* Data Science & Analysis:

~~Big Data:~~

* Minor
Data Science:

* Minor
Big data:

Data science & analysis:

Mathematical fundation of data science

1) Linear algebra : Vector & matrixes, matrix operations, graph

Appⁿ – PCA (principle component analysis, ~~SVD~~, SVD, optimisation alogrithms

2.) Probability & Statistics –

Normal, Binomial and Poisson Distribution

Statistical inference and Baye's theorem

Appⁿ – Predictive modeling

3.) Calculas - Differentiation & integration.
(differential calculas) (integral calculas)
understanding
Appⁿ - ML, distribution function and exception,
area, volume

4.) Optimization → Convex optimization & non-convex optimization.

Appⁿ → SVM, Logistic regression,
Simple ML tecnology and training ML

5.) Discrete ~~mathematics~~ mathematics
Appⁿ → Network analysis, cryptography

6.) Numerical Methods

Information gain/
**7.) ~~Infoo~~ Information theory - Entropy, mutual information.
(measure of randomness)
Appⁿ → Training & Testing

8.) ~~Major theories~~
Measure theory and probability

**9.) Graphical model-

10. Mutual Information –
Feature selection

11. Statistics Processes

Date :

- Data Analytics:

1.) Data collection
2.) Data cleaning
3.) Data transformation
4.) Data Analysis
5.) Data Interpretation
6.) Data visualization

- Types of data analytics:

1.) Descriptive data analytics
2.) Diagnostic data analytics
3.) Predictive data analytics
4.) Prescriptive data analytics

- Tools and technology used:

1.) Python programming
2.) Tabelau
3.) Power - BI
4.) SQL
5.) Big data technology - Apache, Hadoop, Spark (for processing large data/datasets)

- Applications:

1.) Marketing
2.) Finance
3.) Healthcare

• Data science & Analysis:

very imp → sample & sample ~~selection~~ distribution

*** → Sampling - Sampling involves selecting a sub-set (sample) from a large population to make an inference. It is done because it is impossible to ~~it~~ collect data about everyone in the population.

→ Types of Sampling:

1.) Random sampling - everyone has equal chance of selection

2.) Stratified " - random sample from each strata

3.) Cluster " - entire cluster is randomly selected

4.) Systematic " - every nth member is selected

5.) Convenience " - sample is selected from conveniently ~~assecable~~ assesable.

↦ Sample distribution - is the pro. dis. of given statistics.

∴ key concept - CLT

↦ Hypothesis Testing - is a statistical method used to make a decisions or inference about a population parameter based on sample data. It involves testing & assumptions (hypothesis) about a population.

∴ Steps in hypothesis testing:

Data Science

→ Hypothesis —.

- Null Hypothesis
- alternative Hypothesis.

→ significance level (α)

→ p value (probability value).

S.L = 5%, 1%, 10% → generally.

