Classical & Baysian Statistics

Statistics: Discipline that concerns collection, organisation, analysis, interpretation, and presentation of data.

Applied statistics: Applying statistics to real everyday problems.

Classical Statistics: Set of tools for decision making using hypothesis (Null hypothesis significance testing, NHST).

Bayesian Statistics: Unified approach to statistics, which is the mathematical version of how we usually think and involves the reallocation of credibility across possibilities.

Models: Statements about the operation of nature that purposefully omit many details to simplify complex real-world phenomena and achieve insight.

Simulations: Useful tools, often relying on computer power, to approximate quantities for which an exact solution is very difficult or impossible to determine.

Exploratory Data Analysis: Also called descriptive statistics, it concerns the representation of datasets by characterizing them with numbers (e.g., average) and displaying them graphically.

Quantitative data: Data that can take, at least theoretically, any numerical value within an interval of the number line (measurements).

Qualitative data: Data that can only take a certain number of values, often called factors.

Location parameters: Key figures that address where the observations are located on the measuring scale.

Spread parameters: Key figures that describe how the data scatter around their central location.

Arithmetic mean: Colloquially known as the average or simply mean, it is the sum of all data divided by the number of data points, representing the center of the data.

Median: The value where half of the observations are below or equal to this value, and the other half is equal to or greater than this value. It is more robust than the mean because it is less influenced by extreme values.

Quantile: A generalization of quartiles to any other percentage; it is the value where a specified proportion of observations are less than or equal to it.

Lower quartile (1st quartile): The value where approximately 25% of all observations are less than or equal to this value, and 75% are greater than or equal to this value.

Upper quartile (3rd quartile): The value where approximately 75% of all observations are less than or equal to this value, and 25% are greater than or equal to this value.

Empirical variance: A measure of the variability or spread of observations. The value itself has no physical interpretation as its unit is the square of the data unit.

Empirical standard deviation: The root of the empirical variance, having the same unit as the original data, and representing the average deviation from the mean.

Interquartile range (IQR): A robust measure for data spread, calculated as the upper quartile minus the lower quartile, representing the length of the interval containing about half of the central observations.

Outlier: An observation in a boxplot that is more than one and a half times the interquartile range away from one of the two quartiles.

Histogram: A graphical overview of occurring values where the height of a bar for a class is proportional to the number of observations falling into that class.

Normalized Histogram: A histogram where the bar heights are scaled such that the total area of all bars equals one, meaning bar areas correspond to the proportion of observations in the respective classes.

Skewness (right skewed): A distribution where data is concentrated on the left and flattens out towards the right (the tail is on the right side).

Skewness (left skewed): A distribution where data is concentrated on the right and flattens out towards the left (the tail is on the left side).

Sample space: Contains all possible elementary events or outcomes of a random experiment.

Event: A subset of the sample space, consisting of one or more elementary events.

Stochastic independence: The condition where the outcome of one event has no influence on the outcome of another event.

Random variable: A function that assigns a numerical value to each elementary event of a random experiment.

Realization: A specific numerical value that a random variable takes when a random experiment is performed.

Probability distribution: The assignment of probabilities to all possible realizations of a random variable, often presented as a table for finite sample spaces.

Expected value: The theoretical central location of a distribution, calculated as the weighted mean of all possible values, weighted by their respective probability of occurring.

Standard deviation (Theoretical): The theoretical spread or dispersion of a distribution around its expected value.

Standard error of arithmetic mean: The standard deviation of the average of multiple observations.

Continuous random variable: A random variable whose values can theoretically take any value within a certain interval.

Probability density function: A function describing the probability distribution of a continuous random variable; probabilities correspond only to areas under this function, not the function's values themselves.

Normal distribution: A common continuous, symmetrical distribution, often referred to as a bell curve, characterized by its mean (where the peak is) and its standard deviation (which determines its shape).

Central Limit Theorem: States that the sum and the average of a sufficiently large number of independent and identically distributed random variables, regardless of the original distribution, will be approximately normally distributed.

Hypothesis testing: A standardized procedure to decide whether the observed data statistically supports or rejects a statement about a true population parameter (like the mean).

Null Hypothesis: The assumption that is tested, stating that a population parameter (e.g., the true mean) equals a specific value.

Alternative Hypothesis: The statement accepted if the Null Hypothesis is rejected, indicating a relationship or difference exists (e.g., the true mean is not equal to the specific value).

Significance Level: Denoted by α, it represents the risk one is willing to take of making a wrong decision (Type I error), typically 0.05 or 0.01.

Rejection Range: The area in the distribution of the test statistic where, if the sample result falls, the Null Hypothesis is rejected.

p-value: The probability of obtaining the observed result, or a more extreme one, assuming the Null Hypothesis is true. A small p-value provides strong evidence against the Null Hypothesis.

t-test: A hypothesis test used to determine if a sample mean differs significantly from a hypothesized population mean, without assuming the true standard deviation is known.

t-distribution: A distribution used in the t-test that is similar to the normal distribution but flatter, accounting for the additional uncertainty when the population standard deviation is unknown.

Confidence interval: An interval calculated from data that indicates where the true mean lies with a certain predefined probability (e.g., 95%).

Wilcoxon test: A non-parametric alternative to the t-test, often preferred when data is not normally distributed, assuming only symmetry around the median under the Null Hypothesis.

Paired Samples: Measurements where each observation in one group is directly connected or assigned to an observation in the other group (e.g., testing the same subject twice).

Unpaired (Independent) Samples: Measurements where observations in different groups are unconnected and independent.

Scatter plot: A graphical display used for two-dimensional data, showing observations as points on a coordinate system to visualize relationships between two variables.

Causality: A concept describing a proven cause-and-effect relationship, which cannot be established solely based on statistical dependence shown in a scatter plot.

Regression line: The straight line drawn in linear regression that minimizes the sum of squared residuals, defining the approximate linear relationship between variables.

Residual: The vertical distance between an observed data point and the corresponding point on the regression line.

