"UNIT-01"

DATE: __/__/____
PAGE: _____

Introduction to Big Data Plateform:

⇒ Data is nothing but facts and statistics stored or flee flowing over a network, generally it's raw and unprocessed.
⇒ When data are processed, organised, structured or presented in a given context so as to make them useful, they are called "information".

Data:

⇒ The quantities, character or symbols on which operations are performed by a computer.
⇒ Stored in form of magnetic signals.
⇒ recorded on magnetic, optical or mechanical recording media.

Big Data:

⇒ Big Data is also data but with a huge size.
⇒ Collection of data that is huge in size and yet growing exponentially with time.
⇒ Non of the traditionally data management tools are able to store it or process it efficiently.

Importance:

i) Big data can bring dramatic cost reduction
ii) Substantial improvements in time required to perform a computing task.
iii) New product and service offerings.

Example: New York Stock exchange

1T new data generated every day.

DATE: __/__/____
PAGE: _____

Types of Big Data:

i) Structured
ii) Unstructed
iii) Semi-structured.

Difference between Traditional Data and Big Data:

Tradition Data	Big Data
i) Deals with structured Data.	i) Deals with structured, unstructured or semi-structured data.
ii) Volume ranges from GB to T. (Terabyte).	ii) Volume ranges from Petabytes to Zettabytes.
iii) Generated per hour or per day or more.	iii) Generated per second.
iv) Centralized and ~~mang~~ managed in centralized form.	iv) Distributed and managed in distributed form.
v) Data integration is very easy.	v) Data integration is very difficult.

Good Write

DATE: __/__/____
PAGE: _____

Structured Data:

⇒ Any data that can be stored, accessed or processed in form of fixed format is termed as "structured data".

⇒ An "Employee" table in a database is an example of Structured Data.

Unstructured Data:

⇒ Any data with unknown form or the structure is classified as unstructured data.

⇒ A heterogeneous data source containing a combination of simple text files, images, video etc.

⇒ Example of unstructured data : the output returned by "Google Search".

Semi-structured data:

⇒ Contains both forms of data.

⇒ Actually not defined with a table definition in relational DBMS.

⇒ Example = a data represented in an XML file.

Three Characteristics Difference between Structured, Unstructured and Semi-structured data:

Structured	Unstructured	Semi-structured
i) It is based on Relational database table.	i) It is based on character or binary data.	i) It is based on XML / RDF.
ii) Versioning over tuples, row, table	ii) Versioned as a whole.	ii) Versioning over tuples or graph is possible.
iii) It is schema dependent and less flexible.	iii) It is more flexible and there is absence of schema.	iii) More flexible than structured and less flexible than unstructured data.
iv) Allows complex joining.	iv) Only textual queries possible.	iv) Queries over anonymous nodes are possible.
v) Matured transaction	v) No transaction management	v) Transaction is adapted from DBMS.

DATE: / /
PAGE:

Three Characteristics of Big Data:

i) Volume → Data quantity.
ii) Velocity → Data Speed.
iii) Variety → Data Types.

Challenges faced by conventional systems:

i) Large quantities – impossible to process
ii) Meaningful and collected in real-time.
iii) Multiple sources
iv) Collect correct data.

Conventional System: System consists of one or more zones each having either manually operated call points or automatic detection devices or combination of both.

Challenges:

i) Uncertainity of Data Management landscape.
ii) Big Data Talent Gap.
iii) The talent gap that exists in the industry getting data into the big data platform.
iv) Need for synchronization across data sources.
v) Getting important insights through the use of big data analytics.

Other three challenges:

i) Data
ii) Process
iii) Management

Good Write

DATE: __/__/____
PAGE: _____

Challenges of Big Data:

i) Meeting the need for speed.
ii) Visualization helps organizations perform analyses.
iii) The degree of granularity increases.
iv) Understanding the data.
v) Addressing data quality.
vi) Displaying meaningful results.
vii) Dealing with outliers.

Intelligent Data Analysis: (IDA).

IDA: An interdisciplinary study concerned with the effective analysis of data.
⇒ Used for extracting useful information from large quantities of online data;
⇒ Extracting desirable knowledge or interesting patterns from existing database.

⇒ At a level of abstraction higher than the data, and information on which it is based and can be used to deduce new information and new knowledge.

(*) Goal: Extract useful knowledge, the process demands a combination of extraction, analysis, conversion, classification, organization, reasoning, and so on.