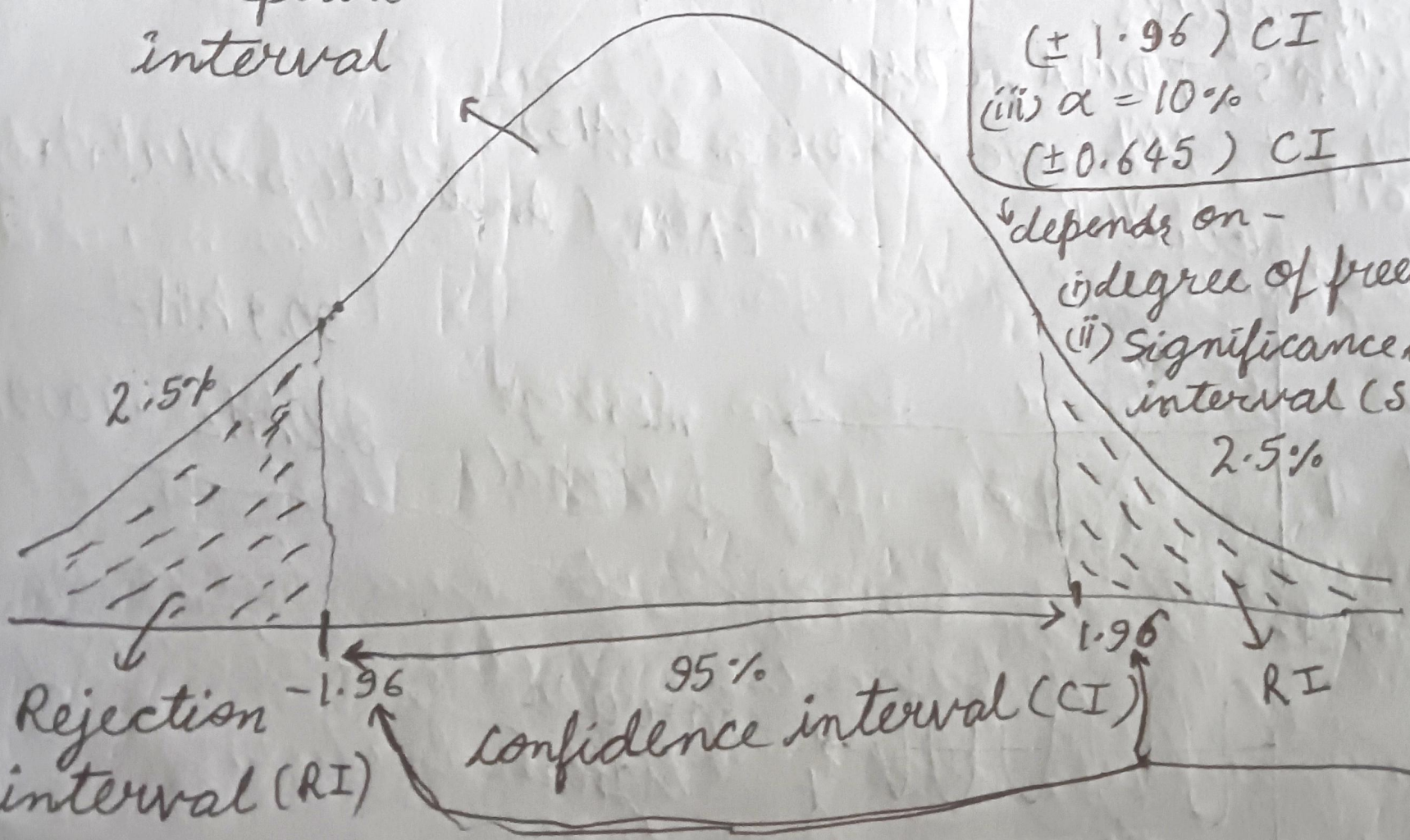
(I) (II) (III)

5% Risk of rejecting null hypothesis.

** → Z-Test : When sample size > 30
Variance is known

(i) $\alpha = 1\%$
(± 2.58) CI
(ii) $\alpha = 5\%$
(± 1.96) CI
(iii) $\alpha = 10\%$
(± 0.645) CI

↳ depends on -
(i) degree of freedom
(ii) Significance level / interval (SI)



$df = n - 1$ ⇒ degree of freedom - cannot be -ve
↓
sample size

If CI is 5% then

• Data Science:

Q:- $n = 100$ (population size), average score = 50

$H_0 = \mu_0 = 50$, $H_1 : \mu \neq 50 \Rightarrow (\mu < 50 \text{ or } \mu > 50)$

$\bar{x} = 30 \Rightarrow$ variance (mean of the sample)

$\sigma = 20$

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

$$Z = \frac{30 - 50}{20/\sqrt{100}} = \frac{-20}{20/10} = \frac{-20}{2} = -10$$

If Z is between confidence interval then it is accepted.

→ T-Test: When sample size < 30
Variance is unknown (standard deviation is given)

standard deviation

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

Q:- $H_0 : \mu_0 \geq 60$ (above 60)

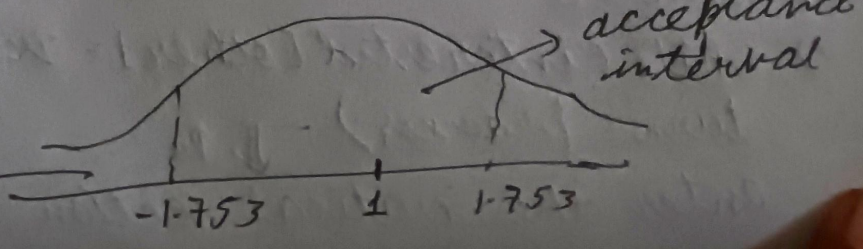
$H : \mu < 60$, $\bar{x} = 70$, $n = 16$, $\alpha = 5\%$, $s = 40$

