Statistics

Data

Types of Data:

Category	Type	Description	Example
Numerical (numbers)	Interval	Numeric scale with meaningful intervals	Temp. in C
	Ratio	Interval, but also with meaningful zero	Height in cm
	Discrete	No arbitrary precision (integers)	Population
Categorical (labels)	Ordinal	Sortable, discrete	Education (high school, uni.)
	Nominal	Non-sortable, discrete	Movie genre (sic-fi, romcom)

Sample vs Population Data:

Population data: Data from all members of a group. E.g. Salaries in a department; Number of lions in a zoo; Facebook group members. Parameters are μ, β, σ^2

Sample Data: Data from some members of a group (hopefully randomly selected). E.g. Average height of Italians; Salaries of teachers; Brightness of stars in Milky Way. Parameters are $\hat{\mu}, \hat{\beta}, \widehat{\sigma^2}$.

- Most statistical procedures are designed either for sample or for population data.
- Applying a procedure to the wrong data type can lead to incorrect results and incorrect interpretations.

Sample Sizes and Names:

N > 1: Actual research with potentially important and generalizable findings. Can suffer from sampling variability, noise and outliers. Appropriate statistical analyses can facilitate interpretation.

- Pilot study
- Proof-of-principle
- Small-scale study
- Large-scale study

N = 1: Interesting stories that may inspire future research, but that (most often) should not be overly trusted or generalized. Highly likely to be noise, outliers or other non-representative data (otherwise why report it?). Statistical inference is difficult or impossible.

- Case study: One patient
- Anecdote: One person

Visualization

Bar Plot:

What kinds of data can be shown in a bar plot?

⇨ Categorical (nominal and ordinal) and Numerical (only discrete)

Conclusions:

⇨ Numerical data must be converted to discrete type for bar plot visualization.
⇨ Bin number and boundaries are important.
⇨ Sometimes the terms *histogram* and bar plot are often used interchangeably.

What are the purposes of error bars?

⇨ Standard deviation, standard error or confidence intervals

Can error bars be misinterpreted or used for evil purpose?

⇨ Yes! (To the first question). It's important to use error bars correctly and explain them clearly.

Descriptive Statistics

Descriptive Statistics: Describing the characteristics of a dataset:

⇨ Mean, median, mode
⇨ Variance
⇨ Kurtosis, skew
⇨ Distribution shape

- ⇨ Spectrum

Note: No relation to population; no generalization to other datasets or groups.

Inferential Statistics: Using features of the dataset to make claims about a population.

- ⇨ P-value
- ⇨ T/F/chi-square value
- ⇨ Confidence intervals
- ⇨ Hypothesis testing

Note: The entire purpose is to relate data features to populations or generalize to other groups!

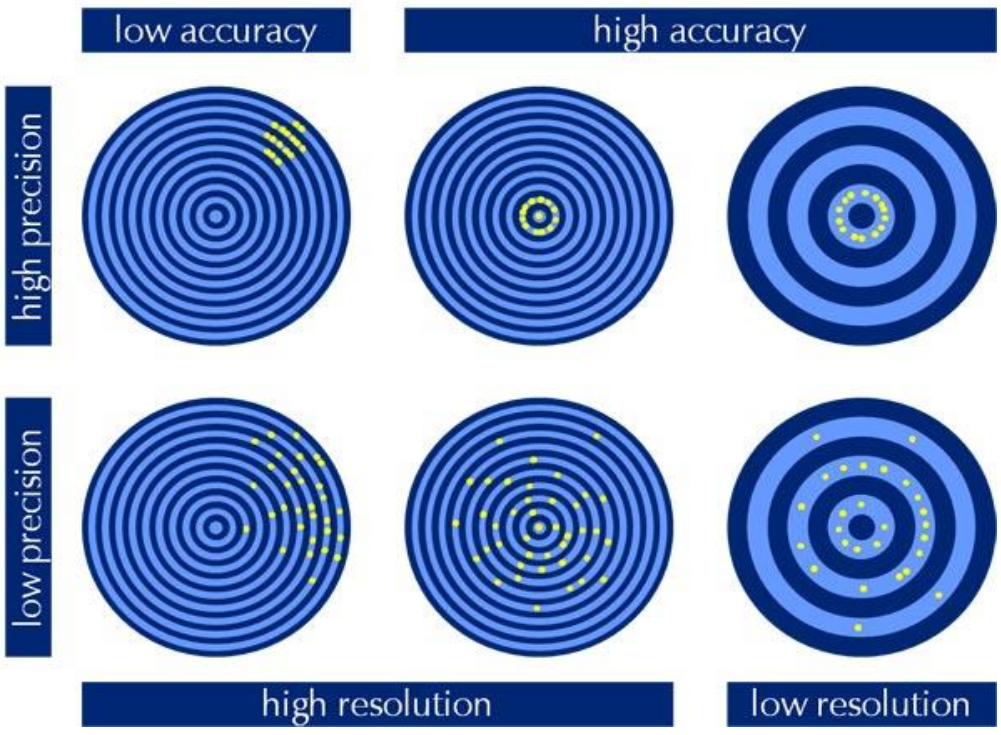
Accuracy, Precision & Resolution:

https://meettechniek.info/measurement/accuracy.html#:~:text=The%20following%20terminology%20are%20often,magnitude%20from%20the%20measured%20value

Accuracy: The relationship between the measurement and the actual truth (inversely related to *bias*).

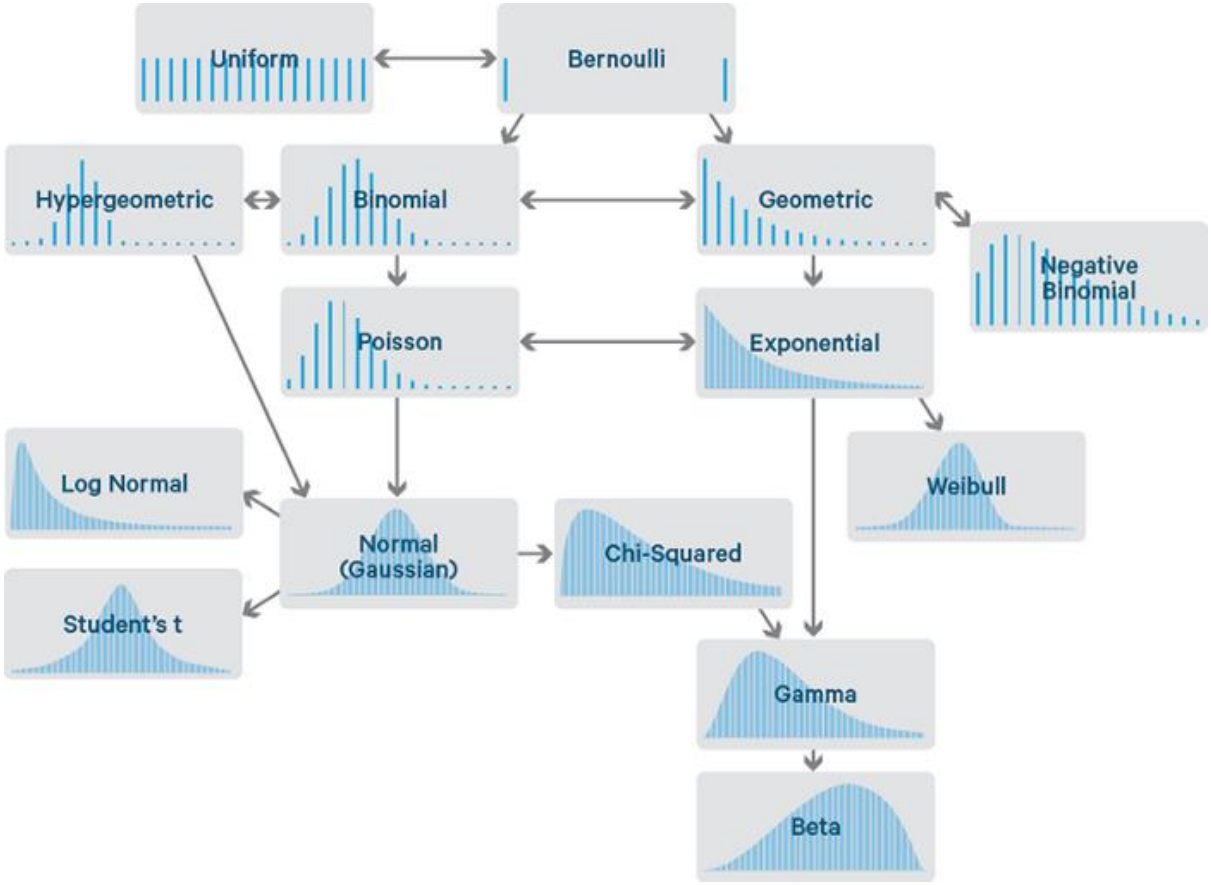
Precision: The certainty of each measurement (inversely related to *variance*)

Resolution: The number of data points per unit measurement (time, space, individual, …)



Data Distributions:

Actually, it is the shape of histogram.



Why cares about distributions?

- ⇨ Most statistical procedures are based on assumptions about distributions.
- ⇨ Knowing these distributions is necessary for appropriately applying statistics.
- ⇨ Data distributions provide insights into nature.
- ⇨ Physical and biological systems are modelled using distributions.

Number of Modes

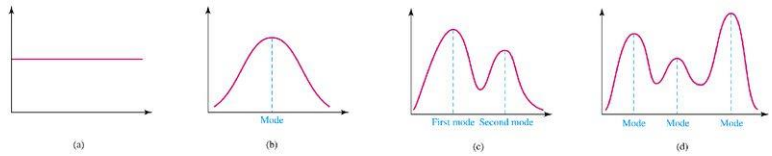


Figure 4.4

Figure 4.4a shows a distribution, called a **uniform distribution**, that has no mode because all data values have the same frequency.

Figure 4.4b shows a distribution with a single peak as its mode. It is called a **single-peaked**, or **unimodal**, distribution.

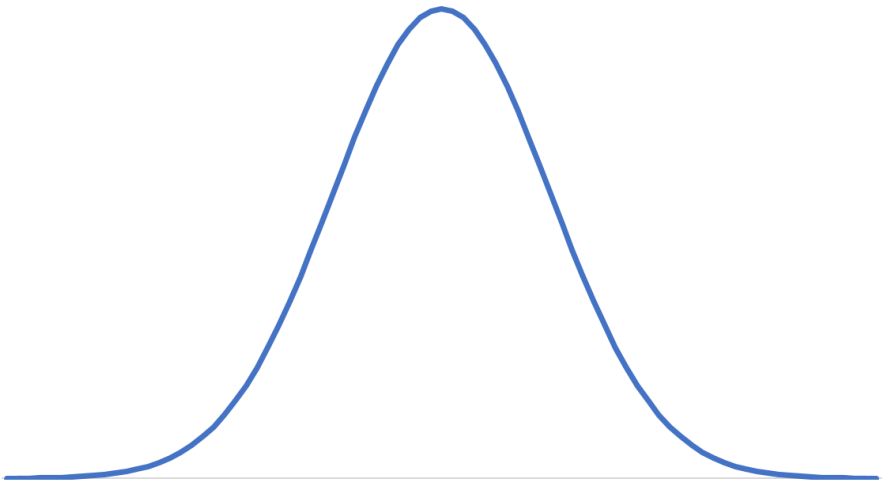
Python Code for different distribution:
https://colab.research.google.com/drive/1QCoJrqhwHTclglRo4IFAyFG1h1J_b7z4?usp=sharing

The Beauty and simplicity of Normal:

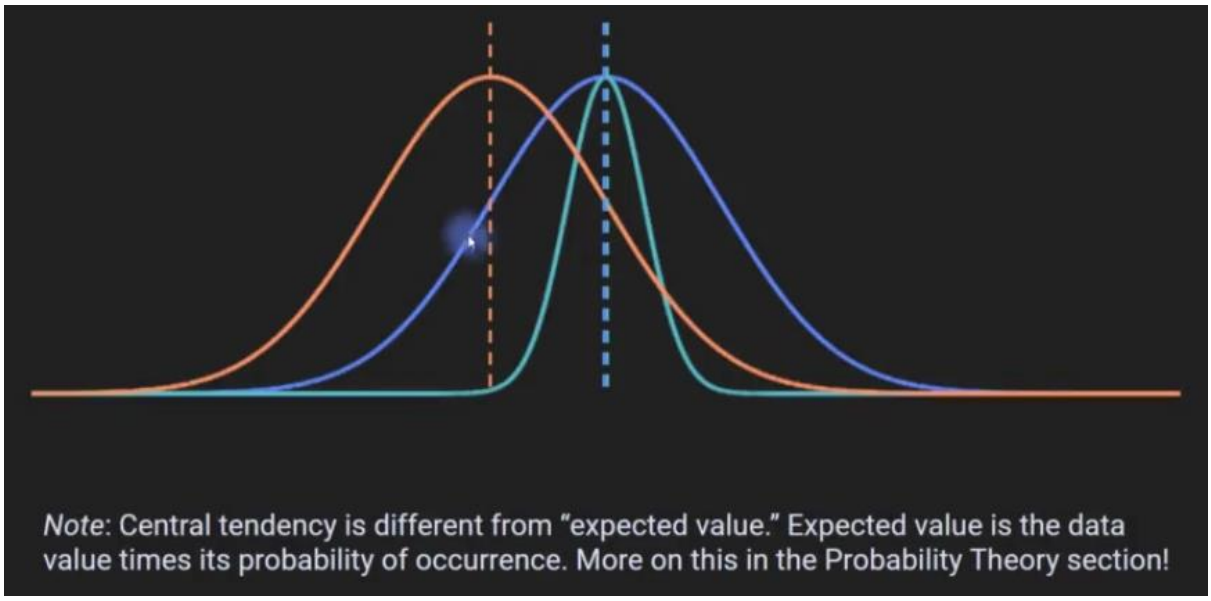
Gaussian Formula:

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

$$=> f(x) = e^{-x^2}$$



Measures of Central Tendency:



✓ **Mean:**

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

- **Suitable for:** roughly normally distributed data.
- **Suitable data types:** interval, ratio
- **Example:** x = [-2, 0, 4, 1, 7]. The mean is 2

Is the mean suitable for discrete data?

⇨ Example: Average US family has 1.9 children.

Is the mean suitable for ordinal data?

⇨ Example: Average course rating of 4.3 stars (out of 1-5 starts; whole star ratings only)

Is the mean suitable for nominal data?

⇨ Example: Average person likes 1.7 ice cream flavour (1 = chocolate, 2 = vanilla, 3 = strawberry)

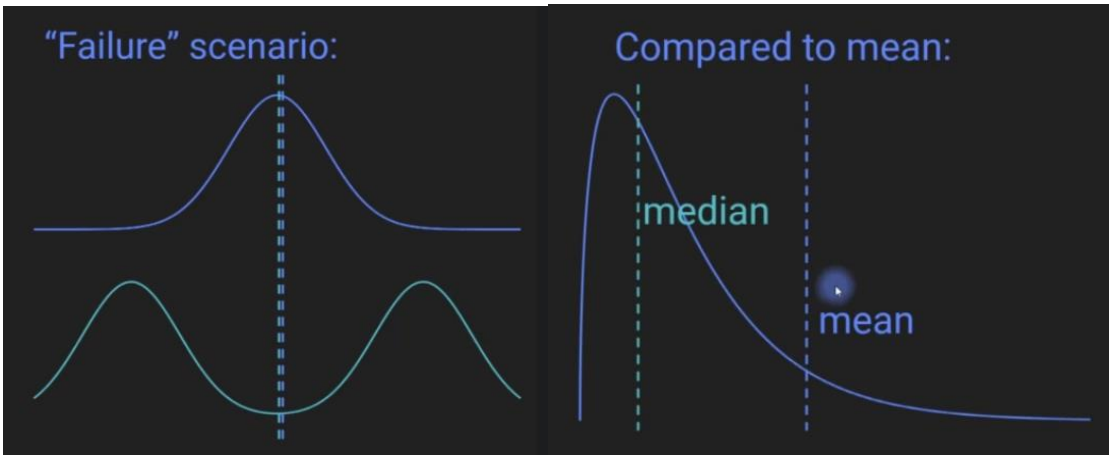
Conclusions:

⇨ The mean is best applied to interval and ratio scale data.
⇨ The mean of discrete and ordinal data can be useful, but must be carefully interpreted.
⇨ The mean is not appropriate for nominal data.

✓ **Median:**

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

- **Suitable for:** unimodal distributions
- **Suitable data types:** Interval, ratio
- **Example:** x = [0, 4, 1, -2, 7]. med(x) = [-2, 0, **1**, 4, 7] or 1. For even number dataset, x = [10, 0, 4, 1, -2, 7]. Med(x) = [-2, 0, **1**, **4**, 7, 10] or 1 and 4 = (1+4)/2 = 2.5



✓ **Mode:**

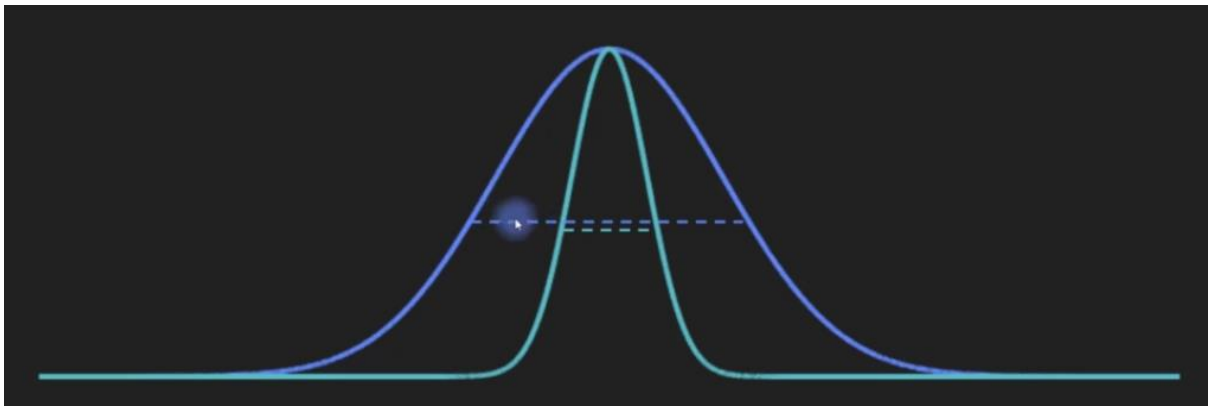
- **Formula:** Most common value
- **Suitable for:** Any distribution
- **Suitable data types:** Any (numerical data should be converted to discrete)
- **Example:** x = [0, 0, 1, 1, 1, 2, 3, 4]; mode(x) = 1
- It is possible that, a dataset can have more than one mode.

✓ **Summary:**

- **Mean:** Average value, sensitive to outliers. Most commonly used measure of central tendency.
- **Median:** Middle value (50% below, 50% above)
- **Mode:** Most common value

Python Code:
https://colab.research.google.com/drive/1shiOFSTpu5jTxgJ_T6C7Ha6dJt7NpvVx?usp=sharing

Measures of Dispersion:



✓ **Variance:**

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

- **Suitable for:** Any distribution
- **Suitable data types:** Numerical, Ordinal (but requires mean)
- **Example:** x = [8, 0, 4, -1, -2, 7]; var(x) = mean([5, -3, 1, -2, -5, 4]^2); var(x) = 16
- **Example:** x = [2, 3, 4, 3, 4, 4]; var(x) = 0.667

Why mean-center $\bar{x}$?

Variance indicates the dispersion around the average; the following two datasets should have the same variance:

d1 = [1 2 3 3 2 1]

d2 = [101 102 103 103 102 101]

Why are the differences squared?

We want the distances to the average; without squaring the variance would be 0.

d1 = [1 2 3 3 2 1]

Mean-centered d1 = [-1 0 1 1 0 -1] => sums to 0!

Why not take the absolute value ("mean absolute difference")?

Squaring: emphasizes large values; is better for optimization (continuous and differentiable); is closer to Euclidean distance; is the second "moment" of the distribution; better link to least-squares regression; other nice properties.

MAD: Also, good; robust to outliers; less commonly used.

Why divided by n-1?

Dividing by N-1 is for *sample variance*

Dividing by N is for *population variance*

Population mean is a theorical quantity, whereas the sample mean is an empirical quantity.

The population mean of a die is 3.5. Roll a die 4 times (sample). The sample mean is 3. How die rolls do you need to see to know all 4 values? => 3. For example: 1, 2, 4, ? Hence, there are N-1 free values, or degrees of freedom.

- ✓ **Standard Deviation:**

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

- ✓ **Fano factor:**

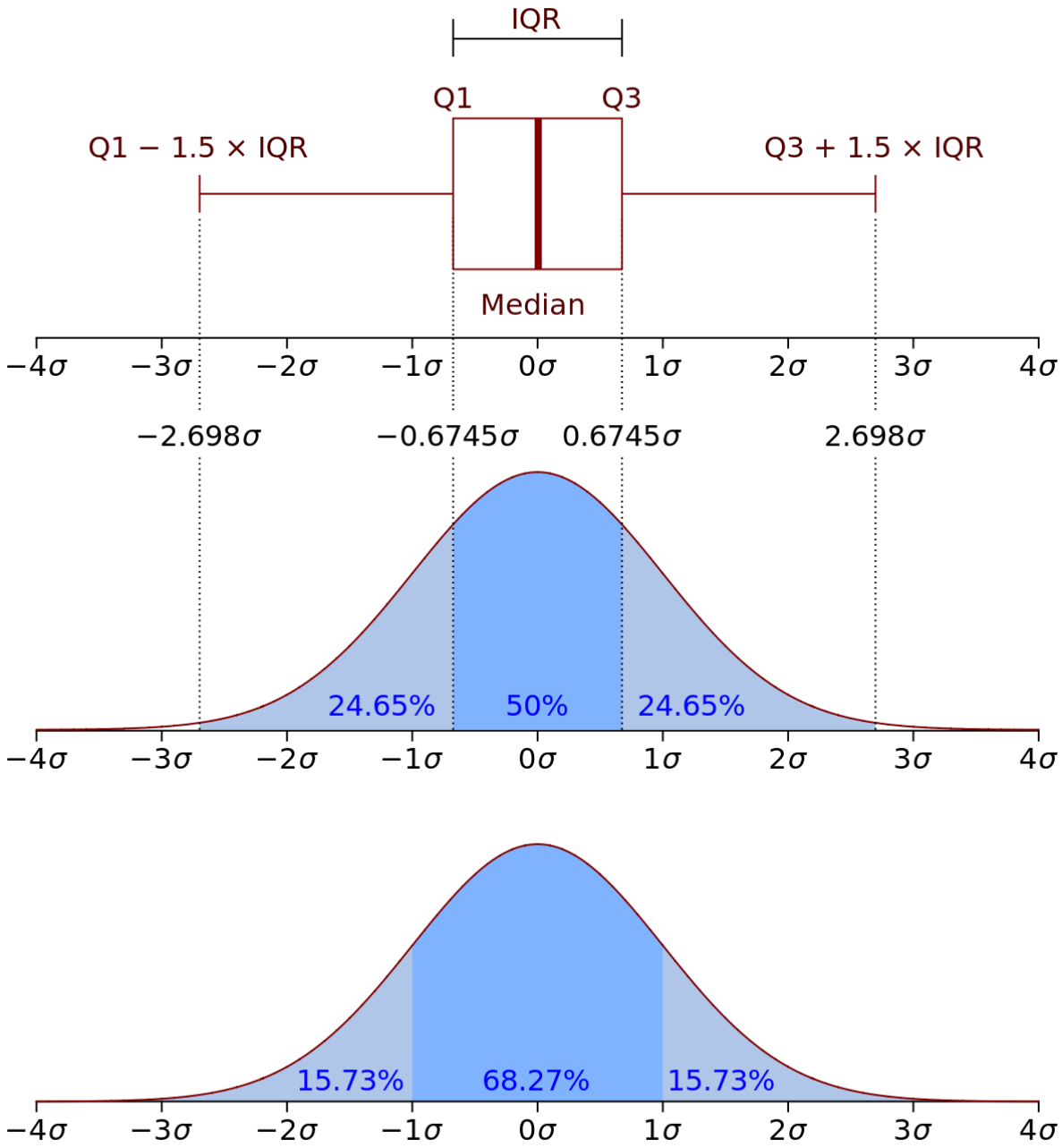
$$F = \frac{\sigma^2}{\mu}$$

- ✓ **Coefficient of variation:**

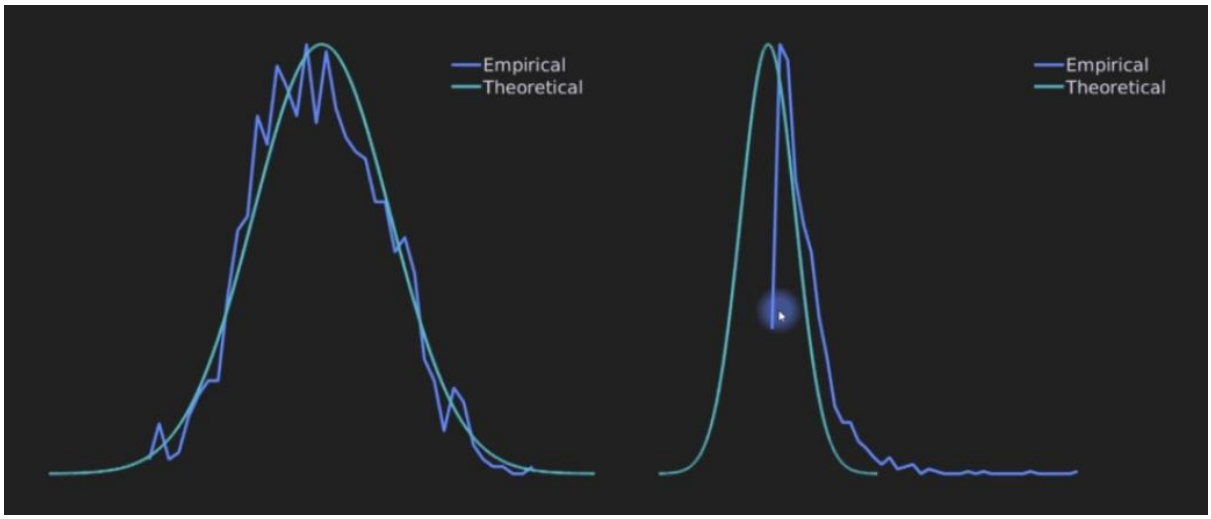
$$CV = \frac{\sigma}{\mu}$$

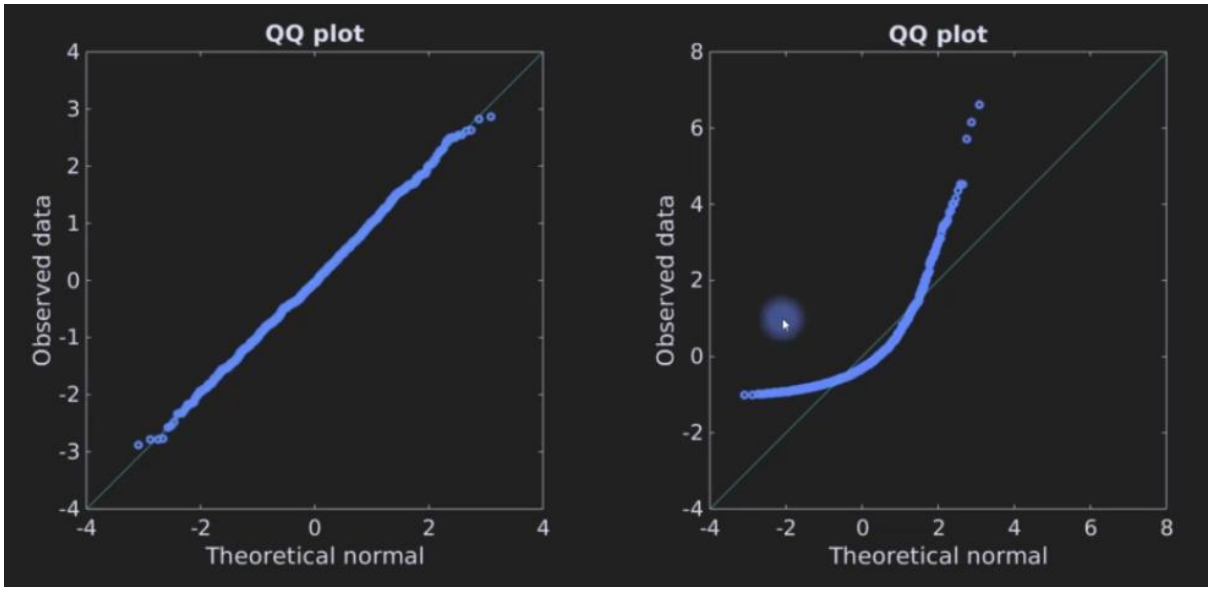
Python Code:
https://colab.research.google.com/drive/1Pv7h9bKBZDidPQrnfvpXB8XZvX9lmcI0?usp=sharing

Interquartile Range (IQR):



QQ (Quartile-Quartile) Plot:





Statistical Moments:

- ✓ **Unstandardized Statistical Moments:**

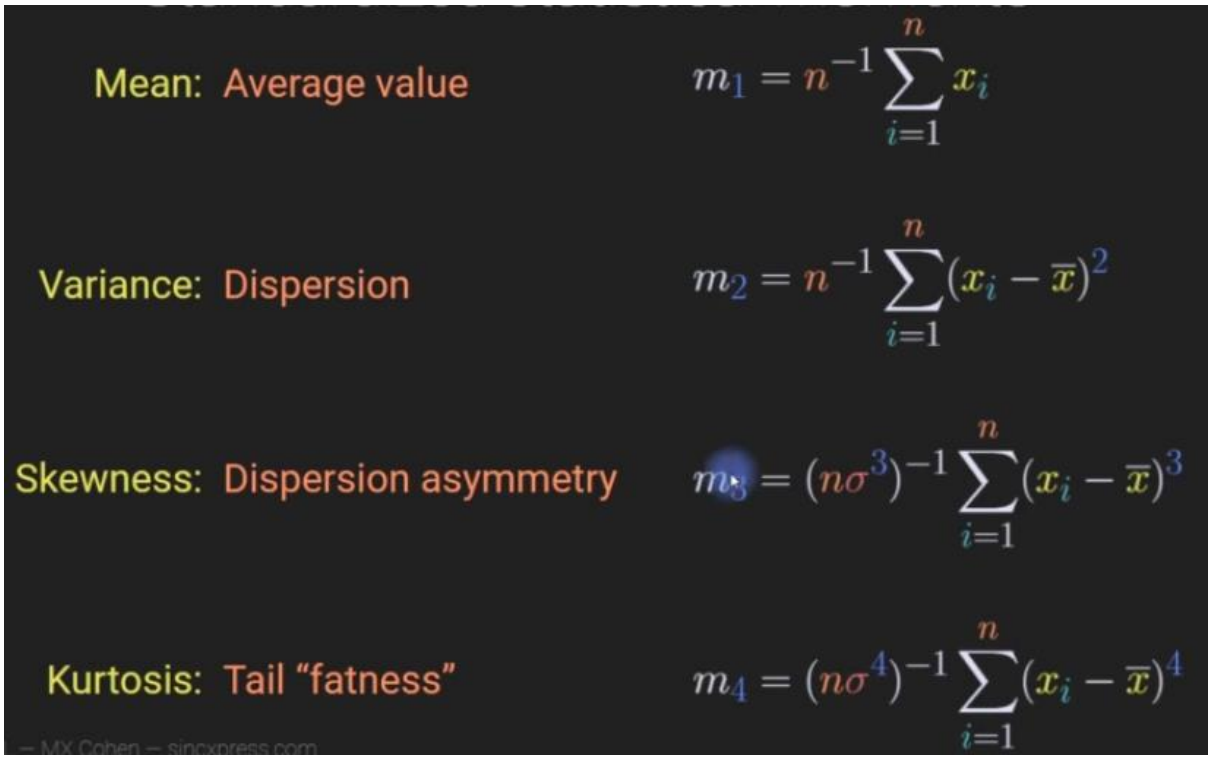
$$(General\ formula)\ m_k = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

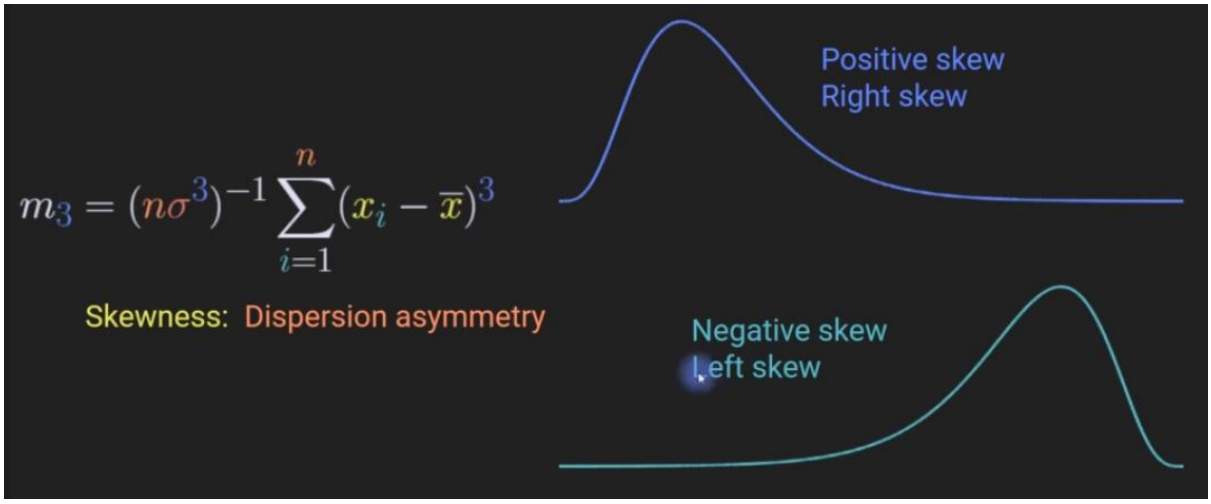
First moment: mean

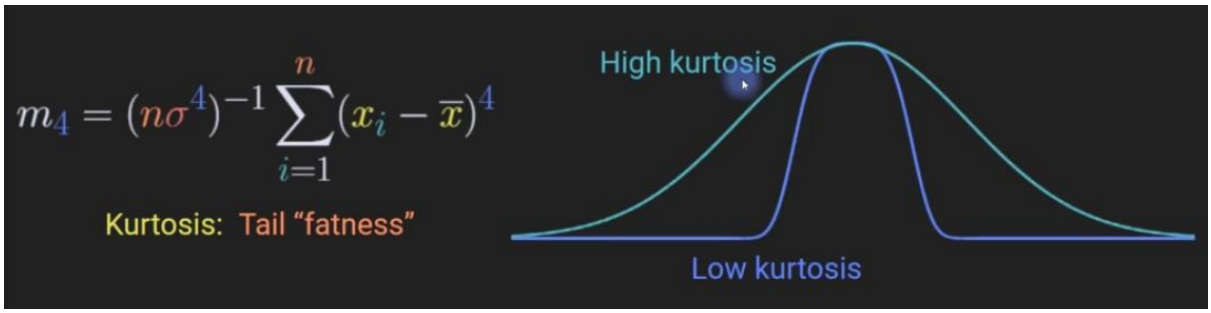
$$m_1 = n^{-1}\sum_{i=1}^{n} x_i$$

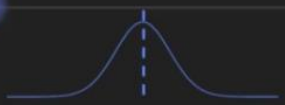
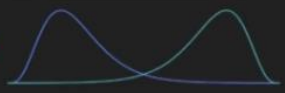
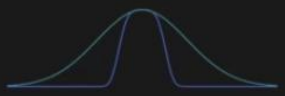
Second moment: variance

$$m_2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$







Moment number	Stats name	Interpretation	Picture
First moment	Mean	Average value	
Second moment	Variance	Dispersion	
Third moment	Skewness	Dispersion asymmetry	
Fourth moment	Kurtosis	Tail "fatness"	

Histograms (Number of bins):

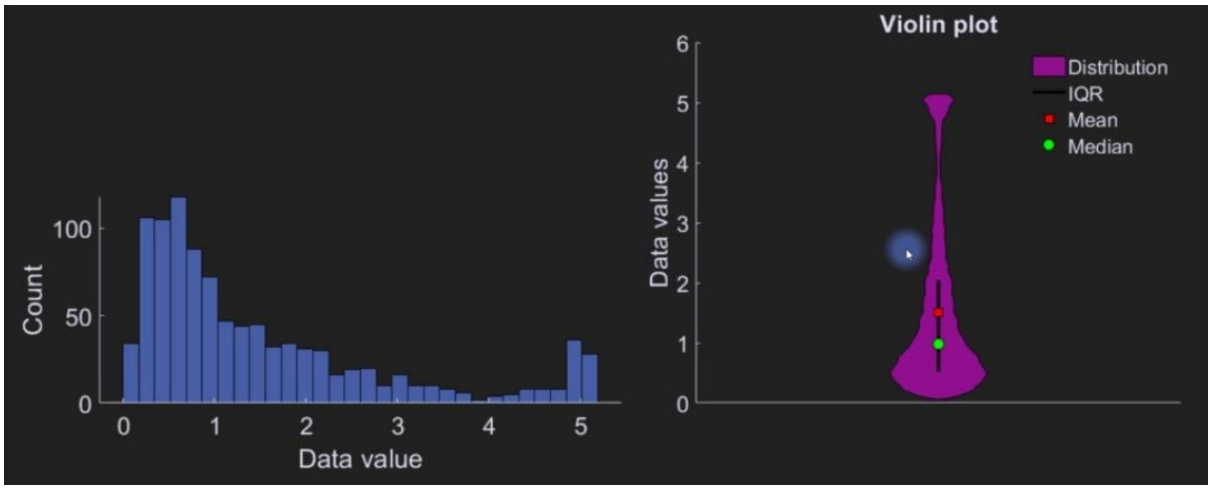
How many bins(k)?

$$k = [\frac{\max(x) - \min(x)}{h\ (width\ of\ bins)}]$$

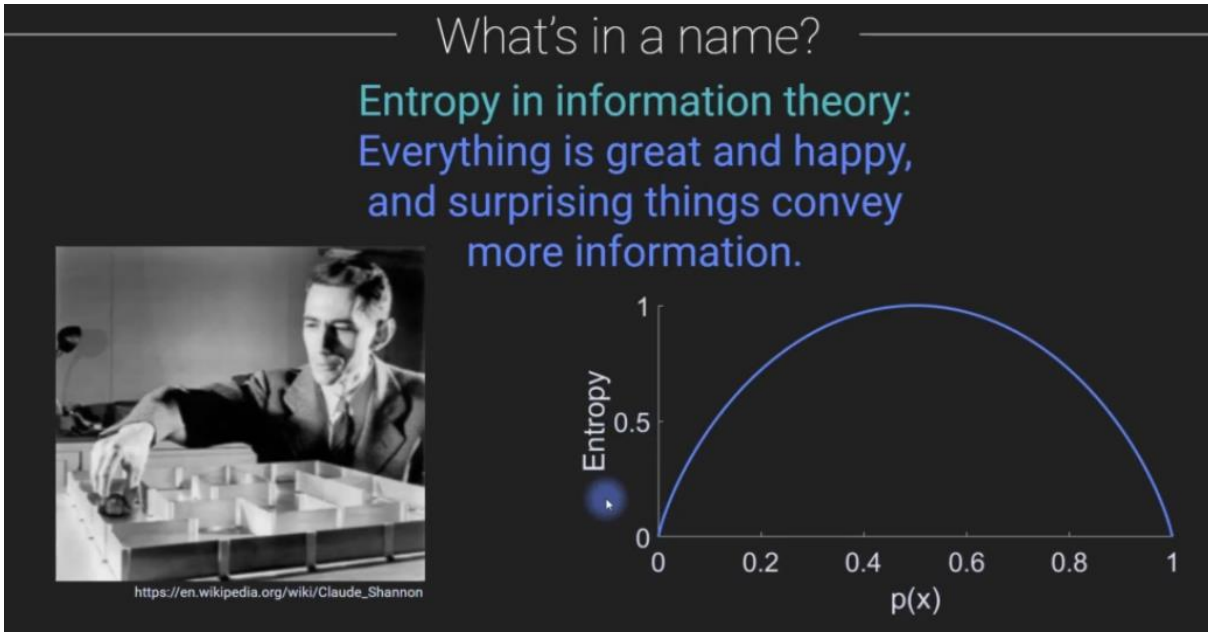
Guideline	Formula	Key advantage
Sturges	$k = \lceil \log 2(n) \rceil + 1$	Depends on data count.
Usually the best one. Probably the most common choice.		
Freedman-Diaconis	$h = 2\frac{\mathrm{IQR}}{\sqrt[3]{n}}$	Depends on data count and on data spread.
Arbitrary	$k = 40$	Easy to use.

