Statistics - Mahesh Anand.

Day-1

Categorical columns - interms of properties

Continuous columns - interms of mean & S.D.

Outlier -

$abs|x - \bar{x}| > 3\sigma$ is considered as outlies - recommended.

treatment methods

i) Remove

ii) Replace with median

iii) Resampling

Extreme values - (outliers which are imp data)

treatment -

i) log transform (if skewed)

if very big range, then log transform compresses it.

ii) exponential - (if data is very narrow)

if the data is is a very narrow range then exponential transform spreads it.

Descriptive stats –

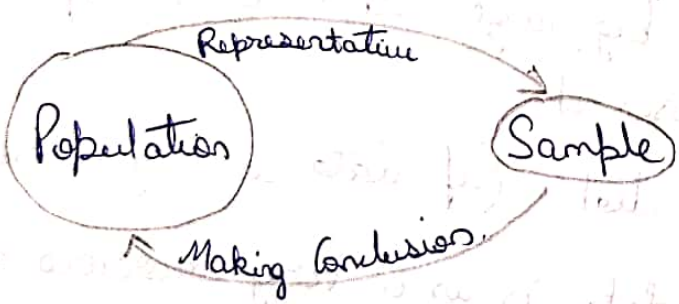
gives some idea about the samples collected. it doesn't talk about the population. → measures of central tendency
→ measures of variance.

Inferential stats –

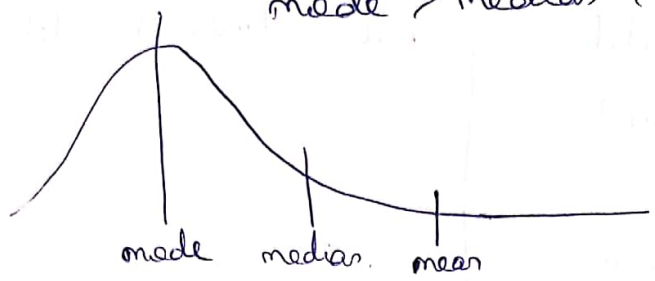
we are doing one step more than descriptive stats and are able to comment on population as the sample as so-and-so characteristics. Makes inference out of the samples collected. From the samples we are infering the population.

t - test - for test of mean (continuous)

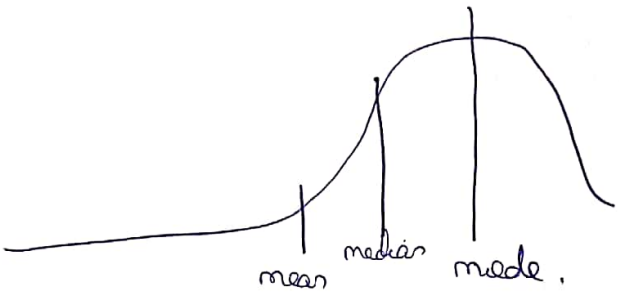
z - test - for test of proportion (categorical)



mode > median > mean.

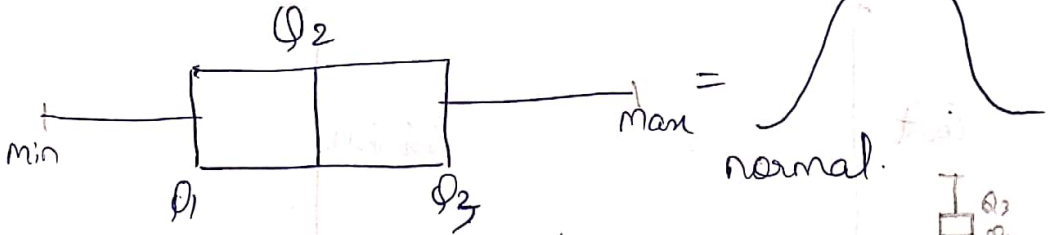


Right skewed



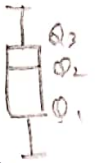
Left skewed.

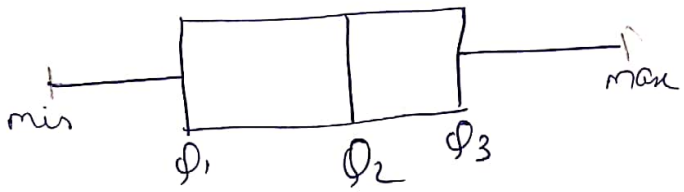
mean moves outwards because of extreme values.



=

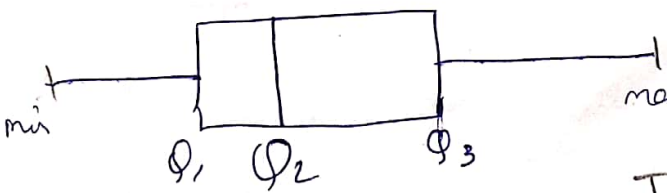
normal.





left skewed.

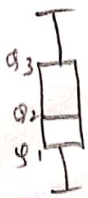
$Q_2 - Q_1 > Q_3 - Q_2$



Right skewed.

~~$Q_1 < Q_3$~~

$Q_2 - Q_1 < Q_3 - Q_2$



Linear model (Parametric)

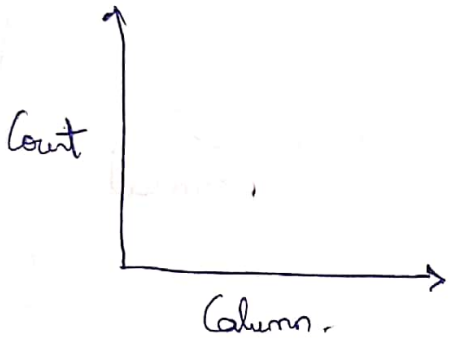
Constraint-

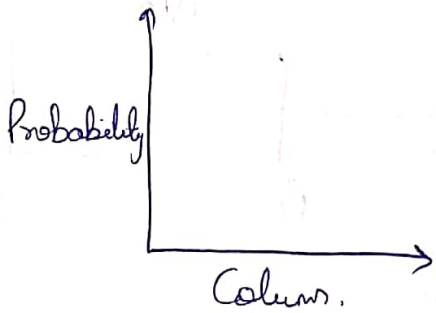
-> Samples are random.

-> Attributes follow normal distribution

Non-linear models

Histogram vs Probability Density function.





* Plot of all the probabilities is probability distribution

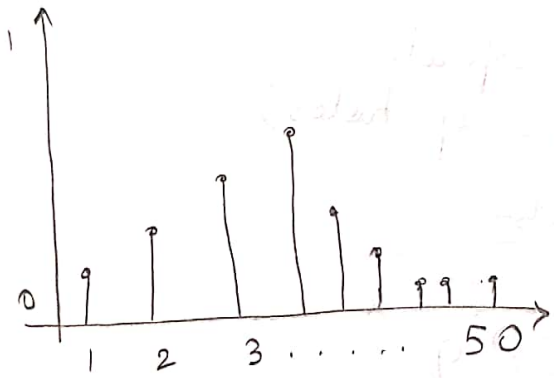
Probability -

1) classical probability → Tossing a coin, dice, Cards etc (already proved.)
2) Empirical prob ⟶ most used → based on historical data.
3) Subjective prob. (not used) → depends on person to person. → domain expert.

* As the individual probability increases, the probability decreases.

Probability distributions

① 5% defective
n = 50 t-shirt.



* Defect
no defect
* more outcomes
Binomial distribution

Discrete probability

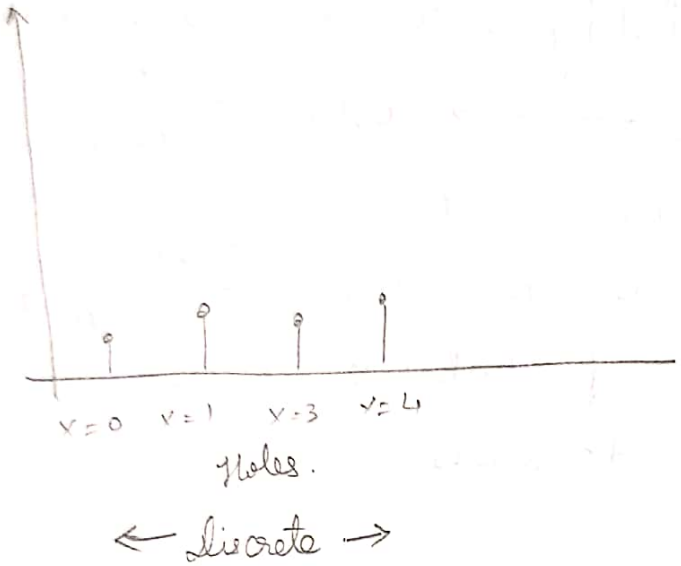
Cannot say 1.5 t shirt ~~difficult~~ defect.

Only whole no's.

② Empirical data

→ avg defects is '7'

n = 50.



← Discrete →

many outcomes

Poissons

distribution

* Difference b/w ① & ②.

① is Binomial distribution as limited no. of outcomes (defect, no defect)

② is Poisson's distribution as the outcomes are infinite (no. of holes)

For Binomial distrib-

$$P(x) = nC_x P^x q^{n-x}$$

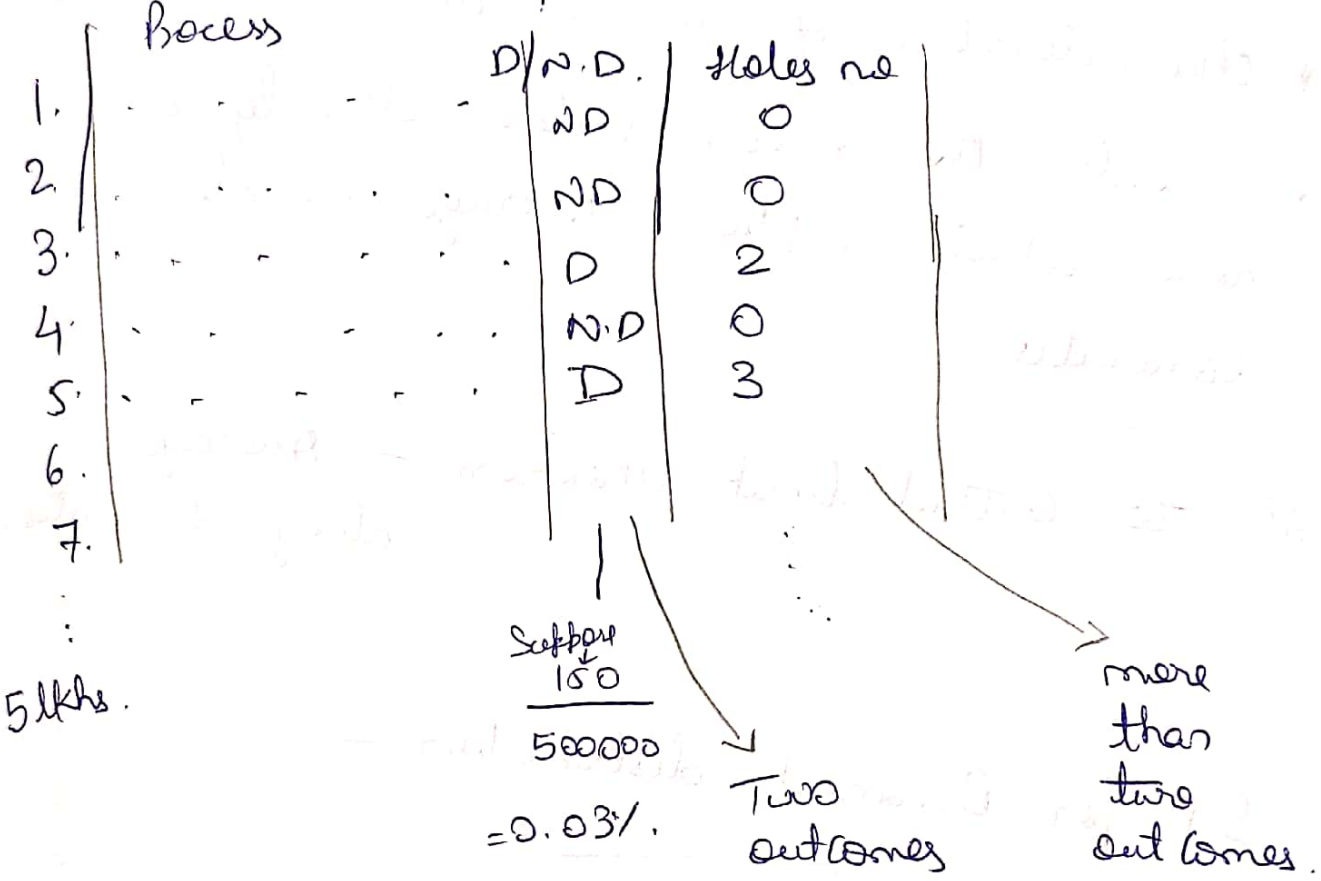
$x = 1, n = 50, P = 0.05$, so $q = 0.95$

