Notes for DS101: Principles of Data Science

Chris Kong(zk2086@nyu.edu), Dingyu Fu(df2331@nyu.edu)

December 17, 2023

Contents

1 Intro to Data Science

1.1 Philosophy of Data Science

1. **What is Science?**
   A set of knowledge that is learned through observations (data).

2. **What is the goal of Science?**
   To understand natural world.

3. **What does "understand" mean?**
   To describe, to explain, and to predict.

4. **What are scientific methods?**
   Deduction, induction, falsification, and experiments.

5. **What does Maths do?**
   Only deduction.

6. **What is Deduction?**
   To make predictions from models. It goes from true premises to logically valid conclusions.
   The **advantage** is: if premises and reasoning is good, the conclusions are certain.
   The **disadvantage** is: nothing new will be learned.

7. **What is Induction?**
   To formulate principles from observations. It goes from observations to general rules/conclusions/principles.
   The **advantage** is: new knowledge will be generated.
   The **disadvantage** is: one can be wrong.

8. **What are experiments?**
   Experiments help us understand causality.

9. **What is Falsification?**
   To falsifying a rule that was theorized through observation(induction). This is the only way we learn knowledge which is counterintuitive.

10. **What is human fallacy?**
    It is that we tend to favor evidence that supports what we expect. This is also called **Confirmation Bias**.

1.2 Basic Linear Algebra

1. **What are vectors?**
   Vectors are stacks of numbers. Linear Algebra is the study of vectors and the operations on them.

2. **What are some of the vector operations?**
   Scalar multiplication, addition.

3. **What is Transpose?**
   It converts rows to columns and vice versa, leaving elements and their order intact.

4. **What is vector subspace?**
   A subspace is the region of the vector space that can be reached by any linear combination of a set of vectors.

5. **What is Linear Dependency?**
   If a vector $\mathbf{v}$ can be expressed as a linear combination of another vector $\mathbf{w}$, these vectors are linearly dependent. If no such combination exists (other than scaling both vectors with 0), they are linearly independent.

6. **What is Dot Product?**
The (vector) dot product of $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$ is defined as:

$$\mathbf{x} \cdot \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i = \alpha$$

This is why sometimes it is called **scalar product**. The **inner product** is a generalized version of the dot product in cases where the vectors are continuous functions.

7. **How do we derive vector length?**
The dot product has immediate geometric applications. For instance, it can be used to determine the length of a vector **a**=[3,4]:

$$\mathbf{a}^T \mathbf{a} = \sum_i \mathbf{a}_i \mathbf{a}_i = |\mathbf{a}|^2 \Rightarrow |\mathbf{a}|^2 = 3^2 + 4^2 = 25 \Rightarrow |a| = 5$$

$|\mathbf{a}|$ is called the **magnitude** of the vector, the **vector length**, and also the **Euclidean norm**.

8. **What is Norm?**
Norms are functions that determine (assign) a length to a vector. **L1** norm: The sum of absolute values. **L2** norm: Square root of the sum of the squared absolute values. In general:

$$|x|_p = \left(\sum |x|^p \right)^{1/p}$$

9. **How is Norm related to the vector?**
The larger the level of norm is, the smaller it gets (closer to the largest value in the vector).

10. **What are some geometric implications of the dot product?**
In addition to the dot product of a vector with itself, one can also take the dot product between different vectors. Generally speaking, the dot product is the cosine of the angle between the two vectors, times their length:

$$\mathbf{a}^T \mathbf{b} = |\mathbf{a}| \cdot |\mathbf{b}| \cdot \cos \theta_{ab}$$

This also have geometric implications as follow:

- If this cosine is 1, the vectors point in the same direction, i.e. they are collinear;
- If this cosine is positive, the angle between the vectors is acute;
- If this cosine is negative, the angle between the vectors is obtuse;
- If this cosine is 0, the vectors are orthogonal: $\mathbf{a} \perp \mathbf{b}$.

11. **What is vector projection?**
Vector projection is the operation of projecting one vector onto another, resulting in a vector that represents the directional component of the first vector in the direction of the second. Projection of a point b onto a line a is:

$$proj_a(b) = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \mathbf{a}$$

12. **Is Matrix multiplication commutative?**
matrix multiplication is *not* commutative. The dimensionality n_A of $A = m_A \times n_A$ and m_B of $B = m_B \times n_B$ has to match ("need match where they touch").

13. **What are spatial matrices?**
Some matrices represent particularly useful transformations. The effect of a matrix multiplication with an orthogonal matrix is to rotate the input vectors without changing their lengths or the angles between them; The effect of a matrix multiplication with a diagonal matrix is to squeeze or stretch the input vectors without changing their direction.

1.3 Basic Probability Theory

1. **What is a probability?**
   Abstract: Weight of empirical evidence. Two ways of understanding it are as follow. **Objective** (Frequentist) interpretation: Relative frequency in the long run. **Subjective** (Bayesian) interpretation: Degree of belief (plausibility/confidence) based on the currently available evidence.

2. **What is sample space?**
   The set of all possible outcomes (what would happen) of a random action. One important property is that: The probabilities of the outcomes in a sample space add up to 1.

3. **What is Addition Rule?**
   If events are mutually exclusive, then we have: $p(A \vee B) = p(A \cup B) = p(A) + p(B)$;
   If events are not mutually exclusive, then we have: $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$.

4. **How can we tell if two events are independent or not?**
   If events are independent of each other: $p(A \wedge B) = p(A \cap B) = p(A) * p(B)$. In other words, the happening (presence) of one event will NOT affect the happening (presence) of the other. That's why we call it independent.

5. **What if two events are not independent?**
   Then we will utilize/introduce conditional probability. If event A and event B are not independent, then the probability that they happen together is given by:

$$p(A \cap B) = p(A) * P(B|A)$$

1.4 Basic Probability Distribution

1. **What is Poisson Distribution?**
   It is a discrete distribution with the following properties:

   - A constant rate of events (over time);
   - Events are happening independently.

   Formally, it is a discrete distribution that describes the # of events that will occur within a fixed period, given the average number of times the event occurs over that period.

2. **What are some properties of Poisson Distribution?**
   One key feature is that: Poisson distribution is fully determined by a single parameter λ, i.e. "rate". The formulation is:

$$P(X) = \frac{e^{-\lambda}\lambda^X}{X!}$$

   - Increase λ: it means the average number of events in the interval is going up. Consequently, the probability distribution becomes more spread out. This means that not only do more frequent occurrences become more likely, but the probability of observing very high counts increases as well. The distribution becomes less skewed and starts to resemble a normal distribution as λ becomes very large.
   - Decrease λ: when λ decreases, the average number of events in the interval goes down. The distribution becomes more concentrated around lower numbers, and the probability of observing higher counts decreases. For very small values of λ, the distribution becomes highly skewed, with a strong peak at zero or near-zero values.

3. **What is Binomial distribution?**
   The Binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent experiments, each asking a yes/no question, and each with its own boolean-valued outcome: success/failure (or yes/no, 0/1, etc.). This distribution is used extensively in statistics, especially for binary outcomes. The Binomial distribution is widely used in scenarios like quality control (defective vs. non-defective items), medical trials (effective vs. ineffective treatment), and any process where the outcome is binary, and the process is repeated a fixed number of times.

4. **What are some key features of Binomial distribution?**
The probability of getting exactly k successes in n trials in a Binomial distribution is given by the formula:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k}$ is the binomial coefficient, calculated as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

5. **What are some characteristics of Binomial distribution?**
Characteristics of the Binomial Distribution:

- The mean (expected value) of the distribution is given by $\mu = np$.
- The variance is given by $\sigma^2 = np(1-p)$.
- The distribution is symmetric if $p = 0.5$, skewed to the right if $p < 0.5$, and skewed to the left if $p > 0.5$.

