287-400)^2 + (118-400)^2}{400}$$

$$= \frac{232,324 + 7569 + 12,769 + 79,524}{400}$$

$$= 804.65$$

Sol:- Null hypothesis H_0 : Experiment supports the theory
H_a : exp doesn't support the theory.

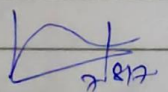
Theory ⇒ A : B : C : D = 9 : 3 : 3 : 1

	Experim E_i	expected T_i (9:3:3:1)
A	882	900
B	313	300
C	287	300
D	118	100
		1600

$$\chi^2 = \frac{(882-900)^2}{900} + \frac{(313-300)^2}{300} + \frac{(287-300)^2}{300} + \frac{(118-100)^2}{100}$$

$$= 4.73$$

DOF = 3



At 5% significance, critical point = 7.815

4.7 < 7.817.

Inf:- Can't reject null hypothesis
The experiment support theory.

Pb 3:- Records taken of no of male and female births in 800 families having 4 children are given as:

male	F	families
0	4	32
1	3	178
2	2	290
3	1	236
4	0	64

Test whether the data is consistent with the hypothesis that male & Fem births are equally likely.

Sol:- Null H_0 : M & F births are equally likely
H_a : Not equally likely

Fam	Expected
32	160
178	160
290	160
236	160
64	160

prob. of male birth is same as female birth
$p(m) = p(f) = 1/2$

$\chi^2 = 128^2 + 18^2 + 130^2 +$

$P(m) = P(f) = 1/2$ Total families $= 800$.

In a family of 4 children → No. of trails $= 4$

There may be 4 male, 3 / 2 / 1 female

Binomial distribution

$n = 4$

Max. 4 male

$P(r = r) = {}^nC_r \cdot p^r \cdot q^{n-r}$

$P(r = 0) = {}^4C_0 \cdot (1/2)^0 (1/2)^4 = 0.0625$

$P(r = 1) = {}^4C_1 (1/2)^1 (1/2)^3 = 0.25$

$= 0.375$

$P(r = 4) = {}^4C_4 (1/2)^4 (1/2)^0 = 0.25$

$= 0.0625$

Prob. that there is 1 male child.

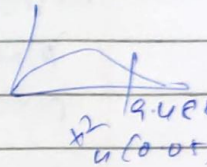
actual	Exp. no. of families
32	$800 \times 0.0625 = 50$
178	$800 \times 0.25 = 200$
290	$800 \times 0.375 = 300$
236	$800 \times 0.25 = 200$
64	$800 \times 0.0625 = 50$

$\chi^2 = \frac{18^2}{50} + \frac{22^2}{200} + \frac{10^2}{200} + \frac{36^2}{200} + \frac{14^2}{50}$

$= 6.48 + 2.42 + 0.333 + 6.48 + 3.92$

$= 19.63$

Critical Point $= 9.488$.



Reject null hypothesis that male & female birth are not equally likely.

Pb 5] The dist of typing mistakes committed by a typist is given as.

Mistake/page	0	1	2	3	4	5
No. of pages	142	156	69	27	5	1

Assuming that dist to be random find out the exp. no. of pages containing 0, 1, 2, 3, 4, 5 mistakes resp. Test a hypothesis that there is no sign. difference among the observed & expected no. of pages containing the mistakes.

Sol:

H_0 :

H_a :

Poisson distribution,

$$P.M.F \quad P(X=r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!} \qquad r = 0, 1, 2 \ldots \infty$$

λ = Avg. rate of occurance.

Tot. No of mistakes = $5\times1 + 4\times5 + 3\times27 + 2\times89 + 1\times156$

$= 5 + 20 + 81 + 138 + 156 = 400$

Tot. pages = 400.

Mistake / page = $1 \leftrightarrow \lambda$.

$$P(X=0) = \frac{e^{-1} \cdot 1^0}{0!} = 0.3678$$

$$P(X=1) = \frac{e^{-1} \cdot 1^1}{1!} = 0.3678$$

$= 0.184$

$= 0.0613, \quad 0.015$

$$P(X=5) = \frac{e^{-1} \cdot 5^5}{5!} = 0.003$$

Expected no. of pages containing 0 mistakes = 0.3678 × 400

$= 147.$

:

$\chi^2 = 1.45$

$\chi^2_{0.05}(5) = 11.07$

The no. of mistakes that occur randomly can be calculated using poisson dist. So, poisson dist is best fit.

