Introduction to Statistics

Defn: Stats is the science of collecting, organizing and analysing data.

Data: Facts or pieces of information

Eg: Heights of students in classroom
Salary of people in society.

⇒ Types of Stats

① Descriptive Stats
② Inferential Stats

→ ① Descriptive Stats

⊙ It consists of organizing, summarizing, and visualizing data.

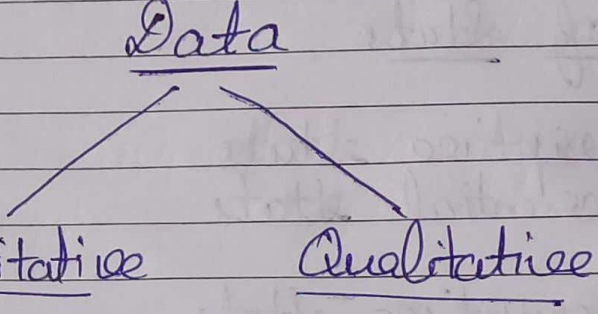
① Measure of Central Tendency
② Measure of dispersion (V, Sd, Z-Score)
③ Different types of distribution.

Eg: Histogram, pdf, pmf, Gaussian, log normal, Exponential, Binomial, Bernoulli, Poisson.

→ ② Inferential stats

⊙ It consists of using data you have measured to form conclusion.

① Z-test
② T-test
③ Chi square test
④ Anova test



⇒ Sampling Techniques

① Random Sampling

⊙ Most common in behavioral research

⊙ Equal chance of being selected in the sample.

② Stratified Sampling

⊙ Dividing the population into subgroups or strata based on certain characteristics or attributes.

- w.r.t age, gender gender, income etc.

(3) Convenience sampling

- Participants are selected based on availability and willingness to take part.

(4) Cluster sampling

- Divide a population into clusters
- Eg: Districts, schools and randomly selecting the clusters

(5) Snowball sampling

- ~~Esi~~ Existing study subjects are used to recruit more subjects into the sample.

(6) Purposive sampling

- Selecting samples based on the judgement of the survey taker or researcher.

(7) Systematic sampling

- It is a probability sampling method where researchers select ~~numbers~~ members at a regular interval.

⇒ Scale of Measurement

1) Nominal scale data
2) Ordinal scale data
3) Interval scale data
4) Ratio scale data

1) Nominal scale data:

⊙ It is a type of qualitative data that labels variables without using numbers.

⊙ Eg: Mode of transportation, genotype, blood type, ~~zip~~ zip code, gender, race, ~~etc~~ etc...

2) Ordinal scale data:

⊙ It is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them.

⊙ Eg: Ranking of school students, Rating of waiters in restaurants.

3) Interval scale data:

- It is a type of quantitative data that measures values on a scale with equal distances between each value.
- Eg: Temperature in Fahrenheit or Celsius, pH measure, IQ and SAT scores, etc...

4) Ratio scale data:

- It is a type of quantitative data that measures variables on a scale with equal intervals between values and a true zero.
- Eg: Height, money, age, weight, etc...

⇒ Random Variables

- A Random variable is a variable in statistics that assigns numerical values to the outcomes of sample space. The possible values of a random variable depend on the outcomes of a random phenomenon.

Variable (x, y)

$x + 6 = 8$

$x = 2$

$x = 2$

$y = 6$

$8 = y + x$.

⊙ Random variable is a process of mapping the output of a random process or experiment to a number.

⊙ Ex: → Toss a coin

$$x = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}$$

→ Rolling a dice

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

⇒ Covariance & Correlation

⊙ It is the measure of the relationship b/w two ~~numbers~~ random variables. It measures how much the variable change together, or the variance between them.

⊙ It measures the direction of a relationship b/w two variables.

→ Population Covariance Formula

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

→ Sample Covariance

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

⊙ Advantages:

→ +ve or -ve value

→ Shows the relationship b/w two variables positive / negative.

⊙ Disadvantages:

→ Doesn't have a specific limit value.

② Pearson Correlation Coefficient

⊙

$$r_c = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] - [n\Sigma y^2 - (\Sigma y)^2]}}$$

⊙ Pearson Correlation Coefficient doesn't work with non-linear data. So we will use Spearman correlation.

③ Spearman's Rank Correlation Coefficient

$$N_S = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

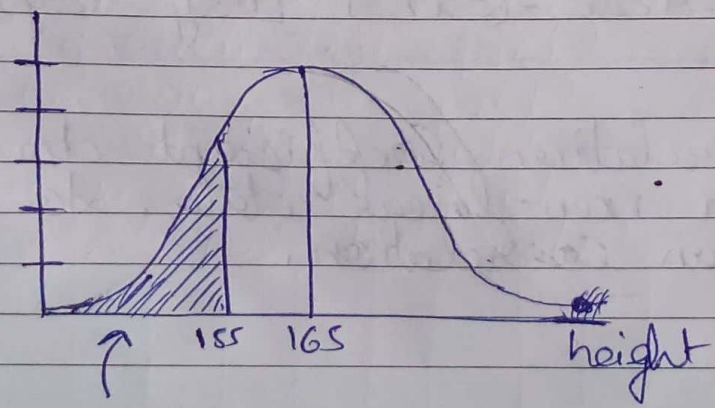
R = Rank based on frequency / order / value

- If 2 features are highly correlated, its ok to drop one of the feature.