6. **What is Cauchy distribution?**
The Cauchy distribution, also known as the Lorentz distribution, is a continuous probability distribution. It is often used in statistics as an example of a distribution with well-defined median and mode but no mean or variance. The formulation is: The probability density function (PDF) of the Cauchy distribution is given by:

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

where:

- x_0 is the location parameter, specifying the peak of the distribution.
- γ is the scale parameter, which specifies the half-width at half-maximum (HWHM).

7. **What are some properties of Cauchy distribution?**

- The Cauchy distribution does not have a finite mean or variance.
- The median and mode of the distribution are both equal to the location parameter x_0.
- The distribution is symmetric around x_0.
- It has heavy tails and a peak at x_0.

8. **What is Normal distribution?**
The Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric around its mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The probability density function (PDF) of the Normal distribution is given by:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where:

- μ is the mean or expectation of the distribution (and also its median and mode).
- σ is the standard deviation.
- σ^2 is the variance.

9. **What are some key properties of Normal distribution?**

- The Normal distribution is symmetric about its mean.

- The shape of the Normal distribution is determined by the mean and the standard deviation.
- The Empirical Rule (68-95-99.7 rule) states that for a Normal distribution, nearly all values lie within 3 standard deviations of the mean.
- The distribution is bell-shaped and is known for its occurrence in a variety of natural phenomena.

2 Characterizing Datasets

2.1 Central Tendency

1. **What are some of the parameters of central tendency?**
   Depending on the properties of the data, one of these is the most suitable measure of central tendency ("averages"):
   - **Mean:** long run expected value;
   - **Median:** central value of ranking data;
   - **Mode:** most common value of a categorical data.
2. **Why do we calculate mean in the way we do?**
   Interpretation: If guessing some value in a dataset, guessing the arithmetic mean minimizes the likely error.
3. **What are some significance of mean?**
   - The sum of deviations from the mean is always zero;
   - The sum of squared deviations from the mean is as small as possible, given the data (minimal).
4. **What are some problems with mean?**
   The mean is easily and perhaps dramatically affected by extreme values, threatening the meaning of mean itself.
5. **What can we do if there are extreme values?**
   Instead of minimizing the sum of the squared deviations ("L2 norm"), one could minimize the sum of the absolute values ("L1 norm"). This way we get the **median**, which is a more robust central tendency parameter. It relies only on rank order (not absolute magnitude).
6. **What is mode?**
   The mode is defined as the most frequent value in a dataset, corresponding to the peak in a histogram (counts per bin), when visualized. Very non-robust when the data is not distributed normally. The number of non-zero elements between the data and any data value is minimized by the mode. So the mode minimizes the L0 norm for a given dataset.
7. **What is Ergodicity and why is it important?**
   It refers to a property of a system where, over a long time, the time spent by the system in some region of the phase space of states is proportional to the volume of this region. Essentially, this means that over a long period, the time average and the ensemble average of a system's properties are equal.

2.2 Dispersion

1. **What is dispersion?**
   It is how much variation we should anticipate around the expected value.
2. **What is the Standard Deviation?**
   It is also written as STD, SD or σ. More explicitly we call it Root Mean Square Error(RMSE) or sometimes Root Mean Squared Deviation(RMSD) because of the way we calculate/define it:

$$\sigma = \sqrt{\frac{\sum_1^n (x_i - \mu)^2}{n}}$$

3. **Why do we care about normal distribution so much?**
   Because it is well behaved, meaning that this distribution can be completely described by two special numbers (mean and standard deviation). Sometimes we call normal distribution the "bell curve" or the "Gaussian distribution".

4. **What's necessary for a normal distribution?**
   1) Factors combining independently. 2) Lots of them.

5. **What is the problem with standard deviation?**
   It is very sensitive with extreme values.

6. **What is MAD?**
   MAD stands for Mean Absolute Deviation. As suggested in its name, we have the function:

$$MAD = \frac{\sum_1^n |x_i = \mu_i|}{n}$$

   It helps us overcome the effect of extreme values, i.e. more robust to outliers. Also, under the same sample size, MAD is always less or equal than STD.

7. **As summary, here is the 4 "chords" of descriptive statistics primarily used in DS**

   - Mean: L2 central tendency
   - Median: L1 central tendency
   - SD: L2 Dispersion
   - MAD: L1 Dispersion

2.3 Association

1. **What is correlation?**
   The point of statistics like a correlation is to characterize how two variables co-vary. The covariance between two random variables X and Y is defined as:

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_1^n (x_i - \mu_x)(y_i - \mu_y) = E[(X - \mu_X)(Y - \mu_Y)]$$

   The correlation coefficient, denoted as $\rho_{X,Y}$ or $\text{corr}(X, Y)$, between two variables X and Y is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

   Expanding the covariance in the numerator, we get:

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}}$$

   This is the correlation coefficient, which measures the linear relationship between X and Y. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

2. **What about other kind of correlations?**
   The above ρ is also called Pearson correlation, which describes **Linear** relationship. What if the data is non-linear? Then we shall introduce **Spearman** correlation and **Chatterjee** correlation.

3. **What is Spearman's Rank Correlation Coefficient?**
   Spearman's Rank Correlation Coefficient is a nonparametric measure of rank correlation that assesses how well the relationship between two variables can be described using a monotonic function.

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

   where d_i is the difference between the ranks of corresponding variables and n is the number of observations.

4. **What is Chatterjee's Correlation Coefficient?**
   Chatterjee's Correlation Coefficient is a measure of association between two variables that is robust to outliers and works well even when classical assumptions of linearity and normality are violated.
$$C = \frac{\sum_{i=1}^{n} \left(I(x_i > x) - I(x_i < x)\right)\left(I(y_i > y) - I(y_i < y)\right)}{n-1}$$
   where I is the indicator function, (x_i, y_i) are the individual data points, and (x, y) are their respective medians.

5. **What to notice?**
   Correlation only captures **linear** relations.

3 Prediction

3.1 Simple Linear Regression

1. **What should we do when we first get the data?**
   We should literally look at it. This is called EDA. Here is the words from BMI:
   "Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions."

2. **What is Linear Regression?**
   When use linear regression, we are exploring the **linear** association between the regressor and the outcome. In general, we would like to know how changes in a varying independent variable (the predictor X) affect the dependent variable (the outcome Y).

3. **How does it work?**
   In one sentence, we are trying to minimize the loss of our fitted line. Imagine fitting a line one our data, this ultimately leads to some loss (deviations from real value). The distance (deviation) between the value that we would expect from the predictor (on the line) and the actual value (what can't be accounted for by the line) is called the residual. Linear regression minimizes the summed squared distances (L2) from the regression line to all individual points.

4. **Why we still use OLS (linear) regression?**
   1) Simple (easy to understand); 2) Regression coefficients are directly interpretable; 3) Is guaranteed to have a unique solution (beta = global minimum of squared deviations); 4) This solution is not hard to find, in fact it can be solved in closed form with linear algebra; 5) So the method is also fast.