Violin Plots:

A violin plot made from histogram.



Shannon Entropy:



Formula for entropy:

$$H = -\sum_{i=1}^{n} p(x_i)\log_2\big(p(x_i)\big)$$

Where,

x = data values

p = probability

What types of data can this formula be applied to?

⇨ Nominal, ordinal, discrete

What if you have interval or ratio data?

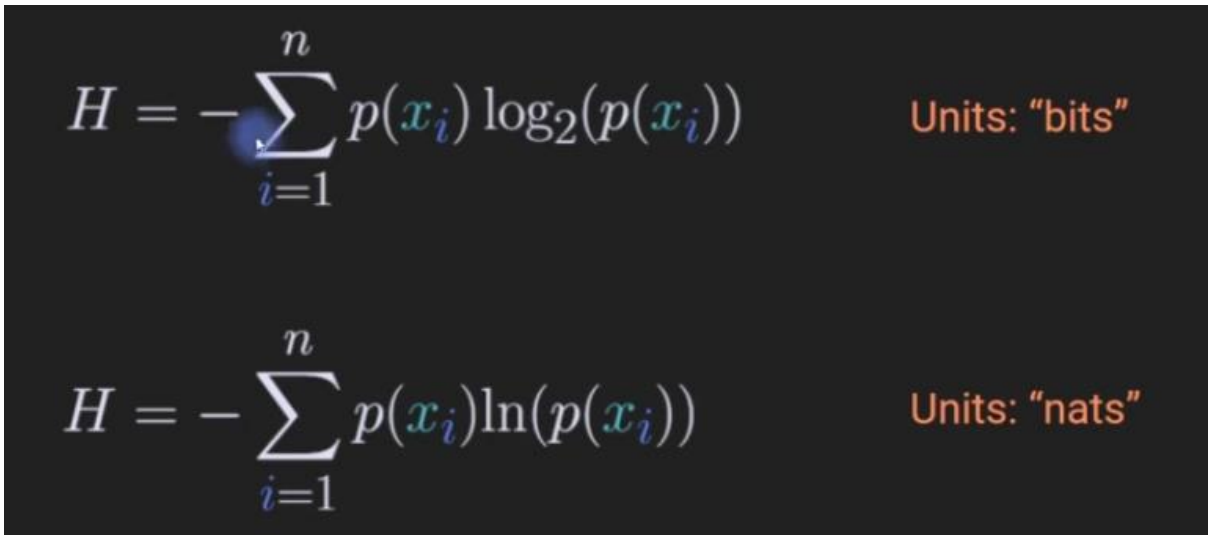
⇨ Convert to discrete by binning (via histogram creation)

- ***Important: Entropy depends on bin size and number!***
- **High entropy means** that the dataset has a lot of variability. Low entropy means that most of the values of the dataset repeat (and therefore are redundant).

How does entropy differ from variance?

Entropy is nonlinear and makes no assumptions about the distribution.

- **Variance depends** on the validity of the mean and therefore is appropriate for roughly normal data.



Normalization & Standardization Data

❖ Z-Score Standardization:

- A value on its own is difficult to interpret; a value relative to its distribution is easy to interpret.
- **Mean-center:** Subtract the average from each individual value.
- **Variance-normalize:** Divide by the standard deviation.
- The units are standard deviations away from the mean of the distributions.
- **Formula:**

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

- Z-transform shifts and stretches, but doesn't change shape.

What is the key assumption that makes the z-transform valid?

⇨ Mean and standard deviation are valid descriptions of the distribution's central tendency and dispersion. That is, the distribution is roughly Gaussian.

❖ Min-Max Scaling:

- Scale to **range of 0 to 1**

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

- Scale to a range ***a*** to ***b***

$$x* = a + \tilde{x}\,(b - a)$$

❖ Outliers:

Where do outliers come from?

- Noisy data
- Noisy or faulty equipment
- Human error (e.g. type)
- Non-cooperative research participant.
- Natural variation

Why are outliers bad?

- Many statistical analyses use squared terms (variance, ANOVA, polynomials, GLM, correlation etc).
- Large outliers become huge when squared.
- Outliers can have more impact with small N.
- Not all outliers are equal. Outliers are worse near the "edges" of the data distribution compared to the "middle".

How to deal with outliers?

- **Strategy 1:** Identify outliers and remove them from the data prior to any analyses.

  Assumption: Outliers are noise or otherwise invalid.

- **Strategy 2:** Leave the outliers in and use robust methods that attenuate the negative impact of the outliers on the results.

  Assumption: Outliers are unusual but valid data.

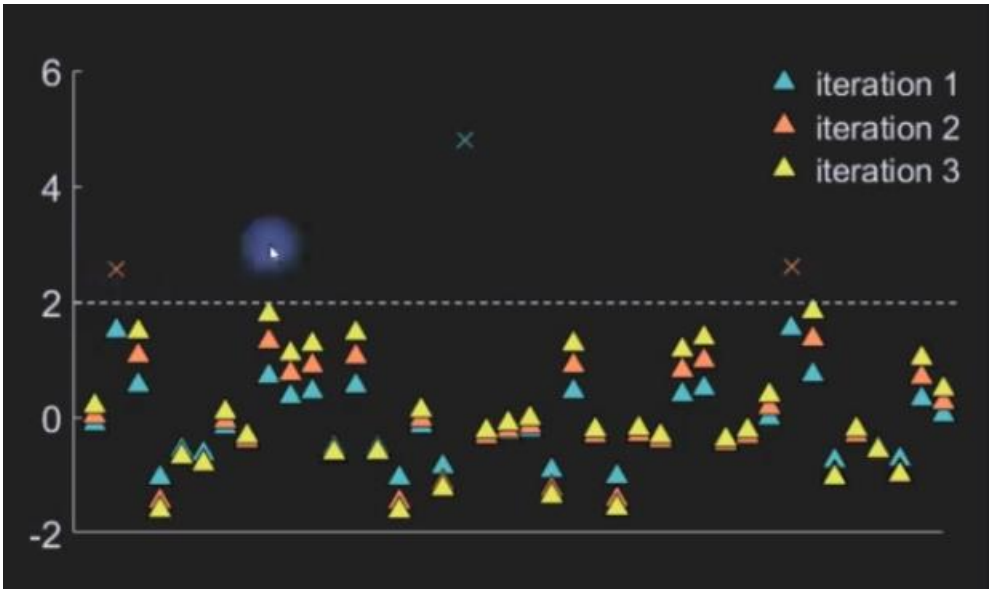
❖ Remove Outliers using Z-Score Method:

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

- Mean-center and standard deviation normalization.
- Z-Score are interpreted as standard deviation units away from the center of the distribution.
- Z-score are valid if the **distribution is roughly Gaussian.**

Iterative Algorithm:

- Convert data to z-score
- A datapoint is an outlier if it exceeds some standard deviation threshold (often 3, but this is arbitrary).
- Remove outliers and repeat until no more outliers.



❖ Remove Outliers using modified Z-Score Method:

For non-normal distributions:

1. Replace "regular" z-score with the modified z-score.
2. Repeat previous methods.
3. Useful for **long-tailed distributions**.

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

$$MAD = median(|x_i - \tilde{x}|)$$

$$\tilde{x} = median(x)$$

MAD = Median Absolute Deviation

❖ Multivariate Outlier Detection:

Euclidean distance:

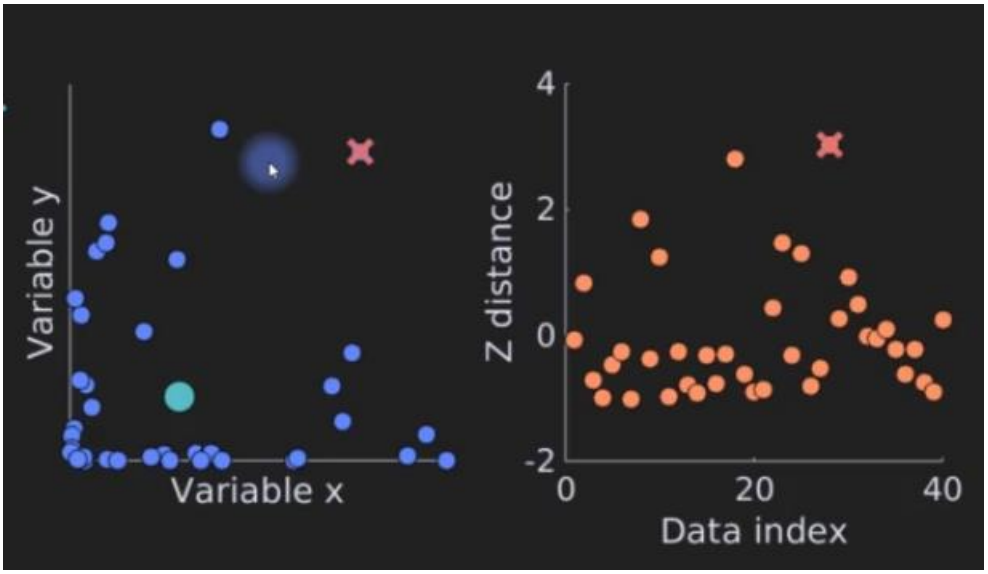
The distance between two points **a** and **b** =>

$$d_{a,b} = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2}$$

Multivariate Algorithm:

- Compute the data mean

- Compute the distance from each point to the mean.
- Convert distances to z-score
- Remove outliers based on threshold, as shown previously.



❖ Removing outliers by data trimming:

Algorithm:

- ⇨ Sort the mean-centered data.
- ⇨ Remove the most extreme k values or the most extreme k%.

Advantages:

- Simple and easy to implement
- Can be effective

Disadvantages:

- Requires subjective threshold
- Can remove non-outliers.

❖ Non-parametric solutions to outliers:

- **Strategy 1:** Identify outliers and remove them from the data prior to any analyses.

Assumption: Outliers are noise or otherwise invalid.

- **Strategy 2:** Leave outliers in and use robust methods that attenuate the negative impact of the outliers on the results.

Assumption: Outliers are unusual but valid data.

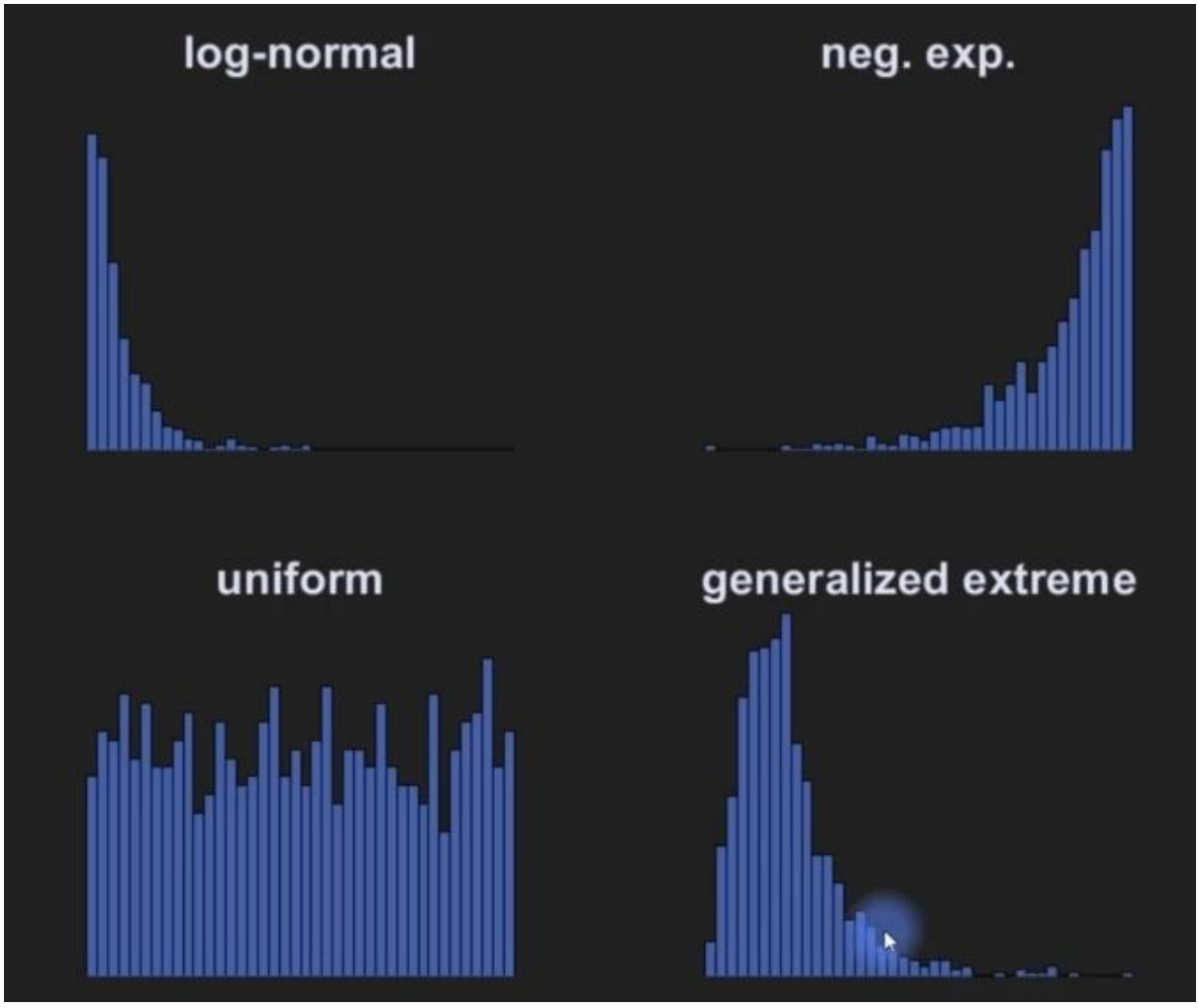
Parametric test	Nonparametric test
• 1-sample t-test • 2-sample t-test • Pearson correlation • ANOVA	• Wilcoxon sign-rank test • Mann-Whitney U test • Spearman correlation • Kruskal-Wallis test • Permutation testing

Why are these nonparametric test methods robust to outliers?

⇨ They are based on medians or ranks, which are insensitive to outliers.

• Nonlinear Data Transformations:

Non-Normal Distributions:



• Many statistical methods are linear or make assumptions about data distributions (e.g., Gaussian)

- Non-extreme data values may be labelled as "outliers" due to nonlinear scaling.
- **Goal:** Transform data to make linear methods valid, or to make data distribution approach Gaussian.

Common nonlinear data transformations:

- Rank-transform
- Logarithm
- Square root
- Fisher-z
- Sign & Co-sign

Mind the interpretation gap:

- Nonlinear methods are not always appropriate (e.g., negative numbers for log or square root).
- Nonlinear transformations alter the spacing between data points as a function of data value.
- Most statistical models are linear, so results must be interpreted in terms of the transformed data, not the original data.

Probability Theory

Probability is a numerical description of how likely an event is to occur or how likely it is that a proposition is true. Probability is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility and 1 indicated certainty. The sum of all probabilities in a set must sum to 1.

Examples of correct & incorrect interpretations:

Statement: There is a 20% probability of rain today.

Interpretation 1: ~~It will rain for 20% of the day -> 0.2*24 = 4.8hours~~ => This is wrong because it is the example of proportion.

Interpretation 2: There is a 1 in 5 chance that it will rain. => If we could repeat the exact conditions of today a very large number of times (approaching infinity), it would rain on 20% of those days.

~~**Interpretation 3:** We can be 20% confident that it will rain today.~~ => Confidence is different from probability: We could be 99% confident of a 20% chance of rain. In this case, 20% is our parameter estimate, and the confidence interval might be [19% 21%].

When do we need probability?

- ⇨ We need probability when there is uncertainty about the outcome of an event.

Probability vs Proportion:

Probability: the likelihood of an event occurring or that a statement is true.

Proportion: a fraction of a whole

Example:

- I spend a total of 5.1 minutes each day brushing my teeth, out of 17 waking hours (1020 minutes).
  - The proportion of my waking day spent brushing my teeth is 5.1/1020 = 0.005(0.5%) → ✓

    - The probability that a randomly selected minute of my day involves teeth-brushing is 0.005. → ✓
    - The probability that I will brush my teeth during the day is 1. → ✓
- I eat 100g of chocolate every Friday.
    - The proportion of my week that contains a chocolate-eating day is 1/7 = 0.143 (14.3%). → ✓
    - The probability that a randomly selected day involves chocolate-eating is 0.143. → ✓
    - The probability that I will eat chocolate during the week is 1. → ✓
- Covid-19 has 3.4% mortality rate (WHO, 3March 2020, worldometers.info)
    - If I get infected, I have a 3.4% change of dying from Covid-19 complications. → ✗ (Because for **conditional probability**. Fatality depends on age. Fatality is conditional on age.)
    - The probability that a randomly selected infected patient will die from Covid-19 complications is 0.034. → ✓
    - The proportion of infected people who die from complications is 0.034. → ✓
- The dataset of flipping the coin, [h t t t h t h h h h]
    - Proportion of heads = 60%
    - Probability of heads = 50%

Computing Probability:

- **Data types for probability:**

Interval	These are not. Convert to below 3
Ratio	
Discrete	Valid for probability
Ordinal	
Nominal	

- Data must have mutually exclusive labels or bins. Because total probabilities must sum to 1 (100%).

$$p_i = \frac{100c_i}{\sum c}$$

The probability of event i is the count of events i divided by the total number of all events.

Example:

A jar containing 40 blue marbles, 30 yellow marbles and 20 orange marbles. What is probability of picking each color at random? (First, determine data validity (data type and exclusivity) and second, compute probabilities)

⇨ Data type: nominal

exclusivity: yes

Pick blue: 40*100/90 = 44.4%.
Pick yellow: 30*100/90 = 33.3%
pick orange: 20*100?90 = 22.2%

Sanity check: 44.4 + 33.3 + 22.2 = 99.9%

Code: Part 1 - Compute Probability

Odds:

- Often discussed in the context of betting or disease. For example, "The odds are **6:1** against."

$$r = n : m = \frac{n}{m}$$

- "Odds" is a ratio of the probability of an event not occurring to the probability of it occurring.

$$r = \frac{1-p}{p}$$

$$p = \frac{1}{1+r} = \frac{1}{1+\frac{n}{m}}$$

Example: What are the odds of drawing a king in a deck of cards?

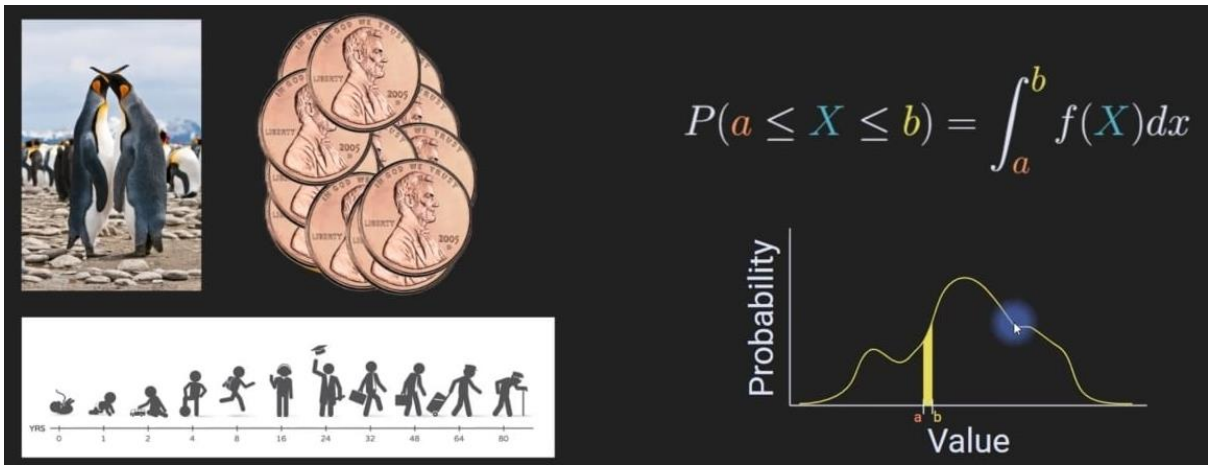
⇨ 12:1

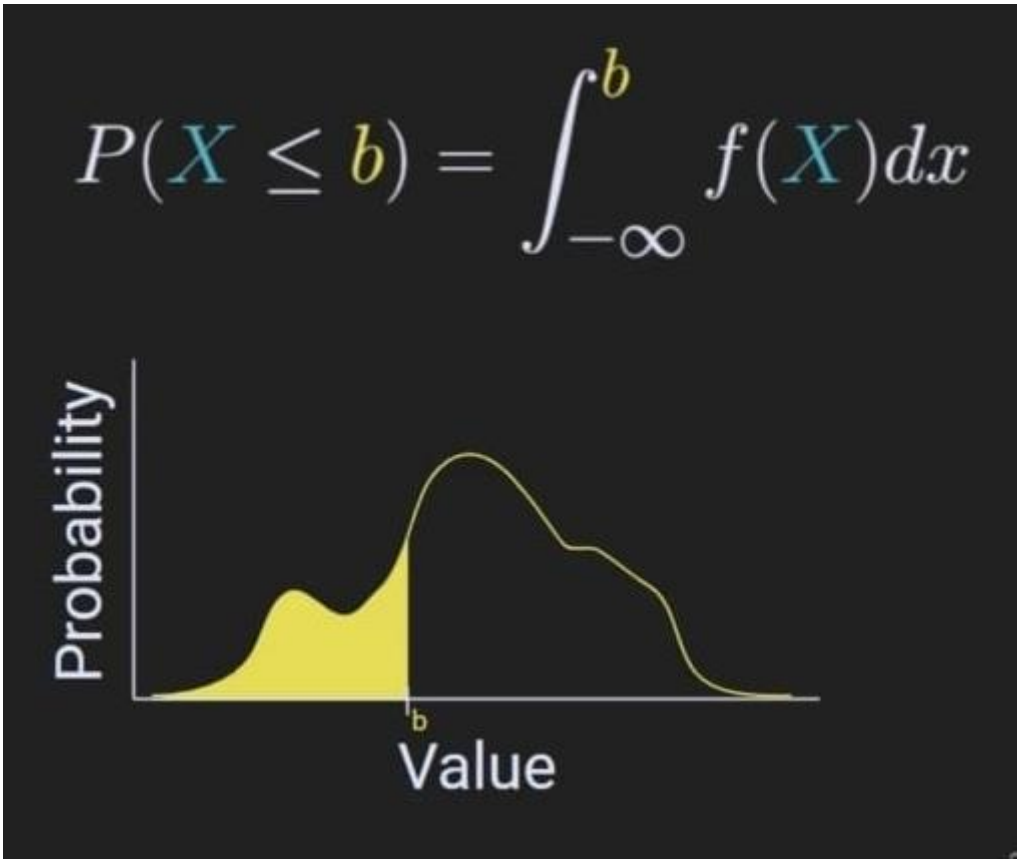
Probability Mass vs Density:

Probability: The chance that an event will occur.

Probability mass: A function that describes probabilities for a set of *exclusive discrete* events. E.g., probability of picking a specific card out of a deck of cards, flipping a coin, picking a marble from a jar etc. This can be visualized by *bar plot* or *histogram*.

Probability density: A function that describes probabilities for a set of *exclusive continuous* events. E.g., height of penguin, age, money (like salary), distance between two places





- A function of variable x defined by the probability of different values of x occurring.

$$f(x) = p(X = x)$$

- The value of the function for each unique element of x is equal to its probability.

$$f(x_i) = p(X = x_i)$$

- Probabilities must be non-negative.

$$p(X = x_i) \geq 0$$

- Events must be exclusive.

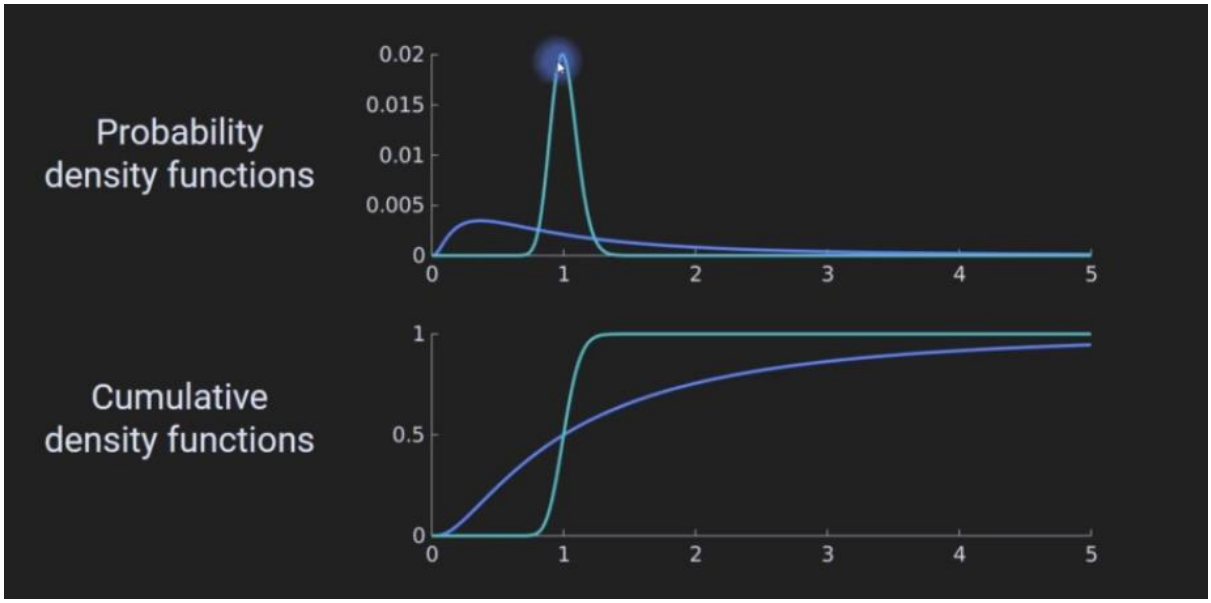
$$p(X \neq x_i) = 0$$

- All probabilities must sum to 1

$$\sum p(X = x_i) = 1$$

Code: Part 2 - Probability Mass Function

Cumulative Probability Distribution(cdf):



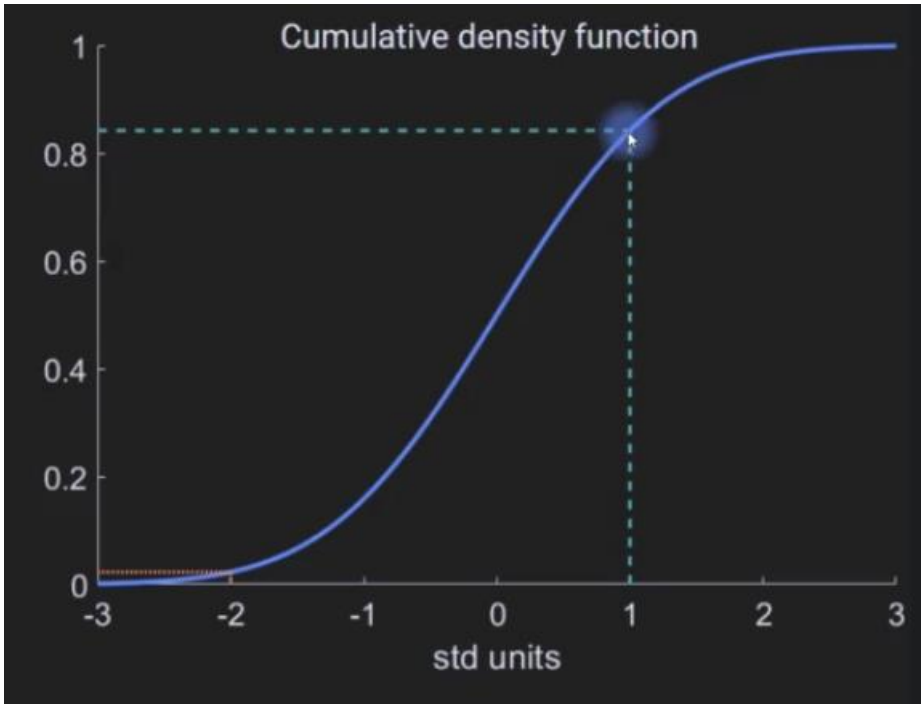
- **A cdf is the cumulative sum (or integral) of the probability distribution (or density).**
- The y-axis value at each x-axis is the sum of all probabilities to the left of that x-value.
- A cdf starts at 0 and increases monotonically to 1. The sum of the cdf is more than 1.

$$C(x_a) = p(X \leq x_a)$$

$$C(x_a) = \int_{-\infty}^{a} p(x_t)dt$$

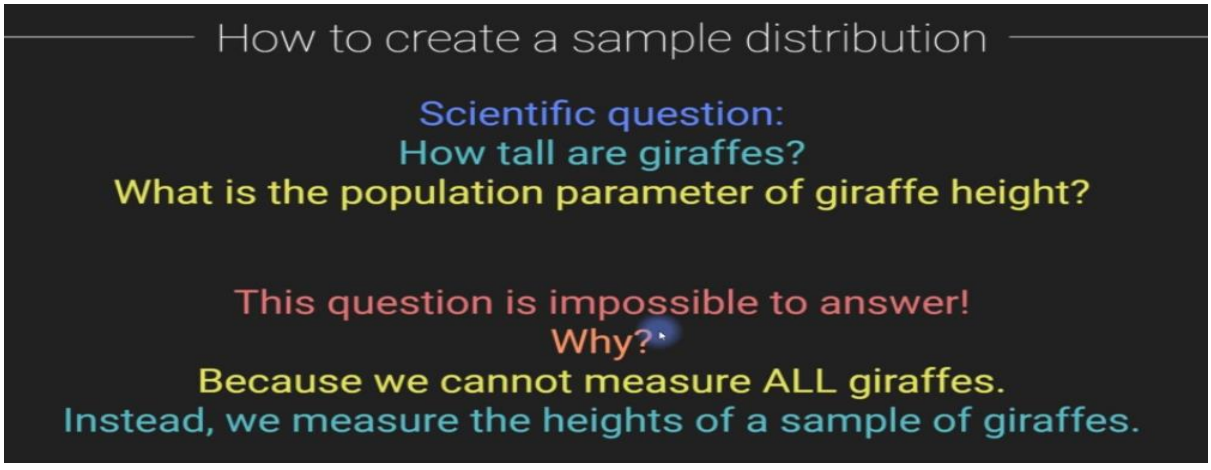
$$C(x_a) = \sum_{i=1}^{a} p(x_i)$$

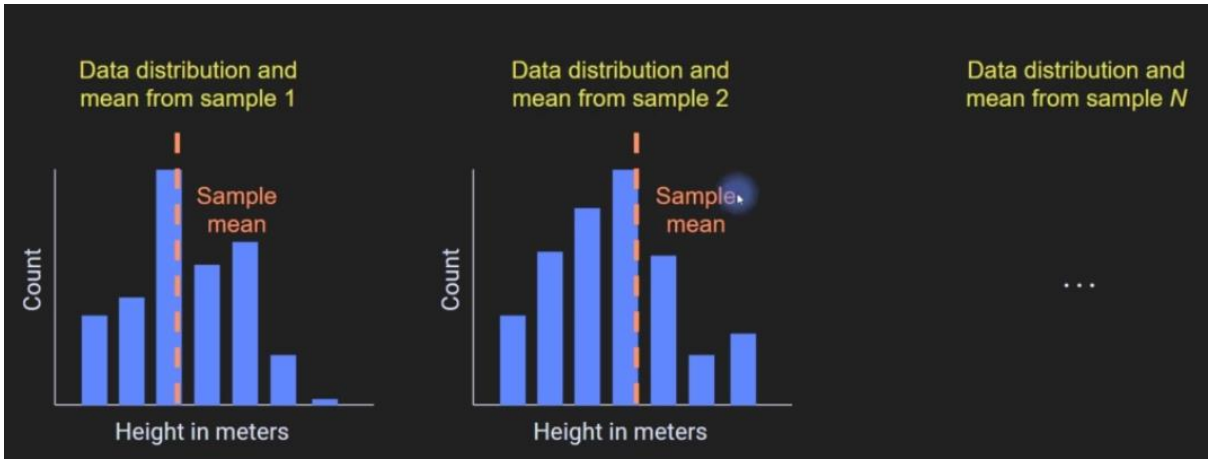
- CDF's are used to evaluate the probability of obtaining a value up to X or at least X.
- **Example:** What is the probability of getting at least 1 std higher on the SATs than average?
- **Example:** What is the probability of an elephant weighing less than 2 std below the average?

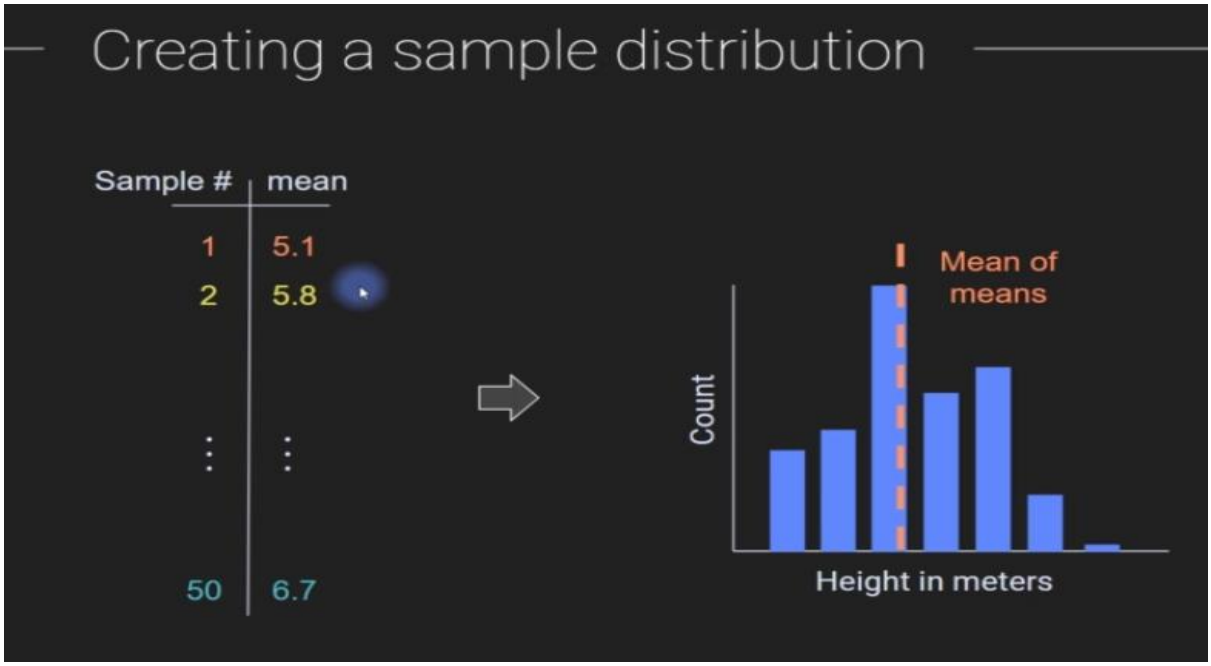


Code: Part 3 - CDF & PDF

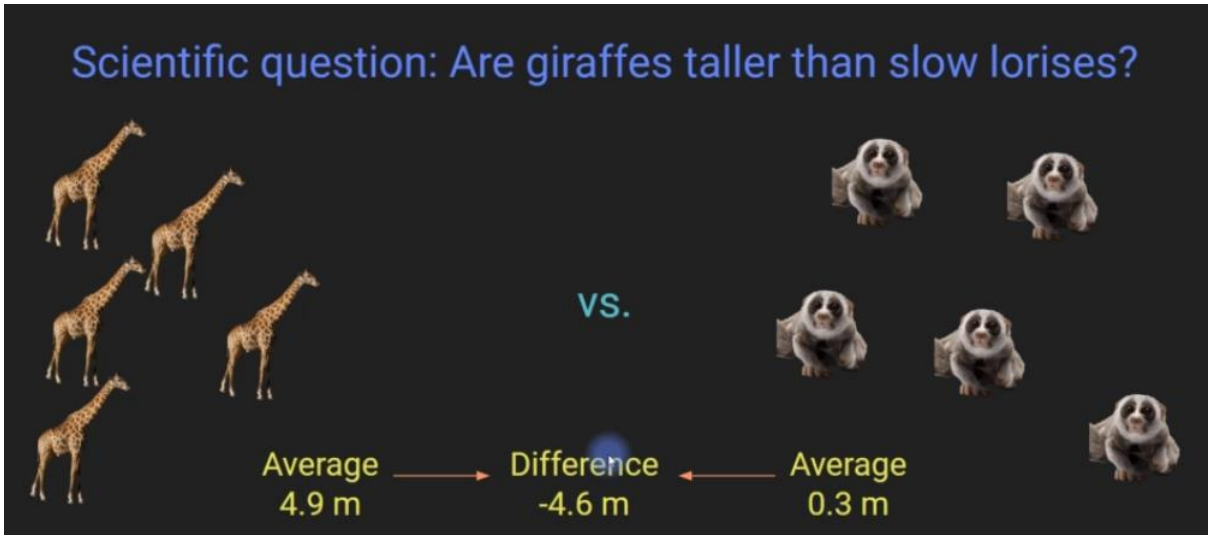
Creating Sample Estimate Distributions:



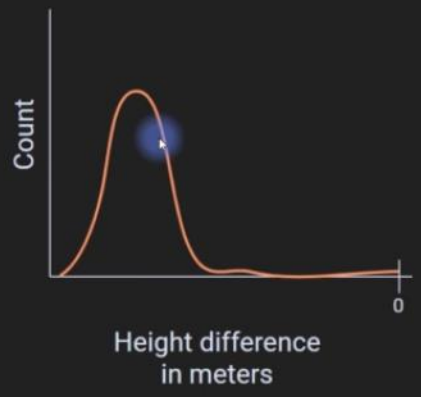




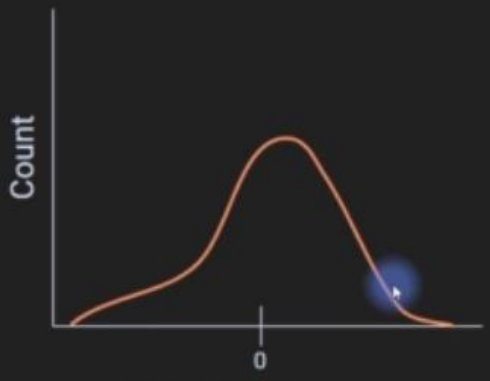
- A distribution of data from one sample has one mean. Many means from many samples gives a sample distribution. (Note: "mean" is just an example of a parameter estimate).

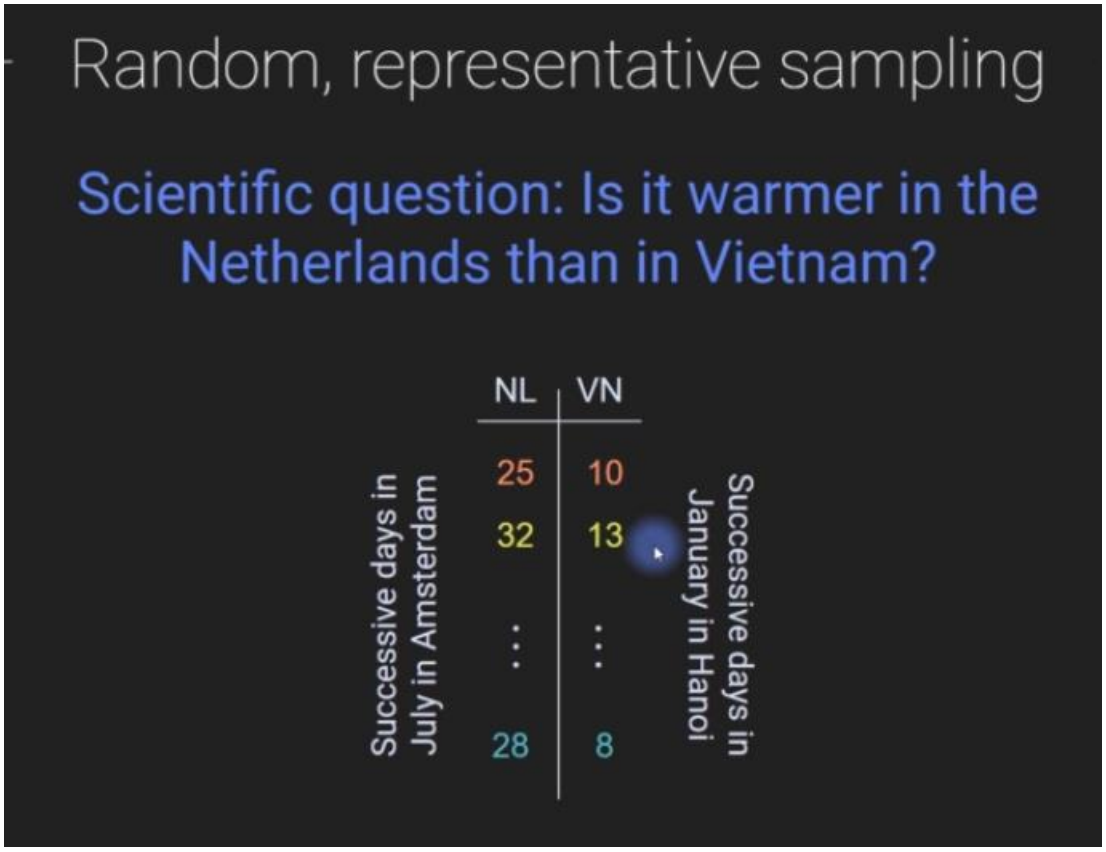


Scientific question: Are giraffes taller than slow lorises?