DATE: __/__/____
PAGE: _____

Uses of IDA:

i) Data visualization
ii) Data pre-processing
iii) Data engineering
iv) Database mining techniques, tools & applications
v) Big data applications.
vi) Evolutionary algorithms.
vii) Machine Learning.
viii) Neural Nets.
ix) Fuzzy logic
x) Statistical pattern recognition.
xi) Post-processing

Importance of IDA:

i) Decision making
ii) Data processing
iii) Epidemiological study.
iv) Multi dimensionality of problems is looking for methods for adequate and deep data processing and analysis.

Illustration of IDA by using Sec5:

⇒ application.names : lists the classes to which cases may belong and the attributes used to describ

⇒ application.data : provides information on the training cases from which Sec5 will extract patterns.

⇒ application.test : provides information on th test case.

Good Write

Nature of data:

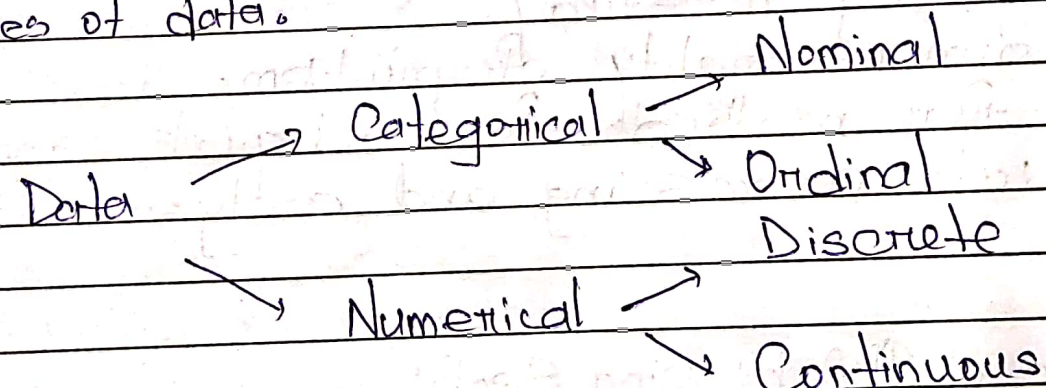
⇒ Set of values of qualitative or quantitative variables; restated pieces of data are individual pieces of information.

⇒ Measured, collected and reported and analyzed

Properties of Data ⇒

i) Amenability of use
ii) Clarity
iii) Accuracy
iv) Essence
v) Aggregation
vi) Compression
vii) Refinement.

Types of data:



1) Nominal: The characteristics of nominal are as follows

- Order: The order of the response or observati does not matter.
- Distance: Nominal scales do not hold distanc The distance between a 1 and 2 is no the same as a 2 and 3.
- True Zero: There is no true or real zero. In nomical, zero is uninterruptable.

Good Write

DATE: __/__/____
PAGE: ____

- Appropiate statistical for nominal scale: mode, count and frequencies.
- Display: histograms or bar chart.

ii) Ordinal Scale: The characteristics are as follows.
- Order: The order of the response or observation matters.
- Distance: Ordinal scales do not hold distance. The distance between first and second is unknown as is the distance between first and third along with all observations.
- True zero: There is no true or real zero. An item, observation or category cannot finish zero.
- Appropiate statistics: mode, count and frequencies.
- Displays: histogram or bar charts.

iii) Interval Scale: Classical interval scale are Likert scale (eg 1 - strongly agree and 9 - strongly disagree).
Semantic interval scale are (1 - dark and 9 - light).
The characteristics are as follows -
- Order: The order of response or observation does matter.
- Distance: Interval scales do not offer distance.
- True zero: There is no zero. However, data can be rescaled in a manner that contains zero.

Good Write

DATE: __/__/____
PAGE: _____

- Appropiate statistics: count, frequencies, mode, median, mean, standard deviation, skewness and kurtosis.
- Displays: histograms or bar chart, line charts and scatter plots.

iv) Ratio scale: Ratio scale appear as nominal scale with a true zero.

The characteristics are as follows:

- Order: The order of the response or observation matters.
- Distance: Have an interpretable distance.
- True Zero: There is true zero.
- Appropiate statistics: count, frequencies mode, median, mean, standard deviation, skewness and kurtosis.
- Displays: histograms or bar chart, line charts and scatter plots.

DATE: __/__/____
PAGE: _____

Analytic process and tools:

There are 6 analytic process:

i) Deployment:
- ⇒ We deploy the result of analysis.
- ⇒ Monitoring and maintenance.
- ⇒ Produce a final report and review the project.

ii) Business understanding:
- ⇒ Determine the business objective.
- ⇒ Assess the situation.
- ⇒ Determine the data mining goals
- ⇒ Produce the project plan.

iii) Data exploration:
- ⇒ Need to gather initial data, describ and explore the data and verify data quality to ensure it contains the data we require.
- ⇒ Data collected from the various sources is described in terms of it's application

iv) Data preparation:
- ⇒ Need to select data as per need, clean it, construct it to get useful information.
- ⇒ Integrate it all

DATE: __/__/____
PAGE: _____

v) Data Modeling:

⇒ Select a modeling technique, generate test design, build a model and assess the model built.