⇒ Probability Distribution Function

① Probability density function

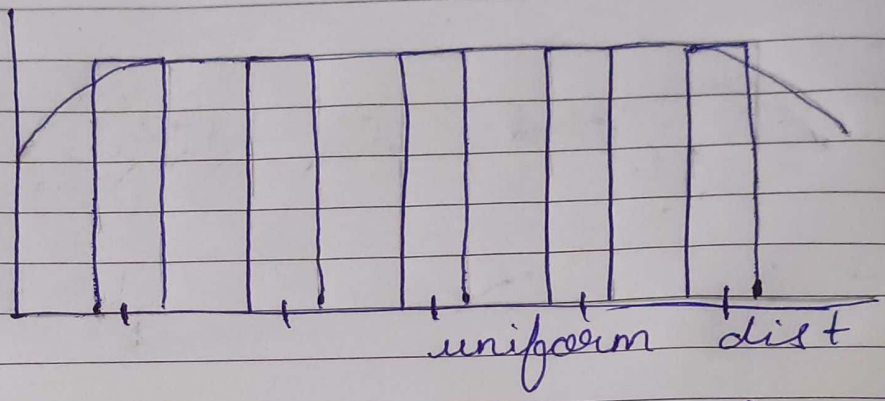
- Continuous function
- Eg: Age, Height [float]



pr(H<=155) pr(H<=155, H>=175)

② Probability mass function

- Discrete values
- Eg: No. of bank acc, Rolling a dice
  int

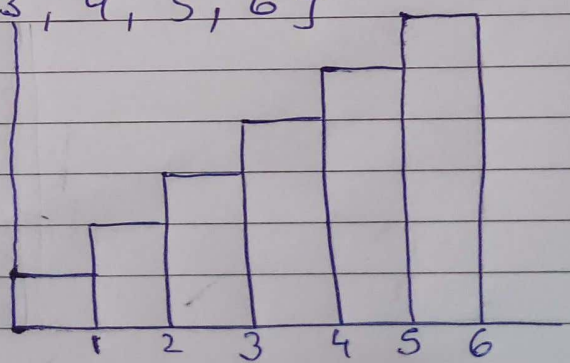


$Pr(x \leq 4) = Pr(x=1) + pr(x=2) + pr(x=3) + pr(x=4).$

$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \boxed{\frac{2}{3}}$

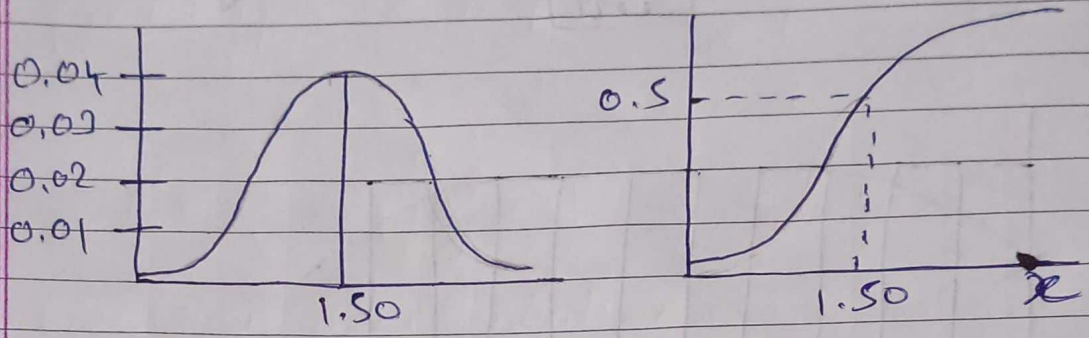
③ Commulative Distribution Function

pmf = {1, 2, 3, 4, 5, 6}



⊙ Adding the previous to the next value.

pdf & cdf



⊙ Probability Density of pdf ⇒ Gradient or derivative of ~~cdf~~ cdf.

pdf is the derivative of cdf
cdf is the integration of pdf

⇒ Types of Probability function

1) Normal dist (pdf)
2) Bernoulli dist (pmf) [binary outcome]
3) Uniform dist (pmf)
4) Poisson dist (pmf)
5) Binomial dist (pmf)
6) Log Normal dist (pdf)

→ Bernoulli Distribution

⊙ Discrete Random Variable (pmf)

⊙ Outcome are Binary

Eg:→ Tossing a coin {H, T}.

Pr (H) = 0.5 = p

Pr (T) = 0.5 = q

→ Whether a person will pass / fail.

pr (pass) = 0.7 = p

pr (fail) = 1 − p = 0.3 = q

→ Binomial Distribution

⊙ Discrete Random Variable

⊙ Every Experiment outcome is Binary

⊙ It represents the probability for x success in n trials, given a success probability p for each trial.

⊙ Eg: Tossing a coin 10 times.

notation → B(n, p)

pr (H) = 0.5 = p

pr (T) = 0.5 = q

parameter → = n ∈ {0, 1, 2, 3, ...} ⇒ no. of trials

$p \in \{0, 1\}$ → Success probability

$q = 1 - p$.

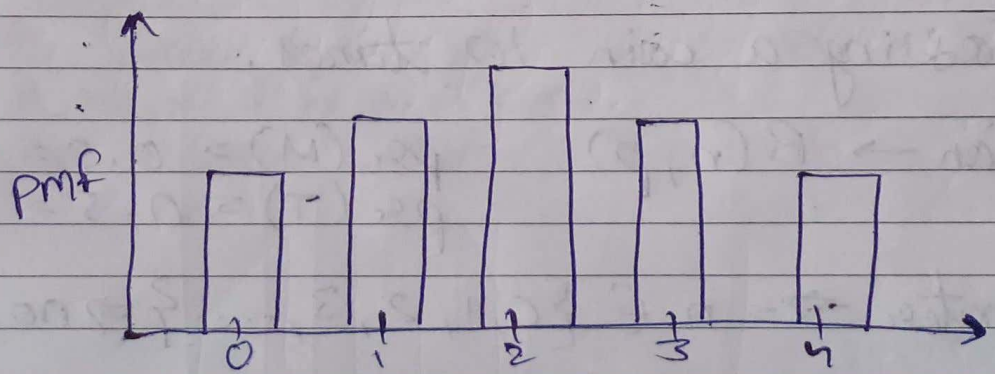
Support = $k \in \{0, 1, 2, \ldots n\}$ ⇒ no. of succes

pmf: $\boxed{pr(k, n, p) = {}^nC_k \, p^k (1-p)^{n-k}}$

$${}^nC_k = \frac{n!}{k!(n-k)!}$$

→ Poission Distribution:

- ⊙ Discrete Random Variable (pmf)
- ⊙ Describes the no. of elements occuring in a fixed time interval.
- ⊙ Eg:- no. of people visiting hospital every hour
  - no. of people visiting banks, at 11 am.



$\lambda = 3$ = Expected no. of event occur at every time interval.

Q. What is the probability of no. of people visiting at 3 pm. $\Rightarrow P(x=3)$

$$= \frac{e^{-3}\, 3^3}{5!} = 0.101 \approx 10\%$$

10% of the people visits the bank at 3 pm.