Scientific question: Are north African giraffes taller than south African giraffes?





Appropriate answer based on sampling of above image: July in Amsterdam is warmer than January in Hanoi.

Larger point: Sample estimates can be generalized only to the population that the sample represents.

Monte Carlo Sampling:

- **Monte Carlo methods:** solve really hard problems by randomly sampling the solution space instead of doing the real work.
- Often used in physics, statistics, deep learning, even in pure math.
- "Monte Carlo sampling" is the same thing as randomly sampling from a population to estimate an unknown population parameter.

Sampling Variability:

- Scientific question: How tall is the average Montenegrin?
- Don't worry about the actual answer. The question is: How do we know the answer?

- **Sampling variability:** Different samples from the same population can have different values of the same measurement.
- **Implication of sampling variability:** A single measurement may be an unreliable estimate of a population parameter.
- **Natural variation:** Often seen in biology (e.g., height, weight) and physics (e.g., earthquake magnitude, number of stars per galaxy)
- **Measurement noise:** The sensors are imperfect (e.g., electrical line noise, measuring µg with a gram-precision scale).
- **Complex systems:** Measuring some factors while ignoring others (e.g., measuring height while ignoring age).
- **Stochasticity (randomness):** The universe is a wild and unpredictable place (e.g., photons hitting a camera lens).

What to do about sampling variability?

⇨ *Take many samples!* Averaging together many samples will approximate the true population mean.
⇨ *Run statistical analysis!* Compute measures of confidence of sample-based parameter estimates.

Code: Part 4 – Sampling Variability

Expected Value:

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

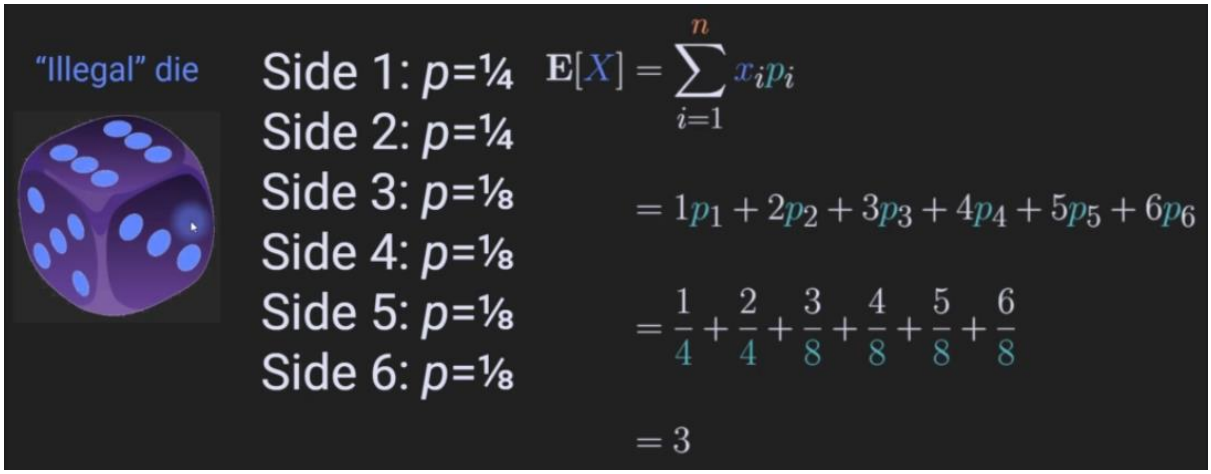
$$E[X] = \sum_{i=1}^{n} x_i p_i$$

When are the two formulas equal?

⇨ When all $p_i = 1/n$, i.e., when all data values are equally likely to occur.

- Average is an empirical sample estimate based on finite data.

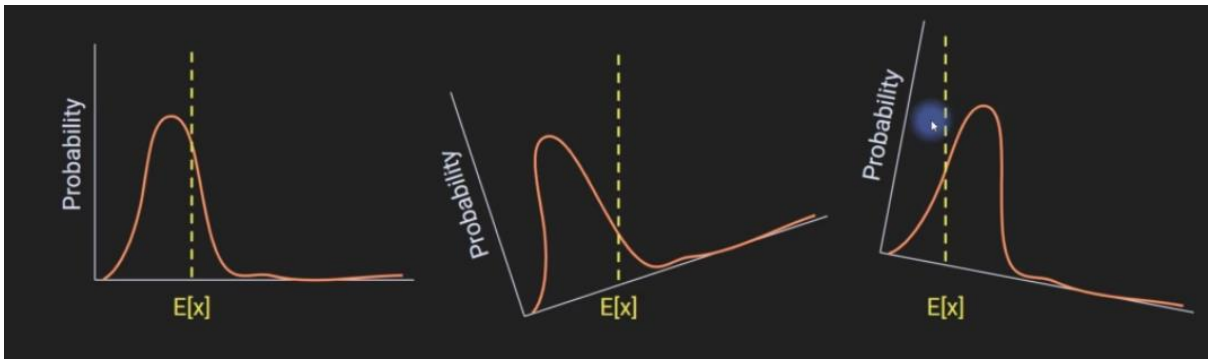
- Expectation value is the expected average in the population, or from a very very large number of samples (approaching infinity).



When do you think the expected value and average are "the same"?

When drawing large and representative random samples from a population.

Expected value is the "balance point" of a probability distribution.



Conditional Probability:

What is the probability that you will pass your statistics course, given that you've been studying for weeks?

What is the probability that the patient has diabetes, given that her parents both have diabetes?

The probability of event A changes based on what you know about event B.

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

$$P(A \cap B) = P(A)P(B)$$

A and B are independent when knowing P(A) provides no information about P(B).

A supermarket hires you to help them sell more toilet paper, by advising them on how to maximize toilet paper marketing. You look through their sales data (hopefully anonymized…) and see that out of 100 random receipts, 42 people bought toilet paper, 60 people bought canned soup, 55 bought candy bars and 24 people bought dried fruit. Which product should t.p. be marketed with? Further inspection reveals that 11 people bought toilet paper and cannot soup, 32 people bought toilet paper and candy bars, and 21 people bought toilet paper and dried fruit.

- ⇨ A = toilet paper, B_1 = canned soup, B_2 = candy bars, B_3 = dried fruit.

$$P(A \cap B_1) = \frac{11}{100} = .11$$

$$P(A|B_1) = \frac{.11}{.6} = .18$$

$$P(A \cap B_2) = \frac{32}{100} = .32$$

$$P(A|B_2) = \frac{.32}{.55} = .58$$

$$P(A \cap B_3) = \frac{21}{100} = .21$$

$$P(A|B_3) = \frac{.21}{.24} = .88$$

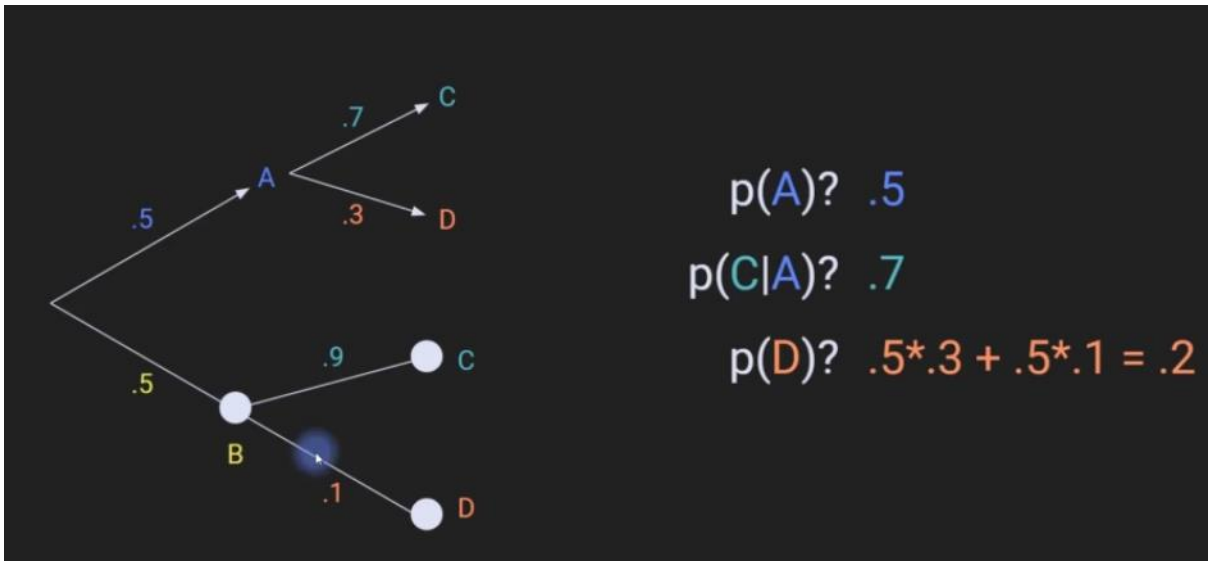
If A and B never co-occur, what is P(A|B)?

If B never occurs, what is P(A|B)?

If A and B are independent, what is P(A|B)?

Code: Part 5 - Conditional Probability

Tree Diagrams for Conditional Probability:



Law of Large Numbers (LLN):

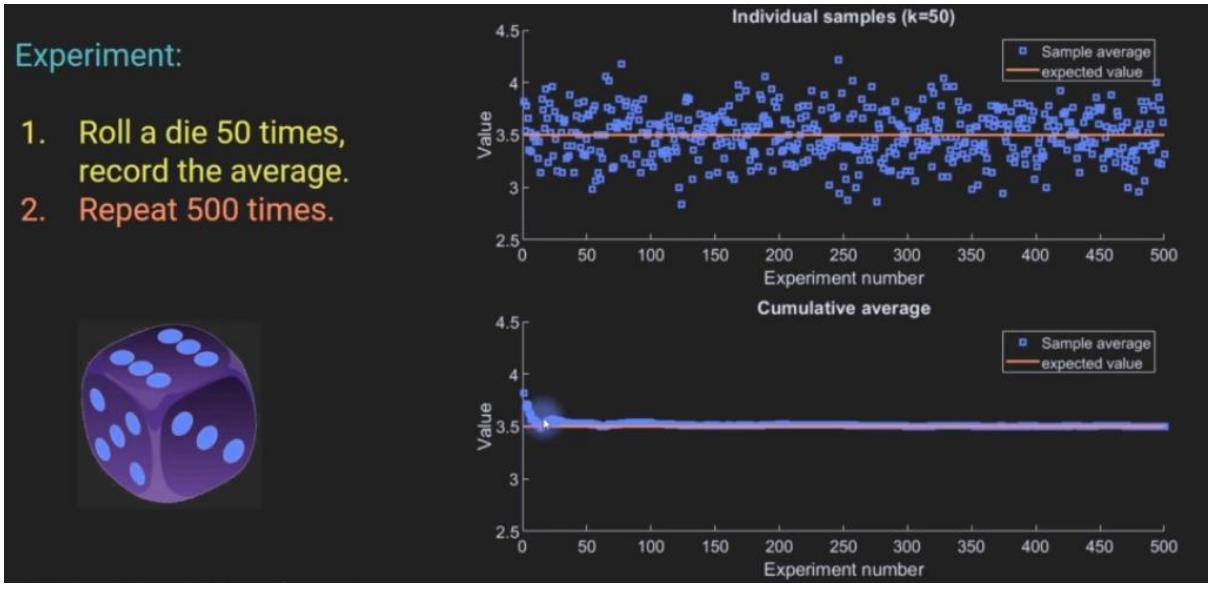
As the number of experiment repetitions increases, the average of the sample means better approximates the population mean.

$$\lim_{n\to\infty} P(|\overline{x}_n - \mu| > \epsilon) = 0$$

- Any one sample (or any one experiment) is sensitive to sampling variability, noise and other sources of non-systematic variation.
- This means that one sample or one experiment is unlikely to provide a good estimated of the true population mean.
- But sampling many times can provide an accurate measure of the true population mean!

How large is Large? What does N need to be to get a reasonable estimate?

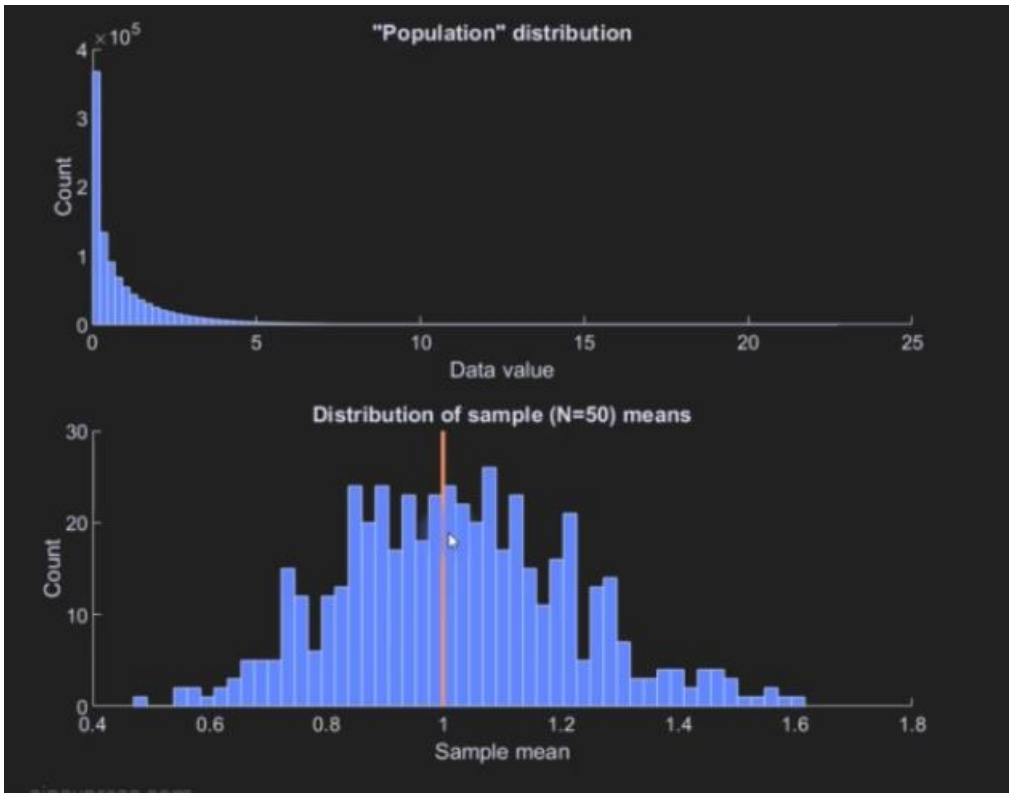
⇨ Unfortunately, this is nearly impossible to say. It depends on many factors, including effect size, noise, system complexity, measurement techniques etc.



Code: Part 6 - Law of Large Numbers

Central Limit Theorem:

- **Part 1:** The distribution of sample means approaches a Gaussian distribution, regardless of the shape of the population distribution.



- **Part 2:** Random samples from independent variables will tend towards a normal (Gaussian) distribution, even if the variables are non-normally distributed. *Note, this depends on sample count, data scale or normalization, and other factors.*

Why the CTL is important?

- ⇨ Many statistical analyses rely on the assumption of normality (roughly Gaussian distribution).
- ⇨ Gaussian distributions are easy to parameterize (mean, variance, skew, kurtosis) and compute confidence intervals around.
- ⇨ Independent components analysis (ICA) is based on the assumption that signals are non-gaussian distributed while random mixture of signals are Gaussian.

Code: Part 7 - Central Limit Theorem

Hypothesis Testing

IVs, DVs and Other Stats Lingo:

- **DV:** Dependent variable (a.k.a. outcome variable). The variable you are trying to explain
- **IV:** Independent variables (a.k.a. explanatory variables). The variables that you hope will explain the DV.
- **Examples:**
  - Effects of soil moisture on plant growth.
  - Effects of time spent on Facebook on irritability.
  - Relation between money spent on cloths and bar hopping.
- **Usually, the IV is on the x-axis and the DV is on the y-axis.**

Model: A model is an equation that explains some features in a dataset.

Residuals: Data (real world) – Model = Residual (error). *Important point: Residuals should be small, but models should be simple! Residuals should be small, but models should be relevant! Model should not be over fitted as well as under fitted also!*

All hypotheses are model comparisons!

A model should be as simple as it can be, and as complicated as it must be.

Hypotheses:

Hypothesis: A falsifiable claim that requires verification, typically from experimental or observational data, and that allows us for predictions about future observations.

Why are hypotheses important?

- ⇨ Hypotheses improve experiment design, critical thinking, and data analysis.
- ⇨ Hypotheses transform loose ideas into concrete and specific claims.
- ⇨ Hypotheses are used to develop new and more accurate theories and to dissolve bad theories.
- ⇨ Most progress in science, engineering and medicine is the result of hypothesis-testing.

A strong hypothesis is:

- ⇨ Clear
- ⇨ Specific
- ⇨ Falsifiable
- ⇨ Based on prior data or theory
- ⇨ Leads to a statistical test
- ⇨ A statement, not a question
- ⇨ A prediction about the direction of an effect
- ⇨ Relevant for unobserved data or phenomena
- ⇨ Relevant for understanding nature

Examples of hypotheses:

- ⇨ **Example 1:**
  - o **Not an hypothesis:** Medical research is important for curing diseases.
  - o **Weak but still an hypothesis:** The medication has an effect.
  - o **Strong:** The medication reduces symptom X in a dose-response fashion.
- ⇨ **Example 2:**
  - o **Not an hypothesis:** Will students pass this course?
  - o **Weak but still an hypothesis:** Studying improves grades.
  - o **Strong:** A combination of self-study and group-study will improve final exam grades by at least 10%.

Null Hypothesis (H_0): The "null hypothesis" is the hypothesis that nothing interesting is happening in the data. In the research, you specify the "**alternative hypothesis (H_a)**" (should be called the "**effect hypothesis**"). In statistical analysis, you test the null hypothesis. E.g.

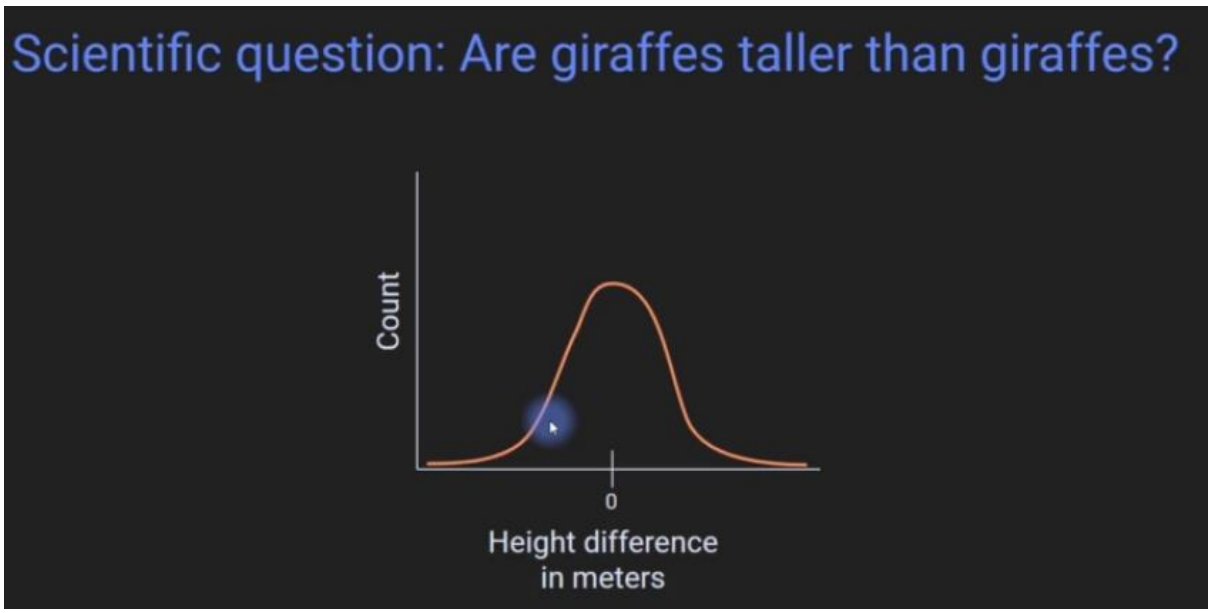
- ⇨ H_a: People will buy more widgets after seeing advertisement X compared to advertisement Y.

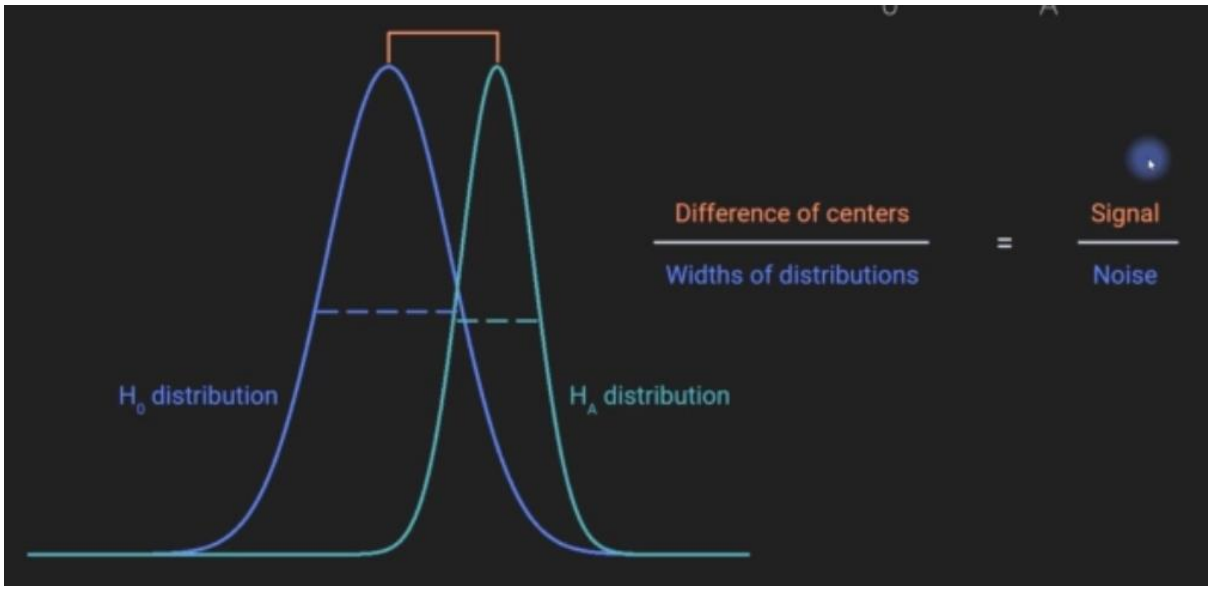
  H_0: The advertisement type has no effect on widget purchases.

Proving hypotheses?

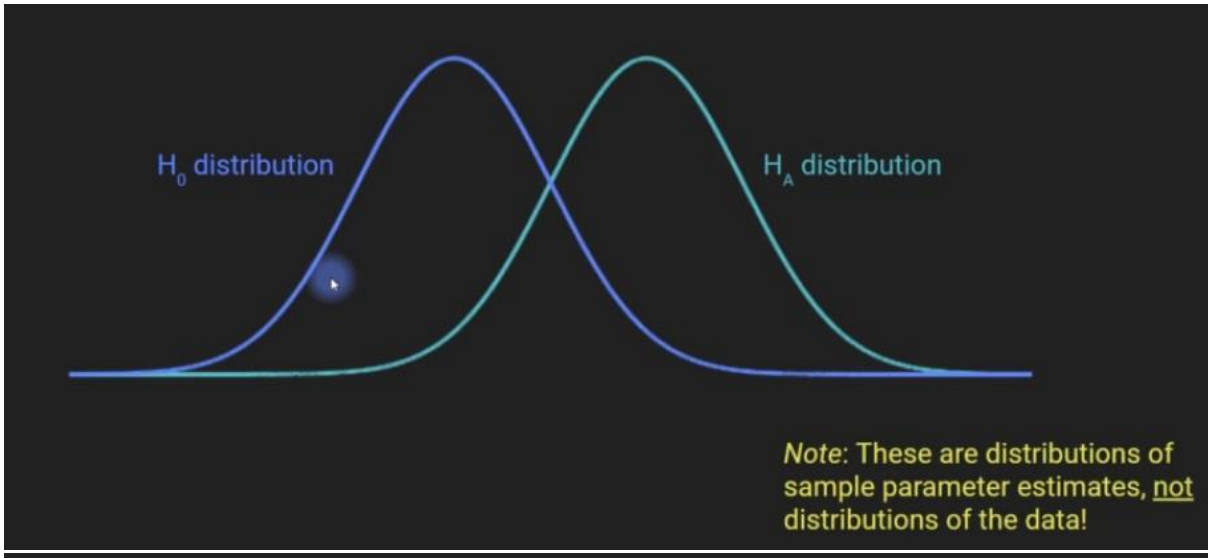
- ⇨ It is not possible to prove a hypothesis.
- ⇨ Hypotheses can be rejected or fail to be rejected (interpreted as being supported.)

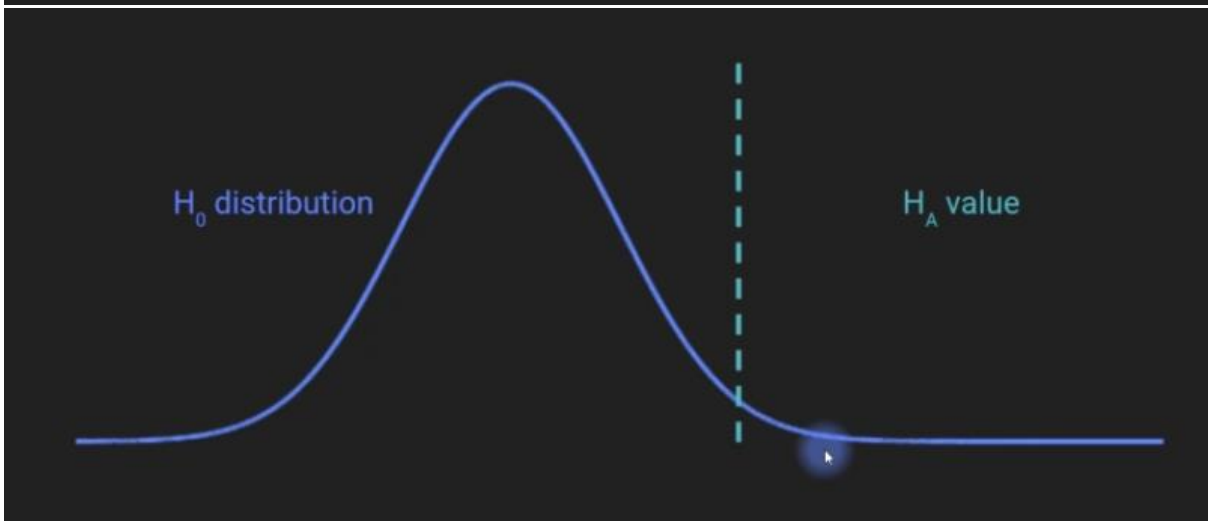
Distributions under the Null and Alternative Hypotheses:

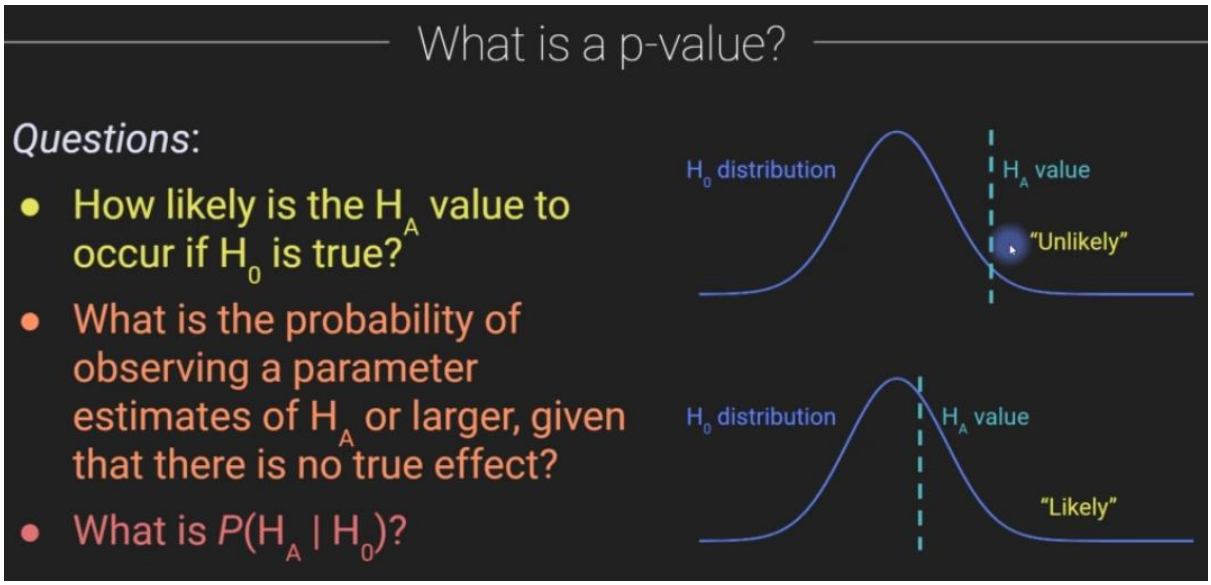


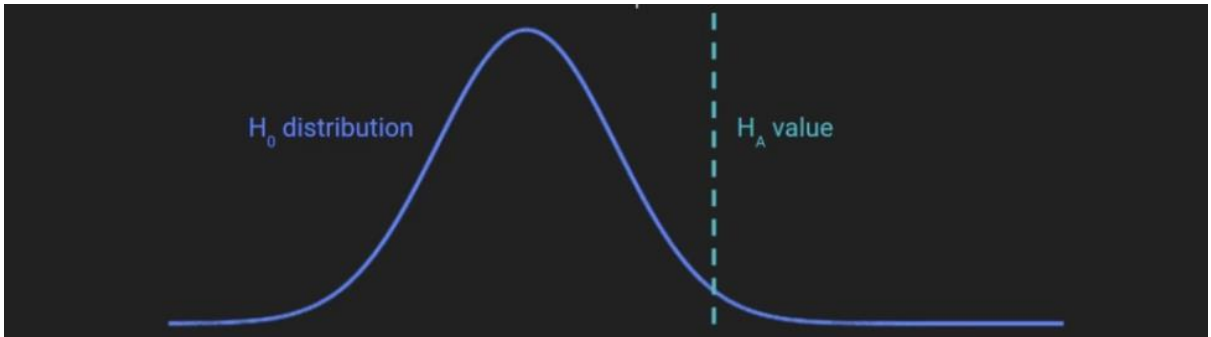


P-Values:


H_0 distribution
H_A distribution
Note: These are distributions of sample parameter estimates, not distributions of the data!

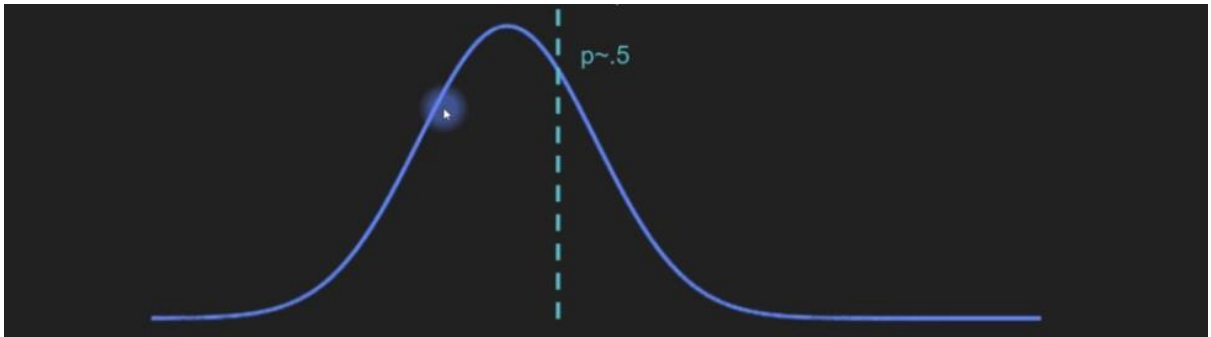

H_0 distribution
H_A value


What is a p-value?
Questions:
- How likely is the H_A value to occur if H_0 is true?
- What is the probability of observing a parameter estimates of H_A or larger, given that there is no true effect?
- What is $P(H_A \mid H_0)$?
H_0 distribution
H_A value
"Unlikely"
H_0 distribution
H_A value
"Likely"



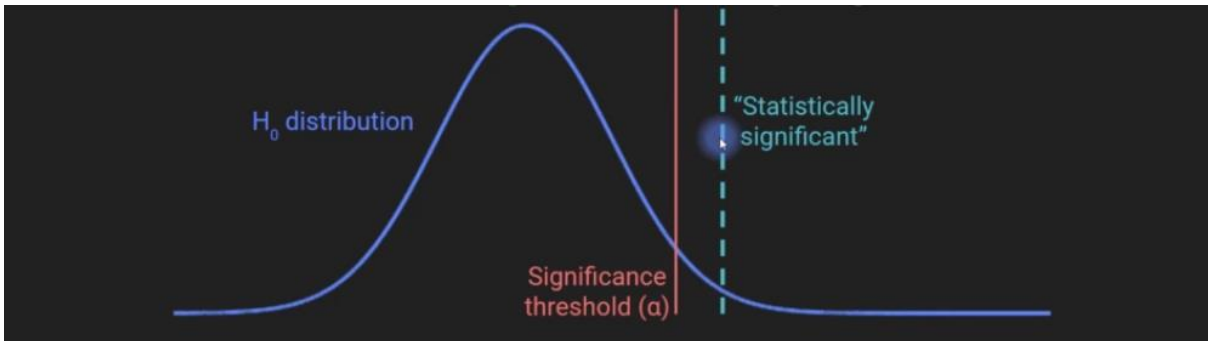
Important concept:

We cannot prove that H_A is true. We can only compute the probability that the test statistic associated with H_A could be observed given that there is no true effect.



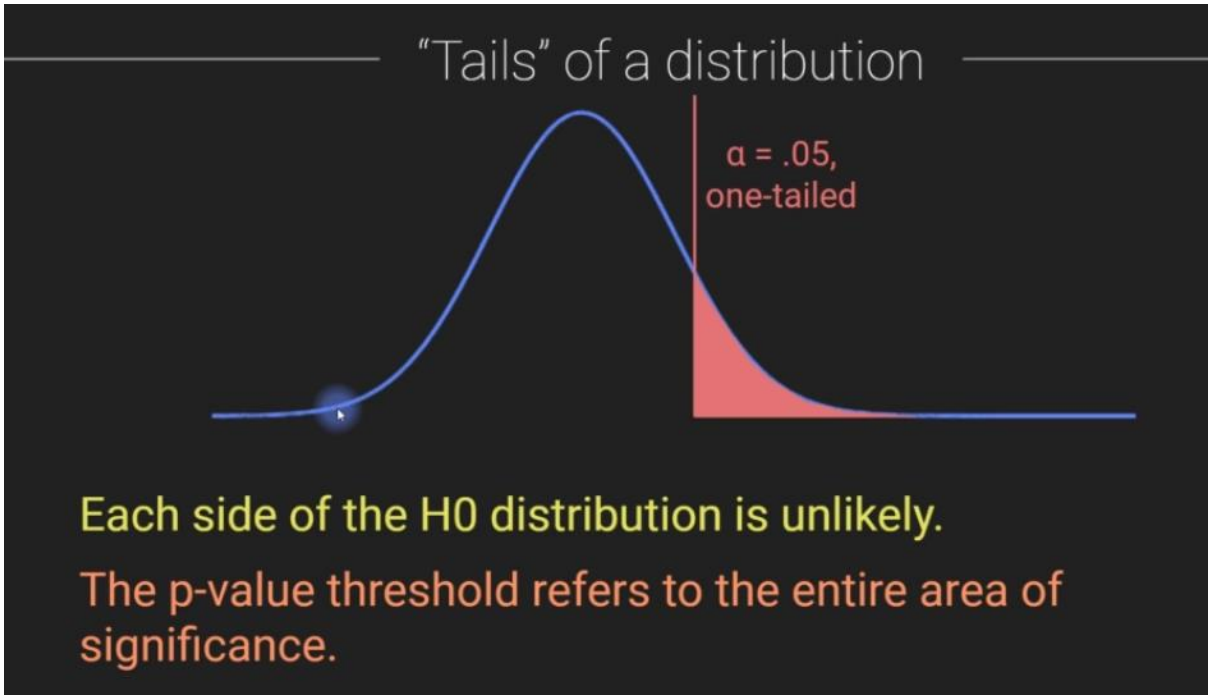
P-values are probabilities. They range from 0 to 1.

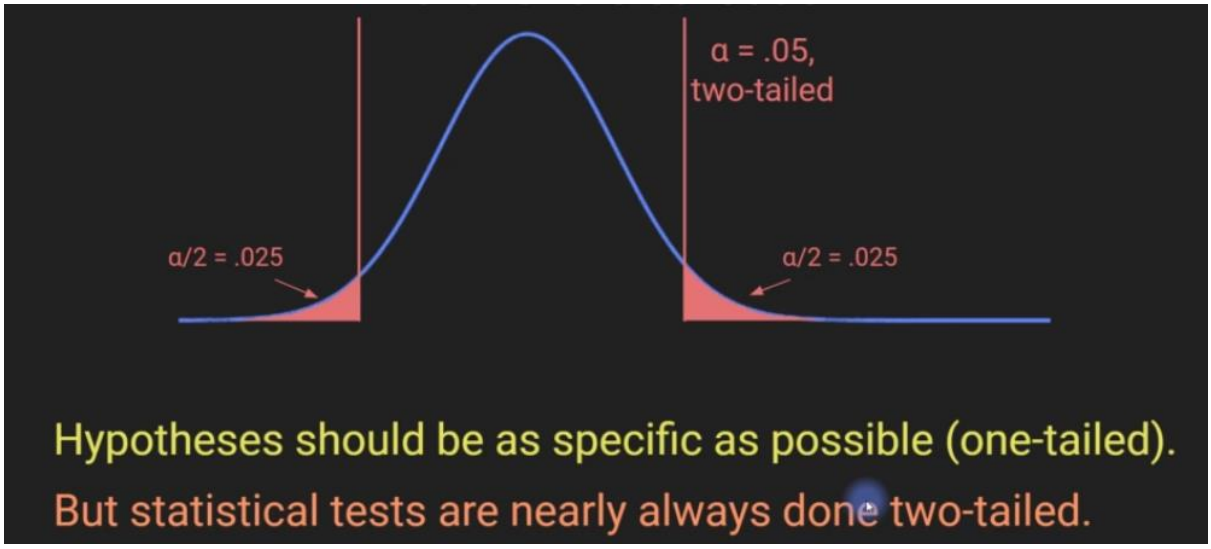
Values closer to zero indicate low probability of $H_A \mid H_0$, and values closer to one indicate high probability of $H_A \mid H_0$.



A finding is called "statistically significant" if the test statistic is greater than a threshold. That is, if $p(H_A) < p(\alpha)$.

