PANIMALAR INSTITUTE OF TECHNOLOGY

JAISAKTHI EDUCATIONAL TRUST
(Affiliated to Anna University, Chennai)

Bangalore Trunk Road, Varadharajapuram,
Poonamallee, Chennai – 600 123

DEPARTMENT OF
ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

II YEAR – IV SEMESTER

AD8403 DATA ANALYTICS

QUESTION BANK

ACADEMIC YEAR 2021-22

AD8403 DATA ANALYTICS

UNIT I

INFERENTIAL STATISTICS I

Populations – samples – random sampling – probability and statistics Sampling distribution – creating a sampling distribution – mean of all sample means – standard error of the mean – other sampling distributions Hypothesis testing – z-test – z-test procedure – statement of the problem – null hypothesis – alternate hypotheses – decision rule – calculations – decisions – interpretations

PART - A

1)What is population?

In statistics, population is the entire set of items from which you draw data for a statistical study. It can be a group of individuals, a set of items, etc. It makes up the data pool for a study.

2)What is a sample?

A sample represents the group of interest from the population, which you will use to represent the data. The sample is an unbiased subset of the population that best represents the whole data.

3) When are samples used?

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical and is unlimited in size. Take the example of a study that documents the results of a new medical procedure. It is unknown how the procedure will affect people across the globe, so a test group is used to find out how people react to it.

4)Difference Between Population and Sample?

Population	**Samples**
All residents of a country would constitute the Population set	All residents who live above the poverty line would be the Sample
All residents above the poverty line in a	All residents who are millionaires would make

country would be the Population	up the Sample
All employees in an office would be the Population	Out of all the employees, all managers in the office would be the Sample

5) DEFINE HYPOTHETICAL POPULATION

A population containing a finite number of individuals, members or units is a class. ... All the 400 students of 10th class of particular school is an example of existent type of population and the population of heads and tails obtained by tossing a coin on infinite number of times is an example of hypothetical population.

6) WHAT IS RANDOM SAMPLINGS

Random sampling occurs if, at each stage of sampling, the selection process guarantees that all potential observations in the population have an equal chance of being included in the sample

8) What Is sampling Distrubtion ?

the sampling distribution of the mean refers to the probability distribution of means for all possible random samples of a given size from some population.

9)WHAT ARE THE TYPES OF SAMPLING DISTRIBUTION?

- Sampling distribution of mean
- Sampling distribution of propotion
- T-distribution

10)Define Sampling distribution of mean

The most common type of sampling distribution is of the mean. It focuses on calculating the mean of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

11)what is meanby Sampling distribution of proportion

This sampling distribution focuses on proportions in a population. Samples are selected and their proportions are calculated. The mean of the sample proportions from each group represent the proportion of the entire population

12)DefineT-distribution

A T-distribution is a sampling distribution that involves a small population or one where not much is known about it. It is used to estimate the mean of the population and other statistics such as confidence intervals, statistical differences and linear regression. The T-distribution uses a t-score to evaluate data that wouldn't be appropriate for a normal distribution.

The formula for t-score is: **t = [x - μ] / [s / sqrt(n)]**

In the formula, "x" is the sample mean and "μ" is the population mean and signifies standard deviation.

13)Define MEAN OF ALL THE SAMPLE MEAN

The mean of the sampling distribution of the mean always equals the mean of the population.

14) Standard Error Of The Mean

The standard error of the mean equals the standard deviation of the population divided by the square root of the sample size

15) What is the Special Type Of Standard Deviation

You might find it helpful to think of the standard error of the mean as a rough measure of the average amount by which sample means deviate from the mean of the sampling distribution or from the population mean.

16)What Is The Hypothesis Testing

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H_0.

17) Hypothesized Sampling Distribution

When you perform a hypothesis test of a single population mean μ using a normal distribution (often called a z-test), you take a simple random sample from the population. ... Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with μ = p and σ=√pqn σ = p q n .