⇒ Empherical Rule of Normal dist

68	95	99.7%
1st sd	2nd sd	3rd sd

⇒ Uniform Distribution

1) Continuous Uniform Distribution

⊙ A Continuous uniform distribution is a probability distribution that takes values within a specified range. It is defined by the two parameters, a, b where a in the lower limit and, b in the upper limit.

Notation = $U(a, b)$

Parameter = $-\infty < a < b < \infty$

$$pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

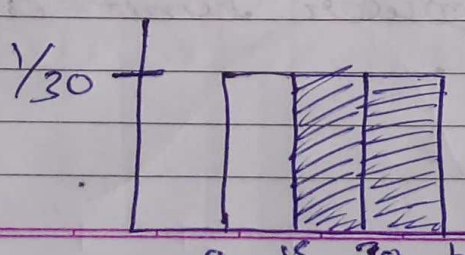
$$cdf = \begin{cases} 0 & \text{for } x < a \\ 1 & \text{for } x > b \\ \frac{x-a}{x-b} & \text{for } x [a, b] \end{cases}$$

$$\text{Mean} = \frac{a+b}{2} \qquad \text{Median} = \frac{a+b}{2}$$

① Example: The no. of candies sold at a shop is uniformly distributed with a max of 40 and a min of 10.

= Probability of daily sales to fall b/w 15 and 30.

$a = 10$
$b = 40$
$Pr(15 \le x \le 30)$



$(x_2 - x_1) \times \frac{1}{b-a}$

$= 30 - 15 \times \frac{1}{40-10}$

$= 15 \times \frac{1}{30} = 0.5$

- $pr(x \geq 30)$

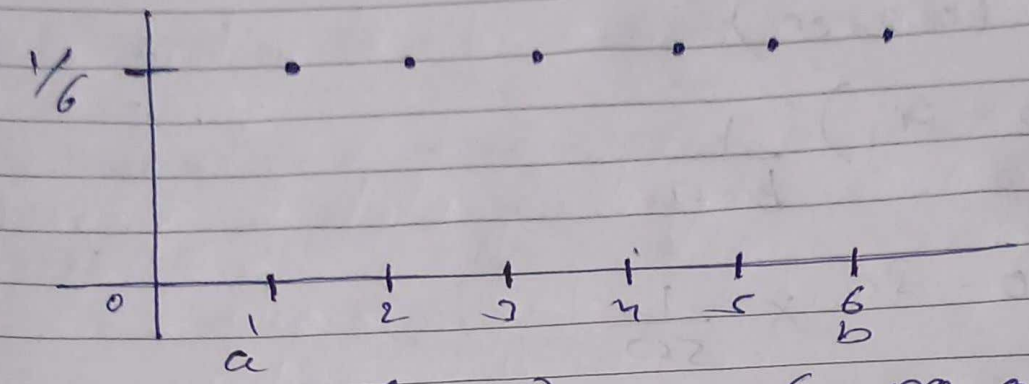
$(x_2 - x_1) \frac{1}{b-a}$

$40 - 30 \times \frac{1}{50}$

$\frac{10}{30} = \boxed{0.33}$

2) Discrete Uniform Distribution

- A discrete uniform distribution is a statistical distribution where the probability of outcomes is equally likely and with finite values.
- In a discrete uniform dist, everyone of a n value has equal probability $1/n$.
- Eg: • Rolling a dice
  • Selecting a card from deck of card
  • Tossing a fair coin.
- Eg: Rolling a dice



Notation: $U(a, b)$ $n = 6$ no. of outcomes

$$\boxed{n = b - a + 1}$$

$P(1) = \frac{1}{n} = \frac{1}{6}$

$$\text{Mean} = \frac{a+b}{2}$$

Parameter a, b $\boxed{b > a}$ $\text{Median} = \frac{a+b}{2}$

$pmf = \frac{1}{n}$

⇒ Standard Normal Distribution

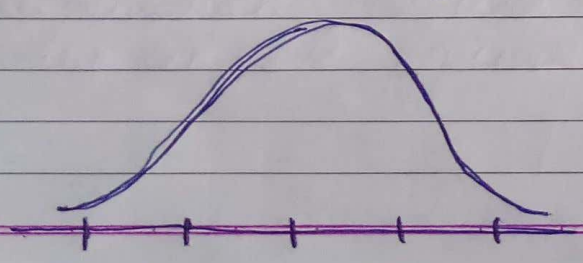
→ Z-score

→ Z-stats

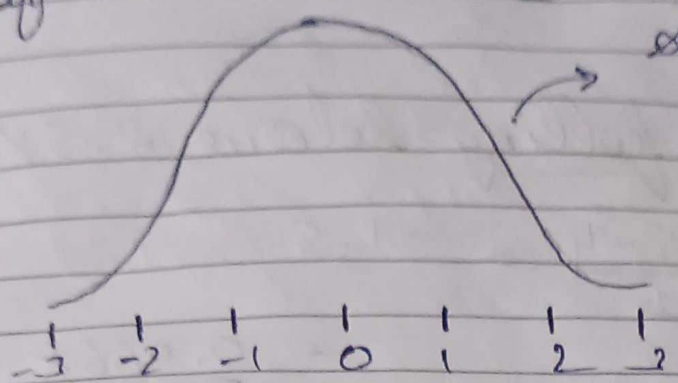
$x = \{1, 2, 3, 4, 5\}$ ⇒ Normally Distributed

$\mu = 3$

$\sigma = 1.414$



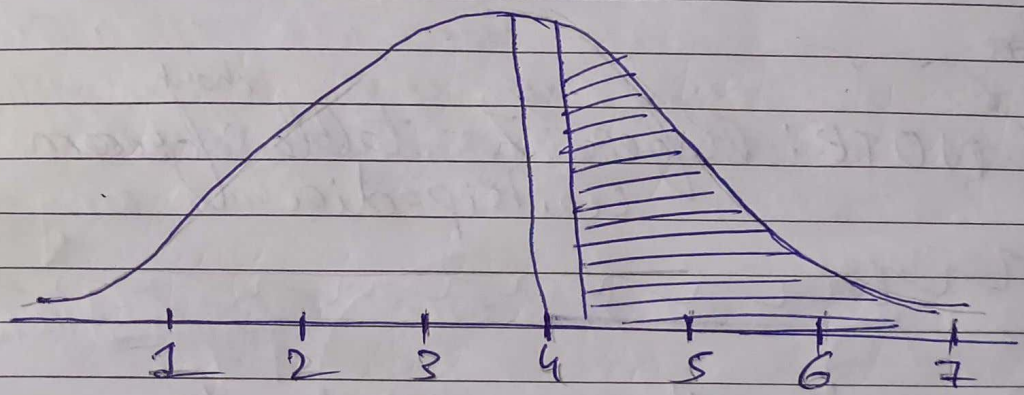
if $\mu = 0$ & $\sigma = 1$



Standard Normal Distribution

$$\Rightarrow \boxed{Z\text{-score} = \frac{x_i - \mu}{\sigma}}$$

Z-table - We use to find out the area under the curve.