5. **What is regression to mean effect?**
   The "regression to the mean" effect is a statistical phenomenon that occurs when a variable that is extreme on its first measurement tends to be closer to the average on its subsequent measurement. This effect is not due to any causal relationship, but rather a result of natural variability in data and the statistical tendency for extreme values to be followed by more central values. Strong regression to the mean effect happens when: 1) If low correlation between x and y; 2) Measurement is extreme

6. **What is regression fallacy?**
   (a) Interventions look more effective than they actually are due to the presence of error
   (b) The second of two successive measurements will necessarily regress toward mean if the first measurement was extreme
   (c) Example: why student doing well on first midterm → true competence(true value)+luck(error) → then in the second midterm student might not perform as well as in the first midterm since no luck(error)

3.2 Experimental Control

1. **What is experimental control?**
   xperiment control refers to the use of specific techniques in scientific research to minimize the effects of variables other than the independent variable. This process is essential for establishing a causal relationship between variables.

2. **What is a confounder?**
   Confounder: a variable that affects both independent X and dependent variables Y so that it seems X cause Y, but in reality X does not lead to Y.

3. **What are the ways to realize "control"?**

   (a) Scientific experiment
   - Key: randomization and intervention
   - Eliminate confounders
   - Implements ceteris paribus(everything else is equal)
   - Systematically vary IV to compare DV
   - Only IV varies; no other factors varing—-allow to
   - infer causality if DV are different
   - Might Not be feasible due to ethic reason
   - Example: A/B test

   (b) Natural experiment
   - No randomization, with intervention and observation
   - Nature does the intervention for you; Compare Natural occurance vs no natural occurance
   - Control is not perfect; can't isolate causal ingredient;might not replicable; it takes a while for the event to take place

4. **What makes experiment special?**
   1) Systematically varying some independent variable (IV), to create several experimental conditions. 2) Randomization: Randomly assigning units of analysis to these conditions – each to receive a particular level of the IV. → Then observation: Measurement of a dependent variable (DV) → Then math: Compare data from different conditions → Then logic.

5. **What is A/B testing?**
   With an A/B test, one element is changed between the original (a.k.a, "the control") and the test version to see if this modification has any impact on user behavior or conversion rates. From a data scientist's perspective, A/B testing is a form of statistical hypothesis testing or a significance test.

3.3 Statistic Control

1. **What is partial correlation?**
   If we already have X and Y, and we know that some Z is confounding the relation from X to Y, then we can compute the partial correlation of X and Y excluding Z.

2. **How is partial correlation computed?**

   (a) Do a simple linear regression, predicting X from Z, this gives us the X that we *can* account for by Z;

   (b) Do another simple linear regression, predicting Y from Z, which gives us the Y that we *can* account for by Z;

   (c) Correlate the residuals of these two regressions, as the residuals are - by definition - what *cannot* be accounted for by Z (in X and Y);

   (d) That correlation *is* the partial correlation $\mathbf{r}_{XY.Z}$.

3. **When to use multiple regression?**
   If more than one predictors matter/ more than one known confounder.
   (a) Control for other variables, attempt to implement ceteris paribus;
   (b) Formulation: $Y = \sum_i \beta_i x_i + \varepsilon$;
   (c) Issues: never know if all confounders are included inside analysis.

4. **How do we assess the linear model?**

   - $\mathbf{R}^2$: The coefficient of determination, defined as: (where $SS_{residual}$ is the sum of squares of residuals and SS_{total} is the total sum of squares (variance of the dependent variable).)

$$\mathbf{R}^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

   - **COD: Coefficient of Determination**
     (a) $SS_{explained}$ = The sum of squares due to regression, calculated as: (where $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of the observed data.)

$$SS_{explained} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

     (b) $SS_{residual}$ = The sum of squares of residuals, calculated as: (where y_i is the observed value)

$$SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

     (c) See COD:

$$COD = \frac{SS_{explained}}{SS_{explained} + SS_{residual}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

   - **RMSE: Root Mean Square Error** - Measures the average magnitude of the residuals or errors, defined as: (where n is the number of observations)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

5. **How do we interpret RMSE?**
   Smaller RMSE values indicate better model performance, as they suggest that, on average, the model's predictions are closer to the actual observed values; while the closer R2 is to 1, the more variance model can account for, the better model. While R^2 gives an indication of how well the model explains the variance, RMSE provides insight into the accuracy of individual predictions.

3.4 Art of Model Design

1. **What is an ideal model?**
   An Ideal model balance between variance accounted for and simplicity.

2. **What are some concerns over simplicity and model?**

   (a) Multicollinearity
      - Definition: when the predictors are themselves correlated (like use lecture attendance and lab attendance to predict grades of students but lecture attendance and lab attendance are correlated);
      - Solutions: use fewer uncorrelated predictors or regulation methods.
   (b) Sparsity/curse of dimensionality: high dimensionality → many predictors → requires a large amoung of data to estimate regression parameters.

(c) Underfitting/Overfitting (to noise)

- Since measurement = true value + errors, Fits well means fitting on the error. Therefore not generalize to a new dataset since errors are different;
- Solution: cross validation by splitting the dataset into two parts; use one part to estimate the model coefficients and the other part to calculate the model error;
- Bias=underfitting: model is too simple to reflect the true relationship/model does not fit well;
- Variance = overfitting: We are looking for models that capture most of the variance and are able to generalize to other datasets.

3. **What is Ridge/Lasso regression?**
Ridge and Lasso regression are techniques used in linear regression for regularization, which involves adding a penalty to the loss function to prevent overfitting and improve model generalization.

- **Ridge Regression (L2 Regularization):** Adds a penalty equal to the square of the magnitude of coefficients. The cost function is:

$$J(\theta) = \text{RSS} + \lambda \sum_{j=1}^{n} \theta_j^2$$

where RSS is the residual sum of squares and λ is the regularization parameter.

- **Lasso Regression (L1 Regularization):** Adds a penalty equal to the absolute value of the magnitude of coefficients. The cost function is:

$$J(\theta) = \text{RSS} + \lambda \sum_{j=1}^{n} |\theta_j|$$

Lasso can lead to sparse models with fewer coefficients; some coefficients can become zero and be eliminated from the model.

4. **Why they work?**
Ridge and Lasso regression work by imposing a penalty on the size of coefficients. This penalty term discourages overfitting by:

- **Ridge:** Shrinking the coefficients, but keeping all the features in the model.
- **Lasso:** Potentially reducing the number of features in the model, as some coefficients can shrink to zero.

These methods are particularly useful when dealing with multicollinearity or when you have more features than observations.

5. **What is cross-validation?**
Cross-validation is a statistical method used to estimate the skill of a machine learning model on new data. It involves partitioning the data into subsets, training the model on some subsets (training set) and evaluating it on the remaining subsets (validation set).

6. **How it works?**
The most common method is k-fold cross-validation:

(a) Split the dataset into k consecutive folds.

(b) For each fold:

i. Treat the fold as the validation set, and the remaining $k-1$ folds as the training set.

ii. Train the model on the training set and evaluate it on the validation set.

iii. Retain the evaluation score and discard the model.

(c) Aggregate the scores from each fold to produce a single estimation.

Cross-validation provides a robust estimate of the model's performance on an independent dataset and helps in tuning model parameters.

4 Inference

4.1 Sampling

1. **What is Law of Large Numbers?**
   If the sampling is independent and representative (each member of the population has an equal chance of being in the sample), then the sample statistics (mean, SD) will approach the population parameters (mean, SD) of the population, as one increases the sample size.

2. **What is CLT?**

   - If sampling randomly and independently, the sample means distribute normally as the sample size increases, regardless of how the underlying population is distributed;
   - Given a sufficiently large random sample from any population with a finite mean (μ) and a finite variance (σ^2), the distribution of the sample means will be approximately normally distributed, regardless of the shape of the population distribution;
   - The mean (average) of the sample means will be very close to the population mean (μ);
   - SEM:The standard deviation (standard error) of the sample means will be approximately equal to the population standard deviation (σ)/sample standard deviation divided by the square root of the sample size (n).