Threshold is arbitrary; common values are $p<.05$ or $p<.01$.





Common misinterpretations of p-values:

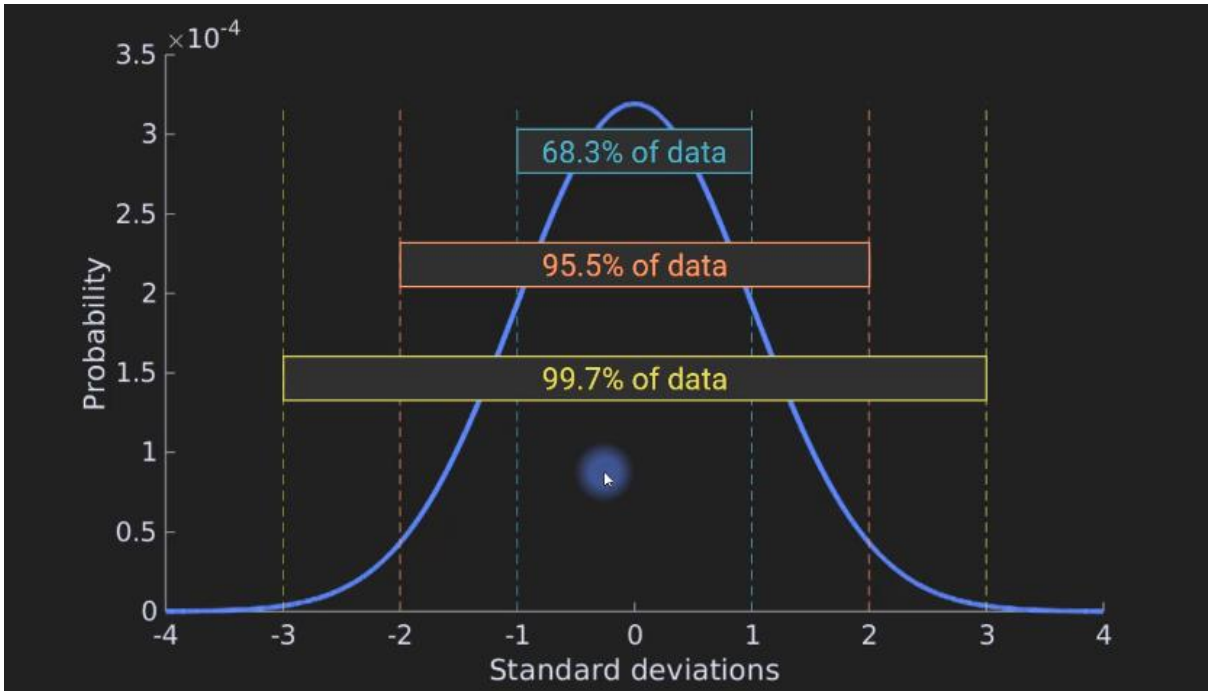
- **Incorrect:** My p-value is 0.02, so the effect is present for 2% of the population.
  Correct: My p-value is 0.02, so there is a 2% chance that there is no effect, and my large sample statistic was due to sampling variability, noise, small sample size or systematic bias.
- **Incorrect:** My p-value is 0.02, so there is a 98% chance that my sample statistic equals to the population parameter.
  Correct: My p-value is 0.02, so there is a 2% chance that there is no effect, and my large sample statistic was due to sampling variability, noise, small sample size or systematic bias.

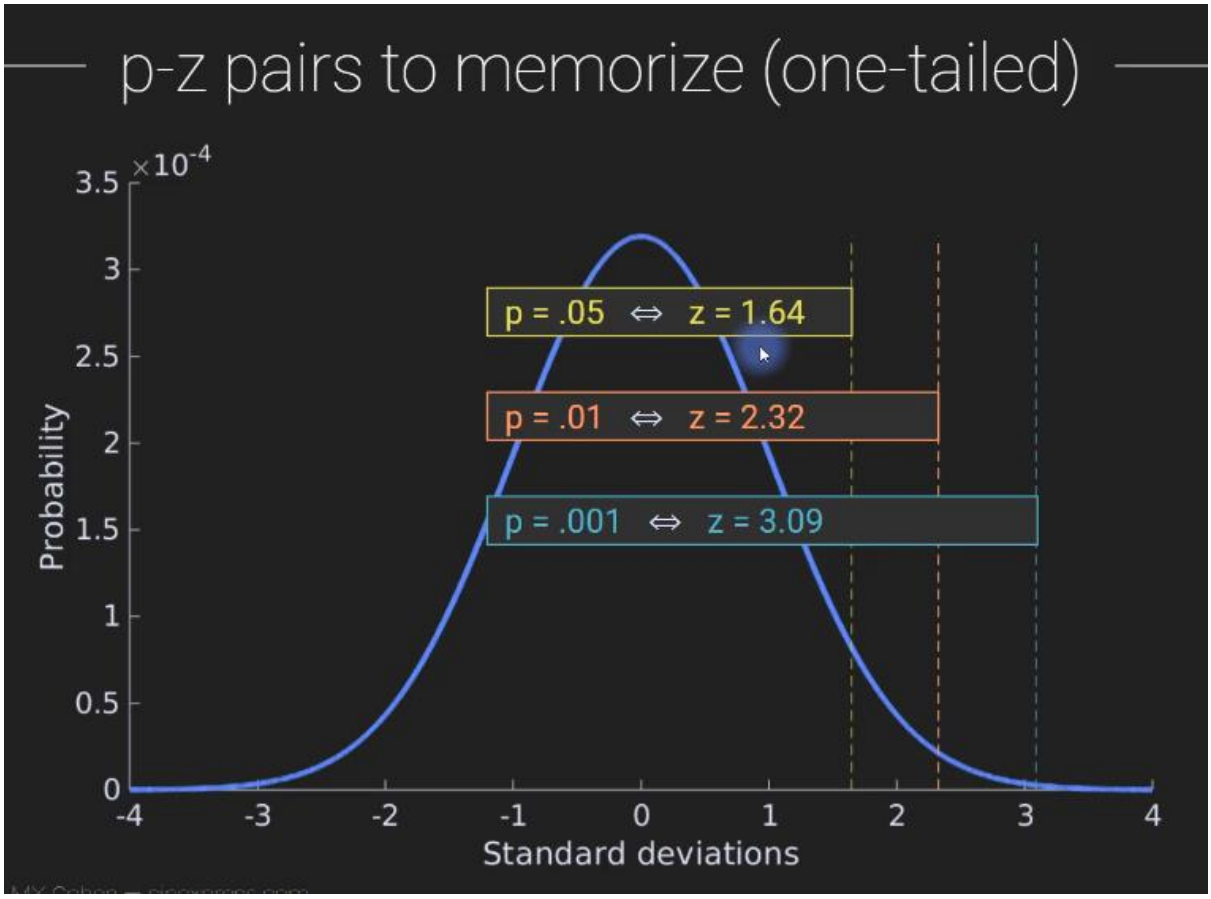
- **Incorrect:** My p-value is smaller than the threshold, so therefore the effect is real.
  Correct: My p-value is smaller than the threshold, so it is unlikely that the effect in the sample would have been observed given the null hypothesis, assuming that the sample is representative of the population.
- **Incorrect:** My p-value is larger than the threshold, so therefore the null hypothesis is true.
  Correct: My p-value is larger than threshold, so it is likely that the effect in the sample would have been observed given the null hypothesis, assuming that the sample is representative of the population. There could be many other explanations for the p-value other than H_0.

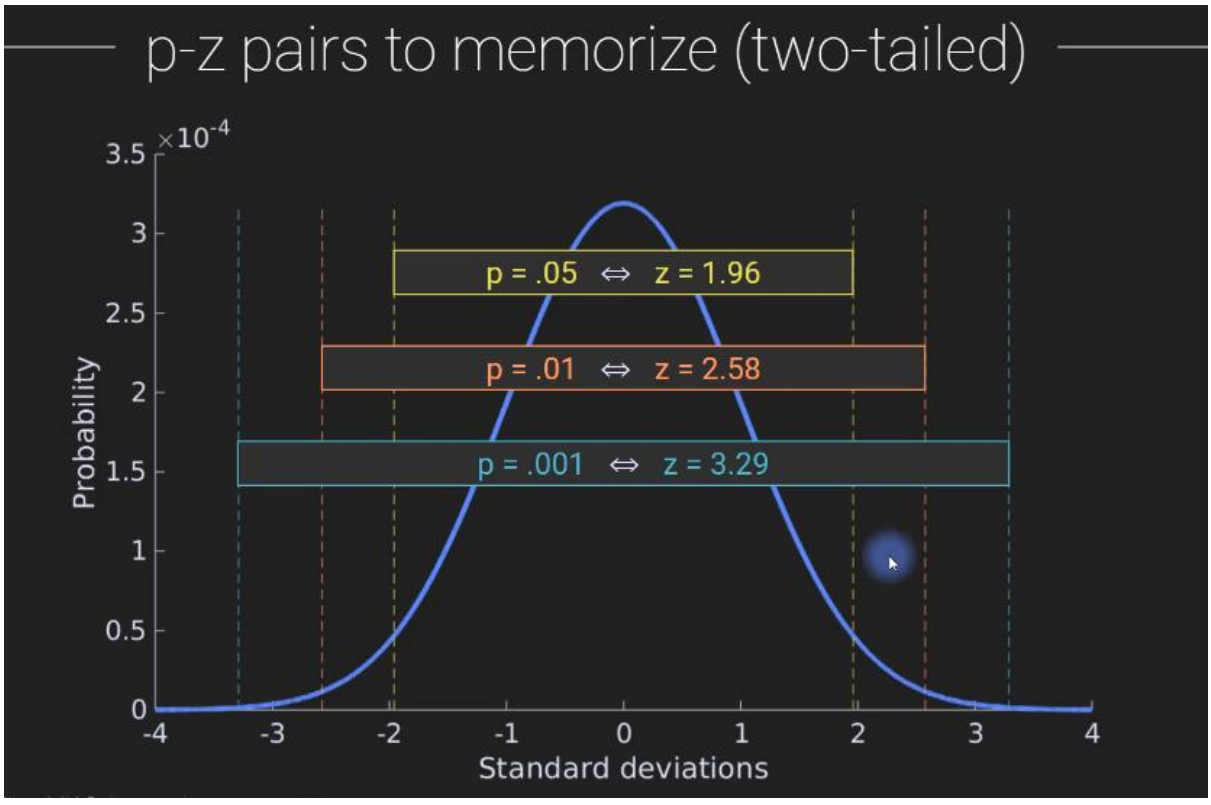
How to compute a p-value?

- There are several ways to compute a p-value, and they depend on the specific statistical test and assumptions made.
- Importantly, the interpretation of a p-value is always the same, regardless of how the p-value was computed.

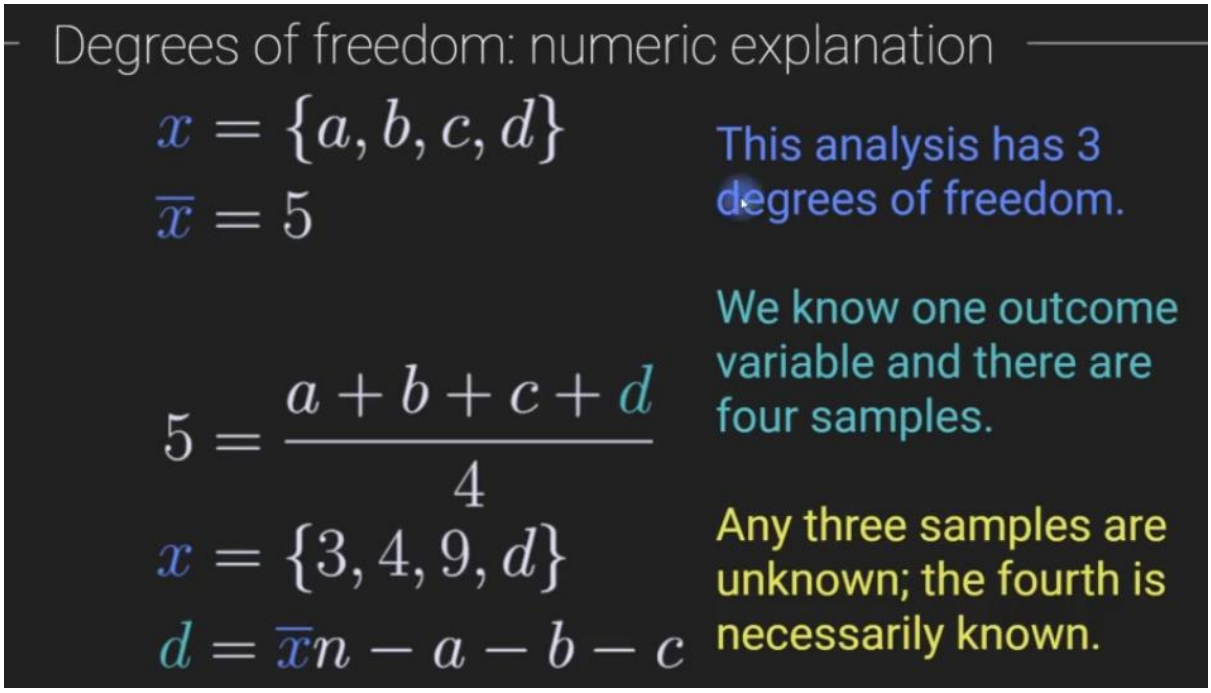
P-Z combinations:

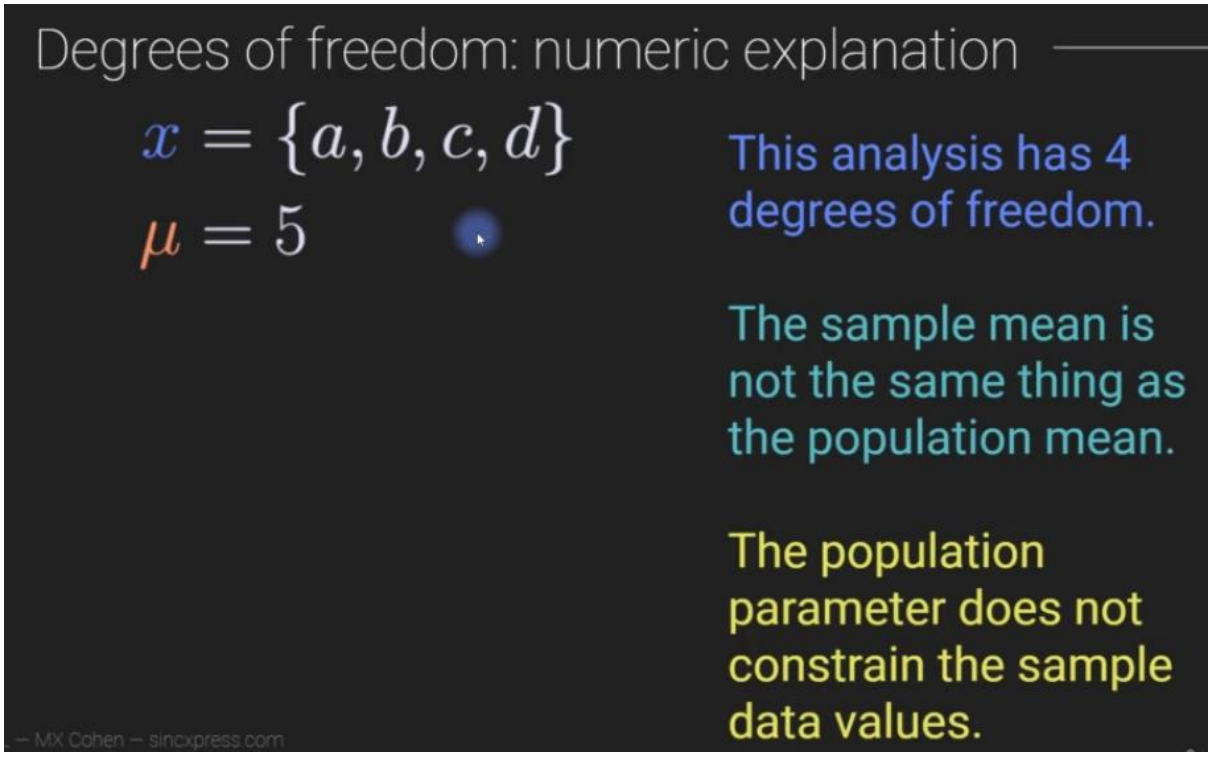



p-z pairs to memorize (one-tailed)
×10^-4
3.5
3
2.5
2
1.5
1
0.5
0
Probability
p = .05 ⇔ z = 1.64
p = .01 ⇔ z = 2.32
p = .001 ⇔ z = 3.09
-4
-3
-2
-1
0
1
2
3
4
Standard deviations

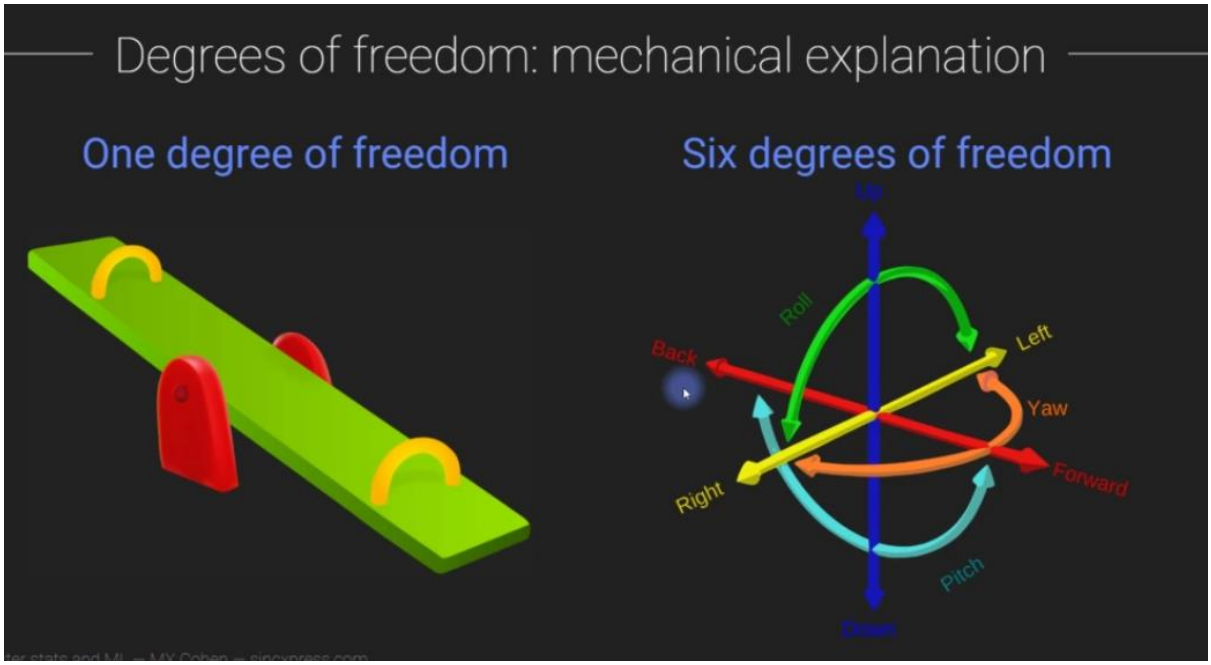

p-z pairs to memorize (two-tailed)
×10^-4
3.5
3
2.5
2
1.5
1
0.5
0
Probability
p = .05 ⇔ z = 1.96
p = .01 ⇔ z = 2.58
p = .001 ⇔ z = 3.29
-4
-3
-2
-1
0
1
2
3
4
Standard deviations

Degrees of Freedom (df):





Where μ is population mean.



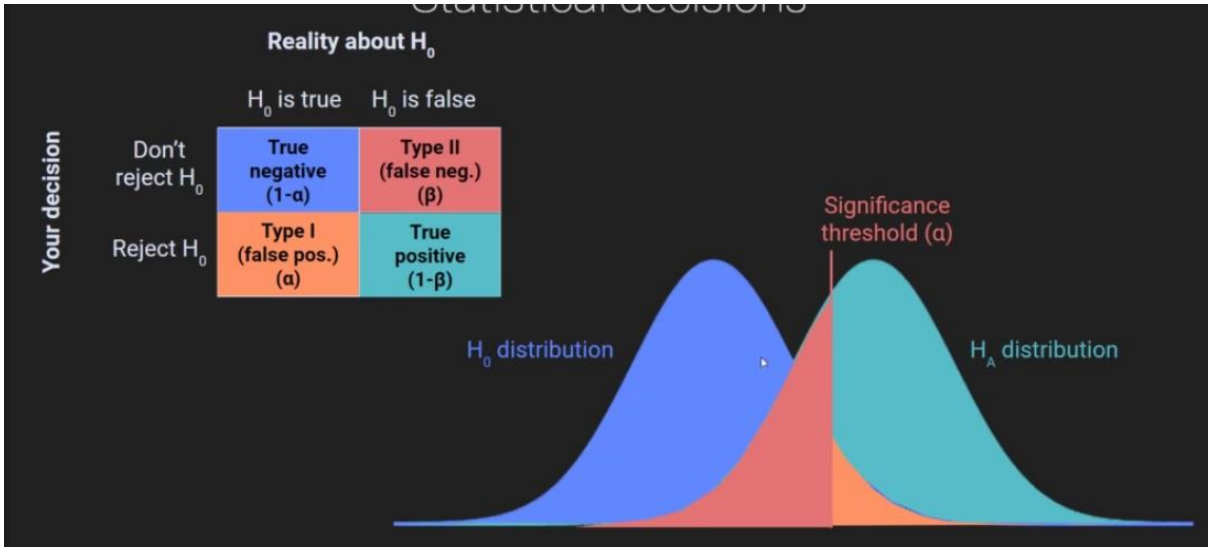
Why are degrees of freedom important?

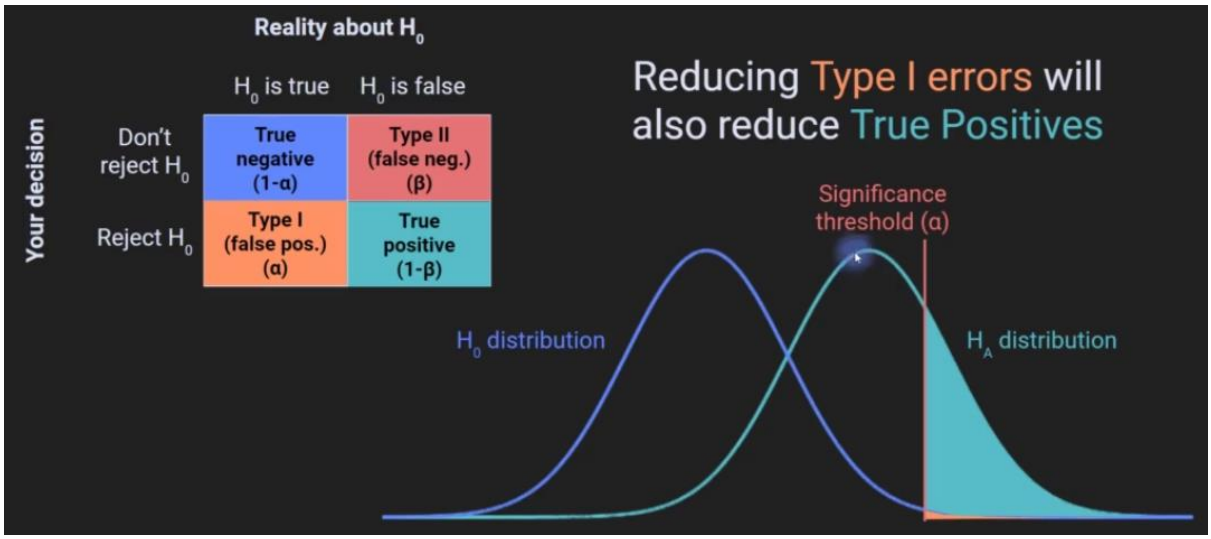
- Degrees of freedom (df) determine the shape of H_0 distributions.
- Higher df generally indicates more power to reject the null hypothesis (related to statistical power).
- df is also a way of checking an ANOVA table for accuracy and understanding.

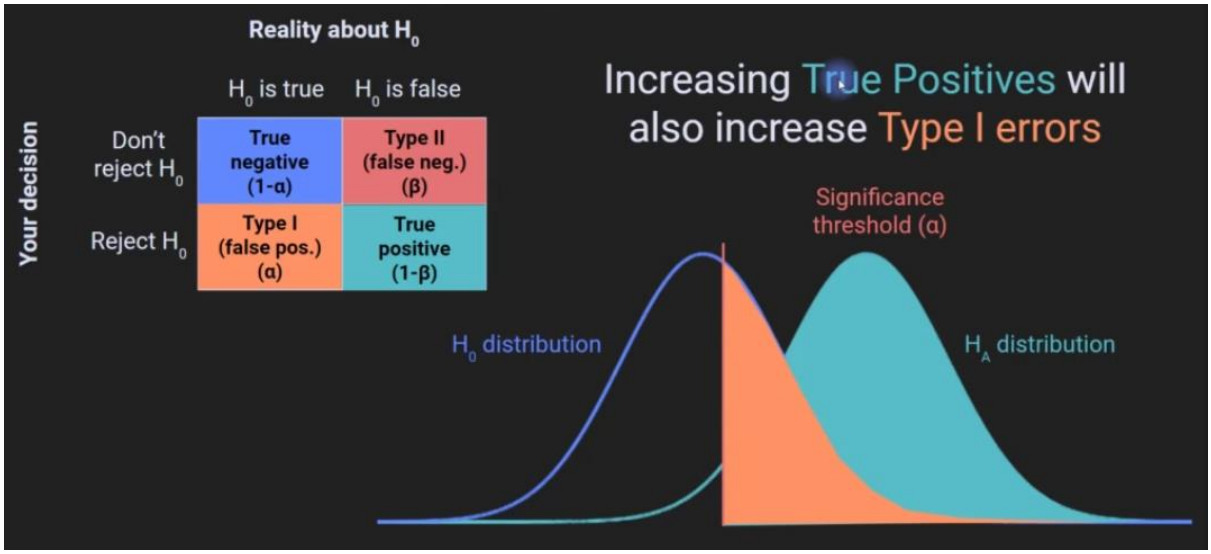
Degrees of Freedom:

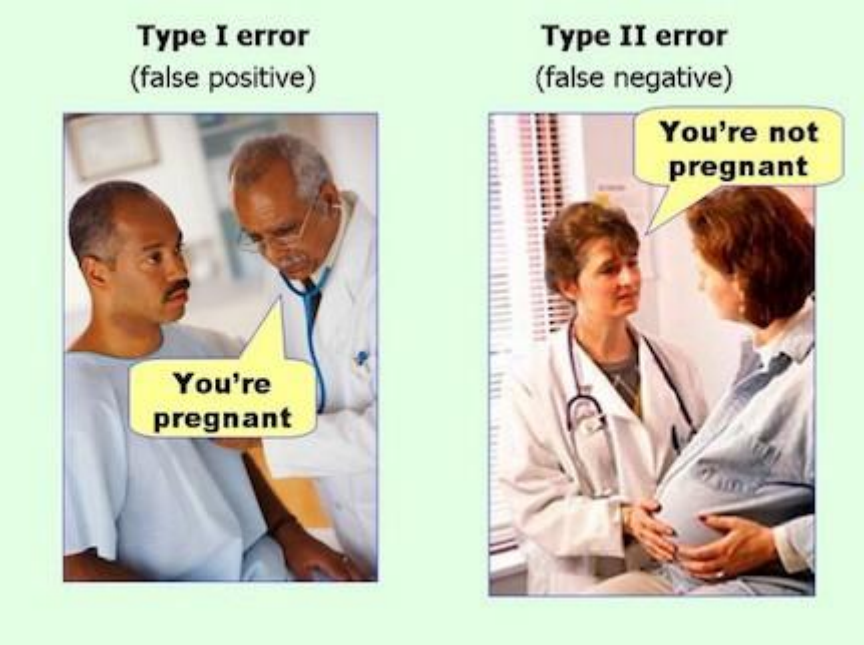
- The number of independent sample values.
- The number of sample data points that can vary.
- The number of data values that are unconstrained by the rest of the data values.
- Generally, $\boldsymbol{df = N - k}$ $\boldsymbol{(N\ data\ points, k\ parameters)}$

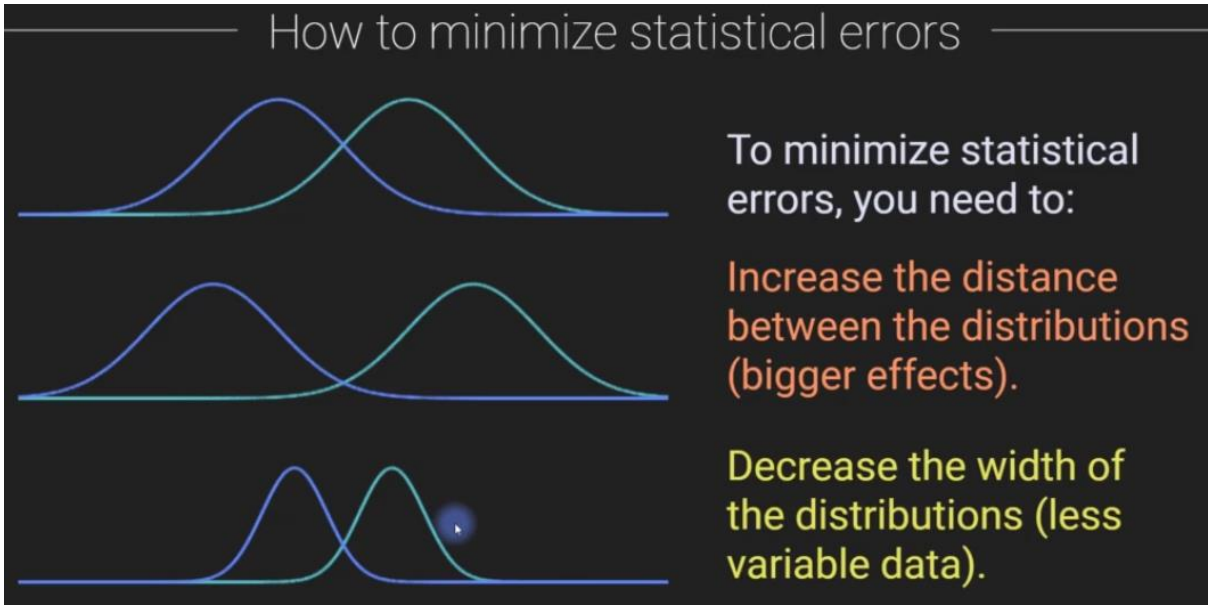
Type 1 & Type 2 Errors:











Parametric vs Non-Parametric Test:

What "non-parametric" doesn't mean: No parameters at all.

Correct meanings: Statistics that are not based on assumptions about underlying distributions (typically, Gaussian).

Statistical inference methods that generate the H_0 distribution from the data, not from an equation.

Parametric Test	Nonparametric Test
• 1-sample t-test • 2-sample t-test • Pearson correlation	• Wilcoxon sign-rank test • Mann-Whitney U test • Spearman correlation

• ANOVA	• Kruskal-Wallis test

Important application of nonparametric statistics: Permutation testing and cross-validation

Parametric statistics:

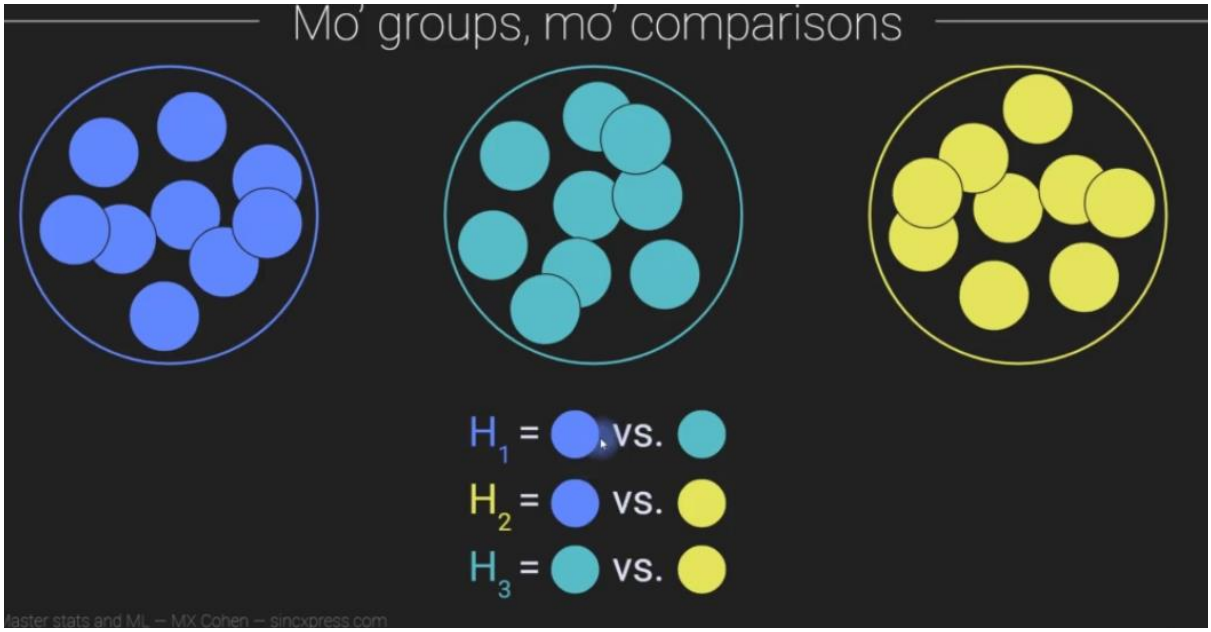
- ⇨ Standard, widely used
- ⇨ Based on assumptions
- ⇨ Assumptions should be tested, (though rarely done)
- ⇨ Can be incorrect when assumptions are violated
- ⇨ Computationally fast
- ⇨ Analytically proven

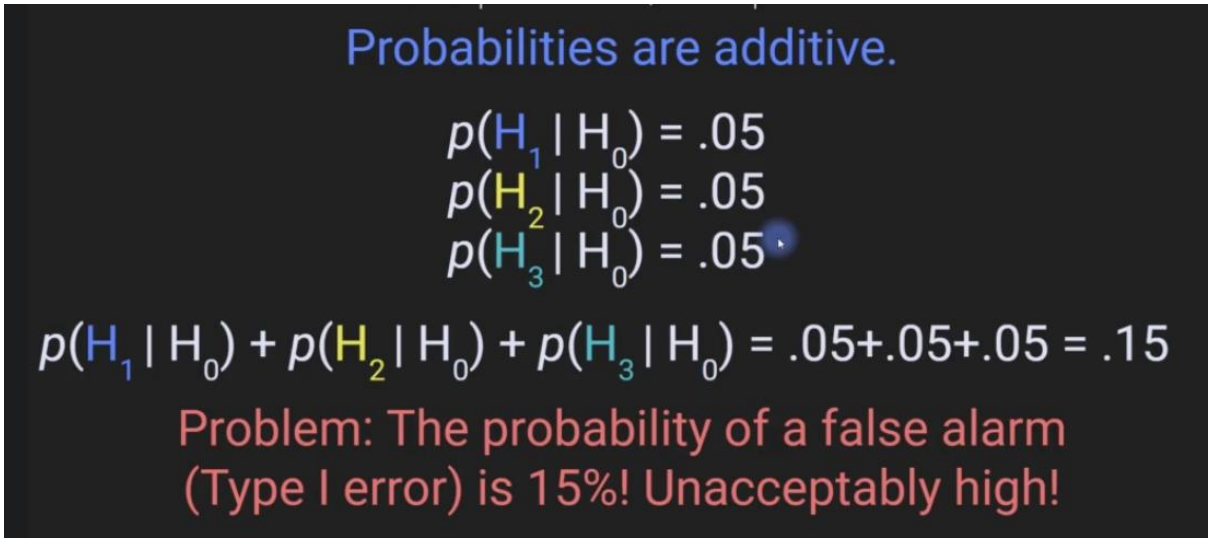
Nonparametric statistics:

- ⇨ Some are nonstandard
- ⇨ "No" assumptions necessary
- ⇨ Can be slow/intensive
- ⇨ Some are sensible algorithms rather than proven methods
- ⇨ Appropriate for non-numeric data
- ⇨ Appropriate for small sample sizes
- ⇨ Some methods give different results each time.

Conclusion: Use parametric methods when possible. Use nonparametric methods when necessary.

Multiple Comparisons & Bonferroni Correction:





Note: This is called the "**Familywise error rate**" (abbreviated FWE or FWER)

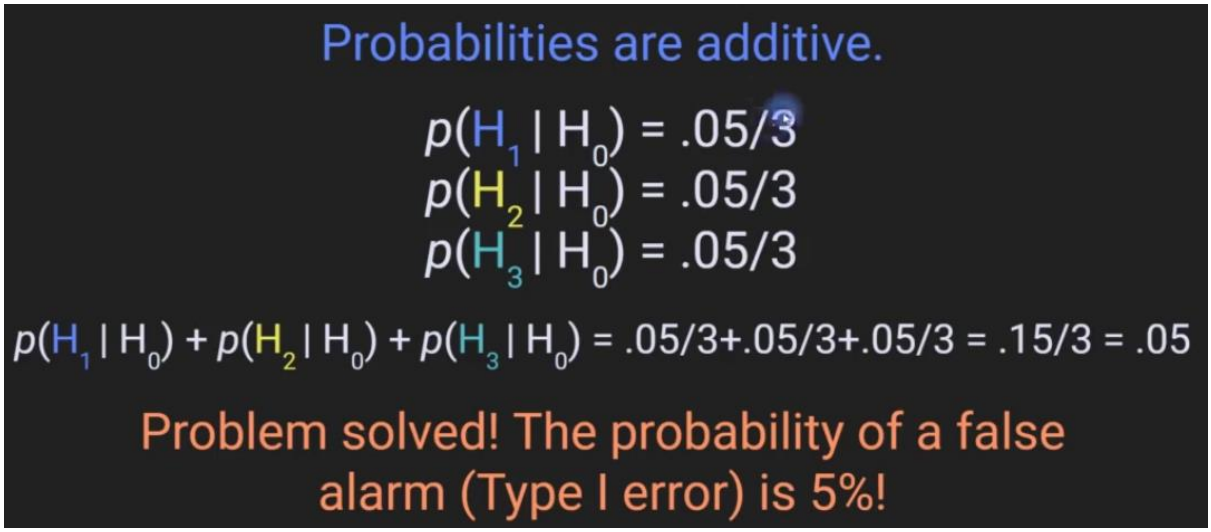
General point: The false alarm rate of *N* tests using a p-value threshold of *a* is *Na*.

Bonferroni Correction:

$$Threshold = \frac{a}{N}$$

a = e.g., 0.005

N = number of tests



Note the difference between the individual α (p=0.0167) and the familywise α (p=0.05).

Statistical vs Theoretical vs Clinical Significance:

Statistical Significance: The probability of observing a test statistic this large given that the null hypothesis is true.

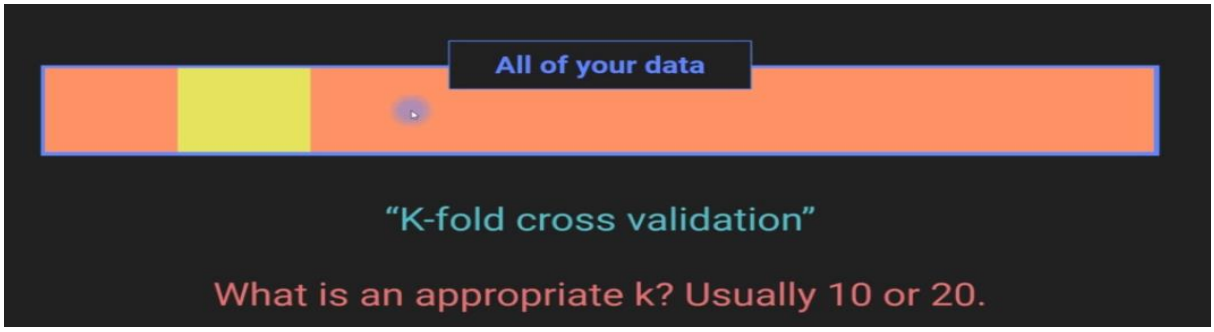
Theoretical Significance: A finding is relevant for a theory or leads to new experiments. This has nothing to do with statistical significance.

Clinical (practical, societal, educational) significance: A finding is relevant for diagnosing or treating a disease.