18) Define Decision Rule

A decision rule specifies precisely when H0 should be rejected (because the observed z qualifies as a rare outcome). There are many possible decision rules, as will be seen in Section 11.3. A very common one, already introduced in Figure 10.3, specifies that H0 should be rejected if the observed z equals or is more positive than 1.96 or if the observed z equals or is more negative than –1.96. Conversely, H0 should be retained if the observed z falls between ± 1.96.

19)Define null hypothesis?

The null hypothesis is a typical statistical theory which suggests that no statistical relationship and significance exists in a set of given single observed variable, between two sets of observed data and measured phenomena.

20) What is Level of Significance

total area that is identified with rare outcomes. Often referred to as the level of significance of the statistical test, this proportion is symbolized by the Greek letter α (alpha) and discussed more thoroughly in Section 11.4. In the present example, the level of significance, α, equals 05.

PART B

1)Explian population and samples. And difference?

2)Describe random sampling?

3)Explian sampling distribution and types?

4) Assume you have taken 100 samples of size 64 each from a population. The population variance is 49.(sampling distribution)

5) Reaction times in a population of people have a standard deviation of 20 milliseconds. How large must a sample size be for the standard deviation of the sample mean reaction time to be no larger than 2 milliseconds?(sampling distribution of a mean)

6) A machine puts an average of 4 grams of jelly beans in bags, with a standard deviation of 0.25 grams. 40 bags are randomly chosen, what is the probability that the mean amount per bag in the sampled bags is less than 3.9 grams. (sampling distribution of a mean)

7) Suppose that the mean height of college students is 70 inches with a standard deviation of 5 inches. If a random sample of 60 college students is taken, what is the probability that the sample average height for this sample will be more than 71 inches?(sampling distribution)

8)Describe null hypothesis test in detail?

9)Explain in detail hypothesis testing and examples?

10)Does the mean of SAT math score for al local freshman differ for all local average of 500?(z test for population mean)

UNIT II

INFERENTIAL STATISTICS II

Why hypothesis tests? – Strong or weak decisions – one-tailed and two-tailed tests – case studies Influence of sample size – power and sample size 46 Estimation – point estimate – confidence interval – level of confidence – effect of sample size

PART A

1)Why hypothesis test?

The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favor of a certain belief, or hypothesis, about a parameter.

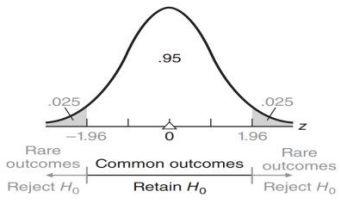
2)what is the importance of standard error?

Standard error statistics measure how accurate and precise the sample is as an estimate of the population parameter. It is particularly important to use the standard error to estimate an interval about the population parameter when an effect size statistic is not available.

3) Define Minimizing Incorrect Decisions

Even though we never really know whether a particular decision is correct or incorrect, it is reassuring that in the long run, most decisions will be correct— assuming the null hypotheses are either true or seriously false

4)Draw Strong Or Weak Decisions;



Proportions of area associated with common and rare outcomes ($\alpha = .05$).

5) what is mean of Rejecting H0 Is a Strong Decision:

the decision to retain H0 implies not that H0 is probably true, but only that H0 could be true, whereas the decision to reject H0 implies that H0 is probably false (and that H1 is probably true).

6) Why the Research Hypothesis Isn't Tested Directly

Even though H0, the null hypothesis, is the focus of a statistical test, it is usually of secondary concern to the investigator. Nevertheless, there are several reasons why, although of primary concern, the research hypothesis is identified with H1 and tested indirectly.

7)Define Lacks Necessary Precision

The research hypothesis, but not the null hypothesis, lacks the necessary precision to be tested directly.To be tested, a hypothesis must specify a single number about which the hypothesized sampling distribution can be constructed.

