Statistics

Regression

Clustering and Dimensionality Reduction

Data

Data Fundamentals

- **Data**: units of qualitative or quantitative information about persons or objects collected via observation.
  - Note: data is different from information—information resolves uncertainty, while data has the potential to be transformed into information post-analysis.
  - Data as a general concept refers to the fact that some existing information or knowledge can be represented in a form suitable for processing.

Data Types

- Data types have two different general meanings:
  - **Data type (computer science)**: involves the format of data storage and has implications on operations and storage space.
  - **Data type (statistics)**: involves the category of data and has implications on the methods used for analysis.
- There are many data types, with more specific definitions than the following definitions, but for now these are frequently used and adequate for topics covered.

Relevant Statistical Data Types

Category	Type	Description	Example
Numerical	**Interval**	Degree of difference	Temperature °C
	Ratio	Interval + meaningful zero	Height
	Discrete	Count (integers)	Population
Categorical	**Ordinal**	Sortable, discrete	Educational level
	Nominal	Non-sortable, discrete	Movie genre

Population vs. Sample Data

- **Population data** μ: data from all members of a group.
- **Sample data** $\hat{\mu}$: data from a subset of members of a group (hopefully random).
- Statistical procedures generally are designed for sample or population data; wrong conclusions can be drawn if the distinction is not clear.
  - Note: most data are sample data in practice, as generalization of populations using sample data is usually the goal of inferential statistics.
- **Anecdotes**: a case study of a rare occurrence, or a sample size of only one; insights may be possible, but poor confidence in ability to generalize should be noted.

Data Visualization

- **Data visualization**: a mapping between the original data and graphic elements in order to determine how attributes of interest vary according to the data.
  - The design of the mapping can have a significant effect on information extracted from data, in both beneficial and detrimental ways.
- Data visualization is a core tool of statistics and generally considered to be a branch descriptive statistics↓; more techniques will be covered in that chapter.

Primer: Visualization Techniques

- Visualizing data can be an art in and of itself, leading to a wide variety of available techniques, i.e., diagram types, in order to better represent the data.
- The following is a rather shallow list of commonly used techniques; in-depth exploration of data visualization will be pursued in other courses.
- **Bar chart**: a representation of categorical data with magnitudes proportional to the values they represent.
  - Displays comparisons among discrete categories vs. a measured value.
  - Subcategories can be displayed in clusters within each category, with colors/patterns used to differentiate them.
  - Ordering of the categories (chart shape) do not typically matter, excluding aesthetic reasons.
- **Histogram**: a representation of the distribution of numerical data via the use of binning.
  - **Binning**: a form quantization of continuous data, wherein small intervals (bins) of the data are replaced with a value representative of that interval.
  - The bins are usually specified as consecutive, non-overlapping intervals of a variable; they must be adjacent and are often of equal size.
  - Histograms of counts are usually better for qualitative inspection of raw data, but can be difficult to compared across data sets.
  - Histograms of proportion are usually better for quantitative analysis, as they are typically easier to compare across data sets, but can take extra effort to create.
- **Scatter plot**: a representation of the relationship between variables, often two or three (2D/3D graphs).
  - Points can be coded via color, shape, and/or size to display additional variables.
  - Often used to investigate correlations between variables.

- **Network graph**: a representation of data as nodes in a network via analysis of specialization of the nodes.
  - Used to discover bridges (information brokers) in a network, relative node influence, and outliers via analysis of how the nodes cluster.
  - Node and tie (connection between nodes) size and color can be used to encode additional information about variables in the data.
- **Pie chart**: a representation of one categorical variable via the division of slices in order to illustrate numerical proportion.
- **Box plot**: a representation of numerical data via analysis of their quartiles.
  - **Quartiles**: a quantile (division point) of data points into four parts, or quarters.
    - Q_1: the middle number between the smallest minimum and the median of the data set; 25% of the data lies below this point.
    - Q_2: the median of the data set; 50% of the data lies below this point.
    - Q_3: the middle value between the medium and the maximum of the data set; 75% of the data lies below this point.
  - Often termed box and whisker plot, as the box represents the 50% of the data, and the two whiskers represent the upper and lower 25% of data.
    - **Interquartile range IQR**: the box, i.e., the difference between upper and lower quartiles; $IQR = Q_3 - Q_1$.
  - Outliers may be plotted as individual points.
  - Useful when examining the variability of samples without making any assumptions about underlying statistical distributions.

Descriptive Statistics

Descriptive Statistics Fundamentals

Descriptive vs. Inferential Statistics

- **Descriptive statistics**: the processes of using and analyzing summary statistics that quantitatively describes or summarizes features of a collection of information.
  - Methods/measures of descriptive statistics:
    - Distribution shape↓
    - Mean, median, mode↓
    - Variance↓
    - Kurtosis, skew↓
  - No relation to population.
  - No generalization to other data sets.
  - Concerned only with properties of observed data.
- **Inferential statistics**: the process data analysis to deduce properties of an underlying probability distribution.
  - Methods/measures of inferential statistics:
    - Probability theory↓
    - Hypothesis testing↓
    - Confidence intervals↓
    - And essentially all of applied statistics.
  - Assumes that the observed data set is sampled from a larger population.
  - Entire purpose is to generalize/relate features to other data sets.

Accuracy, Precision, Resolution

- **Accuracy**: the relationship between the measurement and the actual truth.
  - Inversely related to bias; colloquially interchangeable with accuracy.
- **Precision**: the certainty of each measurement.
  - Inversely related to variance↓
- **Resolution**: the number of data points per unit measurement (e.g., time, space, individual, etc).
- Generally, the goal is accuracy → precision → resolution, but often choice in the matter is not so deliberate.

Primer: Probability Distributions

- The shapes of data distributions are functions of probability theory↓; a more in-depth explanation will be covered later, but for now coverage of common distribution types might be useful.
- Overall, there is one major distinction of distribution type based on data types↑ used, either discrete or continuous.
- **Discrete distribution**:
  - Deals with events that occur in countable sample spaces; contains finite number of outcomes.
  - Summation of values can be done to estimate probability of an interval.
  - Expressed with graphs, piece-wise functions, or tables.
  - Expected values might not be achievable.
  - Common examples:
    - Bernoulli : a model for the set of possible outcomes of any single binary experiment.
    - Binomial : a sequence of n independent Bernoulli experiments; a basis for the binomial test.
    - Uniform : a known, finite number of values are equally likely to be observed.
    - Poisson : a sequence of independent events over a specified interval with a known constant mean rate.
- **Continuous distribution**:
  - Deals with events that occur in a continuous sample space; contains infinitely many consecutive values.
  - Summation of values in order to determine probability of interval not possible; integrals used instead.
  - Expressed with continuous functions or graphs.
  - Common examples:
    - Normal (Gaussian) : used to represent real-valued random variables who are not known.
    - Chi-Squared : the sum of squares of k independent standard normal random variables.
    - Lognormal : distribution of a random variable whose logarithm is normally distributed.
    - Student's t : estimations of the mean using small sample sizes with unknown standard deviations.
- Wikipedia's list of probability distributions

Descriptive Techniques

Measures of Central Tendency

- **Mean** $\bar{x}$: the sum of all measurements x_i divided by the number n of observations in the data set x, i.e.,

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$$

  - Suitable for roughly normally distributed data of continuous data types.

- **Median** $\text{med}(x)$: the middle value of the data, i.e.,

$$x_i, \quad i = \frac{n+1}{2}$$

  - Suitable for unimodal distributions of continuous data types.
  - Odd number of observations with no distinct middle value are usually defined as the mean of the two middle values.

- **Mode**: most common value.
  - Suitable for any discrete distribution, usually used for nominal data types.

Measures of Dispersion

- **Dispersion**: the measure of how distributed, or deviated, data are around a central value.
- **Variance** σ^2, s^2: the primary measure of dispersion, or more explicitly, the expectation of the squared deviation of a random variable from its mean, i.e.,

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  - Suitable for any distribution; better for normally distributed data.
  - Mean centering, i.e., $(x_i - \bar{x})$, is done to capture the dispersion around the average, but not the magnitude of the values themselves.
  - The sum of a mean-centered data set would be zero, thus it is squared.
    - **Mean absolute difference (MAD)**: when the absolute value of mean-centered data is taken instead of the square value.
    - MAD is more robust to outliers, but further from Euclidean distance and less commonly used.
  - Division by $n - 1$ is used for sample variance, as often sample sizes can be small and are considered empirical quantities; n^{-1} is used for population variance (a theoretical quantity).
  - **Standard Deviation** σ: simply the square root of variance, $\sqrt{\sigma^2}$

Statistical Moments

- **Moments**: a quantitative measure related to shape of a functions graph; relates to physics and statistics.
  - Regarding probability distributions, the general formula can be defined as:

$$m_k = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

  - Increments of k define particular moments, i.e.,
    - First moment $k = 1$: expected value, or mean$^{\uparrow}$.
    - Second moment $k = 2$: central moment, or variance$^{\uparrow}$.
    - Third moment $k = 3$: dispersion asymmetry, or skewness.
    - Fourth moment $k = 4$: tail "thickness," or kurtosis.
    - Further moments are possible, but useful applications are less common.
  - **Skewness**: a measure of asymmetry of a probability distribution of a real-valued random variable about its mean.
    - Can be positive, zero, negative, or undefined.
    - Negative skew: an indication that the tail is on the left.
    - Zero skew: an indication that tails balance out; can be true for both asymmetric and symmetric distributions depending on kurtosis.
    - Positive skew: an indication that the tail is on the right.
  - **Kurtosis**: a measure of the thickness/curvature of the tail of a probability distribution is; an indication of deviation/outliers.
    - Univariate normal distributions have a kurtosis of 3, leading to a common basis.
    - Platykurtic < 3: a term for low kurtosis, indicating that a lesser degree of deviations or outliers is observed.
    - Leptokurtic > 3: a term for high kurtosis, indicating that a greater degree of deviations or outliers is observed.
    - **Excess kurtosis**: kurtosis minus 3, often colloquially termed as kurtosis; an indication a greater degree outliers compared to a normal distribution.