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{70 - 60}{40/\sqrt{16}} = \frac{10}{10} = 1$$

for -

$df = n - 1 = 16 - 1 = 15$, $\alpha = 5\%$ then,

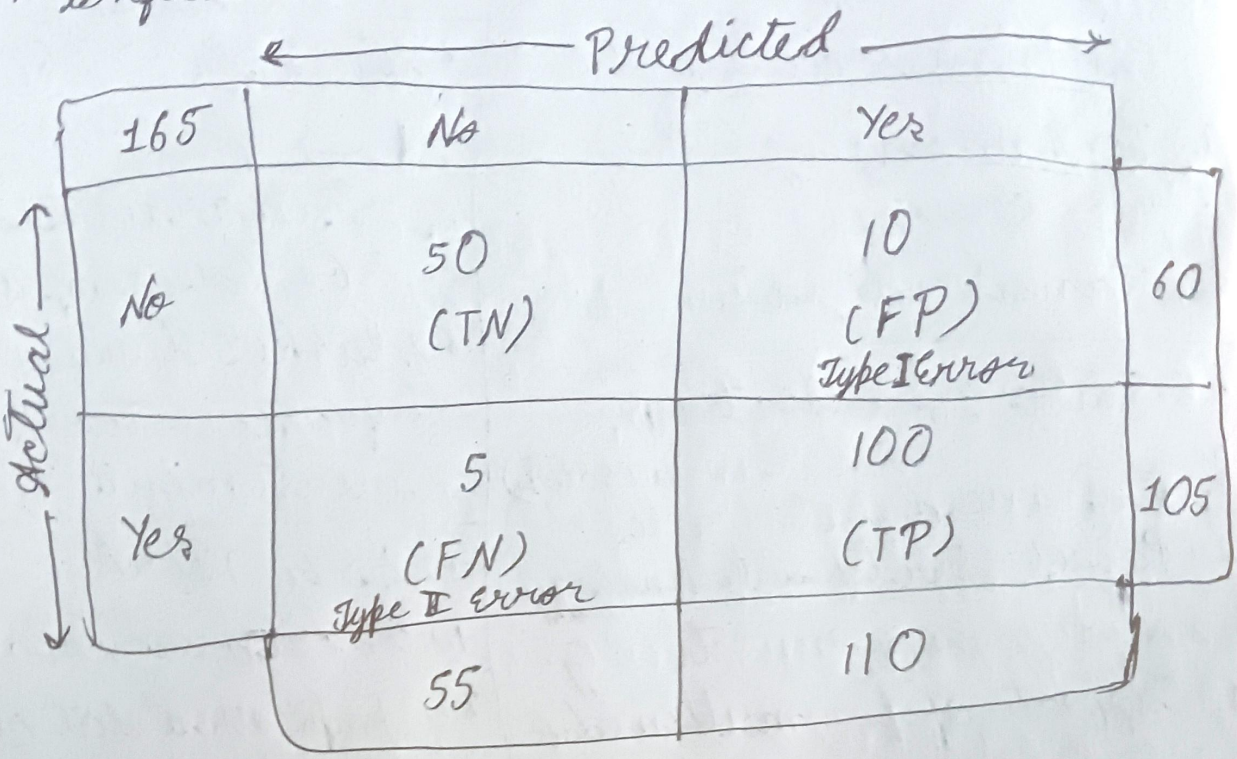
$TC = \pm 1.753$



• Data Science:

→ Training & Testing -

→ ***Major Confusion Matrix: Numerical



(i) Accuracy = $\frac{TP+TN}{\text{Total } (TP+TN+FN+FP)}$

(ii) Precision = $\frac{TP}{TP+FP}$

(iii) Recall = $\frac{TP}{TP+FN}$

(iv) F1-score = $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

- Data Science & Analytics:

1.) Classification
2.) Regression
3.) Clustering
4.) Association

} **Difference