Date.
Page No.

Data Science

EDA Part-2

- **Statistics** →

  - Statistics
    - → Data [Facts/figures]
    - → Information [Knowledge from a certain fact]

- Stats is basically the Science of Uncertainity and the two fundamental ideas in this field are →
  - Uncertainity
  - Variation.

- Statistics
  - Descriptive Statistics — [Presenting, organizing, and summarize]
  - Inferential Statistics — [Drawing Conclusion from a data sample] ← Design/Planning.

Date.
Page No.

- EDA comes under Descriptive Statistics while prediction, forcasting etc comes under Infrencial Statistics.

- Descriptive

  - Describes the data
  - present the data
  - shown in form of Graphs, charts etc
  - Describes what is already known
  - Tools ⇒ mean, median, mode, range, std etc

- Infrential

  - Make Infrence from data
  - predict the data
  - result shown in probability scores,
  - ~~Describes~~ Tries to make conclusion for data.
  - Tools ⇒ hypothesis Test, Analysis of varience etc.

- for Eg ⇒

- The median household income of people aged 25-34 is $35,888 ⟶ Descriptive Statistics

- Drinking decaffinated coffee can raise cholestral levels by 7%. ⟶ Infrencial Statistics.

-

Date.
Page No.

The national avg annual medicine Expenditure per person is $1502

- → Inferencial as they took mean and derived anual costing of medicine.
- → Discriptive is also a good answer as [It may be collected directly]

- Phase - 1 Planning/Designing

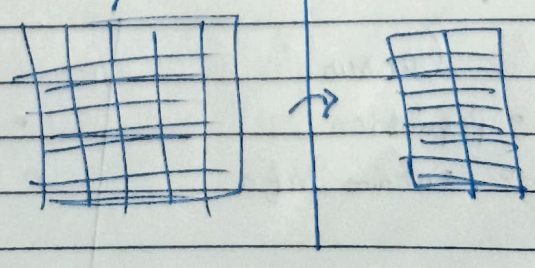
① Purpose for the Analysis.

Purpose → Insights -> Descriptive
Purpose → Result -> Inferencial.

② The Sample to use.

Population vs Sample

Sensus people → fat people.



[How? Sampling Method]

Date.
Page No.

- Observational Study (vs) Experimental Study

Observational Study	Experimental Study
↓	↓
• Data -> Observe	• Compare / Relate
• Observation.	• Inference / Conclusion
	• Recommendation.
• Mostly Cheaper	• More Costly

- Primary (vs) Secondary

Primary	Secondary
• Self made	• Compiled via various sources.
• Paid / Expensive	• Every data on Google is Secondary.
• Time Consuming	• free / Cheap.
• Also called Original Data	• Ready made Data.
• PHD use mostly	• Mostly used data set.
• Stored in SQL / NoSQL	After learning.

- Types of Data

Census	Survey
• Population	• Sampling.
• Contains Info	• Primary data Methods.

Date.
Page No.

- Sampling Techniques →

Probability	Non Probability.
① Simple Random Sampling	① Quota Sampling
② Stratified Random Sampling	② Snowball Sampling (•→•→◉)
③ Cluster Sampling.	③ Judgemental Sampling. [Pros + Cons]
④ Systematic Sampling	④ Convienience Sampling.
⑤ Multistage Sampling	

- Sampling is done to get a small percentage of data that can represent the whole class.

Date.
Page No.

Data Science

Statistics.

Phases ->

- Design & Planning
- Descriptive
- Inferential

Steps for Data Collection ->

⇒ Organize data variable (decide categorical data)
↓
Comes under Feature Eng

→ More Columns ⇒ More cost

→ Conversion of

10 - 100 columns -> 14 - 10 columns
↓
In which 2-3 columns are target columns.