What % of the data is falling down 4.5

$$Z\text{-score} = \frac{x_i - \mu}{\sigma} = \frac{4.5 - 4}{1} = (0.5)$$

$\Rightarrow$ 0.6915

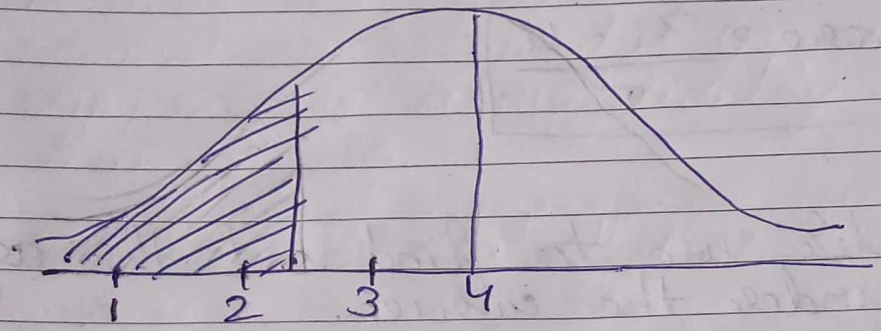
Area under the curve = $1 - 0.6915 = 0.3085 = 30.85\%$

② Percent of data falling below 2.5%.

$$Z\text{-score} = \frac{2.5 - 4}{1} = -1.5$$

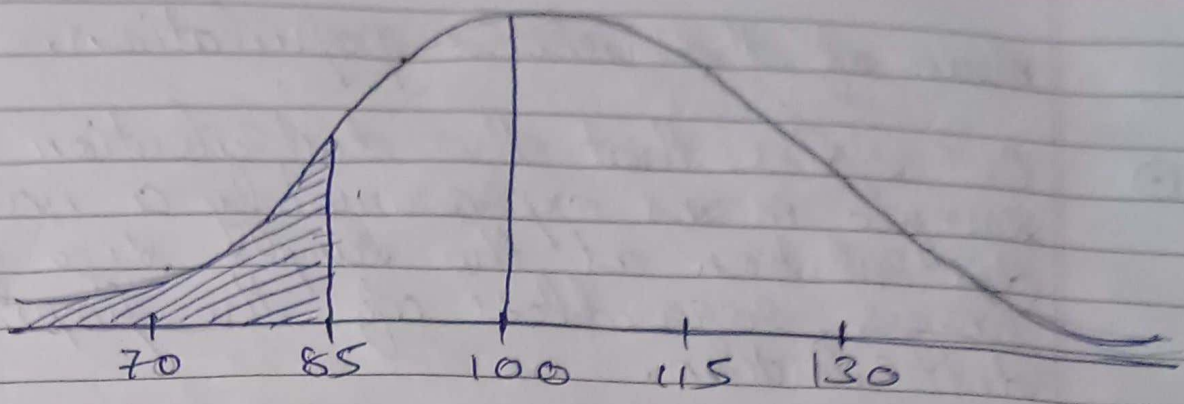
$$= 0.0668$$

$$= \boxed{6.68\%}$$



NOTE: Refer Z-table (chart) from any wikipedia.

In india the avg IQ is 100, with σ of 15. What is the percentage of the population would you expect to have an IQ lower than 85%.



$$Z\text{-Score} = \frac{85-100}{15} = \frac{-15}{15} = \frac{-1}{} = 0.1587$$
$$= \boxed{15.87\%}$$

Area under the curve $= 85\%$.

$$1 - 0.1587$$
$$= 0.8413 = \boxed{84.13\%}$$

Area between 85 & 100

$$= 0.5 - 0.1587$$
$$= 0.3413 = \boxed{34.13}$$

⇒ Central Limit Theorem

⊙ CLT is a statistical theorey that states that when a large sample size has a finite variance, the samples will be normally distributed.

⊙ The theorem also states that the mean of the samples will be approximately equal to the

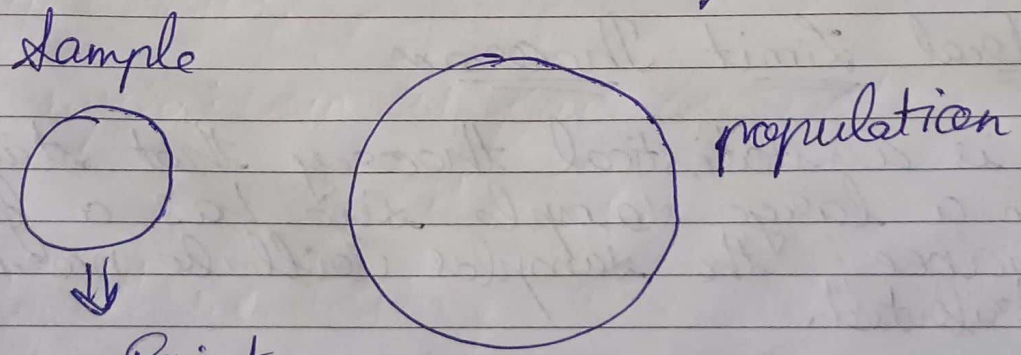
mean of the whole population.

⊙ CLT states that the distribution of sample means approximately a normal distribution as the sample size gets larger, regardless of the population's distribution.

⊙ The theorem holds true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently larger than $n = 30$.