$$P(x=1) = 50C_1 (0.05)^1 (0.95)^{49}$$

for Poissons distribution –

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

e.g- T. shirt manufacturing data.



For discrete values we use Probability mass function (PMF) — Binomial, Poisson.

and for Continuous values we use Probability density function (PDF) — normal

Probability density function -

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

* Characteristics of normal distribution -

→ Towards the mean value density is more, which implies, Average always dominates.

A/C to central limit theorem - Average always dominates.

S.D for Binomial distribution -

$$S.D = \sqrt{np(1-p)}$$

$np = 50 \times 0.15 = 7.5$ 7 ~ 8 will default.

$$S.D = \sqrt{7.5(1-0.15)}$$

$$= \sqrt{6.375}$$

$$S.D. = \pm 2.52$$

Poisson distribution

$$S.D = \sqrt{\lambda}.$$

Normal distribution -

$$\text{mean.} = \frac{\sum_{i=1}^{n} x}{n.}$$

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

$\bar{x}$ = 24 years

σ = 5 year

- spread is more

$\bar{x}$ = 23.6

σ = 2 yr.

- spread is less.

Q On an average 6 customers every 1 min at a bank during 62 hours.

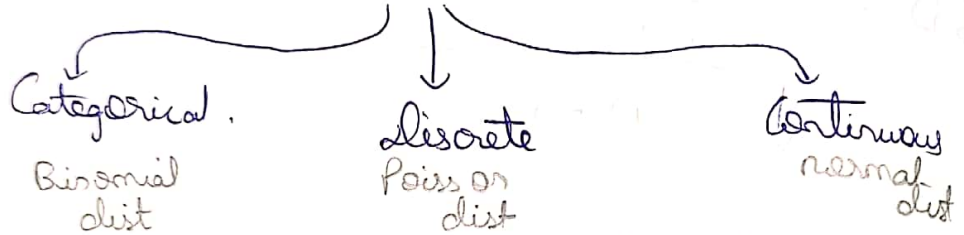
$\lambda = 6$.

what is the prob exactly 4 customers are going at a partic minute

i.e $x = 4$ & $x = 3$.

Day - 2

Random Variable



Categorical - This is also discrete in nature but it has generally only 2 Categories - it is represented as Binomial

e.g - 0/1, M/F, D/ND), Y/N etc, Tossing Coin.

Discrete - This is multi category (more than 2)

e.g - no. of rooms in house. It is represented as Poissons distribution.

eg - Throwing dice.

Continuous - This is of continuous values and follows normal distribution

e.g - Price of house etc.

→ In normal distribution if we Calculate probability for one Single point it will be 0.

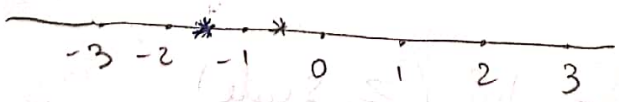
because as it is a Continuous dist. we are Calculating area under the Curve. because at one Single point we Cannot integrate.

$$P(x = 45g) = 0.$$ for normal dist.

→ $$\boxed{Z\ score = \frac{x - \bar{x}}{S.D(x)}}$$

all the values are mapped b/w -3 to 3.

→ In normal distribution, the values above the average → right
below the " → left
as the 0 is mean is normal dist.



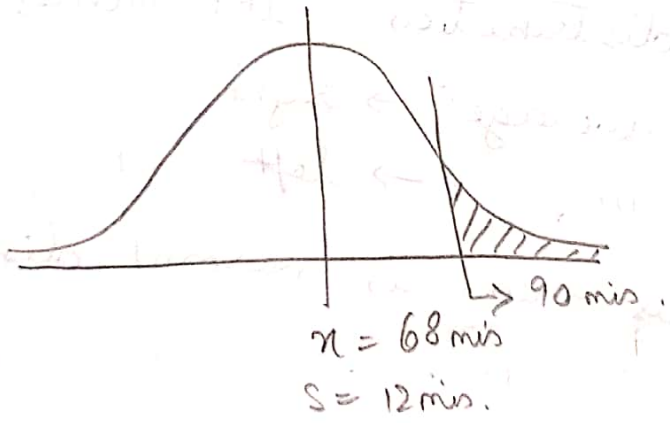
* Stadrdard normal distribution -
is a ritless probability distribution of the Continuous variable.

why? → we Can use Same Scale to analyse different Continuous variables.

* The norm fun. always Calculates area on the L.H.S.

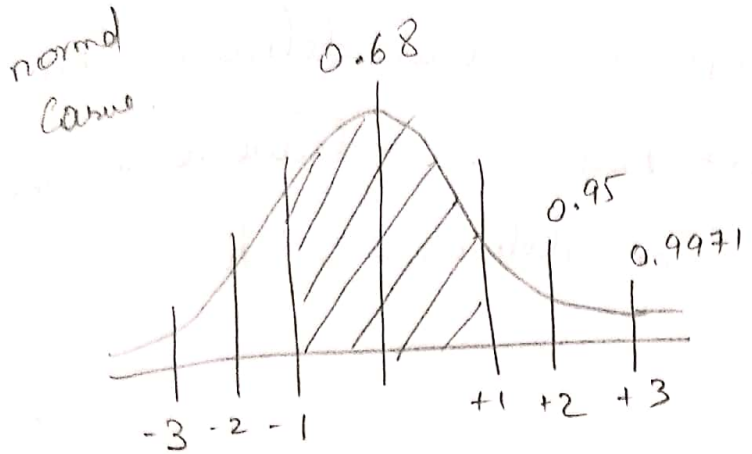
Q - a Survey Conducted, found average 68 mins on phone, S.d is 12 mins.

a) what proportion of S.P users are spend more than 90 mins?



first Convert 90 to z Scale.

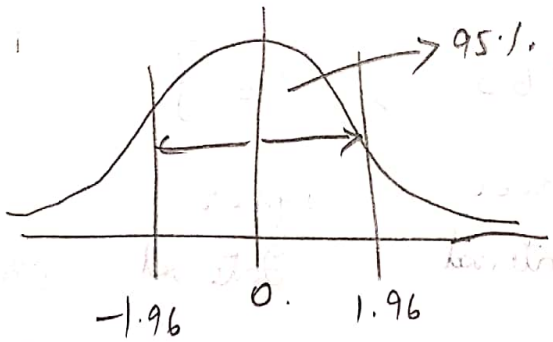
$$= \frac{90-68}{12} = 1.8 \text{ (z Scale)}$$



* Information is Contained in a spread.
  - 3σ to 3σ.

as the spread of the interval increases our Confidence that the inference is Correct for the ~~Sample~~ ~~fro~~ population from Sample will also increase.

Standard error = $\frac{S}{\sqrt{n}}$

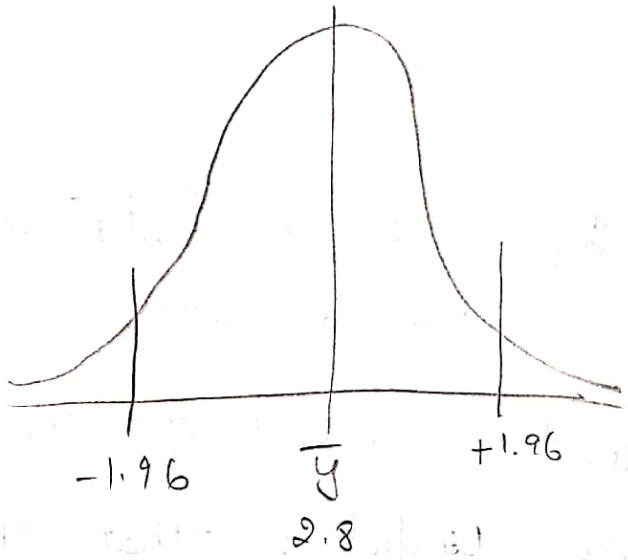


$$\bar{y} \pm \frac{1.96 \times S}{\sqrt{n}}$$

Q- Courier Company, mean delivery less than 3 ~~years~~ hrs. 50 deliveries random, determines mean delivery = 2.8 hrs,

$s = .6$ hrs.

$\alpha = 95\%$



$n = 50$

$\bar{y} = 2.8$ hrs.

$s = 0.6$ hrs.

$$C.I = \bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

$$= \left[2.8 - 1.96 \times \frac{(0.6)}{\sqrt{50}} \;,\; 2.8 + 1.96 \times \frac{(0.6)}{\sqrt{50}}\right]$$

$$= [2.63 \;,\; 2.96]$$

Lower Interval — upper Interval.

→ The interval from 2.634 to 2.966 form a 95% Confidence interval for mean (μ). In other words we are 95% Confident that the average delivery time lies between 2.63 & 2.96 hrs.

Confidence Coefficient	Value of $\alpha/2$	Area in Table 1 $1-\alpha/2$	Corresponding z-value $z_{\alpha/2}$
0.90	0.05	0.95	1.645
0.95	0.025	0.975	1.96
0.98	0.01	0.99	2.33
0.99	0.005	0.995	2.58

90%. 95%. 98%. 99%.
1.645 1.96 2.33 2.58.

→ when Sample Size increase, it becomes a very good representation of the population.

← Instead of $\alpha = 5\%$. use $\alpha = 1\%$.

$$= \left[2.8 - 1.96\left[2.58 \times \frac{(0.6)}{\sqrt{50}}, 2.8 - 2.58 \times \frac{(0.6)}{\sqrt{50}}\right]\right.$$

$$= [2.54, 3.01]$$

1 min = 0.40
120 min = ? x

$x = 80$ min.