Least squares method: The procedure used to estimate the coefficients of a regression model by minimizing the sum of the squared residuals.

Empirical correlation: A dimensionless number between negative one and positive one that numerically measures the strength and direction of the linear dependence between two variables.

Simple linear regression: A model that predicts a quantitative response variable based on a single predictor, assuming an approximately linear relationship.

Response variable: The quantitative output variable that is being predicted in a regression model.

Predictors (Explanatory variables): The input variables used to predict the response variable.

R-squared: A statistical measure between zero and one that indicates the proportion of the variability in the response variable that is explained by the model.

Multiple linear regression: A regression model used when the response variable depends on more than one predictor.

Interaction effect: A term in a regression model indicating that the relationship between one predictor and the response variable changes depending on the value of another predictor.

Dummy variable (Indicator variable): A numerical variable (typically 0 or 1) introduced to represent a qualitative predictor in a regression model.

Baseline (Qualitative Predictor): In qualitative predictors with more than two possible values, this is the level without a corresponding indicator variable, used as the reference point for interpretation.

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

Prior probability: In Bayesian inference, the initial probability or belief assigned to a parameter before observing any data.

Posterior probability: In Bayesian inference, the updated probability or belief about a parameter after incorporating the information from the observed data.

Likelihood function: In Bayesian inference, the function that describes how probable the observed data are for different possible values of the parameter.

Beta distribution: A continuous probability distribution defined between zero and one, frequently used as a prior and posterior distribution for parameters that represent probabilities (like the probability of tossing heads).

Highest Density Interval (HDI): A method for summarizing a distribution by specifying the narrowest interval that contains a specified percentage (e.g., 95%) of the most credible values of the distribution.

Markov Chain Monte Carlo (MCMC): A class of numerical algorithms used to generate random samples from complex distributions, especially used to approximate the posterior distribution in Bayesian analysis.

Metropolis Algorithm: A fundamental MCMC method that generates a representative sample from a target distribution via a random walk in the parameter space.

Region of Practical Equivalence (ROPE): A small range of parameter values surrounding the null value that is considered practically insignificant for the application's purpose, used in Bayesian hypothesis testing.

Computer Science Concept

Algorithm: An exact specification of how to solve a computational problem, requiring every step to be completely specified.

Application Layer: The highest layer in the Internet protocol stack, supporting network applications such as FTP, SMTP, or HTTP.

Binary Search: A search algorithm for a sorted list that compares the target value to the middle element, eliminates the half that cannot contain the target, and repeats the process on the remaining half until the target value is found.

Bit: The fundamental binary unit of data.

Brightness: A visual variable used to convey data, which is good for ordinal data but challenging for numerical data.

Byte: A unit of data consisting of 8 bits.

Client Server Communication Pattern: A model where clients send service requests and receive responses from a centralized server.

Cloud Computing: A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., servers, storage, applications, and services) that can be rapidly provisioned and released.

Color: A visual variable often used in visualizations, for example, to represent data bipolarly.

Computational Problem: Specifies an input-output relationship, defining the input characteristics and the corresponding required output.

CSS (Cascading Style Sheets): Used to define the layout and visual appearance of HTML elements, separate from the content structure.

cut: A utility command used in data processing to select and write specific portions (e.g., fields delimited by a character) of each line from a file to the output.

Divide and Conquer: An algorithm design technique where a large problem is broken up into smaller, easier sub-problems, relying on the solution of the smaller parts to solve the larger one.

Domain Name System (DNS): The system that translates domain names (like hslu.ch) into numerical IP addresses.

Efficient Algorithm: An algorithm designed to run as quickly as possible while using as little memory as possible.

Filter: In data processing, this takes a stream of text or data, performs operations on it, and produces a modified version of that stream as output.

grep: A filter command used to search files for lines matching a specified pattern.

Hosts/End Systems: Connected computing devices running network applications, often referring to client machines.

HTML (Hypertext Markup Language): A descriptive language used to create websites and publish content on the Internet in a simple and standardized way.

HTTP (Hypertext Transfer Protocol): The protocol that ensures content on a Web server can be accessed, interpreted by a Web client, and displayed within a browser.

Hybrid Cloud: A deployment model where the infrastructure is a composition of two or more clouds, which can be private or public, enabling data and application portability between them.

Infrastructure as a Service (IaaS): A cloud service model where the infrastructure layer, including servers, storage, and network, is provided.

IP Address: A unique numerical address assigned to each device in a computer network, consisting of a Network ID and a Host ID.

Linear Search (Sequential Search): A simple search method that sequentially checks each element of a list until a match is found or the entire list has been searched.

Link Layer: The network layer responsible for data transfer between neighboring network elements.

MAC Address: A 48-Bit address used by network switches to uniquely identify devices and support plug-and-play, self-learning functions.

Meta Data: Information about an HTML document, typically contained within the head section.

Meta Characters: Characters used in regular expressions to specify variable characters in a pattern.

Network Layer: The network layer responsible for routing packets from the source to the destination (e.g., using IP).

Objects (Thunkable): Data structures in visual programming that include multiple properties.

Orientation: A visual variable used in visualization, often mapping data values to the direction or rotation of a mark.

Overview first, zoom and filter, then details on demand: A design principle for visualization emphasizing that users should first gain a general overview before exploring specific data through interactive tools.

Packet: A smaller chunk of an application message, of length L bits, into which application messages are broken down for transmission.

Packet Switches: Network devices, such as routers and switches, that forward data packets.

Paragraphs (p): HTML elements used to create blocks of text content.

Pay by use: A key benefit for cloud clients, meaning they pay only for the resources they actually consume instead of provisioning hardware for peak usage.

Physical Layer: The lowest network layer, dealing with the transmission of bits on the wire.

Pipe (|): A mechanism in Linux used to send the output of one command or process to another command or process for further processing.

Position: Considered the most powerful visual variable, using the location of an object in a plot or diagram to encode data.

Private Cloud: A deployment model where the cloud infrastructure belongs to and is operated by only one organization.

