Directorate of Distance and Continuing Education

Manonmaniam Sundaranar University

Tirunelveli-627012, Tamil Nadu.

M.A.ECONOMICS

(First Year)

Statistics for Economists

(SECM13)

Prepared by

Dr. G. Monikanda Prasad

Assistant Professor of Economics

Manonmaniam Sundaranar University

Tirunelveli-627012.

Subject Experts:

Dr. D. Amutha Associate Professor and Head Department of Economics S.T. Mary's College (Autonomous) Thoothukudi	**Dr. C. A. Sham Shankar** Associate Professor and Head Department of Economics ST. Hindu College Nagercoil
Dr. S. Sarasudevi Associate Professor Department of Economics Rani Anna Government College for Women Tirunelveli	**Dr. V. Raja Rajeswari** Associate Professor Department of Economics Sri KGS Arts College Srivaikundam, Thoothukudi District
Dr. M. Neeladevi Associate Professor Department of Economics V.O.C College Thoothukudi	**Dr. P. Mary Thangam** Assistant Professor Department of Economics Sarah Turker College (Autonomous) Tirunelveli
Dr. R. Rajan Babu Assistant Professor Department of Economics ST. Hindu College Nagercoil	**Dr. V. Arockia Amuthan** Assistant Professor Department of Economics Nazareth Margoschis College Pillaiyanmanai, Thoothukudi
Dr. S. Sasikumar Assistant Professor Department of Economics ST. Xavier's College (Autonomous) Palayamkottai, Tirunelveli	**Dr. G. Gnana Elpinston** Assistant Professor Department of Economics Nesamony Memorial Christian College Marthandam

Course Coordinator:

Dr. G. Monikanda Prasad

Course Material Compiled by:

Dr. G. Monikanda Prasad
Assistant Professor of Economics
Manonmaniam Sundaranar University
Tirunelveli – 627 012

STATISTICS FOR ECONOMISTS

Course Objective:

1. To provide a strong foundation in statistical concepts and develop skills in data handling and research.
2. The course facilitates in inferring the intensity of relationship between multiple variables and building appropriate statistical models. The models thus formulated can be tested for their significance and can be used for forecasting

UNIT I : Probability

Probability – Addition – Multiplication Theorems – Conditional Probability – Discrete and Continuous – Random Variables – Mathematical Expectation – Bayes Theorem – Theoretical Distributions – Binomial, Poisson and Normal.

UNIT II : Sampling and Hypothesis Testing

Sampling Theory – Types of sampling – Sampling Distributions – Parameter and Statistic – Testing of Hypothesis – Level of Significance – Types I and Type II Errors – Standard Error - Properties of Estimator.

UNIT III : Testing of Significance Large and Small Sample

Difference between Large and Small Samples – Test of Significance for Large Samples – Test for Two Means and Standard Deviations – Proportion and Confidence Interval – Small Sample Test – t-test – Paired t-test – Chi-square Test – Test of Goodness of Fit.

UNIT IV : Analysis of Variance

F test: Assumption in F test – Analysis of Variance: Assumptions – One-Way and Tow-Way Classifications.

UNIT V : Statistical Decision Theory

Definitions – Concepts - Maximin – Minimax – Bayes Criterion – Expected Monetary Value – Decision Tree Analysis: Symbols – Steps –Advantages and Limitations.

Text Books

1. Gupta S.P., Statistical Methods, Sultan Chand and Sons, New Delhi, 2017.

2. Anderson, Sweeney and Williams, “Statistic for Business and Economics”, Cengage, 2014.

UNIT I

PROBABILITY

Probability – Addition – Multiplication Theorems – Conditional Probability – Discrete and Continuous – Random Variables – Mathematical Expectation – Bayes Theorem Theoretical Distributions – Binomial, Poisson and Normal.

The word probability or a chance is very commonly used in day-to-day conversation and generally people have a vague about its meaning. The theory of probability has its origin in the games of chance related to gambling such as throwing a die, tossing a coin, drawing cards from a pack of cards etc.

Theory of probability was developed in the middle of the 17^{th} century. The names which are associated with probability are Hugenes, Pascal, Format, Berouli, Laplace, Bayes. Today, the theory of probability has been developed to a great extent and extensively used in various subjects. The theory of probability was developed from gambling as it is a game of chance.

Importance of probability

- The whole sampling theory, particularly the principle of law of statistical regularity and the law of inertia of large numbers, was developed on the basis of probability theory.
- It is very useful to solve problems related to betting gambling.
- The decision theory is also constructed on the probability theory.
- Different tests used for testing of hypothesis are derived from probability.

Basic concepts

1. **Random experiment**: An experiment or trial outcome is uncertain is called random experiment. Though an experiment is repeated under the same conditions, individual outcome is not predictable. Such an experiment is called random experiment.

2. **Event:** Any possible outcome of a random experiment is called an event. Performing a random experiment is trail and the occurrence or non-occurrence of something is an event. The occurrence of an event which is inevitable, when a random experiment is performed, is called sure event or certain event. On the other hand is the occurrence is impossible, it is called impossible event.
3. **Simple and compound events**: Here the classification is made on the basic of the number of events in question. If only one event is take place at a time, it is called simple event.
4. **Mutually exclusive event**: If the occurrence of the event excludes the occurrence of the alternative, the events are called mutually exclusive. Simply, in mutually exclusive events simultaneous occurrence of events is not possible. Such events are called alternative events or incompatible events.
5. **Equally likely events**: When the change of occurrence of each events is the same, the events are called equally likely events. They are also called equiprobable events. In such events, one events does not occur more often the other.
6. **Independent** events: In independent events, the occurrence of one event doesn't affect and is not affected by the other. For instance, when we toss a coin twice, the result of the second toss will not in any way be affected by the result of the first toss. In other words, the result of the first toss does not affect the result of the second event.
7. **Dependent events:** Two events are said to be dependent, if the occurrence or non-occurrence of one event affects occurrence of the other. In other words, the occurrence of an event affects the result of the subsequent trail, then such events are called department events.

Theorem of probability

Theorems explain functional relationship existing between variables or attributes. The functional relationships or laws are formed to tackle complex situations. The law of probability has also helped to tackle complex situations that arise in the field of probability. There are two important theorems,

1. Addition theorem
2. Multiplication theorem

1.Addition theorem: The addition rule is the simplest and frequently used rule to determine probability. If two events are mutually exclusive, then the probability of happening either 'A' or 'B' is the sum of their separate probabilities. It is also called theorem of total probability.

$$P(A or B) = P(A) + P(B)$$

Proof of the Theorem:

If an event A can happen in a_1 ways andB can happen in a_2 ways, then the number of ways in which either event can happen is $a_1 + a_2$.

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n}$$

$$\frac{a_1}{n} = P(A)$$

$$\frac{a_2}{n} = P(B)$$

$$P(A or B) = P(A) + P(B)$$

Hence, the theorem can be extended to three or more mutually exclusive events.

$$P(A or B) = P(A) + P(B) + \mathrm{P(C)}$$

2.Multiplication theorem: The multiplication law states that "the probability of happening of given 2 events or in different words the probability of the intersection of 2 given events is equivalent to the product achieved by finding out the product of the probability of happening of both the events."

Proof of the Theorem:

If an event A can happen in n_1 ways of which a_1 are successful and the event B can happen in n_2 ways, of which a_2 are successful, then the number of ways in which either event can happen is $n_1 + n_2$.

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} + \frac{a_2}{n_2}$$

$$\frac{a_1}{n_1} = P(A)$$

$$\frac{a_2}{n_2} = P(B)$$

$$P(A \, and \, B) = P(A) \times P(B)$$

Hence, the theorem can be extended to three or more mutually exclusive events.

$$P(A, B \, and \, C) = P(A) \times P(B) \times \mathrm{P(C)}$$

Example:1

One card is drawn from a standard pack of 52. What is the probability that it is either a king or a queen?

Solution

There are 4 kings and 4 queens in a pack of 52 cards.

The probability that the card drawn is a king = $\frac{4}{52}$ and the probability that the card drawn is a queen = $\frac{4}{52}$

Since the events are mutually exclusive, the probability that the card drawn is either a king or a queen = $\frac{4}{52}+\frac{4}{52}=\frac{8}{52}=\frac{2}{13}$

Example:2

A man wants to marry a girl having white complexion-the probability of getting such a girl is one in 20, handsome dowry-the probability of getting this is one in fifty, westernized manners and etiquettes-the probability here is one in hundred. Find out the probability of his getting married to such a girl when the possession of these three attributes is independent.

Solution:

Probability of a girl with the complexion = $\frac{1}{20}=0.05$

Probability of a girl with handsome dowry = $\frac{1}{50}=0.02$

Probability of a girl with westernized manners = $\frac{1}{100}=0.01$

Since the events are independent, the probability of simultaneous occurrence of all these qualities

$=\frac{1}{20}\times\frac{1}{50}\times\frac{1}{100}=0.05\times0.02\times0.01$

$=0.00001$

Conditional Probability

In probability theory, **conditional probability** is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) is already known to have occurred.This particular method relies on event B occurring with some sort of relationship with another event A. In this event, the event B can be analyzed by a conditional probability with respect to A. If the event of interest

is *A* and the event *B* is known or assumed to have occurred, "the conditional probability of *A* given *B*", or "the probability of *A* under the condition *B*", is usually written as P(*A*/*B*) or occasionally $P_B(A)$. This can also be understood as the fraction of probability B that intersects with A, or the ratio of the probabilities of both events happening to the "given" one happening (how many times A occurs rather than not assuming B has occurred):

If two events A and B are dependent, then the conditional probability of B given A is

$$P(B / A) = \frac{P(AB)}{P(A)}$$

Example:3

A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.