3. **How to make sample mean to get closer to population mean?**

   (a) Make sample size big enough;

   (b) Sample as many times as you can and get a sampling distribution, then get the mean of the sampling distribution. That mean is closest to the population mean.

4.2 Hypothesis Testing and Statistic Significance

1. **What is the basic form of testing?**
   Use sampling to Compare 2 samples to each other to decide whether these 2 samples from same population or different population; distinguish 2 samples from each other.

2. **How it works basically?**
   Null hypothesis testing framework: assume a null hypothesis/no effect, and data forces us to abandon/not abandon the null hypothesis.

3. **What is "statistically significant"?**

   - Mean: probability of data(not probability of hypothesis)
   - Mean: the probability of observed data is less than the chosen significance level / the observed data is unlikely assuming null hypothesis is true;an observed result is unlikely to be due to chance alone
   - It doesn't mean: the probability that null hypothesis is true or false, substantial, important, the probability that the alternative hypothesis is true/false
   - Importance of understanding this: it is not probability the hypothesis is true or false; If this is false, then you rule out this hypothesis. But statistically significant just means this thing can happen but is rare, it is not wrong. it is probability of data/ we can't rule out anything
   - What scientist want: P(experimental hypothesis—data)
   - What they actually get P(data—Null hypothesis)—we can only calculate this and then use bayes theorem to invert this to get e

4. **What is the framework of significance testing?**

   (a) Start proposing an expeimental hypothesis

   (b) Assume the treatment has no effect

(c) Administer the treatment to the treatment group

(d) Measure outcome

(e) Compare the outcomes in two groups. But note that any difference could be due to chance

(f) Make the decision:

- If the difference in outcomes is plausibly consistent with chance alone, we don't conclude anything about the treatment (we already started by assuming that it is ineffective, so the result changes nothing)
- If the difference in outcomes is too large to be plausibly consistent with chance, drop the assumption that the null hypothesis is true (and conclude that the treatment had an effect)

5. **What is significance level?**
The significance level, often denoted as α, is a threshold set by the researcher to determine the statistical significance of an outcome. It represents the probability of rejecting the null hypothesis when it is actually true (Type I error). Common significance levels are 0.05, 0.01, and 0.001. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

6. **What is a sign test?**
The sign test is a non-parametric test used to determine if there is a statistically significant difference between the medians of two related groups. It does not assume any particular distribution and is used when the conditions for parametric tests are not met. The test considers only the direction of change (positive, negative, no change) and not the magnitude of change. It's often used for matched-pair data or repeated measures on a single sample.

7. **How to interpret type I/II error?**

- **Type I Error:** Also known as a "false positive," this occurs when the null hypothesis is true, but is incorrectly rejected. It is directly related to the significance level. For example, a Type I error occurs if a drug is deemed effective when it is not.
- **Type II Error:** Also known as a "false negative," this occurs when the null hypothesis is false, but is incorrectly accepted. The probability of a Type II error is denoted by β, and $1 - \beta$ is the power of the test. For instance, a Type II error occurs if a drug is deemed ineffective when it actually works.

Understanding Type I and Type II errors is crucial for interpreting the results of hypothesis tests and making informed decisions in research.

4.3 Parametric Tests

1. **What is a parametric test?**
A parametric test is a statistical test that makes certain assumptions about the parameters of the population distribution from which the sample is drawn. These assumptions typically include the presumption that the data follows a normal distribution and that other parameters like variance are known or can be accurately estimated.

2. **In what situation do we need it?**
Parametric tests are appropriate when the sample size is sufficiently large (often considered as 30 or more samples), the data is approximately normally distributed, and the variance is homogeneous. They are used when these conditions are met because they offer more power and precision in hypothesis testing.

3. **What is a t-test?**
A t-test is a type of parametric test used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is particularly useful when dealing with small sample sizes, where the central limit theorem doesn't assure a normal distribution of sample means.

4. **Why cannot we just use z?**
   The z-test is used when the population variance is known and the sample size is large. However, in many practical situations, the population variance is unknown and the sample size is small. In these cases, the t-test is more appropriate as it accounts for the additional uncertainty in estimating the population standard deviation from the sample.

5. **What is degree of freedom?**
   Degrees of freedom in statistics refer to the number of independent values or quantities that can vary in an analysis without breaking any constraints. It is an important concept in the context of statistical tests, as it affects the shape of the test statistic's distribution.

6. **How do we determine degree of Freedom?**
   The determination of degrees of freedom depends on the statistical test. For example, in a simple one-sample t-test, the degrees of freedom are equal to the sample size minus one (n-1). In the case of a two-sample t-test, it's the total sample size of both groups minus two (n1 + n2 - 2).

7. **What is Welch t-test and what is the assumption?**
   The Welch t-test is an adaptation of the standard t-test which is more reliable when the two samples have unequal variances and/or unequal sample sizes. The assumption of the Welch test is that the two populations from which the samples are drawn do not need to have equal variances, making it a more flexible approach than the standard t-test.

4.4 Nonparametric Tests

1. **What are the steps in doing a Nonparametric Test?**
   The general steps for conducting a nonparametric test are:

   (a) Formulate the null and alternative hypotheses.

   (b) Choose the appropriate nonparametric test based on the type of data and the hypothesis.

   (c) Calculate the test statistic using the given data.

   (d) Determine the p-value or critical value from the appropriate distribution.

   (e) Compare the p-value to the significance level (usually 0.05) to decide whether to reject the null hypothesis.

   (f) Interpret the results in the context of the research question.

2. **What is the chi-square test and what is the assumption?**
   The chi-square test is a nonparametric test used to determine if there is a significant association between two categorical variables. Its assumptions include:

   - The sample data are randomly selected.
   - The variables are categorical (nominal or ordinal).
   - Expected frequencies in each category should be sufficiently high (usually at least 5).

3. **What is the Mann-Whitney U test and what is the assumption?**
   The Mann-Whitney U test is a nonparametric test used to compare differences between two independent groups when the dependent variable is ordinal or continuous but not normally distributed. The key assumption is that the two groups are independent and that the observations are ordinal or continuous.

4. **What is the Kolmogorov-Smirnov test and what is the assumption?**
   The Kolmogorov-Smirnov test is a nonparametric test used to determine if two samples are drawn from the same distribution. It can also be used to compare a sample with a reference probability distribution. The main assumption is that the data are continuous.

5. **What is the process of null hypothesis testing step by step?**
   The process of null hypothesis testing typically involves:

   (a) State the null hypothesis (H0) and the alternative hypothesis (H1).

(b) Choose a significance level (α, commonly set at 0.05).

(c) Select the appropriate test and compute the test statistic.

(d) Determine the p-value or critical value for the test statistic.

(e) Compare the p-value with the significance level or compare the test statistic with the critical value to make a decision about the hypotheses.

(f) Interpret the results and draw conclusions in the context of the study.