28/10/21

If there is a significance diff b/w expected level and the given (observed) level.

If observed = expected then chi-square will be 0. If the diff is more and more chi-square will be higher.

chi-square one handed test (right handed test)

Test of independence

$H_0: P_1 = P_2$ (based on proportion)

$H_a: P_1 \neq P_2$ (1) (2)

If $H_0: P_1 = P_2 = P_3 = P_4$ (Multiple populations in terms of proportion)

$H_a: P_1 \neq P_2 \neq P_3 \neq P_4$

Proportions are not same in at least 2 pops. so, we can test multiple times taking 2 pops at a time.

The prob of null hypothesis is high

In this case we can use chi-square test.

Eg:-

	NE	South	Central	WestCoast	Total
Present Method	68	75	57	79	279
New Method	32	45	33	31	141
Total	100	120	90	110	420

Area is divided into 4 populations. Students applied for interviews. Agency is thinking about new method of selecting candidates.

Sample sizes are NE 100, SE 120, Central 90, WestCoast 110

In first sample 68 out of 100 choose present method and 32 choose new method

sample proportion $\hat{p}_1$ = present 0.68, new 0.32

From samples we can only study proportions.

H_0 : Proportion of stds referring the method is same across regions ($P_1 = P_2 = P_3 = P_4$)

H_a : ~~If~~ $P_2 \neq P_3 \neq P_4$

If we aggregate all the samples then we will get sample for population.
like 100+120+90+110 = 420 → sample size.
bigger

Total No. of candidates referring present method: 229, new method = 141.
Proportion of candidates preferring present method = 0.66, new method = 0.33

These proportions can be used as an estimate proportion of candidates.
So, in NE region.

present	66	66.4	Expected frequency
New	32 (actual)	33.6	
C			
SE 75		79.7	
US		40.3	

Central	57	59.8
	33	30.2
Westcoast	79	73.1
	31	36.9

chi square, $\chi^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_p}$

$= \frac{(68-66.4)^2}{66.4} + \frac{(32-33.6)^2}{33.1} + \frac{(75-79.7)^2}{79.7} +$

$\frac{(45-40.3)^2}{40.3} + \frac{(57-59.8)^2}{59.8} + \frac{(33-30.2)^2}{30.2} +$

$\frac{(79-73.1)^2}{73.1} + \frac{(31-36.9)^2}{36.9}$

$= 0.038 + 0.076 + 0.2771 + 0.548 + 0.137 +$

$0.2591 + 0.426 + 0.943$

$\chi^2 = 2.36$.

As here we have table Degrees of Freedom will be different.

As per the first row DOF = 4 (no. of elements) − 1 = 3

" " " " column, DOF = 2 − 1 = 1

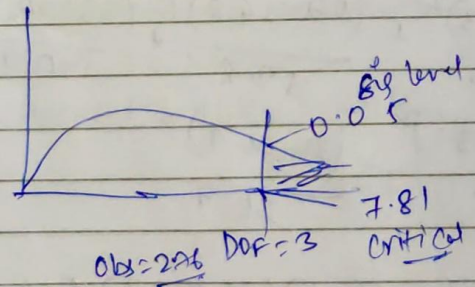
New table consists of 4 × 2 = 8 elements.

DOF for this table = 3 × 1 = 3

In this case

$H_0 = P_1 = P_2 = P_3 = P_4$

$H_a = P_1 \neq P_2 \neq P_3 \neq P_4$



failed to reject null hypothesis.

Proportion of candidates in three regions are not significantly different. Preference of methods is independent in these 4 regions

Two attributes, one is method and other is region. Here both the attributes doesnt influence the preference.

Pb 4 ⇒

Hrs spent vs Avg. grade.

