Akash Kamerkar

Must-Known 60 Statistics Questions for Data science Interview Preparation : A Freshers' Guide

Basic + Medium+Advance

Akash Kamerkar

20 Easy level Statistics Questions

Akash Kamerkar

1.What is the difference between mean and median?

The mean is the average of a set of values, while the median is the middle value when the data is arranged in ascending or descending order.

Akash Kamerkar

2.What is standard deviation?

Standard deviation measures the spread or dispersion of a dataset around the mean.

Akash Kamerkar

3.What is a correlation coefficient?

A correlation coefficient measures the strength and direction of the linear relationship between two variables.

Akash Kamerkar

4.What is the difference between a population and a sample?

A population is the complete set of individuals or objects of interest, while a sample is a subset of the population used to make inferences about the entire population.

Akash Kamerkar

5.What is a p-value?

A p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed result, assuming the null hypothesis is true.

Akash Kamerkar

6.What is the Central Limit Theorem?

The Central Limit Theorem states that, for a large sample size, the sampling distribution of the sample mean will be approximately normally distributed regardless of the shape of the population distribution.

Akash Kamerkar

7.What is the difference between Type I and Type II errors?

Type I error occurs when we reject the null hypothesis when it is true, while Type II error occurs when we fail to reject the null hypothesis when it is false.

Akash Kamerkar

8.What is the difference between a parametric and non-parametric test?

Parametric tests assume certain properties about the population distribution, while non-parametric tests do not make any assumptions about the distribution.

Akash Kamerkar

9.What is the difference between a one-tailed and a two-tailed test?

A one-tailed test is directional and tests for a difference in a specific direction, while a two-tailed test is non-directional and tests for a difference in either direction.

Akash Kamerkar

10.What is the difference between covariance and correlation?

Covariance measures the linear relationship between two variables, whereas correlation measures both the strength and direction of the linear relationship.

Akash Kamerkar

11.What is the law of large numbers?

The law of large numbers states that as the sample size increases, the sample mean approaches the population mean.

Akash Kamerkar

12.What is a sampling distribution?

A sampling distribution is the probability distribution of a sample statistic, such as the mean or standard deviation, based on multiple samples taken from the same population.

Akash Kamerkar

13.What is the purpose of hypothesis testing?

Hypothesis testing is used to make inferences about a population based on a sample and to determine whether there is enough evidence to support or reject a claim about the population.

Akash Kamerkar

14.What is the difference between a null hypothesis and an alternative hypothesis?

The null hypothesis represents the default assumption or claim, while the alternative hypothesis represents the claim that we are trying to find evidence for.

Akash Kamerkar

15.What is a confidence interval?

A confidence interval is an estimate of a population parameter that provides a range of values within which the true population parameter is likely to fall, along with a level of confidence.

Akash Kamerkar

16.What is the difference between a t-test and a z-test?

A t-test is used when the population standard deviation is unknown, or the sample size is small, while a z-test is used when the population standard deviation is known, or the sample size is large.

Akash Kamerkar

17.What is the difference between probability and odds?

Probability represents the likelihood of an event occurring, while odds represent the ratio of the probability of success to the probability of failure.

Akash Kamerkar

18.What is the difference between a random variable and an observation?

A random variable is a variable that can take on different values with certain probabilities, while an observation is a specific value of the random variable.

Akash Kamerkar

19.What is the difference between a continuous and a discrete random variable?

A continuous random variable can take on any value within a specified range, while a discrete random variable can only take on specific, isolated values.

