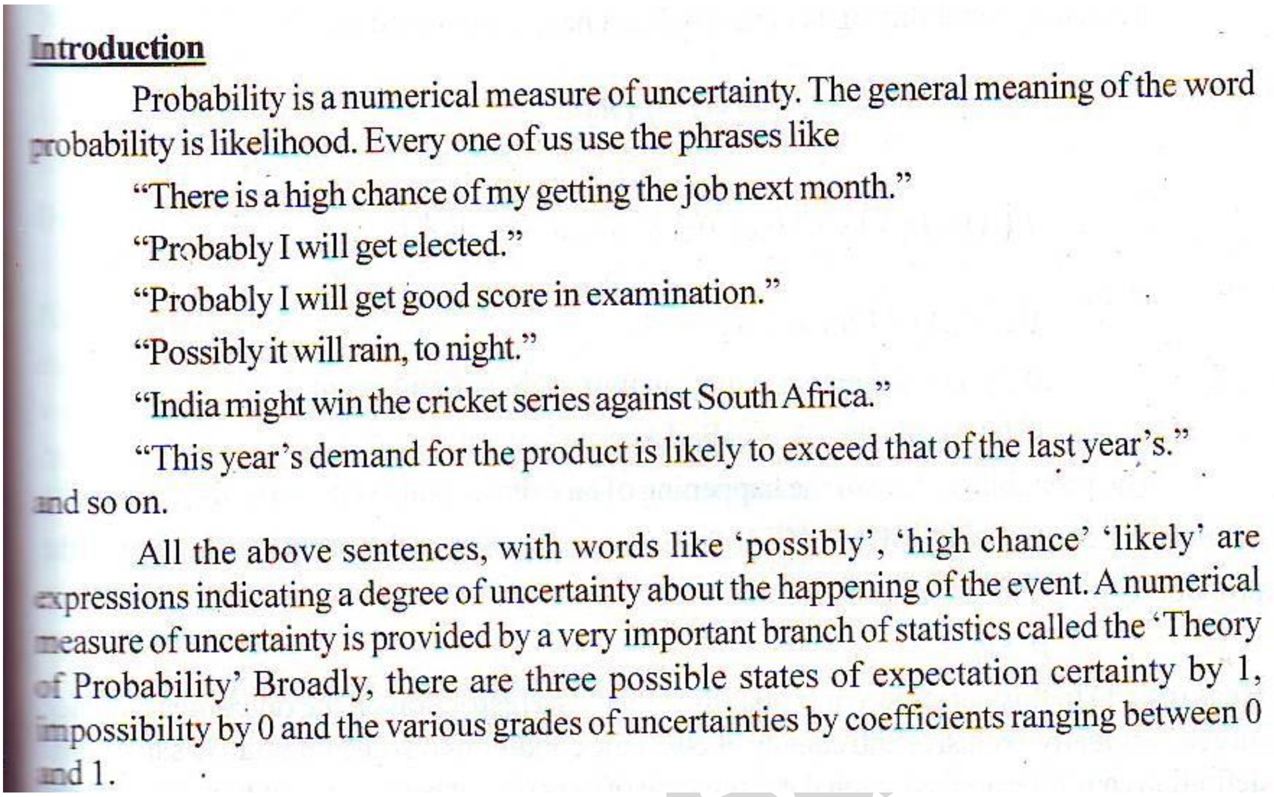
SYLLABUS

Class – B.COM/BBA IV year

Subject –Advanced Statistical Analysis

UNIT – I	Theory of Probability and Probability Distributions: Approaches to calculation of probability. Marginal, joint and conditional probabilities; Probability rules; Bayes theorem; Expected value and standard deviation of a probability distribution; Standard probability distributions Binomial, Poisson, and Normal.
UNIT – II	Statistical Decision Theory: Decision-making process. Payoff and Regret tables. Decision rules under risk and uncertainty; Expected value approach and EVPI; Marginal analysis; Decision-tree analysis.
UNIT – III	Sampling Distributions and Estimation: Sampling concepts; Types of sampling techniques; Sampling distribution of means and proportions; Central Limit Theorem. Point and interval estimation; Properties of a good estimator; Confidence intervals for means; Confidence intervals for proportions; Sample size determination.
UNIT IV	Hypothesis Testing: Steps of hypothesis testing. One and two-tailed tests. Type 1 and type IIErrors; Power of a test; Calculation and use of p-value. One Sample Tests: Means and proportions. Two-sample Tests: Tests for difference between means - Independent samples, Small samples; Dependent samples; Testing of difference between proportions.
UNIT – V	Analysis of Variance and Non-Parametric Tests: F-test of equality of variances; One-factor ANOVA; Chi-square test for Independence and for Goodness-of-fit. Sign test, One-sample runs test.

`UNIT-I



Introduction

Probability is a numerical measure of uncertainty. The general meaning of the word probability is likelihood. Every one of us use the phrases like

"There is a high chance of my getting the job next month."

"Probably I will get elected."

"Probably I will get good score in examination."

"Possibly it will rain, to night."

"India might win the cricket series against South Africa."

"This year's demand for the product is likely to exceed that of the last year's."

and so on.

All the above sentences, with words like 'possibly', 'high chance' 'likely' are expressions indicating a degree of uncertainty about the happening of the event. A numerical measure of uncertainty is provided by a very important branch of statistics called the 'Theory of Probability' Broadly, there are three possible states of expectation certainty by 1, impossibility by 0 and the various grades of uncertainties by coefficients ranging between 0 and 1.

Terms used in Theory of Probability

Random Experiment : Suppose we toss a coin or throw a die or draw a card from a pack of playing cards. In these examples there are a number of possible outcomes which can occur but there is an uncertainty as to which one of them will actually occur. Such experiments are called random experiment. A random experiment may be defined as an experiment which when repeated under essentially identical conditions does not give unique results but may result in any one of the several possible outcomes. These outcomes are known as events.

Properties of Random Experiment

i) It can be repeated.

ii) The number of possible outcomes is more than one.

iii) In a single trial out of the several outcomes one and only one outcomes can happen i.e. one outcomes is certain to happen.

iv) The outcomes of an individual trial connot be predicted.

Trial : By Trial we mean performing of an experiment

Outcome : The result of a trial in a random experiment is called an outcome.

Sample Space : Each performance in a random experiment is called a trial. The result of a trial in a random experiment is called an outcome, an elementary event, or a sample point. The totality of all possible outcome (i.e., sample points) of a random experiment constitutes the sample space. Suppose $e_1, e_2, e_3, \ldots\ldots\ldots e_n$ are the possible outcomes of a random experiment E. We associate with each experiment E, a set $S = \{e_1, e_2, \ldots.. e_n\}$ of possible outcomes with the following properties :

i) Each element of S denotes a possible outcomes of the experiment.

ii) Any trial results in an outcomes that corresponds to one and only one element of the set S.

Thus a set of all possible outcomes in a random experiment is called sample space. It is denoted by S. For example,

i) If a dice is thrown, the sample space will be : $S = \{1, 2, 3, 4, 5, 6\}$.

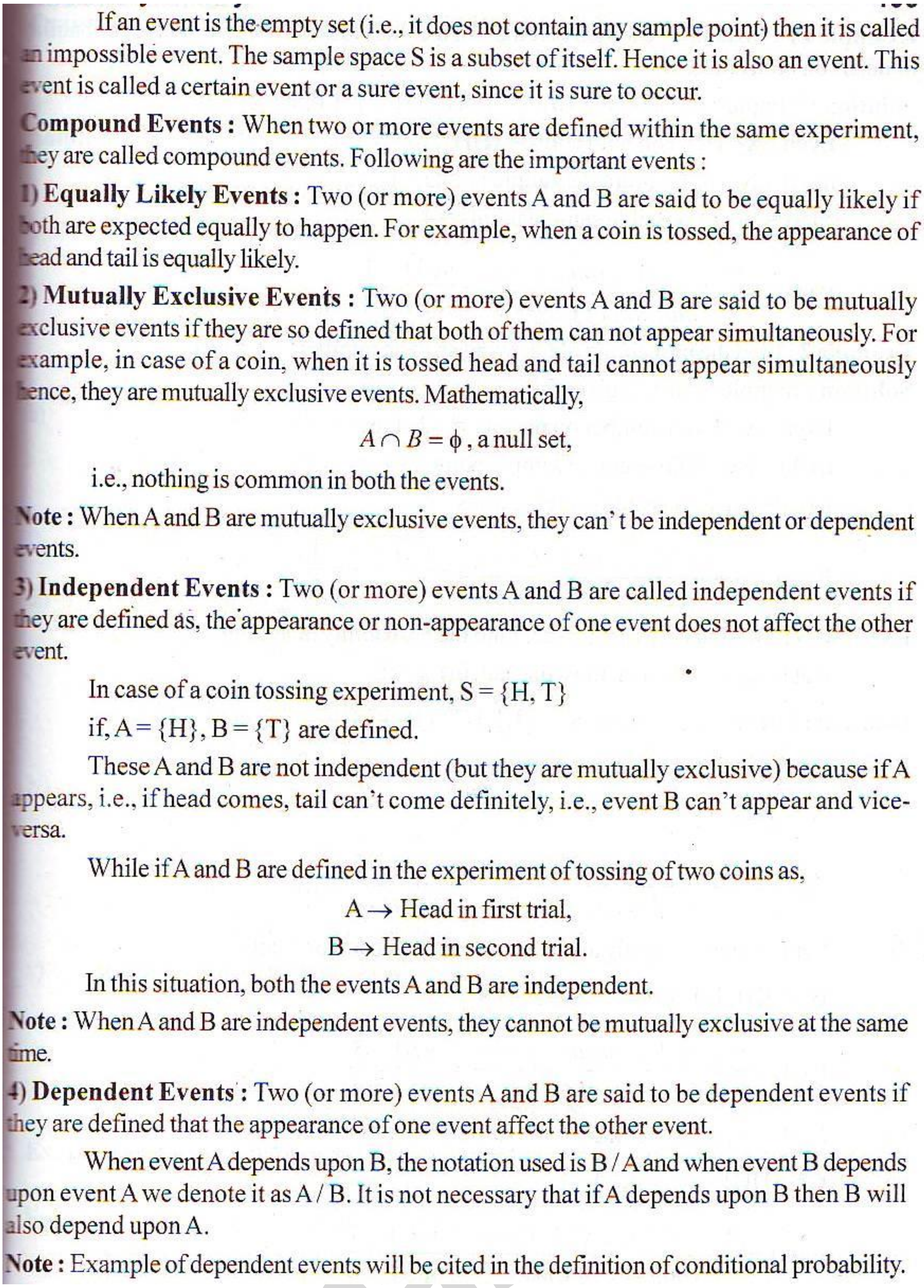
ii) When a single coin is tossed : $S = \{H, T\}$.

iii) When two coins are tossed simultaneously or one coin is tossed repeatedly twice, following will be the sample space : $S = \{HH, HT, TH, TT\}$.

iv) Similarly, in case when either three coins are tossed or one coin is tossed repeatedly thrice, the sample space will be :

$$S = \{HHH, HHT, HTH, THH, TTT, TTH, THT, HTT\}.$$

Event : Any subset of a sample space is called an event. If an event contains only one sample point, then it is called a simple event or an elementary event.



If an event is the empty set (i.e., it does not contain any sample point) then it is called an impossible event. The sample space S is a subset of itself. Hence it is also an event. This event is called a certain event or a sure event, since it is sure to occur.

Compound Events : When two or more events are defined within the same experiment, they are called compound events. Following are the important events :

1) **Equally Likely Events :** Two (or more) events A and B are said to be equally likely if both are expected equally to happen. For example, when a coin is tossed, the appearance of head and tail is equally likely.

2) **Mutually Exclusive Events :** Two (or more) events A and B are said to be mutually exclusive events if they are so defined that both of them can not appear simultaneously. For example, in case of a coin, when it is tossed head and tail cannot appear simultaneously hence, they are mutually exclusive events. Mathematically,

$$A \cap B = \phi \text{, a null set,}$$

i.e., nothing is common in both the events.

Note : When A and B are mutually exclusive events, they can' t be independent or dependent events.