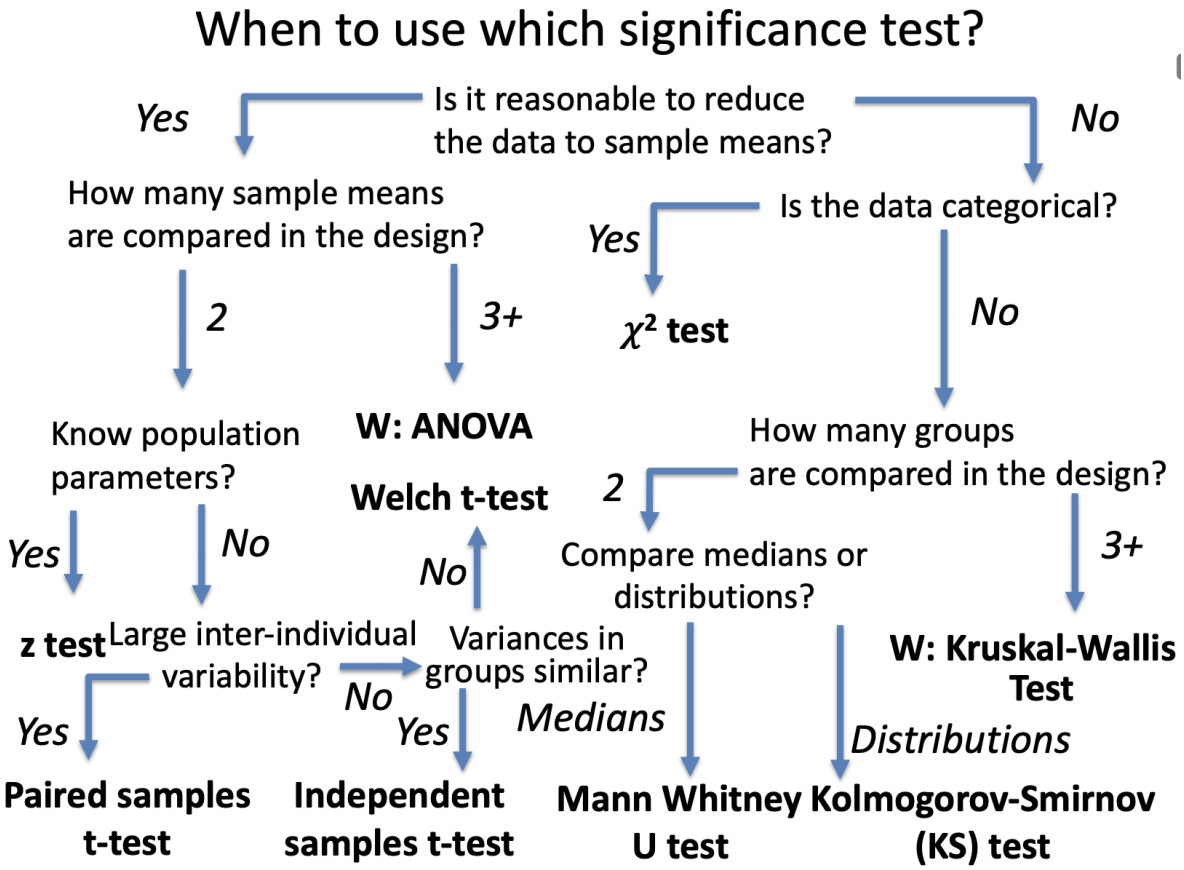

Figure 1: Choose the Right Test Statistics

5 Beyond p and r

5.1 Resampling Methods

1. **What is a Permutation Test?**
   A permutation test, also known as a randomization test, is a type of nonparametric statistical test that evaluates the significance of the difference between groups. It involves rearranging the observed data points to create new samples that could have been observed under the null hypothesis.

2. **When should we use it?**
   Permutation tests are used when the assumptions of traditional parametric tests (like normality or equal variances) are not met. They are particularly useful for small sample sizes or for data with unknown distributions. The test is applicable for comparing means, medians, variances, and other statistics.

3. **How does it work?**
   The process involves:
   (a) Combining all data points from the different groups into a single dataset.
   (b) Randomly redistributing (permuting) these combined data points into new groups.
   (c) Calculating the test statistic (e.g., difference in means) for each new group.
   (d) Repeating the process a large number of times to create a distribution of the test statistic under the null hypothesis.
   (e) Comparing the observed test statistic to this distribution to determine its significance.

4. **What is Bootstrapping?**
   Bootstrapping is a resampling method used to estimate the distribution of a statistic by sampling with replacement from the original dataset. It allows for estimating the sampling distribution of almost any statistic using random sampling methods.

5. **When should we use it?**
   Bootstrapping is particularly useful when the theoretical distribution of a statistic is complicated or unknown. It is widely used for estimating confidence intervals, standard errors, and significance tests in cases where standard methods are not applicable or reliable, especially in complex data structures.

6. **How does it work?**
   The bootstrap method involves:
   (a) Randomly drawing a sample from the original data with replacement, typically of the same size as the original sample.
   (b) Calculating the desired statistic (mean, median, variance, etc.) on this bootstrap sample.
   (c) Repeating the process many times (often thousands) to create a distribution of the statistic.
   (d) Using the distribution to estimate the standard error, confidence intervals, or other aspects of the statistic's distribution.

5.2 Effects, Power, and Confidence

1. **Understanding P-values and Their Distributions:**
   The distribution of p-values depends on the truth of the null hypothesis (H0). If H0 is true, p-values follow a uniform distribution. However, if H0 is false, the distribution may appear more like a beta distribution (not exponential). The presence of two peaks in the p-value distribution often suggests data manipulation or 'p-hacking'.

2. **Replication Crisis and Alpha Level Challenges:**
   The replication crisis in science, partly due to the inability to replicate experimental results, indicates issues beyond just setting an alpha level of 0.05. This doesn't guarantee a 5% false positive rate due to practices like p-hacking and flexible stopping, where experiments are stopped as soon as statistical significance is achieved.

3. **HARKing and Potential Solutions:**
   HARKing (Hypothesizing After the Results are Known) is another issue affecting the reliability of results. Solutions include lowering the alpha value, pre-registration of experiments, reporting effect sizes and confidence intervals, and increasing statistical power.

4. **Effect Sizes and Cohen's d:**
   Effect size, particularly Cohen's d, is a measure of the magnitude of the experimental effect. It's calculated as $d = \frac{\bar{x}_1 - \bar{x}_2}{s}$, where $\bar{x}_1$ and $\bar{x}_2$ are the means of two groups and s is the pooled standard deviation. Larger effect sizes allow for easier differentiation between groups, even with smaller sample sizes.

5. **Power Analysis and Type II Errors:**
   Power, the probability of not making a Type II error, is crucial for detecting real effects. A typical benchmark for power is 80%, but in reality, it can vary. Increasing power can be achieved by increasing sample size or effect size. The replication crisis is partly due to power failure, often resulting from small sample sizes or effect sizes.

6. **Confidence Intervals and Their Interpretation:**
   Confidence intervals provide a range in which the true parameter value is likely to lie. For example, a 95% confidence interval means that in 95 out of 100 samples, the interval will contain the true value. It's important to note that a given interval either contains or doesn't contain the true population parameter, not just the probability of containing it.