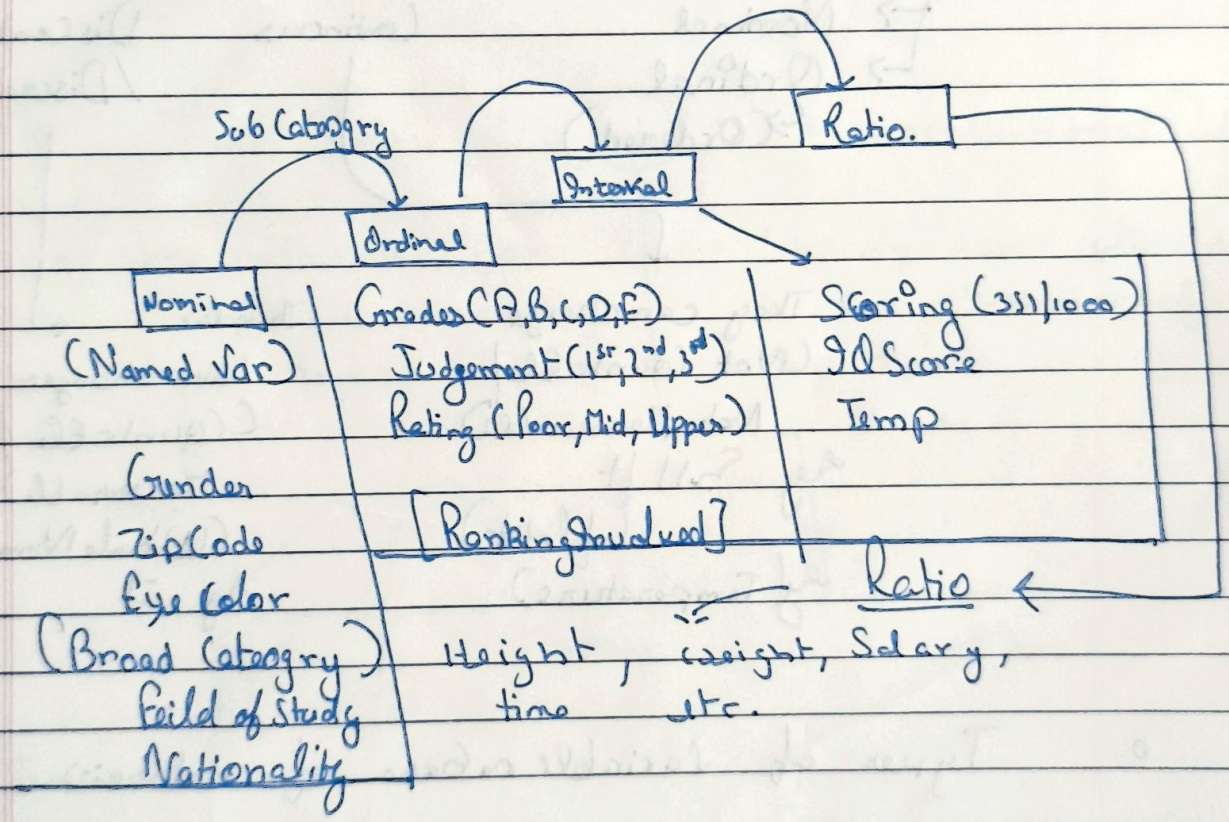
→ The data's purpose is important not the source of origin.

Date.
Page No.

- → Data Scale of Measurement
- → Data type / variable types
- → Types of Study

} Imp

- Scale of Measurement → (Level)



- Sociology ~~and~~ majorly works on Nominal and Ordinal data.

Date.
Page No.

② Data type | ~~Data~~ Variable type.

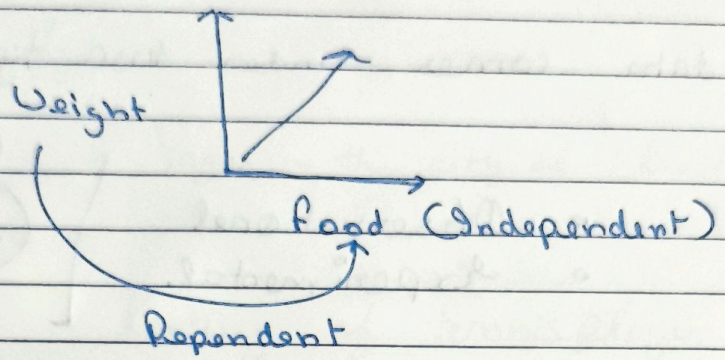
Variable
- Categorical
  - Nominal
  - Ordinal (Ordered)
- Numerical.
  - Continous
    - They can change (Not Countable / Not Measurable)
    - eg Sell ft (float)
    - eg (Temperature)
  - Discontinous / Discrete.
    - The value Never changes (Countable / Measurable) (Whole Number)
    - eg 7

