DATA ANALYSIS AND MODELLING TECHNIQUE

Assignment 1

Name: Utkarsh Gupta **SAP:** 500075374

Batch: B. Tech CSE spz AI & ML B-6 sem-VIII **Roll No.:** R177219194

Unit 1

1. What is meant by dependent and independent variables? (y is dependent, x is independents)

   In statistical analysis, a dependent variable is a variable that is being studied and measured to see how it is affected by changes in other variables. It is also called the response variable or the outcome variable. The value of the dependent variable depends on the value of the independent variable.

   On the other hand, an independent variable is a variable that is used to explain changes in the dependent variable. It is also called the explanatory variable or the predictor variable. The value of the independent variable does not depend on the value of any other variable in the study.

   In a common notation, the dependent variable is denoted by "y" and the independent variable is denoted by "x". The relationship between the two variables can be expressed as y = f(x), where "f" is a function that describes the relationship between x and y. By manipulating the independent variable, researchers can observe how the dependent variable changes.

2. A card is drawn from a pack of 52 cards and then a second card is drawn. What is the probability that both the cards drawn are queen.

   The probability of drawing a queen on the first draw is 4/52, as there are 4 queens in a pack of 52 cards. After the first queen is drawn, there are 3 queens remaining in a pack of 51 cards. Therefore, the probability of drawing a second queen on the second draw, given that the first draw was a queen, is 3/51.

   To find the probability of drawing two queens in a row, we can multiply the probability of the first event by the probability of the second event, if the first event occurred. So:

   Probability of drawing two queens = Probability of drawing a queen on the first draw × Probability of drawing a queen on the second draw, given that the first draw was a queen

   = (4/52) × (3/51)

   = 0.0045 or 0.45%

   Therefore, the probability of drawing two queens in a row from a pack of 52 cards is 0.45%.

3. What is the probability that a leap year, selected at random, will have 53 Sundays?

   A leap year has 366 days, and each week has 7 days. Therefore, the year contains 52 weeks and 2 days.

   To have 53 Sundays in a year, those two extra days in the leap year must both fall on a Sunday. There are 7 possible days for the first extra day to fall on, but if it falls on a Sunday, then the second extra day can only fall on a Thursday, and vice versa. So there are only 2 possible ways for both extra days to fall on a Sunday and give 53 Sundays in the year.

   Therefore, the probability of selecting a leap year with 53 Sundays is 2 out of the total number of leap years. Leap years occur every 4 years, except for years that are divisible by 100 but not by 400. So, in a span of 400 years, there are 97 leap years. Therefore, the probability of selecting a leap year with 53 Sundays is:

   2/97 ≈ 0.0206 or about 2.06%

4. What is the probability of throwing a number greater than 3 with an ordinary dice?

   An ordinary dice has six sides numbered 1 through 6. To find the probability of throwing a number greater than 3, we need to count the number of sides on the dice that have a number greater than 3, which are 4, 5, and 6.

   Since each side of the dice has an equal chance of landing facing up, the probability of throwing a number greater than 3 is equal to the number of favourable outcomes divided by the total number of possible outcomes.

   The total number of possible outcomes when rolling a dice is 6, since there are 6 sides. The number of favourable outcomes is 3, since there are 3 sides with numbers greater than 3.

   Therefore, the probability of throwing a number greater than 3 with an ordinary dice is:

   Number of favourable outcomes / Total number of possible outcomes = 3/6 = 1/2 = 0.5

   So, the probability of throwing a number greater than 3 with an ordinary dice is 0.5 or 50%.

5. 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 the balls drawn are black.

   The probability of drawing a black ball on the first draw is 3/8, since there are 3 black balls out of a total of 8 balls in the bag. After the first ball is drawn, there are 2 black balls left out of 7 remaining balls in the bag. Therefore, the probability of drawing a second black ball on the second draw, given that the first ball was black, is 2/7.

   To find the probability of drawing two black balls in a row, we can multiply the probability of the first event by the probability of the second event, if the first event occurred. So:

   Probability of drawing two black balls = Probability of drawing a black ball on the first draw × Probability of drawing a black ball on the second draw, given that the first draw was black

= (3/8) × (2/7)

= 6/56

= 3/28

Therefore, the probability of drawing two black balls in a row from a bag containing 5 white and 3 black balls, without replacement, is 3/28.