8) Describe Supported by a Strong Decision to Reject

Because the research hypothesis is identified with the alternative hypothesis, the decision to reject the null hypothesis, should it be made, will provide strong support for the research hypothesis, while the decision to retain the null hypothesis, should it be made, will provide, at most, weak support for the null hypothesis

9)Define ONE-TAILED AND TWO-TAILED TESTS

Before a hypothesis test, if there is a concern that the true population mean differs from the hypothesized population mean only in a particular direction, use the appropriate one-tailed or directional test for extra sensitivity. Otherwise, use the more customary two-tailed or nondirectional test

10) What is Two-Tailed Test with example

Generally, the alternative hypothesis, H1, is the complement of the null hypothesis, H0. Under typical conditions, the form of H1 resembles that shown for the SAT example, namely,

$$H1: \mu \neq 500$$

This alternative hypothesis says that the null hypothesis should be rejected if the mean reading score for the population of local freshmen differs in either direction from the national average of 500. An observed z will qualify as a rare outcome if it deviates too far either below or above the national average. Panel A of Figure 11.2 shows rejection regions that are associated with both

tails of the hypothesized sampling distribution. The corresponding decision rule, with its pair of critical z scores of ±1.96, is referred to as a two-tailed or nondirectional test.

11) what is One-Tailed Test (Lower Tail Critical)

Now let's assume that the research hypothesis for the investigation of SAT math scores was based on complaints from instructors about the poor preparation of local freshmen. Assume also that if the investigation supports these complaints, a remedial program will be instituted. Under these circumstances, the investigator might prefer a hypothesis test that is specially designed to detect only whether the population mean math score for all local freshmen is less than the national average. This alternative hypothesis reads:

$$H1: \mu \leq 500$$

12) **What is One-Tailed Test (Upper Tail Critical)**

Panel C of Figure 11.2 illustrates a one-tailed or directional test with the upper tail critical. This one-tailed test is the mirror image of the previous test. Now the alternative hypothesis reads:

$$H1: \mu > 500$$

and its critical z equals 1.65. This test is specially designed to detect only whether the population mean math score for all local freshmen exceeds the national average. For example, the research hypothesis for this investigation might have been inspired by the possibility of eliminating an existing remedial math program if it can be demonstrated that, on the average, the SAT math scores of all local freshmen exceed the national average

13) Define Consequences of Reducing Standard Error

As can be seen by comparing Figure 11.5 and Figure 11.6, the reduction of the standard error from 2.5 to 1.5 has two important consequences:

1. It shrinks the upper retention region back toward the hypothesized population mean of 100.

2. It shrinks the entire true sampling distribution toward the true population mean of 103.

14) Define Power curve

A graph showing power as a function of some other variable; specifically a graph of the power output of a vehicle or aircraft against engine speed. 2 figurative Chiefly Business. The current thinking or trend. 3Statistics. A graphical representation of the power function of a statistical test.

15) For a one-tailed or directional test with the lower tail critical

H0: μ ≥ SOME NUMBERS

H1: μ < SOME NUMBERS

16)For a one-tailed or directional test with the upper tail critical,

H0: μ ≤ SOME NUMBERS

H1: μ > SOME NUMBERS

17) What are four possible outcomes for any hypothesis test:

■ If H0 really is true, it is a correct decision to retain the true H0.

■ If H0 really is true, it is a type I error to reject the true H0.

■ If H0 really is false, it is a type II error to retain the false H0.

■ If H0 really is false, it is a correct decision to reject the false H0.

18) Define POINT ESTIMATE

A point estimate for μ uses a single value to represent the unknown population mean.

19)What is mean by CONIDENCE INTERVAL (CI) FOR μ

A confidence interval for μ uses a range of values that, with a known degree of certainty, includes the unknown population mean.

20) Define EFFECT OF SAMPLE SIZE

The larger the sample size, the smaller the standard error and, hence, the more precise (narrower) the confidence interval will be. Indeed, as the sample size grows larger, the standard error will approach zero and the confidence interval will shrink to a point estimate. Given this perspective, the sample size for a confidence interval, unlike that for a hypothesis test, never can be too large.

PART B

1)Describe hypothesis test in detail.

2)Explain strong weak decision in detail.

3)Brefily explain one tailed and two tailed test.