Visualizations Revisited

- **Q-Q (quantile-quantile) plot**: a graphical method for comparing two probability distributions by plotting their quantiles against each other.
  - **Quantile**: cut points dividing the range of probability distributions into continuous intervals with equal probabilities, e.g.,
    - Percentiles: 0–100
    - Quartiles: 0–4
    - Quantiles: 0–x
  - The points of similar distributions will lie approximately on the line $y = x$;
  - However, other linear relations are possible, meaning points may not necessarily lie on the line $y = x$.
  - Provides a mean for comparing location, scale, and skewness of similarities of differences in two distributions.
- **Histogram bin number** k: the is no "best" number of bins, different bin sizes can reveal different features of the data, but there are several methods of determining k;
  - Determination via suggested bin width h:

$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil$$

  - Sturges' formula: derived from binomial distribution; assumes approximately normal distribution:

$$k = \lceil \log_2(n) \rceil$$

  - Freedman-Diaconis' rule: method of determining h using interquartile range (IQR); often method of choice:

$$h = 2\frac{\mathrm{IQR}(x)}{\sqrt[3]{n}}$$

  - Arbitrary ≈ 42: often intuitive guesses are sufficient and yield useable results:
- **Violin plot**: similar to a box plot, but rotated with addition of a kernel density plot on each side.
  - **Kernel density plot**: essentially a smoothing estimation based on finite data samples.
  - Statistical and IQR moments can be conveniently shown, sometimes with asymmetric comparisons of similar data sets (rather than a mirrored version).

Introduction to Normalization

- **Normalization of ratings (feature scaling)**: adjusting values measured on different scales to notionally common scale, often prior to averaging.
  - Often in more complicated cases, the adjustments are meant to bring the entire probability distribution of adjusted values into alignment.
- **Normalized values (normalization)**: creation of dimensionless, or scaled, versions of samples with the intention of minimizing the effect of gross anomalies/outliers↓.
- There are many types of normalization techniques in statistics, each with their own respective applications based on data types and distribution shapes.
  - For now, only standard score and min-max scaling will be covered, with others introduced at more appropriate times.

Z-Score Standardization

- **Z-score (standard score)**: the number of standard deviations σ↑ by which the value of a raw score x_i is above or below the mean $\bar{x}$↑, i.e.,

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

- Application of z-normalization is best done on data that is roughly Gaussian↑.
- The z-score is dimensionless, as units cancel out, leading to main application wherein data of different scales can be meaningfully compared.

Min-Max Scaling

- **Rescaling (min-max normalization)** x': the simplest method of rescaling the range of features, either from $[0, 1]$ or $[-1, 1]$; the general formula for $[0, 1]$ is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Rescaling to any arbitrary range $[a, b]$:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Outliers

- **Outlier**: a data point that differs significantly from other observations, potentially due to a variety of reasons either, due to the cause of experimental error in observations, random noise, unexplained/surprising phenomena, or simply by natural variability.
- Outliers can cause serious errors in statistical analysis, as many methods square terms, leading to potentially huge errors.
  - Often extremely detrimental impacts on small sample sizes are observed, as significance of the outliers decrease with increasing sample size.
- **Leverage**: a measure of how far away the independent variable values of an observation are from those of other observations.
  - Outliers are worse near the "edges" of the data, compared to the "middle," as outliers further away increase the leverage.
  - Lower leverage has less influence on statistical analysis, and in particular, it is a large factor in regression analysis↓.
- There are two main strategies for dealing with outliers, either:
  - Identify and remove outliers prior to analysis; assuming outliers are noise or invalid.
  - Keep outliers in and use robust methods that attenuate the negative impact of outliers; assume outliers are unusual but valid.
    - Robust methods of retention will be examined when more appropriate.
- Despite strategy chosen, outliers ought to be investigated; sometimes outliers might be an important aspect of the data.

Removing Outliers

- There are many methods of removing outliers, here use of the z-score↑ is explained. Again, more in-depth examinations of methods will be examined when appropriate.
- First, data must be converted to a normalized metric, e.g., the z-score.
- Next, a threshold must be determined that marks data points for suspect, dealing with them either methods of truncation or winsorization.
  - **Truncation (trimming)**: complete removal, with possible replacement of NaN placeholder to maintain indexing.
  - **Winsorization (clipping)**: replacement outlier with the nearest or a less suspect "alternative" value.
  - A variety of methods of determining such threshold can be used, even such methods lead to potentially arbitrary choices; 3 is often a default starting point.

- Finally, suspect data are dealt with iteratively until no other data pass the given threshold.
- Note, the z-score is generally only useful for roughly Gaussian distributions↑, however, a modified z-score using the median can be applied for non-normal distributions, i.e.,

$$z_i = \frac{0.6745(x_i - \text{med}(x))}{\text{med}(|x_i - \text{med}(x)|)}$$

  - 0.6745 is a normalization factor equal the standard deviation units of Q_3↑ of a Gaussian distribution.
- Deletion of data is generally avoided, with only clear indications of measurement error being the reason to do so.
- Multivariate data sets are dealt in similar way, where the only difference is that the mean of the data set is taken by calculating the Euclidean distance between all points in the set, then applying the method(s) described above.

Probability Theory

Probability Fundamentals

- **Probability**: a measure of the likelihood that an event will occur; used to quantify attitudes towards propositions whose truth are not certain.
  - Quantitatively, probability is a number between 0 and 1, which is often expressed as a percentage.
- **Probability theory**: the axiomatic formalization of probability; widely used in many fields of study from math to philosophy.
- **Probability space** $(\Omega, \mathcal{F}, P)$: a formal construct consisting of three elements that provides a model for a random process.
  - **Sample space** Ω: the set of all possible outcomes.
  - **Event space** $\mathcal{F}$: all sets of outcomes; all subsets of the sample space.
  - **Probability function** $P(E \in \mathcal{F})$: the assignment of a number between 0 and 1 that represents the probability of each event E in event space.
- **Proportion**: the measure of certainty; a fraction of a whole or the relation between two varying quantities.
  - Proportion *could* involve random variables, so depending on how the question is asked, then proportion could be the same as probability, but ultimately they are not interchangeable.
- **Odds**: the ratio of the number of events that produce an outcome to the number of events that do not; essentially probability reframed in potentially more efficient way.

Probability Theory Axioms

- **First axiom**: the probability of an event is a non-negative number real number, i.e.,

$$P(E) \in \mathbb{R}, \quad P(E) \geq 0 \qquad \forall E \in \mathcal{F}$$

- **Second axiom**: the assumption of unit measure; the probability that at least one elementary event in the entire sample space will occur is 1, i.e.,

$$P(\Omega) = 1$$

- **Third axiom**: the assumption of σ-additivity, wherein any countable sequence of disjoint sets[↓] $E_1, E_2, \ldots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

  - Thus, only discrete data[↑] are valid for probability; continuous data must be converted to discrete forms in order to be valid.

Independent and Mutually Exclusive Events

- **Stochastically independent**: when an event does not affect the probability of another, i.e.,

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$$

  - Two random variables are independent if the realization of one does not affect the probability distribution of the other.
  - **Pairwise independent (weak notion)**: two specific events in a collection that are independent of each other.
  - **Mutually independent (strong notion)**: when each event is independent of any combination of other events in the collection.
  - Often the stronger notion is simply termed independence, as it implies the weaker version, but not the other way around.
- **Mutually exclusive (disjoint)**: two events that cannot occur at the same time, i.e.,
  - Probability of both:

$$P(A \text{ and } B) = P(A \cap B) = 0$$

  - Probability of either: ↓

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$$

- **Collectively exhaustive (jointly)**: when at least one event must occur while exhausting all other possibilities at a given time, or that their union must coverall the events within the entire sample space, i.e.,

$$A \cup B = \Omega$$

Primer: Conditional Probability

- **Conditional probability**: the probability of some event A, given $|$ the occurrence of some other event B, i.e.,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

  - Note, $P(A \mid B)$ typically differs from $P(B \mid A)$, falsely equating the two often results in errors, termed the base rate fallacy.
- **Bayes' theorem**: probability of an event based on prior knowledge of conditions that might be related to the event; inference using conditional probability, i.e.,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- More on Bayesian statistics may or not be explored in greater depth in this course.

Probability Functions

- Again, a probability function↑ is the assignment of a number of probability space↑ (sometimes denoted X, $\mathcal{A}$, P, respectively).
- **Probability distributions**: the product of variations of the probability function based on given event space and properties of data types.
  - As mentioned in the primer↑ to this topic, probability distributions are generally divided into two classes based on data, i.e., either discrete or continuous.
- **Probability mass function (PMF)**: a function that gives the probability that a discrete random variable is exactly equal to some value.
  - The function $p : \mathbb{R} \to [0, 1]$ is defined formally as:

$$p(x_i) = P(X = x_i) \qquad -\infty < x < \infty$$

  - The associated probability values must follow the Kolmogorov axioms↑, which means all possible values must be positive and sum up to 1, implying all other probabilities must be 0, i.e.,

$$p(x_i) > 0, \qquad \sum p(x_i) = 1, \qquad p(x) = 0 \quad \forall x \neg x_i$$

  - Thinking of probability as mass helps avoid mistakes since physical mass is conserved, as is total probability for all hypothetical outcomes of x.
  - Major associated distributions include Bernoulli and Binomial↑ distributions, but geometric distributions deserve a mention as well;
    - Geometric distribution : a description of the number of trials/failures needed to get to one/first success.
- **Probability density function (PDF)**: a function that describes relative probabilities for a set of exclusive↑ continuous events, i.e.,

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

- **Cumulative density function (CDF)**: the PDF can also be described as the cumulative sum of continuous probabilities up to a particular point, i.e.,

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \quad \text{or (in practice)} \quad C(x_a) = \sum_{i=1}^{a} p(x_i)$$

  - Note: every CDF is non-decreasing and right-continuous
  - Note: the sum of CDF is > 1.

Sampling

- **Sampling distribution**: the probability distribution of a given random variable when derived from a random sample size n.
  - Useful to be considered as the statistic for all possible samples from the same population of a given sample size; is dependent on the underlying distribution of the population.
- Sampling a subset is often an easier and faster way to estimate an entire population, providing a potentially major simplification to statistical inference.
- **Expected (mean) value** μ, $E[X]$, $\bar{X}$: the expected mean of the population↑, or in case of random sampling, the expected mean of numerous samples, i.e.,

$$\bar{X} = \sum_{i=1}^{k} x_i p_i$$

  - Where X is a random variable with a finite number outcomes x_i occurring with respective probabilities p_i.
  - Thus, the expected value is the weighted sum, with the probabilities as weights.
- **Sampling variability**: different samples from the same population can have different values of the same measurement.
  - A single measurement may be an unreliable estimate of a population parameter.
  - Potential to randomly select outliers are the main source of sampling variability, but cannot be avoided.
  - Thus, natural variation, measurement noise, and failing to under complexity of phenomena are all sources of variability.
- **Sampling frame**: the source material, data, or device from which a sample of drawn; ideal frame qualities include:
  - The units have logical, numerical identifiers.
  - The units can be found again, or resampled.
  - The frame is organized, systematically.
  - The frame has additional information about the units and for the potential use of more advanced sampling frames.
  - Every element of the population of interest is present.
  - Every element of the population is only present once.
  - No elements outside the population of interest are present.
  - The data is keep up to date, accepting new information.