6. Three athletes A, B and C are participating in the Olympics. A is twice as likely to win as B and B is twice as likely to win as C. What are the probabilities of their winning?

Let us assume that the probability of athlete C winning is x. Then, the probability of athlete B winning is twice as much as the probability of athlete C winning, which is 2x. Similarly, the probability of athlete A winning is twice as much as the probability of athlete B winning, which is 4x.

The sum of the probabilities of all three athletes winning must be 1, since one of them must win. Therefore:

$x + 2x + 4x = 1$

Simplifying the equation, we get:

$7x = 1$

$x = 1/7$

So, the probability of athlete C winning is 1/7.

The probability of athlete B winning is twice as much as the probability of athlete C winning, which is 2/7.

The probability of athlete A winning is twice as much as the probability of athlete B winning, which is 4/7.

Therefore, the probabilities of athlete A, B, and C winning are 4/7, 2/7, and 1/7 respectively.

7. Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result?

   a. 1% of women have cancer.

   b. 90% of women who have cancer test positive on mammograms.

   c. 8% of women will have false positives.

To answer this question, we can use Bayes' theorem, which relates the probability of a hypothesis given the data (posterior probability) to the probability of the data given the hypothesis (likelihood), the prior probability of the hypothesis, and the probability of the data regardless of the hypothesis (normalizing constant). In this case, the hypothesis is whether a woman has cancer, and the data is whether she has a positive mammogram result.

Let us denote the following probabilities:

P(C) = probability that a woman has cancer = 0.01

P(Pos|C) = probability of a positive mammogram given that a woman has cancer = 0.9

P(Pos|¬C) = probability of a positive mammogram given that a woman does not have cancer (false positive rate) = 0.08

We want to find the probability that a woman has cancer given a positive mammogram result, which is denoted as P(C|Pos). Using Bayes' theorem, we have:

P(C|Pos) = P(Pos|C) * P(C) / P(Pos)

where P(Pos) is the probability of a positive mammogram, regardless of whether a woman has cancer or not. This can be calculated using the law of total probability:

P(Pos) = P(Pos|C) * P(C) + P(Pos|¬C) * P(¬C)

where P(¬C) is the probability that a woman does not have cancer, which is 1 - P(C) = 0.99.

Plugging in the values, we get:

P(Pos) = P(Pos|C) * P(C) + P(Pos|¬C) * P(¬C)

= 0.9 * 0.01 + 0.08 * 0.99

= 0.0871

Now, we can calculate the posterior probability using Bayes' theorem:

P(C|Pos) = P(Pos|C) * P(C) / P(Pos)

= 0.9 * 0.01 / 0.0871

= 0.103

Therefore, the probability that a woman has cancer given a positive mammogram result is 0.103, or about 10.3%.

8. Explain the Central Limit theorem and state the merits, demerits, and uses of standard deviation with a basic example.

The Central Limit Theorem (CLT) is a statistical concept that states that the distribution of sample means approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution. In other words, if we take multiple samples of the same size from a population, the distribution of the means of these samples will follow a normal distribution, with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of the sample size. This theorem has several important implications in statistics.

Merits of Standard Deviation:

- It provides a measure of the spread of the data around the mean.

- It is widely used in inferential statistics to estimate the variability of the population parameter based on a sample.
- It can be used to compare the variability of two or more datasets.
- It is used in hypothesis testing to determine whether the difference between two means is statistically significant.

Demerits of Standard Deviation:

- It can be affected by outliers, which can cause it to be an inaccurate measure of spread.
- It does not provide information about the shape of the distribution.

Example of Standard Deviation:

Suppose we want to compare the heights of two groups of students, A and B. We measure the height of 10 students from each group and calculate the standard deviation of each group. We find that the standard deviation of group A is 2 inches, and the standard deviation of group B is 3 inches. This tells us that the heights in group B are more spread out than in group A. We can use this information to compare the variability of the two groups and determine whether the difference in mean height is statistically significant.

Unit 2

1. Explain descriptive statistics with an example and diagrammatic representation.

Descriptive statistics is a branch of statistics that deals with the summary and presentation of data in a meaningful way. It involves the use of numerical and graphical methods to describe the main features of a dataset, such as its central tendency, variability, and distribution.