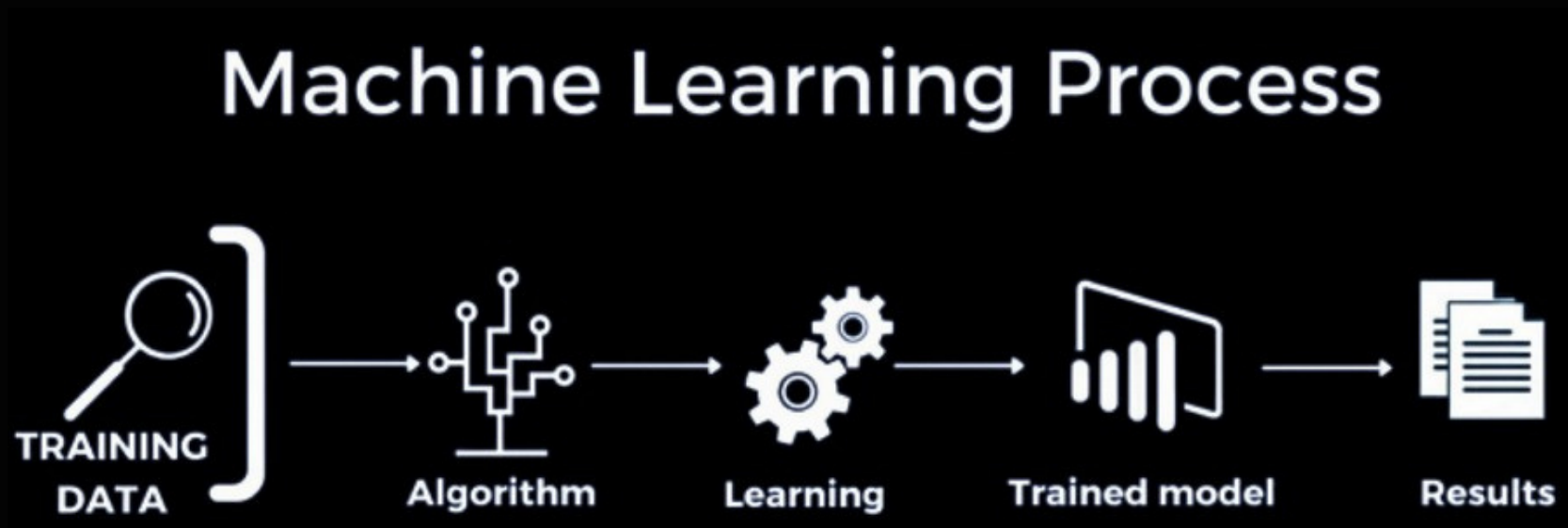
Akash Kamerkar

20.What is the difference between parametric and non-parametric regression?

Parametric regression assumes a specific functional form of the relationship between the independent and dependent variables, while non-parametric regression makes no assumptions about the functional form.

Akash Kamerkar

20 Medium level Statistics Questions



Akash Kamerkar

1.What's the difference between Likelihood and Probability?

Likelihood refers to the probability of observing the data given a specific parameter value, while probability refers to the likelihood of an event occurring.

Akash Kamerkar

2.What is the difference between Type I and Type II errors in hypothesis testing?

Type I error occurs when we reject the null hypothesis when it is true, while Type II error occurs when we fail to reject the null hypothesis when it is false.

Akash Kamerkar

3.What is the bias-variance trade-off in machine learning?

The bias-variance trade-off refers to the trade-off between the model's ability to fit the training data well (low bias) and its ability to generalize to new, unseen data (low variance).

Akash Kamerkar

4.Explain the concept of multicollinearity in regression analysis.

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can lead to unstable or unreliable coefficient estimates.

Akash Kamerkar

5.What is the difference between parametric and non-parametric statistics?

Parametric statistics make assumptions about the population distribution, while non-parametric statistics do not rely on specific distributional assumptions.

Akash Kamerkar

6.Explain the concept of overfitting in machine learning.

Overfitting occurs when a model performs well on the training data but fails to generalize to new, unseen data. It usually happens when a model is too complex and captures noise or random fluctuations in the training data.

Akash Kamerkar

7.What is the purpose of feature scaling in machine learning?

Feature scaling is used to bring different features or variables onto a similar scale to prevent certain features from dominating others and to ensure fair comparisons during model training.