Propagation Delay: The delay related to the length of the physical link and the speed at which signals travel through the medium.

Protocol (Communication Protocol): Rules defining the format and order of messages sent and received among network entities, and the resulting actions taken upon transmission or receipt.

Public Cloud: Clouds owned by service providers (e.g., AWS, Azure) who charge organizations for the use of resources.

Recursion: A problem-solving method where the solution depends on solutions to smaller instances of the same problem, technically involving methods that call themselves.

Regular Expressions (Regex): A powerful means for pattern matching, used to match strings that share common characteristics.

Router: A network device that forwards packets based on IP addresses using a routing algorithm and a forwarding table.

Size (Length, Area, Volume): A visual variable used to encode data, although the perception of area versus length can make accurate decoding difficult.

Software as a Service (SaaS): A cloud service model where the application (packaged software) is provided.

Switch: A network device that uses MAC addresses and a switch table to learn the location of senders for forwarding frames.

Throughput: The rate (bits per time unit) at which bits are transferred between a sender and receiver.

Transmission Delay: The time needed to transmit an L-bit packet into a link, calculated by dividing the packet length (L) by the link transmission rate (R).

Transport Layer: The network layer responsible for process-to-process data transfer, including identifying lost or damaged packets.

URL (Uniform Resource Locator): A reference to a web resource (like a web page or image) that specifies its location in a computer network.

Value (Big Data): One of the 5 Vs, referring to the value or utility that the data generates.

Variables (Thunkable): Elements used in visual programming to store and manage data like numbers or texts.

Variety (Big Data): One of the 5 Vs, referring to the different types of data, such as structured, unstructured, or semistructured data.

Velocity (Big Data): One of the 5 Vs, referring to data generated at a high speed.

Veracity (Big Data): One of the 5 Vs, referring to the correctness or accuracy of the data.

Visual Variables: Components of visual encoding such as Position, Size, Color, and Shape used to represent data features graphically.

Visualization: The compact graphical presentation and user interface for manipulating large numbers of items, used to enable users to make discoveries or decisions.

Volume (Big Data): One of the 5 Vs, referring to the sheer size of the data.

^: Beginning of the line.

$: End of the line.

|: A pattern alternative.

.: An arbitrary character.

: Quote of the directly followed meta character for using it as a normal character.

*: Placeholder for 0 or more appearances of the directly preceding regular expression.

+: Placeholder for one or multiple appearances of the directly preceding regular expression.

[ ]: Defines a character class; matches exactly one of the mentioned characters.

( ): Grouping.

{n}: Exactly n times.

{n,m}: Between n and m times.

\w: Matches any word character (a-z, A-Z, 0-9, _).

\W: Matches any non-word character.

\s: Matches any whitespace.

\S: Matches any non-whitespace.

\d: Matches any digit (0-9).

\D: Matches any non-digit.

Design of experiment

Abduction: The process of knowledge gain that starts with data, where a new explanatory theory is formed by a sudden mental leap, suggesting that something may be.

Activity bias I: A source of bias in social media where data is collected only from users who are active at the time of the study or who are active only once or a few times.

Activity bias II: A source of bias in social media where a few users are very active, while most users use social media only passively, causing the data obtained to often relate to a specific topic or person or be limited to a certain period.

Alternative hypothesis (HA): The research hypothesis to be tested that postulates the presence of a certain effect, such as a difference, in the population.

ANCOVA (Analysis of Covariance): An analysis procedure used when one or more metric variables (covariates) are included in an ANOVA model alongside categorical factors.

ANOVA (Analysis of Variance): Statistical procedures developed by R. A. Fisher for analyzing data and testing structures, where the variance of one or more dependent variables is explained by the impact of one or more factors.

A/B testing: An experiment, also known as bucket testing or split-run testing, often used in web environments to compare two or more versions of a website or app in parallel to determine which achieves a better result, such as a higher click-through rate.

Arbitrary sampling: A non-probabilistic sampling procedure where selection is not based on a random mechanism but by decisions, leading to an unknown selection probability and often being used in populations that are difficult to find, such as in snowball sampling.

Balancing: Refers to the balance in the combination of factor levels in experimental designs.

Bandit algorithms: Advanced methods in A/B testing that allow several treatments to be tested dynamically and simultaneously, drawing faster conclusions than conventional study designs by dynamically allocating more data traffic to the variants showing the highest success.

Bias: A deviation between the true mean in the population and the mean estimated from the sample; in social media, it means the population being examined does not correspond to the defined population.

Blinding: A technique used in experiments to avoid distortions or changes in behavior caused by knowledge about the treatment, ensuring the comparability of groups.

Blocking: Arranging experimental units into homogeneous groups (blocks) based on one or more influential nuisance variables in order to reduce variability (secondary variance).

Case Control Study: A type of observational study that examines how cases and controls differ regarding their previous exposure to a factor.

Causality: A cause-and-effect relationship discovered in a process through a trial or experiment, implying that a manipulation of a particular factor results in a specific outcome.

Census: A full survey that examines all cases associated with the population, requiring high effort and rarely used in empirical social research.

Click-through rate (CTR): A metric defined as the ratio of users who click on a specific link to the number of total users who view a page, email, or advertisement, commonly used to measure the success of online advertising.

Cluster sampling: A probabilistic sampling procedure where elements are selected at a higher level (cluster) to solve logistical problems, and all units within the selected clusters are surveyed.

Cohort study: An observational study in which two or more groups of the same cohort are observed over a defined period to compare groups exposed to an influencing factor with those who are not.

Computational Science (Third Paradigm): The third paradigm of science, focusing on the simulation of complex phenomena.

Confounding: Occurs when a factor (confounder) that is not directly investigated is associated with both the independent variable and the dependent variable, thus causing the observed relationship between the two variables.

Construct validity: The effectiveness of the measurement methods in precisely capturing the intended theoretical construct, which is addressed through operationalization.

Controllable factors: Influencing factors whose strength can be adjusted within defined limits, synonymous with Independent Variables (IV).

