Glossary

Advanced Data Analytics

Terms and definitions from Course 4

A

A/B testing: A way to compare two versions of something to find out which version performs better

Addition rule (for mutually exclusive events): The concept that if the events A and B are mutually exclusive, then the probability of A or B happening is the sum of the probabilities of A and B

B

Bayes' rule: (Refer to **Bayes' theorem**)

Bayes' theorem: A math formula for stating that for any two events A and B, the probability of A given B equals the probability of A multiplied by the probability of B given A divided by the probability of B; also referred to as Bayes' rule

Bayesian inference: (Refer to **Bayesian statistics)**

Bayesian statistics: A powerful method for analyzing and interpreting data in modern data analytics; also referred to as Bayesian inference

Binomial distribution: A discrete distribution that models the probability of events with only two possible outcomes: success or failure

C

Central Limit Theorem: The idea that the sampling distribution of the mean approaches a normal distribution as the sample size increases

Classical probability: A type of probability based on formal reasoning about events with equally likely outcomes

Cluster random sample: A probability sampling method that divides a population into clusters, randomly selects certain clusters, and includes all members from the chosen clusters in the sample

Complement of an event: In statistics, refers to an event not occuring

Complement rule: A concept stating that the probability that event A does not occur is one minus the probability of A

Conditional probability: The probability of an event occurring given that another event has already occurred

Confidence interval: A range of values that describes the uncertainty surrounding an estimate

Confidence level: A measure that expresses the uncertainty of the estimation process

Continuous random variable: A variable that takes all the possible values in some range of numbers

Convenience sample: A non-probability sampling method that involves choosing members of a population that are easy to contact or reach

D

Dependent events: The concept that two events are dependent if one event changes the probability of the other event

Descriptive statistics: A type of statistics that summarizes the main features of a dataset

Discrete random variable: A variable that has a countable number of possible values

E

Econometrics: A branch of economics that uses statistics to analyze economic problems

Empirical probability: A type of probability based on experimental or historical data

Empirical rule: A concept stating that the values on a normal curve are distributed in a regular pattern, based on their distance from the mean

F

False positive: A test result that indicates something is present when it really is not

I

Independent events: The concept that two events are independent if the occurrence of one event does not change the probability of the other event

Inferential statistics: A type of statistics that uses sample data to draw conclusions about a larger population

Interquartile range: The distance between the first quartile (Q1) and the third quartile (Q3)

Interval: A sample statistic plus or minus the margin of error

Interval estimate: A calculation that uses a range of values to estimate a population parameter

L

Literacy rate: The percentage of the population in a given age group that can read and write

Lower limit: When constructing an interval, the calculation of the sample means minus the margin of error

M

Margin of error: The maximum expected difference between a population parameter and a sample estimate

Mean: The average value in a dataset

Measure of central tendency: A value that represents the center of a dataset

Measure of dispersion: A value that represents the spread of a dataset, or the amount of variation in data points

Measure of position: A method by which the position of a value in relation to other values in a dataset is determined

Median: The middle value in a dataset

Method: A function that defines and performs behaviors like computation

Mode: The most frequently occurring value in a dataset

Multiplication rule (for independent events): The concept that if the events A and B are independent, then the probability of both A and B happening is the probability of A multiplied by the probability of B

Mutually exclusive: The concept that two events are mutually exclusive if they cannot occur at the same time

N

Non-probability sampling: A sampling method that is based on convenience or the personal preferences of the researcher, rather than random selection

Nonresponse bias: When certain groups of people are less likely to provide responses

Normal distribution: A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped

O

Objective probability: A type of probability based on statistics, experiments, and mathematical measurements

P

Parameter: A characteristic of a population

Percentile: The value below which a percentage of data falls

Point estimate: A calculation that uses a single value to estimate a population parameter

Poisson distribution: A probability distribution that models the probability that a certain number of events will occur during a specific time period

Population: Every possible element that a data professional is interested in measuring

Population proportion: The percentage of individuals or elements in a population that share a certain characteristic

Posterior probability: The updated probability of an event based on new data

Prior probability: The probability of an event before new data is collected

Probability: The branch of mathematics that deals with measuring and quantifying uncertainty

Probability distribution: A function that describes the likelihood of the possible outcomes of a random event

Probability sampling: A sampling method that uses random selection to generate a sample

Purposive sample: A method of non-probability sampling that involves researchers selecting participants based on the purpose of their study

Q

Quartile: A value that divides a dataset into four equal parts

R

Random experiment: A process whose outcome cannot be predicted with certainty

Random seed: A starting point for generating random numbers

Random variable: A variable that represents the values for the possible outcomes of a random event

Range: The difference between the largest and smallest value in a dataset

Representative sample: A sample that accurately reflects the characteristics of a population

S

Sample: A subset of a population

Sample size: The number of individuals or items chosen for a study or experiment

Sample space: The set of all possible values for a random variable

Sampling: The process of selecting a subset of data from a population

Sampling bias: When a sample is not representative of the population as a whole

Sampling distribution: A probability distribution of a sample statistic

Sampling frame: A list of all the items in a target population

Sampling variability: How much an estimate varies between samples

Sampling with replacement: When a population element can be selected more than one time

Sampling without replacement: When a population element can be selected only one time

Simple random sample: A probability sampling method in which every member of a population is selected randomly and has an equal chance of being chosen

Snowball sample: A method of non-probability sampling that involves researchers recruiting initial participants to be in a study and then asking them to recruit other people to participate in the study

Standard deviation: A statistic that calculates the typical distance of a data point from the mean of a dataset

Standard error: The standard deviation of a sample statistic

Standard error of the mean: The sample standard deviation divided by the square root of the sample size

Standard error of the proportion: The square root of the sample proportion times one minus the sample proportion divided by the sample size

Standardization: The process of putting different variables on the same scale

Statistic: A characteristic of a sample

Statistical significance: The claim that the results of a test or experiment are not explainable by chance alone

Statistics: The study of the collection, analysis, and interpretation of data

Stratified random sample: A probability sampling method that divides a population into groups and randomly selects some members from each group to be in the sample

Subjective probability: A type of probability based on personal feelings, experience, or judgment

Summary statistics: A method that summarizes data using a single number

Systematic random sample: A probability sampling method that puts every member of a population into an ordered sequence, chooses a random starting point in the sequence, and selects members for the sample at regular intervals

T

Target population: The complete set of elements that someone is interested in knowing more about

U

Undercoverage bias: When some members of a population are inadequately represented in a sample

Upper limit: When constructing an interval, the calculation of the sample means plus the margin of error

V

Variance: The average of the squared difference of each data point from the mean

Voluntary response sample: A method of non-probability sampling that consists of members of a population who volunteer to participate in a study

Z

Z-score: A measure of how many standard deviations below or above the population mean a data point is