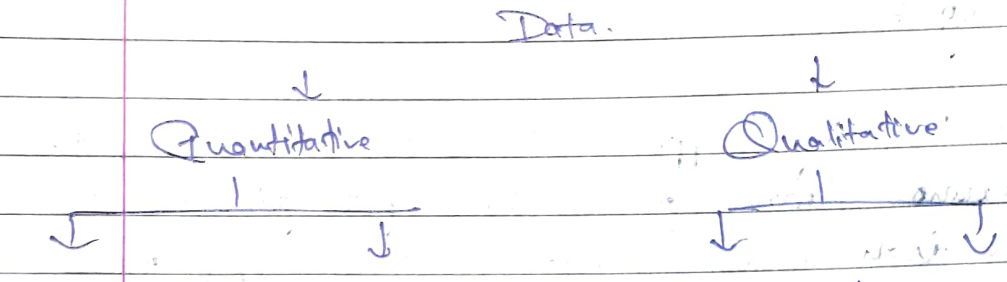
KOSH

Page No.:

Date:

Statistics.




## Qualitative Data.

→ Deals with characteristics & descriptors that cannot be easily measured but can be observed effectively.

Nominal → Data with no inherent order

Ordinal → Data with an ordered series.

## Quantitative Data

→ Deals with nos & things that could be measured objectively

Discrete = AKA Categorical data, holds finite no of possible values. (No of students in class)

Continuous = Data can hold infinite no of possible values.
(Weighing scale.

KOSH

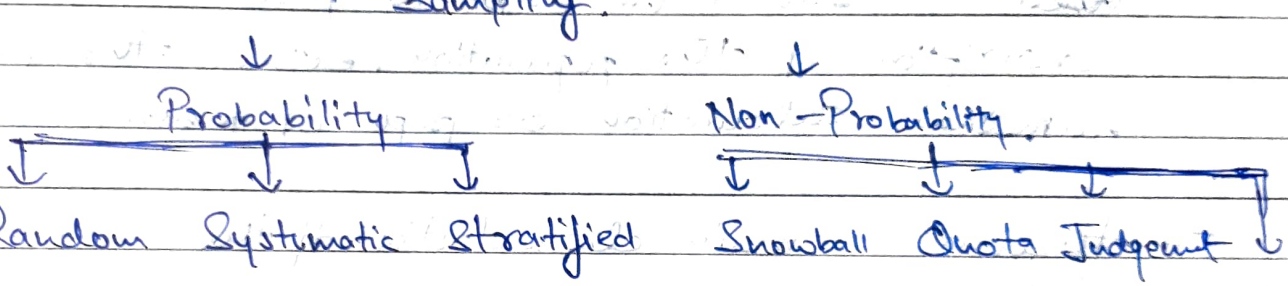
Page No.:

Date:

Population - Collection or set of individuals or objects or events whose properties are to be analysed.

Sample - A subjet of population

Sampling - Is a technique that deals with the selection of individual observations within a population. Used to infer statistical knowledge about population.




Sampling : A sampling is a technique in which samples from a large population are chosen using the theory of probability

Random Sampling : Each member of the population has equal chance being selected in the sample.

Systematic Sampling : Every nth record is choosen to be a part of the sample.

Stratified Sampling : A Stratum is used to form sample. A stratum is a subset of the population that shares at least one common characteristics.

KOSH

Page No.:

Date:

Random sampling is used to select a sufficient no. of subjects from each stratum.

## Types of Statistics

① Descriptive Statistics

→ used to summarise the data to provide description of the population, either through numerical calculation or graph or tables.

→ focuses on main characteristics of the data.

② Inferential Statistics.

→ It makes inferences & predictions about a population based on a sample of data taken from the population in question.

→ It generalizes the large dataset & applies probability to draw conclusion

KOSH

Page No.:
Date:

# Discriptive Statistics

→ It is a method to describe & understand the features of a specific dataset by giving short summaries about the sample & measure the data.

→ Discriptive statistics are broken down into two categories
- Measure of central tendency (rep. summary of Dataset)
- Measure of variability (spread)

Central tendency
- Mean (Avg of all value)
- Mode (Most recurrent)
- Median (Measure of Centre value)

Measure of Spread
- Raye
- IQR
- Variance & Standard Dev (Difference of each elmnt from mean actual value)

Variance is measure of how much a variable differes from
Tell us about the spread of dataset by breaking it in quartile
How spread apart the value in a dataset are.