3) **Independent Events :** Two (or more) events A and B are called independent events if they are defined as, the appearance or non-appearance of one event does not affect the other event.

In case of a coin tossing experiment, S = {H, T}

if, A = {H}, B = {T} are defined.

These A and B are not independent (but they are mutually exclusive) because if A appears, i.e., if head comes, tail can't come definitely, i.e., event B can't appear and vice-versa.

While if A and B are defined in the experiment of tossing of two coins as,

A → Head in first trial,

B → Head in second trial.

In this situation, both the events A and B are independent.

Note : When A and B are independent events, they cannot be mutually exclusive at the same time.

4) **Dependent Events :** Two (or more) events A and B are said to be dependent events if they are defined that the appearance of one event affect the other event.

When event A depends upon B, the notation used is B / A and when event B depends upon event A we denote it as A / B. It is not necessary that if A depends upon B then B will also depend upon A.

Note : Example of dependent events will be cited in the definition of conditional probability.

Conditional Probability

The simultaneous occurrence of two or more events is called a **compound event**. If A and B are two events, then AB denotes the simultaneous occurrence of A and B and P(AB) denotes the probability of the simultaneous occurrence of two events A and B.

The probability of the happening of an event A when the event B has already happened is called the conditional probability and is denoted by P(A/B). Similarly P(B/A) means the probability of the happening of an event B when the event A has already happened.

Two event are said to be independent if the probability of the happening of one does not depend on the happening of the other.

Multiplication theorem of probability

(Theorem of compound probability)

According to this theorem, "If two events are mutually independent, and the probability of the one is p_1 while that of the other is p_2, the probability of the two events occurring simultaneously is the product of p_1 and p_2."

Multiplication Rules

i) **When Events are Independent.** The probability that both independent events, A and B will occur is :

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

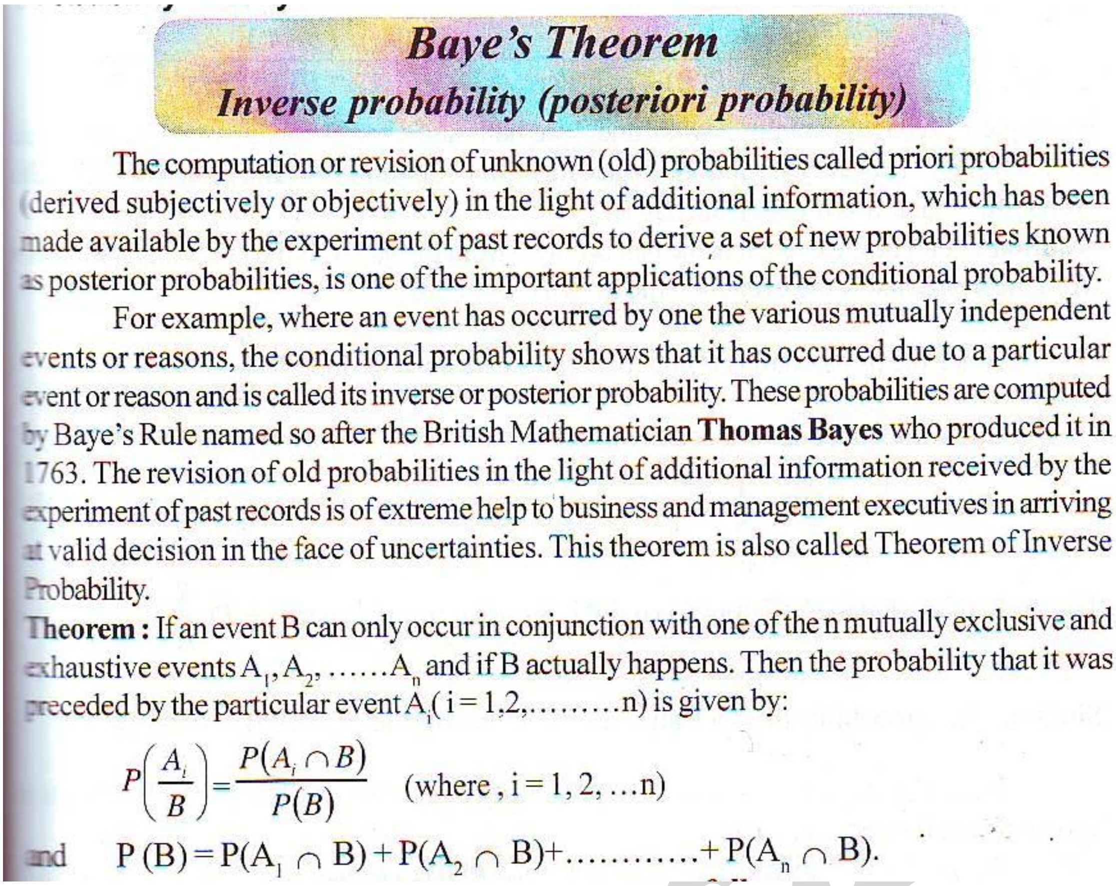
In other words, $P(A \text{ and } B) = P(A) \cdot P(B)$.

ii) **When Events are Dependent.** If events A and B are so related that the occurrence of B is affected by the occurrence of A, then A and B are called dependent events. The probability of event B depending on the occurrence of event A is called conditional probability and is written as P(B/A).

The probability that both the dependent events A and B will occur, is given by:

$$P(A \cap B) = P(A) \cdot P\left(\frac{B}{A}\right)$$

In other words $$P(A \text{ and } B) = P(A) \cdot P\left(\frac{B}{A}\right)$$



Baye's Theorem

Inverse probability (posteriori probability)

The computation or revision of unknown (old) probabilities called priori probabilities (derived subjectively or objectively) in the light of additional information, which has been made available by the experiment of past records to derive a set of new probabilities known as posterior probabilities, is one of the important applications of the conditional probability.

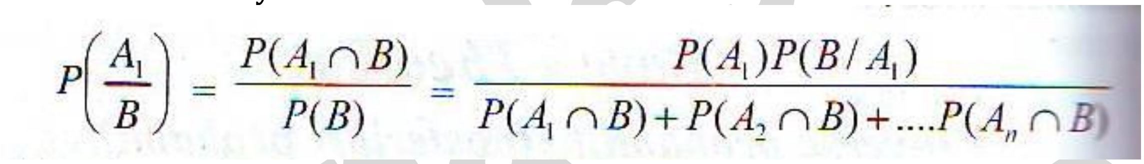
For example, where an event has occurred by one the various mutually independent events or reasons, the conditional probability shows that it has occurred due to a particular event or reason and is called its inverse or posterior probability. These probabilities are computed by Baye's Rule named so after the British Mathematician **Thomas Bayes** who produced it in 1763. The revision of old probabilities in the light of additional information received by the experiment of past records is of extreme help to business and management executives in arriving at valid decision in the face of uncertainties. This theorem is also called Theorem of Inverse Probability.

Theorem : If an event B can only occur in conjunction with one of the n mutually exclusive and exhaustive events $A_1, A_2, \ldots\ldots A_n$ and if B actually happens. Then the probability that it was preceded by the particular event A_i(i = 1,2,.........n) is given by:

$$P\left(\frac{A_i}{B}\right)=\frac{P(A_i \cap B)}{P(B)} \quad \text{(where , i = 1, 2, ...n)}$$

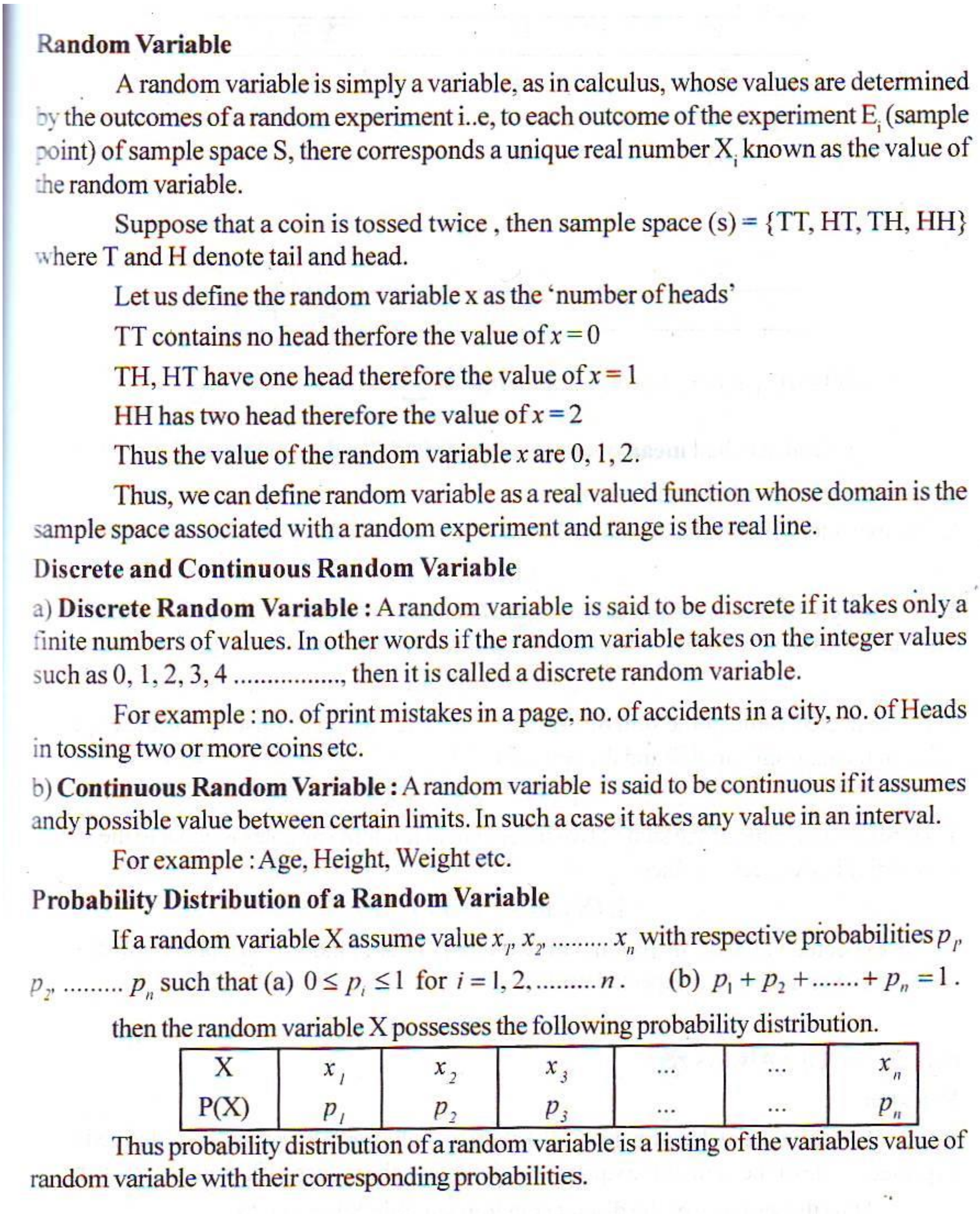
and $P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \ldots\ldots\ldots\ldots + P(A_n \cap B)$.

The Formula for bayes is:



$$P\left(\frac{A_1}{B}\right)=\frac{P(A_1 \cap B)}{P(B)}=\frac{P(A_1)P(B/A_1)}{P(A_1 \cap B)+P(A_2 \cap B)+\ldots.P(A_n \cap B)}$$

.



Random Variable

A random variable is simply a variable, as in calculus, whose values are determined by the outcomes of a random experiment i..e, to each outcome of the experiment E_i (sample point) of sample space S, there corresponds a unique real number X_i known as the value of the random variable.

Suppose that a coin is tossed twice , then sample space (s) = {TT, HT, TH, HH} where T and H denote tail and head.

Let us define the random variable x as the 'number of heads'

TT contains no head therfore the value of $x = 0$

TH, HT have one head therefore the value of $x = 1$

HH has two head therefore the value of $x = 2$

Thus the value of the random variable x are 0, 1, 2.

Thus, we can define random variable as a real valued function whose domain is the sample space associated with a random experiment and range is the real line.

Discrete and Continuous Random Variable

a) **Discrete Random Variable :** A random variable is said to be discrete if it takes only a finite numbers of values. In other words if the random variable takes on the integer values such as 0, 1, 2, 3, 4 ...., then it is called a discrete random variable.