Solution:

Probability of drawing a black ball in the first attempt is $P(A) = \frac{3}{5+3} = \frac{3}{8}$

Probability of drawing a second black ball given that the first ball drawn is black

$$P(\text{B/ A}) = \frac{2}{5+2} = \frac{2}{7}$$

The probability that both balls drawn are black is given by

$$P(\text{AB}) = P(A) \times P(B / A) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$

Discrete and continuous

Discrete data is information that has noticeable gaps between values. Continuous data is information that occurs in a continuous series. Discrete data is made up of discrete or distinct values. Directly in opposition, continuous data includes any value that falls inside a range

Random variables

Key Takeaways: A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. A random variable can be either discrete (having specific values) or continuous.

A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. Random variables are often designated by letters and can be classified as discrete, which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.

Types of Random Variables

Likelihood that any of the possible values would occur. Let's say that the random variable, Z, is the number on the top face of a die when it is rolled once. The possible values for Z will thus be 1, 2, 3, 4, 5, and 6. The probability of each of these values is 1/6

A random variable has a probability distribution that represents the as they are all equally likely to be the value of Z.

For instance, the probability of getting a 3, or P (Z=3), when a die is thrown is 1/6, and so is the probability of having a 4 or a 2 or any other number on all six faces of a die. Note that the sum of all probabilities is 1.

A random variable can be either discrete or continuous.

Discrete Random Variables

Discrete random variables take on a countable number of distinct values. Consider an experiment where a coin is tossed three times. If X represents the number of times that the coin comes up heads, then X is a discrete random variable that can only have the values 0, 1, 2, or 3 (from no heads in three successive coin tosses to all heads). No other value is possible for X.

Continuous Random Variables

Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values. An example of a continuous random variable would be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.

Mathematical Expectations

The mathematical expectation is the events which are either impossible or a certain event in the experiment. Probability of an impossible event is zero, which is possible only if the numerator is 0. Probability of a certain event is 1 which is possible only if the numerator and denominator are equal.

Mathematical expectation value method

In statistics and probability analysis, the expected value is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values.

Bayes Theorem

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

Bayes formula used

The Bayes theorem is a mathematical formula for calculating conditional probability in probability and statistics. In other words, it's used to figure out how likely an event is based on its proximity to another.

One of the most intersting applications of the results of probability theory involves estimating unknown probability and making decisions on the basis of new (sample) information. Since World War II, a considerable body of knowledge has developed known as *Bayesian decision theory* whose purpose is the solution of problems involving decision-making under uncertainty.

The concept of conditional probability discussed above takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to "revise" probabilities based on new information and to determine the probability that a particular effect was due to a specific cause. The procedure for revising these probabilities is known as *Bayes' theorem*.

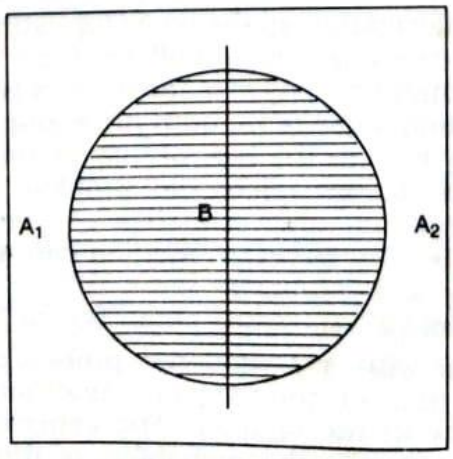
The Bayes' theorem named after the British mathematician Rev. Thomas Bayes (1702-61) and published in 1763 in a short paper has become one of the most famous memoirs in the history of science and one of the most controversial. His contribution consists primarily of a unique method for calculating conditional probabillities. The so-called "Bayesian" approach to this problem addresses itself to the question of determining the probability of some event, A, given that another event, B, has been (or will be) observed, *i.e.*, determining the value of $P(A/B)$. The event A is usually thought of as sample information so that Bayes' rule is concerned with determining the probability of an event given certain sample information. For example, a sample output of 2 defectives in 50 trials (event A) might be used to estimate the probability that a machine is not working correctly (event B) or you might use the results of your first examination in statistics (event A) as sample evidence in estimating the probability of getting a first class (event B).

Bayes' theorem is based on the formula for conditional probability explained earlier. Let :

A_1 and A_2 =The set of events which are mutually exclusive (the two events cannot occur together) and exhaustive (the combination of the two events is the entire experiment ; and

$B = A$ simple event which intersects each of the A events as shown in the diagram below :

Observe the above diagram. The part of B which is within A_1 represents the area "A_1 and B" and the part of B within A_2 represents the area "A_2 and B ".



Then the probability of event A_1 given event B is

$$P(A_1/B) = \frac{P(A_1 \text{ and } B)}{P(B)}$$

and, similarly the probability of event A_2, given B, is

$$P(A_2/B) = \frac{P(A_2 \text{ and } B)}{P(B)}$$

where $P(B) = P(A_1 \text{ and } B) + P(A_2 \text{ and } B)$,

$P(A_1 \text{ and } B) = P(A_1) \times P(B/A_1)$, and

$P(A_2 \text{ and } B) = P(A_2) \times P(B/A_2)$

In general, let $A_1, A_2, A_3, \ldots\ldots\ldots\ldots A_i, \ldots\ldots\ldots, A_n$ be a set of n mutually exclusive and collectively exhaustive events. If B is another event such that $P(B)$ is not zero, then

$$P(A_1/B) = \frac{P(B/A_1)\, P(A_1)}{\sum_{i=1}^{k} P(B/A_1)\, P(A_1)}$$

Example of Bayesian theory

For example, if a disease is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have the disease, compared to the assessment of the probability of disease made without knowledge of the person's age.

Theoretical Distributions

In other words, theoretical distribution is a statistical distribution received by a set of logical and mathematical reasoning from given principles or assumptions. Theoretical distribution is the opposite of distribution derived by real-world data derived by empirical research.

Binomial Distribution

The binomial distribution also known as ‘Bernoulli Distribution’ is associated with the name of a Swiss mathematician James Bernoulli also known as Jasques or Jakob (1654-1705). Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure.

This distribution has been used to describe a wide variety of processes in business and the social sciences as well as other areas. The type of process which give rise to this distribution is usually referred to as Bernoulli trail or as a Bernoulli process.

Properties of the Binomial Distribution

1. The shape and location of binomial distribution changes as P changes for a given n or as n changes for a given p. As p increases for a fixed n, the binomial distribution shifts to the right.

2. The mode of the binomial distribution is equal to the value of x which has the largest probability.

Importance of the Binomial distribution

The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life events. For example, a quality control inspector wants to know the probability of defective light bulbs in a random sample of 10 bulbs if 10 per cent of the bulbs are defective. He can quickly obtain the answer from tables of the binomial probability distribution. The binomial distribution can be used when:

1. The outcome or results of each trial in the process are characterized as one of two types of possible outcomes. In other words, they are attributes.

2. The possibility of outcome of any trial does not change and is independent of the results of previous trials.

The Binomial Distribution

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

Where

$P = \text{Pr}\,obability\ of\ sucess\ in\ a\ \text{single trail}$

$q = 1 - p$

$n = Number\ of\ trails$

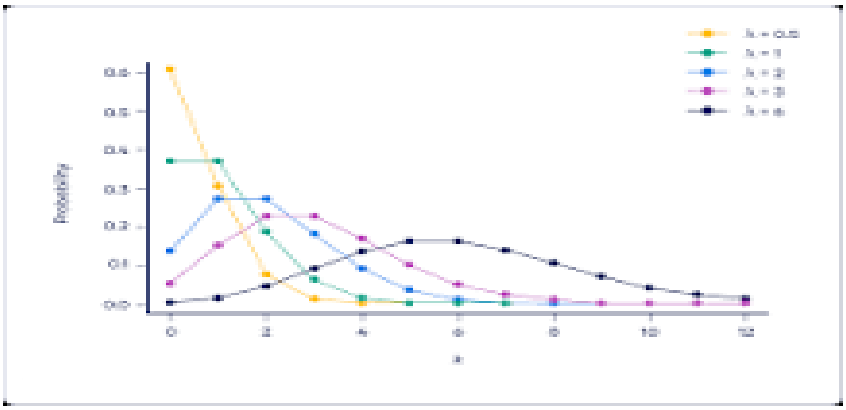
$r = Number\ of\ successes\ in\ n\ trails.$

Poisson Distribution:

Poisson distribution is a discrete probability distribution and is very widely used in statistical work. It was developed by a French mathematician, Simeon Denis Poisson (1781-1840). Poisson distribution may be expected in cases where the chance of any individual event being a success is small.

Unlike a normal distribution, which is always symmetric, the basic shape of a Poisson distribution changes. For example, a Poisson distribution with a low mean is highly skewed, with 0 as the mode. All the data are "pushed" up against 0, with a tail extending to the right.