- Types of variable on base of nature :-

Non Working → Dependent
Working → provide cash
Dependent

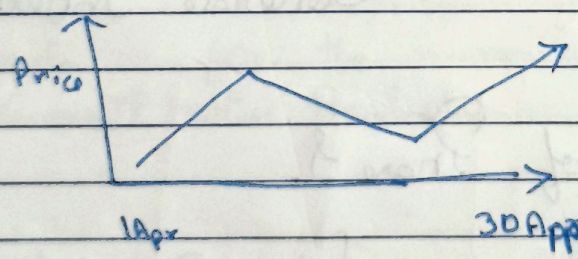
Date.
Page No.

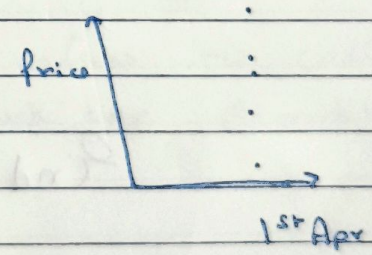
- Dependent Var. (v/s) - Independent Var.



- Time Series (v/s) - Cross Sectional Data.

- Data collected for a specific amount of time
- Set of data collected at the Instance of the time.





Variables are the Part of Data and The Data is of two type.

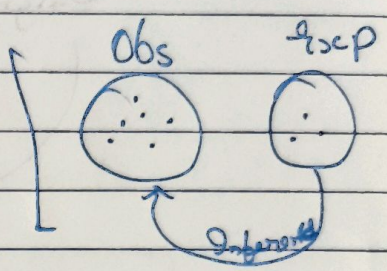
- Primary Data
- Secondary Data.

Date.
Page No.

Types of Study.

All data comes under two types of study

- Observational
- Experimental.



- Observational Study
  ↳ Observe and collect Data

- Experimental Study
  ↳ Observe data and make a conclusion.
  → Generate Inference.

End of Phase 1

[for Data Collection & Planning]

- Phase 2 → Descriptive Statistics.

[Start of Data Analysis]

[EDA]

Date.
Page No.

Statistics

- classify.

  Pages in the city of Cleaveland phonebook.
  - Ratio

  Ranking of tennis player
  - Ordinal.

  Weights of air conditioner.
  - ratio

- When analyzing categorical values it is called Qualitative Analysis, check for Quality.

  while for the numerical values it's called Quantitative Analysis., check for Quantity.

- Types of Data.
  - Primary Data
  - Secondary data.

Date.
Page No.

Phase-II

Descriptive Analysis

- Data Organization
- Data Visualization
- Data Summerization.

→ Data Organization

- Tabulation → Using Tables to store data (structured forms)
- Frequency Distribution → The distribution of frequency across data.

Frequency

Name	hobby	—
XYZ	—	—
ABC	—	—
DEF	—	—
XYZ	—	—

Count of a specific variable / string / Ordinal value etc is called frequency.

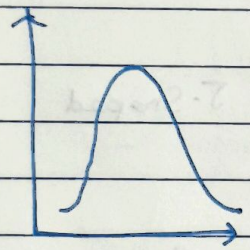
Important when working with categorical values (Ordinal & Nominal)

Date.
Page No.

- Data Visualization

  - Graphical Representation →
    ↳ visual representation of Data.
    for eg.
    Countplot to see distribution.
    Scatterplot for Relation etc.

  - Shapes of Distribution → also can be said as checking of normalization



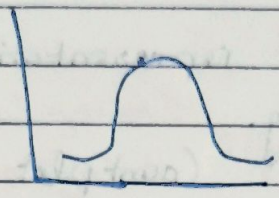
Normal distribution →
- log transformation
- ★ Standard Scaler
- Min Max Scaler
- Square root Transform.