For example : no. of print mistakes in a page, no. of accidents in a city, no. of Heads in tossing two or more coins etc.

b) **Continuous Random Variable :** A random variable is said to be continuous if it assumes andy possible value between certain limits. In such a case it takes any value in an interval.

For example : Age, Height, Weight etc.

Probability Distribution of a Random Variable

If a random variable X assume value $x_1, x_2, \ldots\ldots\ldots x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots p_n$ such that (a) $0 \le p_i \le 1$ for $i = 1, 2, \ldots\ldots\ldots n$. (b) $p_1 + p_2 + \ldots\ldots. + p_n = 1$.

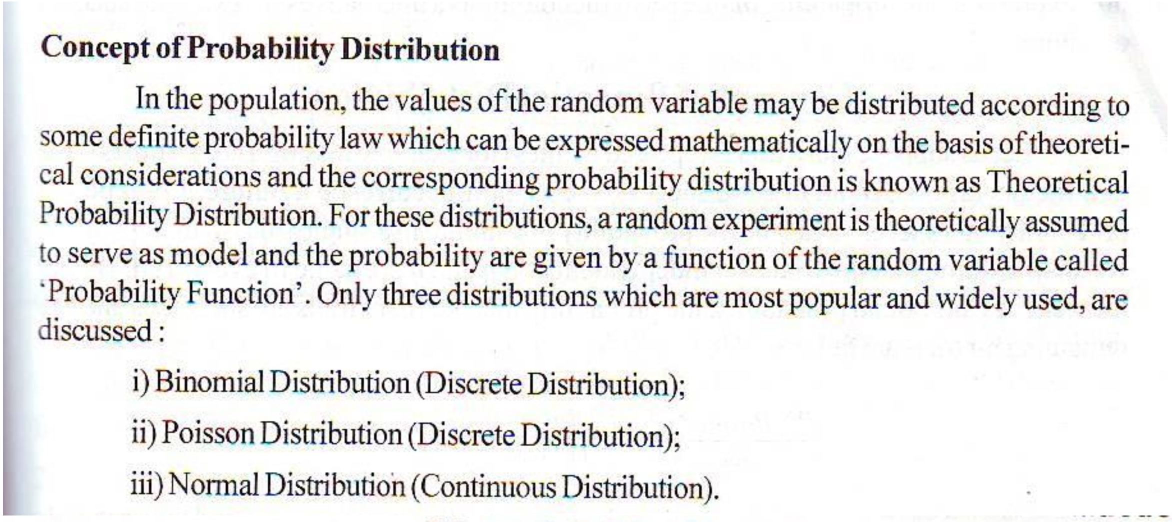
then the random variable X possesses the following probability distribution.

X	x_1	x_2	x_3	...	...	x_n
P(X)	p_1	p_2	p_3	...	...	p_n

Thus probability distribution of a random variable is a listing of the variables value of random variable with their corresponding probabilities.

Expectation : E(x) =$\sum_{i=1}^{n} p_i x_i$

Variance : var(x) = E(x^2) – (E(x))^2



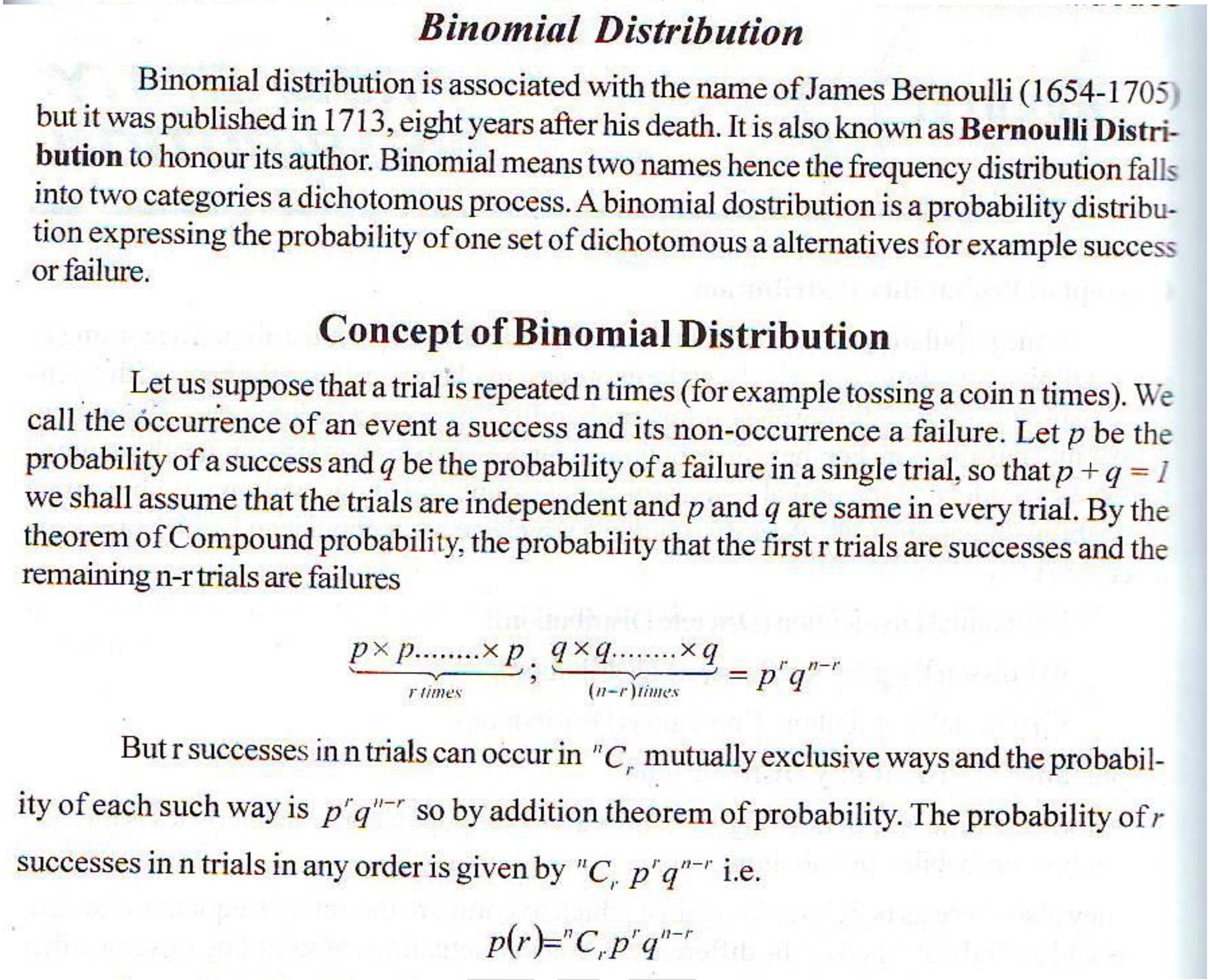
Concept of Probability Distribution

In the population, the values of the random variable may be distributed according to some definite probability law which can be expressed mathematically on the basis of theoretical considerations and the corresponding probability distribution is known as Theoretical Probability Distribution. For these distributions, a random experiment is theoretically assumed to serve as model and the probability are given by a function of the random variable called 'Probability Function'. Only three distributions which are most popular and widely used, are discussed :

i) Binomial Distribution (Discrete Distribution);

ii) Poisson Distribution (Discrete Distribution);

iii) Normal Distribution (Continuous Distribution).



Binomial Distribution

Binomial distribution is associated with the name of James Bernoulli (1654-1705) but it was published in 1713, eight years after his death. It is also known as **Bernoulli Distribution** to honour its author. Binomial means two names hence the frequency distribution falls into two categories a dichotomous process. A binomial dostribution is a probability distribution expressing the probability of one set of dichotomous a alternatives for example success or failure.

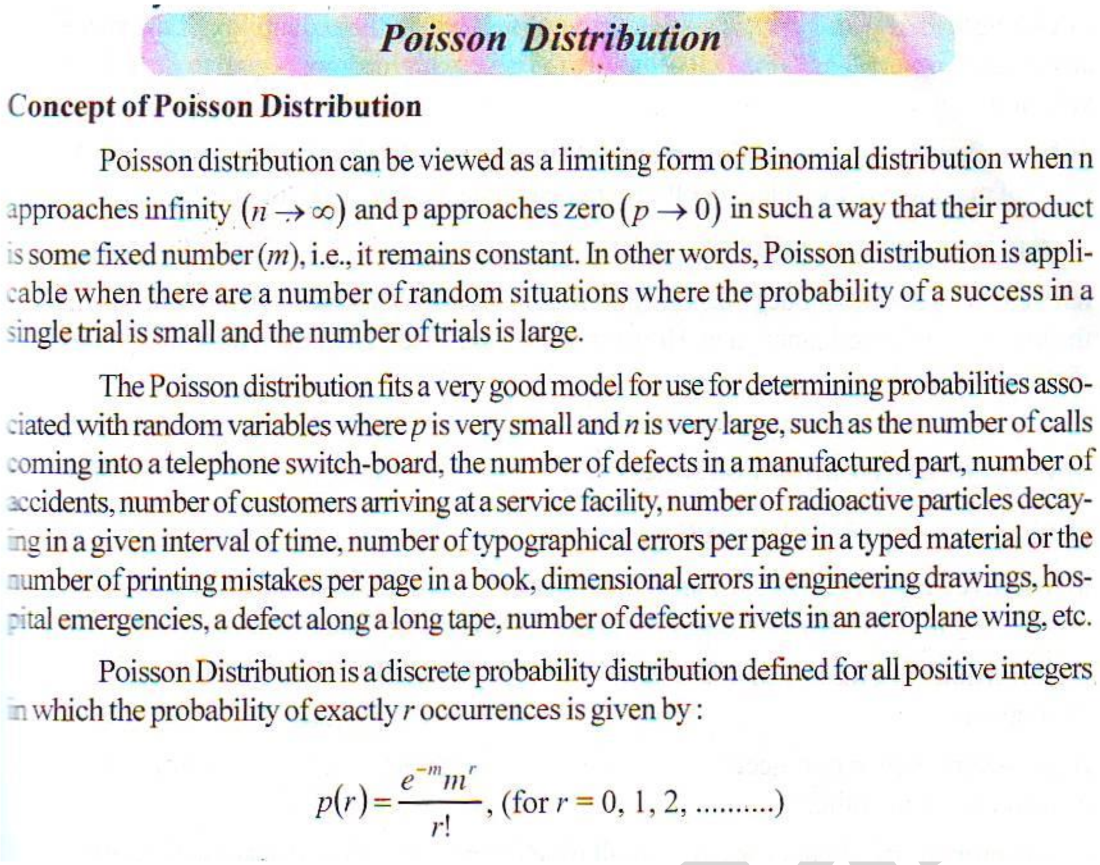
Concept of Binomial Distribution

Let us suppose that a trial is repeated n times (for example tossing a coin n times). We call the occurrence of an event a success and its non-occurrence a failure. Let p be the probability of a success and q be the probability of a failure in a single trial, so that $p + q = 1$ we shall assume that the trials are independent and p and q are same in every trial. By the theorem of Compound probability, the probability that the first r trials are successes and the remaining n-r trials are failures

$$\underbrace{p\times p........\times p}_{r\ times};\ \underbrace{q\times q.......\times q}_{(n-r)\ times} = p^r q^{n-r}$$

But r successes in n trials can occur in ${}^nC_r$ mutually exclusive ways and the probability of each such way is $p^r q^{n-r}$ so by addition theorem of probability. The probability of r successes in n trials in any order is given by ${}^nC_r\ p^r q^{n-r}$ i.e.

$$p(r)={}^nC_r p^r q^{n-r}$$



Poisson Distribution

Concept of Poisson Distribution

Poisson distribution can be viewed as a limiting form of Binomial distribution when n approaches infinity $(n \to \infty)$ and p approaches zero $(p \to 0)$ in such a way that their product is some fixed number (m), i.e., it remains constant. In other words, Poisson distribution is applicable when there are a number of random situations where the probability of a success in a single trial is small and the number of trials is large.