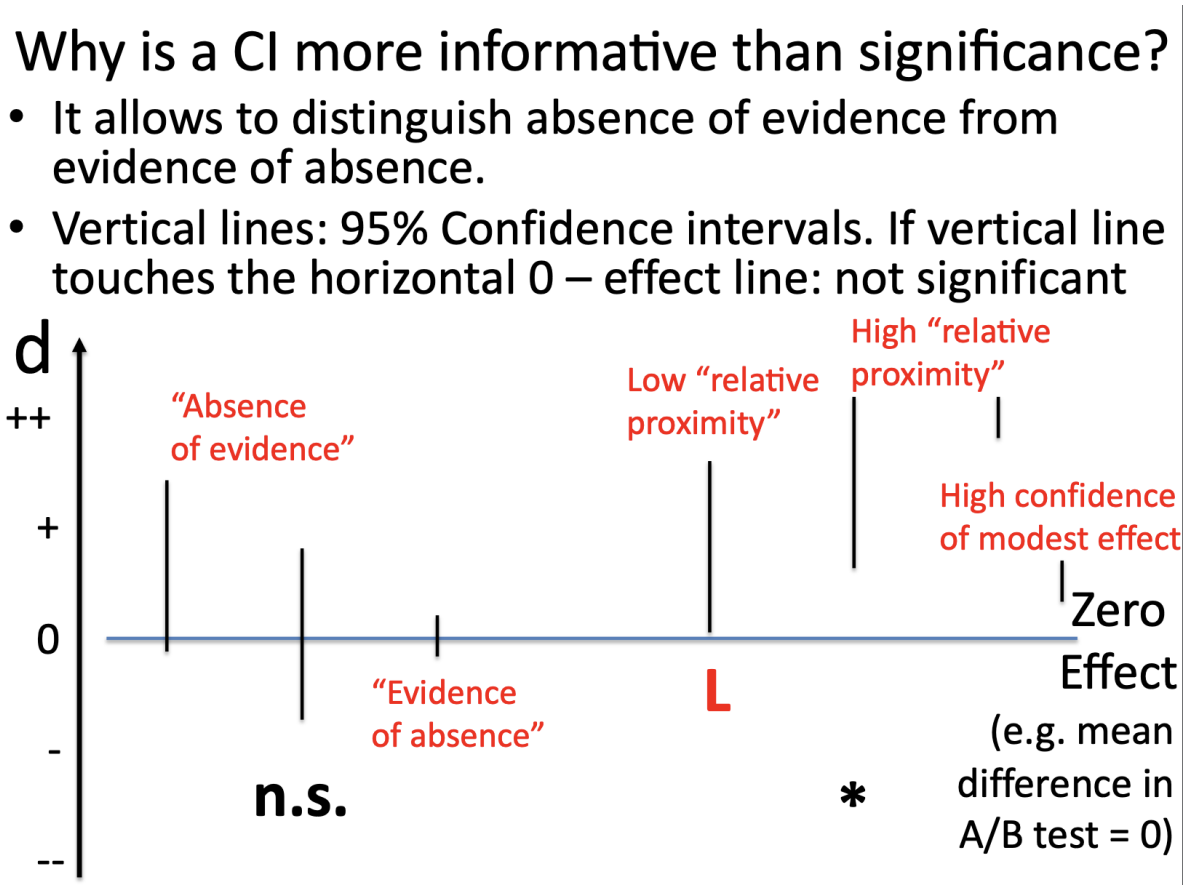

Figure 2: Why is a CI more Informative than Significance?

5.3 Bayes Theorem

1. **Definition of Bayes Theorem:**
   Bayes' Theorem is a fundamental formula in probability theory and statistics that describes how to update the probabilities of hypotheses when given evidence. It is expressed as:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

   where $P(A|B)$ is the probability of hypothesis A given the evidence B, $P(B|A)$ is the probability of evidence B given hypothesis A, $P(A)$ is the prior probability of hypothesis A, and $P(B)$ is the probability of evidence B.

2. **Importance in Statistical Inference:**
   Bayes' Theorem is the cornerstone of Bayesian statistics, allowing for the updating of probability estimates as more data becomes available. It provides a mechanism to incorporate prior knowledge with new evidence, leading to more refined and dynamic probability models.

3. **Prior and Posterior Probabilities:**
   In the context of Bayes' Theorem, the prior probability $P(A)$ represents the initial belief about the hypothesis before observing the evidence. The posterior probability $P(A|B)$ represents the updated belief after taking the evidence into account.

4. **Bayesian vs. Frequentist Approach:**
   Bayesian statistics, underpinned by Bayes' Theorem, differs from the frequentist approach by incorporating prior probabilities and continually updating beliefs as new data is acquired. In contrast, the frequentist approach relies solely on the likelihood of observed data under various hypotheses without considering prior probabilities.

5. **Applications in Various Fields:**
   Bayes' Theorem has wide-ranging applications including in machine learning, medical diagnostics, spam filtering, and decision-making processes in various domains.

6. **Controversies and Limitations:**
   The use of prior probabilities in Bayesian analysis can be subjective, leading to controversy over the choice of priors. However, with sufficient data, the influence of the prior diminishes, and the posterior probability converges on the true value.

5.4 Bayesian Statistics

1. **Definition of Bayesian Statistics:**
   Bayesian statistics is an approach to data analysis and statistical inference that emphasizes the use of probability to represent uncertainty in all aspects of statistical modeling. It integrates Bayes' Theorem to update the probability estimate for a hypothesis as more evidence or information becomes available.

2. **Bayesian Inference:**
   Bayesian inference involves using Bayes' Theorem to update the probability distribution of a parameter as more data is observed. Unlike frequentist inference, which relies on fixed data sets to make predictions, Bayesian methods adaptively update and refine these predictions.

3. **Priors, Likelihood, and Posterior:**
   In Bayesian analysis, the prior distribution reflects the initial belief about a parameter before observing the data. The likelihood represents the probability of observing the data given the parameters. The posterior distribution, which is the result of Bayes' Theorem, combines the prior and the likelihood, yielding updated knowledge about the parameter.

4. **Markov Chain Monte Carlo (MCMC) Techniques:**
   MCMC methods are computational algorithms used in Bayesian statistics to approximate the posterior distribution when it is difficult to compute directly. These methods generate samples from the posterior distribution and can handle complex models with multiple parameters.

5. **Advantages of Bayesian Statistics:**
   Bayesian statistics offers a flexible framework for modeling complex systems and incorporates prior knowledge or expertise. It provides a complete probabilistic description of model parameters and is particularly useful in situations with limited or incomplete data.

6. **Applications and Limitations:**
   Bayesian methods are widely used in various fields such as machine learning, epidemiology, and environmental science. However, the approach can be computationally intensive, and the results can be sensitive to the choice of the prior distribution, especially with limited data.

6 Machine Learning

6.1 Logistic Regression and ROC Curve

1. **What is logistic regression?**
   It is used when we need to predict a binary outcome. Note that the link between predictor and outcome might not be linear.

2. **What is a logistic function?**
   The logistic function is an inverse logit function, which looks like S-shaped. It is also called a "sigmoid" function. See below:

$$logit^{-1}(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

   Note that here the logit function is the natural log of the odds, which is given below:

$$logit(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

   Thus we can derive p as follow:

$$p = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

3. **Interpretation of β_0 (Intercept)**

   - The intercept β_0 shifts the sigmoid curve left or right on the x-axis.
   - It represents the log-odds of the outcome when $X_1 = 0$.
   - A positive β_0 shifts the curve to the left, and a negative β_0 shifts it to the right.
   - The 50% probability threshold is at $-\beta_0/\beta_1$.

4. **Interpretation of β_1 (Slope)**

   - The slope β_1 determines the steepness of the sigmoid curve.
   - A larger absolute value of β_1 indicates a steeper curve.
   - If β_1 is positive, an increase in X_1 increases the probability; the curve slopes upwards.
   - If β_1 is negative, an increase in X_1 decreases the probability; the curve slopes downwards.