Poisson distribution



A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space. The Poisson distribution has only one parameter, λ (lambda), which is the mean number of events.

$$P(r) = \frac{e^{-m} m^{r}}{r!}$$

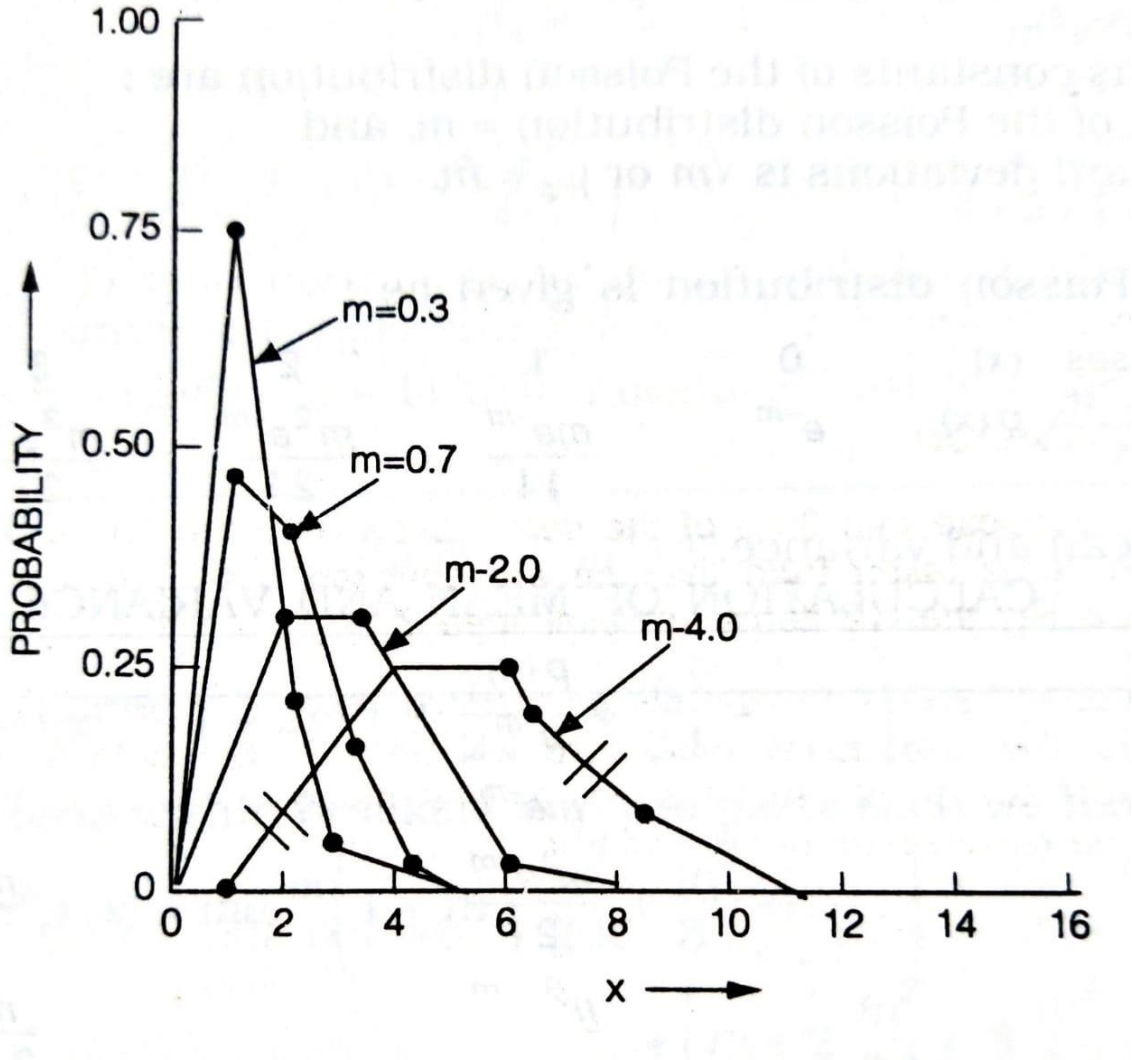
where

$r = 0, 1, 2, 3, 4....$

$e = 2.7183 (\textit{the base of natural Logarithms})$

$m = \textit{the mean of the Poisson distribution}$

The Poisson distribution is a discrete distribution with a single parameter m. As m increases, the distribution shifts to the right. This is explained from the below diagram.



All Poisson probability distributions are skewed to the right. This is the reason why the Poisson probability distribution has been called the probability distribution of area events.

Role of Poisson Distribution:

1. It is used in quality control statistics to count the number of defects of an item.
2. In biology to count the number of bacteria.
3. In physics to count the number of particles emitted from a radioactive substance.
4. In insurance problems to count the number of causalities.

5. In waiting-time problems to count the number incoming telephone calls or incoming customers.
6. Number of traffic arrivals such as trucks at terminals, aeroplanes at airports, slips at docks, and so forth.
7. In determining the number of deaths in a district in a given period, say, a year, by a rare disease,
8. The number of typographical errors per page in typed material, the number of deaths as a result of road accidents, etc.,
9. In problems dealing with the inspection of manufactured products with the probability that any one piece is defective is very small and the lost are very large
10. To model the distribution of the number of persons joining a queue to receive a service or purchase of a product.

Normal Distribution:

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The binominal and the Poisson distribution described above are the most useful theoretical distributions for discrete variables, i.e., they relate to the occurrence of distinct events. In order to have mathematical distribution suitable for dealing with quantities whose magnitude is continuously variable, a continuous distribution in needed. The normal distribution, also called the normal probability distribution, happens to be most useful theoretical distribution for continuous variables. Many statistical data concerning business and economic problems are displayed in the form of normal distribution. In fact normal distribution is the cornerstone of modern statistics.

Importance of the Normal Distribution

The normal distribution has been long occupied a central place in the theory of statistics. Its importance will be clear from the following points.

1. The normal distribution has the remarkable property stated in the socalled central limit theorem. According to this theorem as the sample size n increases the distribution of mean, $\overline{X}$ of a random sample taken from practically any population approaches a normal distribution.
2. As n becomes large the normal distribution serves as a good approximation of many discrete distributions whenever the exact discrete probability is laborious to obtain or impossible to calculate accurately.
3. In theoretical statistic many problems can be solved only under the assumption of a normal population
4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. The normal distribution is used extensively in statistical quality control in industry in setting up of control limits.

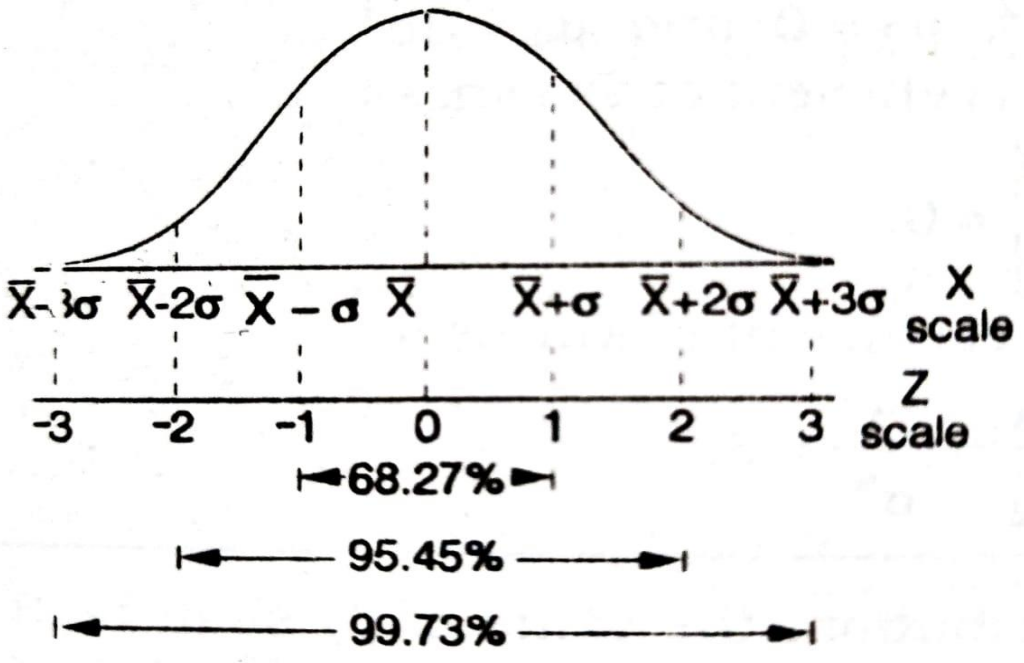
Properties of Normal Distribution:

The following are the important properties.

1. The normal curve is bell-shaped curve and symmetrical in its appearance. If the curves were folded along its vertical axis, the two halves would coincide.
2. The height of the normal curve is at its maximum at the mean. Hence, the mean and mode of the normal distribution coincide. Thus mean, median and mode are equal.
3. There is one maximum point of the normal curve which occurs at the mean. The height of the curve declines as we go in either direction from the mean.

4. The curve approaches nearer and nearer to the base but it never touches it, i.e., the curve is asymptotic to the base on either side. Hence its range is unlimited or infinite in both directions.
5. Since there is only one maximum point, the normal curve is unimodal, i.e., it has only one mode.
6. The points of inflexion, I, e., the points where the change in curvature occurs are $\bar{X} \pm \sigma$.
7. As distinguished from Binomial and Poisson distribution where the variable is discrete, the variable distributed according to the normal curve is a continuous one.
8. The first and third quartiles are equidistant from the median.
9. The mean deviation is 4^{th} or more precisely 0.7979 of the standard deviation.
10. The area under the normal curve distributed as follows:

$a)\, Mean \pm 1\sigma \,\text{cov}\, ers\, 68.27\%\, area; 34.135\%\, area\, will\, lie\, on\, either\, side\, of\, the\, mean$
$b)\, Mean \pm 2\sigma \,\text{cov}\, ers\, 95.45\%\, area$
$c)\, Mean \pm 3\sigma \,\text{cov}\, ers\, 99.73\%\, area$



Significance of the Normal Distribution

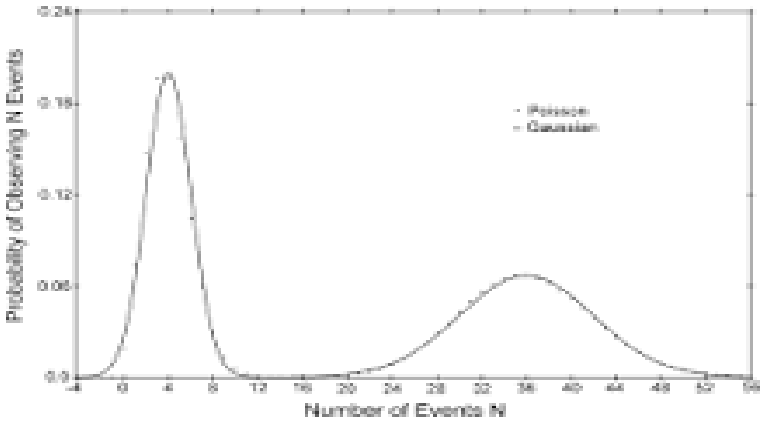
Normal distribution is mostly used for the following purposes.