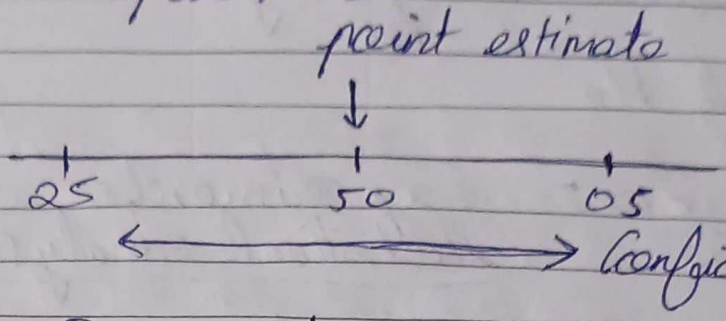
* Inferential Statistics

1) Point Estimate: It is an observed numerical value used to estimate an unknown ~~parameter~~ population parameter. (single numerical value)



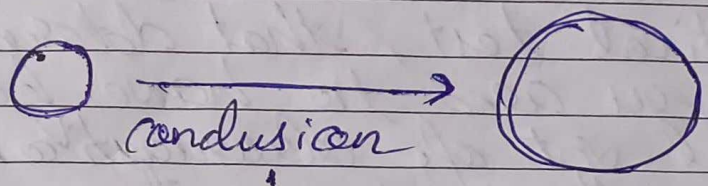
$\bar{x}$ = estimate → μ ⇒ population mean

2) Interval Estimate : Range of values used to estimate the unknown population parameter.



⇒ Hypothesis & Hypothesis Testing Mechanism

Inferential stats → Conclusions



→ Hypothesis testing.

→ Hypothesis testing mechanism

person crime →

① Null hypothesis (H_0)

⇒ The person is not guilty.

→ Assumptions you are ~~bat~~ beginning with

② Alternate Hypothesis (H_1)

The person is guilty →

→ Opp of H_0

→ Evidence - Experiments
(Statistical Analysis)

Z test, t test, Annova test etc.

⇒ P-value

⊙ P value is a number calculated from a statistical test, that determine how likely you are to have found a practical set of observations if the null hypothesis were true. P values are used in hypothesis testing to decides whether to reject the null hypothesis.

⊙ Example: Coin is fair or not

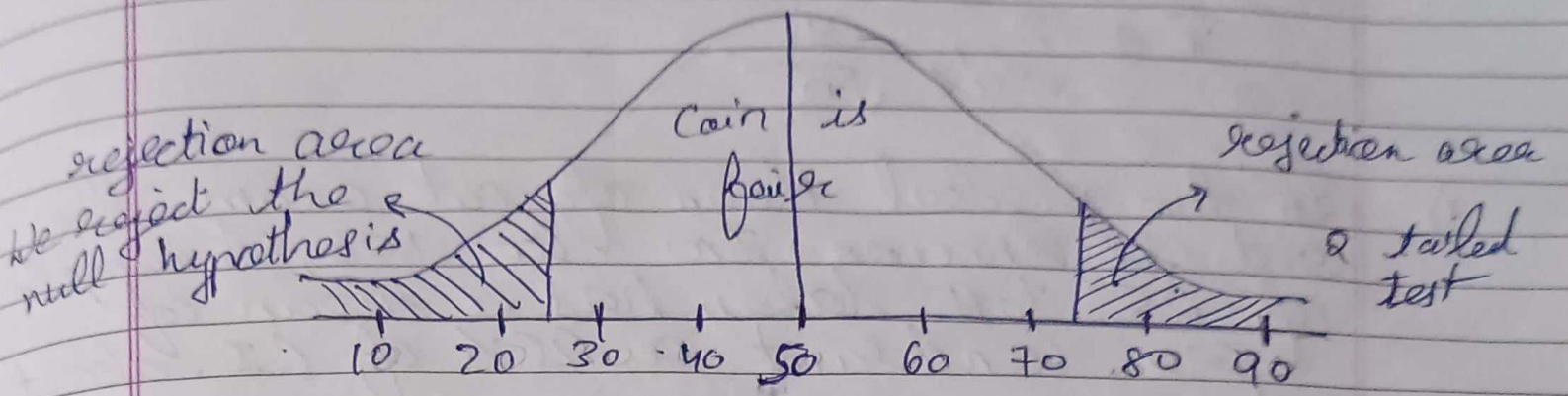
{H, T}.

→ Hypothesis Testing

1) Null hypothesis H_0 → Coin is fair

2) Alternate hypothesis H_1 → Coin is not fair.

3) Experiment → Testing.

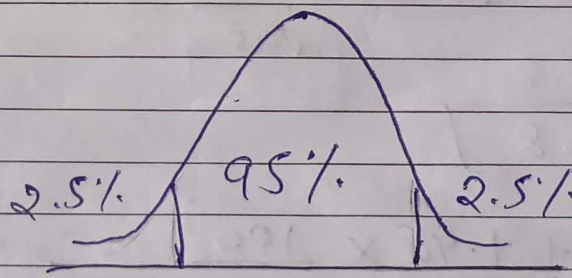


Confidence interval = 95%

Significance value = α 1 - C.I.
= 1 - 0.95 = 0.05

if p < Significance value
we reject the null hypothesis
else:
we fail to reject the null hypothesis.

⇒ Confidence Interval & Margin of Error



C.I. = 95
α = 0.005

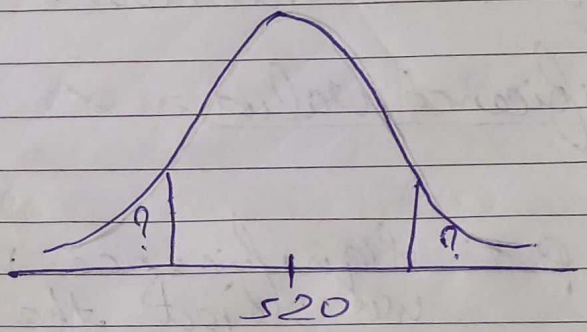
We construct a confidence to help estimate. What is the actual value of unknown population mean.

Point estimate ± margin of error

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

1) In the regional Mean of cat exam, the s.d is known to be 100, A sample of 25 test taken has a mean of 520. Construct a 95% of C.I. about the mean.

