STATISTICS

1. What is Statistics?

Statistics is a science of data. it involves

- Collecting
- Classifying
- Analyzing
- Summarizing
- And interpreting numerical information

Statistics is used in several different disciplines (both Scientific & Non Scientific) to make decisions and draw conclusions based on data.

2. What are 2 branches of Statistics?

Descriptive Statistics & Inferential Statistics

Descriptive Statistics: Descriptive Statistics utilize numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present the information in a convenient form that individuals can use to make decisions. The main goal of descriptive statistics includes both numerical measures (e.g. the mean or the median) and graphical displays of data

Inferential Statistics: Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data. Some examples of inferential statistics might be a z statistics or a t-statistics

3. What is the Central Limit Theorem?

As the sample size of sample mean increases, the distribution of those means becomes more and more like a normal distribution, regardless of the shape of the original population.

In other words, as the sample size is bigger the more reliable estimates and make our data behave more predictably.

4. What is hypothesis testing?

Hypothesis testing is a statistical method used to make inferences about a population parameter based on sample data.

5. What is Confidence Interval?

A Confidence interval is a range of values calculated from sample data that is likely to contain the true population parameter with a certain level of confidence. It gives us an idea of the precision of our estimate,

For eg. If we calculate a 95% confidence interval for the mean, it means we are 95% confident that the true population mean falls within that interval

6. What is P Value?

The p-value is like a probability score that helps us decide if our findings are just by luck or if they're actually meaningful. A low p-value means our results are probably real, not just random chance. So, if we see a low p-value, it gives us confidence that our findings are significant.

7. What is mode and median?

Mode: Mode is the number that appears most frequently in a set of data.

Median: The median is the middle value when a set of data is arranged in order from ascending order. If there are two middle numbers, the median is the average of those two.

8. What is ANOVA?

Analysis of Variance (ANOVA) is a statistical technique which is used to check if the means of two or more groups are statistically different from each other.

It checks the impact of one or more factors by comparing the means of different samples.

9. What is the Chi Square Test?

Chi - Square Test of independence test whether two categorical variables are independent, that is whether there exists a relationship between two categorical variables

10. What is the Z test?

The Z-test is a statistical test used to determine if the mean of a sample is significantly different from a known population mean when the population standard deviation is also known.

11. What is the normal distribution?

Normal distribution, often referred to as the bell curve, is a common probability distribution that is symmetric and shaped like a bell. It shows how data is spread out around the average, and it's commonly used to understand and analyze information in many different areas

12. What is population and sample?

Population refers to the entire group of individuals, objects, or events that you're interested in studying.

Sample is a smaller subset of the population that you actually collect data from. Its often impractical to gather data from every single member of a population, so we take sample instead, and do test analysis on the sample.

13. What is inter quartile range?

The inter quartile range is a measure of statistical dispersion, specifically a measure of the spread of data within a dataset.

14. What is standard deviation?

The standard deviation gives us an idea of how much the values in a dataset vary from the average, helping us understand the spread or dispersion of our data.

15. What is one tail and 2 tailed hypothesis testing?

One-tail test is used when the hypothesis specifies the direction of the effect, either 'greater than' or 'less than'

For eg. If we're testing whether a new medicine improves test scores, the one-tail test would focus on whether the medicine makes scores 'greater than' a certain value or 'less than' a certain value, but not both.

Two-tail test is used when the hypothesis does not specify the direction of the effect, only that there is a difference,

For ef. If we're testing whether a coin is fair (equally likely to land heads or tails), the two-tail would check if there's a difference in either direction: more heads or more tails.

SUPERVISED MACHINE LEARNING

16. What is bagging and boosting?

Bagging: That often considers homogeneous weak learner, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process

It helps reduce overfitting and variance in the model by introducing randomness and diversity among the base models.

Boosting: That often considers homogeneous weak learners, learns them sequentially in a very adaptive way(a base model depends on the previous ones) and combines them following a deterministic strategy.

It aims to improve both bias and variance by iteratively focusing on difficult-to-classify instances, resulting in a strong learner.

17. What is Precision and Recall?

Precision: Precision is the conditional probability that the actual value is positive given that the prediction by the model is positive. In a sense it measures the accuracy of positive predictions made by the model.

Recall: Recall is the conditional probability that the predicted class is positive given that the actual class is positive. In a sense it measures the ability of the model to find all the positive instances in the dataset.