	A	B	C	D	F.	Tot
<5 hrs	13	10	11	16	5	55
	6.875					
5-10 hrs	20	27	27	19	2	95
	11.875					
10-15 hrs	9	27	71	16	32	155
	19.375					
>20 hrs	8	11	41	24	11	95
	11.875					
Tot	50	75	150	75	50	400

$H_0 : P_1 = P_2 = P_3 = P_4 = P_5$

$H_1 : P_1 \neq P_2 \neq P_3 \neq P_4 \neq P_5$.

Proportion of stds who listen to < 5 hrs

$P_1 = \frac{55}{400} = 0.1375$ → 55/400

$P_2 = 0.2375$, → 95/400

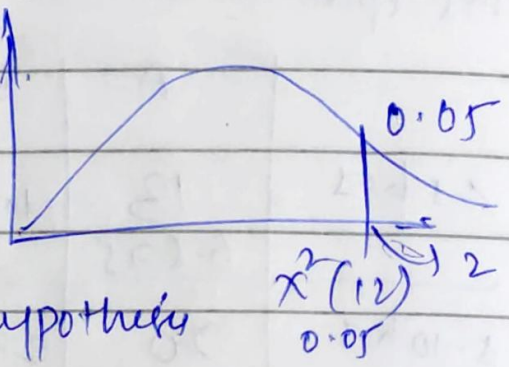
$P_3 = 0.3875$, → 155/400

$P_4 = 0.2375$ → 95/400

$\chi^2 = 63.83$ DOF $= (5-1) \times (4-1) = 12$

5% sig

Critical, $\chi^2 = 21.$



$\chi^2 >$ critical

reject the null hypothesis

$P_1 \neq P_2 \neq P_3 \neq P_4 \neq P_5$

Alt. hypothesis is true.

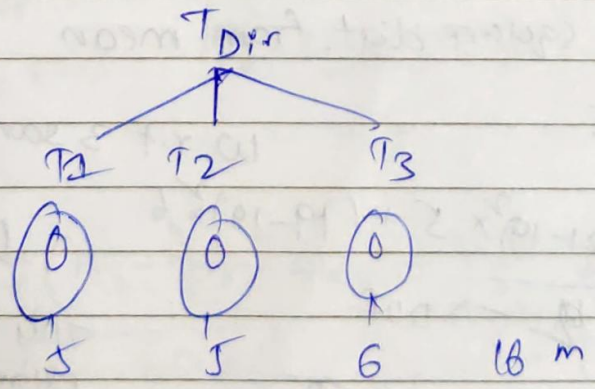
Proportions are dependent.

Getting avg. grade is dependent on no. of hrs listening to music

One attribute (getting grades) dependant on other attribute (listening to music).

Suppose if we have 2 populations A and B when we can we say both the populations are same?
when mean and variance together are same
(std. dev)

To test for the significance of the diff among more than 2 sample means.
Training director wanted to evaluate 3 different training methods to determine whether there were any differences in effectiveness of the training methods.
After completion of the training period, she chose 16 new employees assigned at random to 3 training methods.
Counting the production o/p by 16 trainees, she summarized the data and calculated the mean production of the trainees.



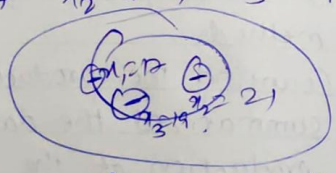
To determine the grand mean, any of the method can be followed.

Grand Mean = $\frac{(15+18+19+22+11)+(22+27+18+21+17)+(18+20+19+16+22+15)}{15} = 19$.

Method 1	M2	M3
17	22	18
⋮	⋮	⋮
11	17	15
85	105	114

Grand Mean = 19.

$\bar{x}_1$ (Mean of Sample 1) = 17 $\bar{x}_2 = 21$ & $\bar{x}_3 = 19$.



$H_0: M_1 = M_2 = M_3$

Variance :→ Avg. square dist from mean

$= \frac{\text{Total. Sq. Dist}}{\text{number}}$

w.r.t 3 sample mean

$\sigma^2 = \frac{5(17-19)^2 + (21-19)^2 \times 5 + (19-19)^2 \times 6}{\leftarrow \quad 12 \quad \rightarrow \text{DOF}}$

DOF = 3−1 = 2

Total distance of 3 samples we have taken elements in terms of samples

$\frac{20+20}{12} = 40/2 = 20$

This can be an estimated variance in terms of sample mean

29/10/21

ANOVA

Each of the sample is drawn from a normal population and that each of the population has the same variance.