1 hr 20 min.

Session - 3

$n = 100$

$\bar{y} / \bar{x} = 4.5$

$S = 1.2$.

$$C.I = \bar{y} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$= \left[4.5 + 1.96 \times \frac{1.2}{\sqrt{100}}, \quad 4.5 - 1.96 \times \frac{1.2}{\sqrt{100}}\right]$$

$$= [4.735, 4.2648]$$

If $n > 30$

Then Sample S.D $\sim$ Populaties S.D.

To Calculate C.I -

LCI, UCI = Stats.norm.interval(Ci, loc = xavg, Scale = S)

Ci = 0.95

xavg = mean.

$S = \frac{\sigma}{\sqrt{n}}$ S.E Standard error.

t - distribution -

→ adjusted Std. dist, used for hypothesis testing

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

Q- Hindustan Pencils; Pencils mean length 172 mm, S.d 0.02mm.

Sample taken., $n = 100$, $\bar{x} = 170$ mm

$\alpha = 95\%$.

$x = 172$ mm.

$$C.I = \bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$= 170 \pm 1.96 \times \frac{0.02}{\sqrt{100}}$$

$$= \left[170 - 1.96 \times \frac{0.02}{\sqrt{100}}, \; 170 + 1.96 \times \frac{0.02}{\sqrt{100}}\right]$$

$$= [169.99, 170.003]$$

as the ~~S.D~~ true mean 172 is not in the Sample Confidence Interval, the Sample is not a inferential for population this means, The S.D is little higher than 0.02,

Hypothesis testing -

→ If the claim is alternate hypothesis is on Continuous variable, it will be on mean.

If the claim is on discrete variable (Catago) it will be on the proportion.

Continuous → mean.
Categorical → Proportion.

$H_a : \mu_0 < 3hrs$ (Couries Company eg)

→ $H_0 : \mu_0 \geq 3hr.$

as a data Sci always Contradict

↓ Soft inequality

One tailed test.

$< >$ - Hard inequality

$\leq \geq$ - Soft inequality (for null hypothesis).

Modelling Hypothesis → for modelling.

In H_0: $=$
H_a: $\neq$

we only deal with $=$ or $\neq$

$=$ or $\neq$ two tailed test.

$><$ → one tailed test

In inferential the soft inequality must be given to null hypothesis whereas in modelling hypothesis, the equality should be given to null hyp.

Inferential (one tailed test)	H_0 : $\mu_D \geq 3hr$	
	H_a : $\mu_D < 3hr$.	
Modeling	H_0 : $\mu_1 = \mu_2$	(Two tailed test)
	H_a : $\mu_1 \neq \mu_2$	

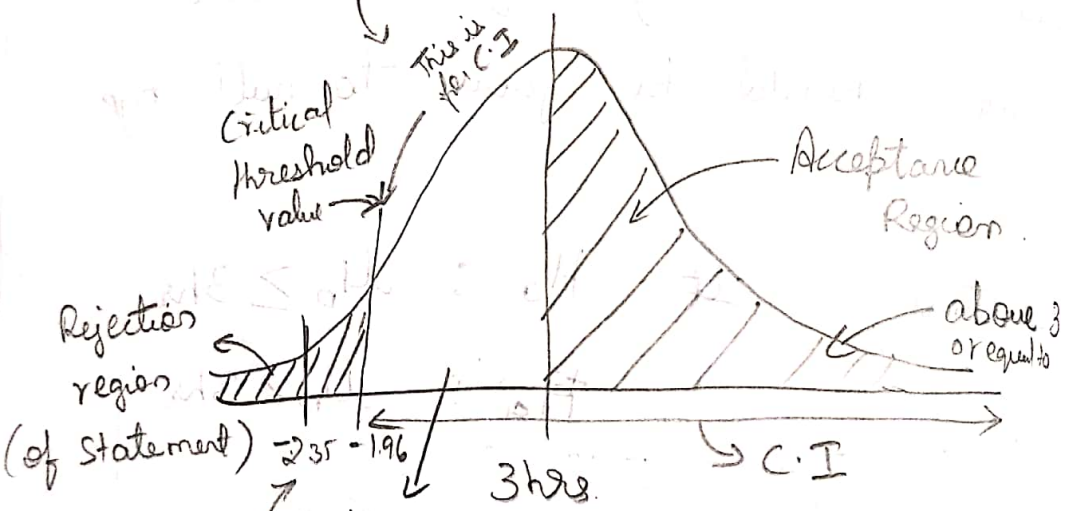
* always start with alternate hypo

→ Q Courier problem, where they had stated that the mean delivery time is less than 3 hrs.

$H_a : \mu_D < 3\text{hrs}.$

$H_0 : \mu_D \geq 3\text{hrs}.$

↳ graphically.



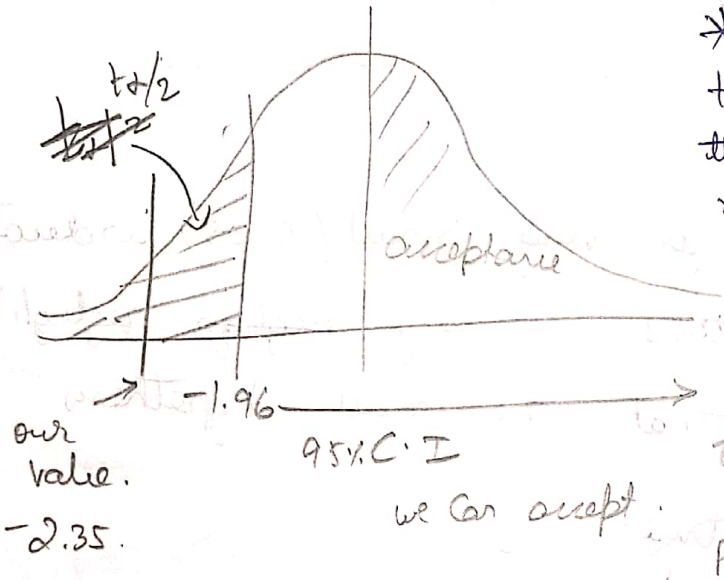
* One Sample t test, or left tail test.

(Critical Value) $\Rightarrow t_{stat} = \frac{\bar{y} - \mu}{s/\sqrt{n}}$

$= \frac{2.8 - 3}{0.6/\sqrt{50}} = -2.35$

→ The left H.S side is contradicting for the null hypothesis, but not the whole of L.H.S is contradicting, we are ready to accept till one point. i.e 95%. we are accepting. Only 5%. we are reserving for the rejection.

* For 95% C.I, minimum -1.96 the acceptance should lie away.



* Here as the t critical lies after the acceptance region, the null hypothesis is rejected.

t stat - distance on x-axis

Pvalue - area under the curve.

If t stat lies within the -1.96 critical value is t stat. and the area w.r.t to critical value it is the p-value.

t stat → distance on x axis
Pvalue → area under curve.

Rule for Rejecting null hypothesis.

	95 %	99 %
tstat	tstat >1.96	tstat >2.58
Pvalue	Pvalue <0.05	Pvalue <0.01

---

p-value is a probability of null hypothesis being true

---

* Pvalue is very Small (0.05) indicates that there is very ~~negligible~~ negligible chance that the null hypothesis being true.

* 5% error, the rule or region is relaxed as compared to 1% error which is very strict.

at what level of Confidence → client will provide.

Continued –

Delivery Company example –

Calculate pvalue for tstat Score.

Stats.norm.cdf(-2.35) = 0.0092. → Pvalue.

This function uses normal dist.

which is less than 0.05.

For accurate Pvalue, use t dist.

∴ There is less than 5% chances for the null hypothesis being true. Here the remaining 95% supports the claim made by the Courier Company.

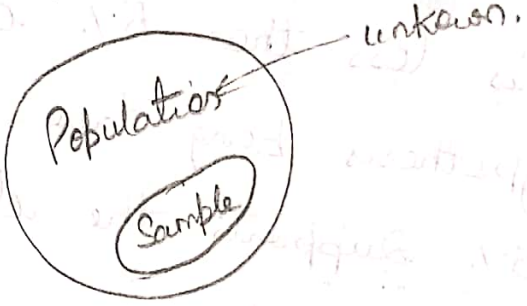
∴ We fail to accept the null hypothesis.

So a/c to the alternate hypothesis, the claim made by the Courier Company is true i.e they make all their deliveries within 3 hours.

* To ~~as~~ get accurate p-value, we should use the t-distribution. In the previous case we are using the normal distribution.

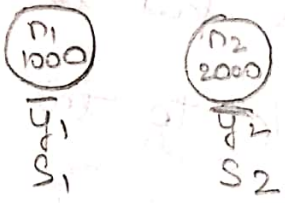
t test → test of mean.

One Sample →



Two Sampl t test →

* Comparing two groups.



$$t_{stat} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

when

One Sample t test has 3 Case –

1) Left tail $<$
2) Right tail $>$
3) Two tail $=$

Two Sample t test has only one –

1) Equal to $=$
2) Not equal to $\neq$

In two Sample t test –

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

→ Functions for t-test →

& for ~~t~~ critical value -

→ t test_1 Samp.

→ we need Samples so we have to generate Samples with random values

* randn → will ger (0 to 1), $\mu = 0$, SD = 1

* randint → integers.

* If we want to generate a Sample with required S.D and mean we just have to reverse the Z Sale

$$Z_{sale} = \frac{x - \bar{x}}{S.D(x)}$$

$$Z_{sale} \times S.D(x) + \bar{x} = x.$$

S.D of x.

for delivery HR example →

$\bar{x}$ = ~~0.6~~ 2.8 $\mu = 3$

S.D = 0.6.

n = 50

For generated Sample in jupyter. $\mu = 3$

$\bar{x}$ = 2.81

S = 0.61.

Python funtion to Cal. t statistic ∴ Sample t test

```
t_test_stat, pval = ttest_1samp(S_std, 3)
```

* This will give us t statistics and p-value.

Sample

Population mean.

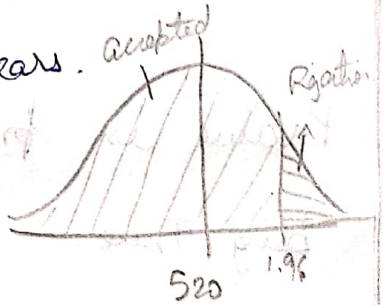
Q Soyabean e.g -

$\bar{y}$ = ~~520~~ 573 $\mu = 520$.

S = 124

n = 70.

Company is Claiming yield is greater than 520. Compared to last 2 years.