$\sigma = 100$
$n = 25$
$\bar{x} = 520$
$CI = 95\%$
$\alpha = 0.05$



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

$Z_{0.025} = -1.96$

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

$= \boxed{480.8}$

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

$= \boxed{559.2}$

I am 95% confident that the mean of scores lies b/w 480.8 and 559.2

⇒ Hypothesis testing & Statistical Analysis

① Z-test } avg value
② t test }
③ Chi square → categorical
④ Anmova → Variance

① Z-test

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

① State null & alternate hypothesis.

② At a 95% confidence level, is there enough evidence to reject the null hypothesis.

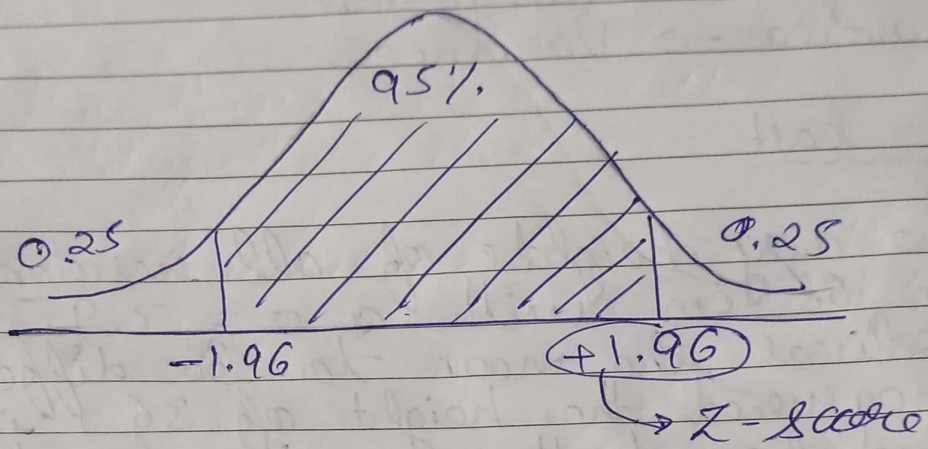
Ans. $\mu = 168$ cm
$\sigma = 3.9$
$n = 36$
$\bar{x} = 169.5$ cm

Whenever population sd is given then we should definitely use Z test.

a) Null Hypothesis : H_0 :- $\mu = 168$ cm
b) Alternate Hypothesis : H_1 :- $\mu \neq 168$ cm.

CI = 0.95
$\alpha = 0.05$

Decision Boundary



Statistical Analysis

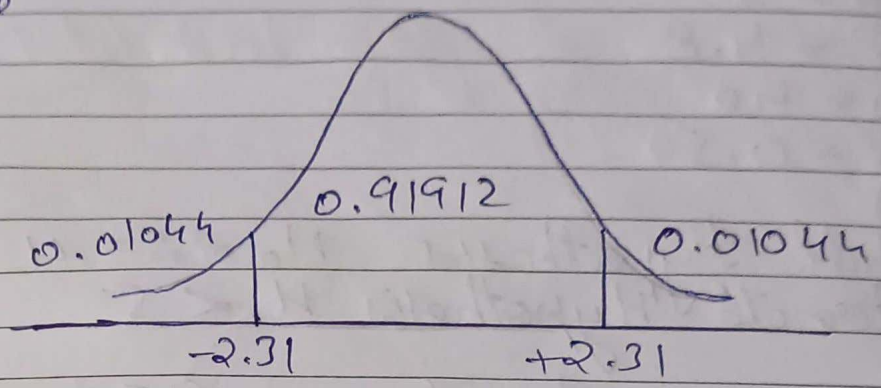
$$Z_{test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \text{Standard error}$$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}} = \boxed{2.31}$$

Conclusion ⇒ If Z-test value is less than -1.96 or greater than 1.96, we reject the ~~null~~ null hypothesis.

$2.31 > 1.96 \Rightarrow$ Rejed the null hypothesis.
The doctor is absolutely right.

$\Rightarrow$ P-value



P value $= 0.1044 + 0.01044 = 0.02088$

if p value < significance value

$0.02088 < 0.05 \Rightarrow$ Reject the null hypothesis.

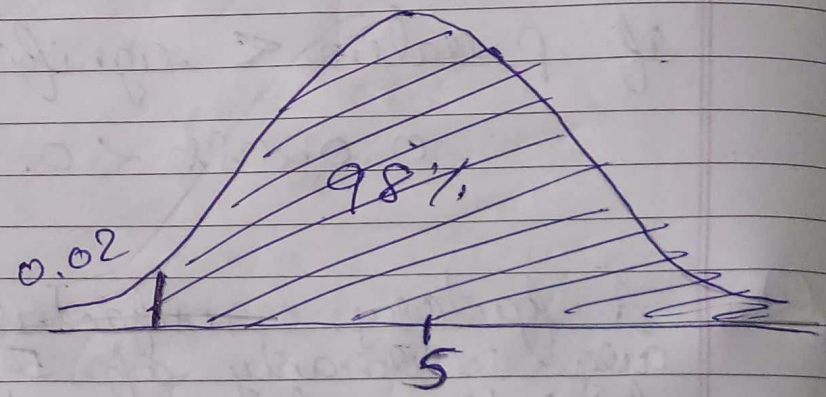
(2) A factory manufacturers bulbs with an avg warranty of 5 years with standard deviation of 0.50. A researchers believes that the bulb will manufacture in less than 5 years. He tests a sample of 40 bulbs and find the avg time to be 4.8 years

a) State null & alternate hypothesis.
b) At a 2% significance level, is there

enough evidence to support the idea that the warranty should be revised?

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

1) Null Hypothesis $H_0 = 5 = \mu$
2) Alternate Hypothesis $\mu_1 < 5$
3) Decision Boundary = C.I. = 0.98
$\alpha = 1 - 0.98 = 0.02$

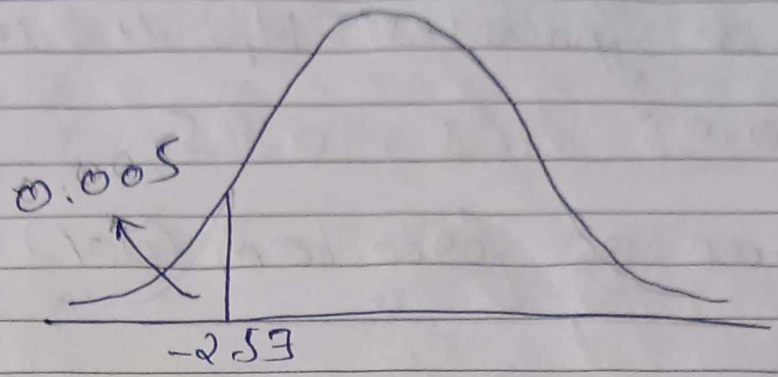


4) Z - test

$$= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.53$$

Z test $< -2.05 \Rightarrow$ Reject the null hypothesis.