Example: Suppose we have a dataset of the ages of 10 people: 20, 22, 23, 25, 26, 28, 30, 32, 34, and 35. We can use descriptive statistics to summarize the main features of this dataset.

Measures of central tendency:

- Mean: The mean age of the 10 people is (20+22+23+25+26+28+30+32+34+35)/10 = 28. This gives us a rough idea of the typical age in the dataset.
- Median: The median age is the middle value when the data is arranged in ascending or descending order. In this case, the median age is 27, which is the average of the two middle values (26 and 28).

Measures of variability:

- Range: The range is the difference between the maximum and minimum values in the dataset. In this case, the range is 35-20 = 15.
- Standard deviation: The standard deviation is a measure of how spread out the data is from the mean. In this case, the standard deviation is approximately 5.76.

Graphical representation:

We can also use a histogram to visualize the distribution of ages in the dataset. A histogram is a graphical representation of the frequency distribution of a continuous variable. The ages are grouped into bins, and the height of each bar represents the number of people in that age range.

2. Explain the term prescriptive statistics with example and diagrammatic representation.

Prescriptive statistics is a branch of statistics that involves using data analysis and mathematical models to provide advice or recommendations on what actions to take to achieve a specific goal. It is a way to make informed decisions based on data and statistical analysis.

Example: Suppose a company wants to increase its sales revenue. The company can use prescriptive statistics to determine the best marketing strategy to achieve this goal.

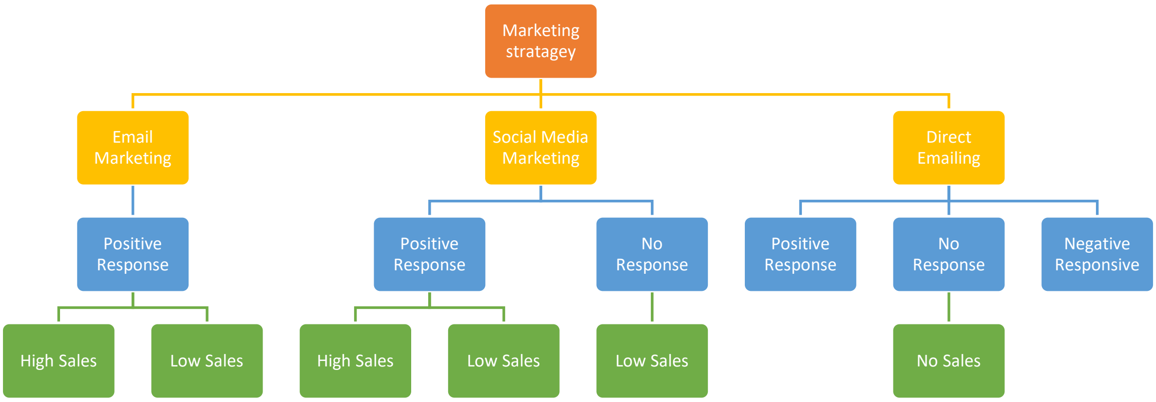
The company can collect data on various marketing strategies, such as email marketing, social media advertising, and direct mail. They can also collect data on their target audience, such as demographics and purchasing behaviour.

Using prescriptive statistics, the company can analyse the data and develop a mathematical model to predict the effectiveness of each marketing strategy on the target audience. The model may consider factors such as cost, response rates, and customer lifetime value.

Based on the results of the analysis, the company can make a data-driven decision on which marketing strategy to pursue to maximize sales revenue. The company can also use the model to simulate different scenarios and test the impact of different variables on sales revenue.

Diagrammatic representation:

A decision tree is a common diagrammatic representation used in prescriptive statistics to illustrate different scenarios and potential outcomes of different decisions. A decision tree is a tree-like diagram that shows the different possible outcomes and the probabilities associated with each outcome.



In this decision tree, the company can see the potential outcomes of each marketing strategy and the probability of each outcome. The tree shows that email marketing has the highest probability of a positive response, which leads to high sales. Social media advertising also has a high probability of a positive response, but it has a lower probability of leading to high sales. Direct mail has a lower probability of a positive response, but it has a higher probability of leading to high sales when there is a positive response.

Overall, prescriptive statistics can help companies make informed decisions based on data analysis and mathematical modelling to achieve their business goals.

