Statistics: It is a branch of Mathematics that deals with the collection, analysis, interpretation, Presentation, and organization of data.

Types of Statistics.

1. Descriptive:

This involves the collection, analysis and Presentation of data to describe and Summarize its main features.

2. Inferential:

This involves using Statistical methods to make predictions and draw conclusions about a larger Population based on a Sample of data.

Data:

Facts or pieces of information that can be measured.

Population $^{(N)}$ and Sample $^{(n)}$:

Elections → Islamabad

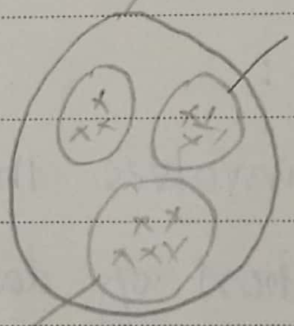
ISL Population

Maximum

⇓

Exit Poll

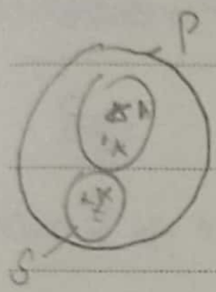
<u>Exit Poll</u>



Sample data

Sampling Techniques :

① Simple Random Sampling :



Every member of the population(N) has an equal chance of being Selected for Your Sample(n).

② Stratified Sampling :

When the Population(N) is Split into non-overlapping groups (Strata).

Eg: Gender < Male / Female — Different type of survey

Age Group < (10-20) / (20,40) / (40,100)

Professions < Doctors / Engineers

Memo No.
Date / /

3. Systematic Sampling:

(N) → n^{th} individual

From the population we take n^{th} individual

Eg: Mall → Survey (covid)

↳ Every 8th person I see → Survey.

4. Cluster Sampling:

Where the population is divided into smaller groups or clusters, and a sample is selected from each cluster.

Variable:

A variable is a property that can take on any value.

Eg:

variable — value.

height = 170

Types:

① Quantatative:

Measured Numerically. { Add, Sub, Multiply, divide

-: height, weight....

② Qualitative / categorical:

Eg: Gender [M F

Based on some characterictics we can drive categorical Variable.

Eg: IQ

0-10, 10-50, 50-100

Less IQ, Medium IQ, Good IQ

Quantitative

Discrete | Contineous

Eg: Whole number

- No of Bank Acc:
- No of family memb:

Eg: Height {172.5, 129.2}

Variable Measurement Scale:

4 types of Measured Variable

1. Nominal data: {Categorical data} → classes
   e.g: Color, Gender, Types of flower
2. Ordinal data: the order of the data matters not values.
   e.g: Rank matters instead of marks.
3. Interval data: Order matters, Value also matter, natural Zero is not Present.
   e.g Temp: 70-80 80-90 90-100 σ won't make difference
4. Ratio data: {Assignment}
   eg: Rose, lily, sunflower, lily

Flower	frequency
lily	3
sunflower	1

Memo No.

Date / /

Bar Graph vs Histogram

Bar is used for discrete data, wheraas, histogram used for contineous data.

• Pdf - Smothing of Histogram

Arithmetic Mean for population and Sample

Mean (average)

Population (N) Sample (n)

Formula:

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N} \qquad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- Useful when the data is numeric and has no extreme value or outliers

Central Tendency:

- Mean
- Median
- Mode

Refer to the measure used to determine the center of the distribution of data.

Memo No.
Date / /

Median:

Median is middle value in dataset. values arranged in lowest to highest.

. It is useful when you have skewed data with extreme values or outliers that might skew the mean.

. It is also useful when data is ordinal.

Mode: The mode is the value that appears most frequently in a dataset.