The Poisson distribution fits a very good model for use for determining probabilities associated with random variables where *p* is very small and *n* is very large, such as the number of calls coming into a telephone switch-board, the number of defects in a manufactured part, number of accidents, number of customers arriving at a service facility, number of radioactive particles decaying in a given interval of time, number of typographical errors per page in a typed material or the number of printing mistakes per page in a book, dimensional errors in engineering drawings, hospital emergencies, a defect along a long tape, number of defective rivets in an aeroplane wing, etc.

Poisson Distribution is a discrete probability distribution defined for all positive integers in which the probability of exactly *r* occurrences is given by :

$$p(r) = \frac{e^{-m} m^r}{r!}, \text{ (for } r = 0, 1, 2, \text{ ....)}$$

Normal Distribution

Concept of Normal Distribution

Normal or Guassian distribution is the most important continuous probability distribution in Statistics and is defined by the probability density function (or simply density function):

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} ;$$

The normal distribution is the most versutile of all Theoretical distributions. It is found to be useful in statistical inferences, in characterising uncertainties in many real life processes, and in approximating other probability distributions. Quite often we face the problem of making inferences about processes based on limited data. Limited data is basically a sample from the full body of data is distributed, it has been found that the Normal Distribution can be used to characterise the sampling distribution. This helps considerably in Statistical Inferences. Heights, weight and dimensions of a product are some of the continuous random variables which are found to the normally distributed. This Knowledge helps us in calculating the probabilities of different events in varied situations, which in turn is useful for decision, making.

The normal distribution was discovered first by **De Moivre** in 1733. **Laplace** also knew about it almost at the same time. It is also associated with the name of **Gauss** and is known as **Gaussian Distribution** .

The graphical shape of normal distribution called the normal curve, is the bell shaped smooth symmetrical curve as shown in T curve is asymptotic in both directions to the x-axis and depends on the two parameters mean and standard deviation of the normal curve distribution.

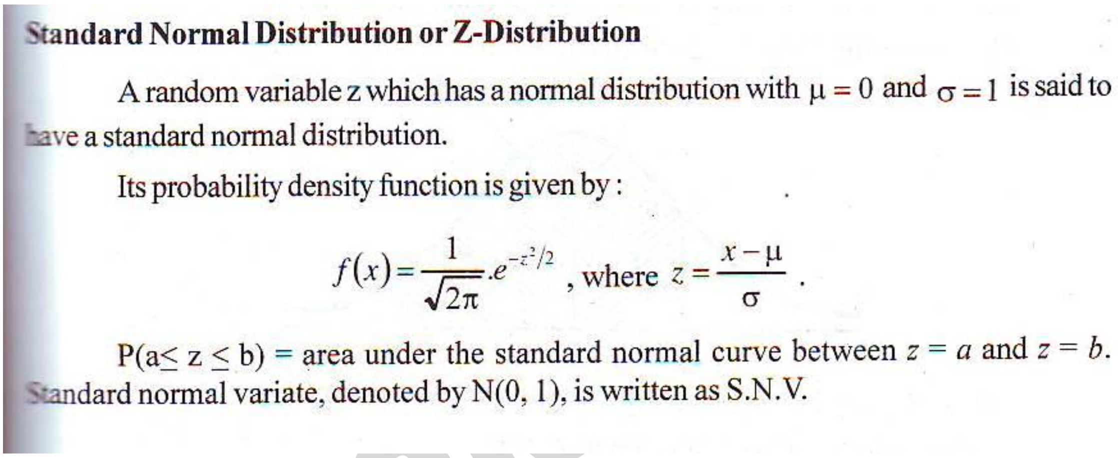
The equation of a normal curve is of exponential type. If X is a normal random variable with mean μ and standard deviation σ, then the equation of the normal curve is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where x can take any value in the range $(-\infty, +\infty)$ and π and e are two mathematical constants having approximate values, $\frac{22}{7}$ and 2.718 respectively.

μ and σ, the mean and standard deviation are known as parameter. The normal distribution with a mean μ and variance σ^2 may be denoted by the symbol N(μ, σ^2).

$p(x)$ is called the probability density function or simply density function.



Standard Normal Distribution or Z-Distribution

A random variable z which has a normal distribution with $\mu = 0$ and $\sigma = 1$ is said to have a standard normal distribution.

Its probability density function is given by :

$$f(x) = \frac{1}{\sqrt{2\pi}}.e^{-z^2/2} \text{, where } z = \frac{x-\mu}{\sigma} .$$

P(a$\leq$ z $\leq$ b) = area under the standard normal curve between $z = a$ and $z = b$. Standard normal variate, denoted by N(0, 1), is written as S.N.V.

Unit II

decision making process in statistical decision theory

The **decision-making process** means the steps people follow to make a choice or solve a problem. It involves thinking about options, weighing pros and cons, and picking the best solution.In **Statistical Decision Theory**, the decision-making process involves systematic steps to analyze and choose the best course of action in uncertain situations. The process typically includes the following steps:

1. Define the Problem

- Clearly identify the decision to be made.
- Understand the alternatives or choices available.

2. Identify Objectives

- Determine the goals or outcomes that the decision aims to achieve.
- For example, minimizing costs or maximizing profits.

3. Identify Alternatives

- List all possible actions or decisions.
- For instance, choosing between two investment options.

4. Gather Relevant Information

- Collect data related to the problem.
- Include probabilities, costs, benefits, and any prior knowledge.

5. Model the Problem

- Use a decision-making framework like **decision trees** or **Bayesian analysis**.
- Define states of nature (uncertain future events).
- Identify the outcomes associated with each decision.

6. Assign Probabilities

- Estimate the likelihood of each state of nature occurring.
- Use historical data or subjective judgments.

7. Evaluate Payoffs

- Assess the benefits or costs associated with each decision and state of nature.
- Represent these payoffs in a matrix or table.

8. Choose a Decision Criterion

- Select a rule for making the decision, such as:
  - **Maximin**: Choose the option with the best worst-case outcome.
  - **Maximax**: Choose the option with the best possible outcome.
  - **Expected Value (EV)**: Choose the option with the highest average payoff based on probabilities.
  - **Minimax Regret**: Minimize the maximum regret across decisions.

9. Make the Decision

- Based on the chosen criterion, select the best alternative.

10. Implement the Decision

- Put the chosen course of action into practice.

11. Monitor and Review

- Evaluate the decision's effectiveness.
- Update the model if more data or insights become available.

By following these steps, statistical decision theory helps to make informed decisions under uncertainty, balancing risks and rewards systematically.

Payoff table

A **payoff table** in statistics is a table that shows the possible outcomes (or payoffs) for different decisions under various conditions. It helps in analyzing and comparing options to make better decisions, especially in situations involving uncertainty or risk.

Key Points:

1. **Rows:** Represent different decisions or strategies.
2. **Columns:** Represent possible future states or events (like success, failure, or market conditions).
3. **Cells:** Contain the payoffs (profits, costs, or losses) for each decision under each condition.

Payoff table helps in choosing the decision that gives the best outcome based on probabilities or preferences.

Regret Tables

regret tables (also called **opportunity loss tables**) are used in decision-making under uncertainty. A regret table shows how much we "regret" not choosing the best possible decision in each scenario.

Key Concepts:

- **Regret**: The difference between the payoff of the best decision and the payoff of the chosen decision for a particular scenario.
- It helps evaluate the consequences of choosing suboptimal decisions.

The regret table helps in choosing decisions that minimize the worst-case regret.

Decision rules under risk and uncertainty

Decision rules under risk and uncertainty are methods used to make choices when outcomes are not fully predictable.

1. **Under Risk**: Probabilities of outcomes are known.
   - **Expected Value (EV)**: Choose the decision with the highest average payoff, considering probabilities.
   - **Expected Utility**: Accounts for preferences and risk tolerance, choosing based on maximum utility.
   - **Risk-Adjusted Return**: Considers the trade-off between risk and reward.
2. **Under Uncertainty**: Probabilities of outcomes are unknown.
   - **Maximin Rule**: Choose the option with the best worst-case payoff (pessimistic).
   - **Maximax Rule**: Choose the option with the best possible payoff (optimistic).
   - **Minimax Regret Rule**: Minimize the maximum regret of not choosing the best option.
   - **Laplace Criterion**: Treat all outcomes as equally likely and choose based on average payoff.

These rules guide rational decision-making when dealing with unpredictable or partially predictable situations.

Expected Value (EV) Approach

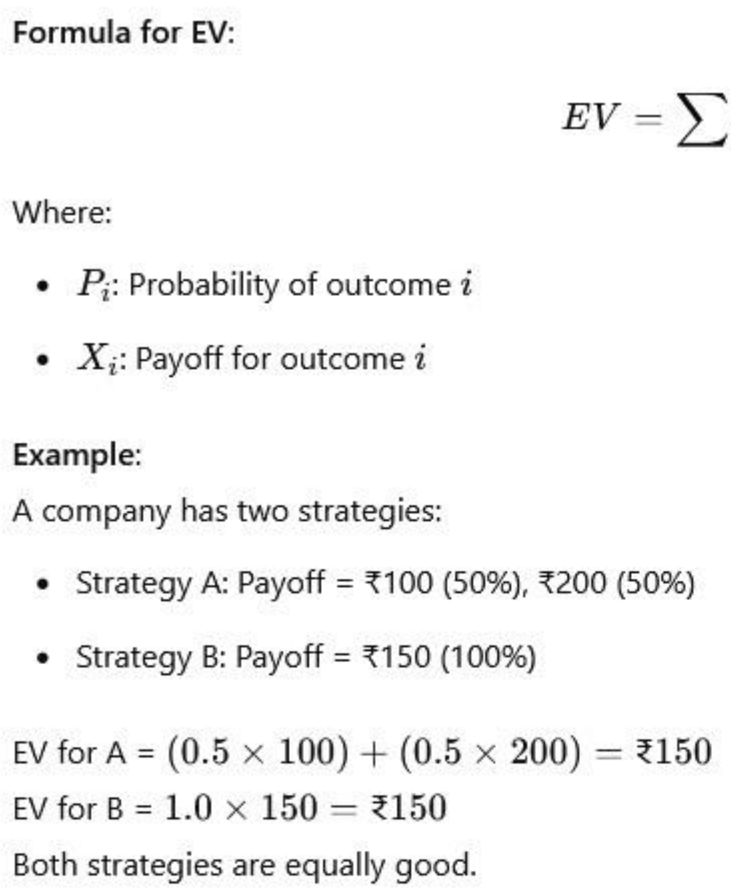
The **Expected Value (EV) approach** is a decision-making method under risk. It calculates the average payoff for each decision, considering the probabilities of different outcomes. The decision with the highest EV is preferred.

The **Expected Value (EV) approach** is a decision-making method used when outcomes are uncertain but probabilities are known. It calculates the average expected payoff of different choices by considering both the probability and the payoff of each possible outcome.

To apply the EV approach:

1. **List all possible outcomes**: Identify the different outcomes for each decision.
2. **Assign probabilities**: Determine the probability of each outcome occurring.

3. **Calculate payoffs**: Determine the payoff or value for each outcome.
4. **Multiply probabilities by payoffs**: For each outcome, multiply the payoff by its probability.
5. **Sum the results**: Add the products of all outcomes to get the expected value.



Formula for EV:

$$EV = \sum(P_i \times X_i)$$

Where:

- P_i: Probability of outcome i
- X_i: Payoff for outcome i

Example:

A company has two strategies:

- Strategy A: Payoff = ₹100 (50%), ₹200 (50%)
- Strategy B: Payoff = ₹150 (100%)

EV for A = $(0.5 \times 100) + (0.5 \times 200) = ₹150$

EV for B = $1.0 \times 150 = ₹150$

Both strategies are equally good.

Expected Value of Perfect Information (EVPI)

EVPI measures the value of having perfect information about future outcomes before making a decision. It quantifies how much better decisions could be with certainty.

EVPI (Expected Value of Perfect Information) is a concept in decision analysis that quantifies the value of having perfect information before making a decision. It represents the maximum amount a decision-maker would be willing to pay to eliminate uncertainty in a situation. EVPI helps in determining whether acquiring additional information is worth the cost.