Control group: The group in an experiment that is not exposed to the condition hypothesized to have some causal effect, serving as a baseline for comparison.

Conversion rate: The percentage of visitors who complete a set goal (e.g., sign-up or purchase), calculated as the total number of conversions divided by the total number of visitors.

Correlational Study: An observational study used to determine whether and how two or more variables are related, typically collected at a single point in time and then correlated.

Counterbalancing: A technique used in studies with repeated measures to ensure that each stage of the treatment occurs equally frequently at each stage of the study, thus eliminating carry-over effects like learning or fatigue.

Covariate adjustment: A method for controlling secondary variance by including nuisance variables as covariates in the statistical model to account for their effects.

Coverage error: A form of non-sampling error where part of the population cannot be identified, resulting in undercoverage, or non-members of the population are included, resulting in overcoverage.

Critical rationalism: A position in scientific theory (Karl Popper) holding that findings are derived deductively based on observations, and a theory can never be finally verified, only falsified.

Cross Sectional Study: A descriptive observational study where variables of interest are collected and described at a specific point in time, typically via a survey.

Deduction/Deductive approach: The process of drawing conclusions from theory and applying them to empirical data to test or falsify hypotheses (conclusion from the general to the specific).

Dependent variable (DV): The input changed by the process; the result of the test or experiment, also referred to as output, target variable, or target value.

Descriptive statistics: Methods limited to describing the data within a sample without allowing for conclusions to be drawn about the entire population.

Design of Experiments (DoE): A strategic plan determining the methodological design of a study, including the definition of experimental conditions, sampling procedures, sample size, and statistical analysis methods.

Design weight: Weighting factors used in statistical analyses to adjust data to more accurately represent the population by compensating for inequalities in sampling design, such as uneven selection probability.

Double-blind: A stage of blinding where both the test persons and the persons carrying out the tests have no knowledge of group membership.

Effect size (ES): A number measuring the strength of the relationship between variables in a statistical population or a sample, indicating the importance or strength of the study results.

Empirical research: Research where knowledge can be gained only through observation, experiment, and experience, based on the Latin word for experience.

Error variance: Variation in the output caused by measurement errors and random processes.

Experimental Science (First Paradigm): The first science paradigm, focusing on the description of natural phenomena.

Experimental group: The group in an experiment that is exposed to the condition hypothesized to have some causal effect.

Experiment (Procedure): A procedure carried out to support, refute, or validate a hypothesis, providing insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated.

External validity: Exists when experimental results from a sample can be generalized to the entire population, increasing with the increasing naturalness of the study environment (Field over Laboratory).

Factorial design (Full): An experimental design where all possible combinations of factor levels are varied to determine all main effects and interactions.

Falsification: The process in critical rationalism where, if derived hypotheses are refuted by data, the existing theory is challenged and must be revised.

Found data (Administrative data): Data obtained administratively and technically, typically collected for documentary and administrative purposes rather than directly for research.

Fractional factorial design: An experimental design where only a selected, balanced part of the possible combinations of factors are varied to reduce effort, meaning interactions can only be partially measured.

Greco-Latin square: An extension of the Latin square design that allows for the study of four factors.

Independent variable (IV): Influencing factors whose strength can be adjusted within defined limits, also referred to as controllable factors.

Inductive approach/Induction: The process of drawing conclusions from specific empirical findings and applying them to general scientific theories (conclusion from the specific to the general), often aimed at arriving at and confirming new theories.

Inferential statistics: A set of methods for drawing conclusions about the population based on information obtained from a sample, typically using statistical hypothesis tests.

Interaction: Occurs in experiments with two or more independent variables when the effect of one factor depends on the levels of the other factor.

Internal validity: Exists when changes observed in the dependent variables (DV) can be definitively attributed to the manipulation of the independent variables (IV), increasing with decreasing impact of nuisance variables.

Key performance indicator (KPI): A quantifiable metric calculated from measured values that shows how effectively the most important company goals are achieved.

Latin square: A fractional factorial design that reduces the required sample size compared to full factorial designs by studying the main effects of factors without observing all combinations of treatment levels, provided no interactions are present.

Longitudinal studies: Studies that determine developments and changes over time, classified either as trend studies (different samples) or panel studies (same sample).

Made data: Data generated by researchers, typically as part of a study, with the primary goal of answering research questions.

Mean-squared error (MSE): A measure of the deviation (bias) of the estimator (sample mean) from the true mean in the population.

Measurement: The method for obtaining one or more measured values that can be assigned to a quantity.

Methodology: The framework focusing on the underlying considerations, decisions, and justifications of the approach used in scientific research projects to gain knowledge.

Multicollinearity: A statistical problem occurring when the independent variables correlate strongly, leading to inaccurate standard errors and inconsistent coefficient estimates.

Multifactorial (designs): Experimental designs where several independent variables (factors) with two or more levels each act on a single dependent variable (Analysis: Multi-factorial ANOVA).

Multi-stage sampling: A combination of single-level sampling procedures using different selection units.

Non-controllable factors: Influencing factors whose strength cannot be determined, but that may be measured (e.g., body weight) or whose strength cannot be determined and that cannot be measured (e.g., random fluctuations), also called nuisance variables.

Non-probabilistic (non-random, purposive, arbitrary) sampling procedures: Selection of elements not based on a random mechanism, resulting in an unknown or zero selection probability for individual elements, and forbidding conclusions to be drawn about the population.

Non-sampling error: The difference in the mean value between the defined ideal population and the real population that cannot be attributed to deficiencies in random selection.

Null hypothesis (H0): Postulates the opposite of the alternative hypothesis, namely the absence of an effect.

Nuisance variables/factors: Influencing factors that are not controllable and cause secondary variance, not being the main focus of the study.

Objectivity (Measurement): A property of a measurement instrument given when the results are independent of the personnel conducting the measurement and the calculation methods used.

Observation: A result of study in which a variable is measured.

Observational study: A research design where the treatment or exposure is not planned or manipulated, suitable for forming hypotheses but typically not for postulating causality.