$H_a : \mu > 250$

$H_0 : \mu \leq 250$

$t_{stat} = 3.41$ $P_{val} = 0.001$

∴ as tstat > 1.96 and Pval < 0.05

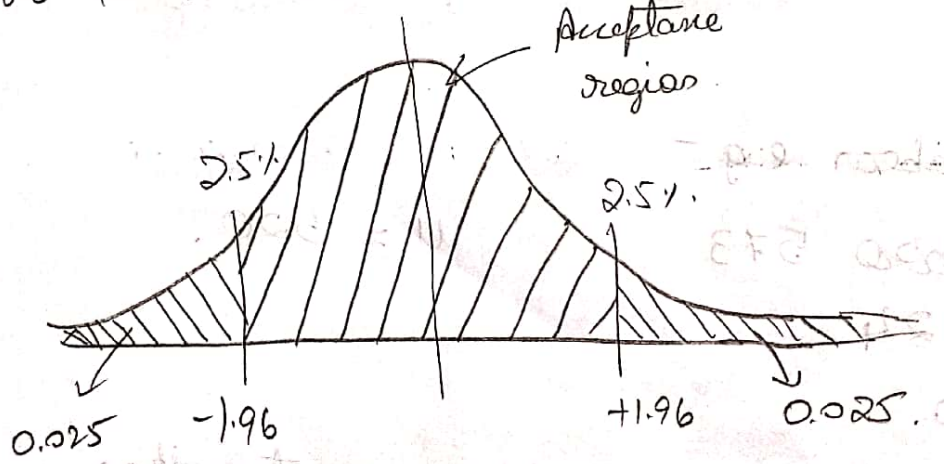
we reject null hypothesis.

error — To reject Ho →

	Pval. α Pvalue <	$\alpha/2$	tstat. >
90%.	0.1	0.05	1.64
95%.	0.05	0.025	1.96
98%.	0.02	0.01	2.33
99%.	0.01	0.005	2.58.

Remember

Two tailed test-



Pvalue is probability of null hypothesis being True.

* If mean and median values are lying closer, the data is normal.

Q class data -

$n = 30$

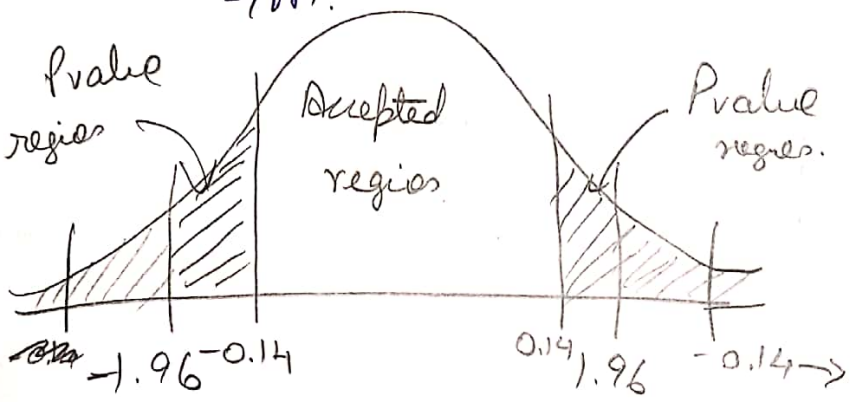
$\bar{y} = 24.95$

$s = 2.48.$

$H_0: \mu = 25yrs$

$H_a: \mu \neq 25yrs.$

$$t\text{stat} = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{24.95 - 25}{2.48/\sqrt{30}} = -0.1424$$



Bob ~~This~~ is is Tuf not.

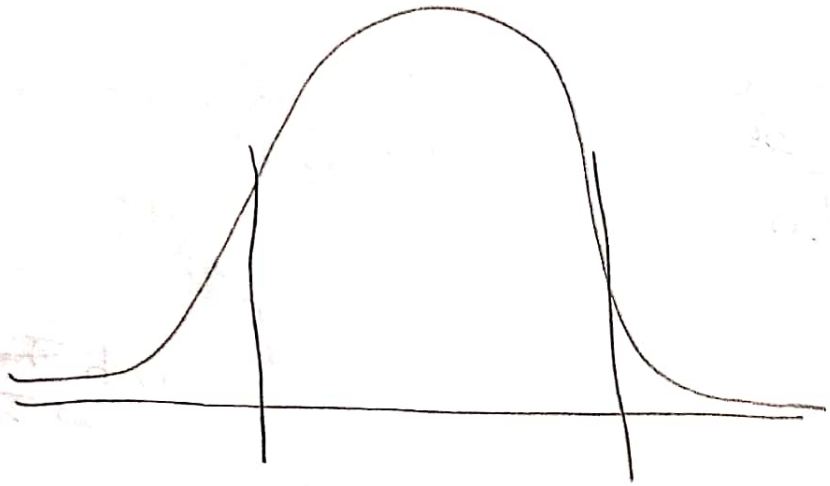
Here $0.14 < 1.96$ and Pvalue is 0.88.

This means that There is 88% chance of Ho being true.

So we accepte Ho.

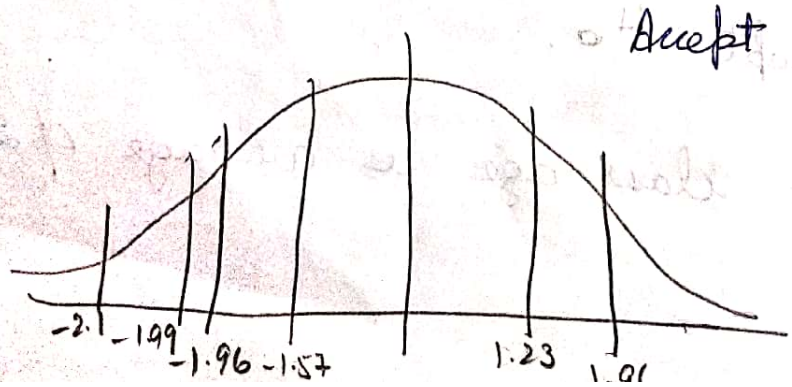
Therfore, the class age is average of 25yrs.

* In the two sided t-test, we have to check the error on both the sides as we are asking wether it is equal to or not equal to.

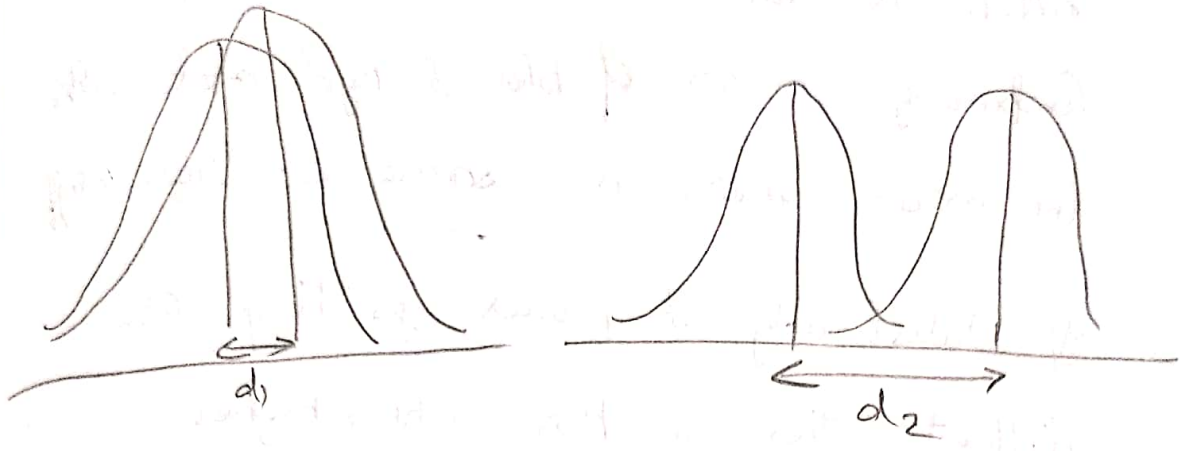


Exercise -

tstat	Pval	Null Hypothesis
-1.57	0.143	Accept Ho (P<0.05
-1.99	0.01	Reject Ho. (P≥0.05)
-2.1	0.0026	Reject Ho. (P<0.05
1.23	0.438	~~Reject Ho. (P≥0.05~~ Accept Ho.



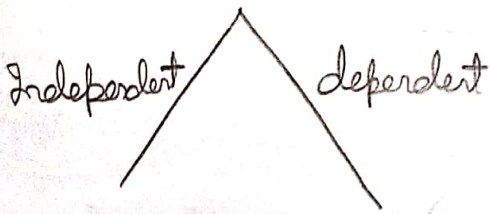
Two Sample t - test -



→ we are gona check wether the mean of two groups are same or significantly differed

→ If the gap between the 2 means are for 95% C.I ie if the gap is more than 1.96 then the means are ~~almost same~~ significantly different.

→ For 2 Sample t test -



If the test of Blore and hyderabad batch is taken and the ~~mean~~ t Comparing mean of blore & hyd and with the mean score is same or Signi diff.

→ If statistically it proves yes they are different, then we know who's higher.

$$H_0 : \mu_{Blor} = \mu_{hyd}$$

$$H_1 : \mu_{Blr} \neq \mu_{hyd}.$$

If Ho accepted, the mean is same, if rejected, then we will know which batch is performing better.

This is an example of <u>is dependent two sample t test.</u>

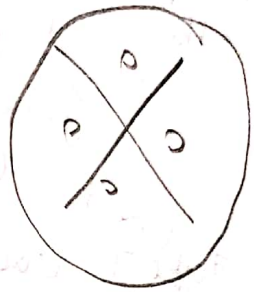
ttest_ind (G1, G2)
↓
Groups.

If A test end of 4 day and found average score is 68 and next week after revision, test taker and mean found out 72.

This is an example of ** dependent two sample t test

*Before and after a process *

ttest_rel (G1, G2)
↳ Group array



* As we are infering from the samples or the population, to play it safe, we never say ~~we never say~~ accept the Ho we always use we failed to reject null hypothesis.

IMP →

* WE REJECT THE NULL HYPOTHESIS.
OR
* WE FAIL TO REJECT THE NULL HYPOTHESIS.

Day 3 stats -

In Reality.

Hypothesis test decision	H_0 : True	H_0 : False
Reject H_0.	Type I - error α - error (wrong decision)	$1-\alpha$ (Correct decision)
Accept H_0.	(Correct decision) $1-\beta$ (power of test)	Type II error β error (wrong decision due to many things) ↓ Costly error.

Type I error (α) - missing out a best Candidate due to the interview process not being good. Cos

Type II error (β) - A worst / bad Candidate gets selected due to bad interview process. (Costly) (risky).

Type I error - when we reject the null hypothesis when it is true

Type II error - when we ~~accept~~ do not reject the null hypothesis when it is false.

e.g -

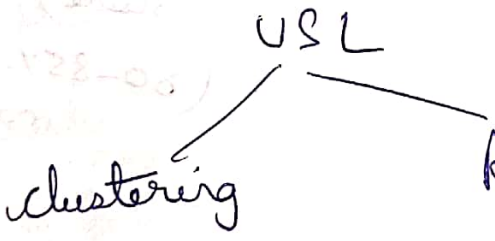
1) Really missing out ona genuine good person and breaking up with her (Type I error)

2) Accepting a really bad person and continuing the relationship (Type II error) (very bad).

---

* Supervised learning model - in this, the outcome variable is already present is the data.

* If there is no outcome data, all are indepen dent variables, in such data we cannot perform supervised modelling thats why we do unsupervised learning.



---

* If they reject the null hypothesis they can be correct or make Type I error.