Calculation:

EVPI is calculated as the difference between:

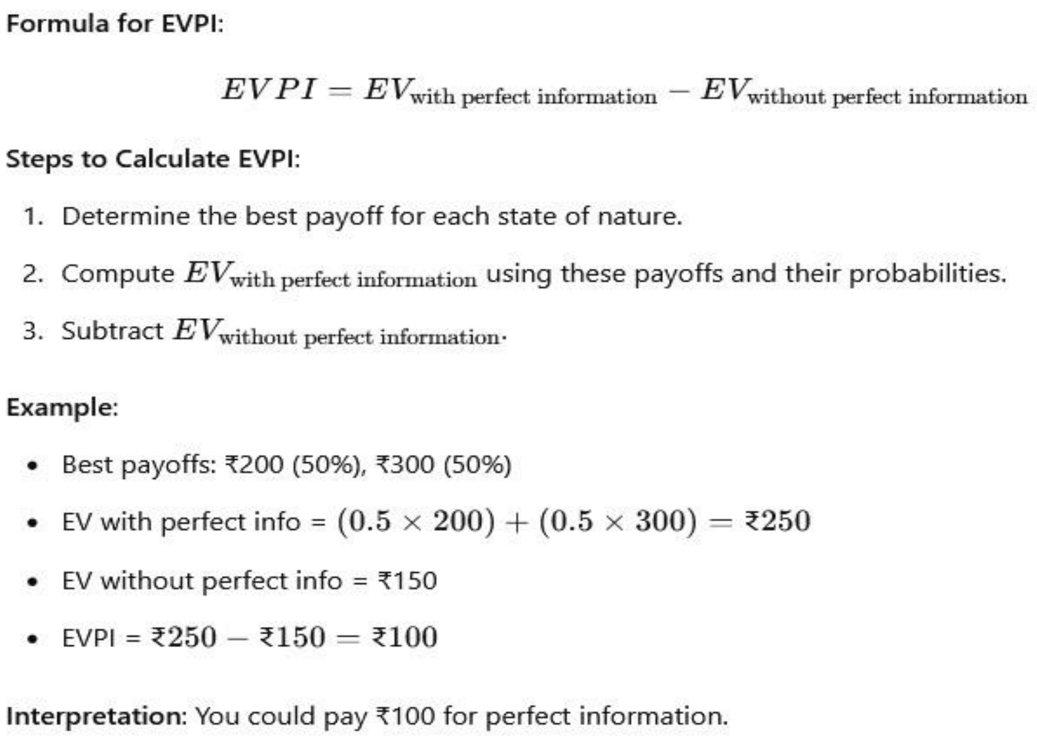
- The **Expected Value with Perfect Information (EVwPI)**: The best possible outcome if the decision-maker knew the future with certainty.
- The **Expected Value under Risk or Uncertainty (EVuR)**: The best decision given the current level of information.

Formula: **EVPI = EVwPI - EVuR**

Interpretation:

- If EVPI is high, it suggests that acquiring more information is valuable.
- If EVPI is low, it may not be worth investing in more information, as the potential benefit is small.

EVPI helps in making informed decisions about whether to invest in gathering more data.



Formula for EVPI:

$$EVPI = EV_{\text{with perfect information}} - EV_{\text{without perfect information}}$$

Steps to Calculate EVPI:

1. Determine the best payoff for each state of nature.
2. Compute $EV_{\text{with perfect information}}$ using these payoffs and their probabilities.
3. Subtract $EV_{\text{without perfect information}}$.

Example:

- Best payoffs: ₹200 (50%), ₹300 (50%)
- EV with perfect info = $(0.5 \times 200) + (0.5 \times 300) = ₹250$
- EV without perfect info = ₹150
- EVPI = $₹250 - ₹150 = ₹100$

Interpretation: You could pay ₹100 for perfect information.

Marginal analysis

Marginal analysis is a method used in decision-making to assess the additional benefits and costs of a decision or action. It focuses on the impact of making small changes in the quantity or level of an activity.

Key Concepts:

- **Marginal Benefit (MB)**: The extra benefit gained from consuming or producing one more unit of a good or service.
- **Marginal Cost (MC)**: The extra cost incurred from producing or consuming one more unit.

Process:

- In marginal analysis, decisions are made by comparing the marginal benefit to the marginal cost.
- If **MB > MC**, it's beneficial to increase the activity.
- If **MB < MC**, it's better to reduce the activity.
- The optimal decision occurs when **MB = MC**, meaning no additional benefits are gained from further action, and no extra costs are incurred.

Marginal analysis is widely used in economics, business, and finance to optimize resource allocation and maximize efficiency.

Decision Tree Analysis

Decision Tree Analysis is a graphical method used in statistics to make decisions under uncertainty. It helps visualize possible outcomes and the consequences of different decisions. The analysis involves breaking down a decision into a tree-like structure with branches representing decisions, outcomes, and probabilities.

Key Elements:

- **Root**: The starting point of the decision process.
- **Branches**: Represent possible decisions or events that may occur.
- **Nodes**: Points where decisions are made or outcomes are observed.
- **Leaves**: Final outcomes or payoffs, often associated with their probabilities.

Process:

1. Start with the initial decision at the root.
2. Branch out by considering possible choices and their outcomes.
3. Assign probabilities to each branch based on likelihood.
4. Calculate the expected value at each node, using probabilities and payoffs.
5. Compare the expected values to make the best decision.

Decision tree analysis helps identify optimal choices by considering all potential outcomes and associated risks.

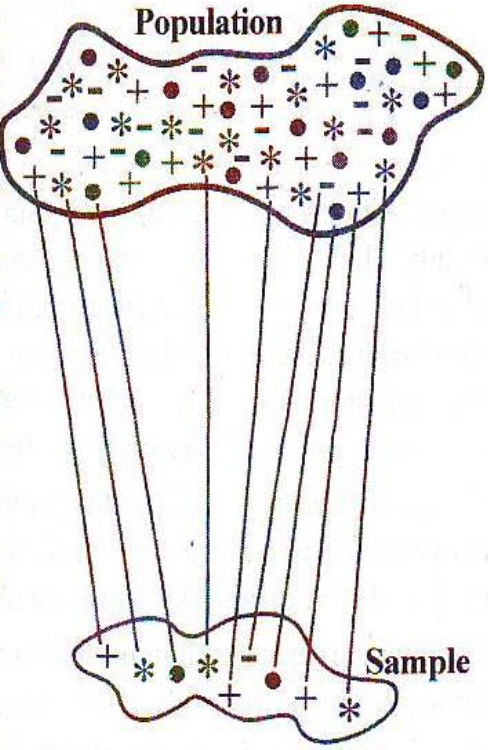
Unit III

INTRODUCTION

Statistical data are obtained either thought a survey or experiment. Statistical surveys are most popular device of obtaining the desired data. A survey is a process of collecting data from existing population units with no particular control over factor that may affect the population characteristics of interest in the study. Experimental data are generally obtained in studies relating to natural or physical sciences. A statistical survey may be either a general purpose or a special purpose survey. So far as general surveys is concerned we may obtain data which are useful for several purpose. The best example of this type of survey is the population census taken every 10 years in India. Such a survey provides information not only about the total population but also about its division into males and females, literates and illiterates, employment and unemployment, age distribution, income distribution, etc. In many cases it is impossible or impracticable to make a complete survey, in such cases we have to depend on sample study only.

MEANING OF UNIVERSE OR POPULATION

In the field of statistics the term 'Universe' or 'Population' The meaning of the totality of cases or items constituting the field or enquiry. Thus, "In statistics, population is the aggregate of objects under study in any statistical investigation". In any statistical investigation the interest usually lies in studying the various characteristics relating to items or individuals belonging to a particular group. This group of individuals under study is known as the population or universe.



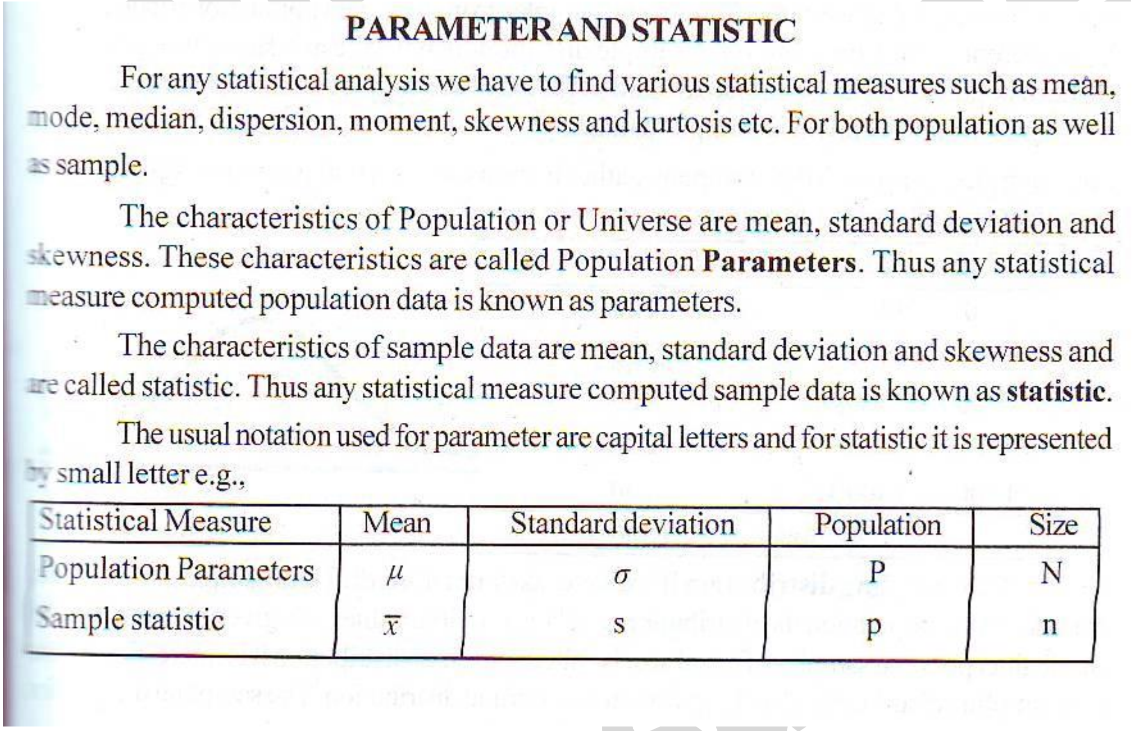
Finite and Infinite Population : The universe or population may be either finite or infinite. A finite population is hereby means a population having a determinable number of items. An infinite population is that in which the number of items cannot be determined.

MEANING OF SAMPLE

A sample is a part of universe or population selected for study to deduce some conclusions about population. According to L.R. Connor

All the items under consideration in any field of inquiry constitute a ‘universe’ or ‘population’. A complete enumeration of all the items in the ‘population’ is known as a census inquiry. It can be presumed that in such an inquiry when all the items are covered no element of chance is left and highest accuracy is obtained. But in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of observations increases. Moreover, there is no way of checking the element of bias or its extent except through are survey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy. Not only this, census inquiry is not possible in practice under many circumstances. For instance, blood testing is done only on sample basis. Hence, quite often we select only a few items from the universe for our study purposes. The items so selected constitute what is technically called a sample.

The researcher must decide the way of selecting a sample or what is popularly known as the sample design. In other words, a sample design is a definite plan determined before any data are actually collected for obtaining a sample from a given population. Thus, the plan to select 12 of a city’s 200 drugstores in a certain way constitutes a sample design.