↳ To see distribution
↳ Histogram
↳ Shapiro wilk test
↳ More Acc.
↳ Kolmogrov Smirnov test

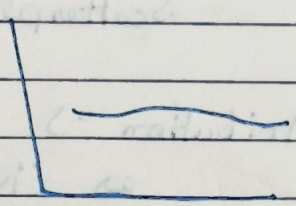
- Types of chart
  - Categorical Data
    - Barchart
    - Pie chart
  - Numerical chart
    - Histogram
    - OGIVE
    - Frequency Polygon

Date.
Page No.

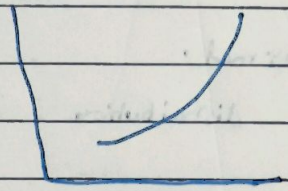
- Shapes of Distribution



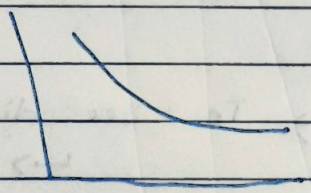
Bell shaped } Normal.



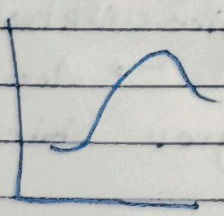
Uniform



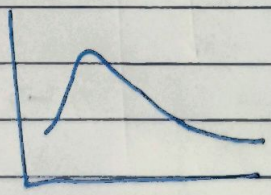
J-Shaped



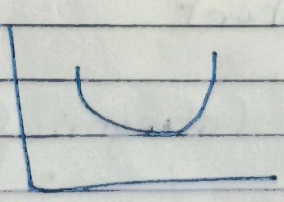
Reverse J Shaped



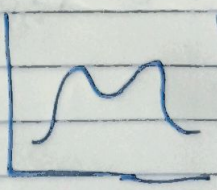
Left Skewed.



Right Skewed



U shaped

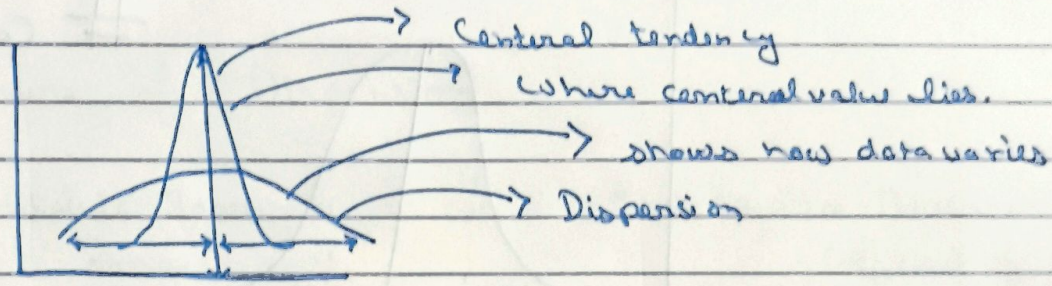


Bimodal

Date.
Page No.

- Summarizing the data.

  - Data is complex, Thus we need values to summarize it's characteristics. Many statistical tests are intended to compare differences in:->
    - Central tendency [typical or avg value]
    - Dispersion [variability in values]



- Central tendency
  - → Mean, Median, Mode
  - → Used for comparing.
- Dispersion
  - → variance, standard deviation, Standard Error etc.

