datacamp

COURSE GLOSSARY

Introduction to Statistics in R

Categorical (qualitative) data: Data that represent discrete groups or categories where values indicate membership rather than magnitude

Central Limit Theorem: The principle that the sampling distribution of the sample mean (and many other summary statistics) approaches a normal distribution as sample size increases, provided samples are independent and identically distributed

Confounding: A situation in which a third variable is associated with both the treatment/exposure and the outcome, potentially producing a spurious association between them

Continuous data: Numeric data that can take on any value within an interval (including fractional values), such as time or height

Correlation coefficient: A numeric measure between –1 and 1 that quantifies the strength and direction of a linear relationship between two numeric variables

Descriptive statistics: Methods for summarizing and describing the main features of a dataset, such as means, medians, counts, and visualizations

Discrete data: Numeric data that consist of countable values, often integers, such as number of pets or number of successes

Hallucination: When a model produces confident but incorrect or fabricated information, often due to gaps or biases in its training data or reasoning process

Hallucination: When a model produces confident but incorrect or fabricated information, often due to gaps or biases in its training data or reasoning process

Interquartile range (IQR): The difference between the 75th and 25th percentiles (third and first quartiles), representing the spread of the middle 50% of the data

Mean: The arithmetic average of a set of numbers, computed by summing the values and dividing by the count of observations

Median: The middle value of a sorted dataset such that half the observations are below and half are above, used as a robust measure of center

Mode: The most frequently occurring value in a dataset, commonly used for categorical variables or to describe repeated values

Numeric (quantitative) data: Data made up of numeric values that represent measurable quantities and can be used in arithmetic operations

Observational study: A study in which researchers observe outcomes without random assignment, allowing analysis of associations but limiting causal conclusions due to potential confounding

Outlier: An observation that is substantially different from the rest of the data, commonly identified as below Q1 – 1.5×IQR or above Q3 + 1.5×IQR

Population: The entire set of items or individuals of interest from which samples may be drawn and about which inferences are made

Probability distribution: A function or table that assigns probabilities to each possible outcome of a random process (discrete) or describes probability density over a continuum (continuous)

Probability: A numeric measure between 0 and 1 (or 0% and 100%) that quantifies how likely an event is to occur

Randomized controlled trial (RCT): An experimental study design in which participants are randomly assigned to treatment or control groups to reduce bias and support causal inference

Sample: A subset of observations drawn from a larger group used to estimate properties of the whole population

Standard deviation: The square root of the variance that quantifies average distance from the mean in the original units of the data

Statistics: The field and practice of collecting, analyzing, interpreting, and presenting data to answer questions and inform decisions

Summary statistic: A single number that describes a characteristic of a dataset (for example, an average or a count) used to summarize or communicate key aspects of the data

Variance: A measure of spread equal to the average squared distance of each observation from the mean, with units squared