Statistics Pocket Dictionary

IMPORTANT STATISTICS TERMS

ramchandrapadwal

A

ANOVA (Analysis of Variance)

A statistical method used to compare means of two or more groups to determine if there are any significant differences between them.

B

Bias

Systematic deviation of results or inferences from the true value due to flaws in data collection or analysis.

Bayesian Inference

A statistical framework that combines prior knowledge and current data to update probabilities and make predictions.

Binomial Distribution

Discrete probability distribution for binary outcomes.

C

Central Limit Theorem

States that the sampling distribution of the mean of any independent, random variable will be approximately normal, regardless of the original distribution.

Coefficient of Variation

A measure of relative variability, calculated as the standard deviation divided by the mean, often expressed as a percentage.

Confidence Interval

A range around a sample statistic that is likely to contain the true population parameter with a certain level of confidence.

Covariance

A measure of the relationship between two variables, indicating whether they tend to increase or decrease together.

Cross-Validation

A technique used to assess the performance of a predictive model by partitioning data into subsets for training and testing.

Continuous Variable

Can take any numeric value within a range.

Categorical Variable

Represents categories or lables.

Chi-Square Test

Hypothesis test for categorical data association.

Confounding Variable

Unaccounted factor affecting the relationship between variables.

D

Data Mining

The process of discovering patterns, relationships, or information from large datasets.

Degrees of Freedom

The number of values in the final calculation of a statistic that are free to vary.

Descriptive Statistics

ummarizing and describing main features of a dataset using measures like mean, median, and mode.

Distribution

The way the values of a variable are spread or distributed across different outcomes.

Data

Information collected for analysis and interpretation.

Discrete Variable

Can only take distinct, separate values.

E

Ensemble Learning

A technique that combines multiple models to improve predictive performance and reduce overfitting.

Exponential Distribution

Continuous probability distribution for time between events in a Poisson process.

F

False Negative

In hypothesis testing, failing to reject a false null hypothesis (Type II error).

False Positive

In hypothesis testing, rejecting a true null hypothesis (Type I error).

Frequentist Statistics

An approach to statistics that relies on the frequency of events in the long run.

H

Hypothesis Testing

A process of making inferences about a population parameter based on a sample of data.

Histogram

Graph depicting data frequency distribution.

I

Imputation

Filling in missing data points with estimated or predicted values.

Interquartile Range (IQR)

The range between the first quartile and third quartile of a dataset; a measure of statistical dispersion.

Inferential Statistics

Techniques drawing conclusions about population from samples.

K

Kurtosis

A measure of the heaviness of the tails of a probability distribution.

L

Linear Regression:

A method for modeling the relationship between a dependent variable and one or more independent variables.

Logistic Regression

A statistical method used to model the probability of a binary outcome.

M

Machine Learning

A field of study that focuses on the development of algorithms that enable computers to learn from and make predictions or decisions based on data.

Mean

The arithmetic average of a set of numbers.

Median

The middle value in a dataset when arranged in ascending order.

Mode

The most frequently occurring value in a dataset.

Multicollinearity

The presence of high correlation between independent variables in a regression model.

Multivariate Analysis

Analyzing and modeling relationships between multiple variables.

N

Normal Distribution

A symmetric, bell-shaped probability distribution commonly observed in nature.

Null Hypothesis:

A statement that there is no significant difference between specified populations or datasets.

O

Outlier

An observation that deviates significantly from other observations in a dataset.

Overfitting

A phenomenon where a model performs well on the training data but poorly on unseen data due to memorizing noise instead of learning general patterns.

P

P-value

The probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming the null hypothesis is true.

PCA (Principal Component Analysis)

A technique used to simplify and reduce the dimensionality of high-dimensional data while preserving its variability.

Pearson's Correlation Coefficient

A measure of the linear relationship between two continuous variables.

Percentile

A value below which a given percentage of observations in a group fall.

Precision

In classification, the proportion of true positive predictions among all positive predictions.

Probability

The measure of the likelihood of an event occurring.

Population

The entire group of items, individuals or data under consideration.

R

Random Forest
An ensemble learning technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

Range
The difference between the maximum and minimum values in a dataset.

Regression Analysis
A statistical technique for modeling the relationship between a dependent variable and one or more independent variables.

Resampling
Techniques like bootstrapping and cross-validation used to estimate the properties of a population from a sample.

Random Variable
A variable with uncertain outcomes.

Regression Coefficient
Slope of the regression line.

R-Squared
Proportion of the dependent variable's variance explained by the independent variable.

Residual
Difference between observed and predicted values.

S

Sample
A subset of a population used to make inferences about the entire population.

Sampling
The process of selecting a subset of individuals or items from a larger population.

Skewness
A measure of the asymmetry of the probability distribution of a real-valued random variable.

Standard Deviation
A measure of the amount of variation or dispersion in a set of values.

Statistical Significance
he likelihood that an observed effect or relationship in data is not due to chance.

Supervised Learning
A machine learning approach where the model is trained on labeled data, learning to map input to output.

Statistics
The study of Data collection, analysis, interpretation and presentation.

T

T-test

A statistical hypothesis test used to determine if there is a significant difference between the means of two groups.

Type I Error

Rejecting a true null hypothesis (false positive).

Type II Error

Failing to reject a false null hypothesis (false negative).

Time Series

Data collected over successive time intercals.

U

Underfitting

Fitting a model too loosely to training data, resulting in poor predictive performance.

Univariate Analysis

Analyzing the distribution and characteristics of a single variable.

V

Variance

The measure of how spread out the data points are in a dataset.

VIF (Variance Inflation Factor)

A measure to detect multicollinearity in regression analysis.

Z

Z-Score

A measure that standardizes data points by subtracting the mean and dividing by the standard deviation.