Akash Kamerkar

8.What is the difference between stratified sampling and cluster sampling?

Stratified sampling involves dividing the population into homogeneous groups and then randomly selecting samples from each group, while cluster sampling involves dividing the population into heterogeneous groups (clusters) and randomly selecting entire clusters for sampling.

Akash Kamerkar

9.What is the difference between precision and recall in binary classification?

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances among all actual positive instances.

Akash Kamerkar

10.What is the purpose of regularization in machine learning models?

Regularization is used to prevent overfitting by adding a penalty term to the loss function, which encourages the model to have smaller parameter values and simpler representations.

Akash Kamerkar

11.Explain the concept of feature importance in a machine learning model.

Feature importance refers to the relative importance or contribution of each feature in a model's predictions. It helps identify the most influential features and understand their impact on the model's performance.

Akash Kamerkar

12.What is the difference between unsupervised and supervised learning?

In unsupervised learning, the model learns patterns and relationships in the data without explicit target labels, while in supervised learning, the model is trained using labeled data to predict or classify new instances.

Akash Kamerkar

13.What is the purpose of cross-validation in model evaluation?

Cross-validation is used to assess the performance and generalization ability of a model by splitting the data into multiple subsets, training the model on some subsets, and evaluating it on the remaining subsets.

Akash Kamerkar

14.Explain the concept of the bias-variance decomposition in machine learning.

The bias-variance decomposition decomposes the expected prediction error of a model into bias, variance, and irreducible error components, providing insights into the model's overall error and its sources.

Akash Kamerkar

15.What is the difference between precision and accuracy in classification metrics?

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, while accuracy measures the proportion of correctly predicted instances (both positive and negative) among all instances.

Akash Kamerkar

16.What is the purpose of the Receiver Operating Characteristic (ROC) curve?

The ROC curve is used to evaluate the performance of a binary classifier by plotting the true positive rate against the false positive rate at various classification thresholds.

Akash Kamerkar

17.Explain the concept of the curse of dimensionality in machine learning.

The curse of dimensionality refers to the challenges and issues that arise when working with high-dimensional data, such as increased computational complexity, sparsity of data, and the need for more data to maintain model performance.

Akash Kamerkar

18.What is the difference between a random forest and a gradient boosting model?

Random forest is an ensemble model that combines multiple decision trees with random feature selection, while gradient boosting builds an ensemble model iteratively by focusing on the samples with higher prediction errors.

Akash Kamerkar

19.Explain the concept of hypothesis testing using p-values.

Hypothesis testing involves making inferences about a population based on sample data. The p-value measures the strength of evidence against the null hypothesis and helps determine whether the observed result is statistically significant.

Akash Kamerkar

20.What is the difference between a chi-square test and a t-test?

A chi-square test is used to compare observed frequencies with expected frequencies in categorical data, while a t-test is used to compare means between two groups in numerical data.

Akash Kamerkar

20 Difficult level Statistics Questions

Akash Kamerkar

1.What is the difference between Type I, Type II, and Type III sums of squares in ANOVA?

Type I sums of squares measure the unique contribution of each variable, Type II sums of squares measure the contribution of a variable after accounting for other variables, and Type III sums of squares measure the contribution of a variable independently of other variables.

Akash Kamerkar

2.Explain the concept of autocorrelation in time series analysis.

Autocorrelation refers to the correlation between observations of a time series with previous observations at different lags. It indicates the presence of patterns or dependencies within the series.

Akash Kamerkar

3.What is the Box-Cox transformation?

The Box-Cox transformation is a method used to stabilize variance in a time series or to make the data conform more closely to the assumptions of a statistical model by applying a power transformation.

Akash Kamerkar

4.What is the Akaike Information Criterion (AIC)?

The AIC is a measure of the relative quality of a statistical model. It penalizes models with more parameters to avoid overfitting and provides a balance between goodness of fit and model complexity.

Akash Kamerkar

5.Explain the concept of latent variables in factor analysis.