PARAMETER AND STATISTIC

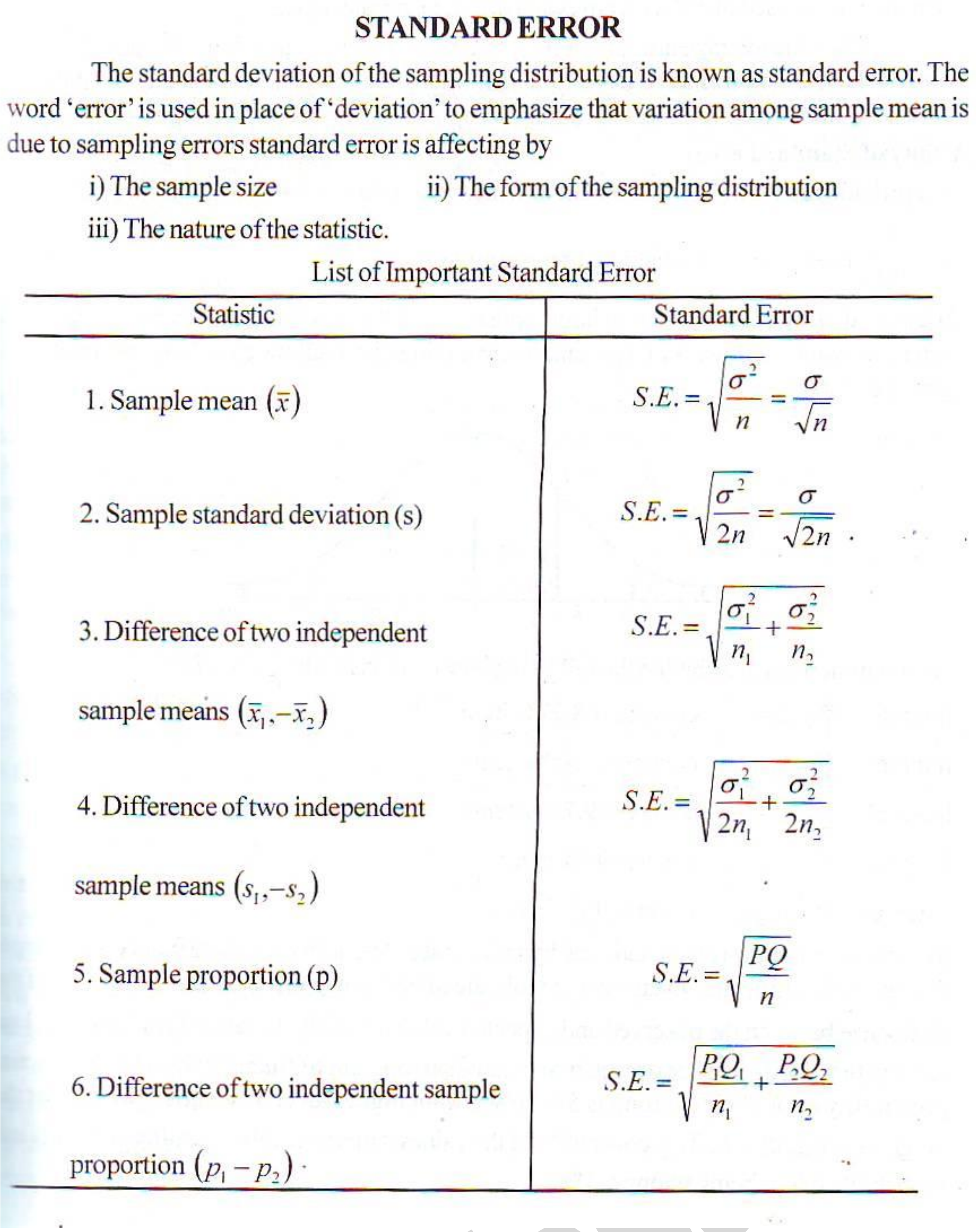
For any statistical analysis we have to find various statistical measures such as mean, mode, median, dispersion, moment, skewness and kurtosis etc. For both population as well as sample.

The characteristics of Population or Universe are mean, standard deviation and skewness. These characteristics are called Population **Parameters**. Thus any statistical measure computed population data is known as parameters.

The characteristics of sample data are mean, standard deviation and skewness and are called statistic. Thus any statistical measure computed sample data is known as **statistic**.

The usual notation used for parameter are capital letters and for statistic it is represented by small letter e.g.,

Statistical Measure	Mean	Standard deviation	Population	Size
Population Parameters	μ	σ	P	N
Sample statistic	$\bar{x}$	s	p	n



STANDARD ERROR

The standard deviation of the sampling distribution is known as standard error. The word 'error' is used in place of 'deviation' to emphasize that variation among sample mean is due to sampling errors standard error is affecting by

i) The sample size ii) The form of the sampling distribution

iii) The nature of the statistic.

List of Important Standard Error

Statistic	Standard Error
1. Sample mean $(\bar{x})$	$S.E.=\sqrt{\frac{\sigma^2}{n}}=\frac{\sigma}{\sqrt{n}}$
2. Sample standard deviation (s)	$S.E.=\sqrt{\frac{\sigma^2}{2n}}=\frac{\sigma}{\sqrt{2n}}$
3. Difference of two independent sample means $(\bar{x}_1,-\bar{x}_2)$	$S.E.=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
4. Difference of two independent sample means $(s_1,-s_2)$	$S.E.=\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
5. Sample proportion (p)	$S.E.=\sqrt{\frac{PQ}{n}}$
6. Difference of two independent sample proportion (p_1-p_2)	$S.E.=\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

METHODS OF SAMPLING

Meaning: - The process of obtaining a sample and its subsequent analysis and interpretation is known as sampling and the process of obtaining the sample if the first stage of sampling.

The various methods of sampling can broadly be divided into:

i. Random sampling method
ii. Non Random sampling method

Random Sampling Method

I Simple Random Sampling: - In this method each and every item of the population is given an equal chance of being included in the sample.

(a) Lottery Method (b) Table of Random Numbers

Merits:

Equal opportunity to each item.
Better way of judgment
Easy analysis and accuracy

Limitations:

Different in investigation
Expensive and time consuming
For filed survey it is not good

II Stratified Sampling:- In this it is important to divided the population into homogeneous group called strata. Then a sample may be taken from each group by simple random method.

Merit:- More representative sample is used.
Grater accuracy
Geographically Concentrated

Limitations: Utmost care must be exercised due to homogeneous group deviation. In the absence of skilled supervisor sample selection will be difficult.

III Systematic Sampling:- This method is popularly used in those cases where a complete list of the population from which sampling is to be drawn is available. The method is to be select k th item from the list where k refers to the sampling interval.

Merits: - It can be more convenient.

Limitation: - Can be Baised.

IV Multi- Stage Sampling: - This method refers to a sampling procedure which is carried out in several stages.

Merit: - It gives flexibility in Sampling

Limitation: - It is difficult and less accurate

Non Random Sampling Method:-

I. **Judgment Sampling: -** The choice of sample items depends exclusively on the judgment of the investigator or the investigator exercises his judgement in the choice of sample items. This is an simple method of sampling.

II. **Quota Sampling: -** Quotas are set up according to given criteria, but, within the quotas the selection of sample items depends on personal judgment.

III. **Convenience Sampling: -** It is also known as chunk. A chunk is a fraction of one population taken for investigation because of its convenient availability. That is why a chunk is selected neither by probability nor by judgment but by convenience.

Size of Sample:- It depends upon the following things:-

Cost aspects. The degree of accuracy desired. Time, etc. Normally it is 5% or 10% of the total population.

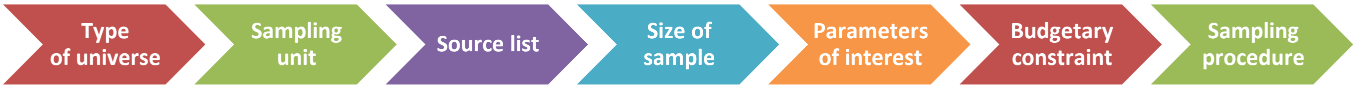
Limitation of overall sampling Method:-

Some time result may be inaccurate and misleading due to wrong sampling.
Its always needs superiors and experts to analyze the sample.
It may not give information about the overall defects. In production or any study.
It Becomes Biased due to following reason:-

(a) Faulty process of selection
(b) Faulty work during the collection of information
(c) Faulty methods of analysis etc.

STEPS IN SAMPLE DESIGN

While developing a sampling design, the researcher must pay attention to the following points:



Type of universe:

The first step in developing any sample design is to clearly define the set of objects, technically called the Universe, to be studied. The universe can be finite or infinite. In finite universe the number of items is certain, but in case of an infinite universe the number of items is infinite, i.e., we cannot have any idea about the total number of items. The population of a city, the number of workers in a factory and the like are examples of finite universes, whereas the number of stars in the sky, listeners of a specific radio programme, throwing of a dice etc. are examples of infinite universes.

Sampling unit:

A decision has to be taken concerning a sampling unit before selecting sample. Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual. The researcher will have to decide one or more of such units that he has to select for his study.

Source list:

It is also known as 'sampling frame' from which sample is to be drawn. It contains the names of all items of a universe (in case of finite universe only). If source list is not available, researcher has to prepare it. Such a list should be comprehensive, correct, reliable and appropriate. It is extremely important for the source list to be as representative of the population as possible.

Size of sample:

This refers to the number of items to be selected from the universe to constitute a sample. This is a major problem before a researcher. The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfills the requirements of efficiency, representativeness, reliability and flexibility. While deciding the size of sample, researcher must determine the desired precision as also an acceptable confidence level for the estimate. The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed. The size of population must be kept in view for this also limits the sample size. The parameters of interest in a research study must be kept in view, while deciding the size of the sample. Costs too dictate the size of sample that we can draw. As such, budgetary constraint must invariably be taken into consideration when we decide the sample size.

Parameters of interest:

In determining the sample design, one must consider the question of the specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population. There may also be important sub-groups in the population about whom we would like to make estimates. All this has a strong impact upon the sample design we would accept.

Budgetary constraint:

Cost considerations, from practical point of view, have a major impact upon decisions relating to not only the size of the sample but also to the type of sample. This fact can even lead to the use of a non-probability sample.

Sampling procedure:

Finally, the researcher must decide the type of sample he will use i.e., he must decide about the the items for the sample. Infact, this technique or procedure stands for the sample design technique to be used in selecting itself. There are several sample designs (explained in the pages that follow) out of which the researcher must choose one for his study. Obviously, he must select which, for a given sample size and for a given cost, has a smaller sampling error.

CRITERIA OF SELECTING A SAMPLING PROCEDURE

In this context one must remember that two costs are involved in a sampling analysis viz., the cost of collecting the data and the cost of an incorrect inference resulting from the data. Researcher must keep in view the two causes of incorrect inferences viz., systematic bias and sampling error. A

systematic bias results from errors in the sampling procedures, and it cannot be reduced or eliminated by increasing the sample size. At best the causes responsible for these errors can be detected and corrected. Usually a systematic bias is the result of one or more of the following factors:

1. Inappropriate sampling frame:

If the sampling frame is inappropriate i.e., a biased representation of the universe, it will result in a systematic bias.

2.Defective measuring device:

If the measuring device is constantly in error, it will result in systematic bias. In survey work, systematic bias can result if the questionnaire or the interviewer is biased. Similarly, if the physical measuring device is defective there will be systematic bias in the data collected through such a measuring device.

3. Non-respondents:

If we are unable to sample all the individuals initially included in the sample, there may arise a systematic bias. The reason is that in such a situation the likelihood of establishing contact or receiving a response from an individual is often correlated with the measure of what is estimated.

4.Indeterminancy principle:

Sometimes we find that individuals act differently when kept under observation than what they do when kept in non-observed situations. For instance, if workers are aware that somebody is observing them in course of a work study on the basis of which the average length of time to complete a task will be determined and accordingly the quota will be set for piecework, they generally tend to work slowly in comparison to the speed with which they work if kept unobserved. Thus, the indeterminacy principle may also be a cause of a systematic bias.

5.Natural bias in the reporting of data:

Natural bias of respondents in the reporting of data is often the cause of a systematic bias in many inquiries. There is usually a downward bias in the income data collected by government taxation department, whereas we find an upward bias in the income data collected by some social organisation. People in general understate their incomes if asked about it for tax purposes, but they overstate the same if asked for social status or their affluence. Generally in psychological surveys, people tend to give what they think is the 'correct' answer rather than revealing their true feelings.

CHARACTERISTICS OF A GOOD SAMPLE DESIGN

From what has been stated above, we can list down the characteristics of a good sample design asunder:

(a)Sample design must result in a truly representative sample.

