Statistics

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 shifted and 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)** $\mathcal{E}$: 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.

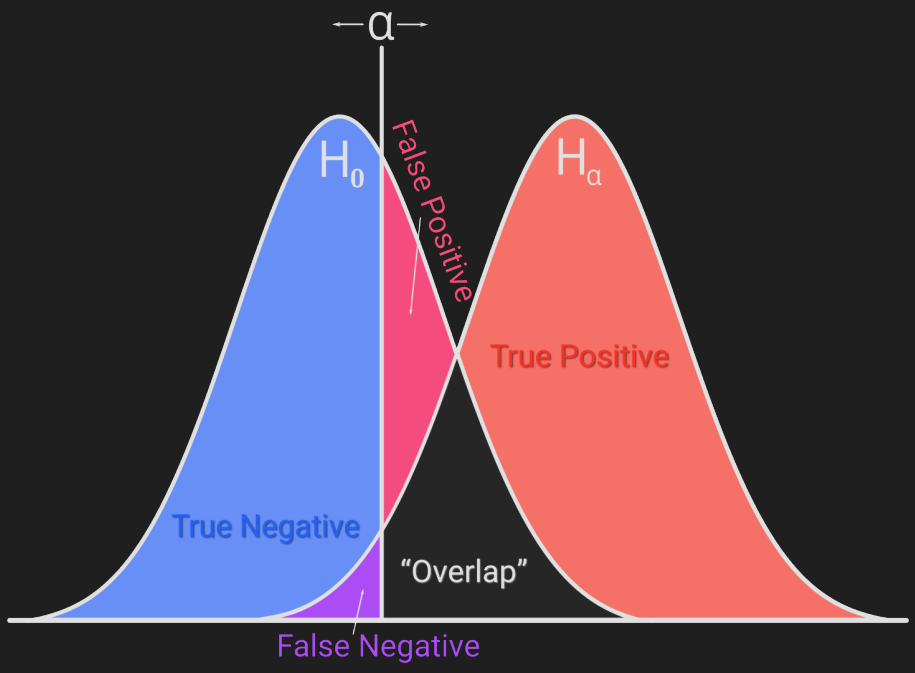
- Common misinterpretations of p-values:
  - ✘ Incorrect:
    - *"My p-value is 0.02, so the effect is present for 2% of the population."*
    - *"My p-value is 0.02, so there is a 90% chance that my sample statistic equals the population parameter."*
    - *"My p-value is smaller than the threshold, 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 $\sigma_1 \leftrightarrow z = \pm 1 = 0.683$
  - 95.5% of the data are within $\sigma_2 \leftrightarrow z = \pm 2 = 0.955$
  - 99.7% of the data are within $\sigma_3 \leftrightarrow z = \pm 3 = 0.997$
- Common p-values parings with standard deviations:
  - One-tailed $\downarrow$
  - $p = 0.05 \leftrightarrow z = 1.64$
  - $p = 0.01 \leftrightarrow z = 2.32$
  - $p = 0.001 \leftrightarrow z = 3.09$
  - $\downarrow$ Two-tailed $\downarrow$
  - $p = 0.05 \leftrightarrow z = 1.96$
  - $p = 0.01 \leftrightarrow z = 2.58$
  - $p = 0.001 \leftrightarrow z = 3.29$

Degrees of Freedom

- **Degrees of freedom (d.f. ν)**: the number of values 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 and cross-validation.
- ✓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.
- 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.
- Cross-validation is of greater importance when dealing with regression↓ and confidence intervals↓.
  - In-depth discussion will occur later, including distinctions between exhausting and non-exhaustive↓ methods.
  - 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 follow 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.
  - Similar to cross-validation↑, as they are both methods of resampling via use of nonparametric statistics.
- 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**: 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.

Primer: Bootstrapping

-

Confidence Intervals: Misconceptions

-

Correlation

Analysis of Variance

Cross-Validation: Revisited

Exhaustive vs. Non-Exhaustive

-

Statistical Power and Sample Sizes

Clustering and Dimension-Reduction

Signal Detection Theory