Latent variables are unobserved variables that represent underlying dimensions or constructs in a dataset. Factor analysis is used to identify and measure these latent variables based on the observed variables.

Akash Kamerkar

6.What is the difference between ridge regression and lasso regression?

Ridge regression and lasso regression are regularization techniques used in linear regression. Ridge regression adds a penalty term based on the squared magnitude of the coefficients, while lasso regression adds a penalty term based on the absolute magnitude of the coefficients.

Akash Kamerkar

7.What is the Kalman filter?

The Kalman filter is an algorithm used to estimate the state of a dynamic system based on a series of noisy observations. It combines the predictions from a mathematical model with the observations to produce optimal estimates.

Akash Kamerkar

8.Explain the concept of Markov Chain Monte Carlo (MCMC) methods.

MCMC methods are a class of algorithms used to sample from complex probability distributions. They rely on Markov chains to generate a sequence of samples that converge to the desired distribution.

Akash Kamerkar

9.What is the difference between causality and correlation?

Correlation measures the statistical relationship between two variables, while causality implies a cause-and-effect relationship, indicating that changes in one variable directly influence changes in the other.

Akash Kamerkar

10.What is the difference between a likelihood ratio test and a Wald test in hypothesis testing?

A likelihood ratio test compares the likelihood of the data under the null hypothesis to the likelihood under an alternative hypothesis, while a Wald test examines the ratio of the estimated parameter to its standard error under the null hypothesis.

Akash Kamerkar

11.What is the concept of a p-value adjustment in multiple comparisons?

When performing multiple hypothesis tests, p-value adjustments (e.g., Bonferroni correction, Benjamini-Hochberg procedure) are used to control the overall false positive rate by compensating for the increased chance of finding significant results by chance.

Akash Kamerkar

12.Explain the concept of a hidden Markov model (HMM).

A hidden Markov model is a statistical model that assumes the underlying process is a Markov chain with hidden states. It is used to model systems with unobserved or partially observable states.

Akash Kamerkar

13.What is the difference between a Bayesian approach and a frequentist approach in statistics?

Bayesian statistics incorporates prior beliefs or knowledge into the analysis, updating them using data to obtain a posterior distribution. Frequentist statistics, on the other hand, relies solely on observed data and does not involve prior beliefs.

Akash Kamerkar

14.What is the concept of high-dimensional data and the challenges associated with it?

High-dimensional data refers to datasets with a large number of variables or features. Challenges include increased computational complexity, increased risk of overfitting, sparsity of data, and the curse of dimensionality.

Akash Kamerkar

15.Explain the concept of a Dirichlet distribution.

The Dirichlet distribution is a multivariate probability distribution defined on the simplex, commonly used as a prior distribution in Bayesian analysis of categorical data or proportions.

Akash Kamerkar

16.What is the concept of a copula in multivariate analysis?

A copula is a function that describes the joint distribution of a set of random variables while preserving their individual marginal distributions. It is used to model dependence structures in multivariate data.

Akash Kamerkar

17.Explain the concept of instrumental variables in econometrics.

Instrumental variables are used in regression analysis to address endogeneity or omitted variable bias. They are variables that are correlated with the independent variable of interest but are not directly correlated with the error term.

Akash Kamerkar

18.What is the concept of a survival analysis and the Kaplan-Meier estimator?

Survival analysis is used to analyze time-to-event data, such as time until failure or time until an event occurs. The Kaplan-Meier estimator is a nonparametric method used to estimate the survival function from censored data.

Akash Kamerkar

19.What is the concept of hierarchical modeling in statistics?

Hierarchical modeling, also known as multilevel modeling, is a statistical framework that allows for modeling nested or clustered data with multiple levels of variation.

Akash Kamerkar

20.What is the difference between a chi-square test and a t-test?

A chi-square test is used to compare observed frequencies with expected frequencies in categorical data, while a t-test is used to compare means between two groups in numerical data.