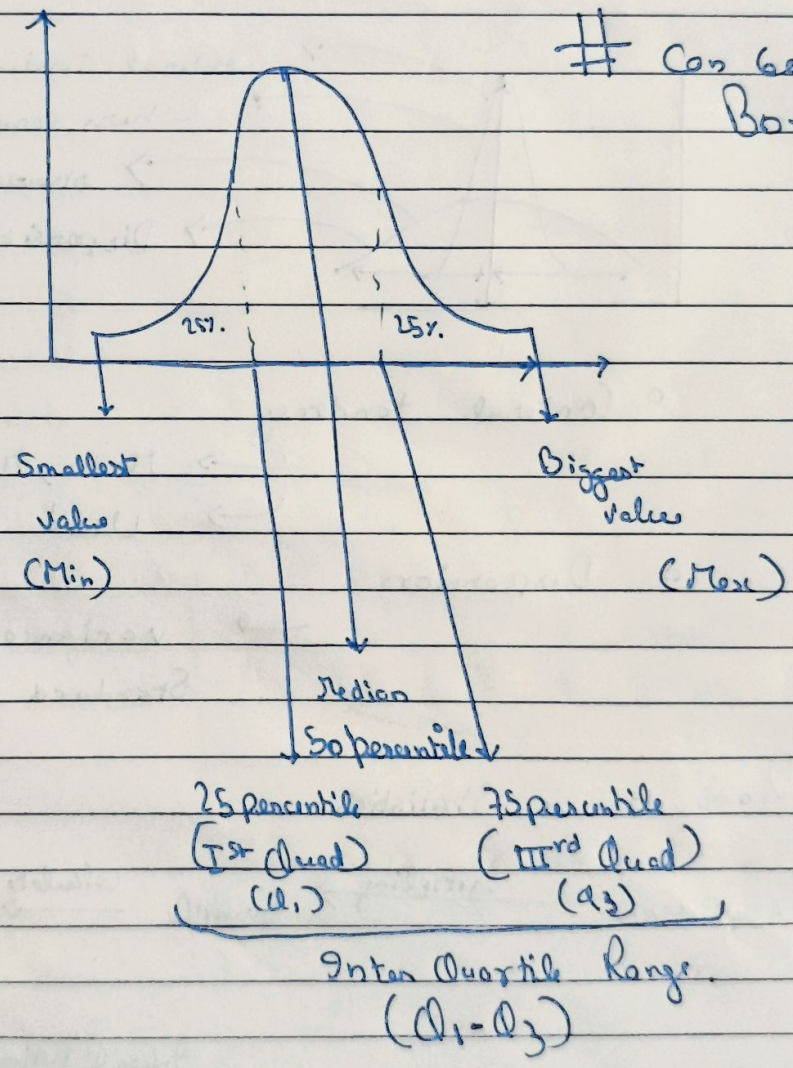
- Parameter vs Statistics.

Population $\xrightarrow{\text{Sampling}}$ Sample $\xrightarrow{\text{Calculate}}$ Statistic (Estimator) [Attribute of Sample]

What we want to know > Parameters (Infer (Estimate)) [Attribute of Population]

Date.
Page No.

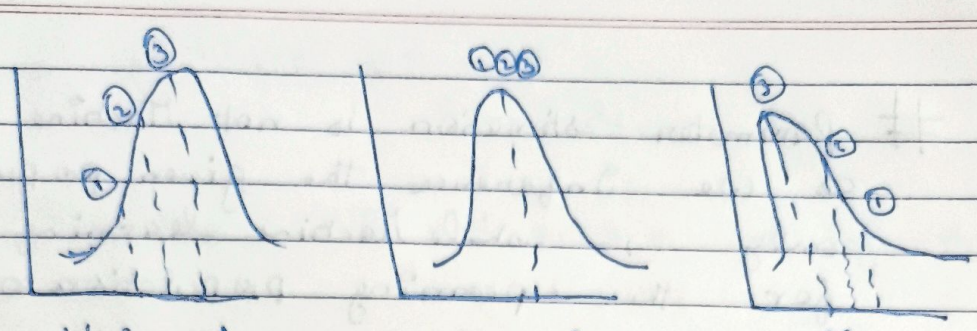
- **Mean** → avg value
- **Median** → Arrange in order
  - get Middle value
- **Mode** → The value that occurs most often.



Can be seen in Boxplot.

- Empirical Relationship:

Date.
Page No.



① Mean ② Median ③ Mode.

- Measure of Dispersion.

- Absolute Measures
  Range,
  Quartile Deviation,
  Mean Absolute Deviation,
  Standard deviation

- Relative Measures
  - Coefficient of Variance
  - Coefficient of Quartile Deviation
  - Coefficient of Mean Dispersion.

- Inferential Statistics ->

Using Sample Statistic to Inference a hypothesis for the given Population.

① Parameter Estimation
② Hypothesis Testing.

Date.
Page No.

Parameter Estimation is not Machine Learning as we Inference the given population only, while Machine learning Inference for the upcoming population only.