* If they do not reject the null hypothesis they can be correct or make a Type II error.

Type I
True but rejected

Type II
False but accept

* Healthy → 400
* disease → 100

& classification model as 2 outcomes.

while labelling

	label
abnormal case	→ 1
normal case	→ 0

Actual	Predicted 0	Predicted 1
Healthy 0	(1-α) 380	20 (α) → They are healthy but it is predicted disease
400		
disease 1	(β) 40	(1-β) 60 → model Sensitivity.
100	↓ This is people having disease is actual but model predicts healthy (very dangerous)	↓ 60/100 = 60% not happy. (very imp score) (80-85% max)

model Specificity.

Specificity - at what rate the model is predicting healthy as healthy.

Sensitivity - at what rate the model predicts disease as disease (1st priority)

ROC - Reciever Operating Characteristics

*

	0	1
0	TN	FP → α.
1	FN	TP
	β	

True Negative
False Positive
False negative.
True Positive

TP → Sensitivity — Power of test.

TN → Specificity.

$$\text{Accuracy} = \frac{(1-\alpha) + (1-\beta)}{\text{Total records.}}$$

* Even though the overall accuracy of the model can be good, that doesn't mean that the model is good, the sensitivity is the of the model is very goo important.

Here → only $\frac{60}{100} = 60\%$ Sensitivity is bear.

but overall accuracy = $\frac{380 + 60}{500} = \frac{440}{500}$

$= 88\%$.

Still this model is not performing well for this data.

Statistical test.

- Test of mean (t stat)
  - One Sample t test.
  - Two Sample t test.
  - ANOVA
- Test of properties. (Z stat)
  - one Sample properties
  - Two Sample properties.
  - Chi Sq Test.

$$t_{stat} = \frac{\bar{y} - \mu}{S/\sqrt{n}}$$

$$\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$H_a : \mu > 81$

$H_o : \mu \leq 81$

tstat is reffered to as Z test if the population S.D is given to us.

$$t_{stat} = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$$

(population S.D.)

One Sample t test → $H_a : \mu > 81$, $H_o : \mu \leq 81$

two Sample t test → $H_o : \mu_{ind} = \mu_{sing}$, $H_a : \mu_{ind} \neq \mu_{sing}$

$P < 0.05$

reject H_o. There is signif diff in Salary.

One Sample proportion test → H_a : $P_{diabetic} > 25\%$.
H_o : $P_{db} \leq 25\%$.

(1D)

Two Sample proportion test → proportion of girls in blore to ~~boys~~ girls is equal to chennai.

(2D)

chi Square test (multi dim) → prop of girls in blore is equal to
prop " " in chennai is equal to
" " " in hyderabad.

Two Sample 't' test
↙ SameSample
↳ Paired (before & after Same Sample) n=Same
↳ Unpaired (Independent) (ttest_ind.)

---

* If data strictly follows normal distribution go with parametric test.
Otherwise go with non parametric test.

---

test for normality.

Shapiro test –

```
From Scipy.stats import shapiro.
```

H_0 : Data = normal

H_1 : Data ≠ normal.

= If e.g – P = 0.837

fails to reject null hypothesis as $P > 0.05$

∴ The data is normal.

we can continue with parametric test.

* Even If one group is not normal, then we have to go with non parametric test

non parametric test → Two Sample 't' test.

- paired → Wilcoxon
- unpaired (Independent) → mannwhitneyu.

```
from Scipy.stats import mannwhitneyu
                        wilcoxon.
```

* For Two Sample t test,

first with shapiro test check for normality. If the data is normally distributed, go with parametric tests depending on what kind of sample -

If data = normally dist.

paired t test → t.test_rel.

Unpaired t test → t.test_ind

Even if one of the group is not following normal distribution, go with non-parametric test.

If data not normally dist.

paired t test → wilcoxon.

unpaired t test → mannwhitneyu.

Ses - 2

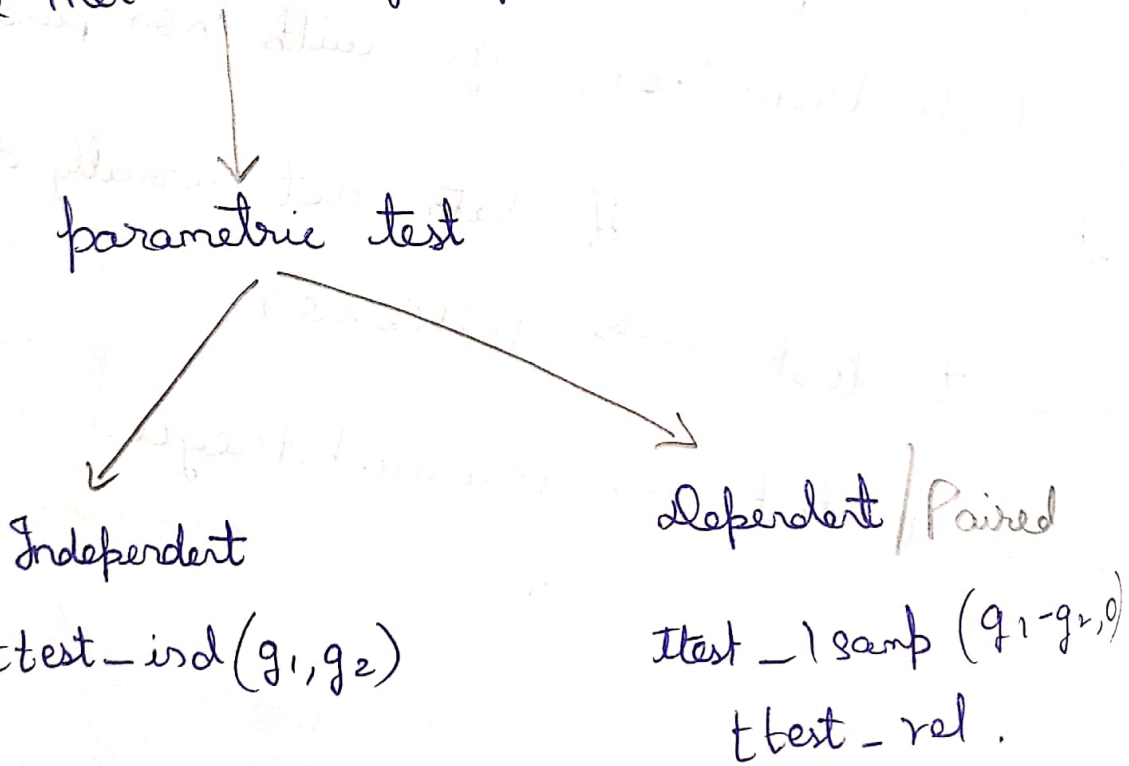
* Start with assumption H_0 is True.

For Two Sample test -

Conditions to be met -

1) Random variable (random Sample)
2) Data is normal. (Shapiro test)
3) Test of variance $\sigma_{g_1}^2 = \sigma_{g_2}^2$

only if the above three Conditions are met will go for parametric test.



If any one Condition fails, we do non parametric

non - parametric test

Independent (mannwhitneyu)

Dependent (upaired) (wilcoxon)

* If any one of the 2 Conditions fails.

→ If data is normal then to check the variance, then do 'Levene test'.

→ If data is not normal, just to gain extra insight, we do 'Bartlett test'.

Test of Variance

Data is normal.

Data is not normal.

'Levene'

Bartlett

$H_0: \sigma_{g_1}^2 = \sigma_{g_2}^2$

$H_a: \sigma_{g_1}^2 \neq \sigma_{g_2}^2$

← null hypothesis & alt hyp for variance test.

H_0:

(HR.txt), names = [' ' ' ' ']

[ft.csv, index_col = 0])

Test of Normality -

H_0: Data = Normal

H_a: Data ≠ Normal.

Test of Variance - levene / Bartlett

$H_0: \sigma_{g_1}^2 = \sigma_{g_2}^2$

$H_a: \sigma_{g_1}^2 \neq \sigma_{g_2}^2$

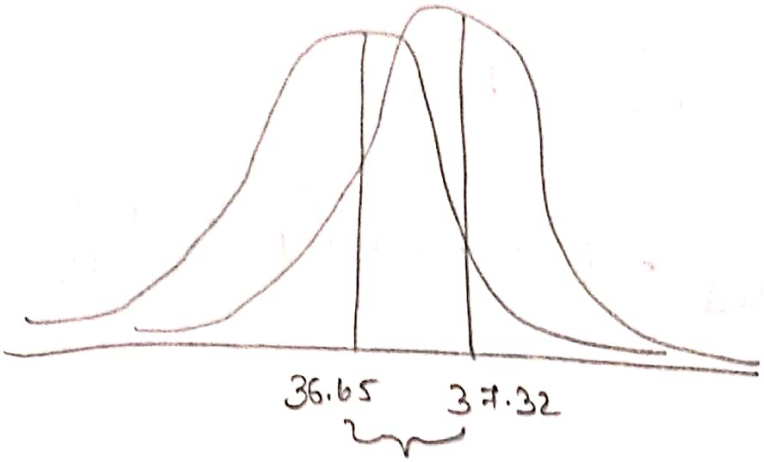
Test of mean - ttest_ind / mannwhitneyu

$H_0: \mu_{g_1} = \mu_{g_2}$

$H_a: \mu_{g_1} \neq \mu_{g_2}$.

Average age of male = 36.65

" " female = 37.32



acceptable range should be within 1.96.

37.32 - 36.65 = 0.67.

Session 3 -

* We only reject null hypothesis or fail to reject null hypothesis.

95%	$P < 0.05$	Reject null hypothesis
99%	$P < 0.01$	Reject Ho.

Proportion test -

$$Z_{data} = \frac{P_1 - P_2}{\sqrt{P_{pooled} * (1 - P_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$P_{pooled} = \frac{x_1 + x_2}{n_1 + n_2}$$

Q - asses the proportion of female employees attrition is equal to male employee attrition.

$H_0 : P_{FA} = P_{MA}$

$H_a : P_{FA} \neq P_{MA}$.

more females are leaving or more male are leaving. or irresp of gender they are are equal.

* Attrition is not dependent on gender, its independent.

↳ statement.

* The moment there are no numerical columns in the test, t test is ruled out.

P_1 - Prop of male emp. leaving Comp.

P_2 - " " female " " "

$$P_1 = \frac{150}{882}$$

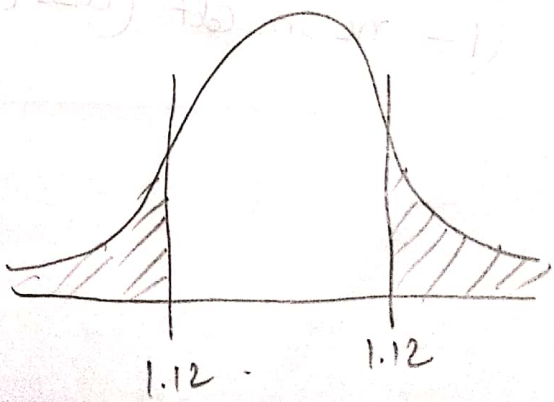
$$P_2 = \frac{87}{588}.$$

$$P_{pool} = \frac{(150+87)}{(588+882)}$$

$$= 0.16.$$

$$Z_{data} = \frac{P_1 - P_2}{\sqrt{P_{pool} * (1-P_{pool}) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z_{data} = 1.129$$



As this is done manually we are use the normal distributios to get an aproximate p value. later on we have built in fun to calc the same

$$(1 - \text{norm.cdf}(z)) * 2$$ for bot sides.