Sampling Methods

- **Probability sample**: a sample wherein every unit in the population has a chance ($P > 0$) of being selected in the sample, and the probability can be accurately determined.
- There are numerous methods of sampling, not all of which will be covered, but the various ways have the following two things in common:
  - Every element has a known nonzero probability of being sampled.
  - Involves random selection at some point.
- Factors that contribute to choice between methods:
  - Nature, quality, and availability of auxiliary information of the data.
  - Accuracy requirements, and need to measure accuracy.
  - Degree of expected analysis, cost, and operational concerns.
- **Simple random sampling**: all subsets of a sampling frame have an equal probability of being selected.
  - Minimizes bias, simplifies analysis.
  - Variance between individual results within the sample is a good indicator of variance of overall population, leading to easy estimations of accuracy.
  - Subject to sampling error, and implicit bias can go unnoticed due to data collection methods.
- **Systematic (interval) sampling**: method of arranging the study population according to an ordering scheme, then selecting the starting element randomly and progressing at a specified interval.
  - Useful if the arrangement of the data correlated with the variable of interest.
  - Some arrangements can introduce periodic biases, potentially leading to samples unrepresentative of the overall population.
  - Can be hard to quantify the accuracy, even if it can be more accurate and efficient than simple random sampling.
- **Stratified sampling**: organization of data into discrete categories, or "strata" where each stratum is treated like an independent population and randomly sampled from.
  - Helps avoid errors due to methods of data collection that may lead to subpopulations being overrepresented, causing to inaccurate generalizations if combined into one population.
  - Can be expensive, hard to select for relevant stratification variables, and is not useful when no homogenous subgroups.

- Other (less common?) methods of sampling include: probability proportional to size sampling, cluster sampling, multistage sampling, quota sampling, voluntary sampling, snowball sampling, accidental sampling, and panel sampling.
- **Monte Carlo methods**: a broad class of computational algorithms that rely on repeated random sampling.
  - Relies on properties of randomness to solves difficult problems that are deterministic in principle but not necessarily in practice.
  - Used in optimization functions, numerical integration, and generation draws from a probability distribution.
  - Might be covered later, and probably will be included under Bayesian statistics, if that is covered in-depth.

Law of Large Numbers and Central Limit Theorem

- **Law of large numbers (LLN)**: describes the result of performing the same experiment many times, wherein the average↑ approaches the expected value↑.
  - There is a weak strong law, that essentially state the same thing, but with slight difference; the strong law contains a more elaborate, but not covered proof.
- **Weak law of large numbers**: with a sufficiently large sample size, then for any nonzero margin specified there will be a high probability that the average observations fall within the margin, i.e.,

$$\lim_{n\to\infty} P\left(|\bar{x}_n - \bar{X}| > \epsilon\right) = 0 \qquad \epsilon = x \in \mathbb{R} \mid x > 0$$

- **Strong law of large numbers**: as $n \to \infty$, then the probability that the average converges to the expected value is 1, i.e.,

$$P\left(\lim_{n\to\infty} \bar{x}_n = \bar{X}\right) = 1$$

- Essentially, this implies that the mean of sample means is often a useful estimate of a population mean. This helps lead to the central limit theorem, which is a critical bridge between classical and modern probability theory.
- **Central limit theorem (CLT)**: when independence random variables are added, then their properly normalized sum tends to converge towards a normal distribution even if the original variables are not normally distributed.
  - **"All roads lead to Gauss:"** another way of stating the CLT, i.e., the distribution of sample means approaches a Gaussian distribution, regardless of the shape of the population distribution.

Hypothesis Testing

Hypothesis Testing Fundamentals

- Reviewing dependent and independent variables (parameters):
  - **Dependent variable** y: the variable you are trying to explain; the output of a function.
  - **Independent variables** x_n: the variables that potentially explain the dependent variable; the input(s) to a function.
  - Often the assumptions about the relationship can effect what is assumed to be the independent and dependent variables; interpretations can be difficult.
- **Models**: a simplified system made of the composition of concepts which are used to help know, understand, or simulate a subject the model represents.
  - **Residual (error)** ϵ: the degree that features not explained by variables that make up the composition of models.
  - Residuals should be small (accurate), but models should also be simple (useful); finding the balance between these two goals is a major part of statistics/science.
- **Alternative (effect) hypothesis** H_a: a proposed explanation for a phenomenon;a falsifiable claim that requires verification, typically from experimental data, and that allows for predictions about future observations.
  - Most formal hypotheses connect concepts by specifying the expected relationships between propositions, leading to expected differences.
  - Hypothesis testing is used to develop better theories via the rejection of previous theories; most progress in science is the result of hypothesis testing.
  - A strong hypothesis is:
    - Falsifiable—ideally testable, makes a criticizable prediction.
    - Parsimonious—limits excessive entities; application of "Occam's razor."
    - Scoped—clear, specific, applicable; a statement, not a question.
    - Fruitful—may explain further phenomena, aids in understanding.
- **Null hypothesis** H_0: the default hypothesis that a quantity to be measure is zero.
  - Typically, a quantity being measure is the difference between two situations, thus support for the alternative hypothesis is gained via rejection of the null hypothesis.
  - Testing the null hypothesis is a central task in hypothesis testing and the modern practice of science; weak evidence fails to reject the null hypothesis.
  - Criteria for excluding the null hypothesis will be covered in more depth when discussing confidence intervals↓.

Basis of Inferential Statistics

- Essentially, the basis of inferential statistics relies on the comparison between sample distributions↑ under the null and alternative hypotheses.
- In most cases, population data↑ is not attainable, instead, use of the central limit theorem↑ allows for the expected value↑ to be found via use of repeated sampling.
  - H_0 **distribution**: the distribution created due to sampling variability↑ under the null hypothesis, i.e., the differences between the expected mean value and sampled mean value, centered around 0.
    - Results from a formula based on assumptions, degrees of freedom↓, and type/nature of particular tests being performed.
  - H_a **distribution**: the distribution of differences due to the alternative hypothesis, rejection of the null hypothesis is likely to occur if observations reflect this distribution and not the H_0 distribution.
    - Results from empirical observations, gathered data and sampling methods↑.
- Quantifying the differences between the H_0 and H_a requires normalization, i.e.,

$$\frac{\text{Difference of centers}}{\text{Widths of distributions}} = \frac{\text{Central Tendency}}{\text{Dispersion}} = \frac{\text{Signal}}{\text{Noise}}$$

- The investigation of the ratio between signal-to-noise is essentially all of inferential statistics; fitting data into workable frameworks contains the majority of the work.

P-Value

- ***p*-value**: the probability↑ of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypotheses is correct, i.e.,
  - How likely is the H_a value to occur if H_0 is correct?
  - What is the probability of observing a parameter estimate of H_a or larger, given there is no true effect?

$$p(H_a \mid H_0)$$

  - Small p-value $\rightarrow$ outcome is very unlikely to occur under the null hypothesis.
- **Significance level** α: the somewhat arbitrary threshold whereby a study would reject the null hypothesis, typically $\alpha \leq 0.05, 0.01,$ or 0.001
- **Statistically significant**: when $p \leq \alpha$; significance can have other interpretations↓.
- Either side of a distribution is unlikely; two-tailed distributions need to split α.
  - Hypotheses should aim to be one-tailed, but this is often not feasible.
- p-values are often misinterpreted, sometimes even intentionally abused, and an important topic in metascience.

- Common misinterpretations of *p*-values:
  - ✘ Incorrect:
    - *"My p-value is 0.02, so the effect is present for 2% of the population."*
    - *"My p-value is 0.02, so there is a 98% chance that my sample statistic equals the population parameter."*
    - *"My p-value is smaller than the threshold, therefore the effect is real."*
  - ✓ Correct:
    - "My *p*-value is 0.02, therefore there is a 2% chance that there is no effect and my sample statistic was due to sampling variability, noise, small sample size, and/or systematic bias."
- Recall that the z-score$^{\uparrow}$ is a dimensionless measure of standard deviations $\sigma_x{}^{\uparrow}$ from the mean; the relation between *p*- and z-values can be useful to memorize.
- Given a Gaussian distribution, z-proportion (above/below) values are:
  - **68.3%** of the data are within $\sigma_1 \leftrightarrow$ **z =** $\pm\mathbf{1} = \mathbf{0.683}$
  - **95.5%** of the data are within $\sigma_2 \leftrightarrow$ **z =** $\pm\mathbf{2} = \mathbf{0.955}$
  - **99.7%** of the data are within $\sigma_3 \leftrightarrow$ **z =** $\pm\mathbf{3} = \mathbf{0.997}$
- Common *p*-values parings with standard deviations:

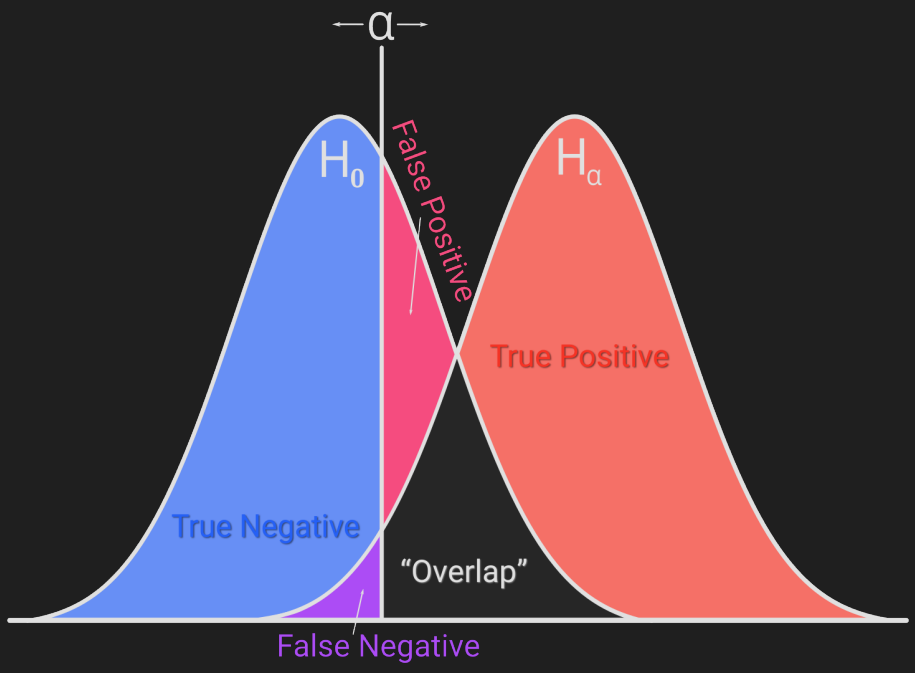
One-tailed ↓	↓ Two-tailed ↓
$p = \mathbf{0.05} \leftrightarrow z = \mathbf{1.64}$	$p = \mathbf{0.05} \leftrightarrow z = \mathbf{1.96}$
$p = \mathbf{0.01} \leftrightarrow z = \mathbf{2.32}$	$p = \mathbf{0.01} \leftrightarrow z = \mathbf{2.58}$
$p = \mathbf{0.001} \leftrightarrow z = \mathbf{3.09}$	$p = \mathbf{0.001} \leftrightarrow z = \mathbf{3.29}$

Degrees of Freedom

- **Degrees of freedom (**df, ν**)**: the number of values (parameters) in the final calculation of a statistic that are free to vary.
  - I.e., the minimum number of independent coordinates that can specify the position of the system completely.
- Degrees of freedom determine the shape of H_0 distributions (often the width).
- Higher degrees of freedom generally indicate more power to reject$^{\downarrow}$ the H_0.
- Can be useful metric for quickly determining relevant accuracy and understanding of experimental designs.
- Generally, $\nu = n - k$; with n data points and k parameters.