1. To approximate of fit a distribution of measurement under certain conditions.
2. To approximate the binomial distribution and other discrete of continuous probability distribution under suitable condition.
3. To approximate the distribution of means and certain other quantities calculated from samples, especially large samples.

Difference between Poisson process and normal distribution

A Poisson distribution model helps find the probability of a given number of events in a time period, or the probability of waiting time until the next event in a Poisson continuous process (where certain events occur randomly and independently but at a rate).

Difference between Poisson and Gaussian curve



The Poisson function is defined only for a discrete number of events, and there is zero probability for observing less than zero events. The Gaussian function is continuous and thus takes on all values, including values less than zero as shown for the $\mu = 4$ case.

Check Your Progress:

Q.No	Short Questions	LOCF Mapping		
1.	Discuss the origin, basic concepts, and statistical importance of Probability theory.	K2	CO1	PO2
2.	Difference between Poisson process and Normal distribution.	K3	CO2	PO3
3.	Write about binomial distribution.	K4	CO1	PO4
4.	Explain the Bayes theorem.	K3	CO3	PO3
5.	Define the Mathematical Expectation.	K4	CO2	PO4
Q.No	**Essay type Questions**	**LOCF Mapping**		
1.	Elucidate the properties of the Normal Distribution and explain its fundamental importance as a tool for economic analysis and inferential statistics.	K2 K3	CO1 CO2	PO2 PO3
2.	Narrate the types of Probability			
3.	Distinguish between Binomial Distribution and Poisson Distribution.	K1	CO3	PO1
4.	Explain the Conditions and Importance of Poisson distribution.	K4	CO3	PO2
5.	Distinguish between Discrete and continuous Random variables.	K5	CO4	PO4

UNIT II

Sampling Theory – Types of sampling – Sampling Distributions – Parameter and Statistic – Testing of Hypothesis – Level of Significance – Types I and Type II Errors – Standard Error - Properties of Estimator.

Sampling Theory

The best way to represent a population is to enumerate its members before selecting a random sample from that population. When properly implemented, this guarantees that the sample will formally represent the population within known limits of sampling error.

Probability Sampling Types

Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

Importance in sampling theory

The idea behind importance sampling is that certain values of the input random variables in a simulation have more impact on the parameter being estimated than others. If these "important" values are emphasized by sampling more frequently, then the estimator variance can be reduced.

Sampling Distribution

A sampling distribution isa probability distribution of a statistic that is obtained through repeated sampling of a specific population. It describes a range of possible outcomes for a statistic, such as the mean or mode of some variable, of a population.

Types of Sampling Distributions

Here is a brief description of the types of sampling distributions:

- **Sampling Distribution of the Mean:** This method shows a normal distribution where the middle is the mean of the sampling distribution. As such, it represents

the mean of the overall population. In order to get to this point, the researcher must figure out the mean of each sample group and map out the individual data.

- **Sampling Distribution of Proportion:** This method involves choosing a sample set from the overall population to get the proportion of the sample. The mean of the proportions ends up becoming the proportions of the larger group.
- **T-Distribution:** This type of sampling distribution is common in cases of small sample sizes. It may also be used when there is very little information about the entire population. T-distributions are used to make estimates about the mean and other statistical points.

Parameter and Statistic

A parameter is a number describing a whole population (e.g., population mean), while a statistic is a number describing a sample (e.g., sample mean).

The difference between parameter and statistic

The key difference between parameters and statistics is that parameters describe populations, while statistics describe samples. You can easily remember this distinction using the alliterations for population, parameter, and sample statistic.

Points	Statistic	Parameter
1	Derived from sample data	Derived from population data
2	Used to estimate population characteristics	Represents population characteristics
3	Subject to sampling variability	Fixed value
4	Provides information about a sample	Provides information about a population
5	Varied values across different samples	Consistent value for the entire population
6	Estimation based on inference techniques	Known or can be determined with complete data
7	Used to draw conclusions about a population	Describes a population
8	Often denoted using Greek letters	Often denoted using English letters
9	Can change with different samples	Remains constant for a specific population
10	Used in hypothesis testing and confidence intervals	Used in defining populations and subgroups

Testing of Hypothesis

Hypothesis testing isa systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

There are three types of hypothesis tests: right-tailed, left-tailed, and two-tailed. When the null and alternative hypotheses are stated, it is observed that the null hypothesis is a neutral statement against which the alternative hypothesis is tested.

Procedure of Testing Hypothesis

The procedure of testing hypothesis is as follows:

1. Set up a hypothesis

The null hypothesis can be thought of as the opposite of the "guess" the researchers made: in this example, the biologist thinks the plant height will be different for the fertilizers. So the null would be that there will be no difference among the groups of plants. Specifically, in more statistical language the null for an ANOVA is that the means are the same.

2. Set up a suitable significance level

The significance level is typically set equal to such values as 0.10, 0.05, and 0.01. The 5 percent level of significance, that is, $\alpha = 0.05$, has become the most common in practice. Since the significance level is set to equal some small value, there is only a small chance of rejecting H_0 when it is true.

3. Setting a test criterion

This involves selecting an appropriate probability distribution for the particular test, that is, a probability distribution which can properly be applied.

4. Doing Computations

A computation is any type of arithmetic or non-arithmetic calculation that is well-defined. Common examples of computations are mathematical equations.

5. Making Decisions

Finally as a fifth step, we may conclude statistical conclusions and take decisions. A statistical conclusions or statistical decision is a decision either to reject or to accept the null hypothesis.

The steps of testing hypothesis

Table of contents

- Step 1: State your null and alternate hypothesis.
- Step 2: Collect data.
- Step 3: Perform a statistical test.
- Step 4: Decide whether to reject or fail to reject your null hypothesis.
- Step 5: Present your findings.

The advantages of hypothesis

A hypothesis can help you to formulate a specific and testable research problem, and to design an appropriate method to collect and analyze data. A hypothesis can also help you to establish a clear direction and focus for your research, and to communicate your expectations and assumptions to your readers or audience.

Hypothesis test to use

A z-test is used to test a Null Hypothesis if the population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population variance is unknown.

Characteristics of the hypothesis:

- The hypothesis should be clear and precise to consider it to be reliable.
- If the hypothesis is a relational hypothesis, then it should be stating the relationship between variables.
- The hypothesis must be specific and should have scope for conducting more test.

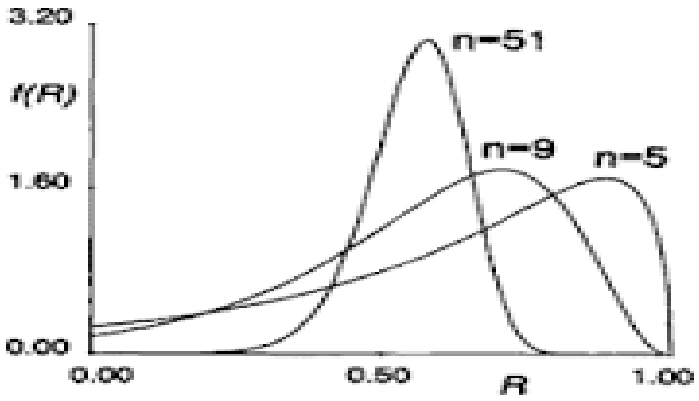
Level of Significance

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true. The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error.

The 4 levels of significance

The null hypothesis is a hypothesis that states that the data has no result, association between variables, or disparity between variables. There are four levels in statistics that are organized by level of complexity and precision. They are nominal, ordinal, interval, and ratio.

5 significance level



The significance level is typically set equal to such values as 0.10, 0.05, and 0.01. The 5 percent level of significance, that is, $\alpha = 0.05$, has become the most common in practice. Since the significance level is set to equal some small value, there is only a small chance of rejecting H_0 when it is true.

Two Types of Errors in testing of Hypothesis

1. The hypothesis is true but our test rejects it (Type I error).
2. The hypothesis is false but our test accepts it (Type II error).
3. The hypothesis is true and our test accepts it (Correct Decision).
4. The hypothesis is false and our test rejects it (Correct Decision).

Type I and Type II Errors

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

Table of Type I and Type II Error

The relationship between truth or false of the null hypothesis and outcomes or result of the test is given in the tabular form:

Error Types	When H_0 is True	When H_0 is False
Don't Reject	Correct Decision (True negative) Probability = $1 - \alpha$	Type II Error (False negative) Probability = β
Reject	Type II Error (False Positive) Probability = α	Correct Decision (True Positive) Probability = $1 - \beta$

Type I and Type II Errors Example

Check out some real-life examples to understand the type-I and type-II error in the null hypothesis.

Example 1: Let us consider a null hypothesis – A man is not guilty of a crime.

Then in this case:

Type I error (False Positive)	Type II error (False Negative)
He is condemned to crime, though he is not guilty or committed the crime.	He is condemned not guilty when the court actually does commit the crime by letting the guilty one go free.

Example 2: Null hypothesis- A patient's signs after treatment A, are the same from a placebox.

Type I error (False Positive)	Type II error (False Negative)
Treatment A is more efficient than the placebo	Treatment A is more powerful than placebo even though it truly is more efficient.

STANDARD ERROR

Standard error is the approximate standard deviation of a statistical sample population. The standard error describes the variation between the calculated mean of the population and one which is considered known, or accepted as accurate.

An example of using standard error

Player Number	Height (in)	mean-measurement	
1	75	-3	9
2	70	2	4
3	69	3	9
4	68	4	16
5	68	4	16
6	72	0	0
7	72	0	0
8	73	-1	1
9	73	-1	1
10	74	-2	4
11	74	-2	4
12	73	-1	1
13	75	-3	9

Add this column

sum of (mean-measurement)^2 = 74

For example, you would construct a 95% confidence interval by adding and subtracting 1.96 times the standard error from the sample mean. Therefore, the 95% confidence interval for high school basketball player height would be 70.65 inches to 73.35 inches.

Standard Error

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ ← Standard deviation

n ← Number of samples

Standard error is calculated by dividing the standard deviation of the sample by the square root of the sample size. Calculate the mean of the total population. Calculate each measurement's deviation from the mean.