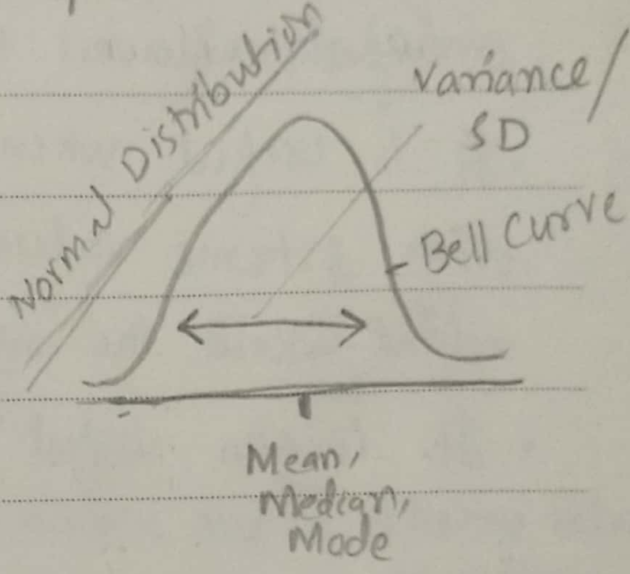
. It is useful when you have categorical or nominal data, meaning that values cannot be ordered and are not numeric.

. It is also useful to know most common or popular value in dataset.

Memo No. ________
Date / /

Measure of Dispersion with Variance and SD

means spread



Mean

Population(N)

$$\mu = \sum_{i=1}^{N} \frac{(x_i)}{N}$$

Sample(n)

$$\bar{x} = \sum_{i=1}^{n} \frac{(x_i)}{n}$$

Measure of dispersion help to provide a more complete understanding of the distribution of data, which can be important for making informed decision and drawing accurate conclusion from data analyses.

Variance:

Population (N)

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

Sample (n)

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\boxed{n-1}}$$

Degree of freedom

Eg:

$X = \{1, 2, 2, 3, 4, 5\}$

x	$\bar{x}$	$(x_i - \bar{x})$	$(x - \bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71

$\bar{x} = 2.83$ 10.84

$S^2 = \frac{10.84}{6-1}$

Sample Variance

$= 2.168$

{Spread of the data}

Mo Tu We Th Fr Sa Su

Memo No.

Date / /

Sample Standard Deviation:

$$S.d = \sqrt{\text{variance}}$$

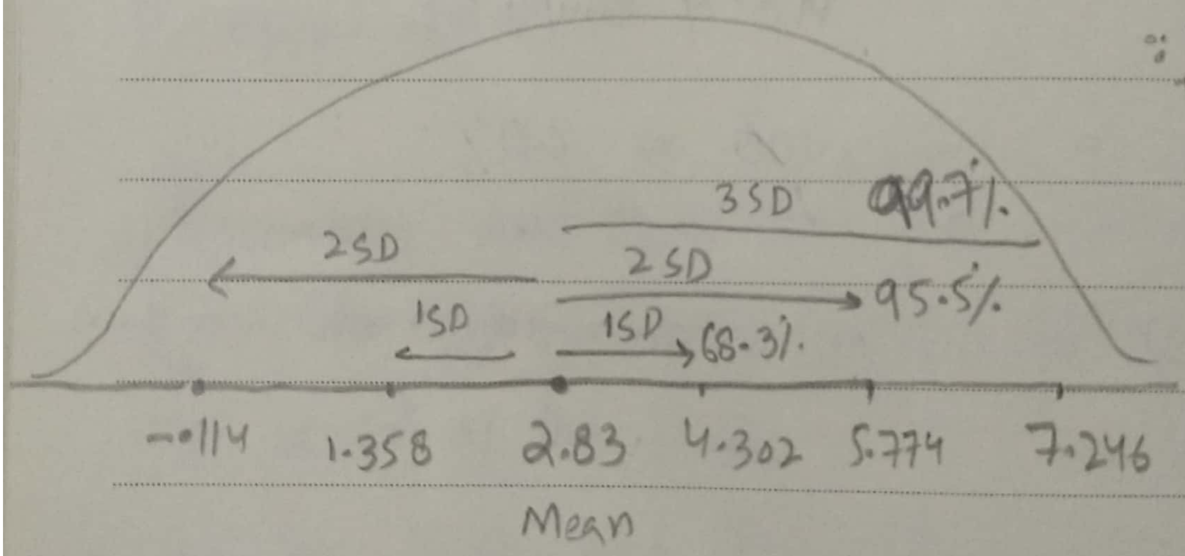
$$= \sqrt{2.168}$$

$$= 1.472$$

Dataset:

$\bar{x} = 2.83 \qquad S = 1.472$

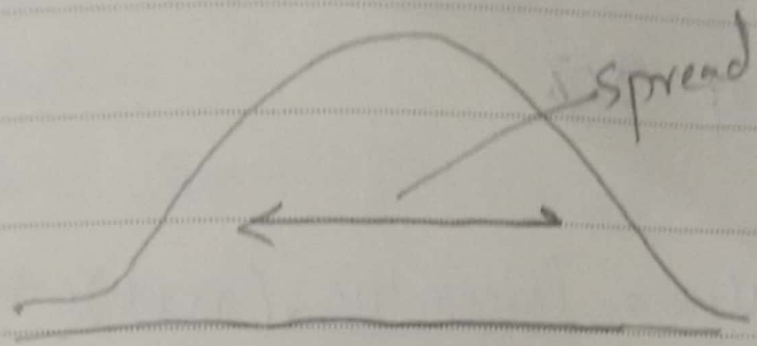
$$\begin{array}{r} 2.83 \\ +1.472 \\ \hline 4.302 \end{array}$$



Left Direction Right Direction

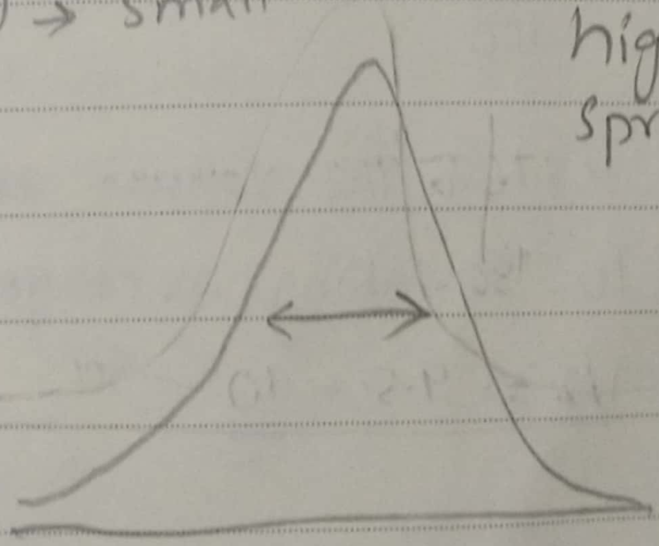
- Variance decides spread of data.

Variance → Big Number
S.d → "



- Variance → Small Number
S.d → small

hight → big
spread → small



- Percentile and Quartiles:

Percentile: A Percentile is a measure of Position that divide a dataset into 100 equal Parts.

Quartiles: Quartiles are Special case of Percentiles that divide a dataset into 4 equal Parts.

∴ Percentiles and Quartiles are useful for identifying outlier and understanding the Spread of the data.

Diff b/w Percentage and Percentile:
Percentage refers to a Proportion of a whole, while Percentile refers to a Position within a dataset.

Percentile

Ex: 1 Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12.

What is the Percentile ranking of 10?

Steps:

1. First Sort the data in ascending or descending order.

Let $x = 10$

$$x = \frac{\text{no: of values below } x}{\text{no of Sample size}} \times 100$$

$$= \frac{16}{\cancel{20}_{1}} \times \cancel{100}^{5} \Rightarrow 80\%$$

80% of the entire distribution are less than 10.

Continue: Percentile And Quartiles:

Ex: 2

What value exist at Percentile ranking of 80%

Formula:

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{80}{100} \times (21) = 16.8 \text{ index position}$$

At index 16.8 the values are 9 and 10. So taking average.

$(9+10)/2 \Rightarrow 9.5 \approx \underline{10}$ — 80%.

Quartiles helps us to identify outliers.

25% → 1st Quartile , Median

75% → 3rd Quartile , Maximum

Mo Tu We Th Fr Sa Su

Memo No. 17
Date / /

→ Five number Summary and Boxplot.