⇒ Analyze relationships between various selected objects in the data.

⇒ test cases are built for assessing the model and model is tested and implemented.

vi) Data evalution:

⇒ At the end of analysis, we evaluate the quality of the measurements and results comparing them to our original design criteria.

Analysis and Reporting:

- Analysis = The process of exploring data or reports in order to extract meaningful insights which can be used to better understand and improve business performance.

- Reporting? = The process of organizing data into informational summaries in order to monitor how different areas are performing.

DATE: __/__/____
PAGE: _____

Difference between Analysing and Reporting:

Analysing	Reporting
i) Provides what is needed	i) Provides what is asked for.
ii) Typically customized.	ii) Typically standardized.
iii) Involves a person.	iii) Does not involve a person.
iv) Extremely flexible.	iv) Fairly flexible.
v) Transforms data and information into insights.	v) Translates raw data into information.
vi) Interpret the data at a deeper level and providing actionable recommandations.	vi) Helps companies to monitor their online business and alerated to when data falls outside of expected range

Modern Data Analytic tools:

i) Hadoop: helps in storing and analyzing data.
ii) MongoBD: used on datasets the change frequently.
iii) Talend: used for data integration & management.
iv) Cassandra: a distributed database used to handle chunks of data.

Good Write

DATE: __/__/____
PAGE: _____

v) Sparks: used for real time processing and analysing large amount of data.
vi) STORM: an open source real time computational system.
vii) Kafka: a distributed streaming problem that is used for fault-tolerant storage.

Statistical Concepts:

⇒ Data consists of information coming from observations, counts, measurements or response.
⇒ Statistics is science of collecting, organizing, analysing and interpreting data in order to make decisions.
⇒ Population is the collection of all outcomes, response, measurements, or counts that are of interest.
⇒ Statistics is a piece of data from a portion of a population.

Basic statistical operations:

i) Mean: A measure of central tendency for Quantitative data i.e. the long term average.
ii) Median: A measure of central tendency for Quantitative data i.e. the half-way point.
iii) Mode: The most frequently occurring (discrete), or where the probability density function peaks.

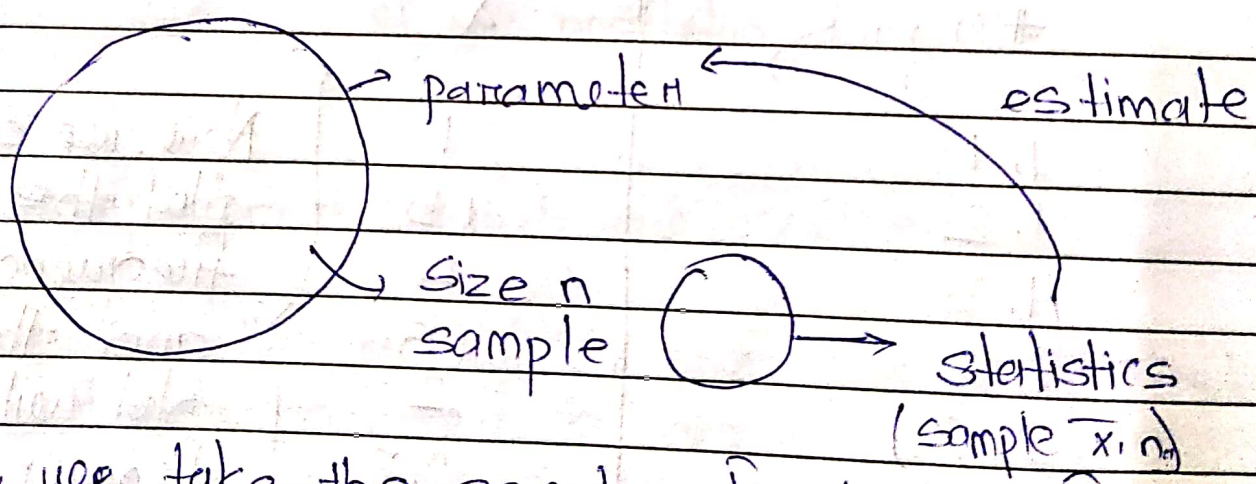
Sample Distribution:

Sample: A sample is "a smaller" (but hopefully representative) collection of units from a population used to determine truths about that population.