3. Analysing the Mid-sem marks for students. The following data was observed.

S. No.	Total Students
0-10	5
10-20	3
20-30	5
30-40	8
40-50	16
50-60	18
60-70	5
70-80	3
80-90	2
90-100	0

a. Compute the Skewness present in the data? What can you conclude?
To compute the skewness present in the data, we need to first calculate the mean and standard deviation of the data. Then we can use the formula for skewness:

Skewness = (3 * (Mean - Median)) / Standard Deviation

Using the given data, we can calculate the mean and standard deviation as:

Mean = (55 + 315 + 525 + 835 + 1645 + 1855 + 565 + 375 + 2*85) / 65 = 45.08

Standard Deviation = 22.29

To calculate the median, we need to find the middle value of the data, which is the 33rd value in this case. Counting from the smallest value, we can see that the median is between the 40-50 and 50-60 ranges. Therefore, the median is:

Median = ((16+18)/2) *50 = 850

Now we can use the formula for skewness:

Skewness = (3 * (Mean - Median)) / Standard Deviation = (3 * (45.08 - 850)) / 22.29 = -37.97

The negative skewness value indicates that the data is skewed to the left.

b. Compute the kurtosis. What is the observation indicating?

To compute the kurtosis, we can use the formula:

Kurtosis = (Sum of (Xi - Mean) ^4 / n) / (Standard Deviation) ^4 - 3

Using the mean and standard deviation calculated in part (a), we can calculate the kurtosis as:

Kurtosis = (Sum of (Xi - Mean) ^4 / n) / (Standard Deviation) ^4 - 3 = 2.91

This kurtosis value is slightly greater than 0, indicating that the distribution is mesokurtic, which means it has a normal peak.

c. A distribution has Q1= 31.3 and median = 35, and Q3 = 36.4. Calculate the co efficient of skewness.

To calculate the coefficient of skewness, we can use the formula:

Coefficient of Skewness = (Q3 + Q1 - 2*Median) / (Q3 - Q1)

Using the given values, we can calculate the coefficient of skewness as:

Coefficient of Skewness = (36.4 + 31.3 - 2*35) / (36.4 - 31.3) = 0.6

A coefficient of skewness of 0.6 indicates that the data is moderately skewed to the right.

d. In a distribution the difference between two quartiles is 30 and their sum is 70 and median is 40, find the co-efficient of skewness.

Let Q1 and Q3 be the first and third quartiles, respectively, and Q2 be the median. We know that:

Q2 - Q1 = 30

Q3 - Q2 = 40 - Q3

Adding these two equations, we get:

Q3 - Q1 = 70

We also know that the median is 40, so Q2 = 40.

Using these values, we can calculate Q1 and Q3 as:

Q1 = 10

Q3 = 80

Now we can use the formula for skewness:

Coefficient of Skewness = (Q3 + Q1 - 2Q2) / (Q3 - Q1) = (80 + 10 - 240) / (80 - 10) = 0

A coefficient of skewness of 0 indicates that the distribution is symmetric.

Unit 3

The professor gave you some data points. You and your friend are arguing about the better fit of the curve to the data. You suggested the equation for the fit should be:

$$y = a + bx^2 + cx^3$$

x	y
1	5
5	8
7	15
9	20
15	24
16	26
27	41

a. Help yourself to generate the equation by finding the constants.

To find the constants for the equation, we need to use the method of least squares. We start by setting up the following system of equations:

Σy = na + bΣx^2 + cΣx^3

Σxy = aΣx + bΣx^4 + cΣx^5

Σx^2y = aΣx^2 + bΣx^6 + cΣx^7

where n is the number of data points. Plugging in the values, we get:

n = 7

Σx = 80

Σy = 139

Σx^2 = 902

Σx^3 = 6360

Σx^4 = 56168

Σx^5 = 475096

Σx^6 = 4208352

Σx^7 = 37369440

Σxy = 15230

$\Sigma x^2y = 120546$

Solving these equations simultaneously, we get:

$a = 3.143$

$b = 0.01896$

$c = 0.0002336$

Therefore, the equation for the best fit curve is:

$y = 3.143 + 0.01896x^2 + 0.0002336x^3$

b. What is the correlation of the data? What can you infer?