* So for this we get p value = 0.258.

→ SO, for this we fail to reject the H_0.

Thus H_0 holds good for this.

which means, ~~the~~ There is no signi proportion of femal att is equal to prop of male att.

even if we get negative statistic value or negative Z data. —

$$(1 - \text{norm.cdf}(\text{abs}(z\text{data}))) * 2$$

* Verify Gender is having any influence is headache.

* In python ↑.

Built in functions →

CHI – SQUARE TEST for Goodness of Fit.

$$\chi^2_{data} = \sum \frac{(O-E)^2}{E}$$

$$\text{expected freq} = \frac{(\text{row total})(\text{Colums total})}{\text{grand total}}.$$

* Proportion test for more than 2 groups is done by chi square test.

E.g (data) → observed Count.

	Healthy	Mild	Severe	Total
Male Sample 1	410	340	250	1000
female. Sample 2	95	85	70	250
Total	505	425	320	1250

Females –

$P_{FH} = \frac{95}{250}$ (Female Healthy)

$= 0.38 = 38\%$

$P_{FM} = \frac{85}{250} = 0.34 = 34\%$ (mild)

$P_{FS} = \frac{70}{250} = 0.28 = 28\%$ (Sever)

males –

$P_{MH} = \frac{410}{1000} = 0.41 = 41\%$

$P_{MM} = \frac{340}{1000} = 0.34 = 34\%$

$P_{MS} = \frac{250}{1000} = 0.25 = 25\%$

Consider females –

* In order for the null hypothesis to be true ~~what~~ is the proportion required is called the expected Count.

If there is a large difference in the expected Count and original observed Count, then it indicates that the the alternate hypothesis holds good. as the difference is greater. If the original observed Count itself to needed not much Correct i e observed Count is almost equal to expected the (O-E) becomes 0, then Ho ~~Stan~~ holds good, there is no significant diff.

Expected Count.

$\frac{1000 \times 505}{1250} = 404$	$\frac{1000 \times 425}{1250} = 340$	$\frac{1000 \times 320}{1250} = 256$	1000
$\frac{250 \times 505}{1250} = 101$	$\frac{250 \times 425}{1250} = 85$	$\frac{250 \times 320}{1250} = 64$	250
505	425	320.	1250 → GT

* we are not modifying the original dat

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(404-410)^2}{404} + \frac{(340-340)^2}{340} + \frac{(256-250)^2}{256} +$$

$$\frac{(101-95)^2}{404} + \frac{(85-85)^2}{85} + \frac{(64-70)^2}{64}$$

$$\chi^2 = 1.1486.$$

* The p-value for Chi-Square dist is diff form z & t dist.

Built in function —

The data should be in Cross tab or D/F.

For 2- groups -

chi2_Contingency (DF/CT)

From Scipy.stats import chiSquare, chi2_Contingency.

n row = degrees of freedom. dof = n.of grop-1
= 3-1
= 2.

chivalue, P-value, dof, EC = chi2 – Contigency (Cross Tab)

dof → degrees of freedom

EC → expected Count

Type of headache.

	female	male.
aura	$\frac{1593}{3545}$ 44.9%.	$\frac{117}{607}$ 19.2%.
mixed	$\frac{291}{3545}$ 8.2%.	$\frac{166}{607}$ 27.3%.
No aura	$\frac{1661}{3545}$ 46.8%.	$\frac{324}{607}$ 53.3%.

This is the cross tab we have in python –

	female	male.
Aura	1593	117
mixed	291	166
No aura	1661	324
	3545	607

* males are highly sensitive with mixed type of migraine. as we can see females are only 8.2%. whereas male 27.3%..

* whereas females are ~~affe~~ sensitive to aura type female – 44.9%. male – 19.2%.

Questions -

Heart data - (2) disease

1) Gender influences heart ~~data~~ ? (2)

↳ Two Sample proportion test

2) Effect of exercise (2) with Heart disease (2)

↳ Two Sample proportion test.

3) Any sig difference in cholestral (continuous) wrt exercise (2)

↳ we have to mean of cholestrol exercise and non exercise people

↳ two Sample ~~is~~ t test.

HR data -

1) Any particular dept having high att rate, χ^2

2) Any particular education. χ^2

3) check the prop of M/F across all depts same?

df['Education'].value_counts() → one dimensional.

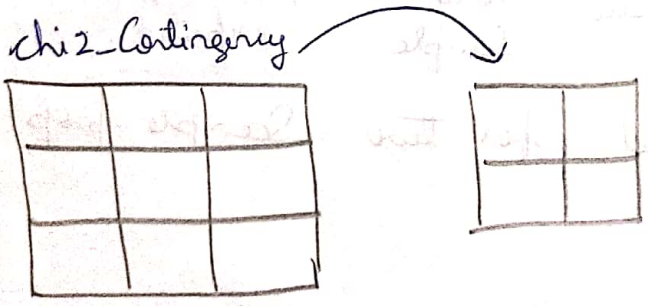
chisquare → for one dimensional.

→ It is going to tell wether the proportions of no's is same across all the group or different.

→ If we put chisquare is a single column suppose department, then it will tell us the proportion of depar the individual department is same or not.

54
50
56
52
54

It will tell the count is proportion or not.



One variable but multi Categ we doing with samples.

1 - Samp proportion test

25% exp percentage of Indian women are diabetic in India.

$H_0 = 0.25$

$H_a \neq 0.25$.

we perform Sampling $n = 500$

Dia = 200

H = 300.

Func-

```
from statsmodels.stats.Proportion
import proportions_ztest
```

Proportions_ztest(200, 500, 0.25)

- 200 → diabetic
- 500 → Total Sample
- 0.25 → percentage expected.

* This can be used for two Sample prop test also.

Counts - (x_1, x_2)

Obs - (n_1, n_2)

One Sample -

1) Proportion_z test (Actual Count, tot Sample, claim on population)
   → to check claim made on popula

Two Sample -

2) Proportion_z test (Obs, Count).
   Obs ← actual Count
   Count → yes, attacks, exercise
   Total Sample.

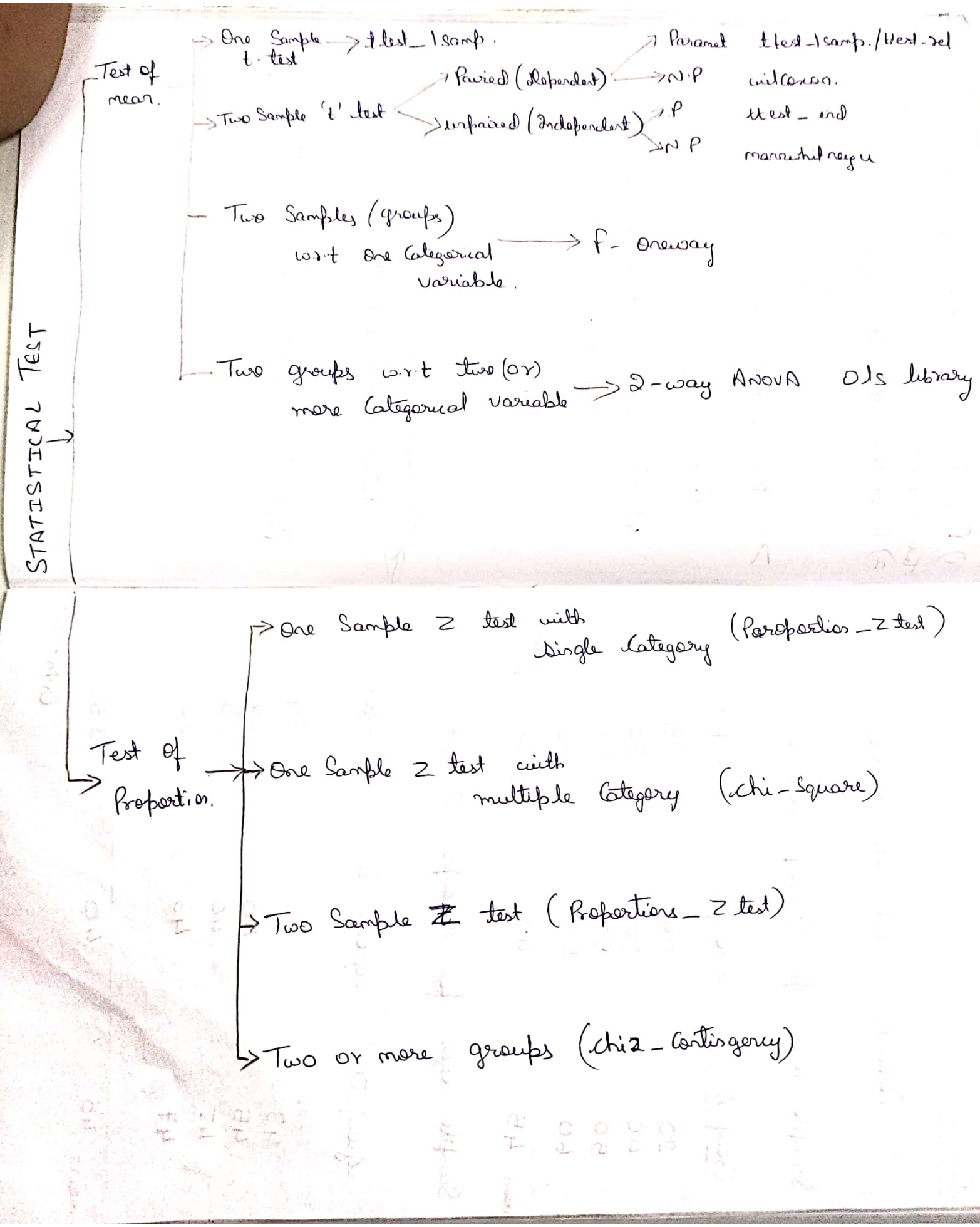
→ Sample, proportion test.

3) chi2_Contingency (CT)

4) chiSquare (Value Counts Single Category)
   (Sampling).

→ To check wether the sample is a representation of the population.

Proportion_ztest (affected, Total).



Day 4

Session 1 -

ANALYSIS OF VARIANCE

Sample ages for Groups A, B, C.

Group A	Group B	Group C
30	25	25
40	30	30
50	50	40
60	55	45
mean 45	40	35

Global mean = 40

Sample ages for Groups D, E, F

Group D	Group E	Group F
43	37	34
45	40	35
45	40	35
47	43	36
mean - 45	40	35.

G.M = 40

Mean Square Treatment (MSTR)

↳ Between sample variability.

Between sample variability / within Sample variability

$$MSTR = \frac{\sum n_i (\bar{x} - \bar{\bar{x}})^2}{k-1} \text{ or } \frac{SSTR}{dof.}$$

n_i = Sample no.