Two types of Hypothesis testing ->

- Null hypothesis
- Alternate hypothesis / Alternative hypothesis

> Null hypothesis -> Here everything has some value. [Statistical Analysis]

Eg Height of A = Height of B

> Alternate hypothesis -> Opposite of Null hypothesis.

Eg. Height of A != Height of B.

- Statistical Analysis -> 2 (value) -> t test, z test. (Comparision)

3 -> ANOVA

Date.
Page No.

Let height of A = 5'7"
" " " B = 5'9"

∴ Thus we will not accept Null hypothesis
↳ (Not right way)

Sol -> We reject Null hypothesis
[correct]

This means Alternate hypothesis was accepted.

- Confidence Interval -> If the answer produced confidently or not.

Confidence Interval = (estimation point) [+/-]
(critical value) * [standard error]

for eg it is used in mean

Confidence Interval = 5'5" ± 0'2"
⟹ Error
= 5'7", 5'3"

Date.
Page No.

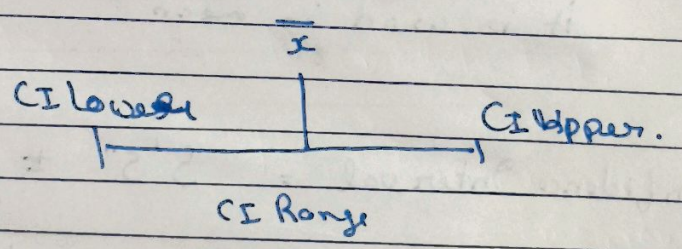
Confidence Interval = $\bar{x} \pm z * (s/\sqrt{N})$

$\bar{x}$ = Mean
z = The chosen z value
s = Sample standard deviation
N = Sample size.

z = based on Confidence level.

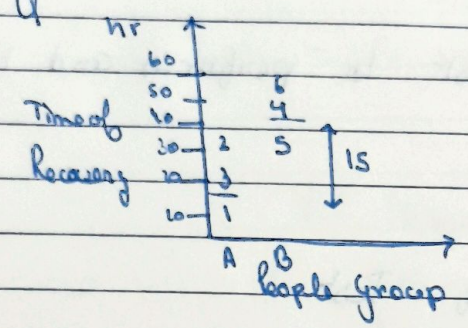
	Confidence Level	z-value
Common [	0.90	1.645
	0.95	1.96
	0.99	2.58.

- Confidence Interval shows the Range of the sample.



Date.
Page No.

- Hypothesis testing.



A -> Took Red pill
B -> Took blue pill

When comparing we compare the mean

∴ It is clear that Blue pill takes more time and Red pill takes less time to help in recovery.

This is a hypothesis that is made by the scientist.

- Hypothesis should be small, clear and concise.

- Thus it Rejects the Null hypothesis

Effect of A = Effect of B
(Inwards)
[There is no difference in Effect of pills A and B. (hr)]

Date.
Page No.

, Now we test the hypothesis if it is true or false.

- Which Statistical test to perform and how?

Types of Test

Parameteric Testing

- More Reliable
- Must meet the assumption

Non Parameteric Testing.

- Less Reliable
- Calculates the Rank of data
- No need to meet the assumptions.

- Step - 1: Normality Test.

(If [p < 0.05] (Rejected)

- Shapiro Wilk test
- Kolmogrov-Smirnov Test.

qqnorm plot is also useful (New)

Shapiro Test ->

H_0 => data = Normal

↳ Null hypothesis

Rejected. (If p < 0.05)

Date.
Page No.

(2) Homogenity Test

Variance of variable in data are equal.

Test to be used Levent's Test.

Homogenity, can be seen via box plot where if there is an odd box plot with different shape, then data is not Homogeneous.

(3) Purpose.

Eg If you want to compare or corelate.

- We shall rather do both.

Purpose

Comparison.

Relationship

Pearson's Correlation.
Spearmans Correlation.

Date.
Page No.

- Comparison
  - 'v/s'
  - (difference b/w)

- Relationship
  - The fact how much A feature change affect other features directly.
    - Connection
    - Correlation
    - Causation
    - Prediction.