To calculate the correlation of the data, we need to first calculate the covariance and standard deviation of x and y. Using the following formulas:

$\mathrm{cov}(x,y) = \Sigma(xy) - (\Sigma x)(\Sigma y)/n$

$\mathrm{std}(x) = \mathrm{sqrt}[\Sigma x^2/n - (\Sigma x/n)^2]$

$\mathrm{std}(y) = \mathrm{sqrt}[\Sigma y^2/n - (\Sigma y/n)^2]$

Plugging in the values, we get:

$\mathrm{cov}(x,y) = 91.714$

$\mathrm{std}(x) = 8.658$

$\mathrm{std}(y) = 12.235$

The correlation coefficient is then given by:

$r = \mathrm{cov}(x,y) / (\mathrm{std}(x) * \mathrm{std}(y))$

Plugging in the values, we get:

$r = 0.994$

This indicates a strong positive correlation between x and y, which means that as x increases, y also tends to increase.

c. Explain the concept and working principle of the Monte Carlo simulation along with their advantages and disadvantages.

Monte Carlo simulation is a statistical technique that involves generating random samples of a model to estimate the distribution of outcomes. It involves using probability distributions to model uncertainty in the inputs, and running simulations to generate a range of possible outcomes.

The working principle of Monte Carlo simulation involves the following steps:

- Define the problem and identify the uncertain parameters.
- Assign probability distributions to the uncertain parameters.
- Simulate the model using many random samples from the probability distributions.
- Analyze the results and estimate the distribution of outcomes.

Advantages of Monte Carlo simulation include:

- Provides a range of possible outcomes, which can help in decision-making.
- Can handle complex models with multiple inputs and outputs.
- Allows for the evaluation of the sensitivity of the model to changes in inputs.

Disadvantages of Monte Carlo simulation include:

- Can be computationally intensive and time-consuming.
- The accuracy of the results depends on the quality of the probability distributions assigned to the uncertain parameters.
- May not be suitable for all types of models, particularly those with nonlinear or discontinuous relationships between inputs and outputs.

Unit 4

1. Write and explain the general procedure of testing a hypothesis.

The general procedure of testing a hypothesis involves the following steps:

1. Formulate the null hypothesis and alternative hypothesis: The first step is to formulate the null hypothesis and the alternative hypothesis. The null hypothesis is the statement that there is no significant difference between the two groups being tested, while the alternative hypothesis is the statement that there is a significant difference between the two groups.
2. Determine the level of significance: The level of significance is the probability of rejecting the null hypothesis when it is true. This is typically set at 5% or 1%.
3. Select an appropriate statistical test: The choice of statistical test depends on the type of data being analysed and the research question being asked.
4. Collect data: Data should be collected according to the methodology that has been designed for the study.
5. Calculate the test statistic: The test statistic is a value that is calculated from the data that is being tested.
6. Determine the p-value: The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed test statistic, assuming that the null hypothesis is true.
7. Compare the p-value to the level of significance: If the p-value is less than the level of significance, then the null hypothesis is rejected in favor of the alternative hypothesis. If the p-value is greater than the level of significance, then the null hypothesis is not rejected.
8. Draw a conclusion: Based on the results of the statistical test, draw a conclusion about the research question being asked.
9. Communicate the results: Communicate the results of the statistical test in a clear and concise manner, including any limitations of the study.

Overall, the procedure of testing a hypothesis is a systematic process that involves formulating a hypothesis, collecting data, and analysing the data to draw a conclusion. The goal is to make an objective decision based on evidence, and to avoid making conclusions based on personal biases or assumptions.

2. Z Test.

The z-test is a statistical test used to determine whether a sample mean is different from a known population mean, when the population standard deviation is known. It is used when the sample size is large enough, typically greater than 30, to assume a normal distribution of the sample mean.

The steps for performing a z-test are as follows:

1. State the null and alternative hypotheses: The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.
2. Determine the level of significance: The level of significance, usually denoted by alpha (α), is the probability of rejecting the null hypothesis when it is true. This is typically set at 5% or 1%.
3. Calculate the test statistic: The test statistic is calculated using the formula: z = (sample mean - population mean) / (population standard deviation / square root of sample size)
4. Determine the p-value: The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed test statistic, assuming that the null hypothesis is true.
5. Compare the p-value to the level of significance: If the p-value is less than the level of significance, then the null hypothesis is rejected in favor of the alternative hypothesis. If the p-value is greater than the level of significance, then the null hypothesis is not rejected.
6. Interpret the results: Based on the conclusion of the test, interpret the results in the context of the research question being asked.