Raye = Max($x_i$) − Min($x_i$)

Quartile
- $Q_1 \Rightarrow 25\%^{th}$ value
- $Q_2 \Rightarrow 50\%^{th}$ value
- $Q_3 \Rightarrow 75\%^{th}$ value.

IQR = $Q_3 - Q_1$

Standard Deviation = $(x_i - M)$, $M$ = Mean.

Variance. $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

$x$ = idividual data point
$n$ = No of data points
$\bar{x}$ = mean.

$\mu$ = Mean of Population  $N$ = no of DP in Population.

$\bar{x}$ = Mean of sample  $n$ = no of DP in sample.

Page No.:

Date:

KOSH

⇒ Population Variance is avg of Squared Deviation

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

⇒ Sample variance is is the avg of squared differences from the mean.

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

Standard Deviation is measure of despersion of a set of data from its mean.

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

KOSH

Page No.:
Date:

# Information Gain & Entropy.

Entropy : Measure of uncertainity in data.

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

S = Set of all instances in the dataset.
N = Number of distinct class values
$p_i$ = Probability event.

Information Gain : How much information a particular feature / variable gives us about final outcome.

$$\text{Gain}(A,S) = H(S) - \sum_{j=1}^{V} \frac{|S_j|}{|S|} \cdot H(S_j)$$

$$= H(S) - H(A,S)$$

$H(S)$ — Entropy of whole dataset S
$|S_j|$ — No of instances with j value of a attribute A
$|S|$ — Total no of instances in dataset A.
V — Set of distinct values of an attribute A.
$H(S_j)$ - entropy a subset of instance for an attribute A.
$H(A,S)$ - Entropy of an attribute A.

KOSH

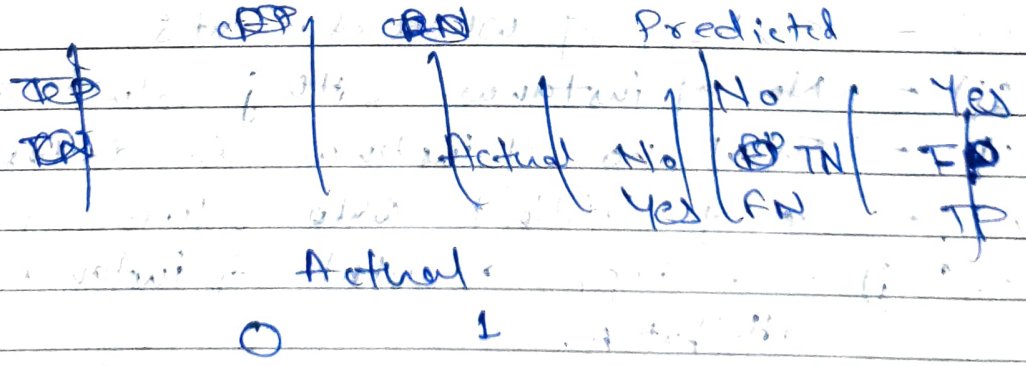
Page No.:
Date:

Use case – Decision Tree.

⇒ Confusion Matrix

- It is matrix used to describe the performance of a classification model or classifier on a set of test data for which the true values are known.
- Represents a tabular representation of Actual vs Predicted values

$$\frac{TP + TN}{TP + TN + FP + FN}$$




| | Actual 0 | Actual 1 |
|---|---|---|
| Predicted 0 | TN | FN |
| Predicted 1 | FP | TP |

KOSH

Page No.:

Date:

# Probability

$$P(x) = \frac{\text{Desired}}{\text{Total}}$$

## Terminologies:

→ Random Experiment: Experiment for which the outcome cannot be predicted with certainity.

→ Sample Space: The entire possible set of outcomes of a random experiment is the sample space (S) of that experiment.

→ Event: Out comes of an experiment

## Event Type:

→ Disjoint: Don't have a common outcome.

→ Non Disjoint: can have common outcome

2:05:45

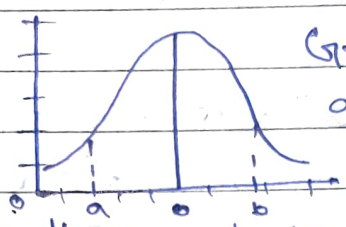
KOSH

Page No.:
Date:

# Probability Distribution

## ① Probability Density function.

→ It is the equation describing a continuous probability distribution




Graph is also know as bell curve.

