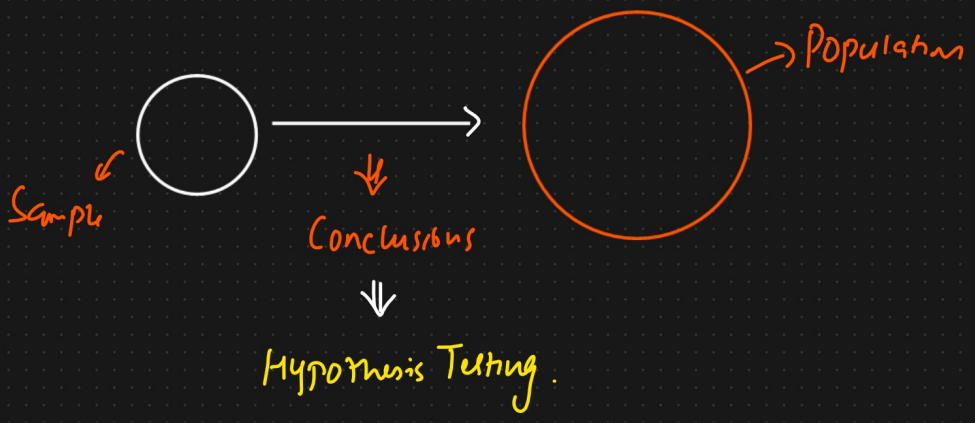
A Hypothesis And Hypothesis Testing Mechanism

Inferential Stats ÷ Conclusion or Inference



Person → crime → Court

Hypothesis Testing Mechanism

① Null Hypothesis (H_0) – Person is not guilty

– The assumption you are beginning with

② Alternate Hypothesis (H_1) – The person is guilty

– Opposite of Null Hypothesis

③ Experiments → Statistical Analysis {p value, significance value}.

→ Correct Proof (DNA, Finger Test)

④ Accept the Null Hypothesis or Reject the Null Hypothesis

Eg: Colleges at District A states its average passed percentages of Students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 90%. with a standard deviation of 4%.

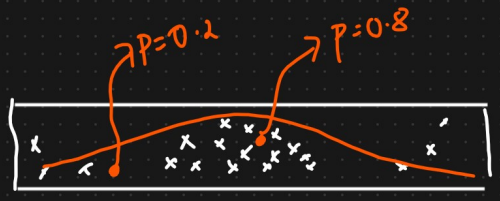
Does this College have a different passed percentage.

Ans) Null hypothesis $(H_0) = \mu = 85\%$

Alternate Hypothesis $(H_1) = \mu \neq 85\%$

Pvalue

The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.



Out of 100 touches, we touch around 20 times in this region

Hypothesis Testing

Eg : Coin is Fair or Not {100 times}

$P(H) = 0.5 \quad P(T) = 0.5$

$P(H) = 0.6 \quad P(T) = 0.4$

$P(H) = 0.7 \quad P(T) = 0.3$

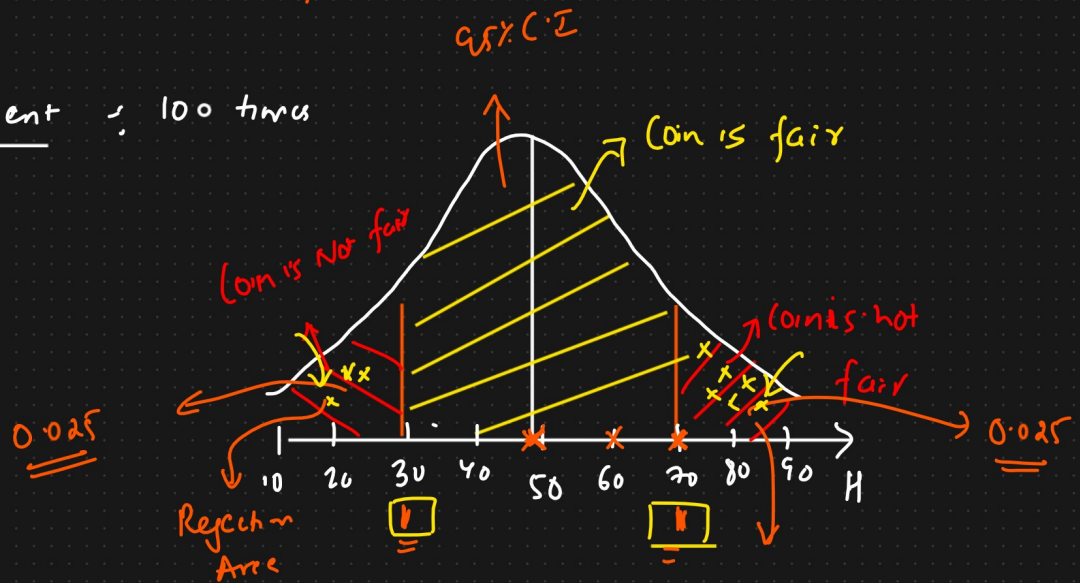
① Null Hypothesis :

