

Questions and Answers

# 1-Mark: (1\*5=5)

1. **What is a Database?**

Data can be stored in a database, which is a structured collection of data that is organized for easy access and manipulation.

1. **Define p-value.**

The p-value is used to decide whether we should accept Null Hypothesis or reject it.

If the p-value is lower than the predetermined significance value (called alpha or threshold value) then we reject the null hypothesis.

1. **What is a tree map?**

The treemap can be composed of nested rectangles in case we have hierarchical categories. For hierarchical categories in data, a tree chart does a good job.

1. **What do you mean by Degree of Freedom in statistics?**

The degree of freedom (df) represents the number of variables that have the freedom to vary in a calculation.

It is calculated by:

df=N−1

1. **When is heat-map used?**

Heat maps are used to show relationships between two variables, one plotted on each axis. By observing how cell colors change across each axis, we can observe if there are any patterns in value for one or both variables.

# 2-Marks: (2\*22=44)

1. **What is data and what are the different types of data?**

Data is information that can be processed by a computer or other electronic system. It can be in the form of numbers, text, images, audio, or any other form of digital information.

There are 2 types of data:

* Structured
* Non-structured

1. **What is Central Limit Theorem in statistics?**

Central Limit Theorem states that when the large sample size has a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.

1. **What are outliers?**

The concept of outliers is closely associated with robustness. Outliers are “abnormal” observations in the sample that seem very unlikely for the assumed distribution model or are remarkably different from the rest of the sample observations. Outliers can be originated from measurement errors, exceptional circumstances, changes in the data-generating process, etc.

1. **What is the difference between Data Science and Analytics?**

***Data science*** is a field that uses scientific methods, processes, and systems to extract knowledge and insights from structured and unstructured data.

***Data analytics*** is a field that focuses on using data to understand and inform decision-making.

1. **Mention some tools and libraries used in Data Science.**

Tools:

* Matplotlib
* Tableau
* Power BI

Libraries:

* Pandas
* Numpy
* Scikit-learn

1. **What is the difference between descriptive and inferential statistics?**

**Descriptive Statistics:**

It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

**Inferential Statistics:**

Inferential Statistics is used to draw inferences beyond the immediate data available. We can answer the following questions with the help of inferential statistics: making inferences about the population from the sample.

1. **What are different types of variables?**

Basically there are 2 types of variables:

1. Measure Variables

* Discrete Variables
* Continuous Variables

1. Categorical variables

* Binary variables
* Nominal variables
* Ordinal variables

1. **What is the difference between Standard variation and variance?**

**Variance:** Variance is the average squared deviation from the mean of a set of data.

It is used to find the standard deviation.

**Standard Deviation:** Standard Deviation shows the variation in data.

If the data is close together, the standard deviation will be small.

1. **What is Probability and how it is different from conditional probability?**

Probability is the likelihood of the occurrence of an event.

Mathematically, the probability that an event will occur is expressed as a number between 0 and 1.

Notationally, the probability of event A is represented by P(A).

Conditional probability is the probability of one event occurring with some relationship to one or more other events.

1. **What is the difference between critical and acceptance region?**

* A critical region is a set of values for the test statistic for which the null hypothesis is rejected.
* The acceptance region is set of values for the test statistic for which the null hypothesis is accepted.

1. **How is parametric test different from non-parametric test?**

A **parametric test** is considered when we have large dataset and mean is given. This helps in making powerful and effective decisions.

A **non-parametric test** is considered regardless of the size of the data set if the median value is better when compared to the mean value.

1. **Under what conditions t-test is used?**

The T-test is an inferential statistic that is used to determine the difference or to compare the means of two groups of samples which may be related to certain features. It is performed on continuous variables.

1. **When should we chi-square test?**

Chi-Square test is used when we perform hypothesis testing on two categorical variables from a single population or we can say that to compare categorical variables from a single population. By this we find is there any significant association between the two categorical variables.

1. **When should we use waterfall chart?**

Waterfall charts are used:

* When you have changes for the measure across time or across different categories.
* To audit the major changes contributing to the total value.

1. **Pie chart and donut chart are mostly similar. Under what conditions should we use pie chart and donut chart?**

Difference between Pie and Donut chart:

* A doughnut chart can contain two different data series in form of two concentric doughnuts, which is not possible in case of a pie chart.
* If there are more than 4 or 5 categories, prefer pie chart. For 2 and 4 categories, go with donut chart.

1. **What is the difference between bar graph and histogram?**

* Bar graphs are used to make comparison between different groups, we can also use it to track changes over time.
* Histograms display the distribution of continuous data. We can use these graphs, when we have continuous measurements and want to find distribution of values or look for outliers.

1. **What are different types of errors in statistics? And what is the difference between them?**

* **Type I error:**

When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha.

* **Type II error:**

When we accept the null hypothesis but it is false. Type II errors are denoted by beta.

1. **How is ANOVA test different from chi-square test?**

* **ANOVA** stands for Analysis of Variance and is used to compare multiple (three or more) samples with a single test. It is used when the categorical feature has more than two categories.
* **Chi-Square** test is used when we perform hypothesis testing on two categorical variables from a single population or we can say that to compare categorical variables from a single population.

1. **What are different types of ANOVA tests?**

There are 2 types of ANOVA tests:

* **One Way ANOVA:**

A one-way ANOVA has one independent variable.

Example: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

* **Two Way ANOVA:**

A two-way ANOVA has two independent variables.

Example: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master’s), and race finishing times in a marathon.

1. **What is the difference between one-tailed and two-tailed test?**

* ONE TAILED TEST

A one-tailed test is a statistical hypothesis 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.

* TWO TAILED TEST

Two-tailed hypothesis tests are also known as non-directional and two-sided tests because you can test for effects in both directions.