Examples:

- **Hypothesis & result:**
  - MMR vaccine causes autism. P = 0.79
  - No statistical significance.
  - Strong clinical and societal significance.
- **Hypothesis & result:**
  - Piles of sand with larger grains collapse sooner. P = 0.001
  - Strong statistical significance
  - Strong theoretical significance (scale-free dynamics)
  - No clinical or societal significance.
- **Hypothesis & result:**
  - People who take Mike X Cohen's course get better grades in their stats courses. $P < 10^{-30}$.
  - Strong statistical significance
  - No theoretical significance
  - Some societal significance
  - No clinical significance
  - Strong individual practical significance.

Cross Validation:



Uses of cross-validation:

- ⇨ Variance of an estimate (e.g., confidence intervals). Also called "jack-knifing". Compare the variance of the parameter estimates over different training set.
- ⇨ Avoid bias in analysis results. Applying the model to data it has never seen. Overfitting the training data will decrease performance on the test data.
- ⇨ Compute classification accuracy. Dominant application in machine learning and deep learning.

- Ideally, the test set is truly independent of the training set. This avoids bias and overfitting in the set.
- Be mindful of whether this is the case in your data. For example, a group of self-selected students from the same school.

Statistical Significance vs Classification Accuracy:

How classification works?

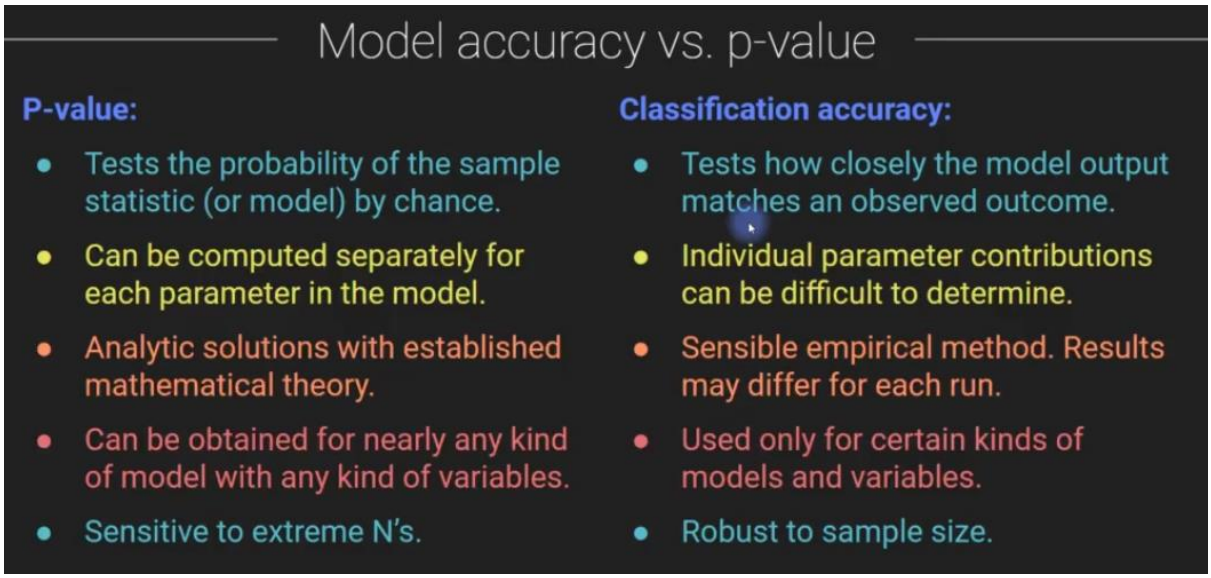
⇨ We want to predict a student will pass or fail the statistics course. The model uses study hours, alcoholic drinks, and web-surfing hours to predict this outcome.

We input the student's behaviour (study, drinking, webbing) into the model and the model gives an output. The model output is not a label; it is a probability e.g., pass 23% / fail 77%. Model output is threshold at 50% to make a binary prediction. This is called *classification*.

Data types for classification:

- ⇨ Classification is mainly appropriate for category data (nominal and ordinal)
- ⇨ Discrete is also possible.

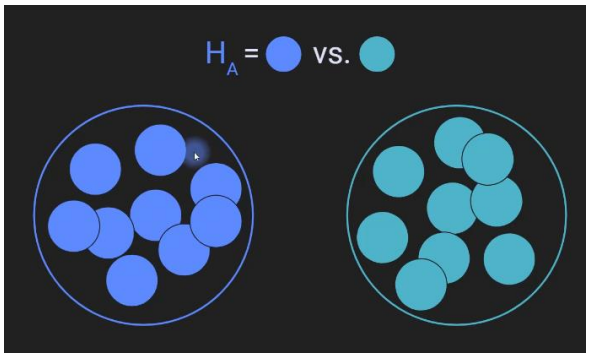
Why not interval or ratio data?

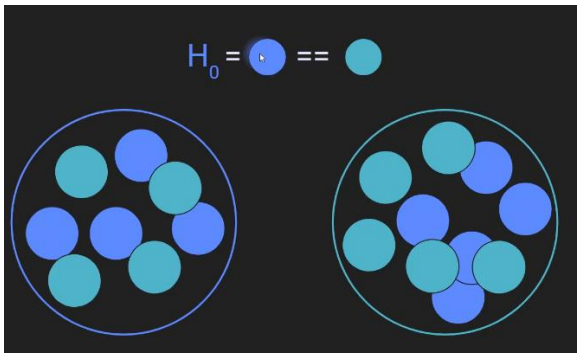


Important: It's not a competition; classification and p-values are different evaluations of a model. They can both be used without conflicting, interfering or misleading.

t-test Family

Purpose & Interpretation:





General formula for t-test:

$$t_k = \frac{\bar{x} - \bar{y}}{s/\sqrt{n}} = \frac{Difference\ of\ means}{Standard\ deviation}$$

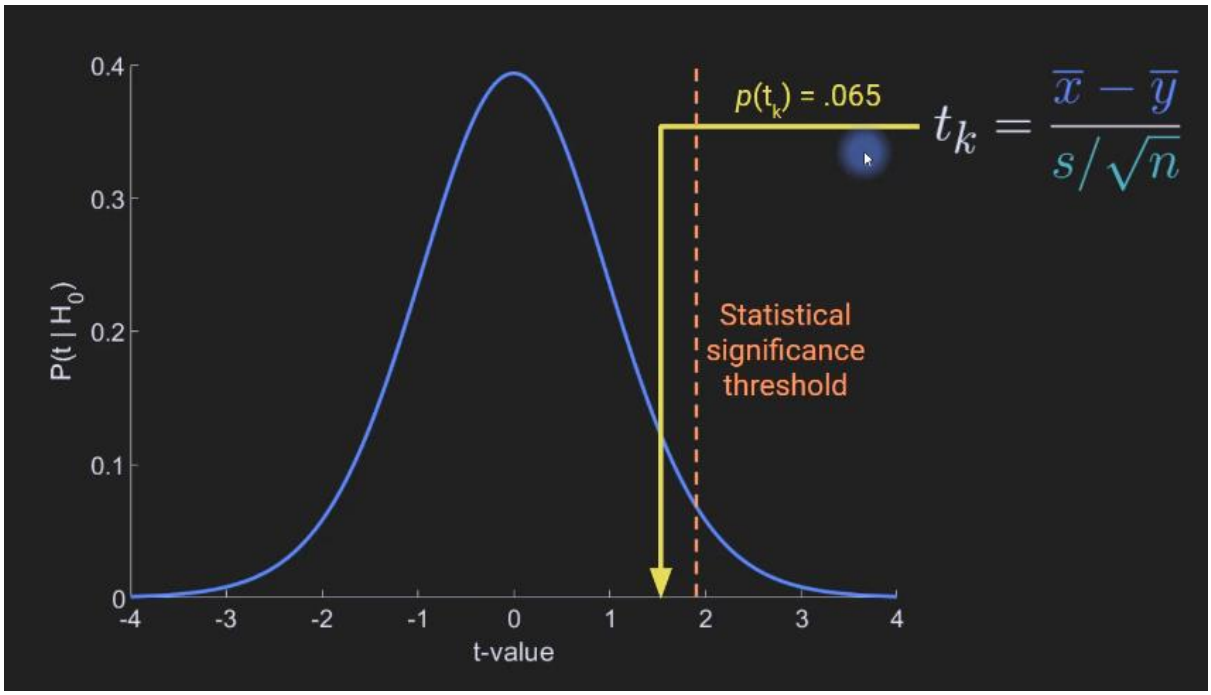
Where,

$\bar{x}$ = average of group 1

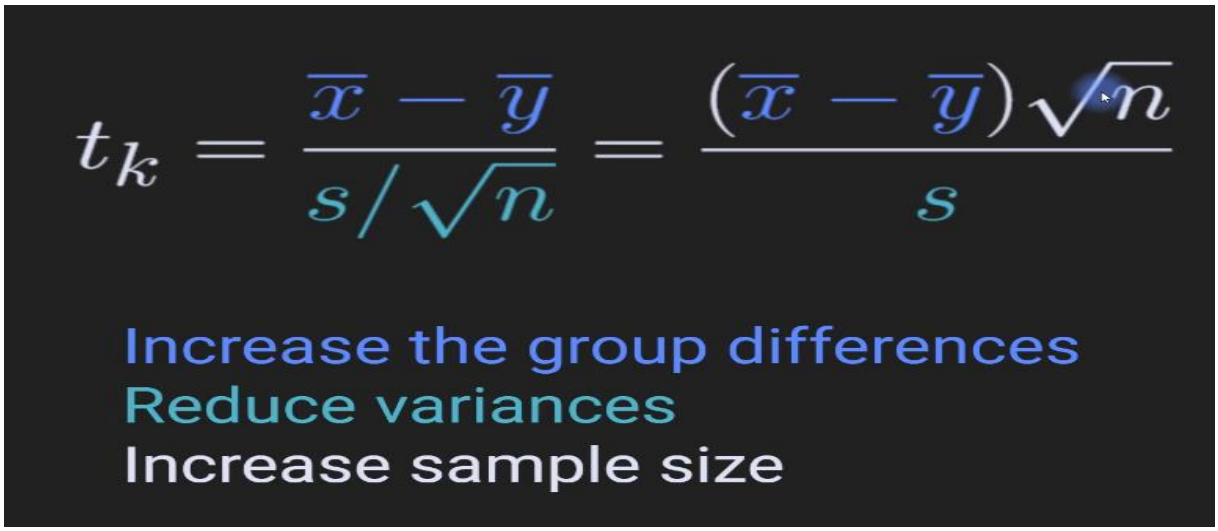
$\bar{y}$ = average of group 2

s = sample standard deviation

$\sqrt{n}$ = number of samples



From the above experiment we can't say that the mean of two groups x & y are same. We can say that we can't find enough evidence to tell that the mean of two groups are different.



One-Sample t-test:

Purpose: Test whether a set of numbers could have been drawn from a distribution with a specific mean.

Example:

- Test whether the IQ of a group of students is significantly different from 100. (Formal statement: Estimate the probability that a certain group of students has an average IQ of 100).

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Where,

$\bar{x}$ = Sample mean

μ = H_0 value

s = Sample standard deviation

n = Number of data points

n-1 = Degree of freedom

One-sample t-test: assumptions:

- Data are numeric (not categorical), ideally interval or ratio (discrete is probably OK)

- ⇨ Data are independent from each other
- ⇨ Data are randomly drawn from the population to which generalization should be made.
- ⇨ Mean and standard deviation are valid descriptors of central tendency and dispersion (i.e., data are approximately normally distributed).

Code: Part 1 - One Sample t-Test

Two-Sample t-test:

Purpose: Test whether two sets of numbers could have been drawn from the same distribution.

Example: Test whether self-reported stress levels changed after 6 weeks of "social distancing". (Formal statement: Estimate the probability that self-reported stress levels before and after 6 weeks of social distancing were drawn from the same distribution.)

Formulas:

- ⇨ There are several two-sample t-test formulas.
- ⇨ The numerator is always the same.
- ⇨ The denominator depends on whether the groups are paired or unpaired, have equal or unequal variance, and have matched of different sample size.

Explanations of t-types:

- **Paired or unpaired:** Whether the two groups of data are drawn from the same or different individuals.
  - **Paired:** The same individuals self-report their stress levels before vs after social distancing. (Same people but measured twice). That's mean the same people helps to create the two samples.
  - **Unpair:** Change in social distancing-related stress in Denmark vs Singapore. That's mean that the two different people helps to create the two samples.
- **Equal or unequal variance:** Whether the two groups have (roughly) equal variance.
  - **Equal variance:** Groups "A" and "B" are Caucasian 20-year-old students from the same university; group "A" studies engineering and group "B" studies computer science. (This is also unpaired t-test)
  - **Unequal variance:** Group "A" is Caucasian 20-year-old students at the same engineering university. Group "B" is a random sample of 20-year-olds from across the country.

- **Equal or unequal sample sizes:** Whether the groups have the same number of values (applies only to unpaired groups).
  - **Equal:** Both groups have N=30
  - **Unequal:** Group "A" is 13 Parkinson's patients; group "B" is 20 matched controls.

Formula for unequal N, unequal variances:

$$t_{df} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_1}}}$$

$$df = n_1 + n_2 - 2$$

Formula for equal N, equal variance, simplify the above equation where $n_1 = n_2$ and $s_1^2 = s_2^2$.

Code: Part 2 - Two Sample t-Test

Wilcoxon Signed Rank (Nonparametric Test):

- Nonparametric t-test **alternative to the one or two sample (paired) t-test.**
- Mainly used when the data do not conform to the normality assumption **that is normal distribution.**
- Tests for differences in medians instead of differences in mean (median is insensitive to outliers).

Test name	When to use
Wilcoxon signed-rank test Signed-rank test	One sample Two dependent (paired) samples
Mann-Whitney U test Mann-Whitney-Wilcoxon U test Wilcoxon rank-sum test	Two independent samples

Algorithm to compute the Wilcoxon test:

⇨ **Step 1: Remove equal pairs:** Remove equal pairs or data points that equal the H_0 value. *Reasoning:* Equal pairs do not contribute to the test either way.

⇨ **Step 2: Rank-transform diffs:**

$$r = rank(|x - y|)$$

Where, x-y are the paired differences. For a one-sample test, y is the H_o value.

⇨ **Step 3: Sum ranks where x > y**

$$W = \sum(r \,.(x > y))$$

- ⇨ **Step 4: Convert to z**

$$Z = \frac{W - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

n is the number of remaining pairs.
Z is normally distributed under H_0 and can be converted to a p-value.

Code: Part 3 - Wilcoxon Signed-Rank for One-Sample or Paired Samples

Mann-Whitney U test (Nonparametric t-test):

- Nonparametric alternative to the **independent two-sample t-test (unpaired samples).**
- Mainly used when the data do not conform to the normality assumption.
- Tests for differences in medians instead of differences in means (median is insensitive to outliers). The two groups **do not need to have the same sample size.**

Algorithm to compute the Mann-Whitney U test:

- ⇨ **Step 1: Note the N's:**

$$x_f = Dataset\ with\ fewer\ points$$
$$x_m = Dataset\ with\ more\ points$$
$$n_f = Smaller\ sample\ size$$
$$n_m = Larger\ sample\ size$$

- ⇨ **Step 2: Pool data and rank:** $r = rank(\{x_f, x_m\})$
  {…} is the set created by concatenating the group with more data points after the group with fewer data points.
- ⇨ **Step 3: Compute U:**

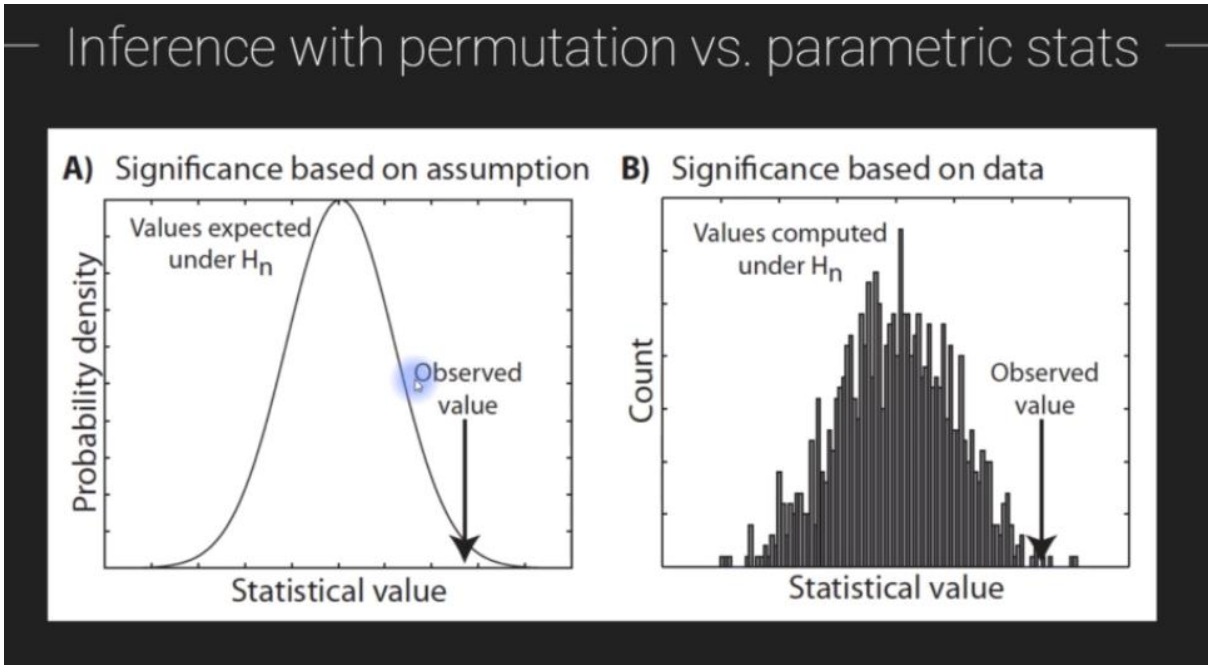
$$\boldsymbol{U} = \sum_{\boldsymbol{i=1}}^{n_f} r_i$$

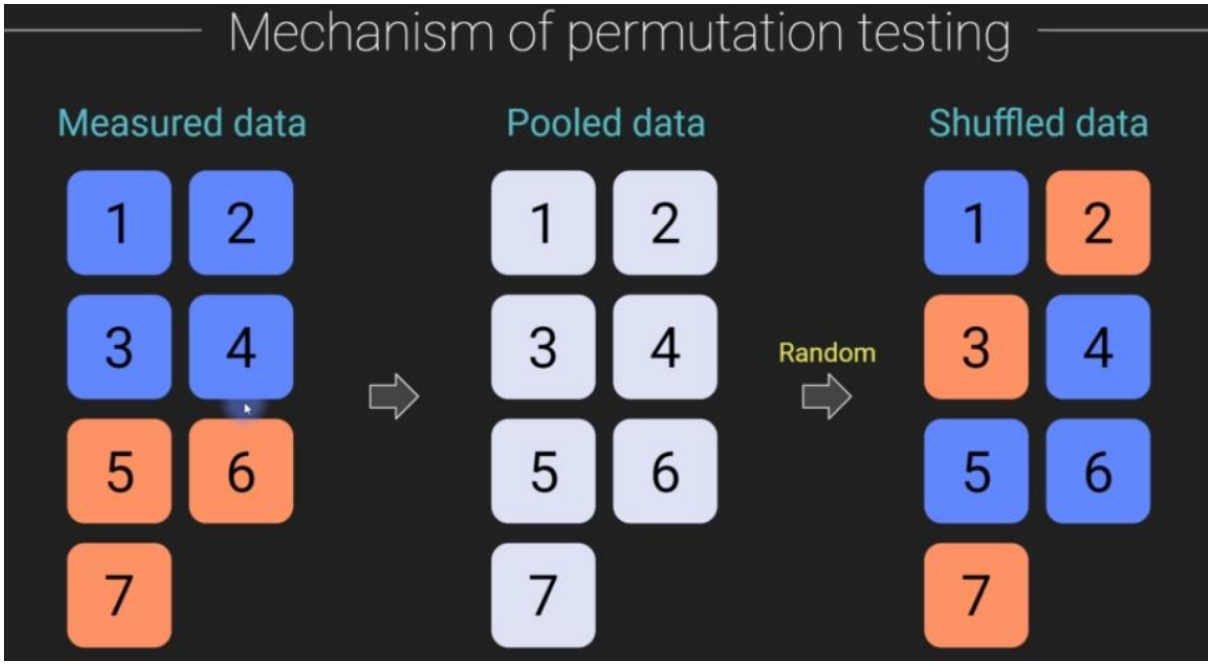
n_f is the smaller of the sample sizes.

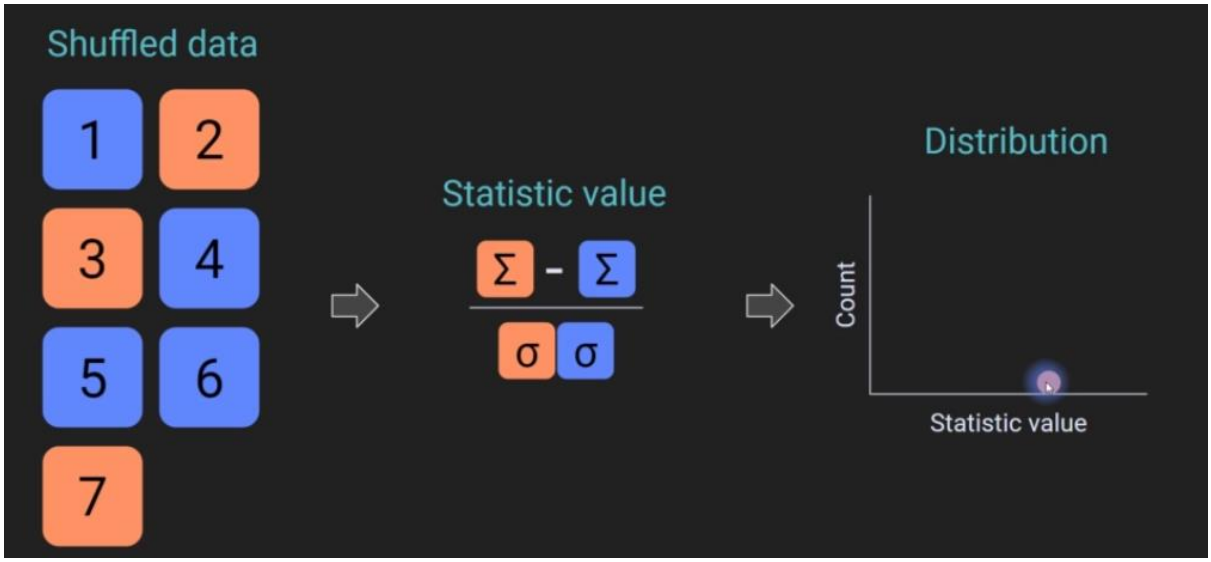
- ⇨ **Step 4: Convert to Z:** Z = … The formula is rather long… Z is normally distributed under H_0 and can be converted to a p-value.

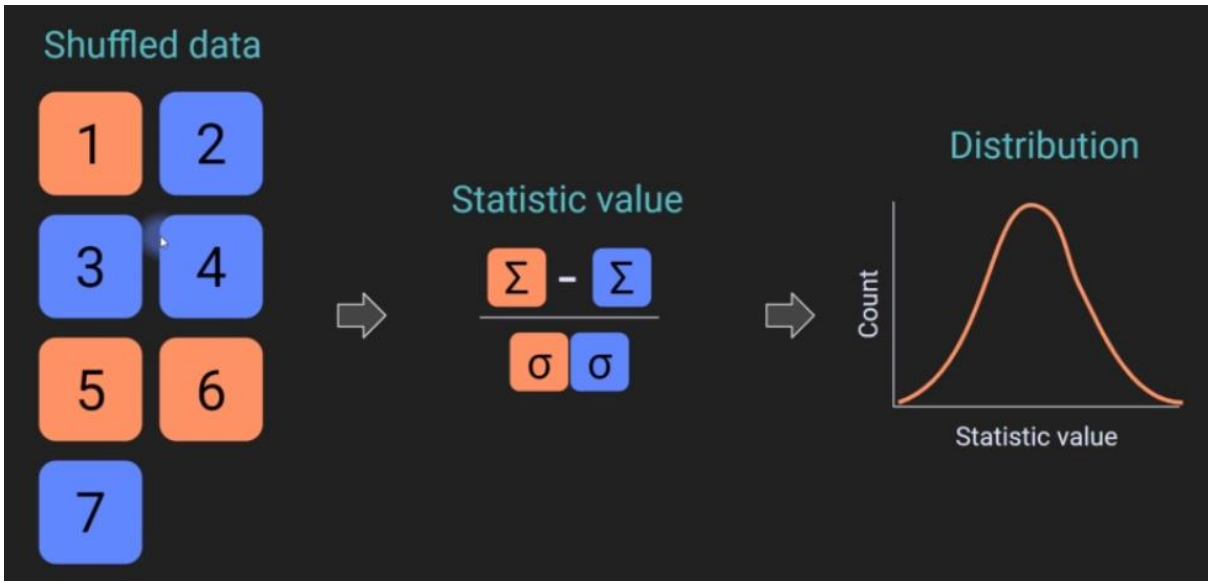
Code: Part 4 - Mann-Whitney U test

Permutation Testing for t-test:

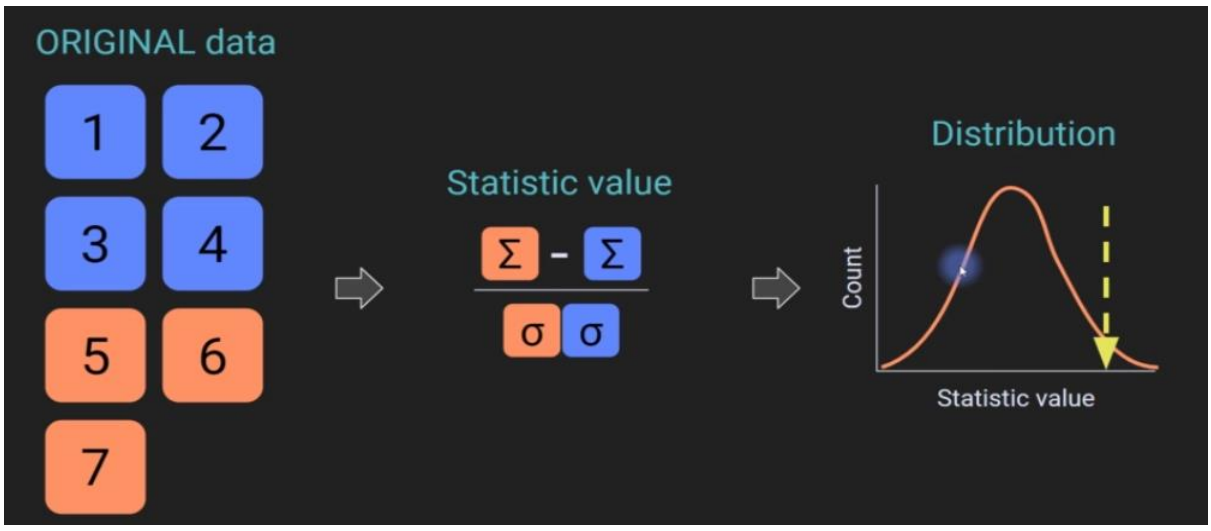


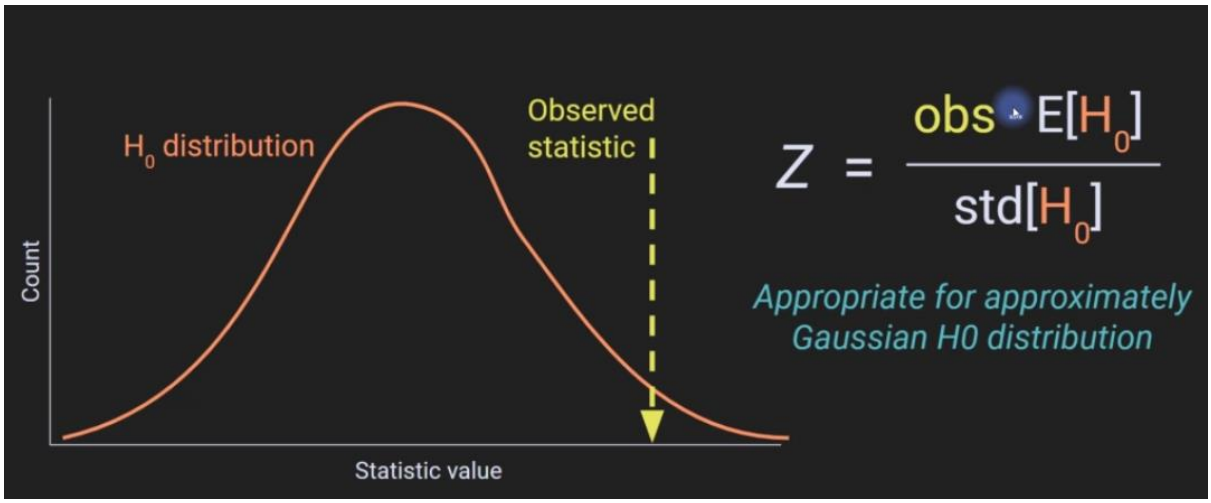


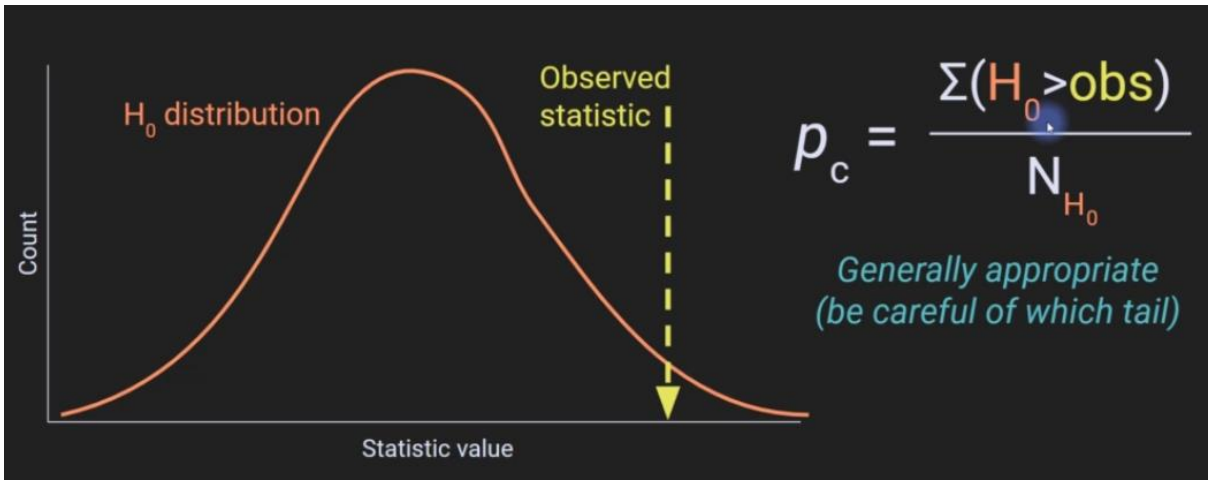




The above slide creates the H_0 distribution. Then compute the below hypothesis test by using the original dataset.



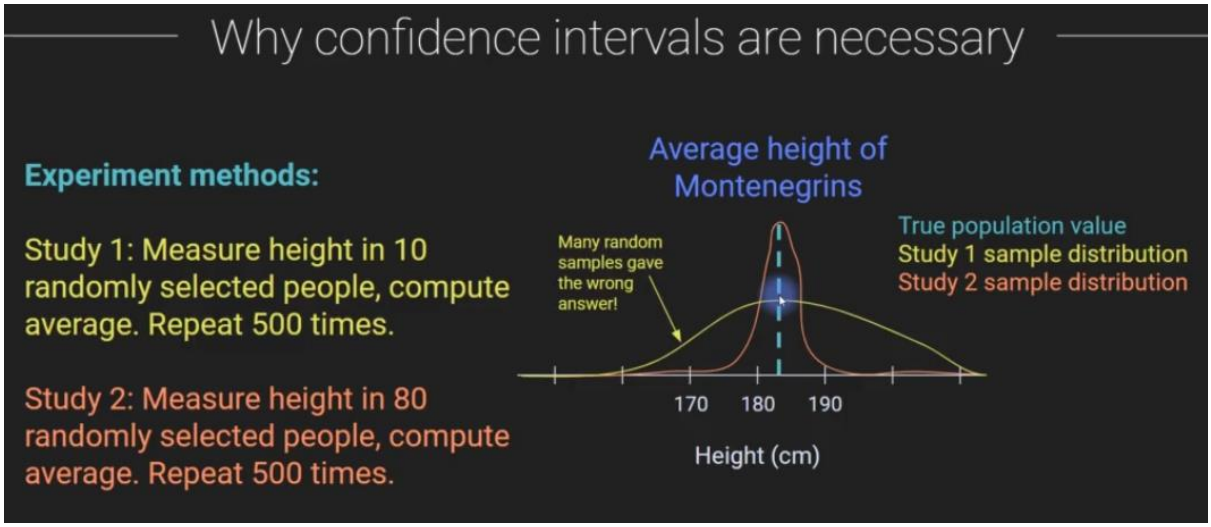


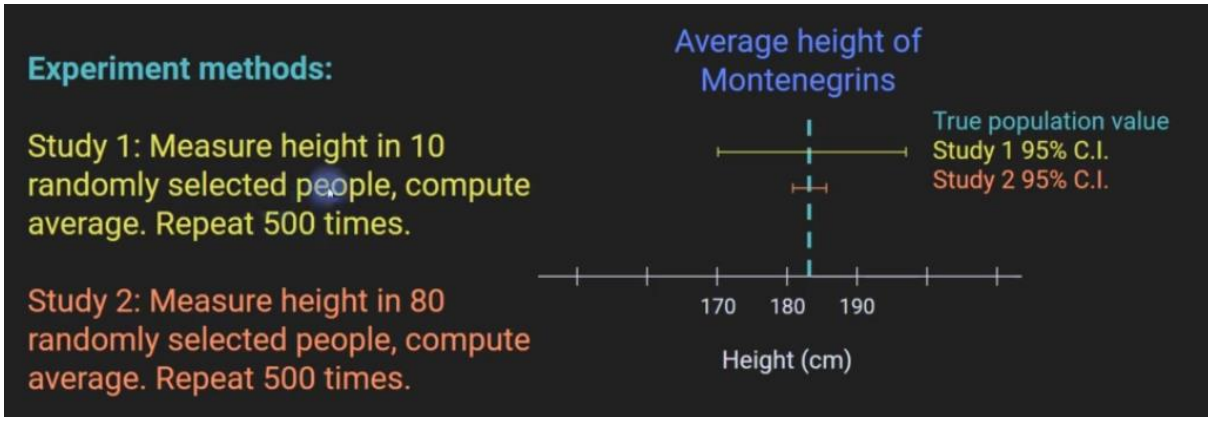


Code: Part 5 - Permutation Testing

Confidence Intervals on Parameters

What are Confidence Intervals:





Confidence Interval (CI): The probability that an unknown population parameter falls within a range of values in repeated samples.

$$P(L < \mu < U) = c$$

Where,

L = Lower boundary of CI

U = Upper boundary of CI

c = Proportion of CI (in the example, 95%)

μ = Population mean or population parameter

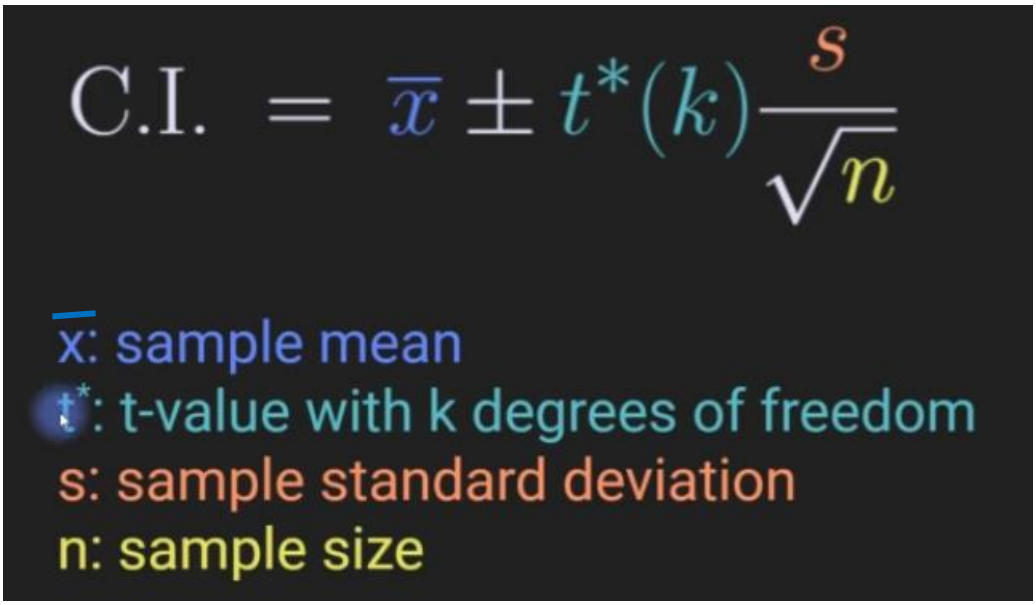
Typical confidence interval probabilities: 95%, 99% or 90%

**The proportion of a large number of samples that will include the population parameter within its confidence interval.*

Factors that influence confidence intervals:

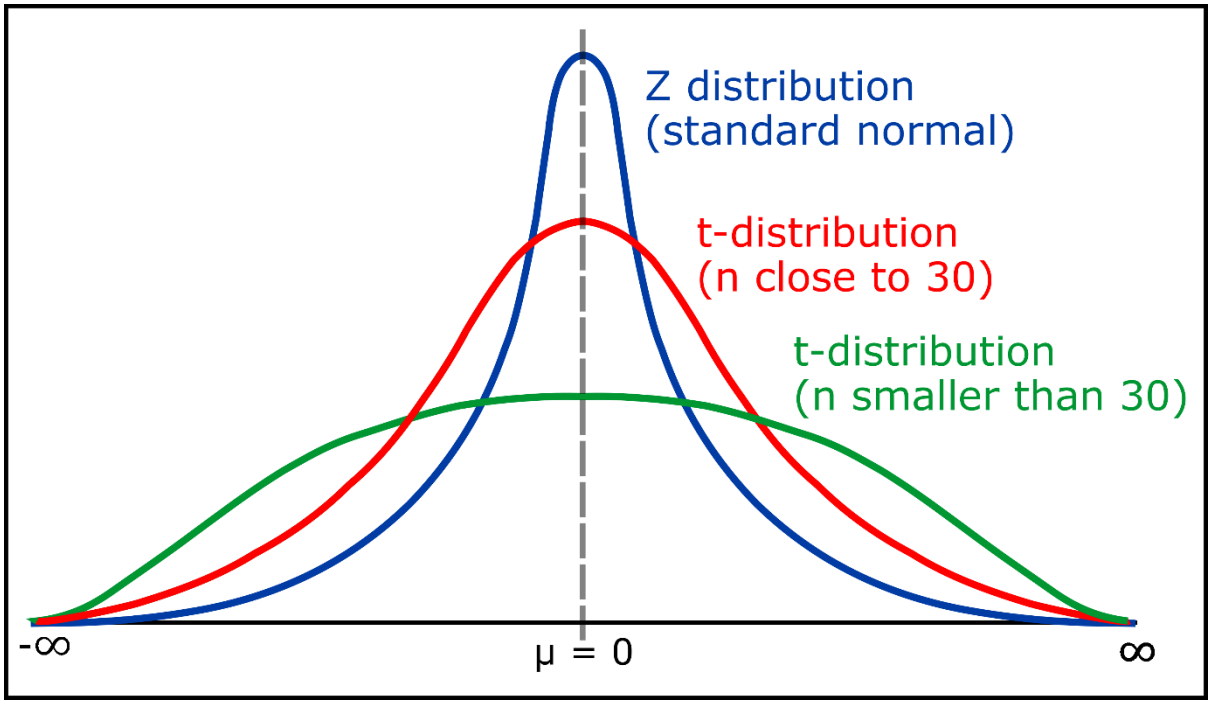
- ⇨ Confidence intervals are influenced by the sample size and variance.
- ⇨ When sample size is larger, confidence intervals are closer together.
- ⇨ When the variance is smaller, confidence intervals are closer together.

