COURSE GLOSSARY

Introduction to Statistics in Python

Binomial distribution: A discrete probability distribution that gives the probabilities of obtaining k successes in n independent trials with a constant success probability p per trial

categorical — nominal & ordinal): A classification of variables where numeric (quantitative) data represent measurable quantities (continuous can take any value in a range, discrete are countable), and categorical (qualitative) data represent group labels which may be unordered (nominal) or ordered (ordinal)

Central limit theorem: The theorem stating that the sampling distribution of the sample mean (and many other statistics) approaches a normal distribution as sample size or number of samples increases, provided samples are independent and identically distributed

Confounding: A situation in which a third variable (a confounder) influences both the explanatory and response variables, creating a spurious or biased association between them

Correlation coefficient: A numeric measure between –1 and 1 (commonly Pearson's r) that quantifies the strength and direction of a linear relationship between two numeric variables

Data types (numeric — continuous & discrete

Descriptive statistics: Methods and measures (like mean, median, counts, and plots) used to describe and summarize the characteristics of a specific dataset without making inferences beyond it

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

Inferential statistics: Techniques that use sample data to make estimates, test hypotheses, or draw conclusions about a larger population

Interquartile range (IQR): The difference between the 75th and 25th percentiles (Q3 – Q1), representing the range of the middle 50% of the data and used as a robust measure of spread

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

Median: The middle value of an ordered dataset such that half the observations are below and half are above, used as a robust measure of center when data are skewed

Mode: The value or category that occurs most frequently in a dataset, commonly used for categorical variables

Normal distribution: A continuous, symmetric, bell-shaped probability distribution fully described by its mean and standard deviation and commonly used as an approximation in many real-world contexts

Outlier: An observation that lies far outside the typical range of the data, often defined as below Q1 – 1.5×IQR or above Q3 + 1.5×IQR, and potentially indicative of error or unusual variation

Poisson distribution: A discrete distribution describing the probability of a given number of events occurring in a fixed interval of time or space when events occur independently at a constant average rate (lambda)

Population: The full set of individuals, items, or measurements about which you want to draw conclusions

Probability distribution: A function or listing that assigns probabilities to each possible outcome of a random process, describing how probability is distributed across outcomes for discrete or continuous variables

Probability: A numeric measure between 0 and 1 (or 0%–100%) that quantifies the chance that a particular event or outcome will occur

Sample: A subset of observations drawn from a larger population used to estimate properties of that population

Sampling with and without replacement: Two sampling schemes where with replacement returns an observation to the pool before the next draw (making draws independent), whereas without replacement removes it (making draws dependent)

Standard deviation: The square root of the variance that measures spread in the same units as the data and indicates how much observations typically deviate from the mean

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

Summary statistic: A single number that captures an important feature of a dataset (for example, an average, count, or percent) used to summarize the data

Variance: A measure of spread equal to the average squared distance of observations from the mean (usually computed with n–1 for a sample), with units that are the square of the original units