H_0 : Coin is fair

② Alternate Hypothesis :

H_1 : Coin is not fair

③ Experiment : 100 times



④ Significance Value : $\alpha = 0.05$

$C.I = 1 - 0.05 = 0.95$

⑤ Conclusion $p <$ Significance value

Reject the Null Hypothesis

else

Fail to Reject the Null Hypothesis

Hypothesis Testing And Statistical Analysis

① Z Test
② t Test
} ⇒ Average ⇒ Z table → Z score And p value
t table

③ CHI SQUARE ⇒ Categorical Data

④ ANNOVA ⇒ Variance

Z test

i) Population std ii) $n \geq 30$

1) The average heights of all residents in a city is 168cm. A doctor believes the mean to be diffuent. He measured the height of 36 individuals and found the average height to be 169.5 cm. with a $\sigma = 3.9$

(a) State null and Alternate Hypothesis

(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

Ans) $\mu = 168cm$ $\sigma = 3.9$ $n = 36$ $\bar{x} = 169.5$

$C.I = 0.95$ $\alpha = 1 - C.I = 1 - 0.95 = 0.05$ //

① Null Hypothesis $H_0 = \mu = 168cms$

② Alternate Hypothesis $H_1 = \mu \neq 168cm$

③ Based on C.I We will draw Decision Boundary

$1 - 0.025 = 0.9750 \Rightarrow$ Z-score

$\Downarrow$

Area $\Rightarrow \pm 1.96$

if Z is less than -1.96 or greater than $+1.96$, Reject the Null Hypothesis.

① Z-test

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

$$Z\text{-score} = \frac{\bar{x} - \mu}{\sigma}$$

$\Downarrow$

$$\frac{\sigma}{\sqrt{n}}$$

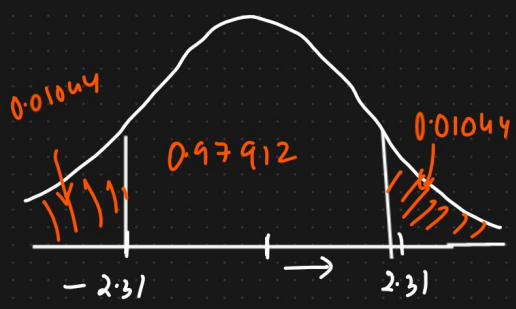
$$= \frac{169.5 - 168}{3.9/\sqrt{36}}$$

$$Z_d = \frac{1.5}{0.65} = \boxed{2.31}$$

Conclusion

$2.31 > 1.96$ Reject the Null Hypothesis (Z-score)

$p < 0.05$



0.98956

$1 - 0.98956 = 0$

② p value $= 0.01044 + 0.01044$

$= 0.02088$ $\quad p < 0.05$

$0.02088 < 0.05 \Rightarrow$ Reject the Null Hypothesis

$\Downarrow$

③ Final Conclusion the Average $\neq$ 168cm

The average height seems to increasing based on sample heights.

(2) A factory manufactures bulbs with a average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will malfunction in less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years.

(a) State null and alternate hypothesis

(b) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

Ans) $\mu = 5 \quad \sigma = 0.50 \quad n = 40 \quad \bar{x} = 4.8$

a) Null Hypothesis $H_0 : \mu = 5$

Alternate Hypothesis $H_1 : \mu < 5$ {1 Tail Test}

b) Decision Boundary

c) Z-test

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}}$$

$$= -2.53$$

Area under curve with Z score $-2.53 = 0.0570$.

P-value $= 0.0570$ $\quad \alpha = 0.02$

Compare P-value with α

$0.0570 < 0.02 \Rightarrow$ False

We accept the Null Hypothesis ==.

We Fail to Reject the Null Hypothesis.