The symbol for standard error

The standard error of a statistic is usually designated by the Greek letter sigma (σ) with a subscript indicating the statistic. For instance, the standard error of the mean is indicated by the symbol: σ_M.

Properties of Good Estimator

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

An example of estimate

We need to estimate how much paint we'll need for the job. The cost of the project has been estimated at/as about 10 million dollars. He estimates that current oil reserves are 20 percent lower than they were a year ago. Damage from the hurricane is estimated (to be) in the billions of dollars.

Check Your Progress:

Q.No	Short Questions	LOCF Mapping		
1.	Difference between parameter and statistic.	K2	CO1	PO2
2.	Explain the Level of significance.	K3	CO2	PO3
3.	Write the steps of testing hypothesis.	K4	CO1	PO4
4.	Give a detail note on various types of samplings.	K3	CO3	PO3
5.	Define Standard Error and explain its role.	K4	CO2	PO4
Q.No	**Essay type Questions**	**LOCF Mapping**		
1.	Distinguish between Type I and Type II error.	K2	CO1	PO2
2.	Briefly discuss the Properties of a Good Estimator.	K3	CO2	PO3
3.	Describe the merits and demerits of Sampling.	K1	CO3	PO1
4.	Explain the relationship between the Sampling Distribution and Standard error.	K4	CO3	PO2
5.	Examine the procedure of testing hypothesis.	K5	CO4	PO4

UNIT III

Difference between Large and Small Samples – Test of Significance for Large Samples – Test for Two Means and Standard Deviations – Proportion and Confidence Interval – Small Sample Test – t-test – Paired t-test – Chi-square Test – Test of Goodness of Fit.

Difference between Large and small samples

When the sample size is under 30, statisticians are supposed to use the Student T distribution instead. It has a much greater chance of being wrong. In statistical context, a sample is considered to be large if it is at least 30. On the other hand, a sample is considered small is it is less than 30.

The difference between a small sample size and a large sample size lies in the number of observations or data points included in each sample. Here's a comparison of the characteristics of small and large sample sizes:

Small Sample Size:

1. Limited Representation: A small sample size may not fully represent the population from which it is drawn. It may not capture the full range of variability and characteristics present in the population.

2. Higher Sampling Error: Small samples tend to have higher sampling error or variability. The observed data points may deviate more from the true population values, leading to less precise estimates or conclusions.

3. Reduced Statistical Power: Small samples may have lower statistical power, making it more challenging to detect significant effects or relationships. This can increase the likelihood of Type II errors (failing to detect true effects).

4. Narrow Confidence Intervals: With smaller sample sizes, the confidence intervals around estimates or statistical parameters tend to be wider, reflecting increased uncertainty or imprecision in the results.

5. Limited Generalizability: The findings or conclusions from a small sample may have limited generalizability to a larger population. The relationships or patterns observed in the small sample may not hold true for the broader population.

Large Sample Size:

1. Improved Representation: A large sample size is more likely to capture the characteristics and variability of the population. It provides a better representation of the overall population, leading to more reliable and generalizable results.

2. Lower Sampling Error: Large samples tend to have lower sampling error or variability. The observed data points are closer to the true population values, resulting in more precise estimates and reduced random fluctuations.

3. Higher Statistical Power: Large samples have higher statistical power, enabling a better chance of detecting significant effects or relationships. This reduces the risk of Type II errors.

4. Narrow Confidence Intervals: With larger sample sizes, the confidence intervals around estimates or statistical parameters tend to be narrower, indicating greater precision and reduced uncertainty in the results.

5. Enhanced Generalizability: Findings from a large sample are more likely to be generalizable to the broader population, increasing the confidence in applying the conclusions to a wider context.

In summary, larger sample sizes generally offer more accurate and reliable estimates, higher statistical power, and increased generalizability. However, the appropriate sample size depends on the research question, desired level of accuracy, effect size, variability, and statistical techniques employed. Statisticians employ power analysis and other methods to determine the sample size needed for specific research studies.

Test of Significance for Large Samples

Before moving to large samples, test of significance has to be seen in brief. Test of significance is performed after framing the hypothesis (tentative statements) at say 1%, 5% and 10% level. The level of significance (denoted as α or alpha) represents the probability of error or chances of making wrong decisions.

Statistical test is used for large sample size

If the frequency of success in two treatment groups is to be compared, Fisher's exact test is the correct statistical test, particularly with small samples.

Test for Two Means and Standard Deviations

The 2-Sample Standard Deviation test compares the standard deviations of 2 samples, and the Standard Deviations test compares the standard deviations of more than 2 samples. In this paper, we refer to k-sample designs with k = 2 as 2- sample designs and k-sample designs with k > 2 as multiple-sample designs.

The t-test for means and standard deviation

The t-test is a test used for hypothesis testing in statistics. Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values. T-tests can be dependent or independent.

Used for standard deviation

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

The t-test with means

A t test is used to measure the difference between exactly two means. Its focus is on the same numeric data variable rather than counts or correlations between multiple variables.

Proportion and Confidence of Fit

Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers. The build a confidence interval for population proportion p, we use: $\hat{p}–z\alpha 2\cdot\sqrt{\hat{p}(1−\hat{p})n<p<\hat{p}+z\alpha 2\cdot\sqrt{\hat{p}(1−\hat{p})n}}$.

Therefore, the 99% confidence interval is 0.37 to 0.43. That is, we are 99% confident that the true proportion is in the range 0.37 to 0.43.

Small sample Test

t, and F χ -tests are some commonly used small sample tests. Unit which are based on $\chi 2$ and F-distributions described in Unit 3 and Unit 4 of this course respectively. This unit is divided into eight sections.

Best for small sample size

If the frequency of success in two treatment groups is to be compared, Fisher's exact test is the correct statistical test, particularly with small samples.

Small sample size sampling

The size of the sample is small when compared to the size of the population. When the target population is less than approximately 5000, or if the sample size is a significant proportion of the population size, such as 20% or more, then the standard sampling and statistical analysis techniques need to be changed.

t-test:

A t-test is a statistical test that compares the means of two samples. It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

The three types of t-tests

There are three forms of Student's t-test about which physicians, particularly physician-scientists, need to be aware: (1) one-sample t-test; (2) two-sample t-test; and (3) two-sample paired t-test.

(**1) one-sample t-test**

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

$$t = \frac{(\overline{X_1} - \mu)\sqrt{n}}{S}$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

(**2) Two-sample t-test**

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not.

$$t = \frac{\overline{X_1} - \overline{X_2}}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\sum (X_1 - \overline{X}_1)^2 + \sum (X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}}$$

Example:1

Two types of drugs were used on 5 and 7 patients for reducing their weight.

Drug A was imported and drug B indigenous. The decrease in the weight after using the drugs for six months was as follows.

Drug A	10	12	13	11	14		
Drug B	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs? If no, which drug should you buy. (forv=10, $t_{0.05}$=2.223)

Hypothesis: There is no significant difference in the efficacy of the two drugs. Apply t-test.

X_1	$(X_1-\overline{X})$	$(X_1-\overline{X_1})^2$	X_2	$(X_2-\overline{X}_2)$	$(X_1-\overline{X}_2)^2$
10	-2	4	8	-3	9
12	0	0	9	-2	4
13	1	1	12	1	1
11	-1	1	14	3	9
14	2	4	15	4	16
			10	-1	1
			9	-2	4
$\sum X_1 = 60$		$\sum (X_1-\overline{X_1})^2 = 10$	$\sum X_2 = 77$		$\sum (X_2-\overline{X_2})^2 = 44$

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{60}{5} = 12$$

$$\overline{X_2} = \frac{\sum X_2}{n_2} = \frac{77}{7} = 11$$

$$S = \sqrt{\frac{\sum (X_1-\overline{X}_1)^2 + \sum (X_2-\overline{X}_2)^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{10+44}{5+7-2}} = \sqrt{\frac{54}{10}} = 2.324$$

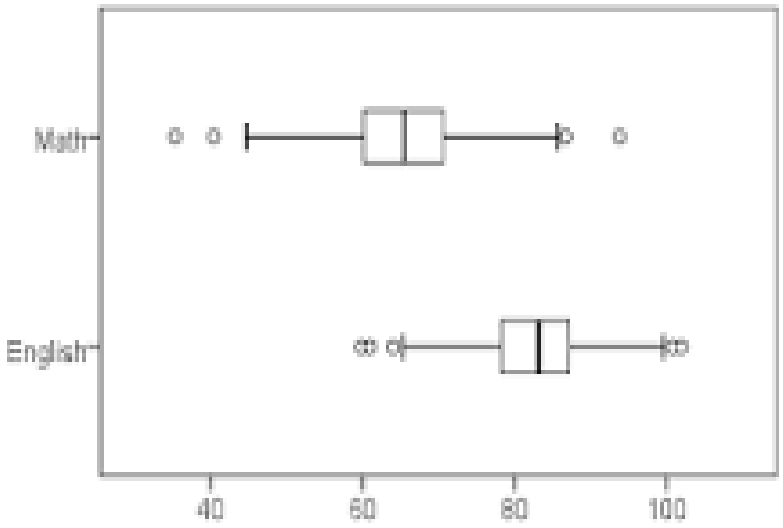
$$t = \frac{\overline{X_1}-\overline{X_2}}{S} \times \sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

$$t = \frac{12-11}{2.324} \times \sqrt{\frac{5\times 7}{5+7}}$$

$$t = \frac{1.708}{2.324} = 0.735$$

The calculated value of t is less than the table value, the hypothesis is accepted. Hence, the hypothesis is accepted. We should buy indigenous drug.

(3) two-sample paired t-test.



The Paired Samples t Test compares the means of two measurements taken from the same individual, object, or related units. These "paired" measurements can represent things like: A measurement taken at two different times (e.g., pre-test and post-test score with an intervention administered between the two time points).