$k-1$ = dof.

for g-ABC -

H_0 : $\mu_{age}(A) = \mu_{age}(B) = \mu_{age}(C)$

H_a : " $\neq$ " $\neq$

for group - DEF

H_0 : $\mu_{age}(D) = \mu_{age}(E) = \mu_{age}(F)$

H_a : $\mu_{age}(D) \neq \mu_{age}(E) = \mu_{age}(F)$

$$MSTR = \frac{4(45-40)^2 + 4(40-40)^2 + 4(35-40)^2}{3-1}$$

$$= \frac{(100 + 0 + 25)}{2} \quad \frac{4(25) + 0 + 4(25)}{2}$$

$$MSTR = 100$$

$$f_{stat} = \frac{MSTR}{MSE} = \frac{\frac{SSTR}{df_1}}{\frac{SSE}{df_2}}$$

Variance of A = 166.66 (σ_A^2)

B = 216.66 (σ_B^2)

C = 83.33 (σ_C^2)

* while using `np.var(A, ddof=1)` always use ddof.

because it calculates for $\frac{\sum_{i=1}^{n} (v_i - \bar{x})^2}{n}$ whereas we need $\frac{\sum_{i=1}^{n} (v_i - \bar{x})^2}{n-1}$.

$$MSE = \frac{SSE}{df_2} = \frac{\sum (n_i - 1) S_i^2}{n_t - k}$$ → 3 groups × 4 samples per group = 12 samples.

* $(n_i - 1) = 3$

$S_i = \sigma_A^2 + \sigma_B^2 + \sigma_C^2$

$$= \frac{3 \times \sigma_A^2 + 3\sigma_B^2 + 3\sigma_C^2}{12 - 3}$$ → groups.

total Samples

$$= \frac{(3 \times 166.66) + (3 \times 216.66) + (3 \times 83.33)}{9}$$

$$= \frac{499.98 + 649.98 + 249.99}{9}$$

$$MSE = 155.55$$

$$\therefore \quad F_{stat} = \frac{MSTR}{MSE}$$

$$= \frac{100}{155.55}$$

$$\boxed{F_{stat} = 0.6428}$$

$P_{value} = 0.548.$

$P > 0.05$

∴ We fail to reject the null hypothesis.

$\therefore \mu_{age}(A) = \mu_{age}(B) = \mu_{age}(C)$

A A A B B B C C C

Python function -

* The data shud be in form of dataframe

model -

```
mod = ols('age ~ group', data = df).fit()
```

(y ~ x)

↑ always left side should be Continuous

for Anova table -

which numerical var a/c to which Cat variable

```
aov_table = sm.stats.anova_lm(mod, typ=1)
print(aov_table)
mod = ols('Continuous ~ Categorical', df).fit()
```

display o/p for anova model

For group- D, E, F

MSTR is same

$$MSTR = 100.$$

Variance of $D = 2.666 (\sigma_D^2)$

$E = 6.00 (\sigma_E^2)$

$F = 0.666 (\sigma_F^2)$

$$MSE = \frac{SSE}{df_2} = \frac{\sum (n_i - 1) S_i}{n_t - K.}$$

$$= \frac{3 \times \sigma_D^2 + 3 \times \sigma_E^2 + 3 \times \sigma_F^2}{12 - 3}$$

$$= \frac{(3 \times 2.66) + (3 \times 6.00) + (3 \times 0.66)}{9}$$

$$= \frac{7.98 + 18 + 1.98}{9}$$

$$MSE = 3.10$$

$$F_{STAT} = \frac{MSTR}{MSE}$$

$$= \frac{100}{@3.10}$$

$$F_{STAT} = 32.25$$

Calculating on python -

Pvalue = 0.00008.

$\therefore$ Pvalue < 0.05

we ~~to~~ Reject the null hypothesis

Hence -

$\mu_{age(A)} \neq \mu_{age(B)} \neq \mu_{age(C)}$

* First we should check wether the anova test is passing or not, if the mean values for all the groups look same, the we cannot predict the group based on age. So, inorder to predict group based on age, there should be differene in mean values of the groups.

* from scipy.stats import f_oneway

The columns must be grouped first

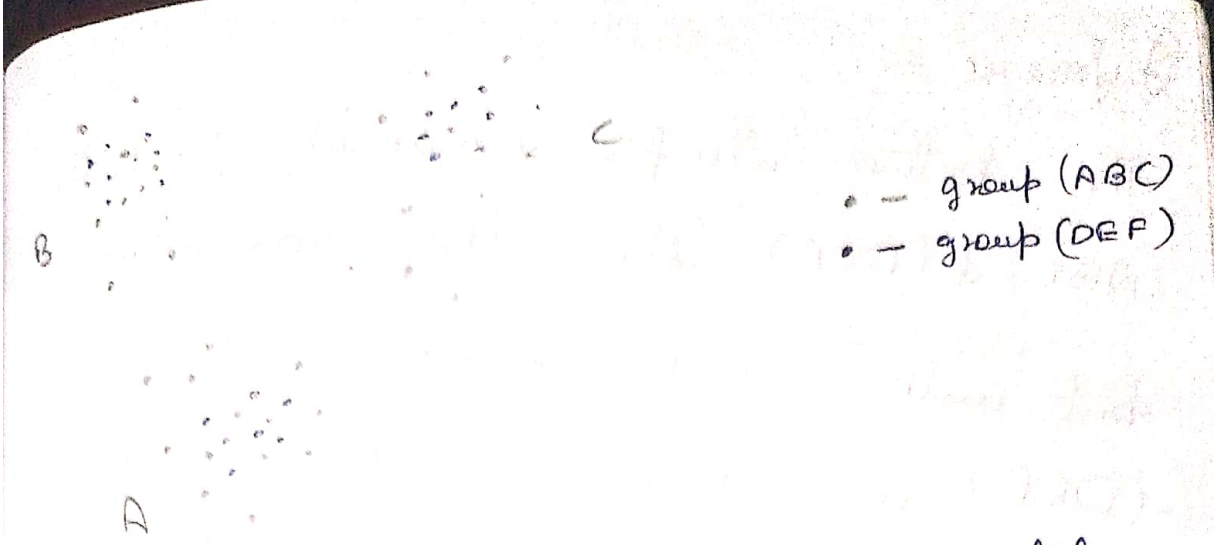
To get complete detail, use anova_lm.

$$F_{stat} = \frac{\text{Between Sample Variation}}{\text{within Sample Variation}} = \frac{MSTR}{MSE}$$

* The first group (A, B, C) mean is same whereas, the second group (D, E, F) mean is different

* The Ho for group 1 got accepted wether The Ho for group 2 got reject because of the more spread of data is group 1, the variance for that is more; whereas the spread of data is group 2 is less as compared to 1.

Inference-



Here we can see that the spread of data of group 1 is higher than the spread of data of group 2

* Since, the datapoints of group 1 have too much of spread is them, there is a high risk of these two groups overlapping each other, they are not significantly different.

but as the spread of data for group 2 is very less, then the chance of them overlapping is very rare, hence ~~they are~~ ~~there is no~~ they are → significant diff.

* If the between variability increases for group 1, then, it can ~~become~~ be concluded that $\mu_A \neq \mu_B \neq \mu_C$.

Inference –

The between sample variability for both (ABC) & (DEF) share the same value, but within sample variability for (DEF) was very compact, because of that their means are far apart.

$$f_{stat}^{ABC} = \frac{\left(\frac{MSTR}{MSE}\right)}{155} = \frac{100}{155} = 0.6428.$$

$$f_{stat}^{DEF} = \frac{100}{3.11} = 32.14.$$

→ Both g1 & g2 are sharing same numerator, which means the dist b/w groups is same (between sample variability is same)

→ But within sample variability for g1 is more for ~~g2~~ than g2 (compact)

→ If the ABC mean should be sign. diff then either the denominator shud shrink or the numerator should increase more.

* check avg. Sal vs Dept.

check average Salaries in all 3 departments

for this we use anova test

	df	Sum_Sq	mean_Sq	F	PR(>F)
					Pvalue
* dof → group	2.0	SSTR 200	MSTR 100.00	Fstat 0.64285	0.5483
Residual	9.0	SSE 1400.0	MSE 155.556	NaN	NaN

Two Way ANOVA

* Simultaneously when we check two categories with mean of a continuous Column, then it is called two way ANOVA.

Salary ~ Dept ← Interactions

Salary ~ Joblevel

(Dept alone cannot find Salary but together is there any change.)

H_0 : There is no interaction

H_a : There is an interaction

$Sal = \beta_0 + \beta_1 \cdot Joblevel$

$Sal = \beta_0 + \beta_1 \cdot Department$ → This is ruled out as we cannot find Salary by dept

$Sal = \beta_0 + \beta_1 \cdot Joblevel + \beta_2 (Department \times Joblevel)$

Interaction

mod = ols (Salary ~ Job_level : Depart, hr).fit()

Anov_tab = Anova_lm (mod, typ = 1)

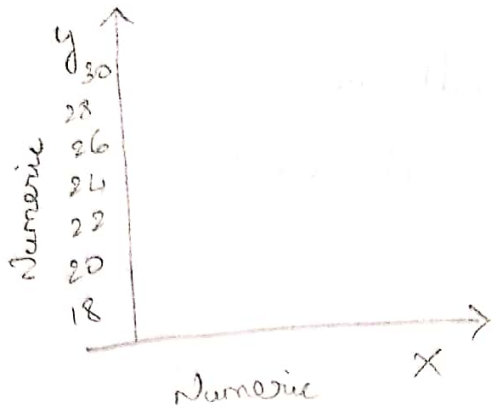
In two way ANOVA, Here, the Ho: There is no interaction b/w the two Columns is getting rejected, here we deduce that, yes, both put together influence the prediction of the Salary.

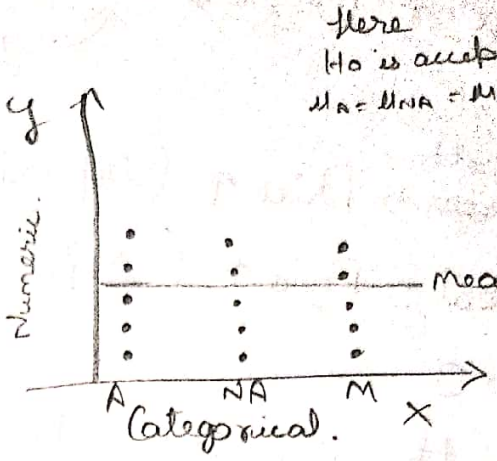
So, If we use departments and job_level as an interaction, it has a better chance at predicting it (Salary) Correctly.

Only the product of depatment and Sales has an influence on predicting Salary.