F.N.S

1. Minimum
2. First Quartile (25%) Q1
3. Median
4. Third Quartile (75%) Q3
5. Maximum

⇒ Boxplot
↓
outliers detect

Dataset: Removing the outliers

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

1. First decided [lower Fence ⟷ higher fence]

lower Fence = Q1 − 1.5 (IQR)   ∵ IQR = Q3 − Q1

Higher Fence = Q3 + 1.5 (IQR)

Q1: 25% = (25/100) × (20) ⇒ 5 (index position)

Q2: 75% = (75/100) × (20) ⇒ 15 (index position)

∵ at index 5, value is 3

at index 15, value is 7

P-T-O

Memo No.
Date / /

Mo Tu We Th Fr Sa Su

So, lower fence ⇒ Q1 − 1.5 (IQR)

⇒ 3 − 1.5 (7 − 3)

⇒ −3

Higher Fence ⇒ Q3 + 1.5 (IQR)

⇒ 7 + 1.5 (7 − 4)

⇒ 13

[−3 ⟷ 13]

In dataset there is no value less than −3 but we one value that is greater than 13, that is 27.

So, 27 is outlier.

* Either I can remove 27 or If I have to plot so I have to use Box plot.

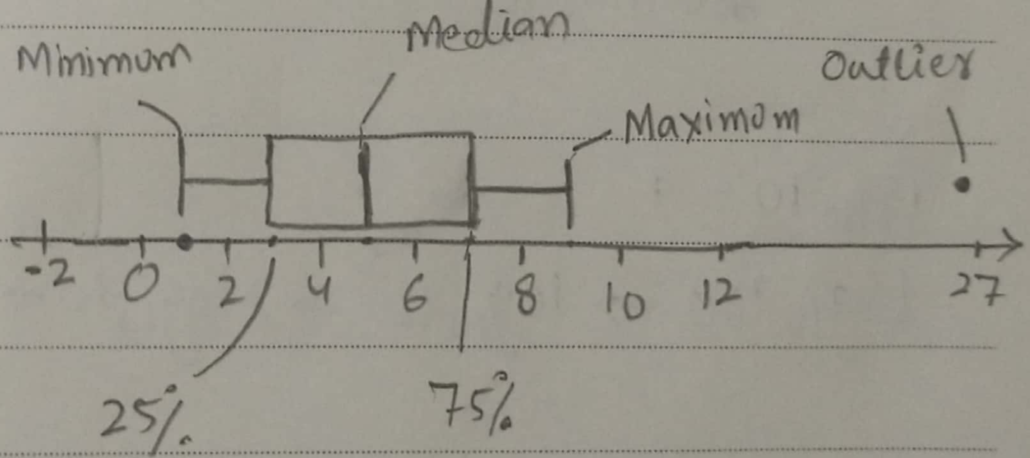
Mo | Tu | We | Th | Fr | Sa | Su

Memo No. ______

Date / /

Box Plot:

1. Minimum = 1
2. First Quartile (25%) Q1 = 3
3. Median = 5
4. Third Quartile (75%) Q3 = 7
5. Maximum = 9



Scanned with CamScanner

Memo No.
Date 10/04/23

Mo | Tu | We | Th | Fr | Sa | Su

Logarithm = logarithm is an inverse of an exponent.

Q. With a initial investment of 5\$ and 5x return, how many years will it take for my money to become 125\$?

$\log_5 125 \Rightarrow 3$

Base investment

$\log_{10} 10 = 1$

$\log_{10} 100 \rightarrow \log_{10} 10^2 \rightarrow 2\log_{10} 10 \rightarrow 2 \times 1 \rightarrow 2$

log y = True helps in comparing values on scale

Memo No.

Date / /

Modified Z-Score

$$MAD = median(|x - median(x)|)$$

$$\text{modified Z Score} = 0.6745 \times \frac{x - median(x)}{MAD}$$

σ – Standard Dev
μ – Mean

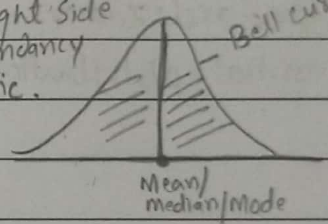
Date:

Distributions

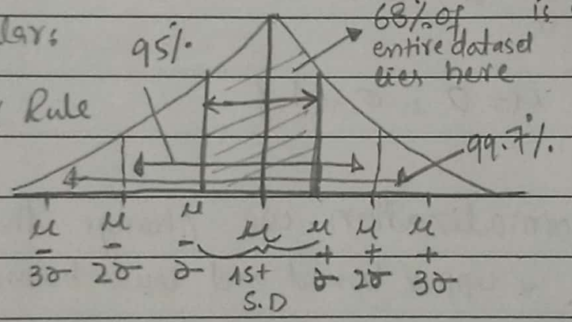
1. Normal Distribution
2. Standard Normal Distribution
   ↳ Z Score
3. Log Normal Distribution
4. Bernoulli Distribution
5. Binomial Distribution

1. Normal Distribution / Gaussian Distribution:
It is used to model continuos data that is symmetric and bell shaped.

left and right side of central tendancy is symmetric.



Emperical formulas:
68%–95%–99.7% Rule



∴ Many real-world phenomena follow a normal distribution, including the height and weight of people.

Z-Score : Z score tells us how many standard deviation a given data point is above or below the mean of the distribution.

$$Z = (x - \mu)/\sigma$$

Standard Normal Distribution: It is a specific type of normal distribution with a mean of Zero and a standard deviation of One.
We can convert normal dist: to standard normal dist: by Z-Score.

Date:

→ Practical Application Standardization:

• DATASET

Age (years)	Salary (Rs)	Weight (kg)
24	40K	70
25	80K	80
26	60K	55
27	70K	45

Standardization: Converting a normal distribution to Standard Normal distribution using Z score is called Standardization.

$$\{ \mu = 0 , \sigma = 1 \}$$

Normalization: In normalization, we change the values between (0 to 1). ∴ upper bound and lower bound.

- To apply normalization, we can Use Min Max Scalar.
- We apply/use normalization in Deep learning.
  CNN → image classification.

Z-table (Standard normal table): It is used to find the area under the curve of the Standard normal distribution for a given Z-score or range of Z-score.

Date:

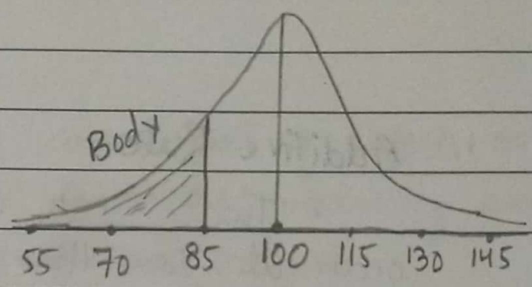
★ Practice Question:

In India the average IQ is 100, with a Std deviation of 15. What Percentage of the population would you have an IQ lower than 85?

Ans=

$Z = \frac{85-100}{15}$

$= -1$



∴ Value of 1 in Z-table : 0.84134

1 - 0.8413 ⇒ 0.1587

15.87% of Population

Date: 7/04/23

Probability

Probability: Probability is a measure of the likelihood of an event.

$$P = \frac{\text{no of way an event can occur}}{\text{no of Possible outcomes}}$$

1. **Additive Rule:**

- Two events are mutual exclusive if the can not occur at same time.

Eg: Rolling a dice → {1, 2, 3, 4, 5, 6}, getting a nomber once at a same time.

- Multiple events can occur at the same time is non Mutual exclusive.

Eg: deck of cards

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

2. **Multiplication Rule:**

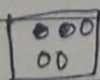
- Independent events

eg: Rolling a dice = {1, 2, 3, 4, 5, 6} ⇒ 1, 1, 2

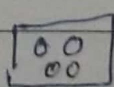
Each and every number is indpendent.

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

- Dependent events



suppose we have 5 marbles, three blue and two black. $P(B) = 3/5$ Now,



$P(B) = 2/4$

First event is dependent on other event.

Date:

$P(A \text{ and } B) = P(A) + P(B|A)$ — conditional Prob.

→ Permution and Combination:

Permutation:

Permutation refers to the arrangement of objects or events in a specific order. It is the number of ways objects or events can be rearranged or ordered.

Formula ⇒ $nP_r = n!/(n-r)!$

n = total no of obj
r = no of obj or events to be arrange

Eg: Arrange 3 out of 5 students in a row to form a Committee.

$$nP_r = \frac{5!}{(5-3)!} \Rightarrow \frac{5 \times 4 \times 3 \times \cancel{2!}}{\cancel{2!}} = 60 \text{ ways.}$$

Combination:

Permutation refers to the Selection of objects or events without regard to their order. It is the number of ways objects or events can be selected without considering thir arrangement or order.