Computing Confidence Intervals:



- ⇨ S gets smaller, CI gets smaller
- ⇨ n gets larger, CI gets smaller

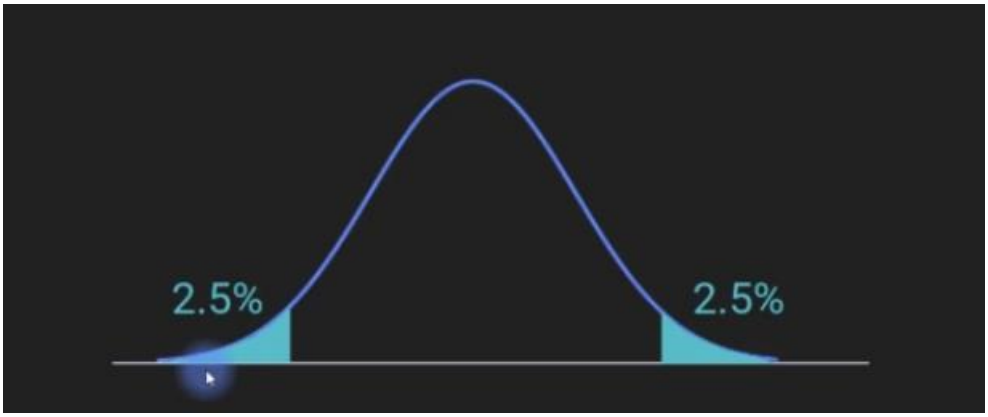
t^*: t comes from t-distribution, not from t-test.

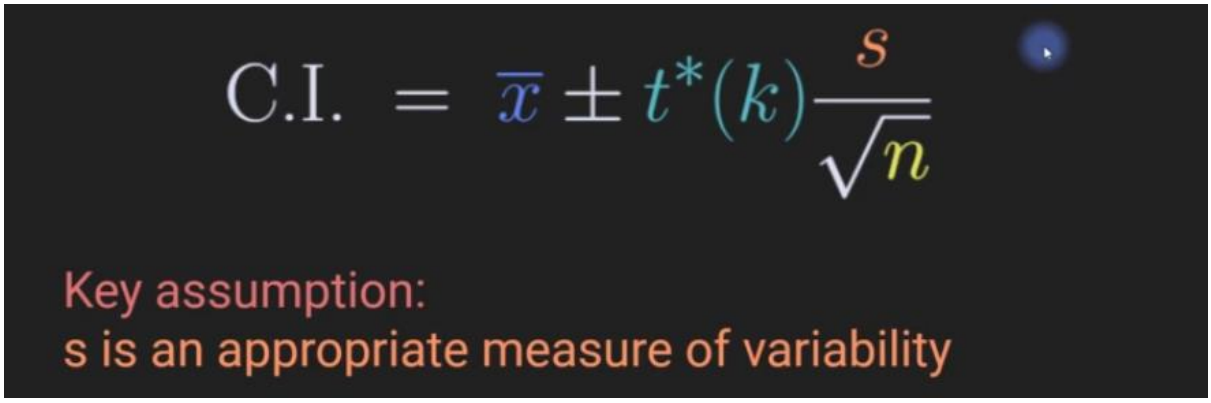


$$\text{C.I.} = \overline{x} \pm t^*(k)\frac{s}{\sqrt{n}}$$

t*: t-value associated with *one tail* of the confidence interval: *(1-C)/2 = tinv((1-.95)/2,n-1)*

example: 95% C.I., n=20 ⇒ t* = 2.093



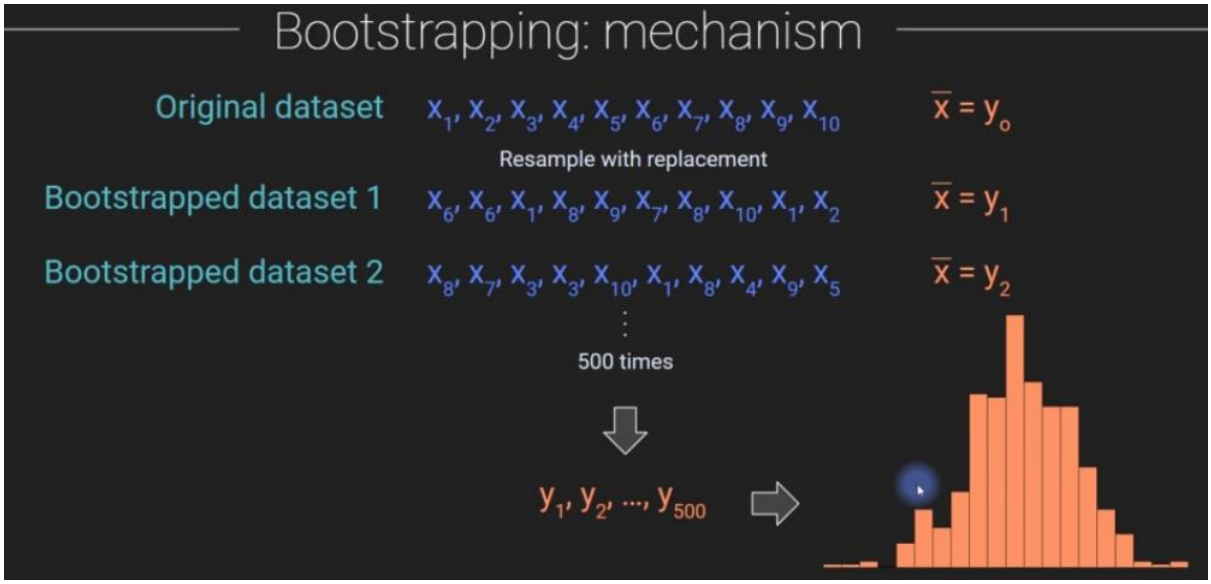


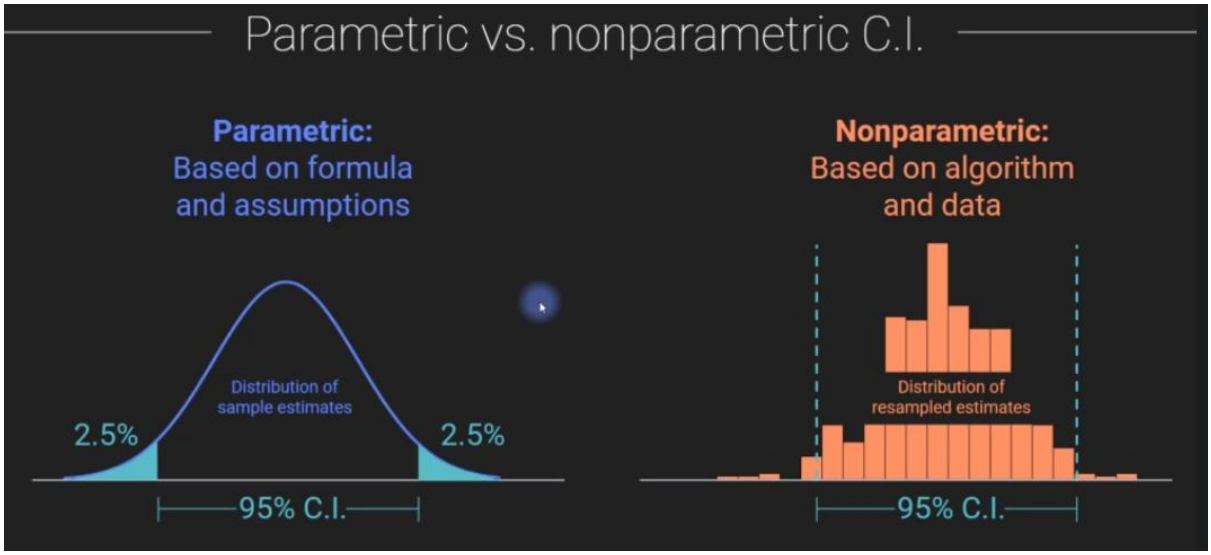
Code: Part 1 - Compute Confidence Intervals by Formula

Confidence Intervals Via Bootstrapping:

Bootstrapping C.I. The Idea:

- Instead of using a formula to compute confidence intervals, compute them directly based on the data.
- This is done by repeatedly randomly resampling from your dataset.
- Thus, pretend that your sample is the population, and the resampling is the sample.





Bootstrapping Advantages:

- ⇨ Works for any kind of parameter (mean, variance, correlation, median etc)
- ⇨ Useful for limited data (e.g., no experiment repetitions)
- ⇨ Not based on assumptions of normality.

Bootstrapping Limitations:

- ⇨ Gives (slightly) different results each time
- ⇨ Can be time-consuming for large datasets
- ⇨ Sample must be a good representation of the population.

Code: Part 2 - Bootstrapping Confidence Intervals

Misconceptions about Confidence Intervals:

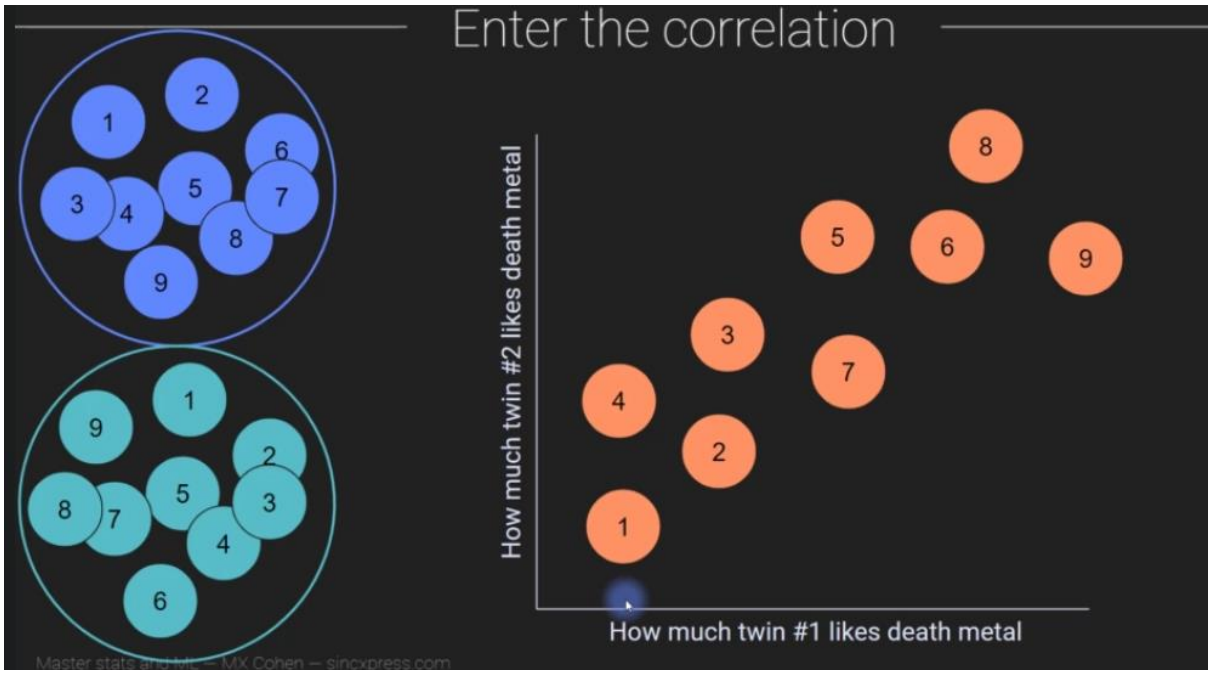
$$P(L < \mu < U) = \bar{x} \pm t^*(k)\frac{s}{\sqrt{n}}$$

- ⇨ **Incorrect:** I am 95% confident that the population mean is the sample mean.
  Correct: 95% of confidence intervals in repeated samples will contain the true population mean.
- ⇨ **Incorrect:** I am 95% confident that the population mean is within the CI in my dataset.
  Correct: See above example. Also, the confidence refers to the estimate, not to the population parameter.
- ⇨ **Incorrect:** 95% of the data are between L and U
  Correct: The CI is not based on the raw data; it's based on the descriptive statistics of the sample data.

⇨ **Incorrect:** Confidence intervals for two parameters overlap; therefore, they cannot be significantly different.
Correct: The CI refers to the estimate of a parameter, not to the relationship between parameters.

Correlation

Motivation & Description of Correlation:

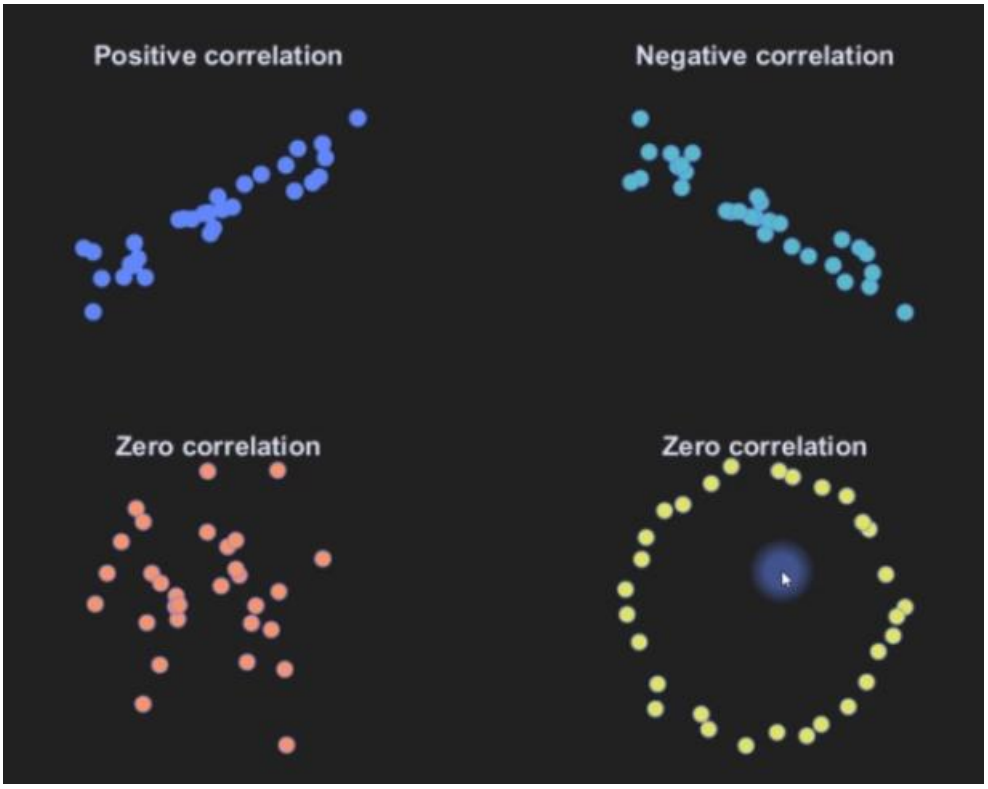


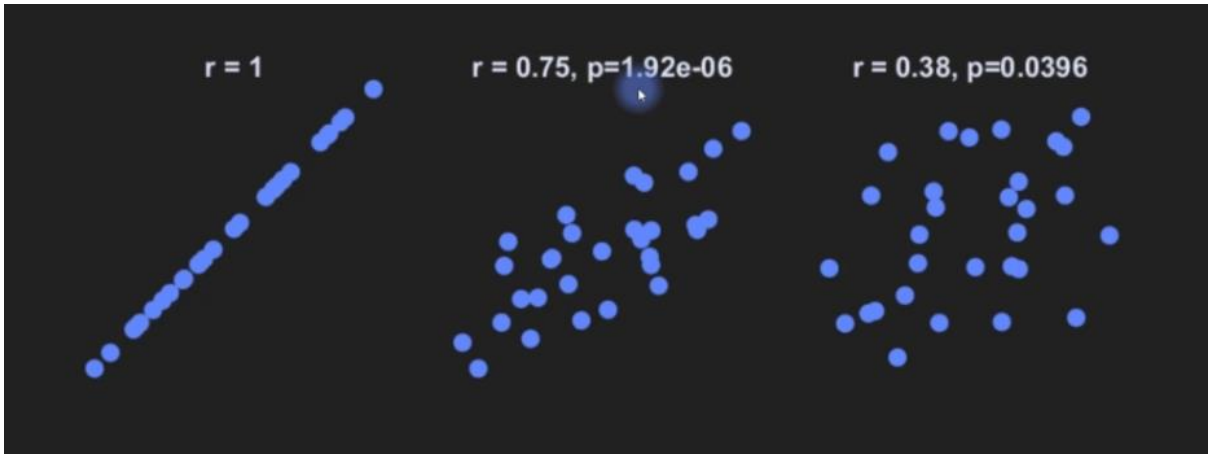
A correlation analysis computes a correlation coefficient (termed "**r**")

The correlation coefficient r is a single number that shows the **linear relationship** between two variables.

The correlation coefficient **varies between -1 and +1**. -1 means a perfect **inverse relationship,** 0 means no relationship and +1 means a perfect **positive relationship.**

The correlation coefficient itself is a continuous measure of correlation strength. A corresponding p-value must be computed to interpret its statistical significance.





Correlation vs causation:

- ⇨ Correlation merely shows a relationship
- ⇨ It does not reveal or imply causality
- ⇨ Causality can be demonstrated by experimental manipulations.
- ⇨ Example: ice cream → shark attacks

Covariance(c) and Correlation(r):

Covariance vs Correlation:

- Covariance is a single number that measures the linear relationship between two variables.
- Correlation is the scaled covariance. Covariance is ranged from -x to +x but correlation lies between -1 to +1. The values of |-x| and |x| are not same.
- Covariance is in the same scale as the original data. It is not normalized. Correlation is normalized to be independent of the data scale.

$$c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

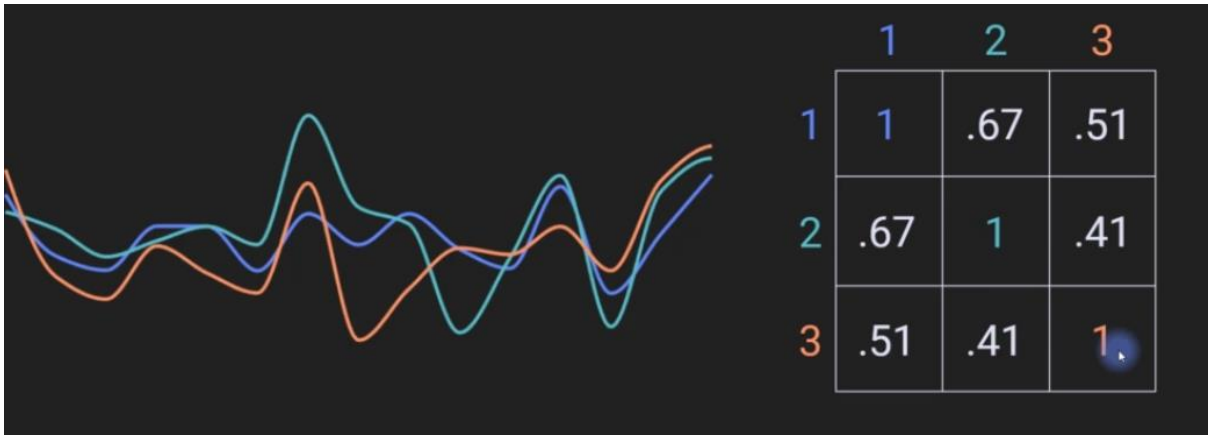
P-value of correlation coefficient:

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

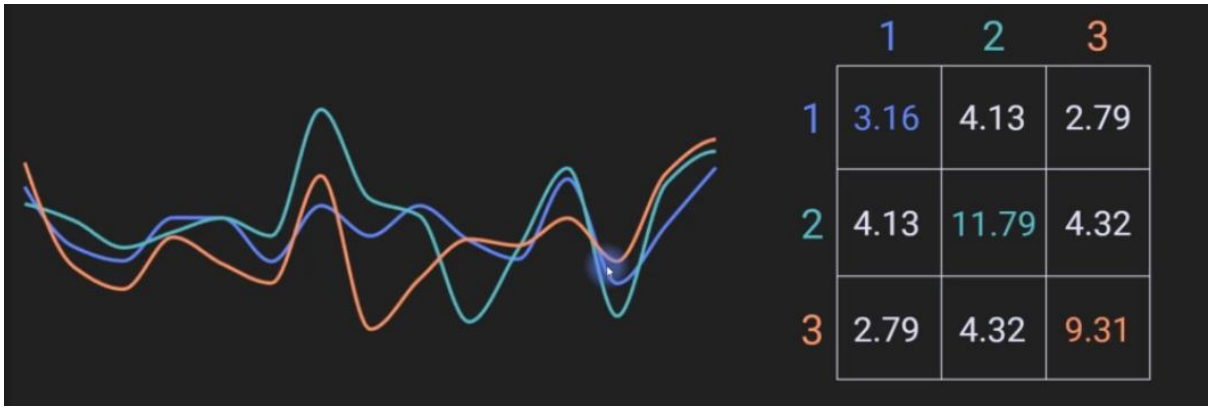
Statistical significance is computed from a t-value that is based on the strength of the correlation and the number of data points.

Code: Part 1 – Correlation coefficient & Part 2 - Simulate Data with Specified Correlation

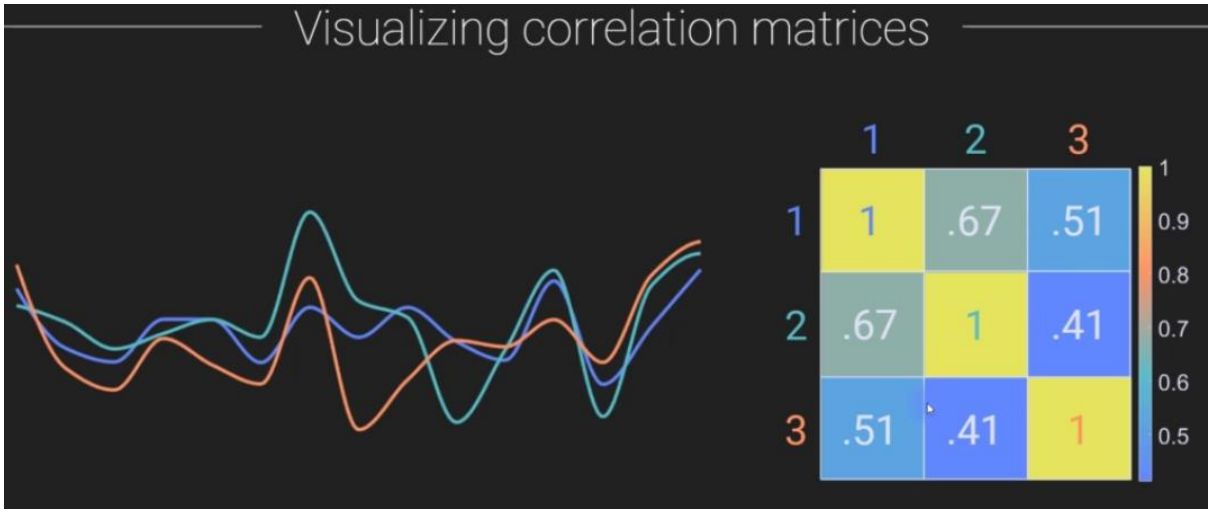
Correlation Matrix:



Covariance Matrix:



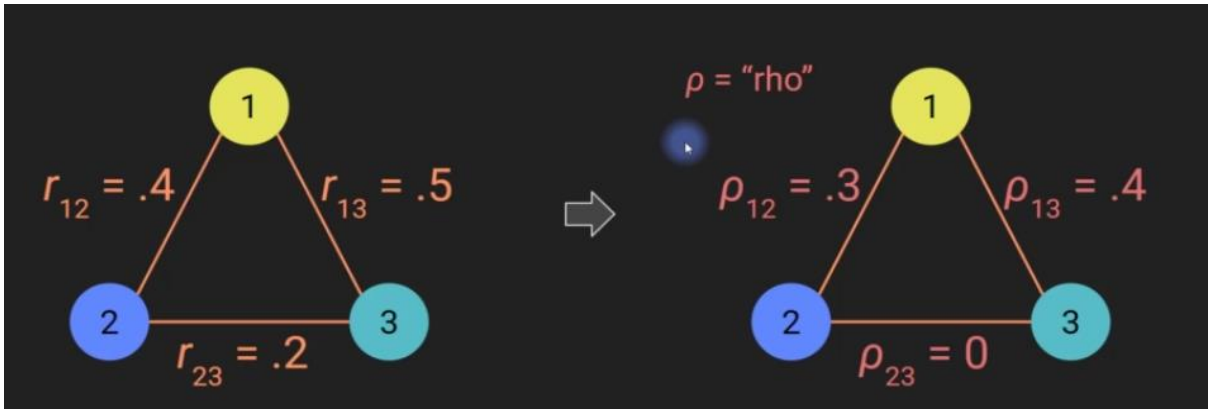
Why are not the diagonal values same or 1 for covariance matrix? Actually, the diagonal values of the covariance matrix tell the variance. Can prove this by the covariance formula where $x_i = y_i$.

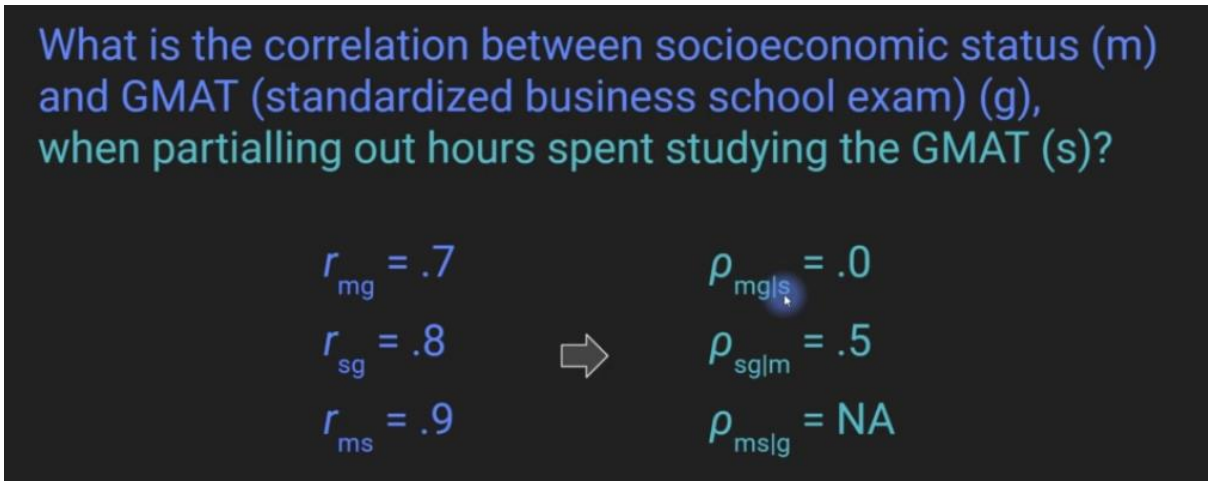


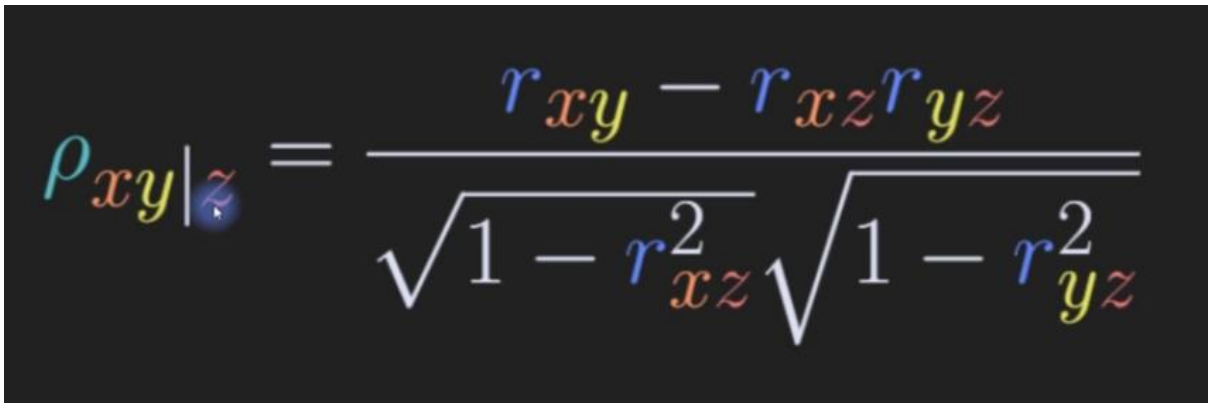
This color visualization helps for those data which have many variables.

Code: Part 3 – Correlation Matrix

Partial Correlation:



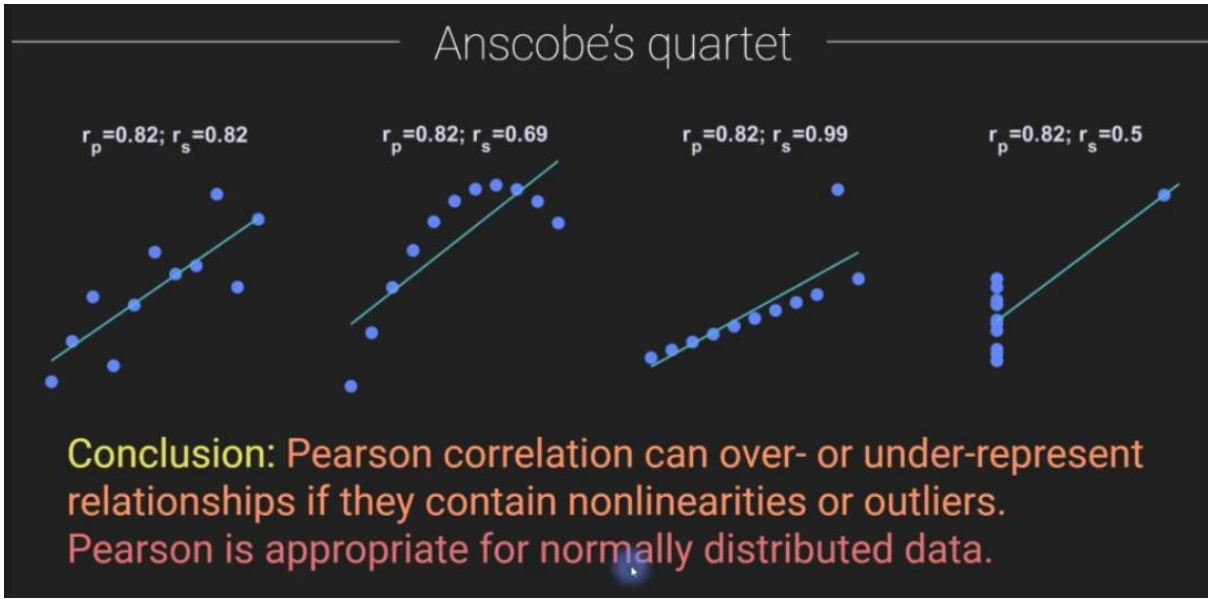




Code: Part 4 - Partial Correlations

The Problem with Pearson(r_p):

Till now, we are doing Pearson Correlation.



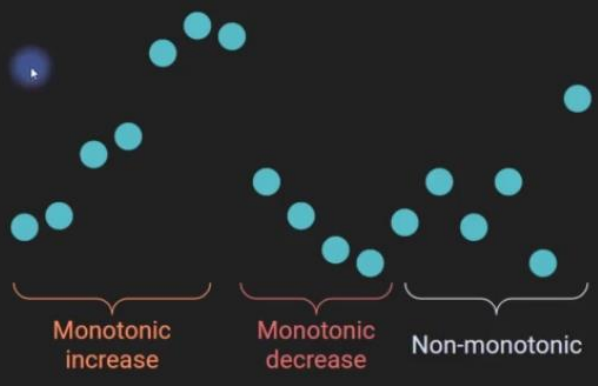
Nonparametric Correlation: Spearman rank correlation (r_s):

- Pearson's t inflates or deflates nonlinear relationships.
- Pearson's r is sensitive to outliers.
- Thus, Pearson's r is appropriate for normal data without outliers.
- Pearson and Spearman converge when data are normally distributed.

Spearman (rank) correlation

Spearman's rho tests for a monotonic relationship, regardless of whether the relationship is linear or nonlinear.

A monotonic relationship tests for increasing or decreasing numbers, regardless of the spacing between numbers.



How to compute the Spearman correlation

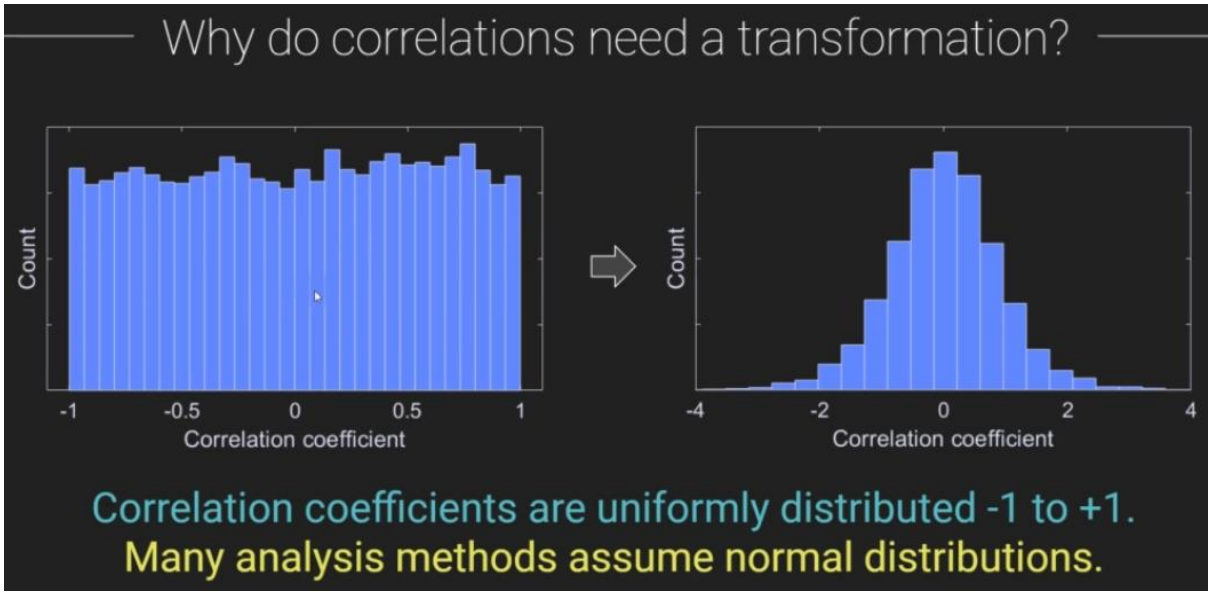
Step 1: Transform both variables to rank.
([3321654, -40, 1, 0] ⇒ [4, 1, 3, 2]).

Step 2: Compute Pearson correlation coefficient on ranks.

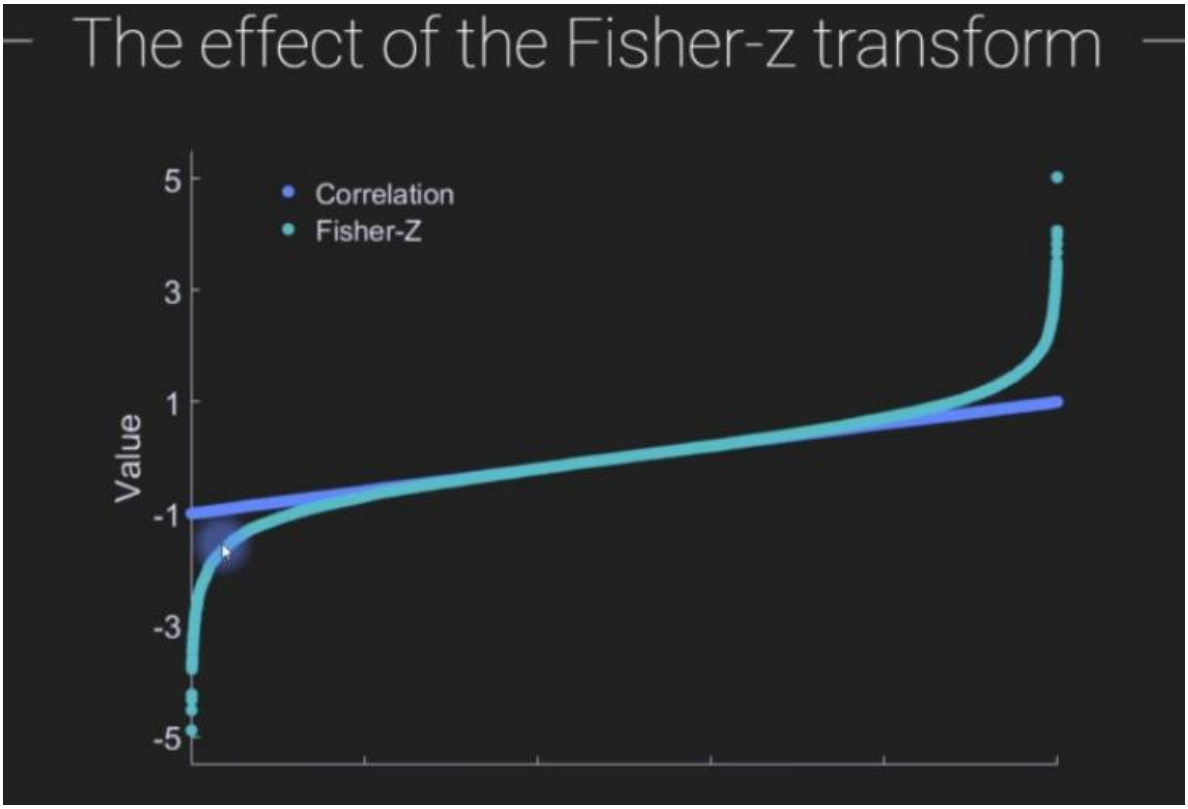
P-value: Same as for "regular" Pearson coefficient.

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

Fisher-Z Transformation for Correlations:



$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r)$$



Code: Part 5 - Spearman Correlation and Fisher-Z

Kendall's Correlation for Ordinal Data:

- Used for ordinal data (numerical meaningful order, but no fixed relationship across the levels).
- **Examples:** education, movie ratings

How it works?

- ⇨ Transform the data to rank concordances (relative signs across the values per variable).
- ⇨ "**Kendall tau-b**" has an adjustment for ties and is most often used.
- ⇨ Interpretation is identical to Pearson and Spearman.

$$\tau = K^{-1} \sum sgn(\widetilde{x_i} - \widetilde{x_{i:}}) sgn(\widetilde{x_i} - \widetilde{x_{i:}})$$

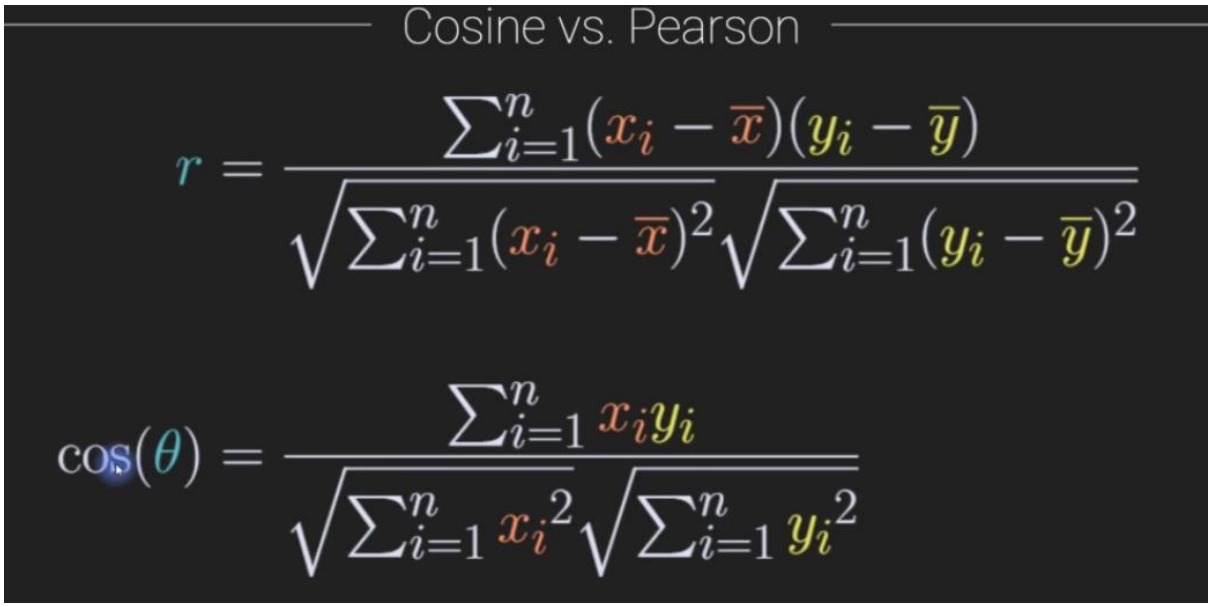
Code: Part 6 – Kendall Correlation

The Subgroups Correlation Paradox:

Have to enter the code link

Code: Part 7 - Subgroups Correlation Paradox (Simpson's Paradox)

Cosine Similarity:



Code: Part 8 - Cosine Similarity

ANOVA

ANOVA (Analysis of Variance):

- The goal of an ANOVA is to determine the effects of **discrete independent variables** (categorical variable)(groups, levels) on a **continuous dependent variable**.
- **Quick example:** Effects of medication type and age group on medication treatment.