4)Define estimation .Expalin in detail about point estimation

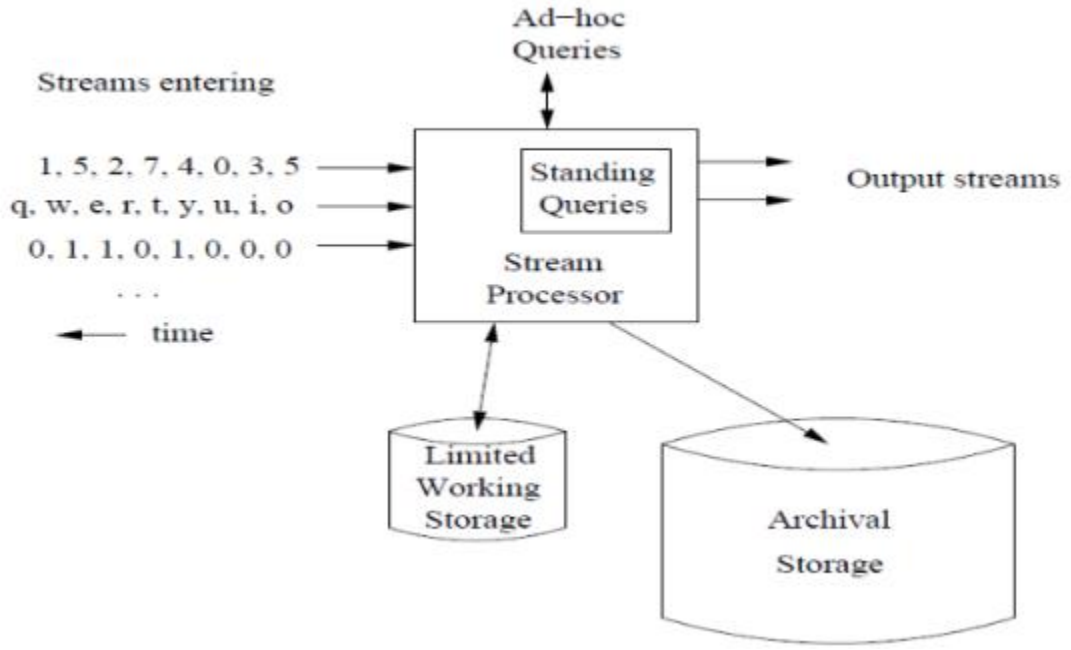
UNIT III

T-TEST

t-test for one sample – sampling distribution of t – t-test procedure – degrees of freedom – estimating the standard error – case studies t-test for two independent samples – statistical hypotheses – sampling distribution – test procedure – p-value – statistical significance – estimating effect size – meta analysis t-test for two related samples

PART A

1. Draw Stream Processing – Architecture



2. Write the Issues & Constraints of stream Processing

• Data stream processing differs from other Big Data Processing because this is mostly real time, not batch processing.

• Data need to be processed on its flight. That is, store & process is not possible. If the data is not processed in the stream then it is lost for good.

• Speed of data stream could be “very high”; in the sense not enough processing capability to process each and our element in the streamVolume of traffic could be “very high”; in the sense

not enough storage to store and process.Every other issue in this area can be traced to Speed and Volume of data.

• There should be provision to handle both ad-hoc & pre-defined queries.

• Reporting need not be real time.

3. Give the Examples of Stream Processing.

Sensor based data collection, Internet traffic targeting a server, Routed packets in a back-bone router

4. What are the Problems in Filtering Streams?

Filtering requires: Matching some key and data values in the streaming data with stored keys. This requires some table lookup – consequently this makes it difficult to scale filtering

5. What does Bloom filter consist of?

Bloom filter consists of: A bit-array of n bits (n buckets), initially all the bits set to 0's. A collection of hash functions h1, h2, . . . , hk. Each hash function maps a "key" value to n buckets, corresponding to the n bits of the bit-array. A set S of m key values.The purpose of the Bloom filter is to allow through all stream elements whose keys are in S, while rejecting most of the stream elements whose keys are not in S.