18. What is the Accuracy formula?

Accuracy is a measure of the overall correctness of a model's predictions and is calculated as the ratio of the number of correct predictions to the total number of predictions made.

$$Accuracy = (Number\ of\ Correct\ Predictions \div Total\ Number\ of\ Predictions) \times 100\%$$

19. What is the Confusion matrix?

A Confusion matrix is a table that is often used to describe the performance of a classification model. It presents a summary of the predictions made by the model compared to the actual labels in the dataset.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Each cell in the matrix represents a combination of predicted and actual classes, and the counts in these cells help evaluate the model's performance.

In simple terms, a confusion matrix provides a clear way to see how well a classification model is performing by showing the number of correct and incorrect predictions for each class.

20. What is Information gain?

The information gain for a feature F is calculated as the difference between the entropy in the segment before the split (S1) and the partitions resulting from the split (S2):

$$InfoGain\ (F) = Entropy\ (S1) - Entropy\ (S2)$$

Information gain helps decision tree algorithms decide which feature to split on first by measuring how much a particular feature reduces the randomness or uncertainty in the dataset when used for splitting.

21. What is Hyperparameter Tuning?

Hyperparameter tuning is the process of finding the best set of hyperparameters for a machine learning model. It's like finding the best settings for a machine learning model to make it perform its best. It's like adjusting the knobs and switches to get the optimal performance out of the model.

UNSUPERVISED MACHINE LEARNING

22. What is Unsupervised ML? Examples of some business problem?

In Unsupervised learning, the model looks for hidden patterns or structures in the data without any guidance or labels provided by humans. It aims to find relationships or groupings in the data on its own.

Eg:

Clustering: Grouping similar data points together. For eg, a retail store might use clustering to group customers based on their purchasing behavior to target them with personalized marketing campaigns.

Anomaly Detection: Identifying unusual or abnormal data points. For instance a cybersecurity company might use anomaly detection to detect unusual network activity that could indicate a potential security threat.

Dimensionality Reduction: Reducing the number of features in the data while retaining its essential information. This can help in visualization and speeding up other machine learning algorithms. For eg. a social media platform might use dimensionality reduction to analyze and visualize user behavior patterns.

23. What is elbow curve?

The elbow curve, also known as the elbow method, is a graphical technique used to determine the optimal number of clusters in a clustering algorithm, such as k-means clustering.

The elbow curve helps us decide how many clusters to use in a clustering algorithm by looking for the ‘elbow’ point on a graph where adding more clusters doesn’t improve the clustering much. It’s like finding the sweet spot where adding more flavors to your ice cream doesn’t make it taste any better.

24. What is PCA?

PCA, or Principal Component Analysis, is a technique used for dimensionality reduction in data analysis and machine learning. PCA is like compressing a big, complex puzzle into a simpler puzzle that still captures the main picture. It helps us focus on the most important aspects of the data while discarding the less important details.

25. What is K-means clustering? What is KNN?

K-means Clustering is a popular unsupervised machine learning algorithm used for clustering data points into groups or clusters based on their similarity. The algorithm aims to partition the data into K clusters, where each cluster is a represented by its centroid (the mean of all points in the cluster). K-means clustering is widely used in various applications, such as customer segmentation, image compression, and anomaly detection.

K-Nearest Neighbors is a simple and intuitive supervised machine learning algorithm used for classification and regression tasks. In classification, given a new data point, KNN predicts its class label based on the majority class of its K nearest neighbors in the training data. In regression, KNN predicts the output value for a new data point by averaging or taking the majority of the output values of its K nearest neighbors. KNN is commonly used in applications such as recommendation systems, handwriting recognition, and anomaly detection.

PYTHON

26. What are some examples of data structures in python?
    - Lists
    - Tuples
    - Dictionaries
    - Sets
    - Arrays

27. What are some examples of data types in python?
    - Integer(int)
    - Float
    - string(str)
    - Boolean(bool)
    - List
    - Tuple
    - Dictionary(dict)
    - Set

28. Difference between lists and tuples?

Lists	Tuples
Denoted as []	Denoted as ()
Mutable	Immutable
Slower than Tuples since it was mutable	Faster than Lists since it was Immutable

29. Is python case sensitive?

Yes, python is case sensitive, which means it distinguishes between uppercase and lowercase letters in variable names, function names, and other identifiers.

SOME OTHER COMMON DATA SCIENCE QUESTIONS:

30. What do you mean by word Data Science?

Data Science is an interdisciplinary field that involves extracting insights and knowledge from structured and unstructured data using scientific methods, processes, algorithms, and systems.

31. What is Data Visualization?

Data Visualization is the graphical representation of data and information using visual elements such as charts, graphs, and maps.

32. Describe in brief the Data Science Process flowchart?