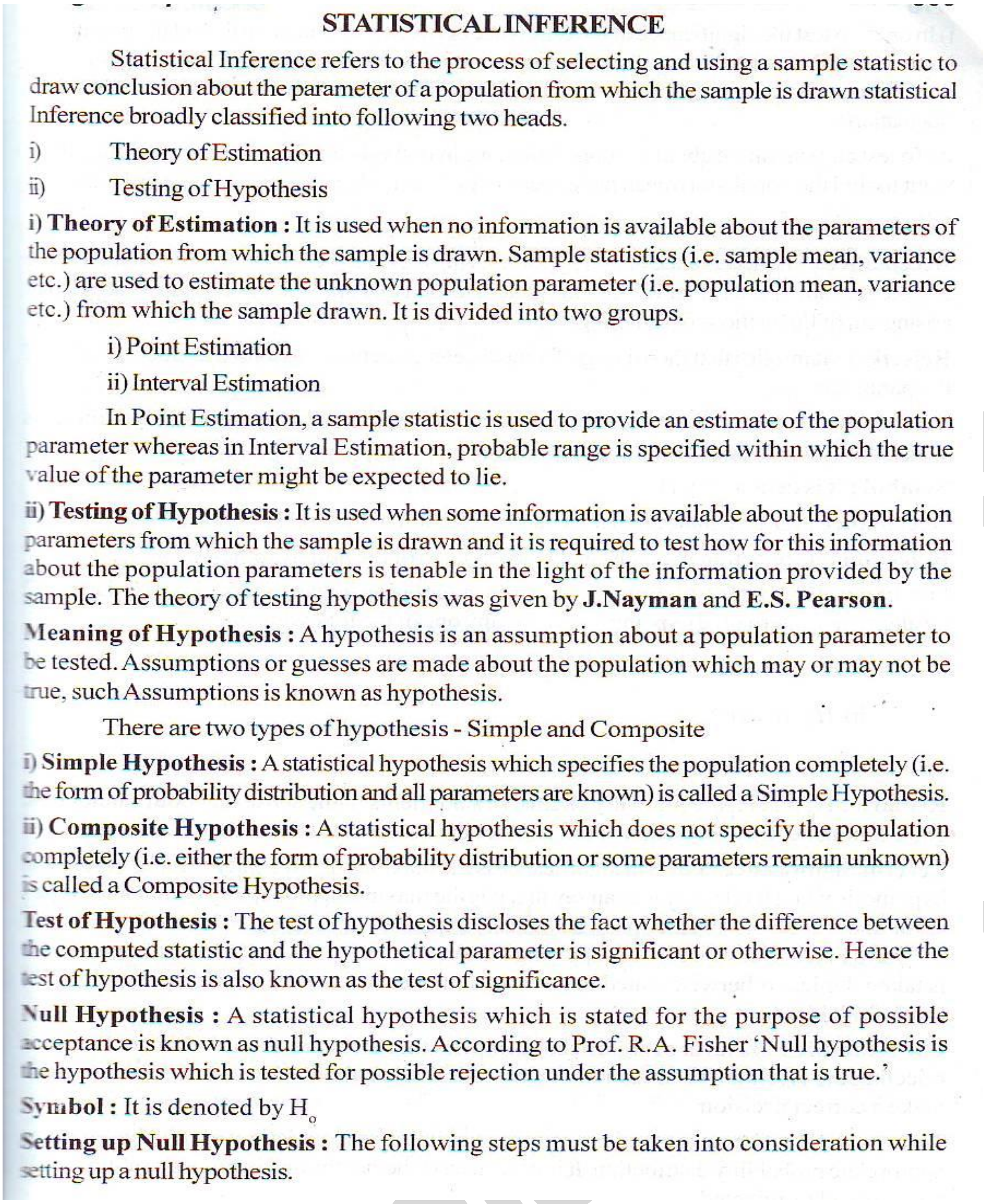
(b)Sample design must be such which results in a small sampling error.

(c)Sample design must be viable in the context of funds available for the research study.

(d)Sample design must be such so that systematic bias can be controlled in a better way.

(e)Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence

Unit IV



STATISTICAL INFERENCE

Statistical Inference refers to the process of selecting and using a sample statistic to draw conclusion about the parameter of a population from which the sample is drawn statistical Inference broadly classified into following two heads.

i) Theory of Estimation

ii) Testing of Hypothesis

i) **Theory of Estimation :** It is used when no information is available about the parameters of the population from which the sample is drawn. Sample statistics (i.e. sample mean, variance etc.) are used to estimate the unknown population parameter (i.e. population mean, variance etc.) from which the sample drawn. It is divided into two groups.

i) Point Estimation

ii) Interval Estimation

In Point Estimation, a sample statistic is used to provide an estimate of the population parameter whereas in Interval Estimation, probable range is specified within which the true value of the parameter might be expected to lie.

ii) **Testing of Hypothesis :** It is used when some information is available about the population parameters from which the sample is drawn and it is required to test how for this information about the population parameters is tenable in the light of the information provided by the sample. The theory of testing hypothesis was given by **J.Nayman** and **E.S. Pearson**.

Meaning of Hypothesis : A hypothesis is an assumption about a population parameter to be tested. Assumptions or guesses are made about the population which may or may not be true, such Assumptions is known as hypothesis.

There are two types of hypothesis - Simple and Composite

i) **Simple Hypothesis :** A statistical hypothesis which specifies the population completely (i.e. the form of probability distribution and all parameters are known) is called a Simple Hypothesis.

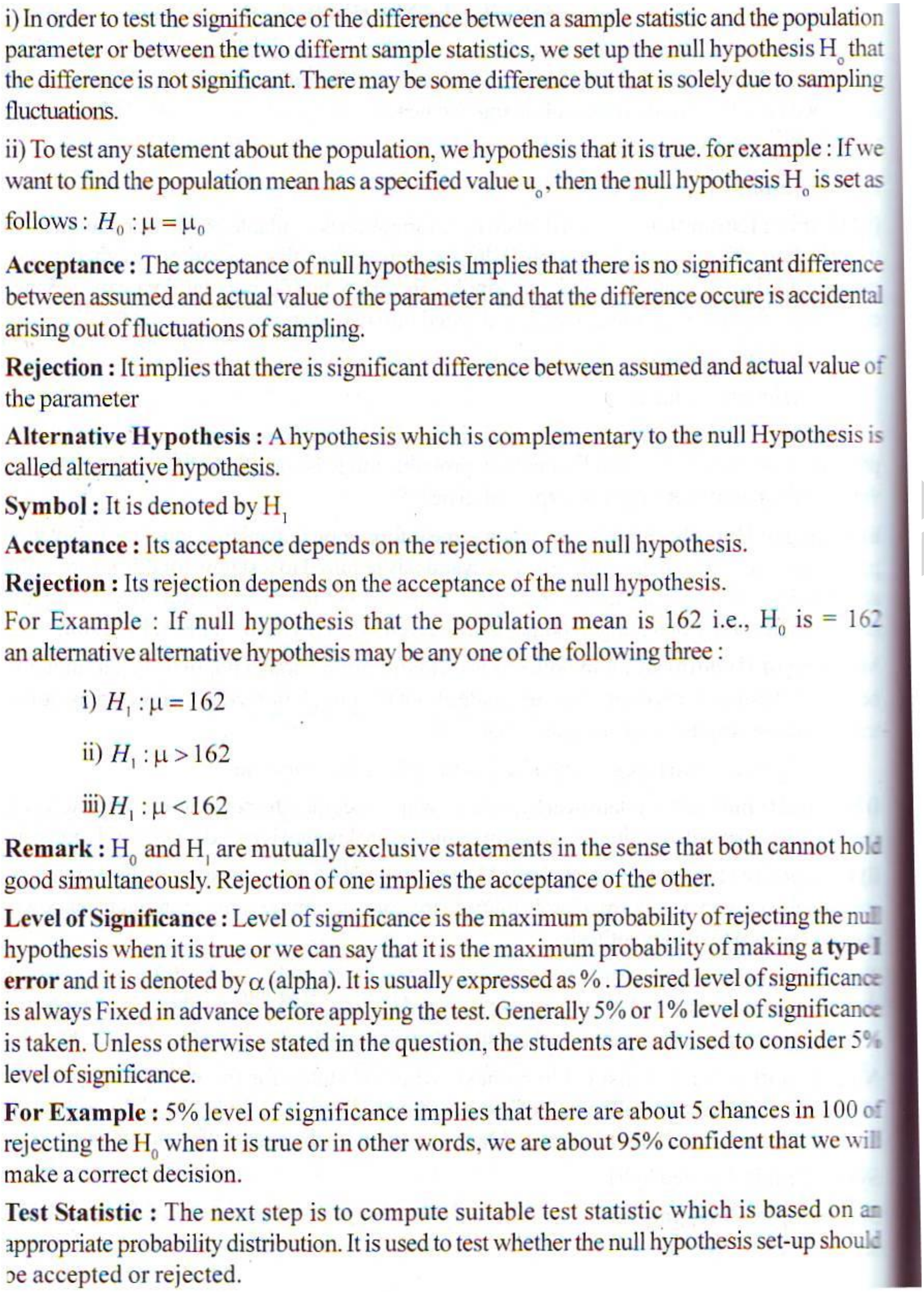
ii) **Composite Hypothesis :** A statistical hypothesis which does not specify the population completely (i.e. either the form of probability distribution or some parameters remain unknown) is called a Composite Hypothesis.

Test of Hypothesis : The test of hypothesis discloses the fact whether the difference between the computed statistic and the hypothetical parameter is significant or otherwise. Hence the test of hypothesis is also known as the test of significance.

Null Hypothesis : A statistical hypothesis which is stated for the purpose of possible acceptance is known as null hypothesis. According to Prof. R.A. Fisher 'Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that is true.'

Symbol : It is denoted by H_o

Setting up Null Hypothesis : The following steps must be taken into consideration while setting up a null hypothesis.



i) In order to test the significance of the difference between a sample statistic and the population parameter or between the two diffent sample statistics, we set up the null hypothesis H_0 that the difference is not significant. There may be some difference but that is solely due to sampling fluctuations.

ii) To test any statement about the population, we hypothesis that it is true. for example : If we want to find the population mean has a specified value u_0, then the null hypothesis H_0 is set as follows : $H_0 : \mu = \mu_0$

Acceptance : The acceptance of null hypothesis Implies that there is no significant difference between assumed and actual value of the parameter and that the difference occure is accidental arising out of fluctuations of sampling.

Rejection : It implies that there is significant difference between assumed and actual value of the parameter

Alternative Hypothesis : A hypothesis which is complementary to the null Hypothesis is called alternative hypothesis.

Symbol : It is denoted by H_1

Acceptance : Its acceptance depends on the rejection of the null hypothesis.

Rejection : Its rejection depends on the acceptance of the null hypothesis.

For Example : If null hypothesis that the population mean is 162 i.e., H_0 is = 162 an alternative alternative hypothesis may be any one of the following three :

i) $H_1 : \mu = 162$

ii) $H_1 : \mu > 162$

iii) $H_1 : \mu < 162$

Remark : H_0 and H_1 are mutually exclusive statements in the sense that both cannot hold good simultaneously. Rejection of one implies the acceptance of the other.

Level of Significance : Level of significance is the maximum probability of rejecting the null hypothesis when it is true or we can say that it is the maximum probability of making a **type I error** and it is denoted by α (alpha). It is usually expressed as % . Desired level of significance is always Fixed in advance before applying the test. Generally 5% or 1% level of significance is taken. Unless otherwise stated in the question, the students are advised to consider 5% level of significance.

For Example : 5% level of significance implies that there are about 5 chances in 100 of rejecting the H_0 when it is true or in other words, we are about 95% confident that we will make a correct decision.

Test Statistic : The next step is to compute suitable test statistic which is based on an appropriate probability distribution. It is used to test whether the null hypothesis set-up should be accepted or rejected.

Test - Statistic

i) Z-test	For test of Hypothesis involving large sample i.e. > 30
ii) t-test	For test of Hypothesis involving small sample i.e. ≤ 30
iii) χ^2-test	For testing the significant difference between observed frequencies and expected frequencies.
iv) F-test	For testing the sample variances.

The region of the standard normal curve corresponding to a pre-determined level of significance that is fixed for knowing the probability of making the type I error or rejecting the hypothesis which is true, is known as the "rejection region" or "critical region". The region of standard normal curve that is not covered by the rejection region, is called "accepted region". When the test statistic computed to test the hypothesis falls in the acceptance region, it is reasonable to accept the hypothesis as it is believed to be probably true.