If the sample size is large - Normality assumption is not required.

If null hypothesis is true, classifying data into 3 columns is unnecessary

3 samples will have their own sample variance,

For prev problem

S_1^2	S_2^2	S_3^2
15, 18, 19, 22, 11	22, 27, 18, 21, 17	18, 20, 19, 16, 22, 15.
Mean = $\frac{85}{5} = 17$	Mean = $\frac{105}{5} = 21$	Mean
$S_1^2 = \frac{2^2+1+2^2+25+36}{4}$	$S_2^2 = \frac{1+36+9+0+16}{4}$	$S_3^2 =$
$= \frac{70}{4} = 17.5$	$= 15.5$	$= 12$

$$S_1^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{(n_1-1)+(n_2-1)+(n_3-1)}$$

$$= \frac{4\times17.5 + 4\times15.5 + 5\times12}{13} = 14.769$$

First one, we got $\sigma^2 = 20$.

Estimated variance we got based on sample means.

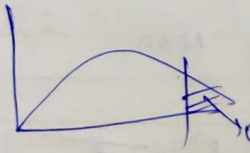
second one, variance within the samples

⇒ when populations are same then there is no significant difference b/w these 2 variances (var. among sample means and var. within/among) samples.

we can use f-test to compare 2 variances. So, that we can know the difference b/w variances.

F-test, $F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{20}{14.769} \rightarrow \begin{matrix} DOF \rightarrow 2 \\ DOF \rightarrow 13 \end{matrix}$

$= 1.354$



$F(2,13) = 3.81 \rightarrow$ critical.

$1.354 < 3.81$

Failed to reject null.

Populations are same. → This is from f-test

variances are same → This is assumption

i.e., means are same ∴ $\mu_1 = \mu_2 = \mu_3$.

Even if we have 20 populations to compare, those 20 populations can be converted into 2 variances i.e., variance amongst sample means and the other is variance within samples.

Problem McDonald,

						Avg
P1	3	4	5.5	3.5	4	4
P2	3	3.5	4.5	4	5.5	4.2
P3	2	3.5	5	6.5	6	4.6
P4	3	4	5.5	2.5	3	3.6
					$(\bar{x})$ =	4.075

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ $\quad H_a: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Variance among samples

$$= \frac{5(4 - 4.075)^2 + 5(4.075 - 4.1)^2 + 5(4.075 - 4.6)^2 + 5(4.075 - 3.6)^2}{3}$$

$$= \frac{5(0.028125) + 5(0.003125) + 5(1.3781) + 5(0.22585) \quad 1.128125}{3}$$

$$\sigma^2 = \frac{2.5874}{3} = 0.845$$

$\sigma_1^2 = 4(0.005624) + 4(0.000624) +$

S^2 → Mean = 4 | S_2^2 | S_3

$S_1^2 = \frac{(4-3)^2+(4-4)^2+(4-5.5)^2+(4-3.5)^2+(4-4)^2}{4} = 0.875$

$S_2^2 = 0.925$

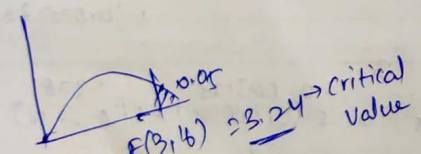
$S_3^2 = 3.425$

$SU^2 = 1.425$

$\sigma_1^2 = \frac{4(0.875)+4(0.925)+4(3.425)+4(1.425)}{16}$

$= 2.21 = 1.6625$

F-value, $F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{0.846}{1.6625} = 0.508$ → DOF = 3, DOF = 16



$F = \sigma_1^2 = \sigma_2^2$

We can not reject null hypothesis.

There is no sig. diff b/w estimated population variance

Hence, population's are same.