⊛ Types of Samples ⇒

i) Probability Samples: Systematic random sample
Stratified random sample
Multistage sample
Multiphase sample
Cluster sample.

ii) Non-probability Samples: Convenience sample
Purposive sample
Quota



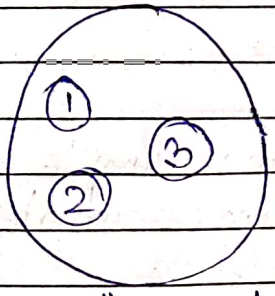
Here, we take the sample of size n from a population; calculate the statistics and estimate the parameter of the population.

Good Write

DATE: __/__/____
PAGE: _____

The distribution / frequency with which we can get different values for the statistics that is trying to estimate this parameter is called "Sampling distribution".

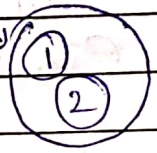
Example



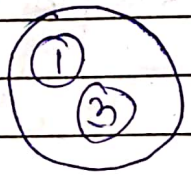
Population parameter here. $\mu = \frac{1+2+3}{3} = 2$

-three balls population

* taking samples like



$\rightarrow \bar{X} = 1.5$



$\rightarrow \bar{X} = 2$

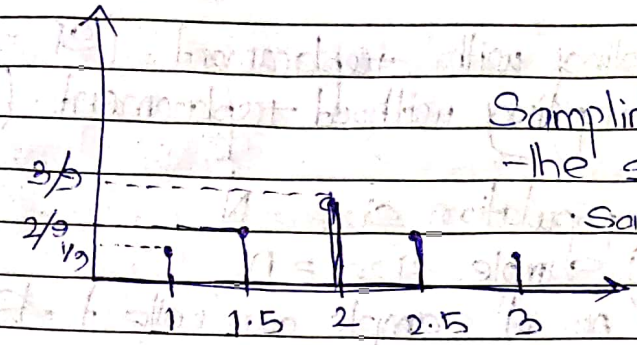
likewise.

#'s pick	$\bar{X}$
1, 1	1
1, 2	1.5
1, 3	2
2, 1	1.5
2, 2	2
2, 3	2.5
3, 1	2
3, 2	2.5
3, 3	3

Now, we can plot the frequencies and the plot will be known as "Sampling distribution" of the sample means.

Good Write

DATE: __/__/____
PAGE: _____



Sampling distribution for the sample mean for sample size of 2.

Specific types of samples:

i) Stratified sample: Ensures that each segment from the population is represented.

ii) Cluster sample: All members from randomly selected segments of a population.

iii) Systematic sample: A sample in which each member of the population is assigned a number.

iv) Convenience sample: Consists only of available members of the population.

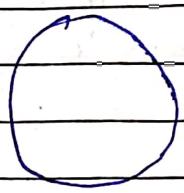
Sampling distributions are used to:

⇒ calculate the probability that sample statistics
⇒ could have occurred by chance.
⇒ to decide whether something that is true of sample statistic
⇒ come to conclusions about a population

DATE: __/__/____
PAGE: _____

① Simple random sampling with replacement (SRSWR)
② Simple random sampling without replacement. (SRSWOR)

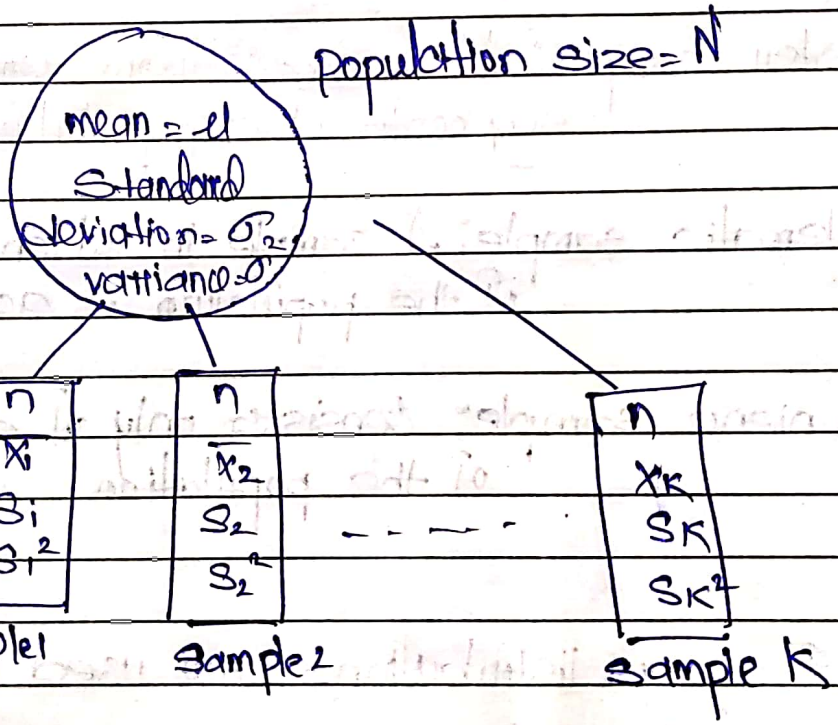
If,



population size = N
Sample size = n
no. of sample can collect = K

For SRSWR, $K = N^n$
For SRSWOR, $K = {}^N C_n$

Standard error:



Sampling distribution of mean

$\bar{x}_1$	$P(\bar{x}_1)$
$\bar{x}_2$	$P(\bar{x}_2)$
:	:
$\bar{x}_K$	$P(\bar{x}_K)$

→ Standard deviation of sampling distribution of mean is "Standard Error" of mean.

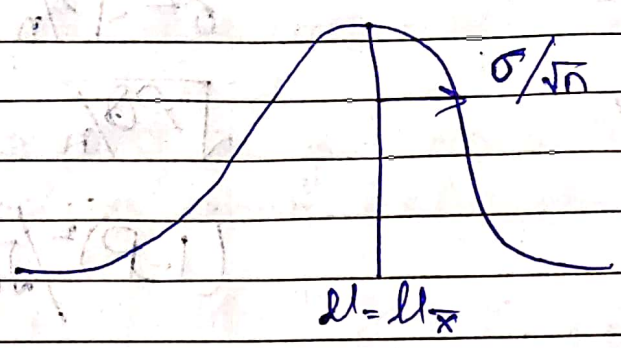
Good Write

DATE: __/__/____
PAGE: _____

* Mean of sampling distribution of mean, $\mu_{\bar{x}} = \mu$ (population mean)

* Standard deviation of sampling distribution of mean or Standard error of mean, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

σ = standard deviation of population
n = sample size



* Uses:

~~1) Setting condif "Confidence Interval"~~

1) To set the "Confidence Interval"
   ↳ (the mean of estimate plus and minus the variation in that estimate).

ii) Test of Significance
   ↳ (the procedure for comparing observed data with a claim the truth of which being assessed).

DATE: __/__/____
PAGE: _____

Statistics	Standard Error
(i) Sample mean, $\bar{x}$	$\sigma/\sqrt{n}$
(ii) Sample standard deviation, s	$\sqrt{\sigma^2/2n}$
(iii) Sample variance, s^2	$\sigma^2\sqrt{2/n}$
(iv) Sample proportion, P	$\sqrt{PQ/n}$
(v) Sample correction, r	$(1-P)^2/\sqrt{n}$
(vi) Difference of two sample mean $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

⇒ The Central Limit Theorem says, "for random sampling, as the sample size n grows, the sampling distribution of $\bar{x}$ approaches a normal distribution.

⇒ The sampling distribution will be normal no matter what the population distribution's shape as long as $n > 30$.

Good Write

DATE: __/__/____
PAGE: _____

Re-sampling:

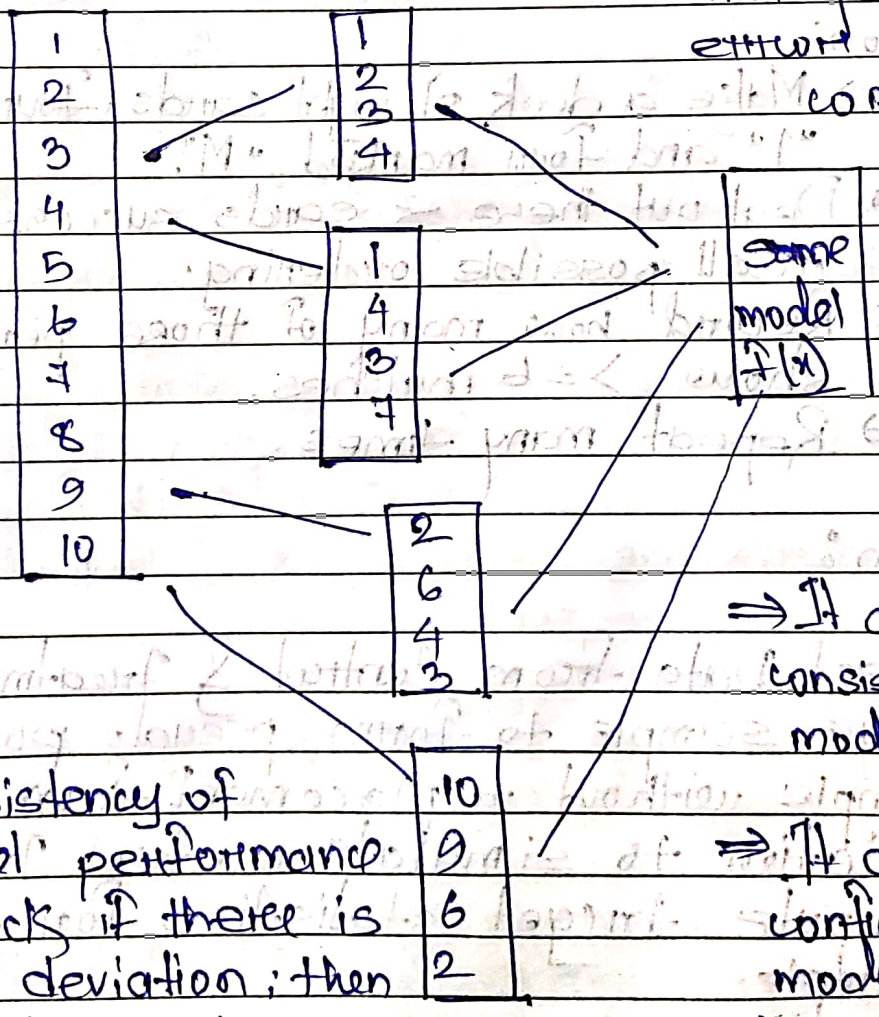
The method that consists of drawing repeated samples from the original data samples.