Student t distribution

In z stats when we perform any analysis using z-score

We require σ (population standard deviation) → is already known

How do we perform any analysis when we dont know the population standard deviation?

⇓

Student's t distribution

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

z table

t stats

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

s = Sample Standard deviation

t table ⇒ t test

Degree of freedom

$$dof = n - 1 = 3 - 1 = 2$$

3 people

T-stats : J test → One Sample t- test.

① In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence? C.I = 95% α = 0.05

Ans) $\mu = 100$ $n = 30$ $\bar{x} = 140$ $S = 20$ C.I = 95% $\alpha = 0.05$

① Null Hypothesis H_0 : $\mu = 100$

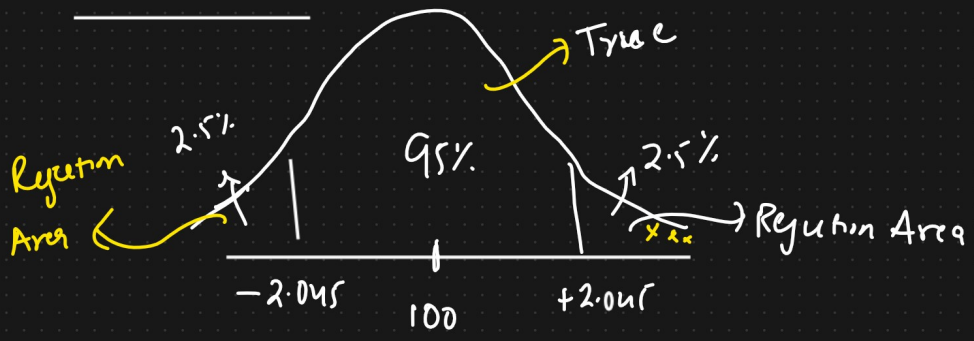
Alternate Hypothesis H_1 : $\mu \neq 100$ {2 Tail Test}

② $\alpha = 0.05$

③ Degree of freedom

$df = n - 1 = 30 - 1 = 29.$

④ Decision Rule



if t test is less than -2.045 or greater than 2.045, reject the null hypothesis

⑤ Calculate Test statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96$$

$$t = 10.96$$

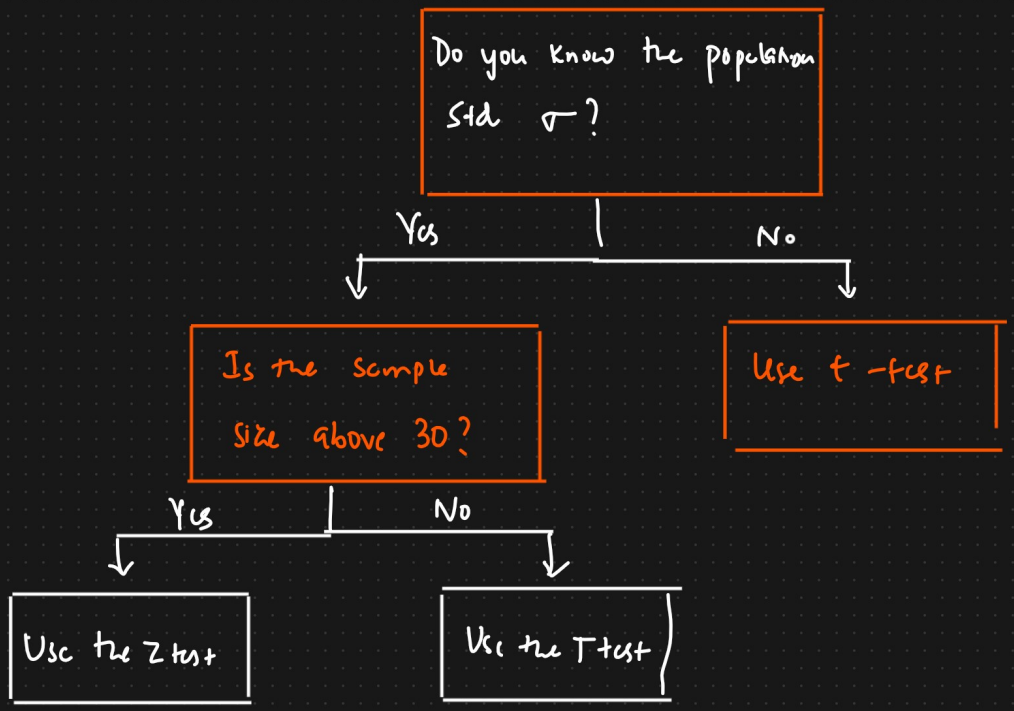
Since

$t = 10.96 > 2.045$ {Reject the NULL Hypothesis}.