Assumption → variances are same. So, means are same

∴ Avg. service time is same

Non-Parametric Tests

Also applicable to a situation where data points instead of actual values acts as ranks.

It can also apply to assumptions.

Eg:- SAT scores of stds at 2 state universities

University A → 1000 1100 800 750 1300 950 1050
1250 1400 850 1150 1200 1500
600 725

B → 920 1120 830 1360 650 725 890
1600 900 1140 1550 550 1240 925 500.

$n_1 = 15$

$n_2 = 15$

$H_0 : M_1 = M_2$ (Assumption → Variances are same)

$H_a : M_1 \neq M_2$

Without assumption we can have parametric test i.e, U test (Mou test)

Sample elements should be expressed in terms of ranks.

Aggregate both universities and rank all the samples

Univ A → 16 18.

B →

What is the total rank of all the elements for both universities R_1 and R_2

$n_1 = 15, \quad n_2 = 15$

$R_1 = 247 \quad R_2 = 218$

Based on these values, $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$$\boxed{M_U = \frac{n_1 n_2}{2} \; ; \; SE = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

$$U = 15 \times 15 + \frac{15(16)}{2} - 247$$

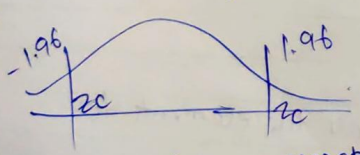
$$= 225 + 120 - 247 = 98$$

Mean, $M_U = \frac{225}{2} = 112.5, \quad SE = \sqrt{\frac{(15)^2(31)}{12}} = 24.1$

U distribution can be approximated (assumed) as z-distribution when $n_1, n_2 > 10$. Else we need to refer U table

Small samples will be used in biology

$$Z = \frac{\bar{x}_1 - M}{\sigma^2} \Rightarrow Z = \frac{U - M_U}{SE} = \frac{98 - 112.5}{24.1} = -0.601$$



-0.601 is b/w Critical value → Failed to reject null hypothesis

U test (Non-parametric) doesn't take assumptions.

We can also write as

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$= 225 + 120 - 218 = 127.$$

$$Z = \frac{127 - 112.5}{24.1} = 0.60$$

1/11/21

Kruskal-Wallis (K) Statistic

2 assumptions needed regarding

1) Normal distribution
2) Variance is same for all the populations

Then we cannot do the anova.

This will do if we have more than 2 pop and the data is converted to rank data.

Eg:- written examination scores of students by 3 diff methods

Video Cassette : 74 88 82 93 55 70 → $n_1 = 6$

$R_1 = 61$

Audio Cassette : 78 80 65 57 89 → $n_2 = 5$

$R_2 = ?$

Class room : 68 83 50 91 84 77 94 81 92 → $n_3 = 9$

$R_3 = ?$

$n = R_1 + R_2 + R_3 = 6 + 5 + 9 = 20$

K-statistic

$$K = \frac{12}{n(n+1)} \sum_{j=1}^{\text{population}} \frac{R_j^2}{n_j} - 3(n+1)$$

In this example $j = 1, 2, 3$.

$$K = \frac{12}{20 \times 21} \left(\frac{(6)^2}{5} + \frac{(}{6} + \frac{}{9} \right) - 3(21)$$

$$= 1.145$$

$H_0 : M_1 = M_2 = M_3$

$H_a : M_1 \neq M_2 \neq M_3$ → at a significance

Just like U-statistic we also have K-statistic

If the size of the sample is < 5 then normal distribution is used.

If the size of sample > 5 then chi square dist. is used.

If we need to verify hypothesis

$n_1 + n_2 + n_3 = n$. → DOF = 2

$\chi^2(2, 0.05) = 5.991$.

$1.145 < 5.991$.

Failed to reject null hypothesis

Correlation and Regression

Degree of similarity b/w 2 variable. Each variable can be expressed in terms of data.

They are called as univariate statistic.

Eg:- Mean, Meadian, variance, SD, mode etc.

Multivariate statistic.

We can have multiple variates explained from same data.

Eg:- Correlation, Regression.

There are measures for multivariate.

Degree of similarity b/w variate can be seen.