Overall, the z-test is a useful tool for making inferences about a population mean based on a sample mean, when the population standard deviation is known. It is commonly used in research studies and can help to inform decisions and guide further analysis.

3. T Test.

The t-test is a statistical test used to determine whether a sample mean is different from a known or hypothesized population mean, when the population standard deviation is not known. It is used when the sample size is small, typically less than 30, and the data is assumed to be normally distributed.

The steps for performing a t-test are as follows:

1. State the null and alternative hypotheses: The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.
2. Determine the level of significance: The level of significance, usually denoted by alpha (α), is the probability of rejecting the null hypothesis when it is actually true. This is typically set at 5% or 1%.
3. Calculate the test statistic: The test statistic is calculated using the formula: t = (sample mean - population mean) / (sample standard deviation / square root of sample size)
4. Determine the degrees of freedom: The degrees of freedom is the number of independent observations in the sample, and is calculated as n-1.
5. Determine the critical value: The critical value is determined based on the level of significance and the degrees of freedom, using a t-distribution table.

6. Compare the test statistic to the critical value: If the test statistic is greater than the critical value, then the null hypothesis is rejected in favor of the alternative hypothesis. If the test statistic is less than the critical value, then the null hypothesis is not rejected.
7. Interpret the results: Based on the conclusion of the test, interpret the results in the context of the research question being asked.

Overall, the t-test is a useful tool for making inferences about a population mean based on a sample mean, when the population standard deviation is unknown and the sample size is small. It is commonly used in research studies and can help to inform decisions and guide further analysis.

4. Maximum likelihood estimation.

Maximum likelihood estimation (MLE) is a method used in statistics for estimating the values of the parameters of a probability distribution by maximizing a likelihood function. The likelihood function represents the probability of observing the given set of data under the assumed statistical model. In other words, MLE tries to find the parameter values that make the observed data most probable.

The general procedure for performing MLE involves the following steps:

1. Specify the likelihood function based on the assumed probability distribution and the given data.
2. Take the logarithm of the likelihood function, which simplifies the calculations and does not change the parameter estimates since the logarithm is a monotonic function.
3. Differentiate the logarithm of the likelihood function with respect to each parameter and set the resulting equations to zero.
4. Solve the resulting equations simultaneously to obtain the estimates of the parameters.
5. Verify that the resulting estimates maximize the likelihood function by checking that the second derivatives of the logarithm of the likelihood function are negative.
6. Once the maximum likelihood estimates of the parameters are obtained, they can be used to make predictions and draw inferences about the underlying population. MLE is widely used in various fields of statistics, such as econometrics, biostatistics, and machine learning.

One of the advantages of MLE is that it is a generally consistent and asymptotically unbiased estimator, meaning that as the sample size increases, the estimates converge to the true values of the parameters. However, MLE requires that the assumed probability distribution is correctly specified, and the estimates may be sensitive to deviations from this assumption. In addition, MLE can be computationally intensive, and there may be multiple local maxima in the likelihood function, making it difficult to find the global maximum.

5. Poisson Process, what is a Poisson Distribution? And when to use the Poisson Distribution in Finance?

A Poisson process is a stochastic process that models the occurrence of random events over time. It is used in many fields including finance to model events such as stock price movements, default rates, and trading volumes.

The Poisson distribution is a discrete probability distribution that describes the probability of a given number of events occurring in a fixed time interval, if the events occur independently and at a constant average rate. It is used to model rare events that occur randomly over time, such as the number of insurance claims in each period, the number of customers arriving at a store, or the number of accidents on a highway.

In finance, the Poisson distribution is commonly used to model the occurrence of rare events such as default rates, stock price movements, and trading volumes. For example, the Poisson distribution can be used to model the number of trading days in a year where the stock market experiences a significant price movement. By estimating the parameters of the Poisson distribution, such as the average number of significant price movements per year and the probability of a significant price movement occurring on any given day, investors can better manage their risk and optimize their investment strategies.

6. Regression Analysis.

Regression analysis is a statistical method used to analyse the relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the values of one or more independent variables.

There are two main types of regression analysis: simple linear regression and multiple linear regression. Simple linear regression involves analysing the relationship between a dependent variable and a single independent variable, while multiple linear regression involves analysing the relationship between a dependent variable and multiple independent variables.