6. Write an Application of Bloom Filtering

Spam filtering in email

7. State Flajolet-Martin Algorithm

Flajolet-Martin algorithm approximates the number of unique objects in a stream or a database in one pass. If the stream contains n elements with m of them unique, this algorithm runs in O (n) time and needs O(log(m)) memory. So the real innovation here is the memory usage, in that an exact, brute-force algorithm would need O (m) memory This is an approximate algorithm. It gives an approximation for the number of unique objects, along with a standard deviation, which can then be used to determine bounds on the approximation with a desired maximum error, if needed.

8. How can the Accuracy of counting be improved?

• Averaging: Use multiple hash functions and use the average R instead.

• Bucketing: Averages are susceptible to large fluctuations. So use multiple buckets of hash functions from the step and use the median of the average R. This gives good accuracy.

9. List some types of Simple Moments.

0th Moment simply calculates the number of distinct elements in a stream

1st Moment simply calculates the frequencies of distinct elements in a stream

2nd Moment calculates the sum of the squares of the frequencies of distinct elements in a stream. The second moment is sometimes called the surprise number, since it measures how uneven the distribution of elements in the stream is.

10. List the steps to be followed when a new element arrives at the stream window for a decaying window.

• Multiply the current sum by 1-c

• Add a(t+1)

11. Write the rules to be followed when representing a stream by buckets

• The right end of a bucket is always a position with a 1

• No position is in more than one bucket

• There are one or two buckets of any given size up to some maximum size

• All sizes must be a power of 2

• Buckets cannot decrease as we move to the left

12. What is Stream Processing?

A stream is a sequence of data elements made available over time , it can be thought of as a conveyor belt that allows items to be processed one at a time rather than in large batches. Streams are processed differently from batch data – normal functions cannot operate on streams as a whole, as they have potentially unlimited data, and formally, streams are co-data, not data.

13. What is random walk hypothesis?

When applied to a particular financial instrument, the random walk hypothesis states that the price of this instrument is governed by a random walk and hence is unpredictable. If the random walk hypothesis is false then there will exist some correlation between the instrument price and some other indicators such as trading volume or the previous day's instrument closing price. If the correlation can be determined then a potential profit can be made.

14. List the types of stock market prediction methods

• Fundamental analysis

• Technical methods

• Internet based data sources

15. What is Stock Market prediction?

It is an act of trying to determine the future value of a company stock or other financial instrument traded on a financial exchange. The successful prediction of a stock's future price could yield significant profit.

16. What do you mean by Fundamental analysis?

Fundamental analysts are concerned with the company that underlies the stock itself. They evaluate company's past performance as well as the credibility of its accounts. Many performance ratios are created that aid the fundamental analyst with assessing the validity of a stock, such as the P/E ratio.

17. What is Technical analysis?

Technical analysts or chartists are not concerned with any of the company's fundamentals. They seek to determine the future price of a stock based solely on the trends of the past price.

18. What do you mean by Internet-based data sources in stock market prediction?

Tobias Preis et al. introduced a method to identify online precursors for stock market moves, using trading strategies based on search volume data provided by Google Trends. Their analysis of Google search volume for 98 terms of varying financial relevance, published in Scientific Reports, suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.

19. How to estimate the number of 1's in a window?

We can estimate the number of 1's in a window of 0's and 1's by grouping the 1's into buckets. Each bucket has a number of 1's that is a power of 2; there are one or two buckets of each size, and sizes never decrease as we go back in time. If we record only the position and size of the buckets, we can represent the contents of a window of size N with O (log2N) space.

20. What is Exponentially Decaying Window?

Rather than fixing a window size, we can imagine that the window consists of all the elements that ever arrived in the stream, but with the element that arrived t time units ago weighted by e-ct for some time-constant c. Doing so allows us to maintain certain summaries of an exponentially

decaying window easily. For instance, the weighted sum of elements can be recomputed, when a new element arrives, by multiplying the old sum by 1-c and then adding the new element.

PART B

1.Explain sampling data in a stream with examples.