Conclusion : Medicaton used has affected the Intelligence

Medication has increased the Intelligence

When To Use T-tests Vs Z-test



Type 1 and Type 2 Errors

Reality : Null Hypothesis is True or Null Hypothesis is False

Decision : Null Hypothesis is True or Null Hypothesis is False

Outcome 1 : We reject the Null Hypothesis when in reality it is false → Good

Outcome 2 : We reject the Null Hypothesis when in reality it is True → Type 1 Error

Outcome 3 : We retain the Null Hypothesis, when in reality it is False → Type 2 Error

Outcome 4 : We retain the Null Hypothesis when in reality it is True → Good

Bayes Statistics (Bayes Theorem)

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem.

Baye's Theorem

Probability → Independent Events
Probability → Dependent Events

① Independent Events

Eg: Rolling a dice

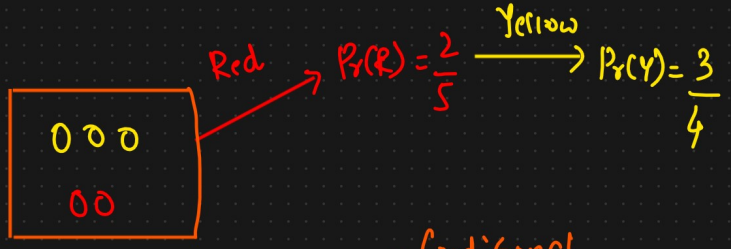
$\{1, 2, 3, 4, 5, 6\}$

$Pr(1) = \frac{1}{6}$ $Pr(2) = \frac{1}{6}$ ---

Tossing a coin

$Pr(H) = 0.5$ $Pr(T) = 0.5$

② Dependent Event



Conditional probability

$$Pr(R \text{ and } Y) = P(R) * Pr(Y/R)$$

$$= 2/5 * 3/4 = \frac{6}{20}$$

$$Pr(A \text{ and } B) = Pr(B \text{ and } A)$$

$$Pr(A) * Pr(B/A) = Pr(B) * Pr(A/B)$$

$$Pr(B/A) = \frac{Pr(B) * Pr(A/B)}{Pr(A)}$$

⇒ Bayes theorem

⇓

$$Pr(A/B) = \frac{Pr(A) * Pr(B/A)}{Pr(B)}$$

A, B = events

Pr(A/B) = Probability of A given B is true

Pr(B/A) = " " B " A is true

Pr(A), Pr(B) = Independent probabilities of A and B

DATASET

Size of House (x_1)	No. of Rooms (x_2)	Location (x_3)	Price (y)
	⇑ Independent		⇑ o/p / dependent

$$Pr(y/x_1, x_2 x_3) = \frac{Pr(y) * Pr(x_1 x_2 x_3/y)}{Pr(x_1, x_2, x_3)}$$

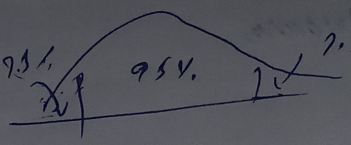
⇓

Bayes Theorem

Section 23 :- Inferential statistics :-

118. Confidence Interval and margin of error.

C.I = 95%



Confidence Interval

Point Estimate

$\bar{x}$ ⟶ μ

Confidence Interval

Point Estimate ± Margin of Error

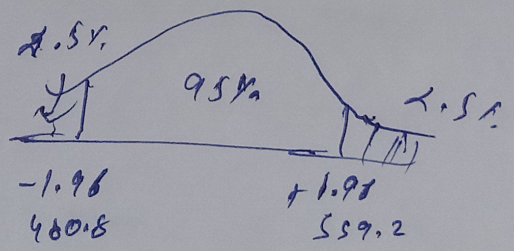
Z test ⇒ $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Ex:- On the verbal section of CAT exam, the standard deviation is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% C.I. about the mean

$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$\alpha = 0.05$

$1 - 0.025 = 0.9750$



$\pm 1.96 \hookrightarrow Z_{0.05}$

Lower C.I = $520 - (1.96) \frac{100}{\sqrt{25}} = 480.8$

Higher C.I = $520 + 1.96) \times \frac{100}{\sqrt{25}} = 559.2$

Conc.! Zom 95% confidence about the mean CAT score is between 480.8 & 559.2

CHI SQUARE TEST

The chi square test for Goodness of fit test claims about population proportions.