⇒ It is a non-parametric method of statistical inference.

⇒ Re-sampling uses experimental methods, to generate the unique sampling distribution.

⇒ Re-sampling is performed to get additional information about model fitness.

example:



⇒ Finding standard error and confidence interval

⇒ Consistency of model performance

⇒ Checks if there is high deviation; then there is something wrong with the model.

⇒ It checks the consistency of model statistics (parameter).

⇒ It checks the confidence of the model so that the model is not overfitted for the data.

Good Write

DATE: __/__/____
PAGE: _____

Re-sampling methods:

There are four major re-sampling methods available:

1) Permutation: "Fisher's tea Taster

- 8 cups of tea are prepared
  - four with tea poured first
  - four with milk poured first
- The cups are presented to her in random order.
  - let's say T T T T M M M M.

Solution:

⇒ Make a deck of eight cards, four marked "T" and four marked "M".
⇒ Deal out these 8 cards successively in all possible orderings.
⇒ Record how many of these permutations show >=6 matches.
⇒ Repeat many times.

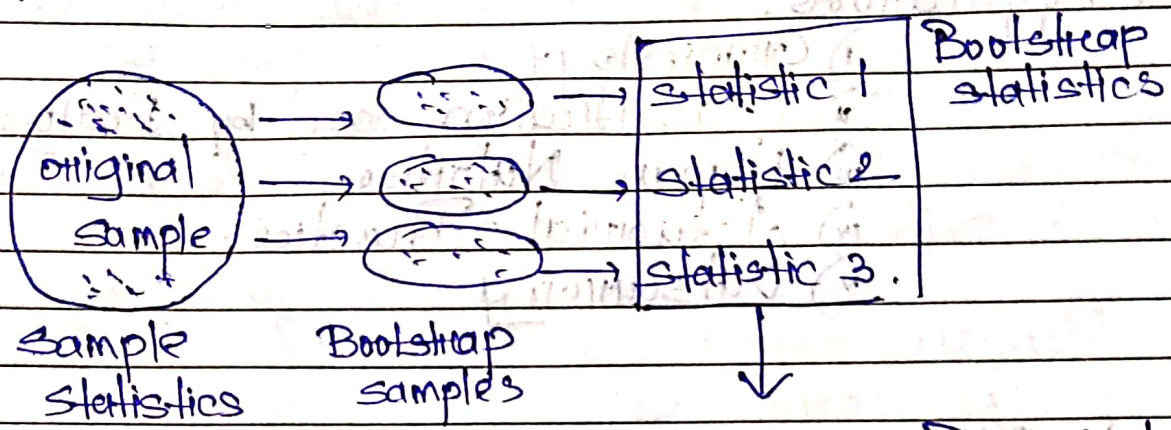
Algorithm:

Step 01: Collect Data from Control & Treatment group
Step 02: Merge sample to form pseudo population.
Step 03: Sample without replacement from pseudo population to simulate control Treatment.
Step 04: Compute target statistics for each example.

DATE: __/__/____
PAGE: _____

1) Bootstrap: Bootstrap generates distinct data sets by repeatedly sampling observations from the original dataset.

⇒ These generated data sets can be used to estimate variability in lieu of sampling independent data sets from the full population.