Setting up an ANOVA in four steps:

⇨ **Step 1: Review the experiment design and make sure ANOVA is the appropriate method.**
Research goal: Test whether Covid-19 medications are effective for different groups
Experiment: Randomly assign patients to receive medication A, B, or placebo. Measure disease severity after 10days. Separate older (>50 years) from younger (<50 years) patients.

⇨ **Step 2: Identify the independent and dependent variables.**
Dependent variable (a.k.a. outcome variable): The variable you are trying to explain.
Independent variables (a.k.a. explanatory variables): The variables that you hope will explain the DV (dependent variable).
IVs: Medication and age group
DVs: Disease severity after 10 days

⇨ **Step 3: Create a table of factors and levels (when possible).**
Factors: The "dimensions" of IVs
Levels: The specific groups or manipulations within each factor.

Factors: Medication, age.
Levels: A, B, placebo (medication); younger, older (age)

		Medication		
		A	**B**	**placebo**
Age group	**Younger**			
	Older			

⇨ **Step 4: Compute the model and interpret the results.**
Main effect: One factor influences the DV even when ignoring all other factors.
Interactions: The effect of one factor depends on the levels of another factor.
Intercept: The average DV is different from zero. The intercept of the ANOVA is usually ignored.

Main effect: young people's symptoms improve faster than older people's symptoms, regardless of medication type.
Interaction: Medication A works better in older people; medication B works better in younger people.
Intercept: Symptoms improve for almost everyone after 10 days.

Examples of when ANOVAs are inappropriate:

⇨ **Research goal:** Test whether people with more Facebook friends have higher self-reported extraversion.
Variables: DV: number of Facebook friends. IV: scores on a personality questionnaire.
ANOVA? No categorical factors.
T-Test? No group(s) to compare.
Correlation? Yes, because we are looking for a linear relationship between two continuous variables.

⇨ **Research goal:** Test whether RSI (repetitive stress injury) is decreased for a group of meditators vs non-meditators.
Variables: IV: group (meditate or not) DV: scores on an RSI index
ANOVA? Only one factor with two levels
T-Test? Yes, because there are two groups to compare.
Correlation? No, because the IVs are discrete groups, not continuous variables.

The way of the ANOVA:

- **<number>-way:** The number of factors.
  Examples:
  - **One-way ANOVA:** Determine the influence of day-of-week on iPhone purchases (7 level 1 factor).
  - **Two-way ANOVA:** Determine the influences of day-of-week and gender (male, female) on iPhone purchases.

Repeated-measures ANOVA:

- **rmANOVA:** If at least one factor involves multiple measurements from the same individual.
  Example:
  - **Research question:** Determine effects of snack type on mood.
  - **Experiment:** Volunteers eat chocolate for 2 days, potato chips for 2 days, and ice cream for 2 days (order randomized).

Balanced vs Unbalanced ANOVA:

- **Balanced:** The same number of data points in each cell.
- **Unbalanced:** Different number of data points across cells. This could happen because of data collection or cleaning.

Balanced ANOVA				
		Medication		
		A	**B**	**placebo**
Age group	**Younger**	20	20	20
	Older	20	20	20
Unbalanced ANOVA				
		Medication		
		A	**B**	**placebo**
Age group	**Younger**	20	23	21
	Older	18	20	20

Dummy-Coding Variables:

- **Dummy-Coding Variables:** Converting categorical variables into numbers. Best applied to two cases with numbers {0, 1}.
- **Lingo:** Gender was entered into the ANOVA as a dummy-coded variable (male=0, female=1).
- **Interpretation:** The effect of the factor shows the change for the "1" variable compared to the "0" variable ("0" acts like a baseline).
- **Example:** Main effect of gender on lipstick use indicates that women use lipstick more than men (cannot claim

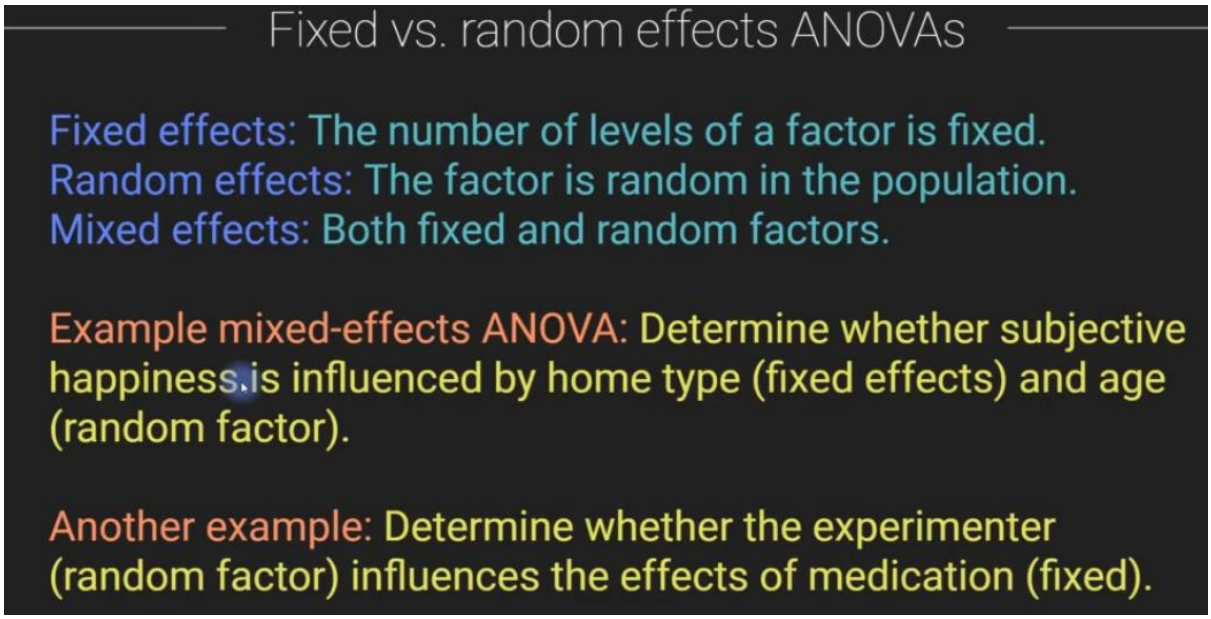
anything about men using lipstick independently of women).

ANOVA vs MANOVA:

- **ANOVA:** Only one DV (as many IVs as appropriate).
- **MANOVA (multivariate ANOVA):** Multiple DVs (as many IVs as appropriate)
- **Example:** Effects of medication type and age on Covid-19 symptoms and total medical expenses.

Fixed vs Random Effects ANOVAs:

- **Fixed effects:** The number of levels of a factor is fixed (e.g., home type: dorm room, apartment, house).
- **Random effects:** The levels of a factor are random (continuous) in the population (e.g., age, salary, nurse)
- **Mixed effects:** Some factors are fixed; others are random.



Assumptions of ANOVA:

- **Independence:** The data are sampled independently of each other in the population to which you want to generalize.
- **Normality:** The "residuals" (unexplained variance after fitting the model) are normally (Gaussian) distributed.
- **Homogeneity of Variance (a.k.a. heteroscedasticity):** Variance within each cell is (roughly) the same.

Nonparametric ANOVA Alternatives:

- **Kruskal-Wallis test (KW-ANOVA):** Alternative to the one-way ANOVA on rank-transformed data.
- **Note:** ANOVAs are generally robust to violations of the assumptions. Nonparametric ANOVAs are rarely used and may cause more of a headache than they are worth.

Sum of Squares:

The null hypothesis is that the means of all groups (all cells in the ANOVA table) are statistically indistinguishable.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

The alternative hypothesis is that the mean of at least one group is different from the mean of at least one other group.

$$H_A : \mu_i \neq \mu_j$$

Computing and Interpreting Sum of Squares:

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

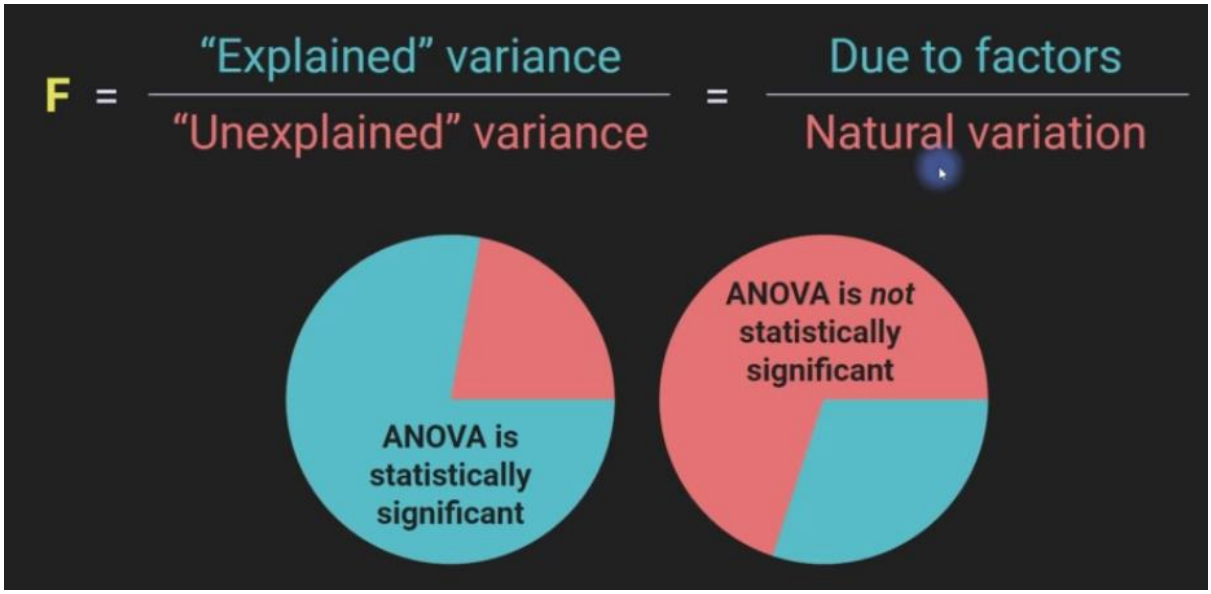
The formula is almost same with variance formula.

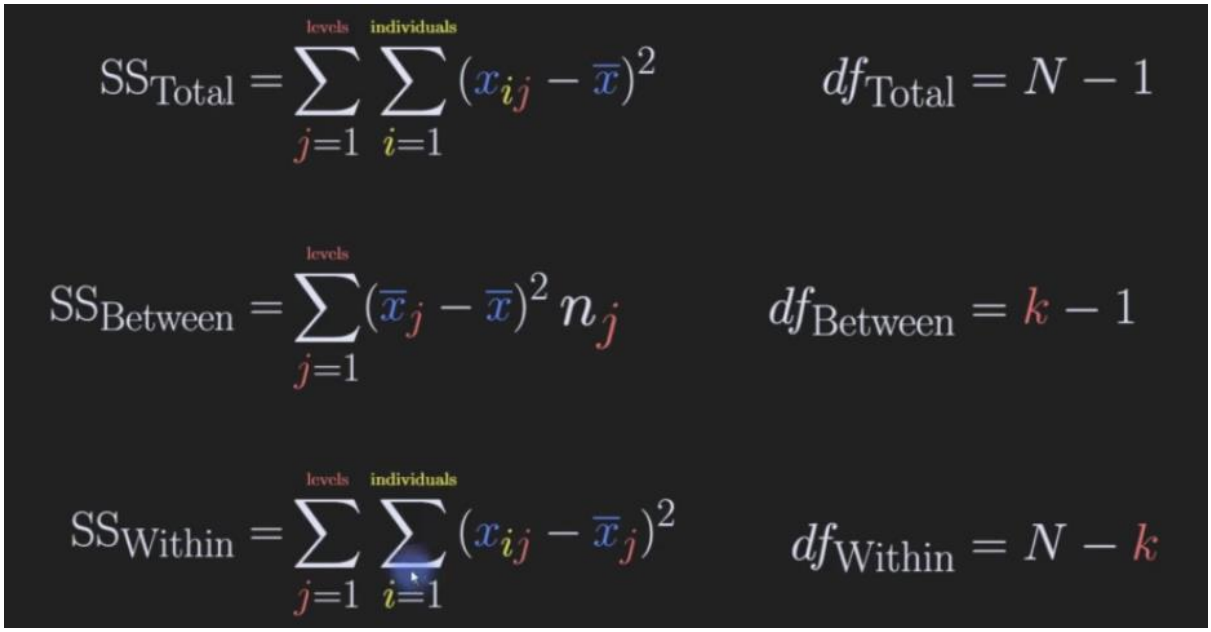
ANOVA as a partition of sum of squares:

The total variation in the dataset is the sum of the variation across individuals within each group and the variance across the different levels.

Total SS = Within-group SS + Between-group SS

Total SS = Error SS + Between-group SS





From the formula 2, n_j is the number of individuals within each level and *k* is the number of levels.

Thus far, the explanations are ***for one-way ANOVA***. This is for simplicity.

The F-Test and ANOVA Table:

Sum of squares to mean square

$$\mathrm{MS}_{\mathrm{Between}} = \frac{\mathrm{SS}_{\mathrm{Between}}}{df_{\mathrm{Between}}} = \frac{\sum_{j=1}^{\text{levels}} (\bar{x}_j - \bar{x})^2 n_j}{k-1}$$

$$\mathrm{MS}_{\mathrm{Within}} = \frac{\mathrm{SS}_{\mathrm{Within}}}{df_{\mathrm{Within}}} = \frac{\sum_{j=1}^{\text{levels}} \sum_{i=1}^{\text{individuals}} (x_{ij} - \bar{x}_j)^2}{N-k}$$

The F-test

$$F_{k-1,N-k} = \frac{\mathrm{MS}_{\mathrm{Between}}}{\mathrm{MS}_{\mathrm{Within}}}$$

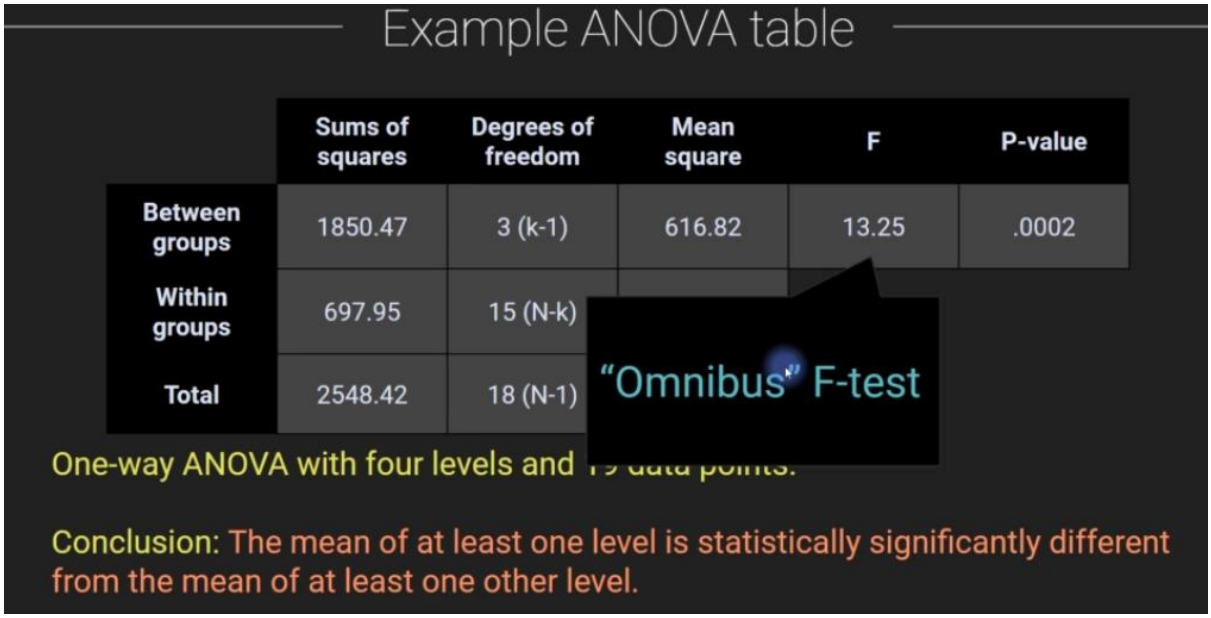
The ANOVA table

Source of variance	Sums of squares	Degrees of freedom	Mean square	F	P-value
Between groups	SS_B	k-1	MS_B	MS_B/MS_W	p
Within groups	SS_W	N-k	MS_W		
Total	SS_T	N-1			

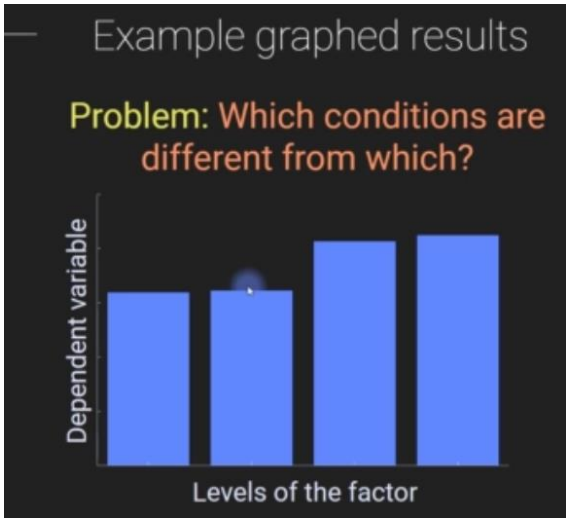
Correct interpretation of p<.05: At least one level (group) is statistically significantly different from at least one other level. Determining which groups differ requires data visualization follow-up t-tests.

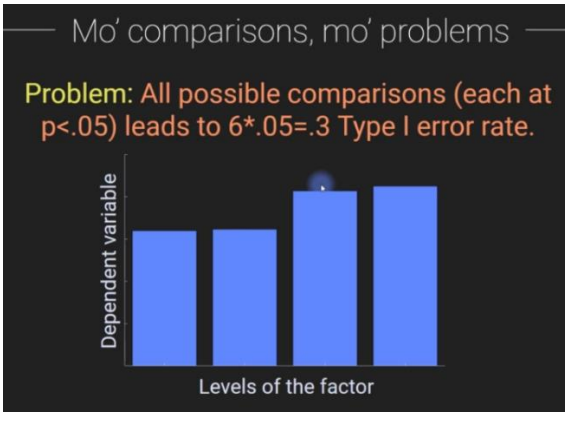
The Omnibus F-test and Post-hoc Comparisons:

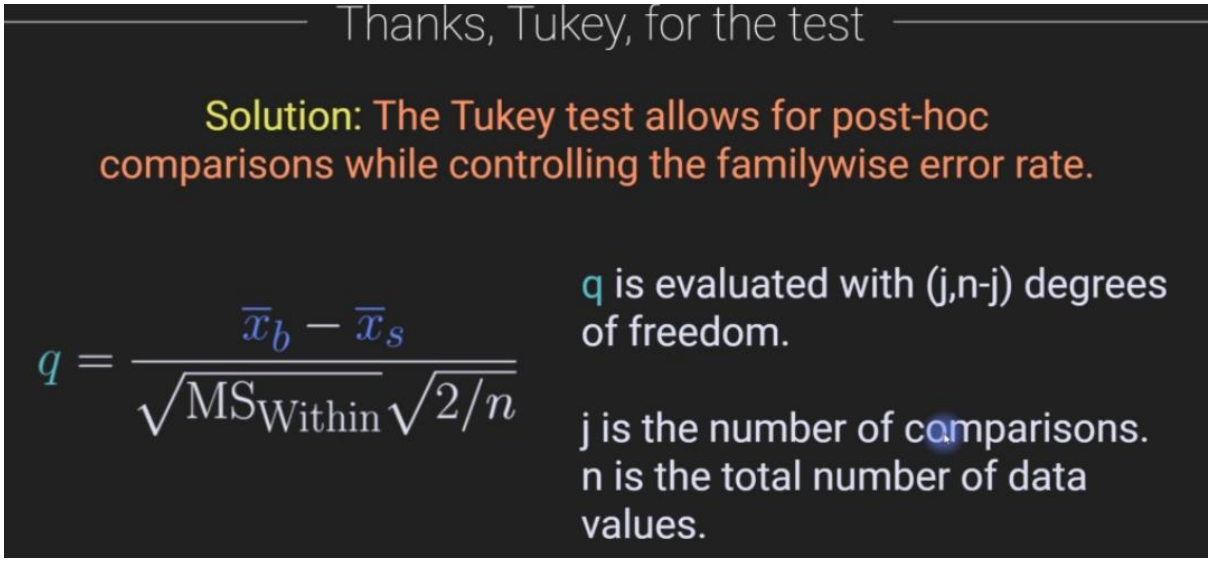
The p value does not tell which level (group) is different from which level (group). And that is the problem as well as the feature of ANOVA. The ANOVA tells that something is different from something else.



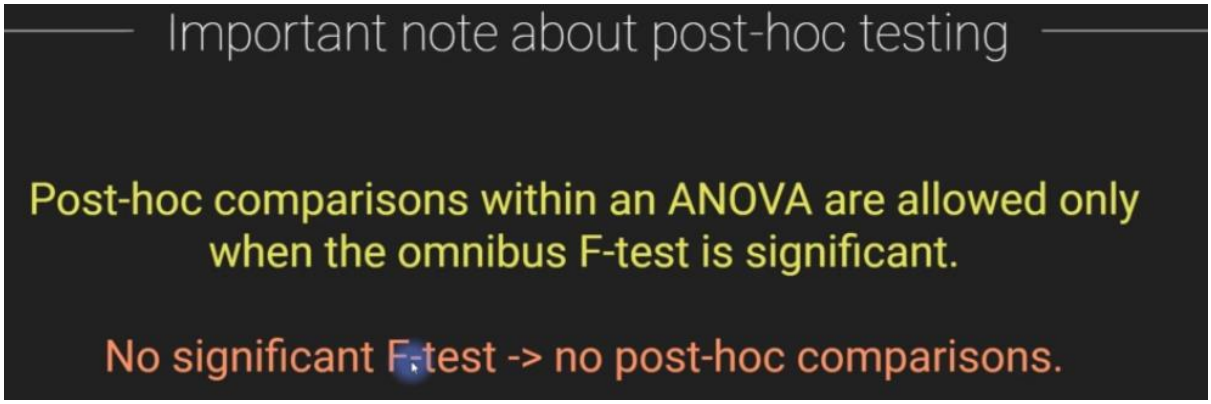
(Within groups, Mean square) = 46.53 (the value is hidden to the above table). The table came from One-Way ANOVA with four levels and 19 data points.







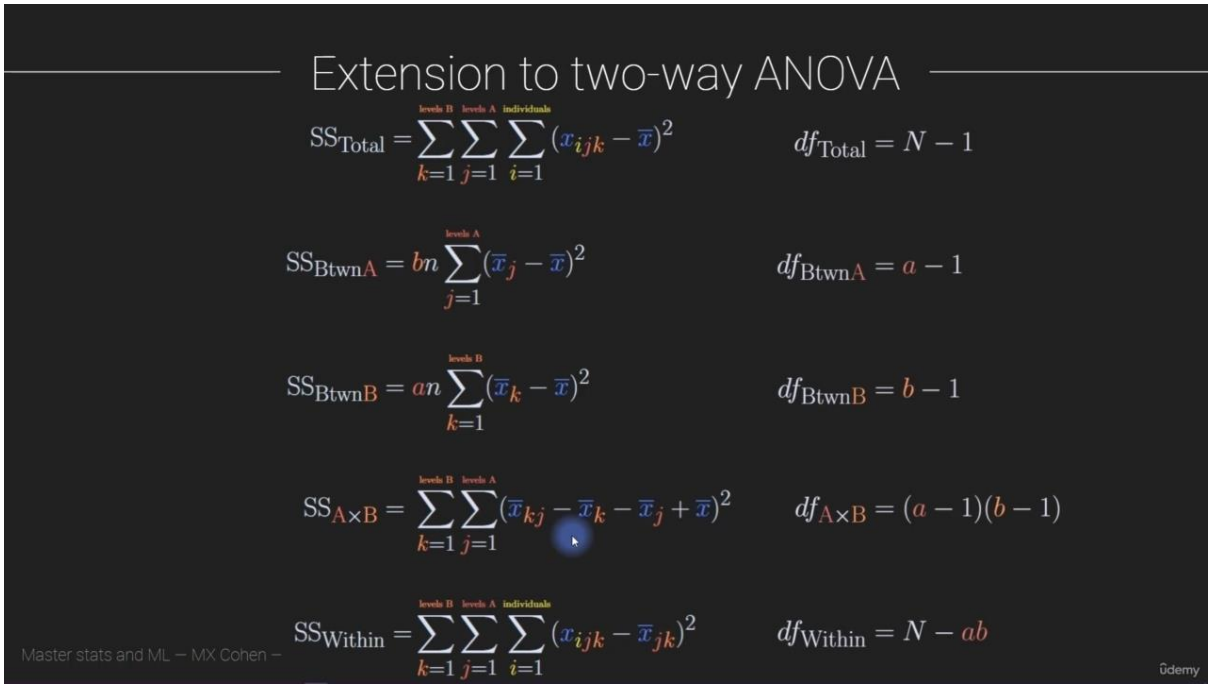
x_b and x_s is the two conditions. One could be one level (group) and another one could be another level (group) with their mean

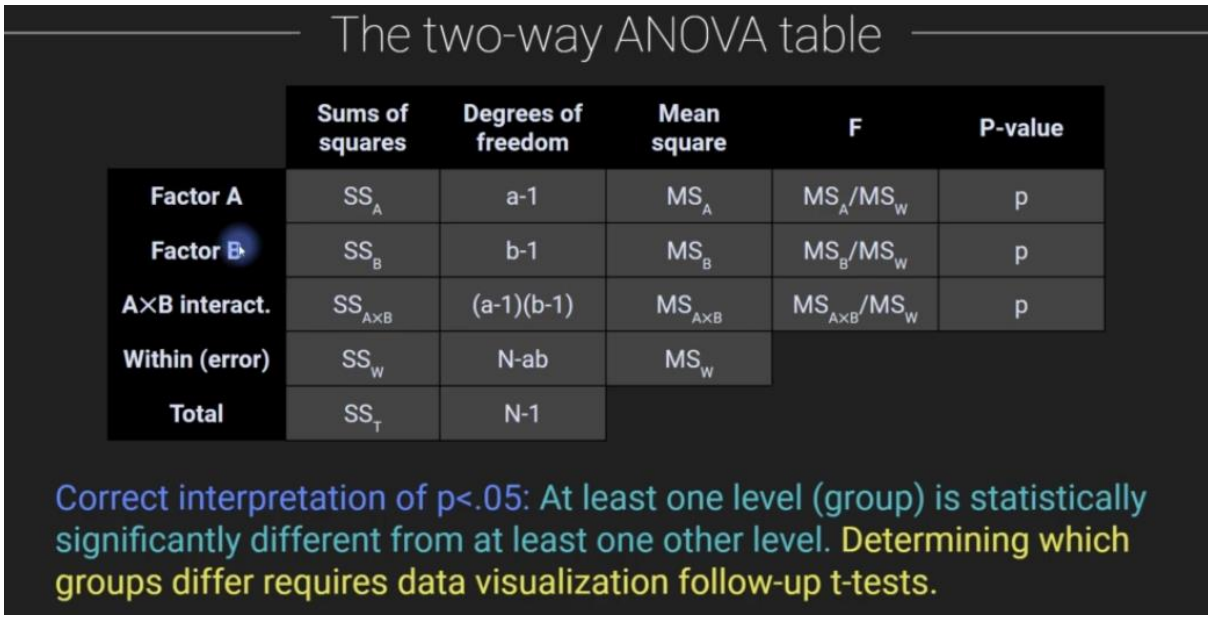


The Two-Way ANOVA:

Partition of Sum of Squares in 2-way ANOVA:

The total variation in the dataset is the sum of the variation across individuals within each group and the variation across the different levels *within each factor and the variation at the interaction between the factors.*

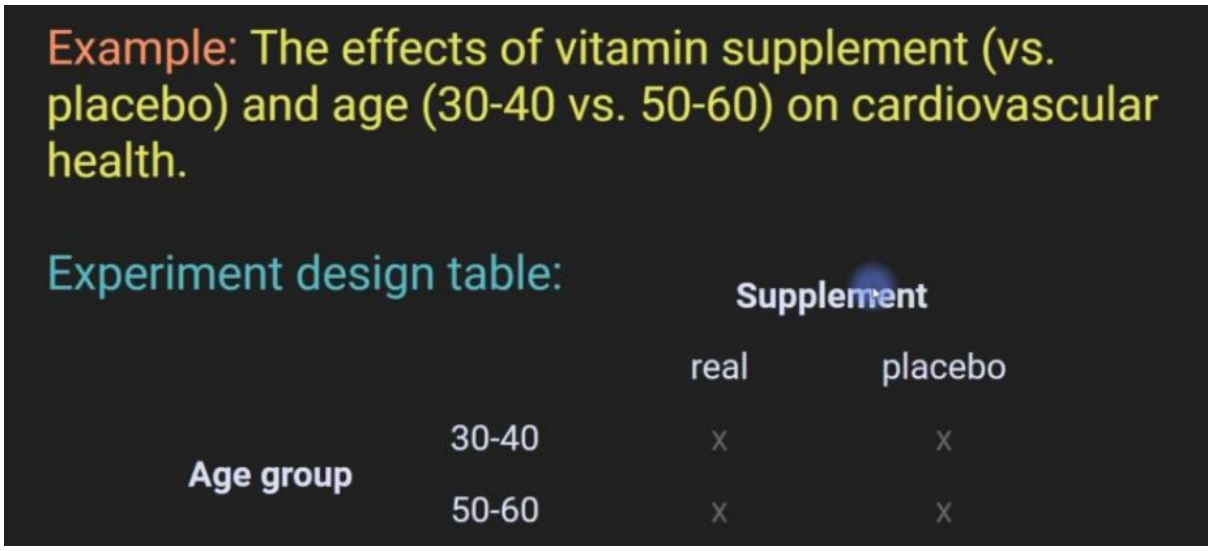


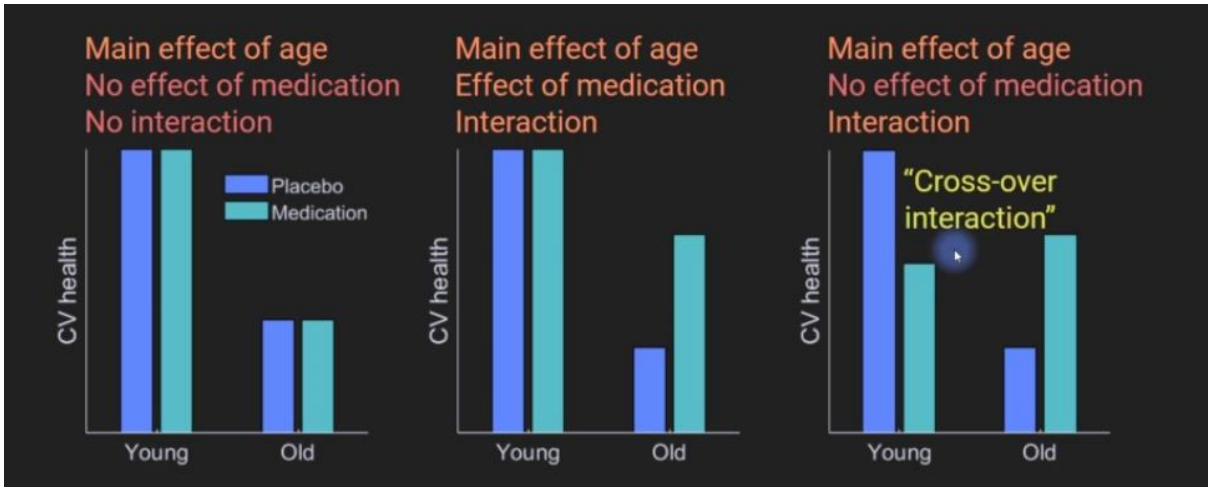


	Sums of squares	Degrees of freedom	Mean square	F	P-value
Factor A	SS_A	a-1	MS_A	MS_A/MS_W	p
Factor B	SS_B	b-1	MS_B	MS_B/MS_W	p
A×B interact.	$SS_{A\times B}$	(a-1)(b-1)	$MS_{A\times B}$	$MS_{A\times B}/MS_W$	p
Within (error)	SS_W	N-ab	MS_W		
Total	SS_T	N-1			

How to interpret the main effects and interactions

- A significant main effect means that the factor influences the data independently of the other factor(s).
- A significant interaction means that the effect of one factor depends on another factor.





Important take-home message 1: Main effects must be cautiously interpreted with a significant interaction.

Important take-home message 2: Data must always be visualized for proper interpretation!

One-Way ANOVA Example:

Research goal: Test self-reported happiness after watching different genres of movies.

Variables: IV: movie genre (horror, romcom, documentary, sci-fi).
DV: happiness rating (1-100).

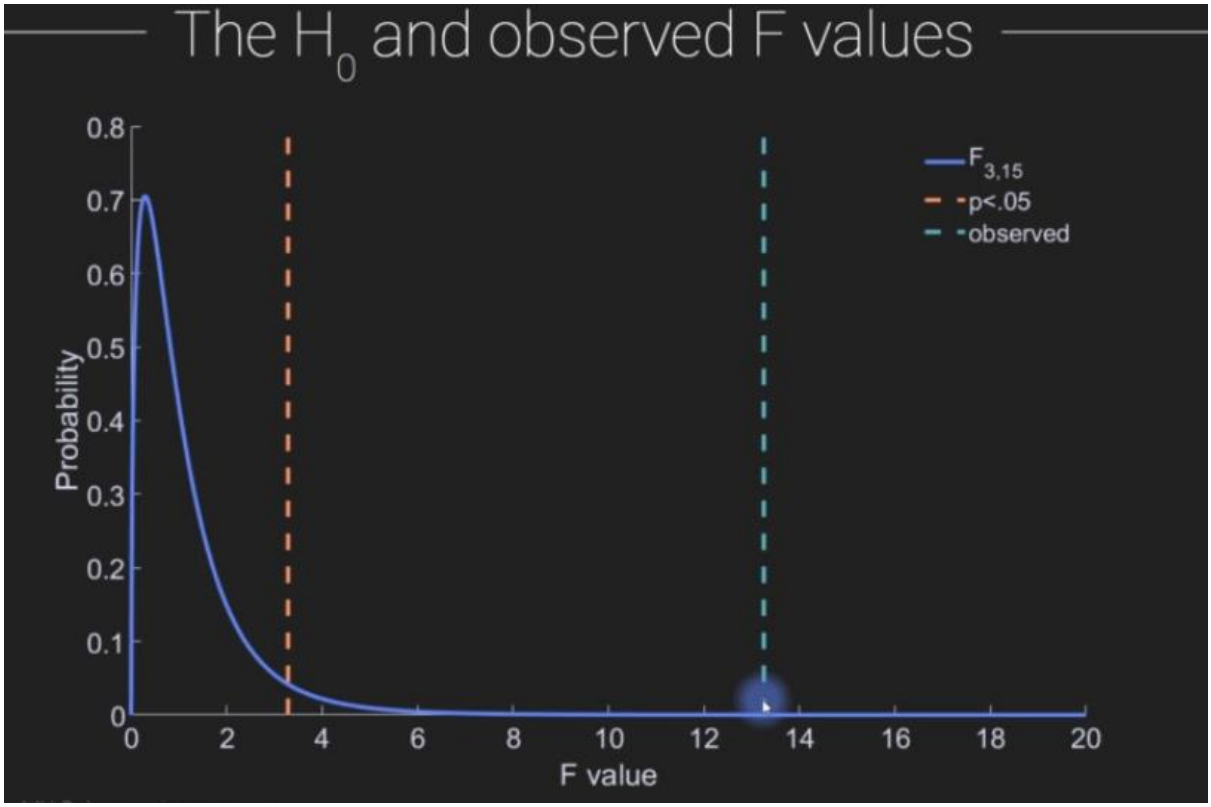
	Movie genre			
	Horror	Romcom	Docu	Sci-fi
N per group	5	5	4	5

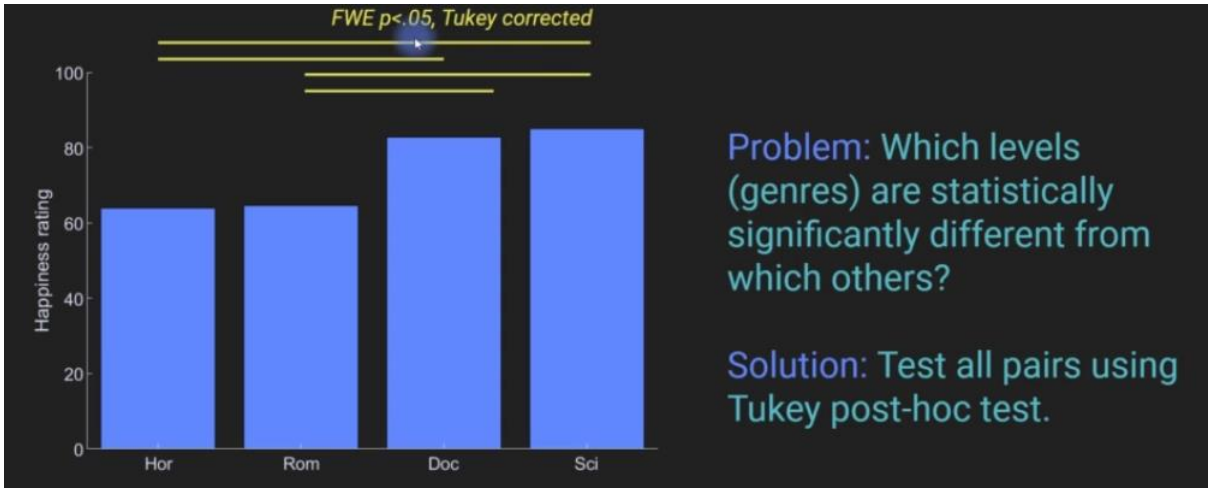
	Movie genre			
	Horror	Romcom	Docum.	Sci-fi
Happiness rating	61	68	82	90
	57	78	84	87
	70	65	90	75
	65	57	75	88
	67	55		85

	Sums of squares	Degrees of freedom	Mean square	F	P-value
Between groups	1850.47	3 (k-1)	616.82	13.25	.0002
Within groups	697.95	15 (N-k)	46.53		
Total	2548.42	18 (N-1)			

Conclusion: The mean of at least one level is statistically significantly different from the mean of at least one other level.

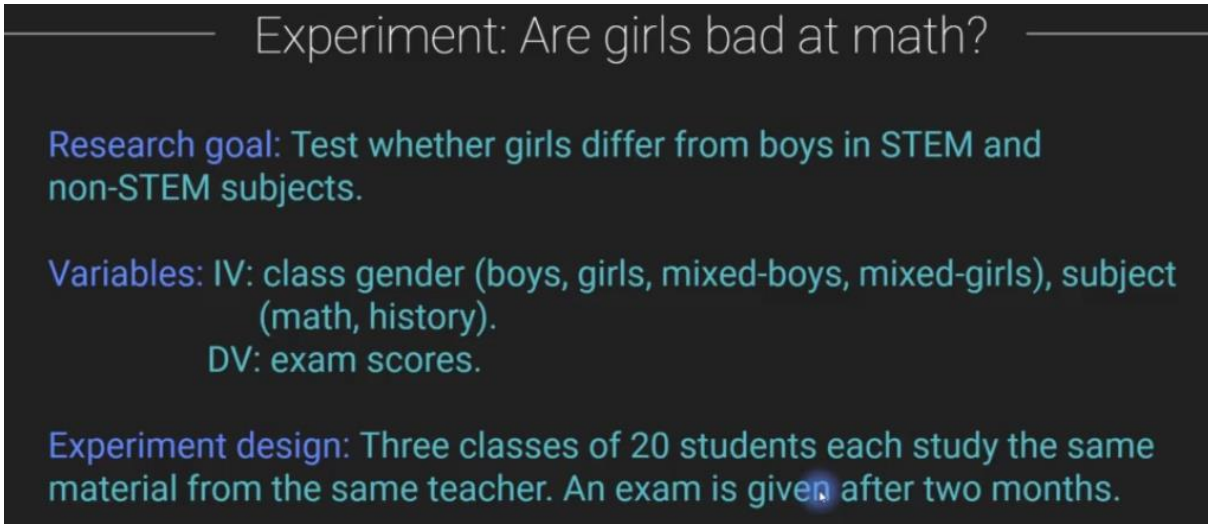
People were happier after at least one genre compared to at least one other genre.