It is a non parametric test that is performed on categorical [ordinal and nominal] data.

There is a population of male who like different color bikes

	Theory	sample
Yellow Bike	1/3	22
Red Bike	1/3	17
Orange Bike	1/3	39
	↓	↳ Observed
	Theory categorical distribution	categorical distribution

(a) Goodness of fit test

Goodness of fit test

In a science class of 75 students, 11 are left handed, does this class fit the theory that 12% of people are left handed.

$\frac{12}{100} \times 75 = 9.$

A)

	O	E
Left handed	11	9
Right handed	64	66
	75	75

2. CHI SQUARE for Goodness of fit

CHI SQUARE for Goodness of fit

In 2010 census of the city - the weight of the individuals in a small city were found to be the following

<50kg	50-75	>75
20%	30%	50%

In 2020, weight of n=500 individuals were sampled. Below are the result.

<50	50-75	>75
140	160	200

Using $\alpha = 0.05$, would you conclude the population differences of weight has changes in the last 10 years?

Sol.

Expected =

<50	50-75	>75
0.2*500 = 100	0.3*500 = 150	0.5*500 = 250

① Null Hypothesis: H_0: The data meets the expectation.
Alternate Hypo: H_1: The data does not meet the expectation.

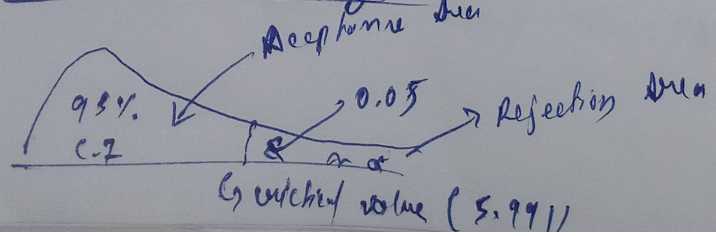
② $\alpha = 0.05$, C.I = 95% (confidence interval)

③ Degree of freedom

df = k-1 = 3-1 = 2

$e.v = df * \alpha$

④ Decision Boundary



If χ^2 is greater than 5.99, Reject H_0

else

we will fail to reject the null hypothesis

③ Calculate chi square test statistics

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$ ← observed, ← expected

2020

$n = 500$

observed

<50	50–75	775
140	160	200

Expected

<50	50–75	775
$0.2 \times 500 = 100$	$0.3 \times 500 = 150$	$0.5 \times 500 = 250$

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250} = 16 + 0.66 + 10 = 26.66$$

$\chi^2 = 26.66$

~~If χ^2 is~~

$26.66 > 5.99$, Reject H_0

Answer

The weight of 2020 population are different than those expected in 2010 paper.

Analysis of variance (ANOVA)

Def^n: ANOVA is a statistical method used to compare the means of two more groups.

ANOVA

① factors (variable)

② Levels.

eg:- medicine (factor)

[Dosage] [mg 10mg 18mg → levels.

mode of payment (factors)

[Levels] UPI Phonepe Imps NEFT

- Analysis of variance (ANOVA)

Assumptions in ANOVA

① Normality of sampling distribution of mean
The distribution of sample mean is normally distributed.

② Absence of Outliers
Outlying score need to be removed from the dataset

③ Homogeneity of variance
Population variance in different level of each independent variable are equal.

$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$

④ Samples are independent and random.

Analysis of variance (ANOVA)

Types of ANOVA (3 Types)

① One way Anova: One factor with atleast 2 levels, these levels are independent.

E.g Doctor wants to test a new medication to decrease headache. They split the participation in 3 conditions [10mg, 20mg, 30mg] Doctor ask the participation to rate the headache [1-10]

Medication → factor

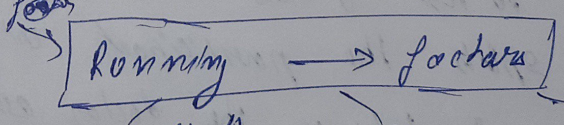
10 mg 20 mg 30 mg

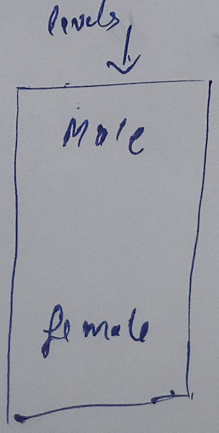
② Repeated measure Anova :- One factor with atleast 2 levels, levels are dependent.