Algorithm: A bootstrap sample is a random sample conducted with replacement

step 01: Randomly select an observation from the original data.

step 02: "Write it down"

step 03: "Put it back" (i.e. any observation can be selected more than once).

step 04: Repeat (step 1-3) N times; N is the number of observations in the original sample.

∴ Final Result: One "bootstrap sample" with N observation.

Good Write

DATE: __/__/____
PAGE: _____

⇒ Bootstrapping makes no assumption regarding the population.
⇒ No normality of error terms.
⇒ No equal variance.
⇒ Allows for accurate forecasts of intermittent demand.
⇒ Applications:
  i) Criminology
  ii) Classification used by Ecologist
  iii) Human Nutrition
  iv) Actuarial Practice.
  v) Outsourcing

Difference between Parametric Bootstrap and non-Parametric Bootstrap.

Parametric	Non-Parametric
i) Generates samples from the (assumed) distribution of the data using the estimated parameter values.	i) Generates its sample by sampling with replacement from the observed data.

DATE: __/__/____
PAGE: _____

Comparison between Bootstrap and Jackknife.

⇒ Bootstrap can be viewed as closely related method of Jackknife.

⇒ Bootstrap is 10 times more computationally intensive than Jackknife
⇒ Bootstrap is conceptually simpler than Jackknife.
⇒ Bootstrap perform well than Jackknife.
⇒ Jackknife is more conservative than Bootstrap.
⇒ Jackknife produces same result each time while Bootstrap gives different result.

iii) Cross Validation: A technique used to protect against overfitting in a predictive model.

⇒ In cross validation, we make a fixed number of folds (or partitions) of data, run the analysis on each fold, then average the overall error estimate.

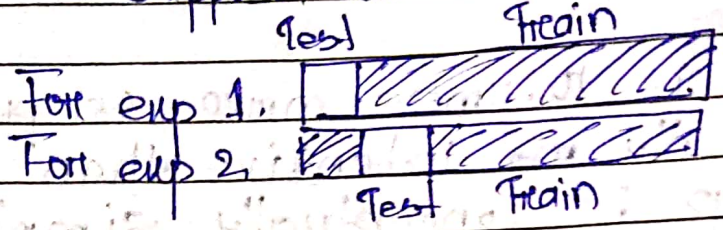
⇒ It is not a re-sampling technique.
⇒ Requires large amount of data.
⇒ Useful in data mining and AI.

DATE: __/__/____
PAGE: _____

Types of Cross Validation:

i) Leave One Out Cross validation:

Suppose, we have 1000 records (data set).

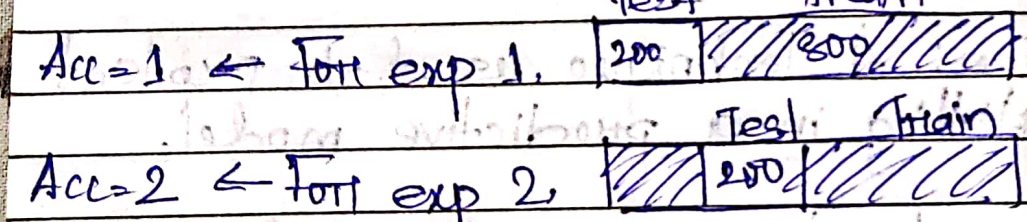


Check each data, so we have to do many iterations and it will have low bias.

ii) K-Fold cross validation:

Suppose, we have 1000 records.

Consider the random value of $K=5$



Iterations will be equal to K.

The minimum accuracy getting in this model will be the minimum of all the accuracy and

The maximum accuracy getting in this model will be the maximum of all the accuracy.

DATE : __/__/____
PAGE : _____

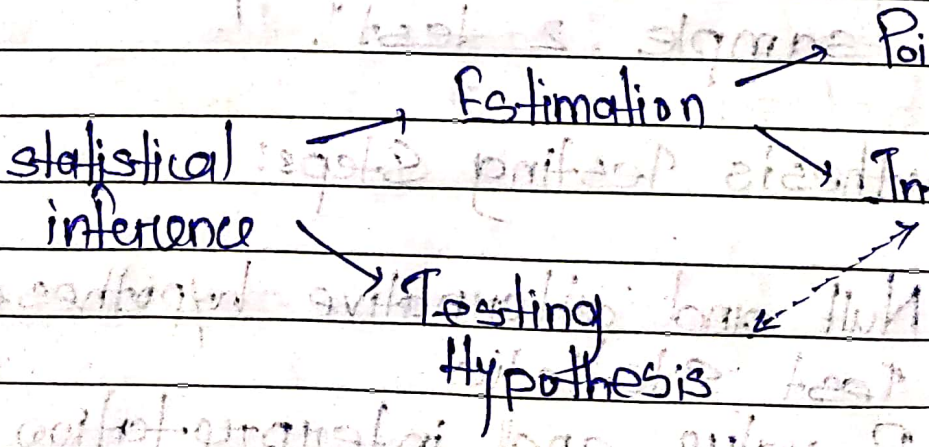
iii) Stratified Cross validation:

No of instances of each class should be proportionate.