Formula ⇒ $nC_r = n!/(r!(n-r)!)$

Eg: Arrange 3 out of 5 students ~~to~~ form a Study group, withouting considering their order.

$$nC_r = 5C_3 = \frac{5!}{3!(5-3)!} \Rightarrow \frac{5 \times 4 \times 3 \times 2!}{3!\,2!} = 10 \text{ ways}$$

Date:

→ P-Value :

p-value is a measure that quintifies the strength of evidence against a null hypothesis in a hypothesis test. P-Value is a probability value ranges from 0 to 1. It is commonly used in hypothesis testing to determine whether theeir is enough evidence to reject a null hypothesis or not. eg Tossing a coin 100 times to get head $P(H) = 100\%$.

if 50 times Head (The coin is fair)

Hypothesis Testing:

1. Null hypothesis (H0) : Coin is fair
2. Alternate hypothesis (Ha): Coin is unfair
3. Experiment
4. Reject or Accept the Null Hypothesis.

Confidence Interval = 1 - 0.05 = 0.95 (95%)

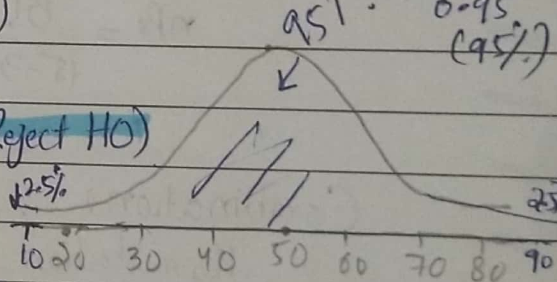
Significance value: 0.05 (or 5%)

if p-value < alpha

means Significant (Reject H0)

if p-value >= alpa

means not significant (Accept H0)



if we get 10 Heads.

Reject H0

~~because 10 head ≠ 0.1~~

We accept Ha

Date:

→ Type 1 and Type 2 error :

- Type 1 Error (False Positive)

A type 1 error occurs when a null hypothesis (H0) is rejected, even though it is actually true. It is denoted by alpha (α), also known as significance level.

e.g : A person awarded death sentence even though he is an innocent.

- Type 2 Error (False Negative).

A Type 2 error occurs when a null hypothesis (H0) is failed to be rejected (accepted), even though it is actually false.

It is denoted by (β).

e.g : A person is not innocent even though he is not awarded death.

. Confusion Matrix :

	P	N	
T	TP	TN	
F	FP	FN	→ T2
	T1		

Date:

→ 1 Tail and 2 Tail Test

2 Tail Test / 1 Tail Test Statement :

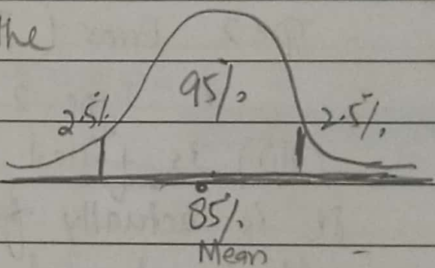
Colleges in Karnatka have an 85% Placement rate. A new college was recently opened and it was found that a Sample of 150 Students had a placement rate of 88% with a Standard deviation of 4%.

Q. Does this college has different Placement rate?

Significance value is $\alpha = 0.05$ (5%)

Confidence interval is (95%). $\therefore 1 - 0.05 = 95\%$

It is 2 Tail test because the rate can be greater or less than 85%

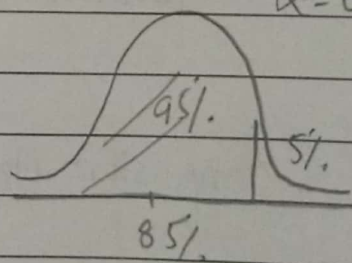


1 Tail Test :

Q. Does this College has a Placement rate greater than 85%?

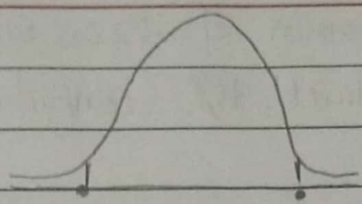
$\alpha = 0.05$

It is one tail Test b/c where focusing on only right side and α value not divided.