Properties:

① Graph of PDF is continuous in range
② Area bounded by the curve of density function & the x axis is equal to 1 (area below the curve)
③ Probability that a random variable assume a value b/w a & b is equal to the area under PDF bounded by a & b.

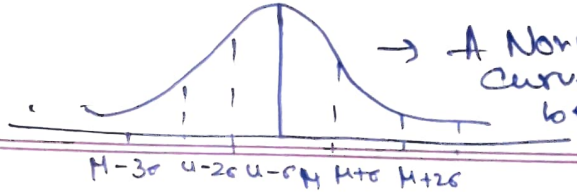
## ② Normal Distribution. (Gaussian Distribution)

→ Normal Distribution is a probability distribution that associates the normal random variable X with cumulative properties

→ It denotes the symmetric properties of the mean. Idea behind is that the data near the mean occurs more frequently than the data away from the mean, or data near the mean represents the entire dataset

→ Graph of Normal Distribution depends on two factors: the Mean & the Standard Deviation.

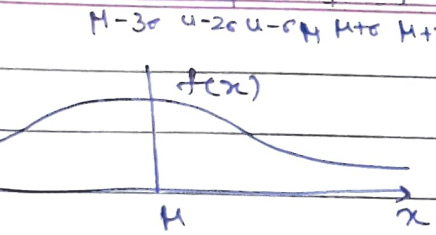
Mean: determines location of center of the graph
Standard deviation: Determines height of the graph.

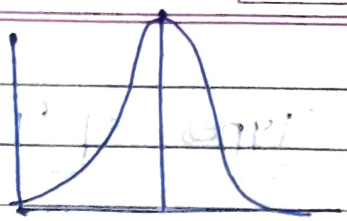


→ A Normal Distribution Curve is also a bell curve.

Page No.:
Date:

KOSH



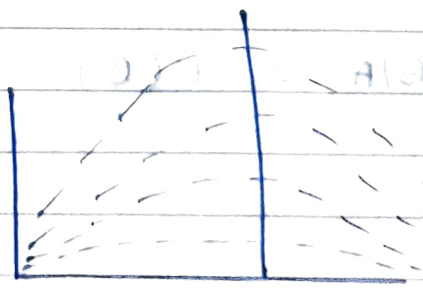
If Standard Deviation is large, curve is short & wide

if Standard Deviation is small, curve is tall & narrow

⇒ Central Limit Theorem (only true for large dataset as small dataset has a lot of variations due to some uncertain factors)

- It states that the sampling distribution of the mean of any independent random variable will be normal or nearly normal, if the sample size is large enough.

- If we have a large population & we divided it into many samples, then the mean of all the samples of the population will be almost equal to the mean of the entire population. Hence, mean is normally distributed. Holds true only for a large data set



Mean of each Sample is almost along the same line.

The accuracy or resemblance depends on two factors:
① No. of sample points considered ② Shape of underlying population ← shape depends on
① Standard Deviation
② Mean of sample.

KOSH

Page No.:
Date:

## ⇒ Types of Probability

① Marginal Probability.
→ Probability of occurence of single event.

Hearts in a deck of card = $\frac{13}{52}$.

② Joint Probability.
→ Measure of two events happening at same time.

Ace of hearts = $\frac{1}{52}$

③ Conditional Probability.
→ Probability of an event based on occurance of a previous event.

An event B will occur given that A has already occured.

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$ (In case of A & B are dependent)

$$P(B/A) = P(B)$$ (In case A & B are Independent)

KOSH

Page No.:
Date:

## ⇒ Baye's Theorem

- Shows the relationship between one conditional probability & its inverse.
- It is the probability of an event occuring based on prior knowledge of conditions that might be related to the same event.

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

P(A|B) is refered to as posterior which means the probability of occurence of A given B.

P(B|A) is refered to as likelihood ratio which measures the probability of occurence of B given A.

P(A) is refered as Prior which represent the actual probability distribution of A.

Eg: we have 3 bowls.

Bowl A = 2 blue 4 red

Bowl B = 8 blue 4 red

Bowl C = 1 blue 3 red.