2. Write in detail about counting ones in a window and decaying window.

3. Describe about any two Real time Analytics Platform (RTAP) Applications.

4. Explain the steps to be followed when a new element arrives at the stream window for a decaying window.

5. Describe in detail the Alon-Matias-Szegedy Algorithm for Second Moments.

UNIT IV

ANALYSIS OF VARIANCE

F-test – ANOVA – estimating effect size – multiple comparisons – case studies Analysis of variance with repeated measures Two-factor experiments – three f-tests – two-factor ANOVA – other types of ANOVA Introduction to chi-square tests

PART A

1.Define F-Test?

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled

2. what is analysis of variance?

Analysis of variance is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher

3. Define effect size estimation.

Effect size estimates provide important information about the impact of a treatment on the outcome of interest or on the association between variables. • Effect size estimates provide a common metric to compare the direction and strength of the relationship between variables across studies.

4. what is mean by multiple comparisons, multiplicity or multiple testing.

the multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values. The more inferences are made, the more likely erroneous inferences become

5. Define ANOVA.

The repeated measures analysis of variance (ANOVA) is an omnibus test that is an extension of the dependent samples t test. The test is used to determine whether there are any significant differences between the means of three or more variables (also called levels

6. Mention a two-factor factorial design

A two-factor factorial design is an experimental design in which data is collected for all possible combinations of the levels of the two factors of interest. If equal sample sizes are taken for each of the possible factor combinations then the design is a balanced two-factor factorial design.

7.Define statistical test in F-test.

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.

8.What are the two- way analysis of variance?

the two-way analysis of variance is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable.

9.What are the types of ANOVA?

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time.

10.Define chi-square test.

The Chi-Square test is a statistical procedure used by researchers to examine the differences between categorical variables in the same population. For example, imagine that a research group is interested in whether or not education level and marital status are related for all people in the U.S

11.What Does the Analysis of Variance Reveal?

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

12. How to Use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups

13. What is the Analysis of Variance in Other Applications

In addition to its applications in the finance industry, ANOVA is also used in a wide variety of contexts and applications to test hypotheses in reviewing clinical trial data.For example, to compare the effects of different treatment protocols on patient outcomes; in social science research (for instance to assess the effects of gender and class on specified variables), in software engineering (for instance to evaluate database management systems), in manufacturing (to assess product and process quality metrics), and industrial design among other fields.

14. What Is a Test?

In technical analysis and trading, a test is when a stock's price approaches an established support or resistance level set by the market. If the stock stays within the support and resistance levels, the test passes. However, if the stock price reaches new lows and/or new highs, the test fails. In other words, for technical analysis, price levels are tested to see if patterns or signals are accurate.

A test may also refer to one or more statistical techniques used to evaluate differences or similarities between estimated values from models or variables found in data. Examples include the t-test and z-test

15. Define **Range-Bound Market Test**.
When a stock is range-bound, price frequently tests the trading range's upper and lower boundaries. If traders are using a strategy that buys support and sells resistance, they should wait for several tests of these boundaries to confirm price respects them before entering a trade.

Once in a position, traders should place a stop-loss order in case the next test of support or resistance fails.

16. What is the **Trending Market Test**.

In an up-trending market, previous resistance becomes support, while in a down-trending market, past support becomes resistance. Once price breaks out to a new high or low, it often retraces to test these levels before resuming in the direction of the trend. Momentum traders can use the test of a previous swing high or swing low to enter a position at a more favorable price than if they would have chased the initial breakout.

A stop-loss order should be placed directly below the test area to close the trade if the trend unexpectedly reverses.

17. Define Statistical Tests

Inferential statistics uses the properties of data to test hypotheses and draw conclusions. Hypothesis testing allows one to test an idea using a data sample with regard to a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. In particular, one seeks to reject the null hypothesis, or the notion that one or more random variables have no effect on another. If this can be rejected, the variables are likely to be associated with one another

18. What Is Alpha Risk?

Alpha risk is the risk that in a statistical test a null hypothesis will be rejected when it is actually true. This is also known as a type I error, or a false positive. The term "risk" refers to the chance or likelihood of making an incorrect decision. The primary determinant of the amount of alpha risk is the sample size used for the test. Specifically, the larger the sample tested, the lower the alpha risk becomes.