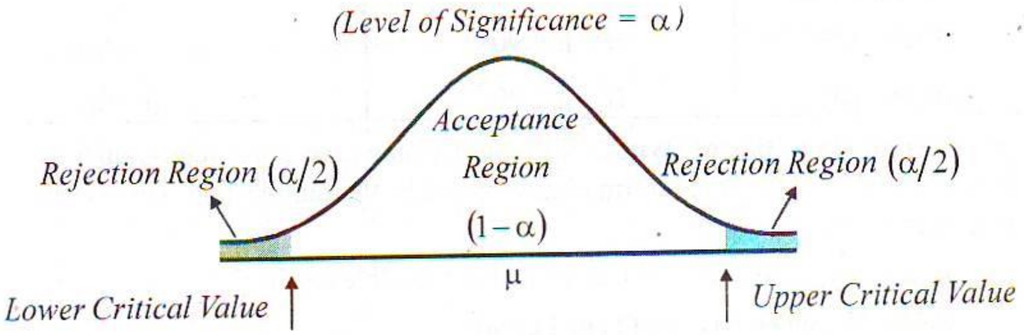
Two tailed test and one tailed test : The critical region may be shown by a portion of the area under the normal curve in two ways.

i) Two Tails

ii) One Tail (right tail or left tail)

i) **Two Tailed Test :** When the test of hypothesis is made on the basis of rejection region represented by both sides of the standard normal curve, it is called a two tailed test or two sides test for example :

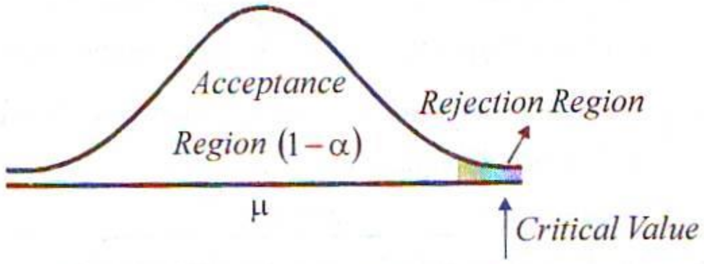
Null Hypothesis $(H_0): \mu = 90$



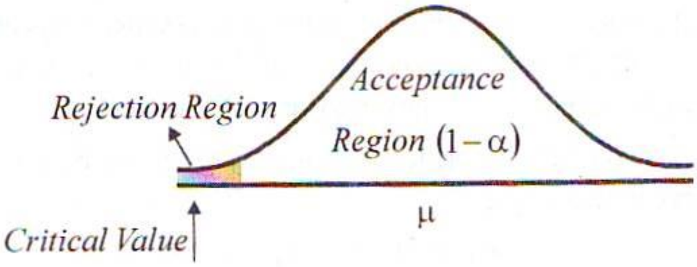
Alternative Hypothesis $(H_1): \mu \neq 90$ (i.e. either $\mu > 90$ or $\mu < 90$)

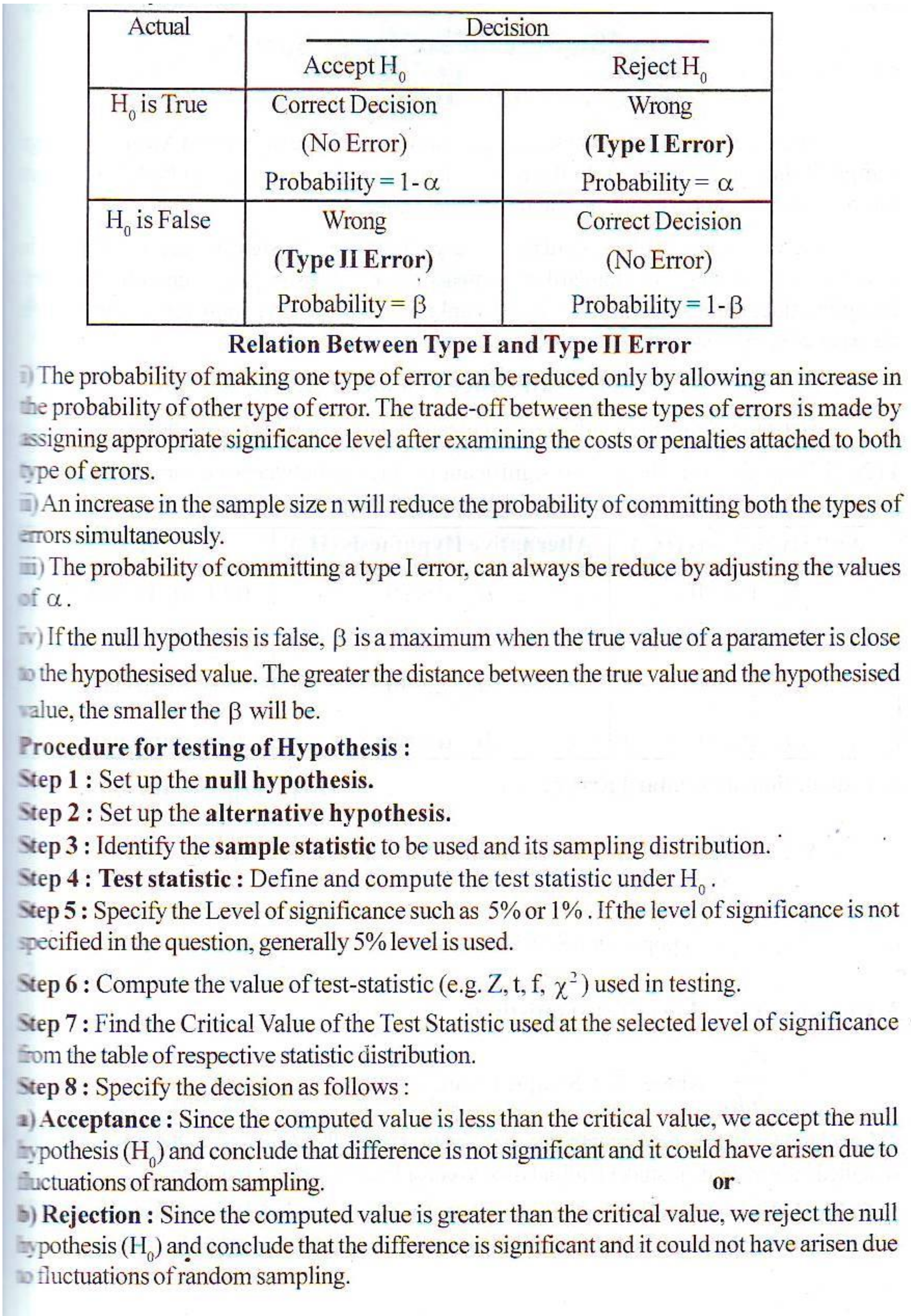
i) **One Tailed Test :** The one tail test is used in cases where it is considered that the population mean is at least as large as some specified value of mean or at least as small as some specified value of mean. There are two types of one tailed tests

i) Right Tailed Test : In the right tailed test the rejection region or critical region lies entirely on the right tail of the normal curve.



i) Left Tailed Test : In the left tailed test the critical region or rejection region lies entirely on the left tail of the normal curve.





Actual	Decision Accept H_0	 Reject H_0
H_0 is True	Correct Decision (No Error) Probability $= 1 - \alpha$	Wrong **(Type I Error)** Probability $= \alpha$
H_0 is False	Wrong **(Type II Error)** Probability $= \beta$	Correct Decision (No Error) Probability $= 1 - \beta$

Relation Between Type I and Type II Error

i) The probability of making one type of error can be reduced only by allowing an increase in the probability of other type of error. The trade-off between these types of errors is made by assigning appropriate significance level after examining the costs or penalties attached to both type of errors.

ii) An increase in the sample size n will reduce the probability of committing both the types of errors simultaneously.

iii) The probability of committing a type I error, can always be reduce by adjusting the values of α.

iv) If the null hypothesis is false, β is a maximum when the true value of a parameter is close to the hypothesised value. The greater the distance between the true value and the hypothesised value, the smaller the β will be.

Procedure for testing of Hypothesis :

Step 1 : Set up the **null hypothesis.**

Step 2 : Set up the **alternative hypothesis.**

Step 3 : Identify the **sample statistic** to be used and its sampling distribution.

Step 4 : Test statistic : Define and compute the test statistic under H_0.

Step 5 : Specify the Level of significance such as 5% or 1%. If the level of significance is not specified in the question, generally 5% level is used.

Step 6 : Compute the value of test-statistic (e.g. Z, t, f, χ^2) used in testing.

Step 7 : Find the Critical Value of the Test Statistic used at the selected level of significance from the table of respective statistic distribution.

Step 8 : Specify the decision as follows :

a) **Acceptance :** Since the computed value is less than the critical value, we accept the null hypothesis (H_0) and conclude that difference is not significant and it could have arisen due to fluctuations of random sampling. **or**

b) **Rejection :** Since the computed value is greater than the critical value, we reject the null hypothesis (H_0) and conclude that the difference is significant and it could not have arisen due to fluctuations of random sampling.

UNIT - V

CHI-SQUARE TEST

$$\chi^2 = \sum \frac{(\text{Observed Value} - \text{Expected Value})^2}{(\text{Expected Value})}$$

Degrees of freedom (df) =(n-1), where n is the number of classes

Analysis of Variance (ANOVA)

Purpose

1) The reason for doing an ANOVA is to see if there is any difference between groups on some variable.

2) For example, you might have data on student performance in non-assessed tutorial exercises as well as their final grading. You are interested in seeing if tutorial performance is related to final grade. ANOVA allows you to break up the group according to the grade and then see if performance is different across these grades.

ANOVA is available for both parametric (score data) and non-parametric (ranking/ordering) data.

Types of ANOVA

One-way between groups

The example given above is called a **one-way between groups model.**

You are looking at the differences between the groups.

There is only one grouping (final grade) which you are using to define the groups.

This is the simplest version of ANOVA.

This type of ANOVA can also be used to compare variables between different groups - tutorial performance from different intakes.

One-way repeated measures

A one way repeated measures ANOVA is used when you have a single group on which you have measured something a few times.

For example, you may have a test of understanding of Classes. You give this test at the beginning of the topic, at the end of the topic and then at the end of the subject.

You would use a one-way repeated measures ANOVA to see if student performance on the test changed over time.

Two-way between groups

A two-way between groups ANOVA is used to look at complex groupings.

For example, the grades by tutorial analysis could be extended to see if overseas students performed differently to local students. What you would have from this form of ANOVA is:

The effect of final grade

The effect of overseas versus local

The interaction between final grade and overseas/local

Each of the **main effects** are one-way tests. The **interaction effect** is simply asking "is there any significant difference in performance when you take final grade and overseas/local acting together".

Use of Multivariate Analysis in Research

1) Many statistical techniques focus on just one or two variables. Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once. Multivariate statistical analysis refers to multiple advanced techniques for examining relationships among multiple variables at the same time. Researchers use multivariate procedures in studies that involve more than one dependent variable (also known as the outcome or phenomenon of interest), more than one independent variable (also known as a predictor) or both. Upper-level undergraduate courses and graduate courses in statistics teach multivariate statistical analysis. This type of analysis is desirable because researchers often hypothesize that a given outcome of interest is effected or influenced by more than one thing.

2) Multiple regression is not typically included under this heading, but can be thought of as a multivariate analysis.

3) Commonly have many relevant variables in market research surveys

– E.g. one not atypical survey had ~2000 variables

– Typically researchers pore over many crosstabs

– However it can be difficult to make sense of these, and the crosstabs may be misleading

• MVA can help summarise the data

– E.g. factor analysis and segmentation based on agreement ratings on 20 attitude statements

• MVA can also reduce the chance of obtaining spurious results

Multivariate Analysis Methods

• Two general types of MVA technique

– Analysis of dependence

• Where one (or more) variables are dependent variables, to be explained or predicted by others

– E.g. Multiple regression, PLS, MDA

– Analysis of interdependence

• No variables thought of as "dependent"

• Look at the relationships among variables, objects or cases

– E.g. cluster analysis, factor analysis