OFAT (One-factor-at-a-time): An experimental design procedure where one factor is varied per run while all others are held constant; easy to implement but cannot recognize interactions.

Open (blinding): The absence of blinding in a study.

Operationalization: A process that defines and specifies subsequent research steps to capture a theoretical construct (like intelligence) by translating it into measurable, observable elements.

Oversampling: The process of intentionally increasing the proportion of elements from a small stratum in the sample to improve the quality of the estimate for that subgroup.

p-hacking: The unethical practice of analyzing data in multiple ways or seeking specific data subsets until a statistically significant p-value (e.g., below 5%) is obtained.

Panel study: A type of longitudinal study conducted at different times using the same sample.

Placebo: A solution or substance given to the control group that has the same appearance as the treatment but is designed to have no effect on the outcome, crucial for blinding.

Population: The set of all potentially explorable elements that share a common characteristic or combination of characteristics, defined before the research begins using geographical, temporal, and factual criteria.

Population validity: The degree to which the results of a study can be generalized from the chosen sample to the whole population.

Power analysis: A statistical technique used to determine the minimum sample size required for a study to detect a certain effect size at a specified significance level and statistical power.

Primary variance: The systematic impact of experimental factors on the change or variation of the output being examined, desirable for the study.

Probabilistic (random) sampling procedures: Selecting elements based on a random mechanism where the selection probability of each element is known and non-zero, permitting valid generalizations to the population.

Quasi-experimental study (Non-randomized controlled): A study where the groups (treatment and control) are not determined by a random process, often due to self-selection or participant preference.

Qualitative research: Empirical research where knowledge is gained through observation, resulting in subjective and interpretive findings (e.g., interviews or case studies).

Quantitative methods/research: Empirical research designed around the principles of critical rationalism, employing systematized procedures for obtaining objective, functionalistic knowledge (e.g., surveys or experiments).

Randomized Controlled Trial (RCT): The highest quality experimental design where subjects are randomly allocated to treatment and control groups, making it highly suitable for hypothesis testing and postulating causality.

Randomization: The random assignment of trial objects to treatment and control groups, aiming to eliminate selection bias and confounding, and ensure group comparability at the start of the study.

Reliability (Measurement): The degree to which a measurement instrument consistently produces the same result each time under comparable conditions.

Repetition: A technique for controlling secondary variance or minimizing error variance by repeating measurements on the same trial objects.

Representativeness: A non-statistical concept referring to a sample accurately reflecting the relevant aspects and key characteristics of the population.

Research design: A strategic plan setting out the broad outline and key features of a research project, including methods of data collection and analysis.

Research methods: Systematized procedures and approaches used for obtaining knowledge.

Scientific theory: A branch of philosophy dealing with the epistemology of scientific knowledge, scientific methods, and research, examining how scientific knowledge can be obtained.

Secondary variance: Variation of the output caused by nuisance variables, representing undesirable, non-systematic influences in the experiment.

Selection Bias: An experimental error where the participant pool is not representative of the target population, often arising from parental consent procedures or self-selection.

Selection error: A form of sampling error where not all elements of the population have the same selection probability.

Simple random sampling (SRS): A probabilistic sampling procedure where elements are randomly selected from the population, ensuring each element has the same probability of being included.

Single-blind: A stage of blinding where the test persons have no knowledge of their group membership.

Situation validity: The degree to which the findings of a study can be applied to different situations, such as different geographical locations or temperatures.

Snowball sampling: An arbitrary sampling technique that uses social networks to recruit participants, particularly effective for populations that are difficult to reach.

Spurious Correlation (Big Data context): Strong, meaningless correlations that arise in large data sets characterized by high dimensionality.

Standard error (of the sample mean): A measure quantifying the variability of sample means among many repeated random samples of the same size, indicating the quality of the estimated mean.

Stepped wedge design: A cluster-randomized study design in which different groups (clusters) switch from a control condition to the intervention at different, staggered times.

Stratified random sampling: A probabilistic sampling procedure where the population is first divided into strata based on known characteristics, followed by random sampling within each stratum (proportional or disproportional).

Systematic non-response: A form of non-sampling error where certain individual elements are systematically missing information because individuals refuse to answer questions.

Theoretical Science (Second Paradigm): The second paradigm of science, focusing on modeling and generalization, such as Newton's Laws.

Total survey error: A concept describing statistical properties of survey estimates, incorporating various error sources like measurement error, coverage error, and sampling error.

Treatment: A term used in experimental design referring to any prescribed combination of values of explanatory or independent variables.

Trend study: A type of longitudinal study conducted at different times using different samples.

Trial and error: A general, unsystematic DoE approach involving changing many factors per run, often lacking reproducibility and scientific rigor.

Triple-blind: A stage of blinding where the test persons, the persons carrying out the tests, and the people doing the analysis all have no knowledge of group membership.

Unbiased estimator: A sample statistic (like the sample mean) that estimates the true mean value in the population without systematic deviation.

Validity (Measurement): The extent to which a measurement instrument actually measures what it was intended to measure, considered the most comprehensive criterion.

Variance (Descriptive statistics): A key figure of a sample (sample variance) that describes the mean square deviation of the individual measured values from the empirical mean.

Volume (Big Data): One of the V dimensions of Big Data, referring to the sheer size or cardinality of the data set, often measured by the relation of the sample size to the population size.

Linear Algebra 1

Linear Algebra: The theory of linear mappings between vector spaces.

Matrix: A rectangular set of ordered elements composed of rows and columns.

Vector: A collection of data, often represented as a vertical bunch of numbers, which is characterized by magnitude and direction.

Scalar: A quantity without direction, which is simply a real number.

Magnitude: The length of a vector, also referred to as the norm.

Unit Vector: A vector whose length is exactly one.

Normalization: The process of obtaining a unit vector that shares the same direction as a given non-zero vector.

Zero Vector: A vector with a norm of zero, which possesses no direction.

Opposite Vector: A vector that has the same magnitude as a given vector but points in the opposite direction.