Statistical Errors

- **False positive (type I error)** $p = \alpha$: an **incorrect rejection** of a true H_0.
  - **True positive** $p = 1 - \beta$: a **correct rejection** of a false H_0.
- **False negative (type II error)** $p = \beta$: an **incorrect non-rejection** of a false H_0.
  - **True negative** $p = 1 - \alpha$: a **correct non-rejection** of a true H_0.
- **"Overlap"**: the area shared between the H_0 and the H_a.
  - Adjustments to the significance level α can bias towards/away from either false negatives/positives, at the cost of increasing the other.
  - Sometimes one error is more costly than the other, however, changing α is a less than ideal way generally arbitrary way to minimize error.
- The best way to minimize error is to minimize signal-to-noise, i.e.,
  - Increase distance between distributions (bigger effects)
  - Decrease the width of the distributions (less variability).



Interpretations of Significance

- **Statistical significance**: the probability of observing a test statistic of a certain magnitude given the H_0 is true.
- **Theoretical significance**: a finding that is relevant for a theory or leads to a new experiment; not directly related to statistical significance.
- **Clinical (practical, societal, educational)**: a finding is relevant for application in a particular field of interest.

Testing Properties

Parametric vs. Nonparametric

- **Parametric statistics**: based on the assumptions wherein the sample data originates from a population that can be adequately modeled by a probability distribution with a fixed set of parameters.
- **Nonparametric statistics**: based on relaxed assumptions surrounding of parametric tests, e.g., underlying distribution less important, presence of outliers, or lower specificity of parameters.
- Generally, there is a nonparametric test related to each parametric test, with particular assumptions relaxed, e.g.,

Parametric	Nonparametric
1-sample t-test	Wilcoxon sign-rank test
2-sample t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
ANOVA	Kruskal-Wallis test

- Important applications of nonparametric statistics with no direct correlate involve permutation testing, cross-validation, and bootstrapping.
- ✓Advantages and ✘ limitations (sometimes) of **parametric statistics**:
  - ✓ Standard, widely used
  - ✓ Computationally efficient/simple
  - ✓ Analytically proven
  - ✘ Based on assumptions
  - ✘ Assumptions can be hard to test
  - ✘ Violations can be inscrutable
- ✓Advantages and ✘ limitations (sometimes) of **nonparametric statistics**:
  - ✓ "No" assumptions necessary
  - ✓ Appropriate for non-numeric data
  - ✓ Appropriate for small sample sizes
  - ✘ Can be "block box" algorithms
  - ✘ Can be inefficient/slow
  - ✘ Results can vary each run
- In general, use:
  - Parametric methods when possible.
  - Nonparametric methods when necessary.

Multiple Comparisons Problem

- **Multiplicity (multiple comparison problem)**: the increase of erroneous inferences when comparing a set of statistical inferences simultaneously, or when inferring a subset of parameters based on the observed values.
  - As more attributes are compared, the more likely it becomes that observed outcome is due to sampling error, as probabilities are additive.
  - E.g., despite all the alternative hypotheses have a statistical significant value individually (5%), together they provide a high rate of type I errors $\alpha\uparrow$, i.e.,

$$p(H_1 \mid H_0) + p(H_2 \mid H_0) + p(H_3 \mid H_0) = 0.15 = \alpha$$

  - The above is just a comparing against the H_0, the problem becomes much worse when including pairwise comparisons between all H_a in the set ($15\% \rightarrow 30\%$ α).
- Common conceptualizations of multiplicity problem can be done via descriptions of errors rates, e.g.,
  - **Family-wise error rate (FWER)** $\tilde{\alpha}$: the probability of making at least one false positive when performing multiple hypotheses tests, i.e.,

$$\tilde{\alpha} = 1 - (1 - \alpha_i)^m \leftrightarrow p(\alpha \geq 1)$$

    - $i =$ per comparison, $m =$ total hypotheses tested
  - **False discovery rate (FDR)** $E[Q]$: the expected proportion Q of false positives relative to total number of true positives $1 - \beta$, i.e.,

$$E[Q] = \frac{\alpha}{(\alpha + (1 - \beta))} \qquad \beta = \text{false negative}$$

- Each conceptualization can have a variety of relevant controlling procedures that are used to correct for multiplicity issues, e.g,
  - **Bonferroni correction**: a conservative method, free of dependence and distributional assumptions, wherein the false positive rate per comparison is simply divided by the total number of hypotheses m tested, i.e.,

$$\alpha_i = \frac{\alpha}{m} \leftrightarrow \text{reject } H_i \text{ if } \leq \frac{\alpha}{m}$$

  - **Šidák correction**: slightly more powerful than Bonferroni, but with small gain and potential to fail when tests are negatively dependent; found via solving the FWER equation, i.e.,

$$\alpha_i = 1 - (1 - \alpha)^{1/m}$$

  - Controlling procedures for false discovery rate not described, I'm not sure relevance as of now—might revisit later.

Primer: Cross-Validation

- **Cross-validation**: a set of model validation techniques for assessing how well statistical analysis will generalize via parcelization of given data.
  - Mainly used to estimate how accurate a predictive model might be in practice for nominal and ordinal data↑ (discrete is also possible).
  - **Training set (known data)**: the portion of given data that a predictive model is used to train on.
  - **Testing set (unknown data)**: the portion of data set aside to later estimate accuracy of the trained model.
  - Falls under the category of **"resampling"** methods, which also includes permutation testing↓ and bootstrapping↓.
- Cross-validation is used on models with one or more unknown parameters, wherein a dataset is used to fit the data to the parameter via optimization.
  - **Optimization**: selection of the best element, with regard to some criterion, from some set of available alternatives.
  - **Overfitting**: when analysis corresponds too closely to a particular dataset, leading to poor predictive performance
  - **Underfitting**: when analysis fails to capture the underlying structure of the data, leading to poor predictive performance.
  - A more in-depth discussion of over and underfitting is covered when discussing nested models↓.
- Cross-validation is often used when dealing with regression↓ and confidence intervals↓.
  - In most methods, multiple rounds of cross-validation are performed using different partitions, with the results being combined over the rounds.

P-Value vs. Classification Accuracy

P-Value	Accuracy
Tests of probability of sample	Model outcome vs. observed outcome
Parameter based scoring	Individual parameters uncertain
Analytical solutions, theoretical	Empirically informed, inconsistent
Works for most model/variable types	Restricted by model/variable type
Sensitive to extreme sample sizes	Robust to sample sizes

T-Tests

- **Student's t-test**: a test statistic that follows a student's t-distribution[↑], i.e., a test for relatively small sample sizes with unknown variance.
  - Common t-tests include the one-sample[↓] and two-sample[↓] tests, often called student's t-test or simply, t-tests.
    - Note: usage student's t-test implies the variances are assumed near equal.
  - Fundamentally, t-tests are often used to determine if the means of two sets of data are significantly different from each other (when $p < t$).
- In general, most t-tests adopt the form based on the signal-to-noise ratio, i.e.,

$$t_k = \frac{Z}{s} = \frac{\bar{x} - \bar{y}}{\sigma / \sqrt{n}}$$

  - Z: difference in means; sensitive to H_a, increasing in magnitude if H_a is not wrong.
  - s: scaling factor, of the standard deviations σ of the sample distribution.
  - n: number of samples. k: degrees of freedom.
  - Thus, increasing the t-statistic can be done via increasing group differences, reducing variances, or increasing sample size.

One-Sample and Two-Sample T-Tests

- **One-sample t-test**: a single test aimed at determining whether a single set of numbers could have been drawn from a distribution with a specified mean, i.e.,

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

  - $\bar{x}$: sample mean. μ_0: specified mean of H_0. σ: sample standard deviation.
  - $k = n - 1$, as the mean is the only unknown value.
- Assumptions for one-sample t-test:
  - Data are numeric, ideally interval or ratio (discrete can work, sometimes).
  - Data are independent of each other and randomly drawn from the population.
  - The parent population (H_0) does not need to be normally distributed, but the $\bar{x}$ is assumed to be (approximately) normally distributed.
- **Two-sample t-test**: an extension of the one-sample t-test, whereby two sets of numbers could have been drawn form the same distribution.
  - The numerator stays the same, but the denominator can change based on group pairing, size, and variance.

- **Paired or unpaired**: whether two groups of data are drawn form the same population, e.g.,
  - Paired: same individuals sampled, overtime.
  - Unpaired: different populations sampled overtime.
- **Equal or unequal variance**: whether two groups have roughly equal variance.
- **Equal or unequal sample size**: whether the groups have the same number of values, only applied to unpaired groups.
- Exact algebraic definitions of each particular case will not be discussed; selection of relevant t-test depends on various combination of above factors and can easily be done in practice using various code libraries.

Nonparametric T-Tests

- **Wilcoxon signed-rank test**: a nonparametric↑ variation of the one-sample or two-sample (paired) t-test.
  - Mainly used when the data are assumed to be not normally distributed; done via testing of medians rather than means.
  - Generally speaking, the test applies the following algorithm:
    - Remove equal pairs.
    - Rank-transform the differences, i.e., $r = \text{rank}(|x - y|)$
    - Sum ranks where $x > y$.
    - Convert to a z-score, which is normally distributed under the H_0, allowing for conversion to a p-value↑.
- Note: the actual process is not covered here, again, when to use tests like these are the important factor here.
- **Mann-Whitney U test (Wilcoxon rank-sum test)**: an alternative to the independent two-sample t-test, wherein the groups do not need to have equal sample sizes.
  - The general algorithm:
    - Note the samples sizes, specifically, determine dataset with fewer points x_f, n_f and dataset with more points x_m, n_m.
    - Pool data and compute rank, i.e., $rank(\{x_f, x_m\})$
    - Compute U score, i.e., $U = \sum_{i=1}^{n_f} r_i$
    - Convert to a z-score, and thus↑, a p-value.