Date:

→ Confidence Inerval:



- Point Estimate:
  The value of an Statistics that estimates the value of a Parameter.

  Inferentian Stats:
  From Sample Mean, to estimate Population mean
  $\bar{x} \longrightarrow \mu$

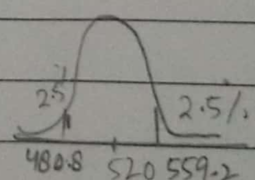
Confidence Interval Formula:

Point Estimate ± Margin of Error

Question: On the Quant test of CAT Exam, the Standard deviation (Population Std) is Know to be 100. A Sample of 25 test takers has a mean of 520 Score. Construct a 95% CI about the mean.

$\sigma = 100 \qquad n = 25 \qquad \alpha = 0.05 \qquad \bar{x} = 520$

① Populatio Std is given
② $n \geq 30$



C.I = Sample Mean ± (Critical value × Standard Error)

$= \bar{x} \pm \left(Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$

$= 520 \pm \left(Z_{0.05/2} \times \frac{100}{\sqrt{25}}\right)$

$= 520 \pm 1.96 \times 20$

$\therefore 1 - 0.025 = 0.975$

$Z_{0.975} = 1.96$

Range (L to U)
480.8 to 559.2

Date:

Question: On the Quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a Standard deviation of 80. Construct 95% confidence interval about the mean.

Data ⇒ $n = 25$ $\bar{x} = 520$ $s = 80$ $\alpha = 0.05$

Condition: if Population Std is not given → t-test.

$$\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

$$520 \pm t_{0.05/2} \times \frac{80}{\sqrt{25}}$$ $\because 1 - 0.025 \Rightarrow 0.975$

$\because$ (df) Degree of freedom $= n - 1 \Rightarrow 25 - 1 = 24$

$$520 \pm 2.064 \times 16$$

$[486.97 \leftrightarrow 553.024]$ Answer.

Date:

→ One Sample Z-test:

1. Population Std is given
2. Sample Size $n \geq 30$

Q: A company claims that the average heighe of its employees is 68 inches. To test this claim, a random sample of 36 employees was taken, and their height was measured. The Sample Mean was found to be 67 inches with a Standard deviation of 3 inches. Conduct a One Sample Z-test with Significance level of 0.05 to determine if there is enough evidence to reject the Company's Claim.

Answer:

Step1: State the hypothesis:

HO: The avg height of Company's employee is 68.

Ha: The avg height of company's employee not equal to 68.

Step2: Given Data

Sample Size(n) = 36

Sample Mean$(\bar{x})$ = 67

Standard deviation = 3 inches

Significance level (α) = 0.05

P.T.O

Date:

Step3: Calculate the test Statistic (Z-Score):

$$Z = (\bar{x} - \mu) / S/\sqrt{n}$$

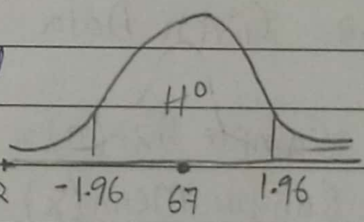
∵ μ = population Mean
$\bar{x}$ = Sample Mean
S = Std deviation
n = Sample Size

$$Z = 67 - 68 / 3/\sqrt{36}$$
$$= -1/0.5$$
$$= -10/5$$
$$Z(\text{score}) = -2$$

Step4: determine Critical Value:
For two-tail test at $\alpha = 0.05$
$\because 1 - 0.05 \Rightarrow 0.975$
From Z-table: critical values are ± 1.96

Step5: Compare the calculated Z-Score with critical value.

Since the Z-Score (-2) is beyond the critical value of -1.96, we reject the Null hypothesis.



Step6: Make a Conclusion:
Based on the result of one Sample Z-test at a Significance level of 0.05, there is enough evidence to reject the company's claim that the avg height of its employees is 68 inches.

Scanned with CamScanner

Date:

→ One Sample T-test

① Unknown Population std

② $n < 30$

Question: Same Question Solved in Z-test with a little changes in Values.