Collinear: Describes two vectors that share the same or opposite direction, meaning one is a scalar multiple of the other, also termed linearly dependent.

Linearly Independent: Describes vectors that are not collinear or parallel.

Components: The unique scalar values that determine a vector in a coordinate system.

Coordinates: The values that specify the location of a point.

Position Vector: A vector defined by having its initial point at the origin and its terminal point at a given point.

Linear System: A set of one or more linear equations.

Coefficient Matrix: The matrix that contains only the coefficients of the variables in a linear system.

Augmented Matrix: The matrix representation of a linear system formed by the coefficient matrix combined with the column vector of constant terms.

Row Echelon Form: A simplified structure of a matrix, often triangular in shape, resulting from Gaussian elimination, used to easily find solutions.

Pivot Element: A non-zero element in a matrix, typically the first non-zero entry in a row of a matrix in row echelon form, used to eliminate coefficients in other rows.

Backward Substitution: The procedure used to solve a linear system that is in triangular form by solving for variables starting from the last equation.

Consistent System: A linear system that possesses at least one solution.

Inconsistent System: A linear system that has no solutions.

Homogeneous System: A linear system where the column vector of constant terms is the zero vector.

Trivial Solution: The solution where all variables are zero, which is always a solution for a homogeneous system.

Rank: The number of pivot elements found in the row echelon form of a matrix.

Nullity: The number of columns in the row echelon form of a matrix that do not contain a pivot element.

Free Variables: Variables associated with columns lacking a pivot element, whose values can be chosen arbitrarily to generate solutions.

Square Matrix: A matrix having an equal number of rows and columns.

Identity Matrix: A square matrix with ones on its main diagonal and zeros elsewhere, serving as the neutral element in matrix multiplication.

Main Diagonal: The entries of a square matrix where the row index equals the column index.

Upper Triangular Matrix: A square matrix in which all entries located below the main diagonal are zero.

Lower Triangular Matrix: A square matrix in which all entries located above the main diagonal are zero.

Transpose: The resulting matrix obtained by swapping the row and column indices of every entry in the original matrix.

Symmetric Matrix: A square matrix that is equal to its own transpose.

Linear Combination: A sum formed by multiplying multiple matrices or vectors by scalars.

Matrix Multiplication: An operation defined by the row-column product, where the resulting entry is the sum of products of corresponding elements from a row of the first matrix and a column of the second matrix.

Commutativity: The property that matrix multiplication generally does not possess, meaning the order of the matrices matters.

Elementary Matrix: A matrix produced by applying a single elementary row operation to the identity matrix.

Row Reduction: The process of using elementary row operations to bring a matrix into row echelon form.

PLU Decomposition: The factorization of a square matrix A into the product of a permutation matrix P, a lower triangular matrix L, and an upper triangular matrix U.

Permutation Matrix: A matrix obtained solely by interchanging rows of the identity matrix, used to account for row swaps in factorization.

Invertible Matrix: A square matrix for which a unique inverse matrix exists, also known as a regular or non-singular matrix.

Inverse Matrix: The unique matrix which, when multiplied by the original invertible matrix, yields the identity matrix.

Singular Matrix: A square matrix that does not have an inverse.

Determinant: A single real number associated with a square matrix that provides a criterion for whether the matrix is invertible, and which can be interpreted geometrically as an oriented volume.

Ill Posed Problem: A mathematical problem in which minor inaccuracies in the input data result in significantly large errors in the output solution.

Python

Abstract Class: A class that provides basic functionalities and from which other classes inherit, but which cannot be instantiated directly.

Access Modifier: The state of visibility for declared variables and methods within a class (public, protected, or private), defining what attributes can be accessed.

And: A boolean operation where the result is True if and only if both operands are True.

Args (*args): The packing operator used in function definitions to accept an arbitrary number of non-keyword (positional) arguments, which are packed into a tuple.

Assert Statement: A debugging tool that checks a condition and raises an AssertionError if the condition is False. It is mainly used for sanity checks during development and testing.

Axes Object: A component of Matplotlib that represents a single or a set of plots within a figure, defining the area where data is plotted.

Base Class: The parent class in an inheritance hierarchy from which other classes (child/derived classes) inherit attributes and methods.

Binary Mode ('b'): A file handling mode used for non-textual files (e.g., images, videos).

Boolean (bool): A basic data type that represents truth values, typically True or False.

Boolean Operations: Operations used to combine two Boolean values, such as and, or, not, and xor.

Break: A keyword used to immediately terminate and exit the current loop (for or while).

Breakpoint: A defined code location where the execution shall stop during debugging.

Buffer (Writing): An intermediate storage location for data being written to a file, which is packaged and written to the file only when the stream is flushed or closed.

Built-in Namespace: One of Python's three namespaces, encompassing objects not defined by the programmer, such as built-in functions like range, def, or len.

Camel Case: A naming style where multiple words are joined without spaces, and each word starts with a capital letter (e.g., helloWorld); generally not conventionally used for Python variables.

Child Class: A derived class that inherits attributes and methods from a parent class.

Class: The assembly instructions or template for constructing objects, describing the properties and methods shared by all instances of that class.

Class Header: The starting declaration of a class definition, introduced by the keyword class and terminated with a colon.

Class Method: A method (function) that belongs to a particular class and typically modifies the class's state. It is defined using the @classmethod decorator.

Class Variable: Variables belonging to the class definition itself, storing information superior to all instances of the class.

Compiler Language: A type of language where the source code is compiled, and an executable file is returned. It allows for code optimization and usually results in compressed executables that are hard to reverse engineer.

Complex Numbers: A basic data type representing numbers with a real and imaginary part (e.g., 5 + 3j).

Concatenation: An operation, often represented by the plus operator (+), used for combining elements of sequences (like strings or lists), typically requiring elements of the same data type.

Conditional Breakpoint: A breakpoint that only stops execution if a specified condition evaluates to True.

Constructor (init()): A special method within a class that is called implicitly by Python when a new object is generated, defining the initial properties (object variables) of the new object.