Primer: Permutation Testing

- **Permutation (randomization) test**: a test of statistical significance wherein the H_0 distribution is obtained via calculation of all possible values of the test statistic under all possible rearrangements of the observed data points.
  - I.e., methods of treatments to the subjects of an experimental design is analysis of that design—if the labels are exchangeable under the H_0, then results should to yield equal significance.
  - Falls under the category of **"resampling"** methods, which also includes cross-validation↑ and bootstrapping↓.
- Permutation tests are mainly used to provide a p-value, generally done via the following methods:
  - **Z-score approach**: simply the difference between observed value and the expected value of the H_0 divided by the standard deviation of the H_0, i.e.,

$$Z = \frac{obs - E[H_0]}{std[H_0]}$$

    - Conversion↑ to p-value is then easily done.
    - The observed value is not contained within the H_0, thus conversion to a z-score is often done case-by-case.
    - Only works for approximately Gaussian H_0 distributions.
  - **Counts approach**: proportion of times that the H_0 was greater than observed value to the number of permutations ran, i.e.,

$$p_c = \frac{\sum(H_0 > obs)}{N_{H_0}}$$

    - Generally appropriate, distribution shape not as significant.
    - Gives p-value directly, must be mindful of tail.
    - Can be more arbitrary than one would like.
- Again, permutation tests are a subset of nonparametric tests meant for unbalanced designs, potentially with mixtures of data types.
- Permutation testing can be computationally expensive, as it is in large part useful thanks to the exploitation of the central limit theorem↑.

Confidence Intervals

- **Confidence intervals (CI)**: the probability that an unknown population parameter θ falls within a range of values in repeated samples, i.e.,

$$p(L < \theta < U) = \gamma$$

  - **Confidence level** γ: a somewhat arbitrary number between 0–1.
    - Similar to significance levels, but instead it represents the consistency of the sampling of parameter in question, rather than the legitimacy of the H_0.
    - Typical values range from 0.95 to 0.99.
- Confidence intervals are influenced by the sample size and variance, with larger sample sizes and smaller variances leading to shorter confidence intervals.
- Analytic method for computing confidence intervals:

$$CI = \bar{x} \pm t^*(k)\frac{\sigma}{\sqrt{n}}$$

  - $\bar{x}$: sample mean, σ: sample standard deviation, n: sample size
  - t^*: t-value with k degrees of freedom; t-value associated with one tail of confidence interval, i.e.,

$$t^*(k) = \text{tinv}\left(\frac{1-\gamma}{2},\ n-1\right)$$

    - Note: σ is assumed to be an appropriate measure of variability↓.

Primer: Bootstrapping

- **Bootstrapping**: an empirical method for estimating measures of accuracy, via use of random sampling with replacement, of a given dataset.
  - Falls under the category of methods termed **"resampling"**, which also includes cross-validation↑ and permutation testing↑.
  - Treats the sample data as pseudo-population data, with each resample being a new pseudo-sample; leads to estimation of any measure of accuracy.
- ✓Advantages and ✘ limitations (sometimes) of bootstrapping:
  - ✓ Works for any kind of parameter.
  - ✓ Potentially useful for limited datasets.
  - ✓ Not based on assumptions of normality
  - ✘ Can be inconsistent
  - ✘ Dependent on quality of representative sample
  - ✘ Can be time/computationally intensive.

Confidence Intervals: Misconceptions

- ✘ Incorrect:
  - *"I am 95% confident that the population mean is the sample mean"*
  - *"I am 95% confident that the population mean is within the CI in my dataset."*
  - *"95% of the data are between the upper and lower bounds of the CI."*
  - *"CIs for two parameters overlap, therefore, they cannot be significantly different"*
- ✓ Correct:
  - *"95% of confidence intervals in repeated samples will contain the true population mean."*
  - The confidence interval is not based on raw data—it's based on descriptive statistics of the sample data.
  - The confidence interval refers to the estimate of a parameter, not the relationship between parameters.

Correlation

Correlation Fundamentals

- **Correlation (dependence)**: a statistical relationship between two random variables or bivariate data.
  - Correlations can indicate a predictive relationship that can be exploited.
  - Presence of correlation is not sufficient to infer a causal relationship, i.e., correlation does not imply causation.
- **Covariance** $\operatorname{cov}(x, y)$: a measure of the joint variability of two random variables, i.e.,

$$\operatorname{cov}(x, y) = E[(x - \bar{X})(y - \bar{Y})]$$

  - Equivocally, using discrete random variables:

$$\operatorname{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})$$

  - If linearity is assumed, then can be simplified to expected values[↑] of their product minus the product of their expected values, i.e.,

$$\operatorname{cov}(x, y) = E[xy] - \bar{X}\bar{Y}$$

- **Correlation coefficient** $[-1, 1]$: a numerical measure of some type of correlation.
  - Correlation is the normalized (dimensionless)[↑] representation of covariance.
  - Often, correlation refers to linear relationships via Pearson correlation, however, there are several measures of correlation based on data types.
    - Not all correlations will be covered, but the most common will be, i.e., the Pearson[↓], Spearman[↓], and Kendall[↓] correlations.
- **Covariance matrix** K_{xx}: a square matrix giving the covariance between each pair of elements of a given random vector, i.e.,

$$K_{xx} = \operatorname{cov}[x, x] = E[(x - \bar{X})(x - \bar{X})^T] = E[xx^T] - \bar{X}\bar{X}^T$$

  - Disclaimer: linear algebra references will not be explained, as long as it was previously covered in my notes on linear algebra.
  - Any covariance is symmetric, positive semi-definite, and its main diagonal contains variances (i.e., the covariance of each element with itself).

Pearson Correlation

- **Pearson correlation coefficient** ρ: a measure of linear correlation between two sets of data; simply the covariance divided by their standard deviations, i.e.,

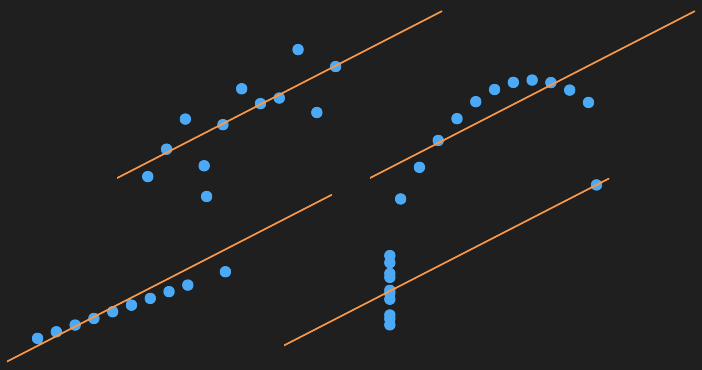
$$\rho_{x,y} = \frac{\text{cov}(x, y)}{\sigma_x \sigma y}$$

  - σ_x, σy: the standard deviation of x and y

- **Sample Pearson correlation coefficient** r: application of Pearson correlation to discrete random sample, assuming paired data:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}}$$

- The Pearson correlation generally does not completely characterize relationships, as it only emphasizes the strength of linear relationships.
  - In particular, if the **conditional mean** $E(Y \mid X)$ of Y given X is not linear in X, then the Pearson correlation will fail to characterize the relationship.
  - E.g., Anscombe's quartet demonstrates this problem, show how outliers or non-linear relationships can heavily impact the linear correlation:



    - All have the same mean, variance, correlation, and regression line, but obviously have different distributions.
  - The conclusion is that Pearson correlation can only fully characterize relationships between variables drawn from multivariate normal distributions.
    - Note: in practice, many distributions can be accurately calculated from normal distributions (not necessarily multivariate) if it has a finite covariance matrix.

- **Cosine Correlation** $\cos(\theta)$: a measure of similar between two non-zero vectors on an inner product space; represented using a dot product over magnitude, i.e.,

$$\cos(\theta) = \frac{\boldsymbol{a}^T \boldsymbol{b}}{||\boldsymbol{a}||\,||\boldsymbol{b}||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

  - If the attribute vectors are normalized (e.g., $\boldsymbol{a} - \bar{\boldsymbol{a}}$), then the measure is equivalent to the Pearson correlation coefficient.

Rank Correlation

- **Rank correlation**: a series of statistical measures of ordinal↑ association, i.e., the relationship between ranking of different ordinal values.
  - **Rank correlation coefficient**: the degree of similarity between two rankings, whereby significance can be assessed.

Spearman Correlation

- **Spearman's ρ, r_s**: a nonparametric↑ measure of rank correlation between two variables via assessment of a monotonic relationship.
  - **Monotonic function**: a function between ordered sets that preserves or reverses the given order,
  - I.e., the Pearson correlation tests for linear relationships, while Spearman's correlation tests for monotonic relationships (linear or not).
  - Can be applied for both continuous and discrete ordinal variables.
- Spearman's ρ is simply defined as the Pearson correlation coefficient between rank variables rv_x, rv_y, i.e.,

$$r_s = \rho_{rv_x, rv_y} = \frac{\text{cov}(rv_x, rv_y)}{\sigma_{rv_x, rv_y}}$$

Determining Significance

- Determining significance of correlations can be done in a number of ways, using permutation testing, the Fisher Z-transformation, or the t-distribution under the H_0.
  - **Fisher z-transformation** $F(r)$: a method of transforming uniformly distributed data ($-1 < r < 1$) into an approximately normal distribution, i.e.,

$$F(r) = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = \arctan(r)$$

    - This creates a new distribution where values near -1 and +1 are stretched out, creating the tails of the normal distribution, as -1 and +1 represent perfect monotonic relationships (typically outliers).
    - Significance (z-score↑) can then be tested when accounting for sample size n:

$$z = \sqrt{\frac{n-3}{1.06}} F(r)$$

- One can use a t-distribution with $n-2$ degrees of freedom under the H_0, based on strength of correlation r and data points n, i.e.,

$$t_{n-2} = \frac{r\sqrt{n-2}}{1-r^2}$$

Kendall Correlation

- **Kendall's τ coefficient**: a measure of ordinal data, wherein the ranks do not have meaningful relationships between levels.
- Kendall τ will be high when observations have relatively similar rank between two variables, and low when relatively not similar, compared to the rest of the data.
- **Concordant**: a pair of observations (x_1, y_1), (x_2, y_2) wherein the signs of both elements are greater than, equal to, or less than the corresponding elements of the other pair, i.e.,
$$\text{sgn}(x_2 - x_1) = \text{sgn}(y_2 - y_1)$$
- **Discordant**: not concordant or inverse signs, i.e., one pair contains a higher value of x then the other pair contains a higher value of y:
$$\text{sgn}(x_2 - x_1) = -\text{sgn}(y_2 - y_1)$$
- **Tau-a**: the strength of associations of the cross tabulations, with no adjustments for ties, i.e.,
$$\tau_a = \frac{n_c - n_d}{n_b}$$
  - $n_c \rightarrow$ number of concordant pairs; $n_d \rightarrow$ number of discordant pairs.
  - $n_b = \binom{n}{2} = \frac{n(n-1)}{2} \rightarrow$ the binomial coefficient for numbers of ways to choose two items from n items.
- **Tau-b**: an adjustment to τ that takes in account ties, i.e.,
$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_b - n_1)(n_b - n_2)}}$$
  - $n_1 = \sum_i t_i(t_i - 1)/2$,
    - $t_i \rightarrow$ number of tied values in the i^{th} group of ties for the first quantity.
  - $n_2 = \sum_j u_j(u_j - 1)/2$
    - $u_j \rightarrow$ number of tied values in the j^{th} group of ties for the second quantity.
- Disclaimer: Tau-c and significance tests for Kendall's τ not covered, as of now.