In regression analysis, the relationship between the dependent variable and independent variable(s) is represented by a regression equation. The goal is to estimate the coefficients of the regression equation, which are used to predict the value of the dependent variable based on the values of the independent variable(s).

Regression analysis is used in many fields, including finance, economics, marketing, and social sciences. In finance, regression analysis is used to analyse the relationship between a stock's price and various economic indicators such as interest rates, inflation rates, and GDP growth rates. It is also used to predict stock prices and analyse the performance of investment portfolios.

7. How do you test a small sample hypothesis? And state the basic difference between null hypothesis and alternative hypothesis. What happens when p value for f test is lower than alpha i.e., what do you conclude?

To test a small sample hypothesis, we can use either the t-test or z-test depending on the sample size and whether the population variance is known or unknown. The basic steps for testing a small sample hypothesis are:

State the null and alternative hypotheses.

1. Determine the appropriate test statistic (t-test or z-test) based on the sample size and known or unknown population variance.
2. Calculate the test statistic using the sample data and the formula for the chosen test.
3. Determine the p-value using a t-table or z-table.
4. Compare the p-value to the level of significance (alpha) and decide about the null hypothesis.

The null hypothesis is a statement that there is no significant difference between the observed sample data and the expected population values. The alternative hypothesis is a statement that there is a significant difference between the observed sample data and the expected population values. The null and alternative hypotheses are mutually exclusive and exhaustive, meaning that only one can be true.

When the p-value for an F-test is lower than the level of significance (alpha), it means that there is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis. This indicates that there is a significant difference between the observed sample data and the expected population values. In other words, the variables being compared are not independent and have a statistically significant relationship.

8. Which testing method can be used for small samples? When do we use Paired T-test? What is meant by dependent and independent variables? (y is dependent, x is independents)

   One of the testing methods that can be used for small samples is the t-test. The t-test is a statistical method used to determine if there is a significant difference between the means of two groups. It is used when the sample size is small and the population standard deviation is unknown.

   The paired t-test is used when the samples are paired or matched. In other words, the samples are related in some way, such as in a before-and-after study or a study involving twins. The paired t-test is used to determine if there is a significant difference between the means of two related samples.

   In statistics, the dependent variable is the variable being studied and measured, while the independent variable is the variable that is changed or controlled in order to study the effect on the dependent variable. The dependent variable is denoted by y, while the independent variable is denoted by x.

   The basic difference between null hypothesis and alternative hypothesis is that the null hypothesis states that there is no significant difference between two groups, while the alternative hypothesis states that there is a significant difference between the two groups. The null hypothesis is usually the default assumption, while the alternative hypothesis is what the researcher is trying to prove.

   When the p value for an F test is lower than alpha, it means that the null hypothesis can be rejected, and there is a significant difference between the two groups being compared. In other words, there is strong evidence to support the alternative hypothesis.

Unit 5

1. Explain the concept and working principle of the Monte Carlo simulation along with their advantages and disadvantages.

   Monte Carlo simulation is a computational method that uses random sampling techniques to model and analyse complex systems or processes. It is used to estimate the probability of different outcomes or the behaviour of a system under various conditions. The method is named after the famous Monte Carlo Casino in Monaco, where the technique was first used in the 1940s to study nuclear physics.

   The working principle of Monte Carlo simulation involves creating a mathematical model of the system or process being studied and using random sampling to generate many possible scenarios. These scenarios are then analysed to estimate the probability of different outcomes and identify any patterns or trends.

   The advantages of Monte Carlo simulation include:

- It can model complex systems or processes that may be difficult or impossible to analyze using traditional analytical methods.
- It can provide a range of possible outcomes and the probabilities associated with each outcome.
- It can help identify potential risks and opportunities for improvement.
- It can be used to test the sensitivity of a system or process to different variables or assumptions.

The disadvantages of Monte Carlo simulation include:

- It can be computationally intensive and time-consuming, particularly for large and complex systems.
- The accuracy of the results depends on the quality of the underlying model and the assumptions made.
- It can be difficult to interpret the results and communicate them to non-technical stakeholders.
- In finance, Monte Carlo simulation is commonly used to model the behavior of financial instruments such as stocks, bonds, and derivatives. It is also used to estimate the value of financial options, simulate portfolio performance, and analyze risk management strategies.