1 ball is drawn from each. Probability of drawing a blue from A given 2 blue is drawn from Bowl. are drawn in total.

Sol

Event 1 = Picking a blue ball from bowl A. ⇒ A

Event 2 = Picking 2 blue balls. ⇒ X.

$$P(A/X) = \frac{P(A \cap X)}{P(X)} \qquad P(A \cap X) = \frac{{}^{6}C_2 + {}^{12}C_2}{{}^{18}C_2}$$

$$P(X) = \frac{({}^{6}C_1 \times {}^{12}C_1) + ({}^{12}C_1 \times {}^{4}C_1) + ({}^{4}C_1 \times {}^{6}C_1)}{{}^{22}C_2}$$

Page No.:
Date:

## ⇒ Inferential Statistics & Point Estimation

Point Estimation is concerned with the use of the sample data to measure the single value that that serves as an approximate value or the best estimate of an unknown population parameter.

Population —Random→ Sample.

Sample —Mean→ $\bar{x}$

$\bar{x}$ —used to→ $\mu$

$\mu$ : Median / Mode / Mean etc. of Population

Estimator = function of the sample used to find out the estimate.
Eg: Sample mean.

Estimate = calculated value from the function
Eg: population mean.

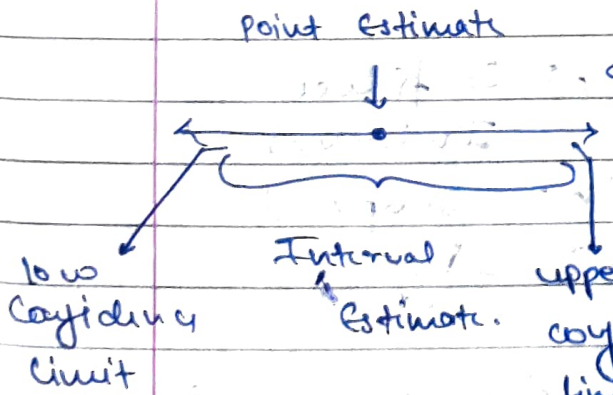
KOSH

Page No.:
Date:

# Method to find Estimate.

1. Method of Moments: form an equation in the sample data set & analyze that similar Equation in the population dataset as well.
   - we take the known facts about the Population sample & extend them to sample & using sample then predict the known parameters such as mean, mode etc.
2. Maximum of Likelihood:
3. Baye's Estimator
4. Best Unbiased Estimators.
5. Interval Estimation Method (Imp)

Interval Estimate: An interval, a range of values used to estimate a population parameter.




: If we want to find a value (Mean mode etc) of a population, we set a range & at that rate the value lies.

So, instead of predicting a point we predict a range. in which a value may occur.

Kosh

Page No.:
Date:

→ Difference b/w point estimate & actual population parameter value is Sampling Error

→ When $\mu$ is estimated, sampling error $= \mu - \bar{x}$

## Confidence Interval

→ It is the measure of your confidence that the interval estimated contains the population parameter

→ It is used to describe the amount of uncertainity associated with a sample estimate of a population parameter.

→ A range of values so constructed that there is specified probability of including the true value of a parameter within it.

## Margin of error - E

→ For a given level of confidence it is the greatest possible distance b/w the point estimate & the value of parameter it is estimating.