Continue: A keyword used to ignore the remaining instructions in the current loop body and immediately jump back to the loop header for the next iteration.

Data Frame: A powerful and efficient tabular data structure, provided by the Pandas library, used for data analysis and manipulation.

Data Type: The kind of value a variable holds (e.g., Integer, String, Boolean). Python assumes the type from the assigned value.

Decorator: A function that wraps another function to extend its functionality, often applied using the @ symbol before the function definition.

Deep Copy: Creating a copy of an object and recursively creating copies of all nested objects within it, ensuring the original object remains completely unaffected by changes to the copy.

Debugging: The process of reproducing errors, finding the underlying bug, determining the root cause, fixing the bug, and validating the fix.

Dictionary: A mutable, ordered collection of data stored as unique key-value pairs. Keys are typically strings.

Dictionary Comprehension: A concise Python syntax for creating dictionaries, requiring the expression result to be in the format of column-separated key-value pairs (key:value).

Elif: A short form for "else if," used to check an additional condition if the preceding if or elif conditions were False.

Enum Data Type: A set of symbolic names bound to unique values, offering improved readability, grouping, and type-safety compared to using raw values.

Enumerate: A built-in function that returns an iterator, providing a tuple containing an incrementing number (index) and the element for each item in a sequence.

Eval() Function: A function that allows evaluation of a Python expression provided as a string input.

Except: The block in a try-except clause that defines how to handle a specific error type that was raised during the execution of the try block.

Execution Flow Control: Keywords (like break, continue, pass) used to manage the flow of a loop or function process.

External Package: Python packages hosted in external repositories, such as PyPI, which can be installed using tools like PIP.

F-String (Formatted String Literal): A convenient Python syntax for string formatting, prefixed with f or F, allowing variables and expressions to be included directly within the string using curly brackets.

File Handling: The process of interacting with files to store data outside the application (e.g., reading or writing data).

File Object: An object created by the open() function, used to access and manipulate the content of a file.

Finally: The optional block in a try-except clause that is always executed, regardless of whether an exception was raised or handled.

Figure Object: A component of Matplotlib that represents the entire canvas or window in which one or more axes (plots) can be drawn.

Float (Floating Point Number): A basic data type representing decimal numbers.

Flush: An operation that forcibly writes the contents of the internal write buffer to the file stream.

For Loop: A control structure used to iterate over sequence objects, accessing and processing each element in a sequential forward order.

Format() Method (String): A string method providing a flexible way for variable alignment within a string using placeholders {}.

Function: A block of code combining multiple instructions that can be executed repeatedly, promoting modularity and readability.

Functional Style: A programming style that discourages functions with side effects, meaning the output should primarily depend only on the input parameters.

Generator: A special function or expression that uses the yield keyword instead of return, returning an iterator object that generates values one at a time.

Global Namespace: One of Python's three namespaces, containing programmer-defined objects that are available and accessible across the entire script (module).

Global Variable: A variable that resides in the global scope and is accessible throughout the whole script.

If-Else Condition: A fundamental control structure that allows the execution of expressions or blocks based on a Boolean condition (True or False).

If __name__ == "__main__": A conditional construct used to ensure that a block of code is only executed when the script is run directly, and not when it is imported as a module into another script.

Import: The action of accessing definitions, variables, functions, or classes contained within another Python module or package.

Indentation: Whitespace used in Python to define the structural boundaries of code blocks (like loops, conditions, or functions), usually following a colon (:).

Inheritance: A concept in OOP where classes are arranged hierarchically, allowing a child class to automatically acquire the attributes and methods of its parent class.

Input Function: A function that reads data from the standard input (e.g., the terminal) and returns it as a string.

Integer (int): A basic data type representing whole numbers.

Integer Division (//): A mathematical operator that calculates how many times the divisor fits into the dividend, returning only the integer part of the quotient.

Intelligent Development Environment (IDE): Software that provides comprehensive facilities for programmers, typically including tools like a debugger, code editor, and version control interface.

Interpreter Language (Script Language): A type of language where code instructions are sent line by line to the interpreter and executed immediately, often lacking advanced compilation optimization.

Iterator: An object used to iterate over sequence elements, managing the process of retrieving elements one after the other.

Jupyter Notebook: An open-source web application widely used for data analysis, allowing documents to be created and shared, containing interactive code, equations, visualizations, and narrative text (Markdown).

Kwargs (**kwargs): The double packing operator used in function definitions to accept an arbitrary number of keyword arguments, which are packed into a dictionary.

Lambda Function: A small, anonymous (unnamed) inline function comprising a single expression, often used as an input argument for other functions (like map or sort).

Len() Function: A built-in function that returns the number of elements in a sequence, considering only the first dimension.

List: A fundamental sequence data type in Python that is mutable (changeable), ordered, and can contain elements of different data types, including nested lists.

List Comprehension: A concise syntax for creating a list based on an iterable, often including conditional logic or complex expressions, providing a shorter alternative to explicit for loops.

Local Namespace: One of Python's three namespaces, containing programmer-defined objects (such as variables or parameters) that are only visible within a specific function block.

Local Variable: A variable defined within a function block and only visible within that block.

Map Function: A built-in function that applies a given function to every item of an iterable and returns a map iterator.

Match-Case: A control structure used to check an expression against multiple discrete patterns or conditions (cases), supporting a default case defined by an underscore (_).

Matplotlib: A comprehensive library for creating static, interactive, and animated visualizations in Python.

Method Resolution Order (MRO): The rule set (an algorithm) Python uses to determine the search order for inherited methods and attributes in classes that use multiple inheritance.

Module: A file (with the .py extension) containing Python definitions and statements, used to organize and structure source code.

Modulo (%): A mathematical operator that returns the remainder of an integer division.

Multidimensional Array Object (ndarray): The primary object type provided by NumPy, representing fixed-size sequences, matrices, or multi-dimensional matrices of numerical values of the same data type.

Multiple Inheritance: A form of inheritance where a child class inherits from more than one parent class.