Step-4 Datatypes :-

- What type of data dealing with.

• Catogorical	• Numerical
Qualitative	Quantitative
No Numerical	Numerical

Date.
Page No.

- Step 5 → Choosing Statistical Tests
  - Three main families.

	① Chi Squared	2 T-test / ANOVA	3 Correlation
Purpose	Comparison	Comparison	Relationship
Data	Categorical only	Categorical and Numerical (t-test)	~~Con~~ Numeric Only

① Chi Squared Test →

Types →

Chi Squared test for homogenity
" " " " Independence

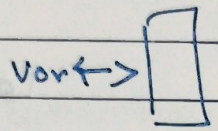
When to use?

- Nothing Effects this
- Can be used with any no. of levels or groups.

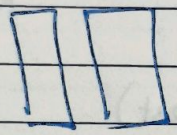
Date.
Page No.

- t-test / Anova

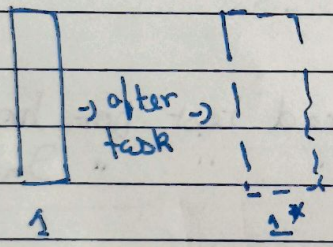
Types ->



- One Sample t-test
  (one sample group with a known mean)



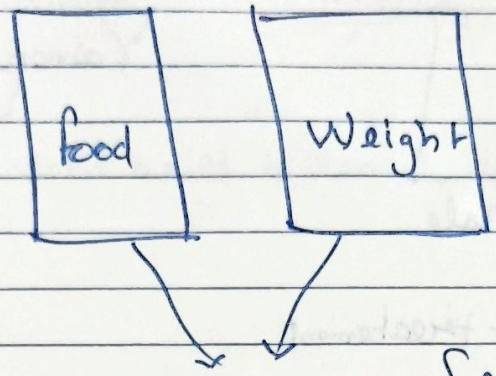
- Two Sample t-test:
  - Unpaired t-test (two diff. grps)
  - Paired t-test (Same grp twice)
- ANOVA (Analysis of variance)
  [3+ levels of groups involved]



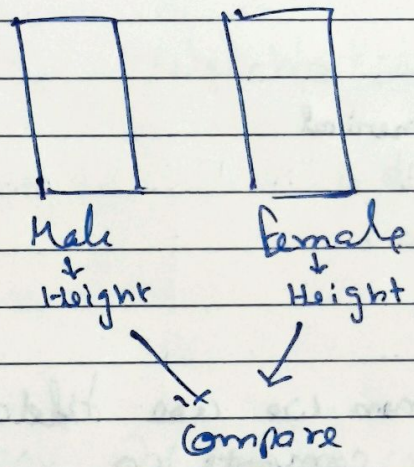
- One Way Anova
  (Even if one group is significant, gives significant results, but doesn't tell you which one)
- Two way Anova
- Repeated measures of ANOVA
  (3+ Paired groups, scale up of Paired t-test).

Date.
Page No.

-)



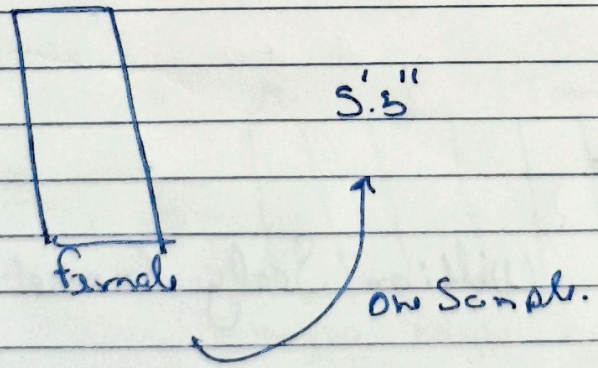
Comparision [Not Sensible]



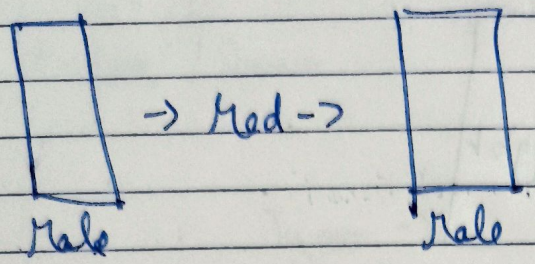
Unpaired t-test

[Sensible]

→ gives pvalue → Then decide the result.
[$p<0.005$] → True [diff]



Date.
Page No.