Chi-square test:

The χ^2 test (chi-square) is one of the simplest and most widely used non-parametric test in statistical work. The symbol χ^2 is the Greek letter Chi. The χ^2 test was first used by Karl Pearson in the year 1900.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

For a Chi-square test, a p-value that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. You can conclude that a relationship exists between the categorical variables.

It is defined as $\chi^2 = \frac{\sum (O-E)^2}{E}$

Where O refers the observed frequencies and E refers to the expected frequencies.

To calculate the expected frequencies

$$E = \frac{RT \times CT}{N}$$

E = Expected frequency

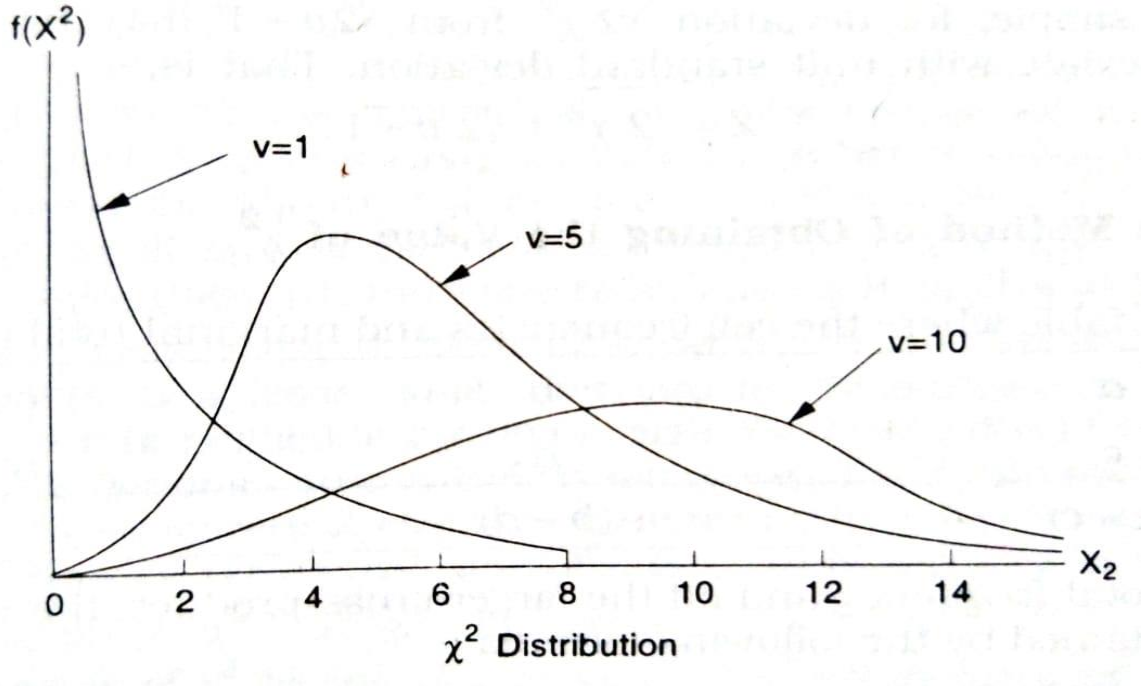
RT = The row for the row containing the cell

CT = the column total for the column containing the cell

N = the total number of observations.

The Chi-Square distribution

The following diagram gives the χ^2 distribution for 1.54 and 10 degree of freedom.



Uses of χ^2 test

1. χ^2 test as a test of independence
2. χ^2 test as a goodness of fit
3. χ^2 test as a test of homogeneity

Example-1

IN an anti-dengue campaign in a certain area, quinine was administered to 812 persons out of total population3248. The number of fever cases as follows:

Treatment	Fever	No Fever	Total
Quinine	**20**	**792**	**812**
No Quinine	**220**	**2216**	**2436**
Total	**240**	**3008**	**3248**

Check the usefulness of quinine in checking dengue.

Hypothesis: There is not effective in checking dengue. Apply χ^2 test.

Solution

Expectation of (AB) $= \frac{A \times B}{N} = \frac{240 \times 812}{3248} = 60$

Likewise find all the expected frequencies.

60	**752**	**812**
180	**2258**	**2436**
240	**3008**	**3248**

O	E	$(O-E)^2$	$(O-E)^2 / E$
20	**60**	**1600**	**26.667**
220	**180**	**1600**	**8.889**
792	**752**	**1600**	**2.128**
2216	**2256**	**1600**	**0.709**
$\sum (O-E)^2 / E$			**38.393**

$$\chi^2 = \sum (O-E)^2 / E$$
$$= 38.393$$

The calculated value of χ^2 is greater than the table value. The hypothesis is rejected. Hence the quinine is useful in checking dengue.

Test and Goodness of Fit

In the previous chapter various tests of significance such as t, F and Z were discussed. These tests were based on the assumption that the samples were drawn from normally distributed populations, or more accurately that the sample means were normally distributed. Since the testing procedure request assumption about the type of the population or parameters.

Though non-parametric theory developed as early as the middle of the nineteenth

Century, it was only after 1945 that non-parametric test came to be used widely. Originated in sociological and psychological research, non-parametric tests today are very popular in behavioural sciences. The following three reasons account for the increasing use of non-parametric tests in business research:

(1) These statistical tests are distribution-free (can be used with any shape of population distribution)

(2) They are usually computationally easier to handle and understand than parametric tests;

(3) They can be used with types of measurements that prohibit the use of parametric tests.

The increasing popularity of non-parametric tests should not lead the reader to form an impression that they are usually superior to the parametric tests. In fact, in a situation where parametric and non-parametric tests both apply, the former are more desirable than the latter.

Check Your Progress:

Q.No	Short Questions	LOCF Mapping		
1.	Difference between Large and Small Samples	K1	CO1	PO1
2.	Explain the Large Sample Size.	K2	CO2	PO2
3.	Explain the Test and Goodness of Fit	K2	CO1	PO2
4.	What are the types of T- test.	K4	CO2	PO4
5.	Two types of drugs were used on 5 and 7 patients for reducing their weight. Drug A was imported and drug B indigenous. The decrease in the weight after using the drugs for six months was as follows. Drug A- 10 12 13 11 14 Drug B- 8 9 12 14 15 10 9 Is there a significant difference in the efficacy of the two drugs? If no, which drug should you buy. (forv = 10, t0.05=2.223)	K2	CO4	PO2

Q.No	Essay type Questions	LOCF Mapping		
1.	Explain the test of significance for large sample.	K3	CO5	PO3
2.	Briefly Explain T test.	K4	CO2	PO4
3.	Discuss the Chi-square test and explain the x^2.	K2	CO1	PO2
4.	For a random sample of 10 persons, fed on diet A, the increased weight in pounds in a certain period were: 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 For another random sample of 12 persons, fed on diet B, the increase in the same period were: 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 Test whether the diets, A and B significantly as regards their effect on increase in weight. Given the following: Test whether the diets, A and B differ significantly as regards their effect on increase in weight. Given the following: Degrees of Freedom: 19 | 20 | 21 | 22 | 23 Value of t at 5% level: 2.09 | 2.09 | 2.08 | 2.07 | 2.07	K4	CO1	PO4
5.	1,000 students at college level are graded according to their I.Q. and their economic conditions. Use chi-square test to find out whether there is any association between economic conditions and the level of I.Q. Economic Condition / IQ: High, Medium, Low, Total Rich: 160, 300, 140, 600 Poor: 140, 100, 160, 400 Total: 300, 400, 300, 1000 (Given for degree of freedom 2, chi-square 5 per cent = 5.99)	K5	CO3	PO4

UNIT IV

F test: Assumption in F test – Analysis of Variance- Assumptions – One-Way and Two-Way Classifications.

Introduction of F Test Formula

The F Test Formula is a Statistical Formula used to test the significance of differences between two groups of Data. It is often used in research studies to determine whether the difference in the means of two populations is statistically significant. It is based on the F Statistic, which is a measure of how much variation exists in one group of Data compared to another. Students who are studying for their Statistics course will need to be familiar with this Formula. Our article will provide a detailed explanation of how to use the F Test Formula. It will also provide examples of how to use it in practice.The use of the F Test Formula is a critical step in any research study, and it is important to understand how to use it correctly. You will be able to find the F Test Formula in most Statistics textbooks.

Definition of F-Test Statistic Formula

It is a known fact that Statistics is a branch of Mathematics that deals with the collection, classification and representation of Data. The tests that use F - distribution are represented by a single word in Statistics called the F Test. F Test is usually used as a generalized Statement for comparing two variances. F Test Statistic Formula is used in various other tests such as regression analysis, the chow test and Scheffe test. F Tests can be conducted by using several technological aids. However, the manual calculation is a little complex and time-consuming.

F Test is a test Statistic that has an F distribution under the null hypothesis. It is used in comparing the Statistical model with respect to the available Data set. The name for the test is given in honour of Sir. Ronald A Fisher by George W Snedecor. To perform an F Test using technology, the following aspects are to be taken care of.

- State the null hypothesis along with the alternative hypothesis.
- Compute the value of 'F' with the help of the standard Formula.
- Determine the value of the F Statistic.
- The ratio of the variance of the group of means to the mean of the within-group variances.
- As the last step, support or reject the Null hypothesis.

Assumptions in F test

- The populations are characterized as having a normal distribution.
- The populations are independent from one another.
- When calculating the F-statistic, the larger variance is used as the numerator, and the smaller variance is used in the denominator.

Assumptions of the F-test

The F-test is based on two assumptions: (1) the samples are normally distributed, and (2) the samples are independent of each other. If these assumptions are fulfilled and H_0 is true, the statistic F follows an F-distribution.