5. **Why "accuracy" alone does not work?**
   However, this will fail in all but the best-case scenarios, as most real datasets will be imbalanced, e.g. we're trying to predict sth relatively rare, e.g. fraud. If only 1% of cases are fraud, a model that simply always predicts "no fraud"would be 99% accurate, and hard to beat. Thus we need the **AUC** method. Most commonly used to assess the quality of classification models: AUC (Area under ROC curve), a.k.a. AUROC

6. **What is AUROC?(1)**
   The Area Under the Receiver Operating Characteristic Curve (AUROC) is a performance measurement for classification problems at various threshold settings. It tells how much a model is capable of distinguishing between classes. Higher the AUROC, better the model is at predicting 0s as 0s and 1s as 1s. An excellent model has AUROC near to the 1, which means it has a good measure of separability. A poor model has AUROC near to the 0, which means it has the worst measure of separability. It is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

7. **What is AUROC?(2)**
   The Receiver Operating Characteristic (ROC) curve is a plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The formula for TPR and FPR are:

- **True Negative Rate (TNR)** or Specificity = $\frac{\text{True Negatives}}{\text{True Negatives+False Positives}}$
- **True Positive Rate (TPR)** or Sensitivity = $\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$
- **False Positive Rate (FPR)** or 1 - Specificity = $\frac{\text{False Positives}}{\text{False Positives + True Negatives}}$

The AUROC is the area under the ROC curve. It provides an aggregate measure of performance across all possible classification thresholds.

6.2 Classification

1. **What is the "blessing of dimensionality"?**
   In 1-D, there is too much overlap to make classification feasible; in 2-D, the subpopulations are much easier to separate.

2. **What is Leaf Impurity?**
   Leaf impurity in decision trees is a key concept used to measure the purity of data in the leaf nodes of a decision tree. This concept can be intuitively understood by imagining a basket filled with balls of different colors. If all balls are of the same color, the basket is considered "pure." Conversely, if the basket contains balls of many colors, it is "impure."

3. **Why do we need Gini Index?**
   In decision trees, we aim for the leaf nodes to be as pure as possible. This ensures higher confidence when classifying new data. To quantify this purity, we use measures like the Gini index.
   The Gini index is a method to measure impurity. Its value ranges from 0 to 1, where 0 indicates complete purity (all instances in the node belong to the same class) and 1 indicates maximum impurity (instances are evenly distributed across all possible classes). The Gini index is calculated using the formula:
   $$\text{Gini Index} = 1 - \sum_{i=1}^{J} p_i^2$$
   Here, p_i is the proportion of instances belonging to class i in the node, and J is the number of classes.

4. **What is the purpose of Impurity/Gini index?**
   In summary, leaf impurity is a measure of data purity in the leaf nodes of a decision tree, and the Gini index is a commonly used method to calculate this impurity. By minimizing the Gini index, the decision tree aims to create purer leaf nodes, enhancing the model's classification accuracy.

5. **What is Tree Pruning?**
   Tree pruning is employed to reduce overfitting and simplify the model in decision tree learning.

6.3 Matrix Factorization

1. **What is Matrix Factorization?**
   Decomposing a matrix into component matrices. Component matrices can be multiplied back into the original matrix or some kind of approximation of it.

2. **Why is it helpful?**
   1) It lowers the dimension of the original data to make it more comprehensible. If we lower it using some techniques, we can make it more intuitive and insightful. 2) Highlights the important patterns and relationships. 3) Widely used in signal and communication fields for size reduction (compression).

3. **How do we find the ideal decomposed matrices?**
   We typically use **Gradient Descent**. It is an optimization algorithm to optimize parameters (minimize loss). For Matrix Factorization, the goal is for the components to multiply to the original matrix, here is how:

   (a) Randomly initialize component matrices;

(b) Multiply them together and calculate the error between original matrix and newly formed matrix;

(c) Take the gradient of the loss and update feature values (components) accordingly.

4. **What is Eigenvalue?**
Eigenvalues represent the magnitude of transformation. Imagine a linear transformation applied to a vector space. Some vectors in this space may change direction, but others may only stretch or shrink. Eigenvalues quantify this stretching or shrinking factor. Mathematically, if A is a matrix representing the transformation, and λ is an eigenvalue, then for some non-zero vector v, the equation $Av = \lambda v$ holds. Here, v is stretched by the factor λ.

5. **What is Eigenvector?**
Eigenvectors, on the other hand, are the vectors that only get scaled (stretched or shrunk) and do not change their direction under a linear transformation. They are the"special" vectors for which the action of the transformation is as simple as possible - just scaling. For the same matrix A and eigenvalue λ, the corresponding eigenvector v satisfies $Av = \lambda v$. This means, applying the transformation A to v results only in scaling v by λ, without rotating or changing its direction.

6. **Why are they important?**
In essence, eigenvalues and eigenvectors reveal the hidden nature of linear transformations, indicating how the space is stretched and which directions remain unchanged in their orientation.

7. **SVD, PCA, and Matrix Factorization**
"SVD" (Singular Value Decomposition), "PCA" (Principal Component Analysis), and"Matrix Factorization" are closely related concepts in mathematics and data analysis but have key differences.

   (a) **Singular Value Decomposition (SVD)**
   - *Definition*: SVD is a method of decomposing any matrix into the product of three specific matrices, formulated as $A = U\Sigma V^*$, where A is the original matrix, U and V are orthogonal matrices, and Σ is a diagonal matrix with singular values.
   - *Application*: SVD has wide applications in signal processing, statistics, computer science, and is used in implementing PCA.

   (b) **Principal Component Analysis (PCA)**
   - *Definition*: PCA is a statistical method for transforming a set of possibly correlated variables into a set of linearly uncorrelated variables, known as principal components, via orthogonal transformation.
   - *Implementation*: PCA can be performed by eigen decomposition of the data's covariance matrix or more commonly, through SVD for numerical stability.
   - *Purpose*: The primary use of PCA is in data dimensionality reduction, retaining the most significant principal components.

   (c) **Matrix Factorization**
   - *Definition*: Matrix factorization is the process of breaking down a matrix into the product of two or more matrices, encompassing methods like SVD and PCA.
   - *Application*: This technique is used in various fields including machine learning, image processing, and recommendation systems.

8. **Relationship and Differences**:

   - *Relationship*: SVD is a method of implementing PCA, both being forms of matrix factorization.
   - *Differences*: SVD is a general decomposition technique, while PCA focuses on finding the main variance directions in data. Matrix factorization is a broader term that includes these methods and more.

6.4 Dimentionality Reduction Methods

1. **What is Principal Component Analysis (PCA)?** Principal Component Analysis (PCA) is a widely used technique in statistical data analysis and machine learning for dimensionality reduction and feature extraction. Let's delve into its core concepts:

2. **How to do PCA?** PCA transforms a set of possibly correlated variables into a set of linearly uncorrelated variables, called principal components. These components are ordered such that the first few retain most of the variation present in all of the original variables.

3. **How PCA Works?**

   (a) **Standardization**: The first step in PCA is usually to standardize the data so that each feature contributes equally to the analysis.

   (b) **Covariance Matrix Computation**: PCA starts with the computation of the covariance matrix to understand how the variables in the dataset are varying from the mean with respect to each other.

   (c) **Eigen Decomposition**: The covariance matrix is then decomposed into its eigenvectors and eigenvalues. The eigenvectors represent the directions of maximum variance, and the eigenvalues represent the magnitude of these directions.

   (d) **Selection of Principal Components**: Principal components are selected based on the eigenvalues. The eigenvector with the highest eigenvalue is the first principal component, and so on.

   (e) **Dimensionality Reduction**: The original data can then be projected onto these principal component axes to reduce the dimensionality of the dataset while retaining most of the original variance.

4. **What are some Benefits of PCA?**

   - Reduces the dimensionality of data, thus simplifying the model.
   - Helps in identifying hidden patterns in the data.
   - Reduces noise and improves the interpretability of the dataset.