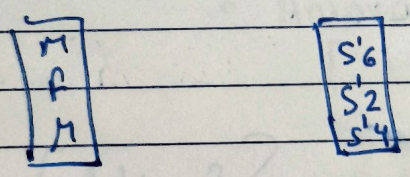


Paired test (t-test)
{ANOVA}

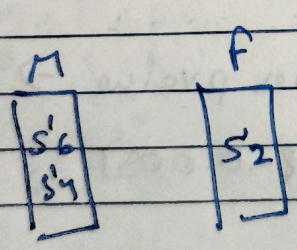
Pre treatment Post treatement

→ How to apply

Categorial Numerical



- When we use tilda (~)
  It converts too



[Height ~ Sex]

- Student's t-test.

by William Sealy Grosset
who worked at
Gunniss Brewery

Date.
Page No.

- Problem -> t-test,

  data should be Normal, Homogenous! But it provides correct variance, correlatibility.

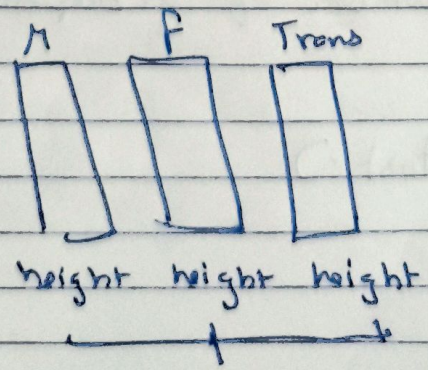
- This was improved by other technique called

  Welch's t-test

  - designed for unequal population variances, but assuming normality is maintained.

- for 2 values we use t-test,
  3 values or more we use ANOVA
  (Analysis of variance)

- ANOVA



To compare we use ANOVA.

Date.
Page No.

- Types of ANOVA ->
  - One way ANOVA
  - Two way ANOVA

- One way gives right answer but doesn't show how?
  To see that we use another test

  Post - ~~Hoc~~ Ho c Test. [gives ranking]
  → Test family.
  - → [Tukey HSD] Imp
  - → Bonferron
  - → Dunet Test

Tukey Key test

gives us the value (pvalue) for individual group with their ranking.

- Two way Anova works like that but we compare two factors among n number of groups.

- Correlation test ->

Date.
Page No.

- Types:-
  - Pearson's Correlation
    [gives us the value of R^2] ← (one Independent, one dependent)
  - Regression
    (one Independent, one dependent)

- Correlation tells us how closely connected two variables are?
- Regression tells us a specific mathematical equation that describes the relationship. (Prediction)

- Important thing:-
  - Your data must be Normally distributed. or should follow a Gaussian Distribution.

If Sample t-test ~~goes~~ should be less than equal to 30, If it goes higher than 30 we use z-test.

Date.
Page No.

- How to progress

(1) Normality $\xrightarrow{\times}$ Normalize Data $\xrightarrow{\times}$ Non Parametric Tests

Normality $\xrightarrow{\checkmark}$ Parametric Tests

Normalize Data $\xrightarrow{\checkmark}$ Parametric Tests

Normalize Data $\xrightarrow{\text{Homogenity.}}$ Parametric Tests

Comparision $\xleftrightarrow{\boxed{\text{Purpose}}}$ Correlation

t-test
- One Sample
- Two Sample
  - Unpaired
  - Paired

ANOVA
- One way ANOVA
- two way ANOVA

- Pearson's correlation
- Regression.

② for Non Parametric tests

Comparision ← Purpose → Correlation

One Sample t test → (One-Sample Wilcoxon Signed Rank test)

Unpaired t-test → (Mann Whitney's U test)

Paired t-test → (Wilcoxon)

ANOVA → (Kruskal-Wallis test)

Pearson's correlation → (Spearman's correlation) & (Kendall's tau)

Regression