F-Test Equation to Compare Two Variances:

In Statistics, the F-test Formula is used to compare two variances, say σ1 and σ2, by dividing them. As the variances are always positive, the result will also always be positive. Hence, the F Test equation used to compare two variances is given as:

$$F = \frac{S_1^{\,2}}{S_2^{\,2}}\ , \qquad S_1^{\,2} = \frac{(X_1 - \overline{X_1})^2}{n_1 - 1}\ ; \qquad S_2^{\,2} = \frac{(X_2 - \overline{X_2})^2}{n_2 - 1}$$

It should note that $S_1^{\,2}$ is always the larger estimate of variance.

$$F = \frac{Larger\ estimate\ of\ \text{variance}}{Smaller\ estimate\ of\ \text{variance}}$$

F Test Formula helps us to compare the variances of two different sets of values. To use F distribution under the null hypothesis, it is important to determine the mean of the two given observations at first and then calculate the variance.

In the above formula, σ2 is the variancex is the values given in a set of datax is the mean of the given Data set n is the total number of values in the Data setWhile running an F Test, it is very important to note that the population variances are equal. In more simple words, it is always assumed that the variances are equal to unity or 1. Therefore, the variances are always equal in the case of the null hypothesis.

The conditions to use F-test

In order to use an ANOVA F-Test, each group must be normally distributed, the groups must have the "same" variance, and the samples must be randomly selected in an independent manner.

Analysis of Variance: Assumptions

There are three primary assumptions in ANOVA: The responses for each factor level have a normal population distribution. These distributions have the same variance. The data are independent.

1. Normality
2. Homogeneity
3. Independence of error

One-way and two-way Classification

The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two. One-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

What Does "One-Way" or "Two-Way Mean?

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test.

- One-way has one independent variable (with 2 levels). For example: *brand of cereal*,
- Two-way has two independent variables (it can have multiple levels). For example: *brand of cereal*, *calories*.

Types of Tests.

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

- One-way ANOVA between groups: used when you want to test two groups to see if there's a difference between them.

- Two way ANOVA without replication: used when you have one group and you're double-testing that same group. For example, you're testing one set of individuals before and after they take a medication to see if it works or not.
- Two way ANOVA with replication: Two groups, and the members of those groups are doing more than one thing. For example, two groups of patients from different hospitals trying two different therapies.

One Way ANOVA

A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution. The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are unequal.

Examples of when to use a one way ANOVA

Situation 1: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea. **Situation 2:** Similar to situation 1, but in this case the individuals are split into groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

Limitations of the One Way ANOVA

A one way ANOVA will tell you that at least two groups were different from each other. But it won't tell you which groups were different. If your test returns a significant f-statistic, you may need to run an ad hoc test (like the Least Significant Difference test) to tell you exactly which groups had a difference in means.

Two Way ANOVA

A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable. With a Two Way ANOVA, there are two independents. Use a two way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two way ANOVA is appropriate.

For example, you might want to find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called **factors** in Two Way ANOVA.

The factors can be split into **levels**. In the above example, income level could be split into three levels: low, middle and high income. Gender could be split into three levels: male, female, and transgender. Treatment groups are all possible combinations of the factors. In this example there would be 3 x 3 = 9 treatment groups.

Assumptions for Two Way ANOVA

1. The population must be close to a normal distribution.
2. Samples must be independent.
3. Population variances must be equal (i.e. homoscedastic).
4. Groups must have equal sample sizes.

An example of a two way classification

For example, one way classifications might be: gender, political party, religion, or race. Two way classifications might be by gender and political party, gender and race, or religion and race. Each classification variable is a called a factor and so there are two factors, each having several levels within that factor.

Example-1

Apply F-test. Two random samples were drawn from the two populations and their values are:

A	66	67	75	76	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance. ($F_{0.05}$= 3.36)

Hypothesis: The two populations have the same variance.

Calculation of F-test:

A	$x_1 = (X_1 - \overline{X_1})$	X_1^2	B	$(X_2 - \overline{X_2})$	X_2^2
66	-14	196	64	-19	361
67	-13	169	66	-17	289
75	-5	25	74	-9	81
76	-4	16	78	-5	25
82	2	4	82	-1	1
84	4	16	85	2	4
88	8	64	87	4	16
90	10	100	92	9	81
92	12	144	93	10	100
			95	12	144
			97	14	196
$\sum X_1 = 720$	$\sum x_1 = 0$	$\sum x_1^2 = 734$	$\sum X_2 = 913$	$\sum x_2 = 0$	$\sum x_2^2 = 1298$

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{720}{9} = 80 \qquad \overline{X_2} = \frac{\sum X_2}{n_2} = \frac{913}{11} = 83$$

$$F = \frac{S_1^2}{S_2^2}$$

$$S_1^2 = \frac{(X_1 - \overline{X_1})^2}{n_1 - 1}, \qquad S_2^2 = \frac{(X_2 - \overline{X_2})^2}{n_2 - 1}$$

$$S_1^2 = \frac{734}{9-1} = 91.75 \qquad S_2^2 = \frac{1298}{11-1} = 129.8$$

$$F = \frac{S_1^2}{S_2^2} = \frac{129.8}{91.75}$$

$$F = 1.415$$

The calculated value of F is greater than the table value. The hypothesis is accepted.

Hence, that the two populations have the same variance.

Check Your Progress:

Q.No	Short Questions	LOCF Mapping		
1.	What are the Assumption in F test.	K1	CO1	PO1
2.	What is analysis of variance? Outline the significance of it.	K2	CO2	PO1
3.	Explain the Assumptions for Two Way ANOVA.	K2	CO3	PO2
4.	Define F- test with its statistic formula.	K5	CO4	PO4
5.	Write the Limitations of the One Way ANOVA.	K2	CO5	PO2

Q.No	Essay Type Questions	LOCF Mapping		
1.	Explain the concept of ANOVA	K4	CO3	PO3
2.	Explain the analysis of variance with suitable technique.	K2	CO2	PO2
3.	Enumerate the types of test.	K6	CO4	PO5
4.	Difference between one-way and two-way classification.	K4	CO4	PO4
5.	The calculated value of F is less than the table value. Two samples are drawn from normal population. From the following data, test whether the two samples have the same variance at 5% level:	K6	CO5	PO5

Sample 1	60	65	71	74	76	82	85	87		
Sample 2	61	66	67	85	78	63	85	86	88	91

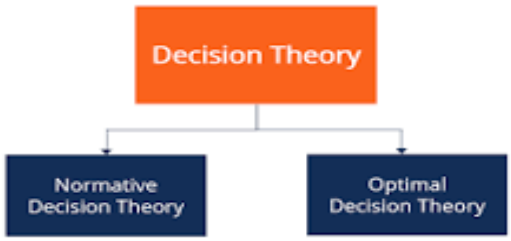
UNIT V

Definitions – Concepts - Maximin – Minimax – Bayes Criterion – Expected Monetary Value – Decision Tree Analysis: Symbols – Steps –Advantages and Limitations.

Statistical Decision Theory

Statistical decision theory is concerned with the making of decisions when in the presence of statistical knowledge (data) which sheds light on some of the uncertainties involved in the decision problem.

The decision theory



Decision theory is the study of a person or agents' choices. The theory helps us understand the logic behind the choices professionals, consumers, or even voters make. The choices come with consequences and are usually discussed in two separate but distinct branches.

The father of decision theory

One of the underlying theories is the 'decision-making theory,' which was first introduced by Herbert A. Simon, the Nobel Prize winner for Economics in 1978. He is best known for his work on corporate decision-making, also called behaviorism.

The three theories of decision-making

The main decision-making theories are Subjective Expected Utility, Heuristics Theory, Attribution Theory, and Prospect Theory. There is a need to research this topic because the more research done on what can influence and facilitate decision-making, the better the organization's growth will be.

(1) Heuristics Theory

Heuristics are mental shortcuts that can facilitate problem-solving and probability judgments. These strategies are generalizations, or rules-of-thumb, that reduce cognitive load. They can be effective for making immediate judgments, however, they often result in irrational or inaccurate conclusions.

(2) Attribution theory

Attribution theory is how we attribute feelings and intentions to people to understand their behaviour. For example, we may unconsciously apply this theory when we see someone shouting on public transport. You may blame their character, assuming they are an angry person.

Prospect theory

Key Takeaways. The prospect theory says that investor's value gains and losses differently, placing more weight on perceived gains versus perceived losses. An investor presented with a choice, both equal, will choose the one presented in terms of potential gains. Prospect theory is also known as the loss-aversion theory...

The concept of statistical decision theory

Statistical decision theory is concerned with the making of decisions when in the presence of statistical knowledge (data) which sheds light on some of the uncertainties involved in the decision problem.

Maximin

A principle of decision theory, that counsels that at least in some circumstance, the right decision is that which maximizes the minimum outcome: i.e., that which makes the worst outcome as good as can be.

The concept of maximin

The maximum of a set of minima. especially : the largest of a set of minimum possible gains each of which occurs in the least advantageous outcome of a strategy followed by a participant in a situation governed by game theory compare minimax.

An example of the maximin principle

His theory was developed to assist a society in ordering its affairs. His ideas have influenced many lawmakers and Supreme Court decisions in the United States. Among many examples are the laws for providing equal access to opportunities for minorities and the disabled.

Bayes Criterion

A Bayesian decision maker proceeds by assigning a numerical utility to each of the possible consequences of an action, and a probability to each of the uncertain events that may affect that utility. The weighted average of the utility with respect to the probability is then used as the criterion of choice.

Decision theory studies the logic and the mathematical properties of decision making under uncertainty. Statistical decision theory focuses on the investigation of decision making when uncertainty can be reduced by information acquired through experimentation. This article reviews the Bayesian approach to statistical decision theory, as was developed from the seminal ideas of Savage. Specifically, it considers: the principle of maximation of expected utility and its axiomatic foundations; the basic elements of Bayesian statistical decision theory, illustrated using standard statistical decision problems; measuring the value of information in decision making and the concept of admissibility.

Decision theory studies the logic and the mathematical properties of decision making under uncertainty, and is applied to a broad spectrum of human activities. Much of it makes use of ideas, approaches, and formalisms that can be categorized as Bayesian.

This discussion focuses on a specific, statistically oriented, definition of decision theory: the investigation of decision making when the state of the world is uncertain, yet further information about it can be obtained through experimentation. We discuss the approach that stems from the theory of Savage (1954), and was developed in depth in the seminal texts of Raiffa and Schleifer (1961) and DeGroot (1970).