Primer: Partial Correlation

- **Partial correlation**: measures of degree of association between two random variables, with the effect of set of controlling random variables removed.
- Correlation coefficients will give misleading results if there is another confounding variable related to both variables of interest.
  - **Confounding**: a variable that influences both the dependent and independent variable, causing spurious associations.
  - **Spurious association**: when two or more variables are associated, but not casually related.
- Controlling for the confounding variables are done via the computation of a partial correlation coefficients.
  - There are a variety of methods, but most commonly linear regression↓ is used to compute partial correlations; other basic methods have significant time complexity.
- Confounding can be defined in terms of a data generating model, where X, Y are the dependent and independent variables respectively, and Z is extraneous variables that casually influences both.
  - To determine this, let $p(y \mid \text{do}(x))$ be probability of the event $Y = y$ under the hypothetical intervention of $X = x$.
  - If the following holds:
    $$p(y \mid \text{do}(x)) = p(y \mid x)$$
    for all values $X = x$ and $Y = y$, then X and Y are not confounded.
  - Essentially, the above is only true if an association is the same as the association measured in a controlled experiment with x randomized.

Subgroup Paradox

- **Simpson's (subgroup) paradox**: a phenomenon wherein a trend appears in several groups of data, but disappears or reverses when groups are combined.
  - The paradox is resolved when confounding variables and causal relations are appropriately addressed.
  - This paradox is often the culprit behind one of many misuses of statistics.
- The issue mainly involves the significance of the subgroups.
  - Often the determination of the subgroups is based on insignificant distinctions, leading to poor inferences due to assumed importance of subgroup correlations.

Analysis of Variance

ANOVA Fundamentals

- **Analysis of variance (ANOVA)**: a collection of statistical models and their associated estimates, based on the law of total variance, aimed at determining the effects of discrete independent variables (IV, X_n) on a continuous dependent variable (DV, Y).
  - **Law of total variance**: if X and y are random variables on the same probability space, and the variance of Y is finite, then

$$\operatorname{var}(Y) = E\left[\operatorname{var}(Y \mid X)\right] + \operatorname{var}(E\left[Y \mid X\right])$$

- **Basic outline for setting up an ANOVA**:
  1. Identify the independent and dependent variables.
     - Sometimes termed the explained and unexplained components of variability.
  2. Determine applicability of ANOVA to the experimental design in question.
     - Needs categorical factors, with two or more levels within each factor.
       - **Factors** λ: the "dimensions" of the IVs.
       - **Levels** ε: the specific groups (means) or manipulations within each factor.
     - Generally used to test differences among at least three levels, as t-tests and correlations are used for two.
  3. Create a table of factors and levels (if possible; when factors > 2, then it's not used).
  4. Perform computation and interpret results.
     - **Main effect**: when one factor primarily influences the dependent variable.
     - **Interactions**: the effect of one factor depends on the levels or another factor.
     - **Intercept**: when the average of dependent variable is different from zero.

Study Designs

- **One-way ANOVA**: used for testing differences of two or more levels with one factor.
- **Factorial ANOVA**: used when there is more than one factor, e.g., two-way ANOVA↓.
- **Repeated measures ANOVA**: used when the same subjects are used for each factor, e.g., in longitudinal studies.
- **Multivariate ANOVA**: when there is more than one dependent variable.
- **Balanced**: the same number of data points (sample size) in each treatment.
  - **Unbalanced**: different number of data points; often increases complexity and reduces both robustness and statistical power.

Classes of Models

- **Fixed-effects (class I)**: when the number of levels of a factor is fixed, i.e., there are discrete, and often static, groups within each factor.
  - Allows estimation of the ranges of dependent variable values that treatments would generate in the population.
- **Random-effects (class II)**: when the levels within a factor are random in the population, i.e., there are random, and often variable, groups within each factor.
  - Some levels can be discretized into discrete, fixed levels, but many cannot.
- **Mixed-effects (class III)**: a mixture of fixed and random effects, i.e., a factorial ANOVA wherein at least one factor is fixed and at least one other is random.
- Note: defining fixed and random effects has proven to be elusive, with non-standardized definitions of each often causing confusion.

Assumptions of ANOVA

- The most common approaches use linear models that relate groups and controls to the independent variables on the dependent variables, which assume:
  - **Independence**: the data are sampled independently of each other in the population to which you want to generalize.
  - **Normality**: the residuals (unexplained variance) are roughly Gaussian.
  - **Homoscedasticity**: the variance of the data in groups should be roughly equal.
- Many problems that do not satisfy the assumptions of the ANOVA can often be transformed to satisfy them.
  - E.g., the Kruskal-Wallis test can be used on rank-transformed data, and unit-treatment additivity can be applied in some cases.
  - However, many uses of ANOVA are generally robust enough to deal with violations; transformation can often be inefficient, leading more work than use.

ANOVA Methods

- In a generic sense, ANOVA is used to compare the H_0 to an H_a;
  - H_0: the null hypothesis, which states the means of all groups μ_k are statistically indistinguishable, i.e,
    $$H_0 : \mu_1 = \mu_2 = \mu_{...} = \mu_k$$
  - H_a: the alternative hypothesis, which is true when at least one group mean is statistically distinguishable from another group mean, i.e.,
    $$H_a = \mu_i \neq \mu_j$$
- Finding out what the difference between the group means, if there is one, requires post-hoc comparisons↓.

Sum of Squares

- The first step in ANOVA is to compute and interpret the measure of dispersion↑ between groups, within groups, and in total, using variance↑ (**sum of squares**), i.e.,
  $$\sigma^2 = \textbf{SS} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$
  - Note: this is sample variance, so the divisor (degree of freedom ν↑) $\frac{1}{n-1}$ is still present, but will be partitioned along with variance in order to interpret the results.
- In order to interpret the dispersion, the total variation SS_T is computed from the combination of within-group variance SS_E and between-group variance SS_B, i.e.,

$$SS_T = SS_E + SS_B \quad \mathcal{E} = \text{levels}^{\uparrow},\ n = \text{subjects in levels},\ N = \text{total subjects}$$

$$SS_T = \sum_{i=1}^{\mathcal{E}}\sum_{i=1}^{\text{n}}(x_{ij} - \bar{x})^2 \qquad \nu_T = N - 1$$

$$SS_B = \sum_{i=1}^{\mathcal{E}}(\bar{x}_i - \bar{x})^2 n_i \qquad \nu_B = \mathcal{E} - 1$$

$$SS_E = \sum_{i=1}^{\mathcal{E}}\sum_{i=1}^{\text{n}}(x_{ij} - \bar{x}_i)^2 \qquad \nu_E = N - \mathcal{E}$$

F-Test

- F**-test**: used for comparing the factors of the total deviation, represents the mean squared MS ratio of the between-group variance and within-group variance, i.e.,
  $$F = \frac{\text{“Explained” variance}}{\text{‘Unexplained’ variance}} = \frac{\text{Due to factors}}{\text{Natural variance}} = \frac{SS_B/\nu_B}{SS_E/\nu_E} = \frac{MS_B}{MS_E}$$

The ANOVA Table

- Results from ANOVA are often displayed together in a table, i.e.,

Variance	SS	ν	MS	F	p-value
Between	SS_B	$\varepsilon - 1$	MS_B	$\frac{MS_B}{MS_E}$	p
Within	SS_E	$N - \varepsilon$	MS_E		
Total	SS_T	$N - 1$			

- Again, the H_a is valid when the p-value is significant$^{\uparrow}$, meaning at least one group mean is distinguishable from at least one other group.
  - I.e., the H_0 is correctly rejected if $p \leq \alpha$.
  - If the H_a is valid, then the F statistic represents the total variability to be investigated between groups, but does not tell us which groups vary; post-hoc comparisons are required to find the source of such variability.
  - The F-test is known to be nearly optimal in the sense for minimizing false negatives$^{\uparrow}$ for a fixed rate of false positives$^{\uparrow}$.

Post Hoc Comparisons

- **Post hoc**: analysis of results after an experiment.
  - Often based on family-wise error rate$^{\uparrow}$ analysis, data visualization$^{\uparrow}$, and t-tests$^{\uparrow}$.
- **Tukey's (range) test**: a single-step multiple comparison procedure to used to find means that are significantly different from each other over a specified range.
  - Tukey's test is common post hoc approach for controlling the family-wise error rate, essentially, it is just a t-test that incorporates such control, i.e.,

$$q_{c,n-c} = \frac{\bar{x}_{\text{max}} - \bar{x}_{\text{min}}}{\sqrt{MS_E}\sqrt{2/n}}$$

    - c: the number of comparisons being made.
    - n: total number of data values.
    - c, $n - c = \nu$: the degrees of freedom, which depends on specified number of comparisons being made (the range) that are relevant.
  - Similar the t-test, if q is larger than the critical value determined by the significance level α, then the means of the comparison are significantly different.