5) P-value:



p value < significance value
We reject the H_0.

⇒ T-test

① In the population, the avg IQ is 100. A team of researchers want to test a new medication to see if it has either a +ve ~~to~~ or -ve 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%.

Ans. If population s.d. is not given, use t-test

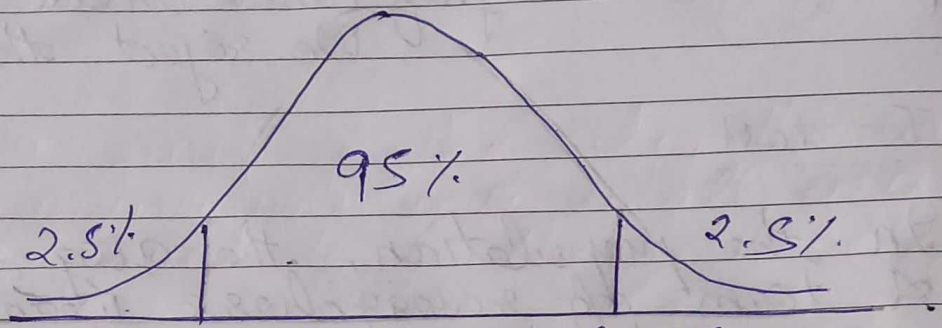
$\mu = 100$, $n = 30$ $\bar{x} = 140$ $S = 20$, C.I. = 0.95
$\alpha = 0.05$

① Null Hypothesis : $H_0 : \mu = 100$
Alternate Hypothesis : $H_1 : \mu \neq 100$ (2 tailed)

② $\alpha = 0.05$ $\quad CI = 0.95$

③ Degree of freedom (dof) $= n - 1$
$= 30 - 1 = \underline{29}$

④ Decision Rule



$\alpha = 0.05$

$t = \pm 2.045$

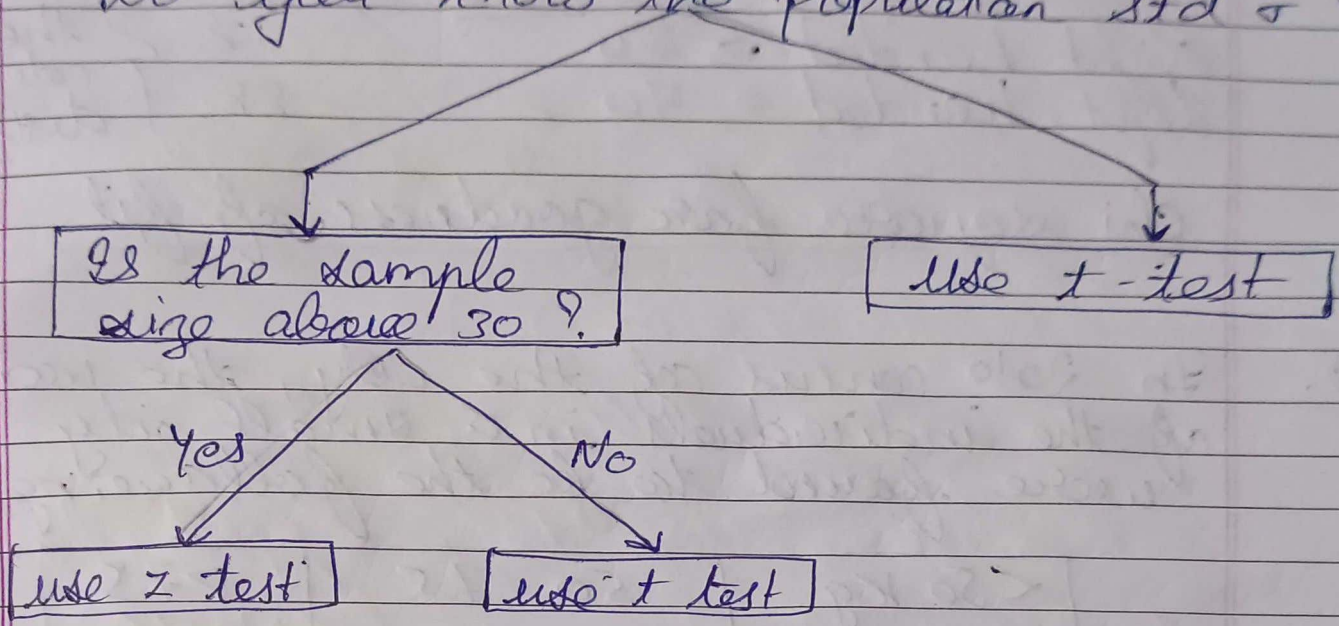
Conclusion: The t-test is less than -2.045 and greater than 2.045. Reject the null hypothesis.

⑤ T-test statistics

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

10.96 > 2.0452. So, we reject the null hypothesis.

⇒ When to use T-tests & z-tests



⇒ Chi-Squared Test

⊙ The chi squared test for goodness of fit claim about population proportion (categorical test).

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

① In a student class of 100 students, 30 are right handed! Does this class fit the theory 12% of people are right handed.

	observed	
Right handed	30	12 } theory
Left handed	70	88 } categorical distribution

Chi square for goodness of fit

Q. In 2010 consus of the city, the weight of the individuals in a small city were found to be the following.

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

In 2010, weight of n = 500 individuals where sampled, below are the results.

<50	50 - 75	>75
140	160	200

using $\alpha = 0.05$, would you conclude the population difference of weights has changed in the last 10 years.

Ans. Expected value

<250 kg	50-75	>75
100	150	250

H_0 → The data meets the expectation.