$Sal = \beta_0 + \beta_1 (Joblevel \times dept)$

Regression -

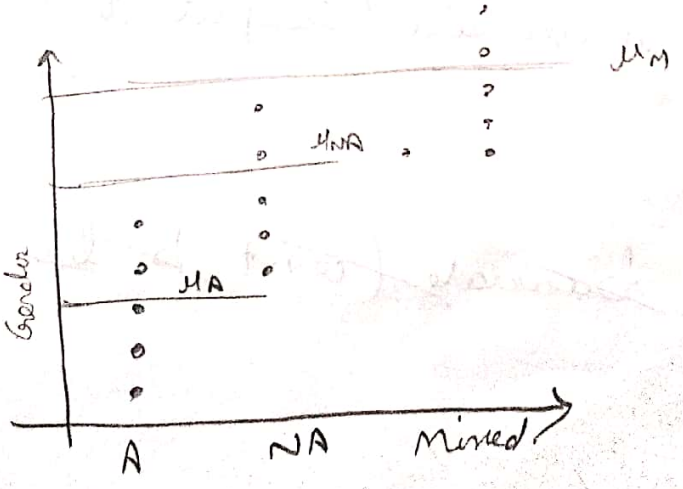




Aura
N Aura
Mixed

Here the mean of all the three samples are same, so, Ho holds good $\mu_A = \mu_{NA} = \mu_M$, so this is a useless variable in predicting.

In order to reject the null hypothesis to, the mean of the 3 Categories used should be different, only then will be able to use it predictor of the dependent var. (age)

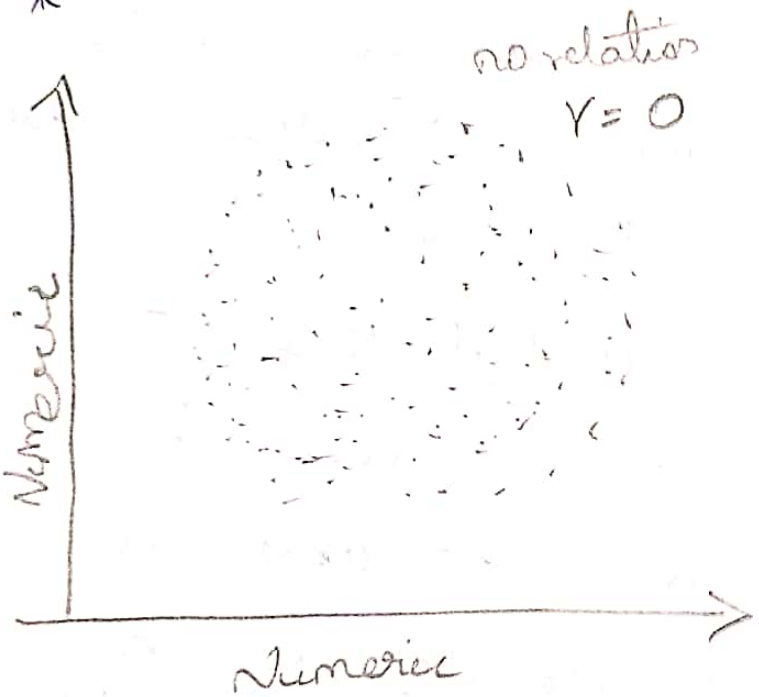


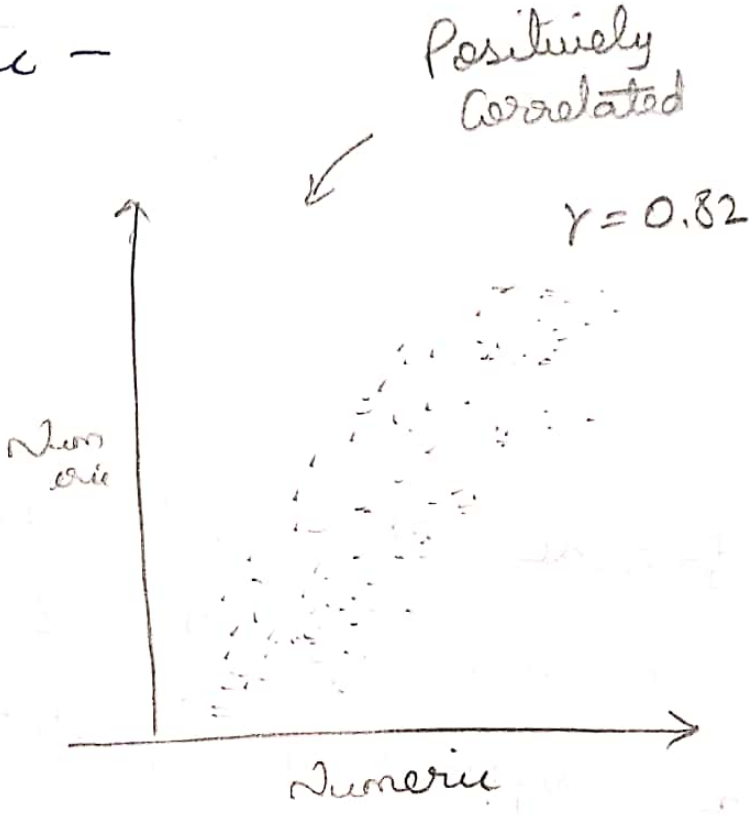
Here Ho rej
$\mu_A \neq \mu_{NA} \neq \mu_M$

Price = f (L, B, L×B)
↑ Interaction (Area)

* length larges → Price ↑ (Separately influence)
* Bread → Price ↑ (Separately influences)
* L×B means area, best include, either length or breadth one, or use the L×B for prediction, Otherwise the data will be redundant.

* Numeric vs Numeric –



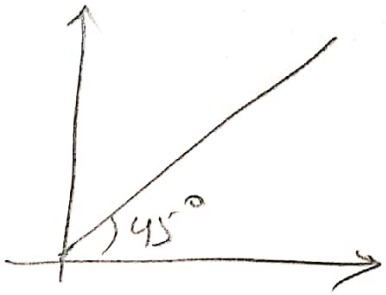


Variance → univariate (respect to one variable)

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} \qquad \frac{x-\bar{x}}{\sigma_x} * \frac{y-\bar{y}}{\sigma_y} \quad (\text{z score form})$$

x	y
3	3
5	5
8	8

$r_{xy} = 1$ self Correlated.

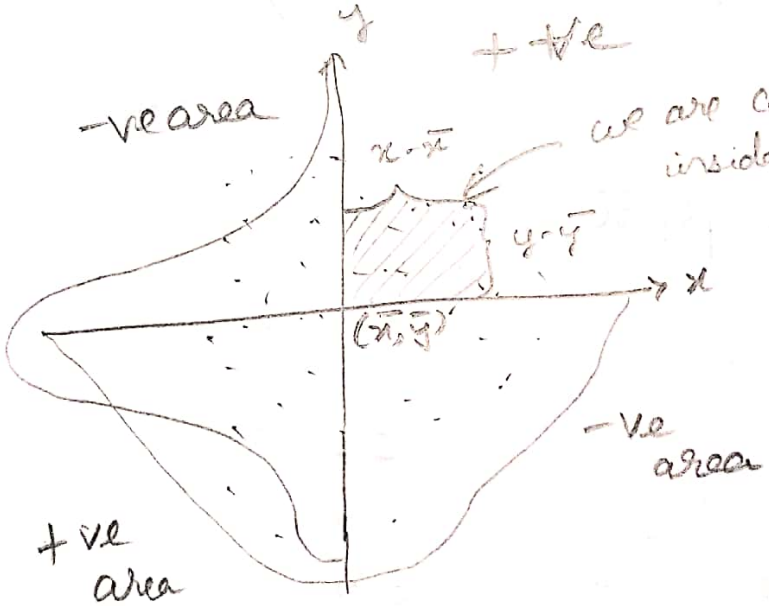


not Practical.

negatively Correlated

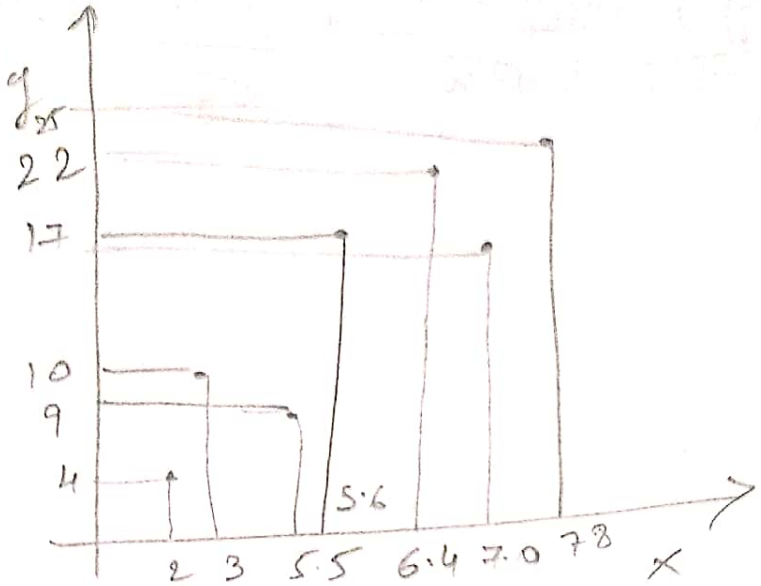
Slope is -ve

e.g horsepower ↑
mileage ↓



we are calc this are inside the regions (rect) for each & every point

So all the areas get Cancelled and $r = 0$.



x	y
2	4
3	10
5.5	9
5.4	17
6.4	22
7.0	16.8
7.8	25

$$\beta_1 = \frac{Cov(x,y)}{Var(x)} = \frac{(x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2}$$

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

$$\beta_1 = \frac{np.Cov(x,y,ddof=1)}{np.var(x,ddof=1)}$$

$\beta_1 = 3.106$

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

x	y	$\hat{y}$
2	4	4.57
3	10	7.68
5.5	9	15.44
5.4	17	1
6.4	22	
7.0	16.8	
7.8	25	

for modelling –

X = 2 dimension

y = 2D / 1 Dim.

X_1	X_2	Y

X = DF [['X_1' + 'X_2']]

Y = DF ['Y']

model = ols('y ~ x_1 + x_2', DF).fit()

y_Pred = model.predict(x)

model.Params

β_0

β_1

β_2

* Statistical analysis of Variablem

X = DF [['x_1', 'x_2']] or DF.drop('y')

y = DF['y']

model = LinearRegr()

model.fit(x, y)

y_pred = model.pred(x)

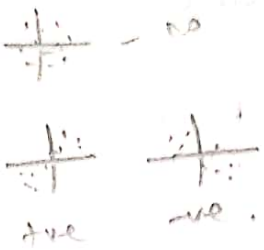
model.Coeff–

* Not Possible to do stat analysis

Hypothesis for model building.

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0.$



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

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

→ denom: large to get good R^2

e.g -

$$Age = \beta_0 + \beta_1 (exp)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \beta_1 = \frac{Cov(x, y)}{Var(x)}$$

If the spread of the variable to be predicted is less, thus it will lead to the poor R^2 value. So, the denominator should be large.

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

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

* The average difference is the error.
  ∠ The average error made by the model is RMSE

* The Smaller The RMSE, better the model.
* Larger the R^2, better the model.

	R^2	RMSE
$y = \beta_0 + \beta_1 Rad.$	35%.	4.25
$y = \beta_0 + \beta_1 Tv$	61%.	3.24
$y = \beta_0 + \beta_1 TV + \beta_1 Rad.$	89%.	1.66

This is the best model.

OLS → to get summary

practical deployment is done using Scikit learn.

RMSE built in func-

```
from sklearn import metrics
metrics.mean_squared_error
```