There should be proportionate amount of "Yes" "No" / "0,1" data in test set or training set.

Statistical Inference:

- Inference: Use a random sample to learn something about a large population.
- Statistical inference: The process of making guesses about the truth from a sample.
- Classification:



DATE: __/__/____
PAGE: _____

Types of Statistical Inference:

i) Confidence Intervals:

Range of values = m is expected to lie within 95% confidence interval.

= 0.95 probability that m will fall within range

= probability is the level of confidence.

Finding:

$$\bar{x} - z_{cv}(\sigma_{\bar{x}}) \leq \mu < \bar{x} + z_{cv}(\sigma_{\bar{x}})$$

lower limit — Upper limit

ii) Test of Significance (Hypothesis testing):

⇒ Sample data to evaluate a hypothesis about a population parameter.

⇒ A hypothesis is an assumption about the population parameter.

⇒ The technique is introduced by considering a "one-sample z test".

⊛ Hypothesis Testing Steps:

a) Null and alternative hypothese.
b) Test Statistic
c) P-value and interpretation
d) Significance level.

Good Write

DATE: __/__/____
PAGE: _____

* Null Hypothesis: A type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations.

* Difference between Null Hypothesis and Alternative Hypothesis:

Null Hypothesis	Alternative Hypothesis
i) It denotes there is no relationship between two measured phenomena.	i) It's a hypothesis that random cause may influence the observed data on sample.
ii) It is represented by H_0.	ii) It is represented by H_a or H_1.
iii) Statement example ⇒ Rohan will win at least Rs. 10000 in lucky draw.	iii) Statement example ⇒ Rohan will win less than Rs. 10000 in lucky draw.

* Prediction Error: Prediction error is the failure of some expected event to occur.

Prediction error = actual outcome − predicted
(Y) (Ŷ)

DATE: __/__/____
PAGE: _____

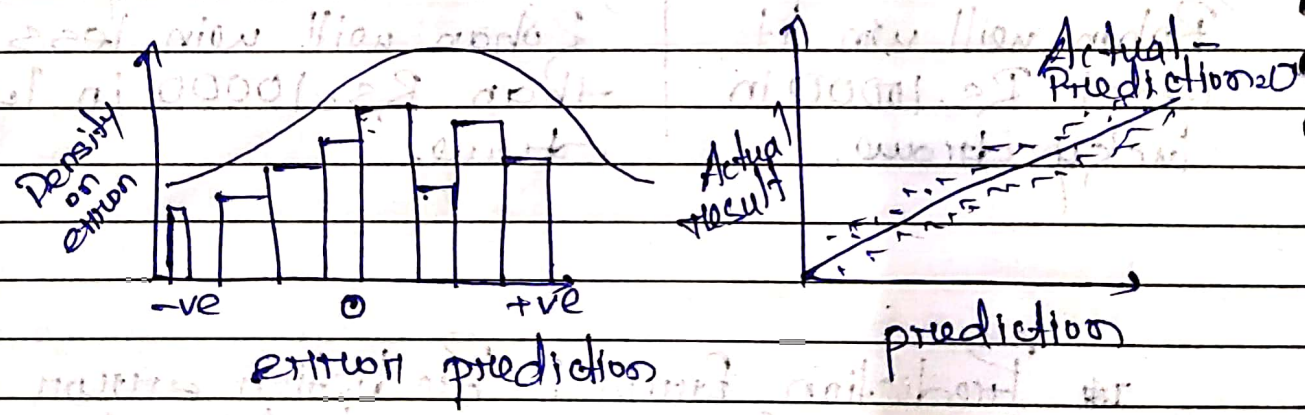
④ Bias = $\frac{\Sigma \epsilon_i}{\text{no. of predictions}}$ | ϵ = errors.

⑤ Root Mean Square Error

$$= \sqrt{\frac{\Sigma (\text{prediction error})^2}{\text{no. of prediction}}}$$

⇒ Bias tells us which direction our predictions are going. It's not good for the magnitude of the prediction.

⇒ RMSE gives us the magnitude of the prediction.



Regressions differing in accuracy of prediction. The standard error of the estimate is a measure of the accuracy of predictions. The regression line is the line that maximizes the sum of squared deviations of prediction.

Good Write