Mutable: A characteristic of an object whose value can be changed after it is created (e.g., lists and dictionaries).

Namespace: The mechanism defining the visibility and uniqueness of a name for every object within a specific context.

Nested List: A list that contains other lists or data structures as its elements.

Nonlocal Keyword: A keyword used within a nested function's local scope to refer to and potentially modify a variable defined in the closest enclosing (parental) local scope, provided that variable is not global.

Not: A Boolean operation that inverts the logical state of its operand (True becomes False, False becomes True).

NumPy: The fundamental Python package for scientific computing, providing efficient operations on large, homogeneous, multidimensional numerical arrays.

Object: An instance of a class; in Python, all data structures (like lists, strings, and integers) are considered objects.

Object Identity (is): A comparison operator that checks if two variables refer to the exact same object in memory (same storage location).

Object Method: A function defined within a class that belongs to a particular object instance and requires the self parameter as a reference to that instance.

Object-Oriented Programming (OOP): A programming paradigm focused on organizing code around objects (instances of classes) rather than functions and logic.

Object Variable: Variables belonging to a specific object instance, defining its unique properties.

Open() Function: A built-in function used for file handling to create a file object that links to a file using a specified file name and mode (e.g., 'r' for reading).

Or: A Boolean operation where the result is True if at least one of the operands is True.

Override: The process where a child class redefines a variable or method inherited from a parent class, using the exact same name.

Package: A system folder that is recognized by Python as a module collection, typically denoted by the presence of an __init__.py script within the folder.

Packing Operator (*): The asterisk operator used to group multiple items into a single sequence (e.g., a tuple).

Pandas: A powerful and widely used Python library providing data manipulation and analysis tools, centered around the DataFrame object.

Parameter: A variable defined in the function header used to receive input values when the function is called. In Python, parameters are passed by reference.

Pass: A null operation (no operation) used as a placeholder where syntax requires a statement but no action is desired (e.g., in an empty loop or function block).

PIP: The Package Installer for Python, a command-line tool used to install and manage external Python packages, typically from PyPI.

Print Function: A built-in function used to send all passed objects to the standard output (e.g., the terminal).

Private Attribute: An attribute intended to be accessed only from within its own class, conventionally indicated by two leading underscores (__).

Property (@property): A decorator used to define a method that can be accessed like an attribute, often employed for encapsulation to control how an object's variables are accessed or modified.

Protected Attribute: An attribute intended to be accessed only from within its own class or any subclasses derived from it, conventionally indicated by a single leading underscore (_).

Public Attribute: An attribute that can be accessed freely by any part of the program using the object reference.

PyCharm: A full-featured IDE developed by JetBrains, popular among Python developers, offering advanced features like debugging and code analysis.

PyPI: The Python Package Index, the official third-party software repository for Python.

Pyplot: A collection of functions within the Matplotlib library used for creating figures and plotting areas.

Python Console: The direct interface to the Python interpreter, allowing immediate command execution and result return.

Python Script: A collection of Python commands stored in a file (usually .py) that are executed consecutively by the interpreter.

Range() Object: A built-in class that efficiently represents an immutable sequence of numbers, defined by a start, stop, and step value, used frequently for controlling loop iterations.

Read() (File Method): A file method that reads a specified quantity of data (in bytes) or the entire content of the file if no size is specified.

Recursion: A programming technique where a function calls itself, often used to solve complex problems or iterate through data structures with undefined dimensions.

Reference: A pointer to the storage location of an object in memory, which is what variables hold and what is passed to functions as parameters.

Return Value: The value produced by a function and sent back to the caller. All Python functions return at least None.

Scope: The region of the code from which a specific object (variable, function, etc.) is accessible.

Seaborn: A high-level Python library built on top of Matplotlib, specifically designed for statistical data visualization, offering built-in themes and color palettes.

Self Attribute: The first parameter of instance methods (by convention named self), which acts as a required reference to the specific object instance the method belongs to.

Sequence: A general term for ordered, indexable data structures in Python, such as lists, tuples, and strings.

Sequence Slicing: The technique of extracting a sub-sequence from an existing sequence (like a list or string) by specifying start, stop, and step indices.

Set: An unordered collection data type that stores only unique and hashable (immutable) elements.

Shallow Copy: Creating a copy of an object where only the main container (e.g., the outer list) gets a new reference, but references to nested mutable objects (e.g., inner lists) are still shared with the original object.

Shorthand If-Statement: A compact conditional syntax used when an expression needs to be evaluated based on a single condition, often used for assignment or immediate expression results.

Static Method: A method defined within a class that is independent of both the object instance and the class state; it relies solely on its input parameters.

String (str): A basic data type representing sequences of text characters, enclosed in single or double quotation marks.

String Formatting: The process of defining the representation, alignment, and inclusion of variables within a string.

Syntax Error: An error resulting from violating the grammatical rules of the programming language (e.g., missing parentheses).

Terminal: A command-line interface window, often included in IDEs, used for interacting with the operating system or executing scripts.

Try-Except: A control structure used for robust error handling, allowing the program to attempt executing "risky" code and gracefully react if a specified exception is raised.

Tuple: A sequence data type similar to a list but immutable (unchangeable) and ordered.

Type Annotation: A convention (not enforced by Python) used to indicate the expected data types for variables, function parameters, and return values, aiding readability and third-party type checkers.

Unpacking Operator (*, **): Operators used to pass elements of a sequence (tuple or list with *) or key-value pairs of a dictionary (with **) as separate, individual arguments to a function.

Variable: A symbolic name used as a reference to a particular storage location holding a value.

Virtual Environment (Venv): An isolated environment for a Python project that manages its own dependencies and keeps the development setup clean, avoiding dependency conflicts.

While Loop: A control structure that executes a block of code repeatedly as long as a specified condition remains True.

Write() (File Method): A file method that sends content (a string) to the file object's internal buffer for eventual writing to the file.

Xor (Exclusive Or): A Boolean operation where the result is True if and only if exactly one of the operands is True.