$$\left[ E = Z_c \frac{\sigma}{\sqrt{n}} \right]$$

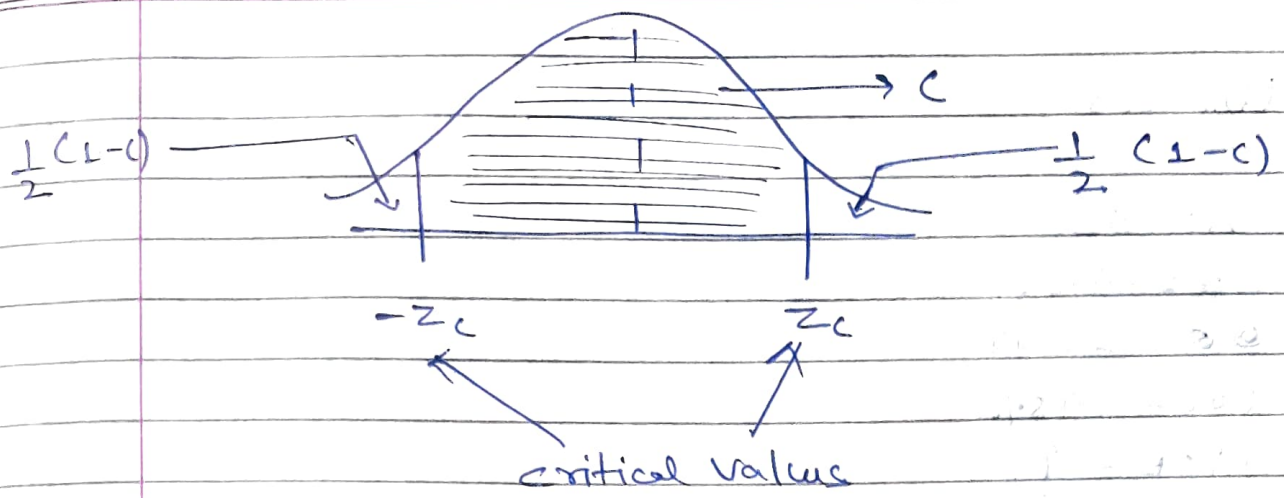
$Z_c$ = Confidence Interval

$n$ = Sample size

$\sigma$ = S.D.

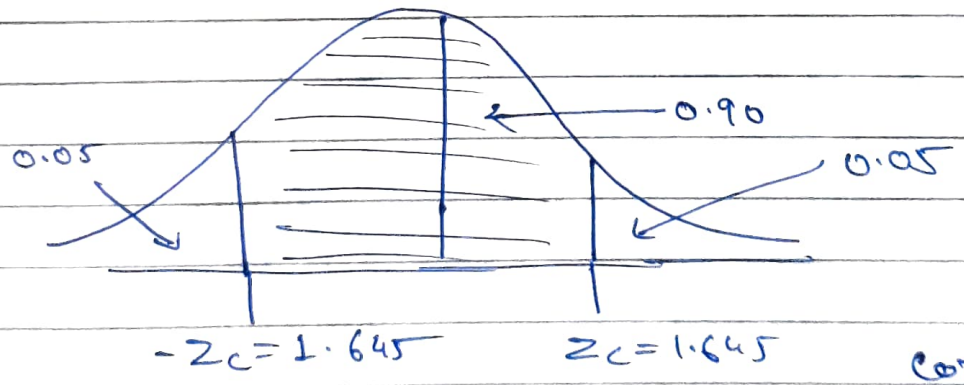
The level of confidence c is the probability that the interval estimate contains the population parameter.

Page No.:
Date:




c is the area beneath the normal curve b/w the critical values. Z-score can be calculated using standard normal table.

↳ level of confidence = 90% i.e. 90% confidence that interval contains population parameter.




Correspondiy
Z score = $\pm 1.645$

Sequence :

1. Sample statistic (used to calculate population parameter)
2. Select confidence level.
3. Margin of error
4. Specify confidence interval

Use Case

$n = 32$

$\bar{x} = 74.22$

$\sigma = 23.44$

(z) LoC = 95%.

(E) MoE = ?

$$\Rightarrow \quad E = z_c \frac{\sigma}{\sqrt{n}}$$

$$E = 1.96 \frac{23.44}{\sqrt{32}} \approx 8.12$$

KOSH

Page No.:
Date:

# Hypothesis Testing

→ Used to check whether hypothesis is accepted or rejected.

## Steps :

① State the hypothesis : Stating null & alternative hypothesis.

② Formulate an analysis plan : This stage involve construction of analysis plan.

③ Analyse Sample Data : Calculation & interpretation of the test statistic as described in the analysis plan.

④ Interpret Result : The application of the decision rule described in the analysis plan.

NULL HYPOTHESIS : Result is no different from assumption (H0)

ALTERNATIVE HYPOTHESIS : Result disapproves the assumption.

Eg: 4 boys Nick, John, Bob, Harry are picked randomly each day. What is the probability that John is not cheating.

Sol P(John picked for a day) = 3/4 = 0.75

P(John picked for 3 days) = 3/4 × 3/4 × 3/4 = 0.42.

P(John —"— 12 days) = $(3/4)^{12}$ = 0.032.

As, P(Event) < 5% (Threshold); John is cheating.