Running → factor

levels →	Day 1	Day 2	Day 3
	8	5	4
	7	4	9

③ factorial ANOVA : Two or more factors (each of which with atleast) 2 levels, levels can be independent and dependent





	Day 1 (level)	Day 2	Day 3
	8	5	4
	9	4	3
	2	4	6

Analysis of variance (ANOVA)

Hypothesis Testing In ANOVA (Partitioning of variance in the Anova)

Null Hypothesis H_0 : $\mu_1 = \mu_2 = \mu_3 \ldots \mu_k$

Alternate Hypothesis H_1 : Atleast one of the sample mean is not equal.

$\mu_1 \neq \mu_2 \neq \mu_3$

Test statistic

$F = \frac{\text{variance b/w samples}}{\text{variance within sample}}$

$H_0 \Rightarrow \bar{x}_1 = \bar{x}_2 = \bar{x}_3$

$H_1 \Rightarrow$ Atleast one sample mean is not equal.

σ^2

Variance within Samples

Variance b/w samples

x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4
$\bar{x} = 3$	$\bar{x}_2 = 19/4$	$\bar{x}_3 = 4$

One Way ANOVA

One factor with atleast 2 levels, levels are independent.

① Doctor wants to test a new medication which reduces headache. They split the participant with 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the to rate the headache b/w [1-10].

Are there any diff. b/w the 3 conditions

alpha = 0.05 %.

15 mg	30 mg	45 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① Define Null and Alternate Hypothesis:

H_0 : $\mu_{15} = \mu_{30} = \mu_{45}$

H_1 : not all μ are equal

② Significance $\alpha = 0.05$ C.L = 0.95

③ Calculate Degree of freedom

$N = 21$ $q = 3$ $n = 7$

N → sample, q → categories

df between $= q - 1 = 3 - 1 = 2$

df within $= N - q = 21 - 3 = 18$

df total $= N - 1 = 20$

df_1, df_2

$(2, 18)$

↓

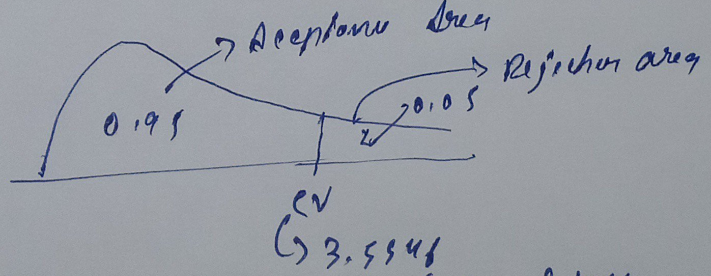
F table

$\alpha = 0.05$

↓

Critical value

④ Decision Boundary



Decision Rule

If F is greater than 3.5546, reject the Null hypothesis

MSC (mean square)

⑤ Calculate F test. statistics

Sum of square (SS) $F = \frac{\text{Variance between samples}}{\text{Variance within samples}}$

	SS	df	MS	F
Between	98.67	2	49.34	
Within	10.29	18	0.54	
Total	108.96	20		

① SS between $= \frac{\Sigma(a_i)^2}{n} - \frac{T^2}{N}$

T_{sum} :- $4+8+3+8+8+9+8 = 57$

3_{sum} : $7+6+6+7+8+7+6 = 42$

4_{sum} : $4+3+2+3+4+3+2 = 21$

$= \frac{57^2+42^2+21^2}{7} - \frac{[57^2+42^2+21^2]}{21}$

$= \boxed{98.67}$

② $SS_{within} = \Sigma y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}$

$\Sigma y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + \cdots$

$= 853$

$= 853 - \frac{[57^2 + 47^2 + 21^2]}{7}$

$= 10.29$

f test $= \frac{MS\ Between}{MS\ within}$ $\quad F = \frac{variance\ b/w\ samples}{variance\ within\ samples}$

$F' = \frac{49.34}{0.54} = 86.56$

If F is greater than 3.5546, Reject the H_0.

$86.56 > 3.5546$ Reject the H_0.