Step 1: State the Hypothesis:

HO: The avg height = 68

Ha: The avg height ≠ 68

Step2: Given Data info

Sample size $(n) = 25$

Sample mean $(\bar{x}) = 67$

Std deviation = 3

Significance level = 0.05

Step3: Calculate the t-score

$t = (67-68) / 3/\sqrt{25}$

$= -1/0.6$

$t(\text{score}) = -1.667$

Step4: Degree of Freedom

$df = n-1$

$= 25-1$

$= 24$

Step4: Critical Value

C.I = 1 - 0.05 = 0.975

df = 24

Critical value is 2.064

Step5: Compare t-score with Critical Value

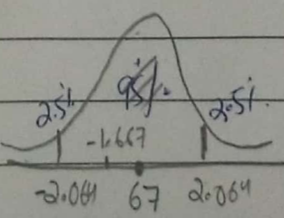
~~-1.667 > -2.064~~

2.064 ≥ 0.05

$t = -1.667 < 2.064$

we accept the HO.

Reject Ha.



Date:

→ CHI Square test: Claims about Population Proportions

It is non parametric test that is performed on categorical (nominal or ordinal data).

Q: In the 2000, indian census, the age of the individual in a small town were found in the following:

less than 18	18-35	>35
20%	30%	50%

In 2010, age of $n=500$ individuals were sampled. Below are the results.

<18	18-35	>35
121	288	91

Using significance value $(\alpha)=0.05$, would you conclude the population distribution of ages has changed in the ~~[illegible]~~ last 10 years.

Ans:

Expected frequencies based on {Population} from 2000.

Age: <18	18-35	>35	n=500
121	288	91	Observed
20% of 500 = 100	30% of 500 = 150	50% of 500 = 250	Expected

Date:

Step ①: State the Hypothesis

HO = The Population distribution of ages have not Changed.
Ha = The Population distribution of ages have changed.

Step ②: State the Significance Value:

$\alpha = 0.05$ (C.I = 95%)

Step ③: Degree of Freedom:

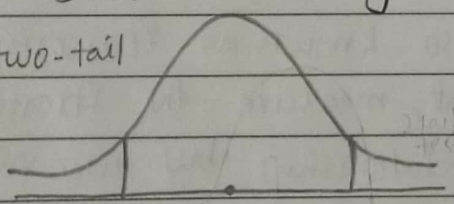
$df = n-1$ ∵ n = no of categories

$= 3-1$

$= 2$

Step ④: Decision Boundry

two-tail



if x^2 is greater than 5.99 reject HO.

Step ⑤: Calculate (Chi-Square test)

$x^2 = \Sigma$ [(Observed frequency − Expected frequency$)^2$ / Exp fre]

$\simeq [(121-100)^2/100] + [(288-150)^2/150] + [(91-250)^2/250]$

$\simeq 232.94$

Step 6: Compare Chi-Square Statistic with Critical value.

$232.94 > 5.99$

{Reject HO}

P-value ≤ Significance value
Reject H0.

Date:

→ Covariance, Pearson Correlation and Spearman Rank Correlation.

1. Covariance: Covariance measures how much two variables change together. It is a measure of the degree to which two variables vary in relation to each other.

+ve	-ve	0
X↑ Y↑	X↑ Y↓	no linear relation
X↓ Y↓	X↓ Y↑	bw X and Y

Formula:

$$Cov(x, y) = \sum\left((x_i - \bar{x})(y_i - \bar{y})\right) / (n-1)$$

2. Pearson Correlation: also known as Pearson's r or simply correlation coefficient, measure the strength and direction of the linear relationship bw two variables.
It range from -1 to 1.

-1	0	1
A perfect negative linear relationship	no linear relationship	A Perfect Positive linear relationship

Formula:

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \ \sigma_y}$$

∵ σ Std deviation

Date:

③ Spearman Rank Correlation: It is used when the relationship b/w variables is not necessarily linear, or when the variables are measured on ordinal or non-Parametric Scale.

$$S(x,y) = \frac{Cov(R(x), R(y))}{R_{\sigma_x} \times R_{\sigma_y}}$$