Two-Way ANOVA

- All the methods of ANOVA discussed so far were only being applied to the one-way ANOVA↑, but such methods can be extended if multiple factors are present.
- Total variation in the dataset is the sum of the variation across: subjects n within each group + levels ε within each factor + interactions between the factors λ_x.
- The methods↑ can be extended to any factorial ANOVA, but here extension to the two-way ANOVA will be demonstrated:

$$SS_T = \sum_{i=1}^{\varepsilon_{\lambda_1}} \sum_{j=1}^{\varepsilon_{\lambda_2}} \sum_{k=1}^{n} (x_{ijk} - \bar{x})^2 \qquad \nu_T = N - 1$$

$$SS_{B\lambda_1} = \epsilon_2 n \sum_{i=1}^{\varepsilon_{\lambda_1}} (\bar{x}_i - \bar{x})^2 \qquad \nu_{B\lambda_1} = \epsilon_1 - 1$$

$$SS_{B\lambda_2} = \epsilon_1 n \sum_{j=1}^{\varepsilon_{\lambda_2}} (\bar{x}_j - \bar{x})^2 \qquad \nu_{B\lambda_2} = \epsilon_2 - 1$$

$$SS_{\lambda_1 \times \lambda_2} = \sum_{i=1}^{\varepsilon_{\lambda_1}} \sum_{j=1}^{\varepsilon_{\lambda_2}} (x_{ij} - \bar{x}_i - \bar{x}_j - \bar{x})^2 \qquad \nu_{\lambda_1 \times \lambda_2} = \nu_{B\lambda_1} \nu_{B\lambda_2}$$

$$SS_E = \sum_{i=1}^{\varepsilon_{\lambda_1}} \sum_{j=1}^{\varepsilon_{\lambda_2}} \sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2 \qquad \nu_E = N - \epsilon_1 \epsilon_2$$

Where ϵ_x = the number of levels within each factor, i.e., $\varepsilon_n \in \lambda_x$

- Which can also be represented in a table:

Variance	SS	ν	MS	F	p-value
λ_1	$SS_{B\lambda_1}$	$\epsilon_1 - 1$	MS_{λ_1}	MS_{λ_1}/MS_E	p
λ_2	$SS_{B\lambda_2}$	$\epsilon_2 - 1$	MS_{λ_2}	MS_{λ_2}/MS_E	p
$\lambda_1 \times \lambda_2$	$SS_{\lambda_1 \times \lambda_2}$	$\nu_{B\lambda_1}\nu_{B\lambda_2}$	$MS_{\lambda_1 \times \lambda_2}$	$MS_{\lambda_1 \times \lambda_2}/MS_E$	p
Within	SS_E	$N - \epsilon_1\epsilon_2$	MS_E		
Total	SS_T	$N - 1$			

- This can be extended, with increasing time complexity.
- Using values from a factorial ANOVA in data visualization is often used to interpret results, as mapping interactions can quickly get out of hand.

Regression

Regression Fundamentals

Model-Fitting

- **Model fitting**: the combination of fixed features and free parameters in such a way that fits experimental data to a mathematical models based on adjustments to the free parameters that attempts to explain the dependent variable.
  - **Fixed features (regressor)** x_n: independent variables imposed on the model based on previous knowledge, understanding, theories, hypotheses, or other evidence; has several other names based on context.
  - **Free parameters** β_n: scalar variables that cannot be predicted precisely or constrained by the model; they must be adjusted or estimated.
  - **Intercept** β_0: the average when all other parameters are 0.
  - **Residual (error)** ϵ: the data not directly observed or fit by the model.
- **Model interpolation**: the prediction of other experimental results given prior experimental data and a fitted model based on those data.
- **The general outline of model-fitting**:
  - Define the equation(s) underlying the model; dependent on data availability.
    - If all the fixed features are discrete↑, then use ANOVA↑.
    - If at least some fixed features are continuous, then use regression.
    - E.g., height h is governed by numerous complex interactions, but a simplistic model can be made to estimate the importance of particular fixed features;

$$h = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$x_1 : \text{sex}, x_2 : \text{parents' height}, x_3 : \text{nutrition}$$

  - Map the data to the model equations, i.e., take the real, or simulated, data and map them to the fixed features.
    - Yields a system of equations with a series of unknown parameters.
    - **Over determined**: a set of equations with more equations than unknowns.
  - Convert the equations into a matrix-vector equation, i.e., the general linear model↓.
    - Sometimes simplified to $\boldsymbol{X\beta} = \boldsymbol{y}$ (linear algebra nomenclature: $\boldsymbol{Ax} = \boldsymbol{b}$).
  - Computer the parameters, e.g., using least-squares↓.
  - Statistical evaluation of the model, i.e., the application of inferential statistics.
    - See model significance↓ and coefficients significance↓.

Least-Squares

- **Least-squares**: a standard approach in regression analysis to approximate the solution of over determined systems by minimizing the sum of the squares of the residuals, i.e.,

$$\|\epsilon\|^2 = \|X\beta - y\|^2$$

  - **Residuals** ϵ: the vector that describes the difference between the observed value y and the estimated value $X\beta$ ($\hat{y}$), i.e.,

$$\epsilon = \hat{y} - y = X\beta - y$$

- Since the design matrix X is an over determined system, then the left inverse can be used to isolate the regression coefficients if X has full column rank, i.e.,

$$\begin{aligned} X\beta &= y \\ (X^T X)^{-1} X^T X\beta &= (X^T X)^{-1} X^T y \\ \beta &= (X^T X)^{-1} X^T y \end{aligned}$$

Model Significance

- **Coefficient of determination** R^2: a measure of how well observed outcomes are replicated by the models; exact definitions vary.
  - In simple linear regression$^{\downarrow}$ it's referred to as r^2, and if the intercept is included, then it's simply the square of the sample correlation coefficient $r^{\uparrow}$.
  - R^2 indicates that there are additional regressors, which means it is equal to the square of the coefficient of multiple correlation.
  - Values range from 0 to 1, with 1 indicating a better fit.
- R^2 can be found by comparing the sum of squares$^{\uparrow}$ of the residuals over the total sum of squares, i.e.,

$$R^2 = 1 - \frac{SS_\epsilon}{SS_T} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

  - There is no real cut-off for a "good fit", often it is used to compare models.
- Further evaluation of significance is similar to the F-test$^{\uparrow}$ used in ANOVA, whereby determination of at least one $\beta \neq 0$ is achieved, else all regressors under $H_0 = 0$.
  - The F-test here uses a comparison $SS_{\text{Model}} = \sum(\hat{y}_i - \bar{y})^2$ over SS_ϵ, with degrees of freedom determined by number of parameters (+ β_0) k, i.e.,

$$F_{(k-1,N-k)} = \frac{SS_M/(k-1)}{SS_\epsilon/(N-k)}$$

  - Each individual β coefficient can then be evaluated using a t-distribution:

$$t_{N-k} = \frac{\beta}{\sigma_\beta} = \frac{\beta}{\sqrt{SS_\epsilon/SS_T}}$$

Standardized Regression Coefficients

- **Standardized (regression) coefficient (weights)**: when the free parameters β of a model have been standardized so that the variances both the dependent and independent variables are equal to 1.
  - Unstandardized β weights can change depending on the scale of the independent variables, leading to near impossible comparisons of variables and studies.
  - There are some cases where unstandardized β weights can reflect important scales of the data, potentially facilitating interpretations.
- Standardized β weights are unitless; the measure refers to effect of one-standard deviations σ ↑ of a regressor x on the dependent variable y, i.e.,

$$\beta^* = \beta_k \frac{\sigma_{xk}}{\sigma_y}$$

  - Note: sometimes, standardization is done with only respect to the regressor.
  - Another method involves z-normalizing ↑ all the variables before regression.
- Comparisons may be easier with standardized weights, but re-scaling based simply on standard deviations may increase difficulty of analysis due to confounding ↑.
  - Additionally, non-normal distributions can potentially make the method of standardization much less representative of the truth.

Statistical Power

- **Power**: the probability to reject the H_0, given that the H_0 is actually false, i.e.,

$$\text{power} = p(\text{reject } H_0 \mid H_0 \text{ is false})$$

  - The power of the test is really just the true positive $1 - \beta$ ↑, where the type II error β probability of a false negative.
- The power cannot be calculated unless the probabilities of all possible values of the parameter that violates the H_0 are known.
- Generally, power refers to a test's power against a specific H_a.
- Power increases with sample size, effect size, and lower significance level α ↑
- Power decreases with higher variability and higher α levels.
- Ultimately, power is more of a useful guideline than a precise numerical value.
- **A priori power**: power analysis conducted prior to a study, typically used to estimate sufficient sample sizes to adequate power; often has many subjective issues with the process, methods not covered here.
- **Post hoc power**: observed power after the study has been completed, proportional to the p-value attained, no new information is gained and can easily be misleading.

Regression Models

- **General linear model**: a compact way of writing several multiple linear regression models using matrix algebra, i.e.,

$$Y = \beta X + \epsilon$$

  - Y: matrix with a series of multivariate measurements, where each column is a set of measurements of one of the dependent (response) variables.
  - X: the design matrix, where each column is a set of observations on independent variables (regressors).
  - β: the matrix of β coefficients (free parameters, scalars) to be estimated.
  - ϵ: the matrix of residuals associated with the model.

Simple and Multiple Regression

  - **Simple linear regression**: the simplest case using a single regressor (plus the intercept[1]) and a single dependent variable over a number of observations i, i.e.,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  - **Multiple linear regression**: the generalization of simple linear regression to the case of more than one regressor X_n.
    - There is a special case (the basic model), where the general linear model is restricted to one dependent variable, i.e.,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \epsilon_i$$

    - As well as the more general multivariate linear regression, where there are more than one dependent variables that share the same set of regressors, i.e.,

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{i1} + \beta_{2j} x_{i2} + \cdots + \beta_{nj} x_{in} + \epsilon_{ij}$$

    - Where all dependent variables are indexed as $j = 1, \ldots, n$
    - Note: multivariate analysis of the general linear model deals with analysis of all the outcomes at once.
      - In contrast, multiple linear regression (multivariable analysis) defined here deals with each outcome independently, with respect to each dependent variable examined.
    - Multiple linear regression often has additional regressors generated from interactions between various regressors in the set, as long as independence is maintained; for now, they can be thought of as new fixed features.

Polynomial Regression

- **Polynomial regression**: a special case of multiple linear regression wherein the relationship between the fixed features and the dependent variable is modeled as an nth degree polynomial, i.e.,

$$y = \beta_0 x^0 + \beta_1 x^1 + \cdots + \beta_k x^k + \epsilon$$

  - The nonlinear relationship is fit to the values of x and the corresponding conditional mean of y, i.e., $E(y \mid x)$
  - The statistical problem is still linear, despite nonlinear fixed features, since the free parameters (β weights) are scalars that keep the model linear.