5. **What are some Limitations of PCA?**

   - PCA assumes linear relationships between features.
   - It is a method that relies heavily on the variance of the data.
   - Important information might be lost during dimensionality reduction.

6.5 Clustering

1. **What is K-means Clustering Algorithm?**
   K-means is a popular unsupervised learning algorithm used for clustering data into a predetermined number of groups. It is widely used in data mining, market segmentation, and image compression. K-means clustering aims to partition a set of observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

2. **How K-means Works?**

   (a) **Initialization**: Start by selecting K initial centroids randomly from the data points.

   (b) **Assignment Step**: Assign each data point to the nearest centroid, creating K clusters.

   (c) **Update Step**: For each of the K clusters, update the centroid by calculating the mean of all points assigned to that cluster.

   (d) **Iteration**: Repeat the assignment and update steps until the centroids no longer change significantly, indicating that the clusters are as optimized as possible given the initial centroid choices.

(e) **Convergence**: The algorithm converges when the assignments no longer change or the changes are minimal.

3. **What are the Benefits of K-means?**
   - Simple and easy to implement.
   - Efficient in terms of computational cost.
   - Effective in clustering data into spherical-shaped clusters.

4. **What are the Limitations of K-means?**
   - The number of clusters K must be specified in advance.
   - Sensitive to the initial choice of centroids.
   - May not perform well with clusters of varying size and density.
   - Not suitable for identifying clusters with non-convex shapes.

5. **How do we select the Optimal Number of Clusters k?**
   Choosing the right number of clusters (k) in clustering algorithms like K-means is crucial. Two popular methods for determining the optimal k are the Elbow Method and the Silhouette Method.

6. **What is Elbow Method?**
   The Elbow Method involves running the clustering algorithm for a range of values of k (say, $k = 1$ to 10), and for each value, computing the sum of squared distances from each point to its assigned center. When these overall distances are plotted against the number of clusters, the "elbow" point, where the rate of decrease sharply changes, represents a good estimate for k.

7. **How Elbow works?**
   (a) Compute clustering for different values of k.
   (b) For each k, calculate the total within-cluster sum of square (WSS).
   (c) Plot the curve of WSS according to the number of clusters k.
   (d) The location of a bend (elbow) in the plot is generally considered as an indicator of the appropriate number of clusters.

8. **What are some Advantages and Limitations?**:
   - Simple and easy to implement.
   - The method can be somewhat subjective as it's sometimes hard to identify the exact elbow point.

9. **What is Silhouette Method?**
   The Silhouette Method measures how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

10. **How Silhoutte works?**:
    (a) For each point, calculate the average distance from the points in the same cluster (cohesion) and the average distance from the points in the nearest cluster (separation).
    (b) The silhouette score for each point is then $\frac{\text{separation}-\text{cohesion}}{\max(\text{separation},\text{cohesion})}$.
    (c) Calculate the mean silhouette score for all points for different values of k.
    (d) The value of k that maximizes the average silhouette score is considered as the optimal number of clusters.

11. **What are some Advantages and Limitations?**:
    - Provides a more objective way to determine the number of clusters.
    - Works well when the clusters are distinctly separated.

- Computationally intensive, especially for large datasets.

12. **What is DBSCAN Clustering Algorithm?**
    DBSCAN is a widely used density-based clustering algorithm that is particularly effective for datasets with noise and clusters of varying shapes and sizes. It stands out for its ability to identify outliers and discover arbitrarily shaped clusters. DBSCAN clusters data based on density. It groups together points that are closely packed together while marking points that lie alone in low-density regions (whose nearest neighbors are too far away) as outliers.

13. **How DBSCAN Works?**

    (a) **Key Parameters**:
    - *Epsilon (ε)*: The radius around a data point to search for neighboring points.
    - *Minimum Points (MinPts)*: The minimum number of points required to form a dense region.

    (b) **Classification of Points**: Points are classified into three types:
    - *Core Points*: A point with at least MinPts within its ε-neighborhood.
    - *Border Points*: A point that has fewer than MinPts within its ε-neighborhood but is in the neighborhood of a core point.
    - *Noise Points*: Points that are neither core nor border points.

    (c) **Cluster Formation**:
    - Start with an arbitrary point and retrieve all points density-reachable from this point based on ε and MinPts.
    - If the point is a core point, form a cluster. If it's a border point, move to the next point.
    - Continue the process until all points are classified into clusters or marked as noise.

    (d) **Handling Noise**: Points classified as noise are excluded from the clusters.

14. **What are some applications of DBSCAN?**
    - Identifying fraudulent activities in banking transactions.
    - Segmenting spatial data like geographic information systems.
    - Analyzing traffic patterns and environmental data.

15. **What are some Benefits of DBSCAN?**
    - Capable of finding arbitrarily shaped clusters.
    - Robust to outliers and noise.
    - Does not require specification of the number of clusters.

16. **What are some Limitations of DBSCAN?**
    - Determining appropriate values for ε and MinPts can be challenging.
    - Not well-suited for high-dimensional data.
    - Varying densities within the same dataset can cause difficulties in clustering.

7 Perspectives

7.1 Truths and Consequences

1. **What is Pseudoreplication?**
   Pseudoreplication refers to the error in statistical analysis where the independence of data points is assumed incorrectly, leading to misleading results.

2. **Why is it serious?**
   It can lead to false positives in scientific studies, affecting the validity and reliability of the conclusions.

3. **What is the law of small numbers?**
   It is the erroneous belief that small samples must reflect the population characteristics, often leading to overgeneralizations.

4. **What is Simpson's Paradox?**
   It occurs when a trend appears in different groups of data but disappears or reverses when these groups are combined.

5. **Why is it happening?**
   This paradox often happens due to confounding variables that are not accounted for in the analysis.

6. **What is Gambler's (a.k.a Hot Hand) Fallacy?**
   It is the mistaken belief that a sequence of random events, like coin flips, can affect the likelihood of future events.

7. **What is Survivorship Bias?**
   It is a logical error of focusing on the surviving subjects of a process and overlooking those that did not survive due to their lack of visibility.

7.2 Natural Language Processing

1. **Why is human language rare?**
   Human language is complex and nuanced, involving intricate syntax, semantics, and pragmatics, making it a unique form of communication.

2. **What is the language trinity?**
   It refers to the three critical components of language: syntax (structure), semantics (meaning), and pragmatics (contextual use).

3. **What is Gricean maxims?**
   These are guidelines for effective communication, including principles like relevance, quantity, quality, and manner.

4. **How is relevance determined?**
   Relevance is determined by the context, the speaker's intentions, and the listener's interpretation and expectations.

5. **What is tokenization?**
   Tokenization in NLP is the process of breaking down text into smaller units, such as words or phrases, for analysis.

6. **How does predictive text work?**
   Predictive text algorithms analyze the context and the user's typing patterns to suggest the next word or phrase.

7.3 The Principled Future of Data Science

1. **What are the principles?**

   - Measurement/Parameter/Test;
   - The right procedure based on the number of DVs;
   - Descriptive statistics;
   - Inferential statistics;
   - Statistical reliability of the conclusions about populations from the sample;

- "New" statistics;
- Machine learning and statistical decision theory.

2. **What are the specific goals of this course?**
    - Find out if Data Science is for you;
    - Build confidence and competence;
    - Learn the principles of Data Science;
    - Planting seeds for advanced topics;
    - "Secret shelf" $\Leftarrow$ Data Science mindset;
    - Appreciation for unknown unknowns.