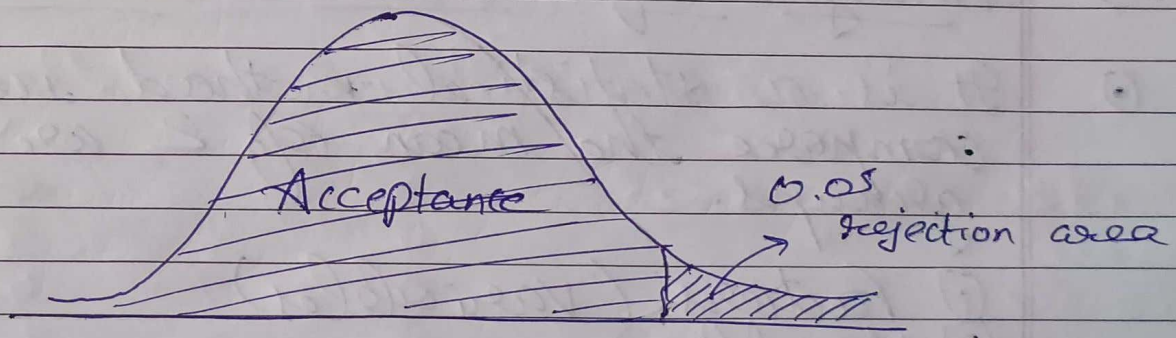
H_1 → The data does not meet the expectation.

$\alpha = 0.05$ CI = 95%.

degree of freedom

$df = k - 1 = 3 - 1 = 2$

Decision boundary → Chi squared test



distribution is always right skewed

critical value = 5.991

If x^2 is > 5.991, we reject the H_0.

Calculate Chi square Test statistics

$$x^2 = \frac{\Sigma (O-E)^2}{E}$$

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

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$= 16 + 0.66 + 10$$

$$\Rightarrow x^2 = 26.66$$

$26.66 > 5.991$, Reject H_0.

⇒ Analysis of Variance (ANOVA)

⊙ It is a statistical method used to compare the mean of 2 or more groups.

① Factors (variables)
② Variables

Eg Medicine → factors

levels → 5mg 10mg 20 mg [damage]

Mode of payment — factors

GPay phonepe Imps NEFT [levels]

→ Types of ANNOVA

① One Way Annova: One factor with atleast 2 levels, then levels are independent.

Eg: Doctor want to test a new medication to decrease headache. They split the participants into 3 condition want to degrees [10, 20, 30].

→ Doctors ask the patients to rate the headache (1-10).

Medication → factor

10 mg	20 mg	30 mg → levels
5	7	2
9	8	7
—	—	—
—	—	—

② Repeated Measures Anova

One factors with atleast 2 levels, levels are dependent.

Running → Factor

Day 1	Day 2	Day 3
8 Km	5 Km	6 Km

③ Factorial Anova

Two or more factors (each of which with atleast 2 levels), levels can be either independent or dependent.

Running → Factor

	Day 1	Day 2	Day 3 ↓
Gender	8	5	6 dependent
Male	7	4	3
Female	6	5	4
	3	2	1

⇒ Hypothesis Testing in Anova

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

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

⇒ F-test Statistics

$$F = \frac{\text{Variation B/w samples}}{\text{Variation within samples}}$$

(1) Doctors want to test a new medication which reduces headache. They splits the participants into 3 condition [15, 30, 45]. Later on the doctor ask the patient to rate the headache between [1 - 10]. Are there any differences between the 3 condition using 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 hypothesis

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

② Alternate Hypothesis:

Atleast one mean is not equal.

③ State Significance value

$\alpha = 0.05 \Rightarrow C.I = 95\%$.

④ Calculate degree of freedom

$N = 21 \quad a = 3 \quad n = 7$

df between $= a - 1 = 2$

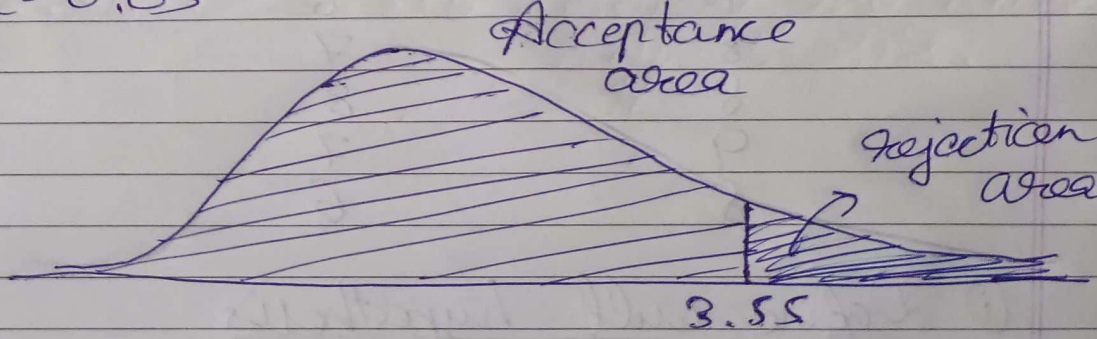
df within $= N - a = 18$

$(2, 18) =$ F test table

df total $= N - 1 = 21 - 1 = 20$.

Decision boundary

$\alpha = 0.05$



if f test is > 3.55, we reject the H_0.

⑤ Calculate F test statistics

	SS	df	MS	F
between	98.67	2	49.34	86.56
within	10.29	18	0.54	
total	108.95	20		

① $SS_{b/w} = \frac{\Sigma (\Sigma a_i)^2}{n} - \frac{T^2}{N}$

15 mg = 9+8+7+8+8+9+8 = 57

30 mg = 7+6+6+7+8+7+6 = 47

45 mg = 4+3+2+3+4+3+2 = 21

T^2

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

$= \boxed{98.67}$

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

$= \Sigma y^2 - \boxed{\frac{57^2 + 47^2 + 21^2}{7}}$

Σy^2 = Sum of square of each example.

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

$= \boxed{10.29}$

(3) $SS_{total} = \Sigma y^2 - \frac{T^2}{N}$

$853 - \frac{125^2}{21} = \boxed{108.95}$

(4) MS

$\frac{\text{Mean Sq between}}{\text{Mean Sq within}}$

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

$86.56 > 3.556$, Reject the null hypothesis.