- Under and overfitting↓ can easily be done using polynomial regression, with a too low or too high order polynomial leading to each case, respectively. Therefore, a method of model selection is needed.
  - **Bayesian information criterion (BIC)**: a criterion for model selection (often used in cross validation↑) among a finite set of models; a lower BIC score is preferred.
  - BIC defined in the case of model order selection:

$$\text{BIC}_k = n \ln(SS_\epsilon) + k \ln(n) \qquad k\text{: n-parameters, } n\text{: n-data points}$$

  - BIC applied to ordered models with increasing degree can be analyzed to find the lowest scoring value.
    - Note: the lowest score is typically found after a large drop, with a slowly increasing score thereafter.

Logistic Regression

- **Logistic regression**: used to model the probability of a binary dependent variable.
  - Outputs with more than two values are modeled by multinomial logistic regression, or ordinal logistic regression if the categories are ordered.
  - Does not perform statistical classification, but can be made into a binary classifier by choosing a cutoff probability value.
- Logistic regression involves a basic regression model the **log-odds (logit)**, i.e.,

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - Finding the probability that $y = 1$ yields the **sigmoid function** $S(x)$:

$$p = S(x) = \frac{1}{1 + e^{-x}} \qquad x = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

    - Using the log-odds increases the dynamic range of the small values, leading to better optimization results when dealing with those smaller probability values.
- Nonlinearities in the coefficients prevent the left-inverse from being used to find the regression coefficient.
    - Instead, iterative methods such as gradient descent are applied to find the set of parameters that make the probabilities best match the dependent variable

Nested Models

- Qualitatively, overfitting and underfitting of models occurs when the models correspond to closely or fail to capture the underlying structure, respectively.
    - All models exist between these two measures; sometimes certain trade-offs can make one direction more favorable, i.e.,

Overfitting	Underfitting
Overly sensitive to noise	Less sensitive to noise
Increased sensitivity to subtle effects	Less likely to detect true effects
Increased estimation complexity	Decreased estimation complexity
Better results with more data	Better results with fewer data

    - Too far in either direction leads to reduced generalizability.
    - **Hidden overfitting**: or "researchers degree of freedom", the result of choices made in data cleaning, organizing, selection, and other choices of surrounding model us, that occur in order to get a certain model to work.
        - Deciding on analysis pipelines to use in advance helps avoid this problem.
        - Exploratory phases of the analysis pipelines should only be done on small samples of the data if needed.
- Quantitatively, there are methods of detecting the fit of a model via comparisons of "full" and "reduced" models.
- **Extended (full) model**: a model that always fits the data better than any reduced model; having more parameters (extending the model) is only justified when the model fit is significantly improved.
- **Nested (reduced) model**: a model that contains a subset of identical fixed features to the full model, both of which explain the same dependent variable.
    - Many models can be nested under the same full model.
    - Nested models can differ by more than one parameter, but often don't.
- Sum of squares↑ and the F-test↑ can be used for model comparison, similar to their usage in model significance↑.

- Namely, the sum of squared errors SS_ϵ between the full model and reduced model can be used to quantitatively measure the model fit to the data, i.e.,

$$SS_\epsilon = \sum (y_i - \hat{y})^2 \qquad y_i\text{: observed} \quad \hat{y}\text{: predicted}$$

$$F_{(p-k,n-p-1)} = \frac{\left(SS_\epsilon^R - SS_\epsilon^F\right)/(p-k)}{SS_\epsilon^F/(p-k-1)}$$

  - SS_ϵ^F: full model SS_ϵ with the full model's number of parameters p.
  - SS_ϵ^R: reduced model SS_ϵ with the reduced model's number of parameters k.
- The main point of focus is the difference between the full and reduced sum of squares $\left(SS_\epsilon^R - SS_\epsilon^F\right)$.
  - F is statistically significant when more parameters improves the model, i.e., the full model is preferred when SS_ϵ^F is small, making F large.
  - F is not significant when reduced model fit is nearly as good as the full model, i.e., the reduced model is preferred when the difference is low, making F small.
- Formally, the H_0 sates that all additional parameters are essentially 0, while the H_a is true if at least one is not, i.e.,

$$H_0 = \beta_{k+1} = \cdots = \beta_p = 0$$
$$H_a = N\left\{\beta_{k+1:p} \neq 0\right\} \geq 1$$

- Post hoc comparisons↑ would need to be performed in case of H_a being true.

Clustering and Dimensionality Reduction

Clustering

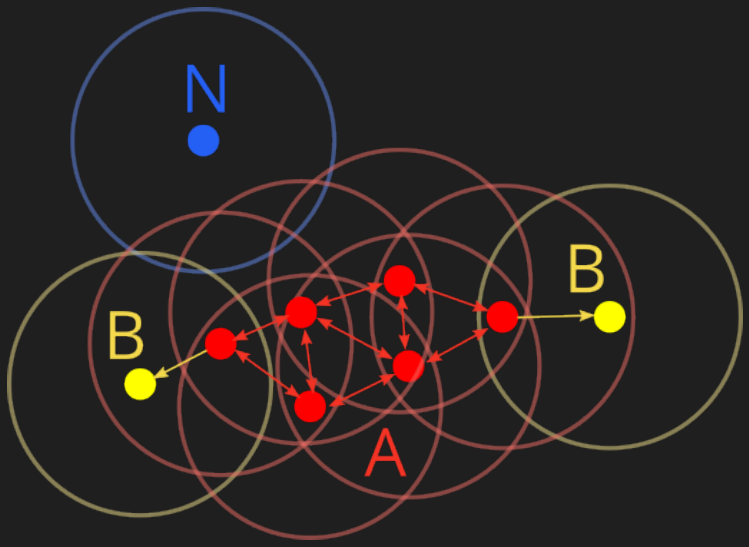
- **Cluster analysis**: the task or grouping a set of objects in such a way that the objects in the same group (cluster) are more similar to each other than those in other groups.
- There is no one specific clustering algorithm, instead there are various algorithms that differ by what constitutes a cluster and how to efficiently find them.
- Distance (to center↓, or to neighbor↓) and density↓ will be the main metrics investigated here, but there are certainly more.

K-Means Clustering

- ***k*-means clustering**: a method of vector quantization that aims to partition *n* observations (multidimensional data) into *k* clusters, wherein each observation belongs to the cluster with the nearest centroid (mean).
  - Cluster membership is defined using distances to the nearest center, where the aim is to minimize within-group and maximizes between-group distances.
- Specifics (or the many alterations) of the algorithm will not be covered, instead, a general outline of the method will be described:
  1. Select *k* (can be very difficult); many potential methods.
  2. Create *k* centroids at random (possibly slightly informed) locations in the dataset.
  3. Compute the sum of squared distances from all data points in the dataset.
  4. Assign each data point to its closest centroid.
  5. Create a new centroid at the average of all data points.
  6. Repeat 3–5 until some measure convergence (how much mean moves each iteration, typically).
- Difficulties with *k*-means:
  - Evaluation of proper *k* is often difficult.
  - Multidimensional clustering is often hard to visualize, making evaluation even harder sometimes.
  - Computation is complex (NP-hard); therefore, each result is often different.
  - Suboptimal when cluster sizes differ greatly.
  - What makes clusters similar may not be Euclidean based.

DBSCAN

- **Desnity-based spatial cluster of applications with nose DBSCAN**: a density based nonparametric↑ algorithm that clusters points into groups that are relatively close together; used as means to label and find clusters in data.
- **Step size** ε: the radius of a point that in combination with specified minimum nearby points m that determine the density of neighborhood of points with respect to other points.
- **Core points** A: a point where at least m points are within distance ε (+ itself).
- **Reachable points** B: when the point is within ε from a core point, but $< m$ required to be a core point.
- **Outliers** N: low-density regions, often considered as noise points.



  - Decreasing $\varepsilon \rightarrow$ increases number of clusters, breaks up clusters.
  - Increasing $\varepsilon \rightarrow$ decreases number of clusters, combines clusters.
  - Decreasing $m \rightarrow$ decreases number of clusters, true clusters may be split up.
  - Increasing $m \rightarrow$ increases number of clusters, true clusters may be ignored.

K-Nearest Neighbor

- **k-nearest neighbor k-NN**: a nonparametric classification method applied to points based on the distance to, potentially and plurality of, its neighboring points.
  - Unlike k-means, data must be labeled; k-NN is used to classify new data points.
  - If $k = 1$, then it's simply assigned to class of nearest neighbor.
- Normalizing the training data can greatly improve accuracy of k-NN, as it relies on distance, which may come in vastly different scales.
- k-NN regression can be used in a similar sense, where out output is the property value of the object based on averages of values of k nearest neighbors.

Dimensionality Reduction

- **Dimensionality (dimension) reduction**: the transformation of data from high-dimensional space into a lower-dimensional space, so that the low-dimensional representation retains meaningful properties of the original data.
  - Raw data in higher dimensions are often sparse, due to the "curse of dimensionality."
  - Analysis in higher dimensions is often computationally intractable, requiring dimension reduction of some kind in order to be feasible.
- Dimension reduction is common in fields that deal with large number of observations and/or variables, e.g., signal processing, speech recognition, neuroinformatics, bioinformatics, etc.
- There are several methods, which are commonly divided into linear and non-linear approaches; for now, only primers with simplified math on principal and independent component analysis will be covered.

Primer: Principal Components Analysis

- **Principal component analysis (PCA)**: the main linear technique for dimension reduction, wherein linear mapping of the data to lower-dimensional space minimizes variance of data in the new representation.
  - The largest eigenvalues (principal components) and corresponding eigenvectors of the data's covariance matrix is used reconstruct (change of basis) a large fraction of the variance of the original data.
  - Often only the first few principal components are used, as they contribute the vast majority to the systems' energy and the rest tend to be noise.
    - These largest values together represent the lower-dimension space that the data will be compressed to.
- Limitations of PCA:
  - Principal components are eigenvectors, thus they are forced to be orthogonal; the dimension of key features often aren't orthogonal.
  - PCA maximizes variance to find distinctions between dimensions, however, sometimes this might maximize unwanted noise and miss important features within the data.
- In conclusion: PCA is great for dimension reduction and visual exploration, especially if the data conform to assumptions, but can be suboptimal when interpreting PCs as factors (unique sources of variance).

Primer: Independent Components Analysis

- **Independent component analysis (ICA)**: a method for separating multivariate signal into additive subcomponents.
  - Performed by assuming that the subcomponents are non-Gaussian signals and are statistically independent of each other.
- Basic outline of ICA:
  - "Whiten" the data, or remove co/variances across different variables (using PCA)
    - Linear dependencies are removed, but shared information is preserved.
  - Rotate axes (oblique) to minimize share information.
    - Unlike PCA, ICA is not limited to orthogonal axes, allowing for independent non-orthogonal axes to be found.