At the core of Bayesian decision theory is the principle of maximization of expected utility. A Bayesian decision maker proceeds by assigning a numerical utility to each of the possible consequences of an action, and a probability to each of the uncertain events that may affect that utility. The weighted average of the utility with respect to the probability is then used as the criterion of choice. In statistical decision theory, this plan can be applied to actions that represent inferences, predictions, or conclusions to be drawn from experimental evidence. Decision theory enables statisticians to bring basic principles of rationality to bear on both the development and the evaluation of statistical methodologies.

History of Decision making

Decision making using probability, utility, and the expected utility principle has deep roots. The idea that mathematical expectation should guide rational choice under uncertainty was formulated and discussed as early as the seventeenth century. During that time, an important problem was finding the fixed amount of money that would be fair to exchange for an uncertain payoff, as when paying an insurance premium. Initially, the prevailing thought was that the fair fixed amount would be the expected payoff. The St. Petersburg game (Jorland 1987) revealed an example where the expected payoff is infinite but where, in the words of Bernoulli (1738), ‘no reasonable man would be willing to pay 20 ducats as equivalent.’ Bernoulli's view was that ‘in their theory, mathematicians evaluate money in proportion to its quantity while, in practice,

people with common sense evaluate money in proportion to the utility they can obtain from it.' Using this notion of utility, Bernoulli offered an alternative approach: he suggested that the fair sum to pay for a game of chance is the moral expectation, which effectively is an example of expected utility.

Another important element of Bayesian decision theory that can be traced to the Age of Enlightment is the notion that, when weighing evidence provided by data, one should often take explicitly into account the consequences of the actions that are to be taken using the data. Condorcet (1785), for example, reasoned that 'The probability that a convicted person is guilty should be in proportion to the probability that an acquitted person is innocent, as is the inconvenience of convicting an innocent in proportion to that of acquitting a culprit.' Later, the use of optimality principles in the evaluation of inferential procedures was present in some of the key developments in mathematical statistics, such as estimation efficiency (Fisher 1925), and power of statistical tests (Neyman and Pearson 1933), an approach that explicitly recognized the link between decision making and hypothesis testing. The broader connection between rational decision making and statistical inference was eventually formalized in great generality by Wald (1949), who is considered the founder of statistical decision theory.

The use of probability to quantify a decision maker's knowledge about uncertain events originates from the work of Ramsey (1931) and de Finetti (1937) on subjective probability. The subjectivist approach provides a coherent foundation to probability statements about a broad typology of events, and can be integrated in a natural way into individual decision making. A strength of the subjectivist view in statistical decision theory is in the use of a single measure of uncertainty for both experimental variation and uncertainty about the state of the world. This provides a simple, general, and self-

consistent formal mechanism for modeling the effects of accruing experimental evidence on knowledge.

Justifications for taking the principle of maximization of expected utility as a general guide to rational statistical inference have also been the subject of extensive investigation. A fruitful approach has been to translate this principle into a set of intuitive axioms in terms of preferences among actions that are satisfied if and only if one's behavior is consistent with expected utility. The critical contribution in this regard is that of von Neumann and Morgenstern (1944), who developed the first and fundamental axiom system. Savage (1954) used the subjectivist approach, together with some of the technical advances provided by von Neumann and Morgenstern and Wald, to develop a general and powerful theory of individual decision making under uncertainty, which provides the logical foundation of current Bayesian decision theory. A sketch of what would become Savage's development is outlined by Ramsey (1931).

DecisionTrees

Decisiontreesareusefulinprovidingavisualdisplayofthesequentialdecisionprocesses and subsequent consequences. It can help in organizing the computationalwork.

Decision tree analysis involves visually outlining the potential outcomes, costs, and consequences of a complex decision. These trees are particularly helpful for analyzing quantitative data and making a decision based on numbers. In this article, we'll explain how to use a decision tree to calculate the expected value of each outcome and assess the best course of action. Plus, get an example of what a finished decision tree will look like.

A decision tree is a flowchart that starts with one main idea and then branches out based on the consequences of your decisions. It's called a "decision tree" because the model typically looks like a tree with branches.

These trees are used for decision tree analysis, which involves visually outlining the potential outcomes, costs, and consequences of a complex decision. You can use a

decision tree to calculate the expected value of each outcome based on the decisions and consequences that led to it. Then, by comparing the outcomes to one another, you can quickly assess the best course of action. You can also use a decision tree to solve problems, manage costs, and reveal opportunities.

Decision tree symbols

A decision tree includes the following symbols:

- Alternative branches: Alternative branches are two lines that branch out from one decision on your decision tree. These branches show two outcomes or decisions that stem from the initial decision on your tree.
- Decision nodes: Decision nodes are squares and represent a decision being made on your tree. Every decision tree starts with a decision node.
- Chance nodes: Chance nodes are circles that show multiple possible outcomes.
- End nodes: End nodes are triangles that show a final outcome.

A decision tree analysis combines these symbols with notes explaining your decisions and outcomes, and any relevant values to explain your profits or losses. You can manually draw your decision tree or use a flowchart tool to map out your tree digitally.

Advantages

When you're struggling with a complex decision and juggling a lot of data, decision trees can help you visualize the possible consequences or payoffs associated with each choice.

Transparent:

The best part about decision trees is that they provide a focused approach to decision making for you and your team. When you parse out each decision and calculate their expected value, you'll have a clear idea about which decision makes the most sense for you to move forward with.

Efficient:

Decision trees are efficient because they require little time and few resources to create. Other decision-making tools like surveys, user testing, or prototypes can take months and a lot of money to complete. A decision tree is a simple and efficient way to decide what to do.

Flexible:

If you come up with a new idea once you've created your tree, you can add that decision into the tree with little work. You can also add branches for possible outcomes if you gain information during your analysis.

Limitations

There are drawbacks to a decision tree that make it a less-than-perfect decision-making tool. By understanding these drawbacks, you can use your tree as part of a larger forecasting process.

Complex:

While decision trees often come to definite end points, they can become complex if you add too many decisions to your tree. If your tree branches off in many directions, you may have a hard time keeping the tree under wraps and calculating your expected values. The best way to use a decision tree is to keep it simple so it doesn't cause confusion or lose its benefits. This may mean using other decision-making tools to narrow down your options, then using a decision tree once you only have a few options left.

Unstable:

It's important to keep the values within your decision tree stable so that your equations stay accurate. If you change even a small part of the data, the larger data can fall apart.

Risky:

Because the decision tree uses a probability algorithm, the expected value you calculate is an estimation, not an accurate prediction of each outcome. This means you must take

these estimations with a grain of salt. If you don't sufficiently weigh the probability and payoffs of your outcomes, you could take on a lot of risk with the decision you choose.

Check Your Progress:

Q.No	Short Questions	LOCF Mapping		
1.	Briefly explain the three theories of decision making.	K1	CO1	PO2
2.	Write a note on the maximin and minimax principle with suitable examples.	K6	CO2	PO5
3.	List the primary advantages of using a decision tree for strategic planning and forecasting.	K1	CO3	PO1
4.	What is EMV? Explain its importance.	K2	CO4	PO1
5.	Explain how probability distributions of outcomes are integrated into these decision – making models.	K5	CO2	PO4
Q.No	**Essay Type Questions**	**LOCF Mapping**		
1.	Enumerate the conditions for maxima and minima.	K1	CO4	PO1
2.	Assess Bayesian decision theory and its uses in decision-making under uncertainty.	K3	CO2	PO3
3.	What is meant by decision tree? Investigate the steps in decision tree analysis.	K5	CO5	PO4
4.	Explain the history of decision making.	K5	CO1	PO4
5.	Discuss and compare the Maximin, Minimax, and Bayes Criteria.	K2	CO5	PO2

Text Books

1.Gupta S.P., Statistical Methods, Sultan Chand and Sons, New Delhi, 2017.

2. Anderson, Sweeney and Williams, "Statistics for Business and Economics",Cengage, 2014.

References:

1. Agarwal. Y.P (2002), "Statistics Methods – Concepts Application and Computation", Sterling Publishers Private Ltd., New Delhi.

2. Vittal P.R.,Mathematical Statistics, Margham Publications.

3. Pillai R.S.N. and Bagavathi V (2010), Statistics, Sultan & amp; Chand Sons, New Delhi.

4. Kanmony, Cyril. J, (2022) Statistical Methods (Scitech Publications, Chennai)

Web Resources

1. https://www.statista.com.

2. https://techjury.net

3. https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm

Course Outcomes:

Upon Completion of this course, the Students will be able to

No.	Course Outcomes	K-Levels
CO1	Summarize the basic Probability rules and understand theoretical distributions.	K1 & K2
CO2	Acquire knowledge on the various sampling methods and testing of Hypotheses	K2 & K3
CO3	Use t test and chi square for analysis	K4
CO4	Understand the importance of one and two way ANOVA	K5
CO5	Know the various Decision making tools available	K6

K1 – Knowledge, K2 - Understand, K3 – Apply, K4 – Analyse, K5 – Evaluate, K6 – Create.

CO-PO Mapping (Course Articulation Matrix)

CO /PO	PSO1	PSO2	PSO3	PSO4	PSO5
CO1	3	2	3	2	2
CO2	3	2	3	3	3
CO3	3	3	3	3	3
CO4	3	3	3	3	3
CO5	2	3	3	2	3
Weightage	14	13	15	13	14
Weighted percentage of Course Contribution to Pos	2.8	2.6	3	2.6	2.8

Level of Correlation between PSO's and CO's

(Suggested by UGC as per Six Sigma Tool – Cause and Effect Matrix)

Assign the value

1 – Low, 2 – Medium, 3 – High, 0 – No Correlation

Compiled by

Dr. G. Monikanda Prasad
Assistant Professor of Economics
Manonmaniam Sundaranar University
Tirunelveli – 627 012