Alpha risk can be contrasted with beta risk, or the risk of committing a type II error (i.e., a false negative).

Alpha risk, in this context, is unrelated to the investment risk associated with an actively managed portfolio that seeks alpha, or excess returns above the market

19. What Is Range-Bound Trading?

Range-bound trading is a trading strategy that seeks to identify and capitalize on securities, like stocks, trading in price channels. After finding major support and resistance levels and connecting them with horizontal trendlines, a trader can buy a security at the lower trendline support (bottom of the channel) and sell it at the upper trendline resistance (top of the channel)

20. What Is a One-Tailed Test?

A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis

PART B

1.Describe Analysis of variance (ANOVA) in detail.

2. Explain detail in F-TEST and Analysis of Variance (ANOVA).

3. What is the testing equality of population (treatment) means explain in detail.

4. Explain in detail about ANOVA TABLE FOR TWO-WAY CLASSIFICATION

5.Define ANOVA. And explain Two-way ANOVA without interaction

6.What is the Two-way ANOVA with interaction in detail.

7. Describe Two-way ANOVA versus one-way ANOVA.

8. What Is Alpha Risk?

UNIT V

PREDICTIVE ANALYTICS

Linear least squares – implementation – goodness of fit – testing a linear model – weighted resampling Regression using StatsModels – multiple regression – nonlinear relationships – logistic regression – estimating parameters – accuracy Time series analysis – moving averages – missing values – serial correlation – autocorrelation Introduction to survival analysis

PART A

1. What Is Predictive Analytics?

The term predictive analytics refers to the use of statistics and modeling techniques to make predictions about future outcomes and performance. Predictive analytics looks at current and historical data patterns to determine if those patterns are likely to emerge again. This allows businesses and investors to adjust where they use their resources to take advantage of possible future events. Predictive analysis can also be used to improve operational efficiencies and reduce risk

2. Understanding Predictive Analytics

Predictive analytics is a form of technology that makes predictions about certain unknowns in the future. It draws on a series of techniques to make these determinations, including artificial intelligence (AI), data mining, machine learning, modeling, and statistics.[3] For instance, data mining involves the analysis of large sets of data to detect patterns from it. Text analysis does the same, except for large blocks of text

3. Predictive models are used for all kinds of applications, including:

- Weather forecasts
- Creating video games
- Translating voice to text for mobile phone messaging
- Customer service
- Investment portfolio development

All of these applications use descriptive statistical models of existing data to make predictions about future data

4. What is mean by Forecasting

Forecasting is essential in manufacturing because it ensures the optimal utilization of resources in a supply chain. Critical spokes of the supply chain wheel, whether it is inventory management or the shop floor, require accurate forecasts for functioning.

Predictive modelling is often used to clean and optimize the quality of data used for such forecasts. Modelling ensures that more data can be ingested by the system, including from customer-facing operations, to ensure a more accurate forecast.

5. Define Credit

Credit scoring makes extensive use of predictive analytics. When a consumer or business applies for credit, data on the applicant's credit history and the credit record of borrowers with similar characteristics are used to predict the risk that the applicant might fail to perform on any credit extended.

6. Define Underwriting

Data and predictive analytics play an important role in underwriting. Insurance companies examine policy applicants to determine the likelihood of having to pay out for a future claim based on the current risk pool of similar policyholders, as well as past events that have resulted in pay-outs. Predictive models that consider characteristics in comparison to data about past policyholders and claims are routinely used by actuaries

7. What is mean by Marketing

Individuals who work in this field look at how consumers have reacted to the overall economy when planning on a new campaign. They can use these shifts in demographics to determine if the current mix of products will entice consumers to make a purchase.