In simple terms Degree of similarity (DOS) can be explained in linear

If we explain in terms of linearly or by referrencing a straight line then we will call as linear correlati

It can be explained using Pearson linear coefficient.

Pearson coefficie is covariance of 2 variants

It can be explained in terms of $[-1, 1]$ ← $P_C = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y}$ 2 sets of data x, y

$$\sigma_x^2 = \frac{1}{N} \sum (x_i - \bar{x})^2$$

$$Cov(x, y) = \frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y})$$

If $P_c = -1$ then

2 variates are perfectly -vely correlated

i.e., If one variate is at max then other is at min.

If $P_c = +1$ then

2 variates are perfectly +vely correlated

i.e., if one variate is at max then other is also at max.

$$S_p = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Spearman Correlation Coefficient

Spearman

↳ Non-parametric equivalent of Pearson coefficient

x y

For pearson both Data set variates can be expressed in terms of ranks. and 'd' is difference in the ranks.

n is no. of ranks.

Regression

If we have 2 variates x and y and we measure the degree of relation

Regression is a functional relation b/w variables.

If 2 variates x and y, can we build a relationship b/w them. i.e., can we express x in terms of y & y in terms of x.

classmate Date Page

If we express as $y = f(x)$

then y is dependent and x is independent variable

Eg:- $y = 5x^2$

i.e., if we have a certain value of 'x' we can calculate y value

Instead of single independent var we can have multiple independent variables

Eg:- $y = 3x_1 + x_2$ → linear relationship.

It can be expressed as either a linear relationship or non-linear form of relationship.

Based on this principle,

If we have multiple variates $x_1, x_2, x_3 \cdots x$ over a period of time they got some time. If one of the variates is dependent variable. then it is expressed as $x_1 = f(x_2, x_3 \cdots)$. Then we have to see what kind of relationship.

	x_1	x_2	x_3	x_4	x_5	
t_0	-	-	-	-	-	$x_1 = f(x_2, x_3, x_4, x_5)$
:	-	-	-	-	-	
t_n						

If the dependent var is a function of a single variable then the regression is a simple regression.

It deals with 1 independent variable.

If we have multiple no of independent variable then it is known as multiple regression. Likewise, the relationship b/w dependent & independent var then it is a linear regression model.

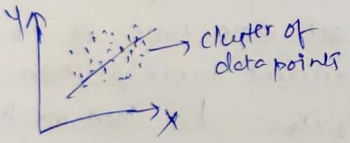
If the relationship is not linear then it is non-linear regression model.

All these are estimated based on data points.

Simple Regression.

2 variates < independent variate
Dependent "

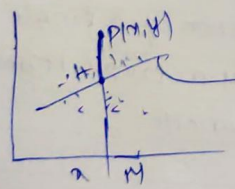
X Y



$Y = f(x)$

If we try to explain in linear then we can draw a line then most of the data points are nearby line/ line passes through most of data point

we need to draw a line such that a point should be on line of the distance b/w the point & line is min.



$y = f(x) = a+bx$.

$H = (x, a+bx)$ $P_H = [y - (a+bx)]$ the -ve root ∠0

if P is below line $P_H = -[(a+bx) - y]$ → 0 ∠

It depends on the point whether above or below line

So, we will take square factor of this to make it +ve

$[y - (a+bx)]^2 = E$ (let)

E is square distance of the point from the straight line.

It is principle of .

We try to minimize the distance b/w point & line we try to minimize 'E'.

To minimize E the necessary condition is first derivative should be 0. So, E has to be differentiated w.r.t 2 variables (slope & intercept)

$\frac{\partial E}{\partial a} = 0$ $\frac{\partial E}{\partial b} = 0$ $a+bx = y$.

b → slope a → intercept

first order derivatives

These 2 First order derivatives will give rise to normal equations in terms of data points

Regression co-efficients.

05/11/21

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \qquad a = \bar{y} - b\cdot\bar{x}$$

a and b are regression coefficients

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

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

If points x, y and $y = a+bx$ → reg eqn (reg. coefficients)

when these equations are solved then we can get regression eqn ⇒ $y = a+bx$.