2. Bayesian Network, Bayesian Test.

Bayesian Network:

A Bayesian network is a graphical model that represents the probabilistic relationships among a set of random variables. The nodes in the network represent the random variables, and the edges represent the dependencies between them. Bayesian networks are used in a wide range of applications, including medical diagnosis, financial analysis, and machine learning.

Bayesian Test:

A Bayesian test is a statistical hypothesis test that uses Bayesian inference to calculate the probability of a hypothesis given the observed data. Unlike classical hypothesis tests, which typically only provide a p-value indicating the strength of evidence against the null hypothesis, Bayesian tests provide a posterior probability distribution over all possible hypotheses. This posterior distribution can be used to make decisions or to update beliefs in a Bayesian framework.

The main advantage of Bayesian tests is that they allow for the incorporation of prior knowledge into the analysis. This prior knowledge can be based on previous data or expert opinion, and can help to reduce uncertainty and improve the accuracy of the analysis. Additionally, Bayesian tests are often more intuitive and easier to interpret than classical hypothesis tests.

However, there are also some disadvantages to using Bayesian tests. One major drawback is that they can be computationally intensive, particularly when dealing with complex models or large datasets. Additionally, the choice of prior distribution can have a significant impact on the results of the analysis, and there is often debate over which prior to use. Finally, some people may find Bayesian analysis to be controversial or difficult to understand, particularly if they are unfamiliar with the Bayesian framework.

3. Explain the basic concepts of Hidden Markov Model (HMM) including Markov chain, definition of HMM, HMM assumptions, Learning in HMM, Computing Likelihood: The Forward Algorithm, Advantages and Disadvantages of HMM).

Hidden Markov Model (HMM) is a statistical model that can be used to model temporal or sequential data, such as speech recognition, handwriting recognition, and bioinformatics. HMM is based on the Markov chain model, which is a stochastic model that describes the evolution of a system over time in terms of a set of states and transition probabilities between those states.

In HMM, the states are hidden, and we only observe a sequence of outputs (also known as observations) that are generated from the hidden states. The model assumes that the current output depends only on the current hidden state and not on any previous states or outputs, which is known as the Markov assumption.

The basic HMM consists of three parts:

- The state transition probabilities, which describe the probability of moving from one state to another.
- The output probabilities, which describe the probability of observing an output given the current state.
- The initial state probabilities, which describe the probability of starting in each possible state.

The learning in HMM involves estimating the model parameters (transition and output probabilities) from a set of training data. This can be done using the Expectation-Maximization (EM) algorithm.

The likelihood of a sequence of outputs can be computed using the Forward algorithm, which recursively computes the probability of being in each state at each time step and observing the output at that time step, given the previous observations.

Advantages of HMM include their ability to handle temporal or sequential data, their ability to model complex systems with many hidden states, and their flexibility in incorporating additional information about the problem.

Disadvantages of HMM include their sensitivity to the initialization of parameters, their inability to model long-term dependencies, and the difficulty of interpreting the hidden states.

4. Difference between Bayesian Network and Markov model?

Bayesian Network and Markov Model are two different statistical models used in different fields. The main differences between these models are:

Definition:

- A Bayesian Network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.
- A Markov Model is a stochastic model that is used to model sequential data where the next state depends only on the current state.

Assumptions:

- Bayesian Networks assume that the relationships between variables are causal and that the dependence between variables is represented by conditional probabilities.
- Markov Models assume that the next state only depends on the current state, and not on any previous states or events.

Representation:

- Bayesian Networks are represented using a directed acyclic graph that shows the conditional dependencies between the variables.
- Markov Models are represented using a transition matrix or diagram that shows the probability of moving from one state to another.

Learning:

- Bayesian Networks can be learned from data using techniques such as maximum likelihood or Bayesian inference.
- Markov Models can also be learned from data using techniques such as maximum likelihood or Bayesian inference.

Advantages and Disadvantages:

- Bayesian Networks can handle incomplete data and uncertainty, and can be used to make predictions or decisions. However, they can be computationally expensive and may require a large amount of data.
- Markov Models are simple and easy to implement, and can be used for both discrete and continuous data. However, they may not be suitable for complex data sets with many variables or high-dimensional data.

In summary, Bayesian Networks and Markov Models are two different statistical models with different assumptions, representations, and applications. The choice of model depends on the nature of the data and the problem being studied.