COURSE GLOSSARY

Introduction to Statistics

Binomial distribution: A discrete distribution giving the probability of a specified number of successes in n independent trials with constant success probability p, with expected value n·p

Box plot: A compact visualization showing a dataset's median, first and third quartiles (the box), whiskers for range or variability, and individual outliers

Categorical (qualitative) data: Data that describe categories or groups, either nominal (unordered categories) or ordinal (ordered categories)

Central Limit Theorem: The principle that the sampling distribution of the sample mean (or many other summary statistics) approaches a normal distribution as sample size increases, given independent random sampling, typically effective for n ≥ 30

Conditional probability: The probability of an event occurring given that another event has already occurred, written P(A|B)

Correlation (Pearson): A numeric coefficient between –1 and 1 that measures the strength and direction of a linear relationship between two variables, where sign indicates direction and magnitude indicates strength

Descriptive statistics: Methods for summarizing or describing the main features of a dataset, such as measures of center and spread, without making inferences about a larger population

Independent event: An event whose probability is unaffected by the occurrence or outcome of another event

git init: A Git command used to create a new repository by initializing the .git directory in the current or specified folder

Inferential statistics: Techniques that use sample data to draw conclusions or make predictions about a larger population, typically involving uncertainty and probability

Interquartile range (IQR): The difference between the third quartile (75th percentile) and the first quartile (25th percentile), representing the spread of the middle 50% of values

Mean: The arithmetic average of a set of numerical values, calculated by summing all values and dividing by the count of values

Median: The middle value of an ordered numeric dataset, with half the observations below and half above, or the average of the two middle values when the count is even

Mode: The most frequently occurring value in a dataset, useful as a measure of center for categorical as well as numeric data

Normal distribution: A symmetric, bell-shaped continuous distribution fully described by its mean and standard deviation, with about 68%, 95%, and 99.7% of values within one, two, and three standard deviations of the mean respectively

Null hypothesis: The default assumption in hypothesis testing that there is no effect or no difference between populations, which statistical tests seek to challenge.

Numeric (quantitative) data: Data represented by numbers that measure quantity and can be continuous (any value on a scale) or discrete/count (whole numbers)

P-value: The probability, assuming the null hypothesis is true, of obtaining a result at least as extreme as the observed sample result, used to assess statistical evidence against the null

Probability distribution: A function or listing that assigns probabilities to all possible outcomes of a random process, discrete or continuous

Probability: A numerical measure between 0 and 1 (or 0%–100%) representing the chance that a specified event will occur

Sampling with replacement: A sampling method where selected items are returned to the population before the next draw, keeping probabilities constant across draws

Sampling without replacement: A sampling method where selected items are not returned to the population, so probabilities for subsequent draws change and events become dependent

Standard deviation: The square root of the variance that measures typical distance of data points from the mean in the original data units

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

Variance: The average of the squared differences between each data point and the mean, which quantifies overall data dispersion in squared units