1. **When is Gauge chart used? Give an example.**

Gauge charts are known as a speedometer chart or a dial chart.

A gauge chart is most commonly used visual tools to represent progressive values, targets or deadlines.

Examples: Gauge charts are extensively used in project management to define deadlines, modules, and related details.

1. **What is the difference between Choropleth Map and Geospatial Map?**

**Choropleth Map:**

Choropleth maps are used to pair data with its geographic location.

They only work when the data set includes location-specific data such as the zip code, county, state, or even country.

**Geospatial Map**

Focus on the relationship between data and its physical location to create insights from it.

Geospatial visualizations highlight the physical connection between data points.

# 3-Marks: (3\*9=27)

1. **Give some real world examples where we use data science.**

* Supply chain optimization in the logistics industry
* Recommendation systems in marketing & advertising
* Weather predictions in the agriculture sector

1. **What are some differences between analysis and analytics?**

|  |  |
| --- | --- |
| **Analysis** | **Analytics** |
| We perform analysis on things that  have already happened in past. | Analytics is working on future |
| The why? How? What? Of happened in past. | Analytics is utilizing machine learning, Statistics, algorithm, models to take better decisions and get  better insight from data |
| What we earn last year | Analytics is defined as a process of  Transforming data into action. |

1. **What do you mean by Mean, Median and Mode?**

* Mean: It is simply the average of all the data (salary) values. Add all the numbers then divide by the amount of numbers.
* Median: It is the value in the middle when the data items are arranged in ascending order.
* Mode: It is the most frequently occurring value in a series of data in case of no repeating values, there would be no mode.

1. **What are different types of probability distributions?**

* Uniform Distribution: A uniform distribution can be used for every case in which any result in a sample space is equally possible.
* Normal Distribution: The "Bell Curve" is a Normal Distribution and some data that follows it closely, but not perfectly (which is usual).
* Binomial Distribution: It is considered a distribution where only two results are possible, such as success or defeat, gain or loss, win or lose, and where the chance of success and failure is the same for all the experiments.

1. **What do you mean by skewness? What is the difference between left and right skeweness?**

Skewness is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.

It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.

Difference between Left and Right skewness

* Left or Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side.
* Right or Positive Skewness means when the tail on the right side of the distribution is longer or fatter.

1. **What is kurtosis? What are different types of kurtosis?**

A measure of peakedness or convexity of a curve is known as Kurtosis.

There are 3 types of kurtosis:

* Mesokurtic (normal curve)
* Leptocurtic (leading curve)
* Platykurtic (flat curve).

1. **What is the difference between covariance and correlation? Which one should we prefer and why?**

* Covariance refers to relationship between two random variables in which a change in the one reflects a change in other variable.

The range of covariance is from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

* Correlation refers to how two variables move in relation to one another.

The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1. When it is close to 0 this means that there is little relationship between the variables and when it is farther away from 0 in positive or negative direction, greater is the relationship between the two variables.

When it comes to making a choice between covariance and correlation to measure relationship between variables, correlation is preferred over covariance because it does not get affected by the change in scale.

1. **What is a Hypothesis? What is the difference between null and alternate hypotheses?**

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is an assumption that we make about the population parameter.

Difference between Null and Alternate hypothesis:

* The null hypothesis predicts that, there is no relationship between the independent variable and dependent variable.
* The Alternate hypothesis predicts that, there is relationship between the independent variable and dependent variable.

1. **What are different types of t-tests?**

There are basically 3 types of t-tests:

* One sample t-test: One sample t-test which tells whether means of sample and population are different.
* Two sample t-test: It is known as Independent t-test. It compares the means of two independent groups and determines whether there is statistical evidence that the associated population means are significantly different.
* Paired t-test: When we want to compare means of different samples from same group or means from the same group at different times.

# 4-Marks:(4\*6=24)

1. **Mention some uses of scatter plot.**

Scatter plots are used to:

* Examine the correlation between two variables.
* Check the outliers
* Create a time series plot.
* Evaluate the fit of a regression model.

1. **When is funnel chart used?**

We can use this chart to show:

* An order fulfillment process.
* A sales process from start to finish.
* Flow of information from top secret to unclassified.
* Knowledge areas from general knowledge to expert knowledge.

1. **When should we use parametric and non-parametric tests?**

**Parametric Tests:**

* This test is used when data is quantitative and continuous.
* The data is of normal distribution.

**Non-Parametric Tests:**

* The data is estimated with different kinds of measurement scales.
* These are used when the distribution of the population is unknown.

1. **What are assumptions for using chi-square test?**

There are certain assumptions before performing a Chi-square test:

1. Both variables are categorical.
2. All observations are independent.
3. Cells in the contingency table are mutually exclusive.
4. Expected value of cells should be 5 or greater in at least 80% of cells.
5. **Name some charts which we use to visualize data?**

Following are the charts which we use to visualize data:

1. Bar chart
2. Pie chart
3. Donut chart
4. Waterfall chart
5. Tree map
6. Histogram
7. Line chart
8. Area Chart
9. **What steps are included in a Data Science process?**
10. **ACQUISITION –**

Data acquisition has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.

1. **EXPLORATION –**

Data exploration refers to the initial step of data processing in which data analysts use data visualization and mathematical methods to define dataset characterizations, such as scale, quantity, and precision.

1. **MUGGING, WRANGLING –**

It is the initial step of preprocessing and refining raw data into content or formats better suited for analysis.

1. **Analysis and Modeling –**

Data modeling is a collection of methods and tools used to explain and analyze how data can be processed, modified, and maintained.

Data analysis allows the analyst to extrapolate insights from huge bulk of data, it is done with the help of statistical and machine-learning techniques.