Active traders, meanwhile, look at a variety of metrics based on past events when deciding whether to buy or sell a security. Moving averages, bands, and breakpoints are based on historical data and are used to forecast future price movements

8. Predictive Analytics vs. Machine Learning

A common misconception is that predictive analytics and machine learning are the same things. Predictive analytics help us understand possible future occurrences by analyzing the past. At its core, predictive analytics includes a series of statistical techniques (including machine learning, predictive modelling, and data mining) and uses statistics (both historical and current) to estimate, or predict, future outcomes

9. What is the Decision Trees

If you want to understand what leads to someone's decisions, then you may find decision trees useful. This type of model places data into different sections based on certain variables, such as price or market capitalization. Just as the name implies, it looks like a tree with individual branches and leaves. Branches indicate the choices available while individual leaves represent a particular decision.

Decision trees are the simplest models because they're easy to understand and dissect. They're also very useful when you need to make a decision in a short period of time

10. Define Regression

This is the model that is used the most in statistical analysis. Use it when you want to determine patterns in large sets of data and when there's a linear relationship between the inputs. This method works by figuring out a formula, which represents the relationship between all the inputs found in the dataset. For example, you can use regression to figure out how price and other key factors can shape the performance of a security

11. Define Neural Networks

Neural networks were developed as a form of predictive analytics by imitating the way the human brain works. This model can deal with complex data relationships using artificial intelligence and pattern recognition. Use it if you have several hurdles that you need to overcome like when you have too much data on hand, when you don't have the formula you need to help you find a relationship between the inputs and outputs in your dataset, or when you need to make predictions rather than come up with explanations.

12. What are the Benefits of Predictive Analytics

There are numerous benefits to using predictive analysis. As mentioned above, using this type of analysis can help entities when you need to make predictions about outcomes when there are no other (and obvious) answers available.[9]

Investors, financial professionals, and business leaders are able to use models to help reduce risk. For instance, an investor and their advisor can use certain models to help craft an investment portfolio with minimal risk to the investor by taking certain factors into consideration, such as age, capital, and goals.[9]

There is a significant impact to cost reduction when models are used. Businesses can determine the likelihood of success or failure of a product before it launches. Or they can set aside capital for production improvements by using predictive techniques before the manufacturing process begins

13. Criticism of Predictive Analytics

The use of predictive analytics has been criticized and, in some cases, legally restricted due to perceived inequities in its outcomes. Most commonly, this involves predictive models that result in statistical discrimination against racial or ethnic groups in areas such as credit scoring, home lending, employment, or risk of criminal behaviour.

A famous example of this is the (now illegal) practice of redlining in home lending by banks. Regardless of whether the predictions drawn from the use of such analytics are accurate, their use is generally frowned upon, and data that explicitly include information such as a person's race are now often excluded from predictive analytics.

14. How Does Netflix Use Predictive Analytics?

Data collection is very important to a company like Netflix. It collects data from its customers based on their behaviour and past viewing patterns. It uses information and makes predictions

based to make recommendations based on their preferences. This is the basis behind the "Because you watched..." lists you'll find on your subscription.

15. What Is Data Analytics?

Data analytics is the science of analysing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption

16. What are the various steps of Data Analysis.

The process involved in data analysis involves several different steps:

1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
3. Once the data is collected, it must be organized so it can be analyzed. This may take place on a spreadsheet or other form of software that can take statistical data.
4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed

PART B

1.How do you solve the least square problem in Python? What is least square method in Python?

2.What is the goodness-of-fit test?

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the table below. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

Day of the Week Employees were Most Absent

	Monday	Tuesday	Wednesday	Thurs
Number of Absences	15	12	9	9

3.One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as in the table.

Number of Televisions	Percent
	10
1	16
2	55
3	11
4+	8

The table contains expected (*E*) percents.

A random sample of 600 families in the far western United States resulted in the data in this table.

Number of Televisions	Frequency
0	66

Number of Televisions	Frequency
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (*O*) frequency values.

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

4.Explain in detail about time series analysis with example.

5. Describe Regression using Stats Models.

6. Explain multiple regression with example

7.What is the nonlinear relationships and types .Difference between linear and non linear relationship

8.Describe logistic regression in deatil

9.Expalin in detail serial correlation and autocorrelation

10.Describe in detail Introduction to survival analysis