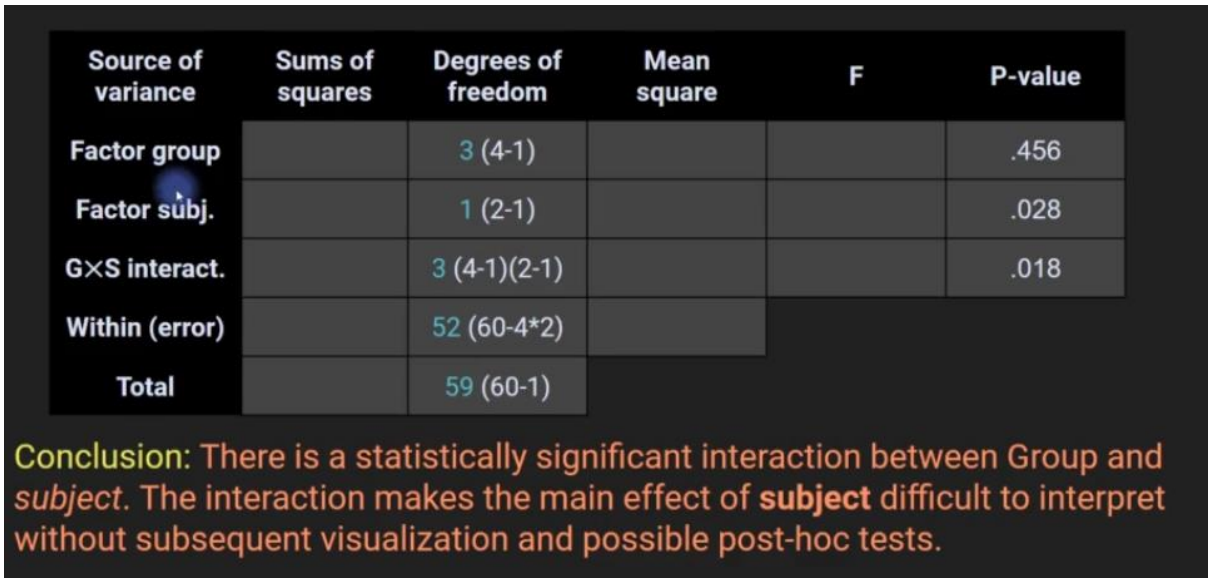




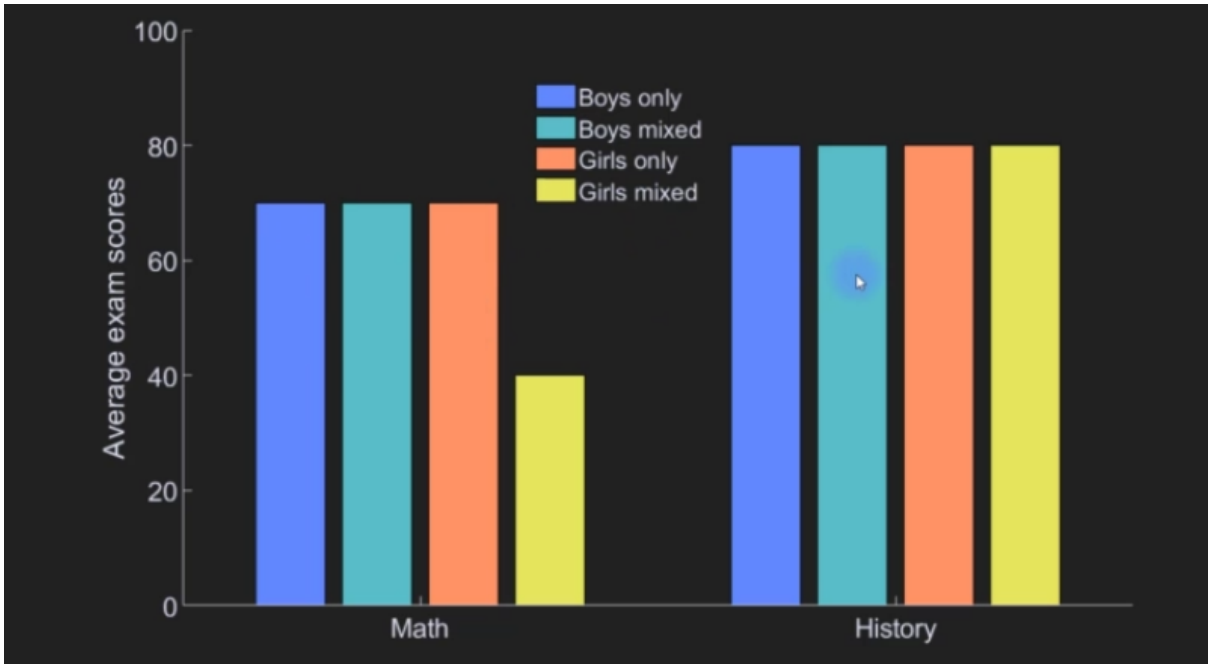
Code: Part 1 - One-way ANOVA (Independent Samples) & Part 2 - One-Way Repeated-Measures ANOVA

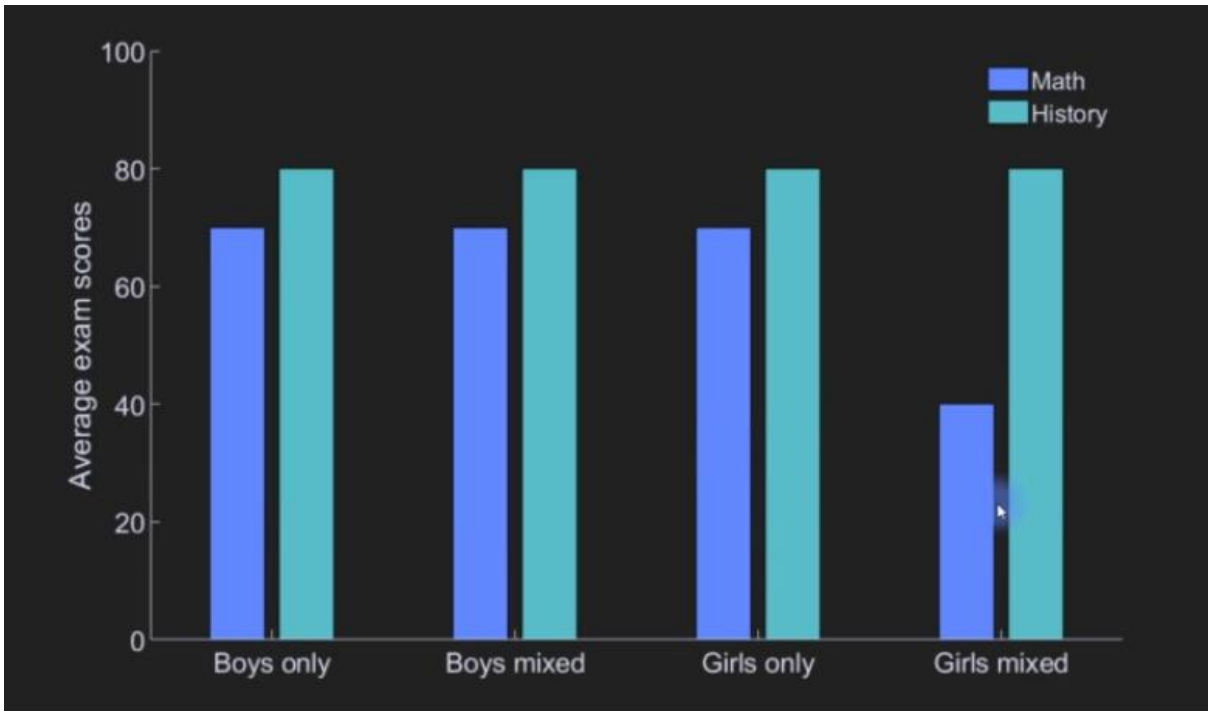
Two-Way ANOVA Example:





For a mixed-effects ANOVA, the Total df is (N-b)(a-1) where a is the within-subjects factor.





Code: Part 3 - Two-Way Mixed-Effects ANOVA

Regression

Introduction to GLM/Regression:

ANOVA vs Regression:

- Use an ANOVA when all IVs are discrete (usually categorical).
- Use a regression when at least some IVs are continuous.

The five steps to model-fitting:

- **Step 1: Define the equation(s) underlying the model.**

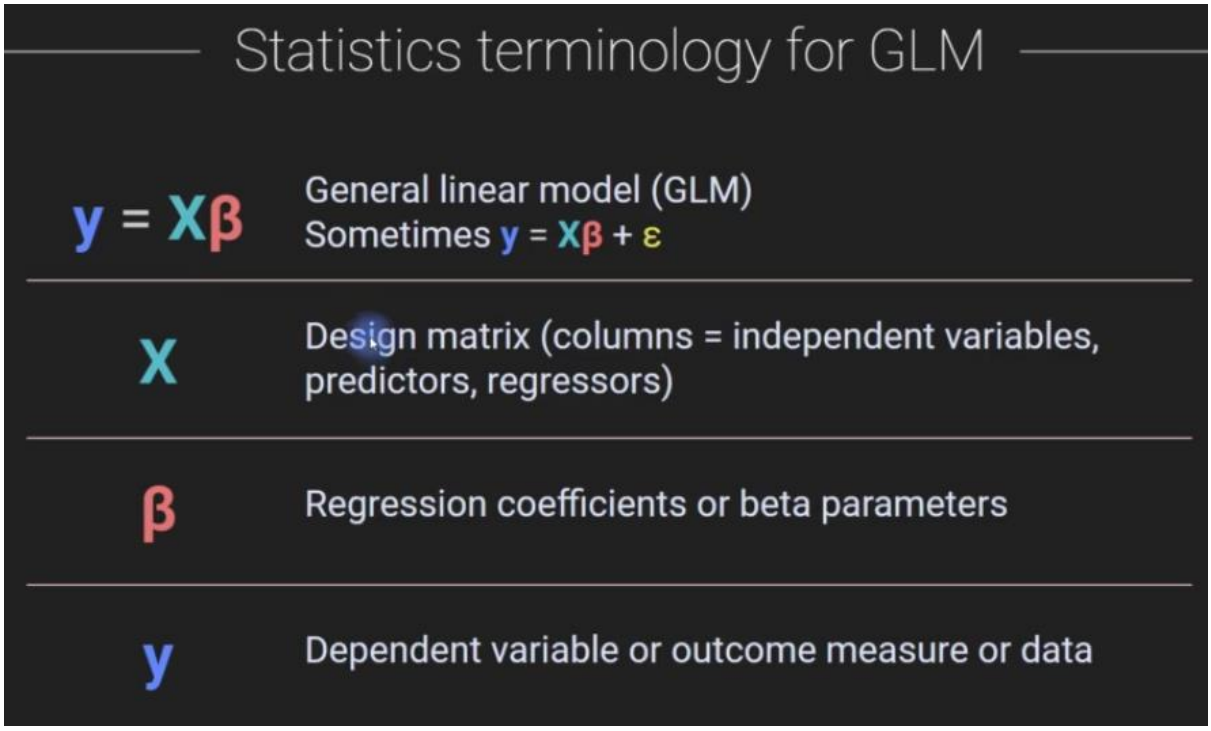
$$h = \beta_0 + \beta_1 s + \beta_2 p + \beta_3 n$$

Where,
h = height
β_0 = intercept (average height when all other parameters are 0)
s = sex (m/f), dummy-coded
p = parents' height
n = childhood nutrition

- **Step 2: Map the data into the model equations.**
- **Step 3: Convert the equations into a matrix-vector equation.**
- **Step 4: Compute the parameters.**
- **Step 5: Statistical evaluation of the model.**
  - **Two question:**
    - **Is the model a good fit to the data?**
    - **Are individual β coefficients statistically significant?**

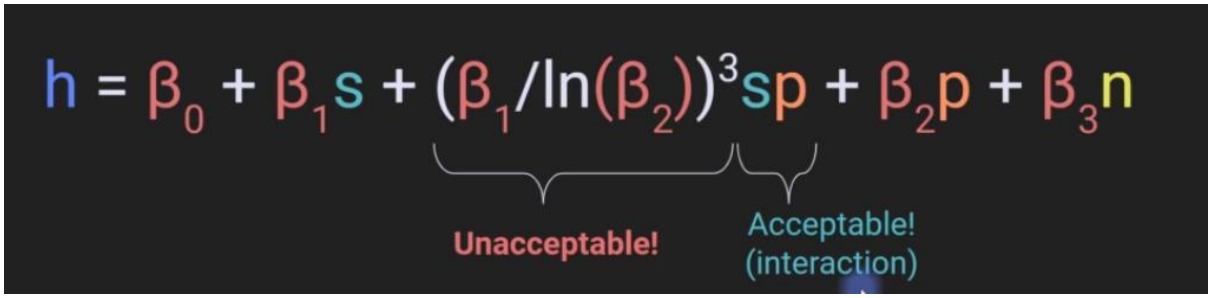
The Intercept:

The intercept term is usually not interpreted and usually not very interesting. But it needs to be in the model because it captures the offset of the data when all other predictors are set to 0.

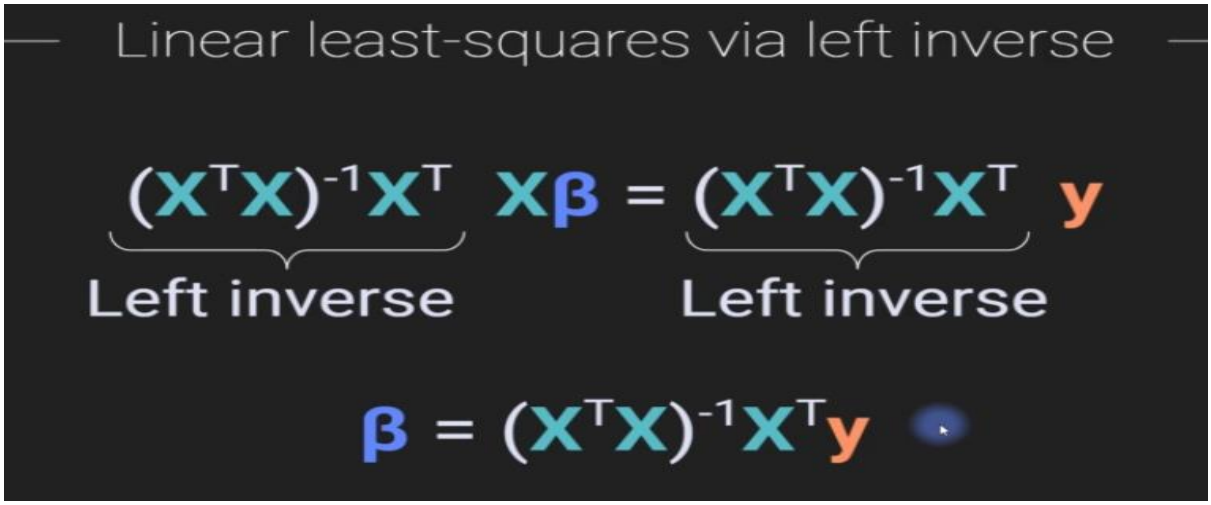


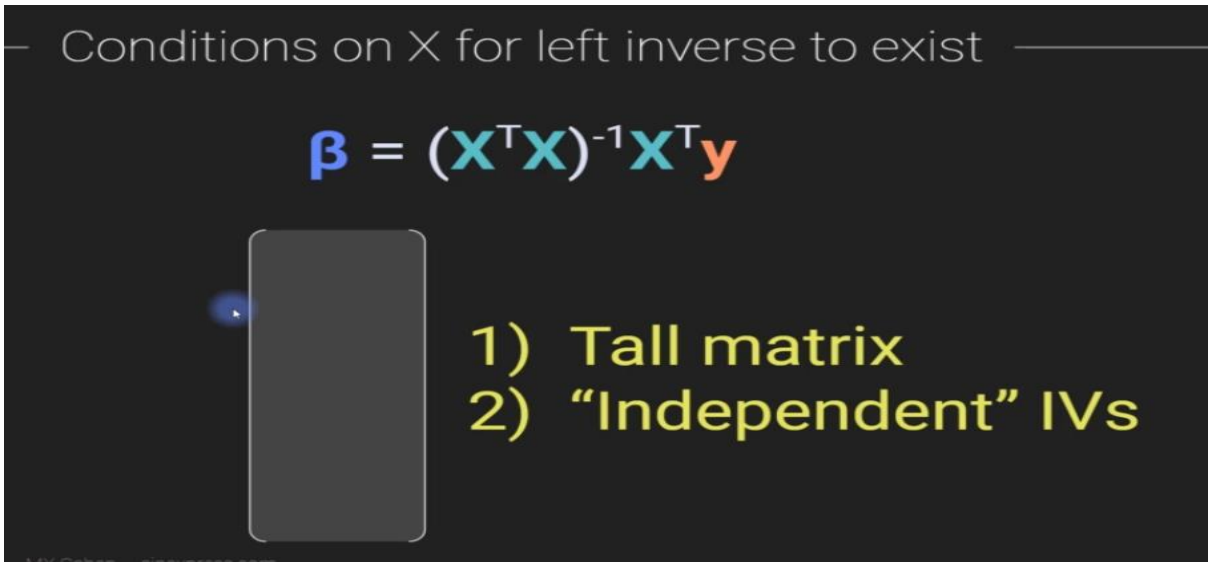
General LINEAR model (GLM):

- GLMs (ANOVA, regression, correlation) are linear models.
- Linear means scalar-multiplication and addition of the regressions.
- Log, square root, powers, trigonometry etc are all nonlinear operations.
- The data can have nonlinearities; only the parameters must be linear in the model.
- Nonlinear data are often linearized to facilitate interpretation.

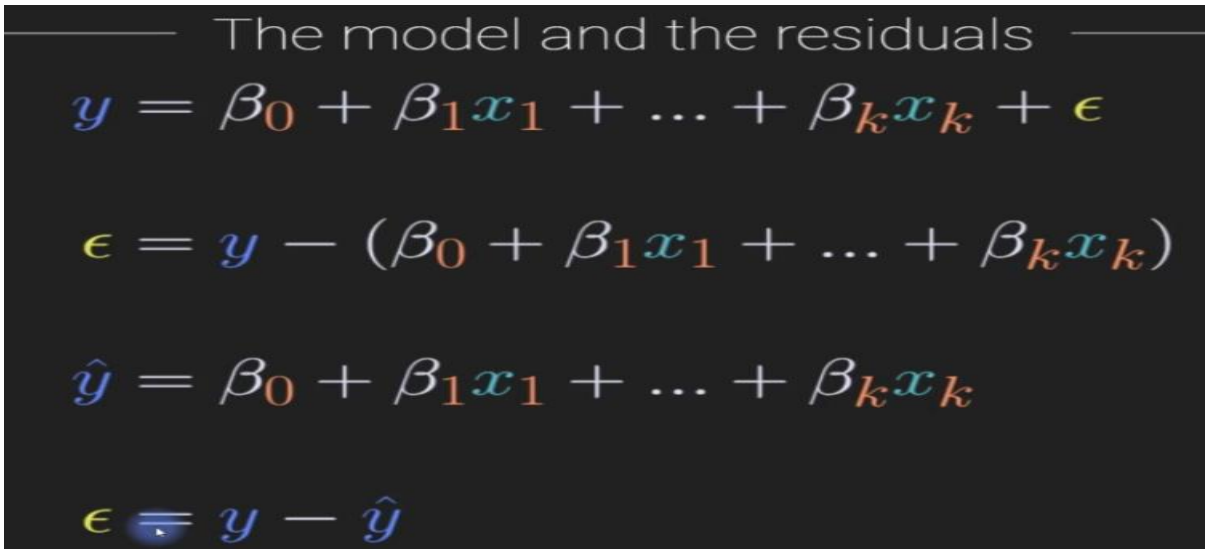


Least Squares Solution to the GLM:





Evaluating Regression Models: R^2 and F:



Evaluating a model fit with R^2

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

R^2 close to 1 means the model fits the data well.
R^2 close to 0 means the model fits the data poorly.
R^2 below 0 is most likely due to an error.

There is no cut-off for a "good" R^2!
R^2 is often used in model comparisons.

Evaluating model statistical significance with F

H_0: $\beta_{1\text{-}k} = 0$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

H_A: At least one $\beta \neq 0$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

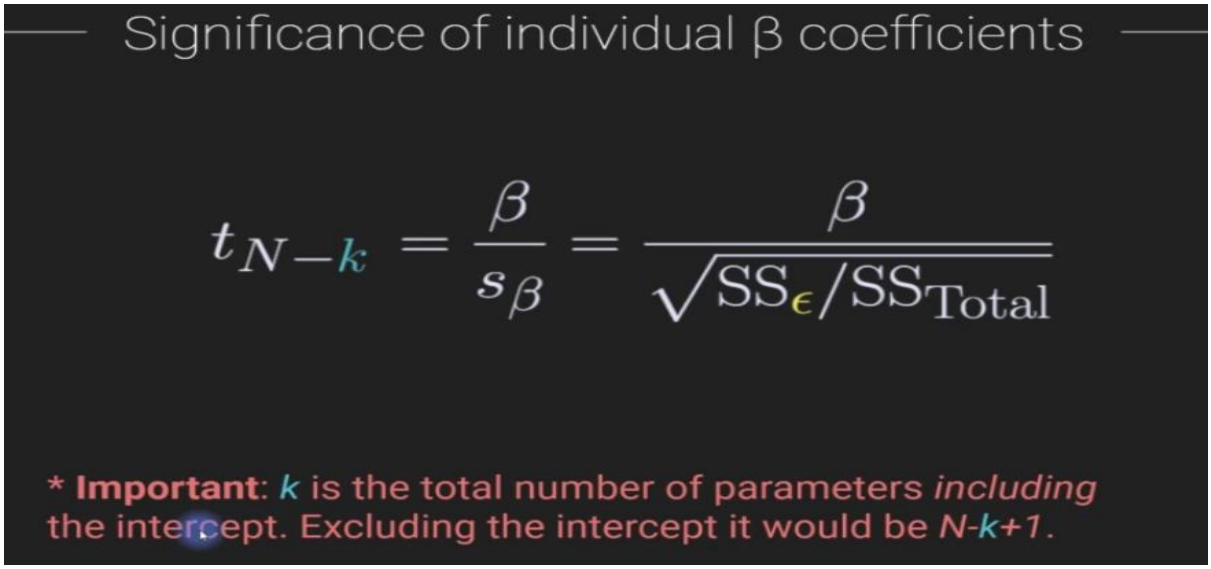
Evaluating model statistical significance with F

$$SS_\epsilon = \sum(y_i - \hat{y})^2$$

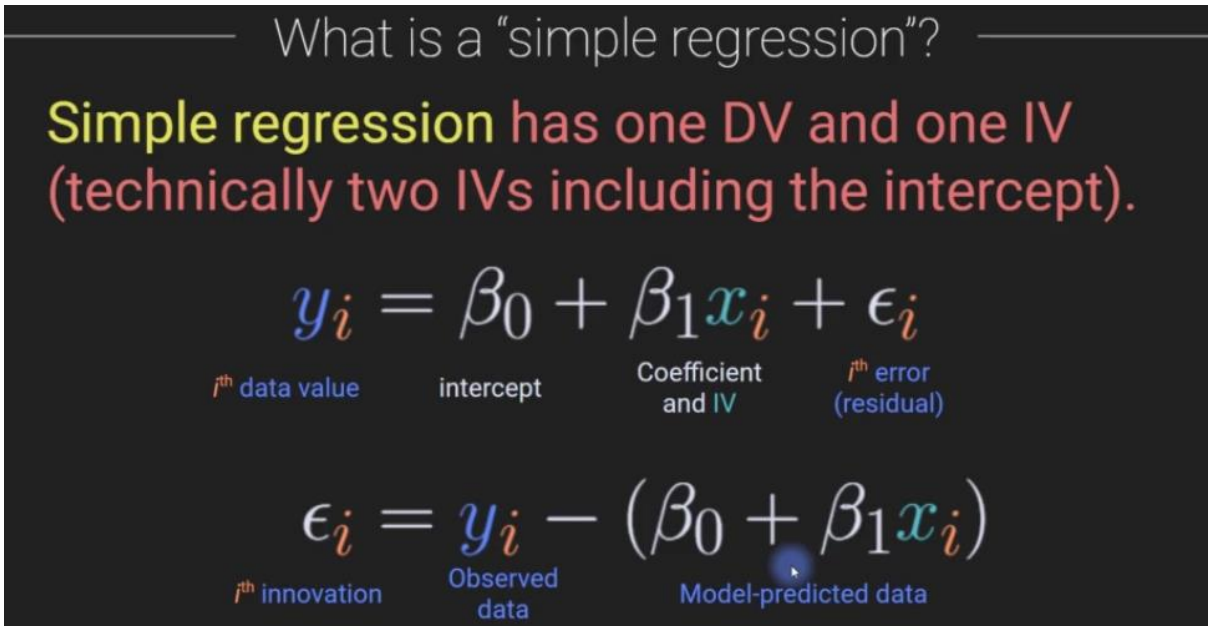
$$SS_{Model} = \sum(\hat{y}_i - \bar{y})^2$$

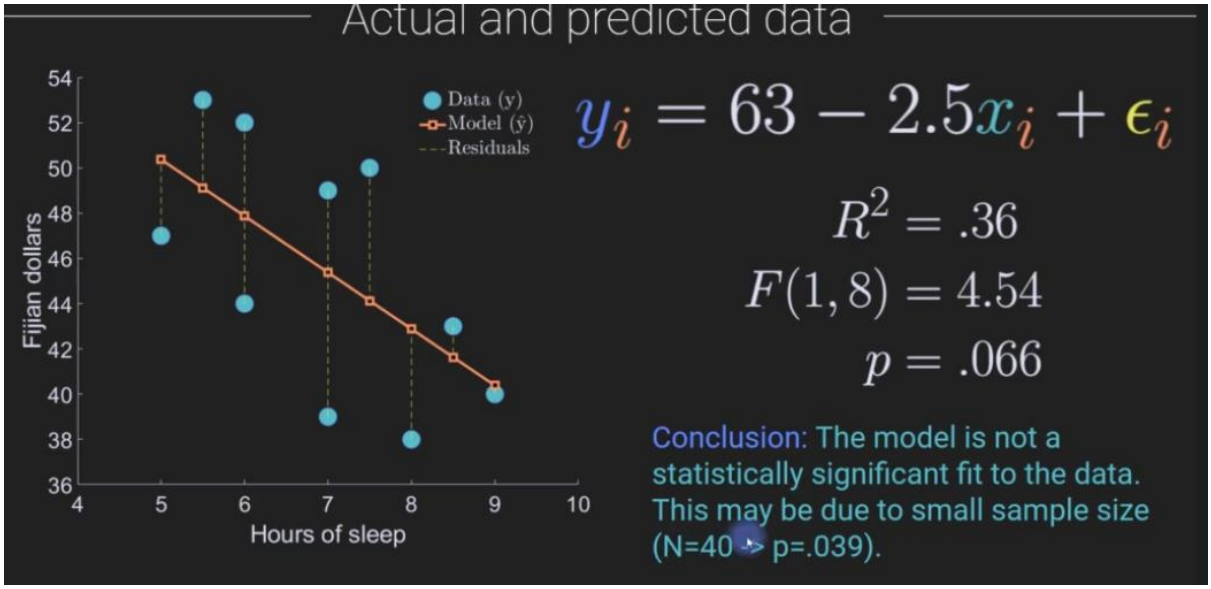
$$F(k-1, N-k) = \frac{SS_{Model}/(k-1)}{SS_\epsilon/(N-k)}$$

Where, k = total number of parameters in the model including intercept.



Simple Regression:





Multiple Regression:

Research question: Do students' grade on a stats exam depend on sleep (s), hours studied (h), and the interaction between them? (N = 30)

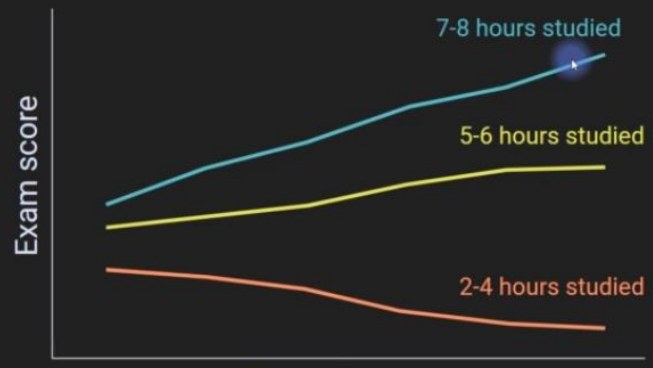
$$y = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 (s.h) + \varepsilon$$

Source of variance	β coeff.	Std. error	t-value (df = N-k = 30-4 = 26)	p-value
Constant	60	0.3	200	.000
Study hours	4.5	1.2	3.75	0.001
Sleep hours	0.8	2.4	0.33	0.373
Study x sleep	5.4	0.8	6.75	0.000

Visualizing the results

Conclusions

- Main effect of study (more study -> higher score)
- No main effect of hours slept.
- Significant interaction (more study is better with enough sleep).



Higher-order interactions?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (x_1 \times x_2) + \beta_5 (x_1 \times x_3) + \beta_6 (x_2 \times x_3) + \beta_7 (x_1 \times x_2 \times x_3) + \epsilon$$

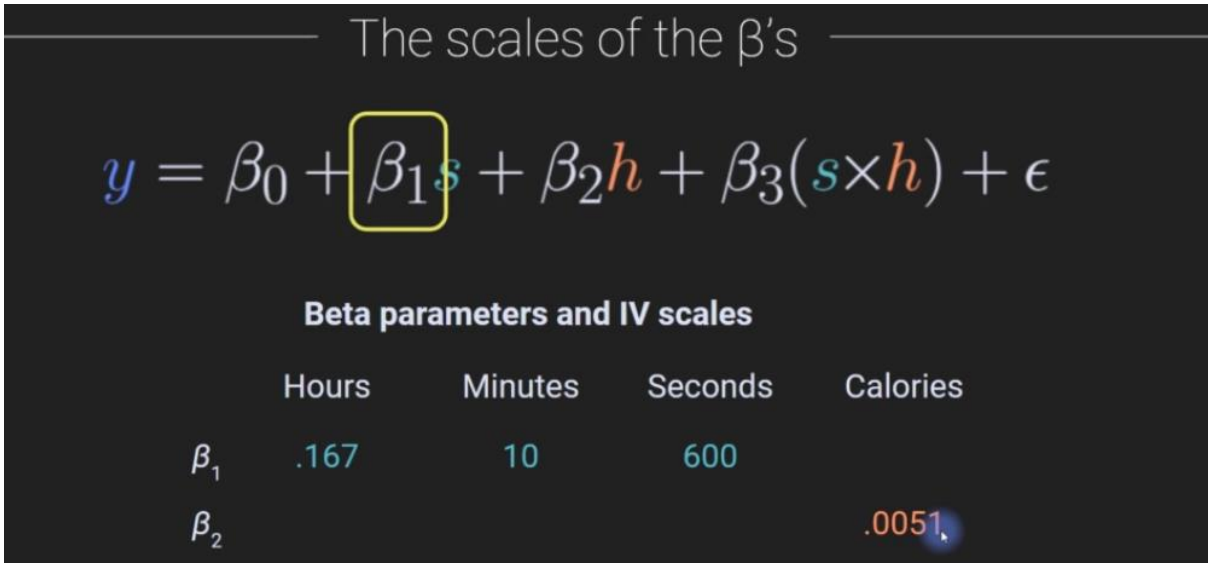
Conclusion: Higher-order interaction terms are mathematically fine, but become difficult to interpret (easier if at least one variable is binary). Use only when necessary!

Precise interpretation of β in a multiple regression

$$y = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 (s \times h) + \epsilon$$

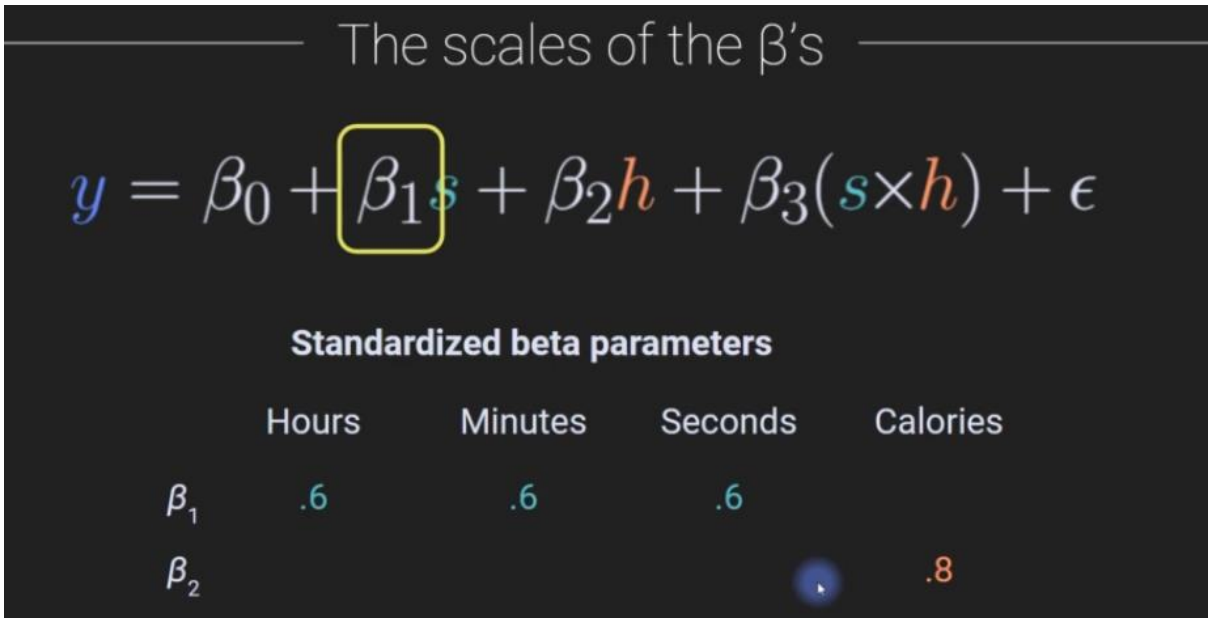
Interpretation: β_2 reflects the effect of a change in *h* on *y* when all other variables are held constant.

Standardizing Regression Coefficients:



Difficulties with interpreting β's:

- Unstandardized β coefficients change depending on the scale of the IV.
- Unstandardized β coefficients can be difficult (or impossible) to compare across variables (and studies…).
- These difficulties motivate a standardization of β coefficients.



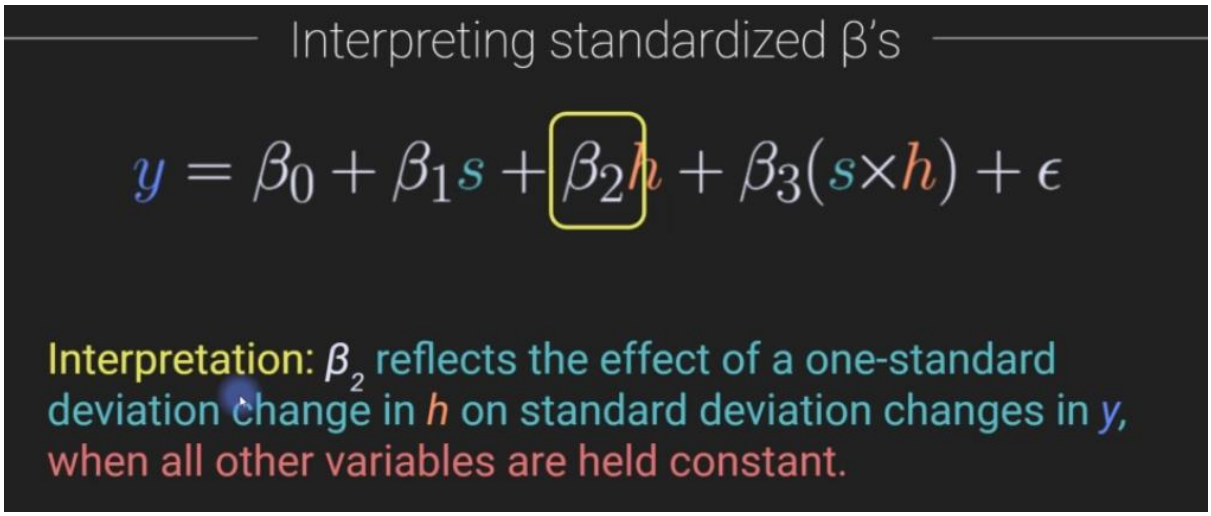
- Unstandardized β coefficients reflect the scales of the data (IV and DV). This can facilitate interpretation but can also stymie comparisons across variables or models.

- Standardized β coefficients are in standard deviation units, unrelated to the scales of the data.
- Both are correct and neither is better; sometimes one is more natural or easier to interpret than the other.
- Importantly, standardization has no effect on the statistics!

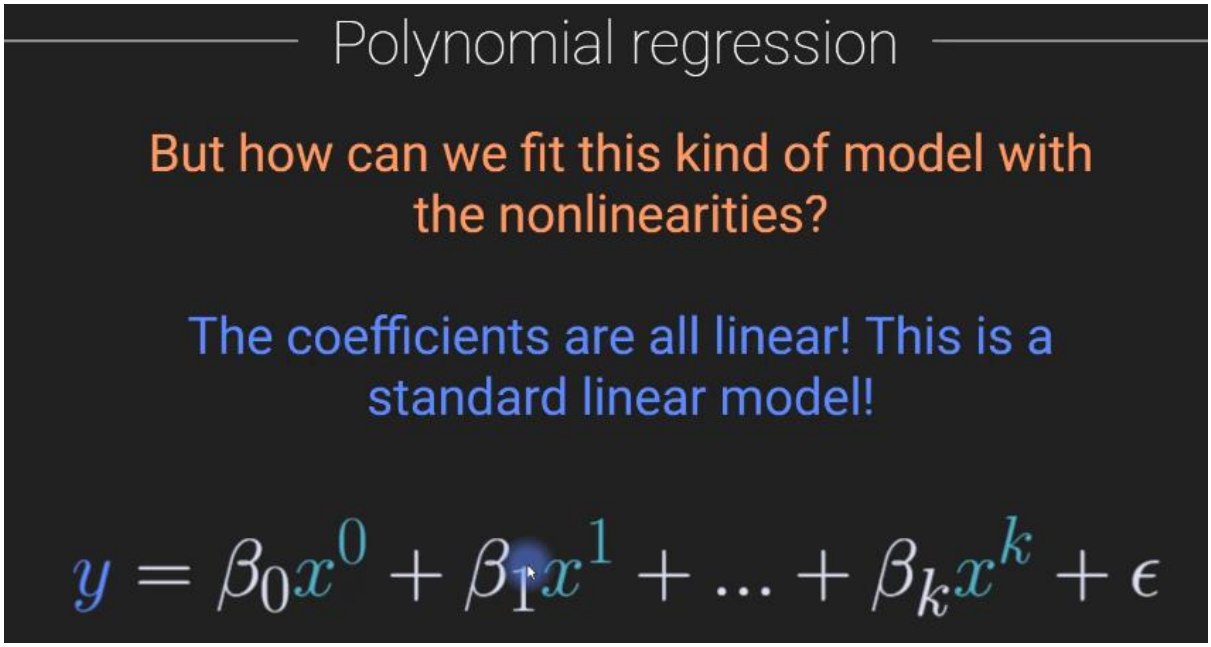
How to standardize regressors:

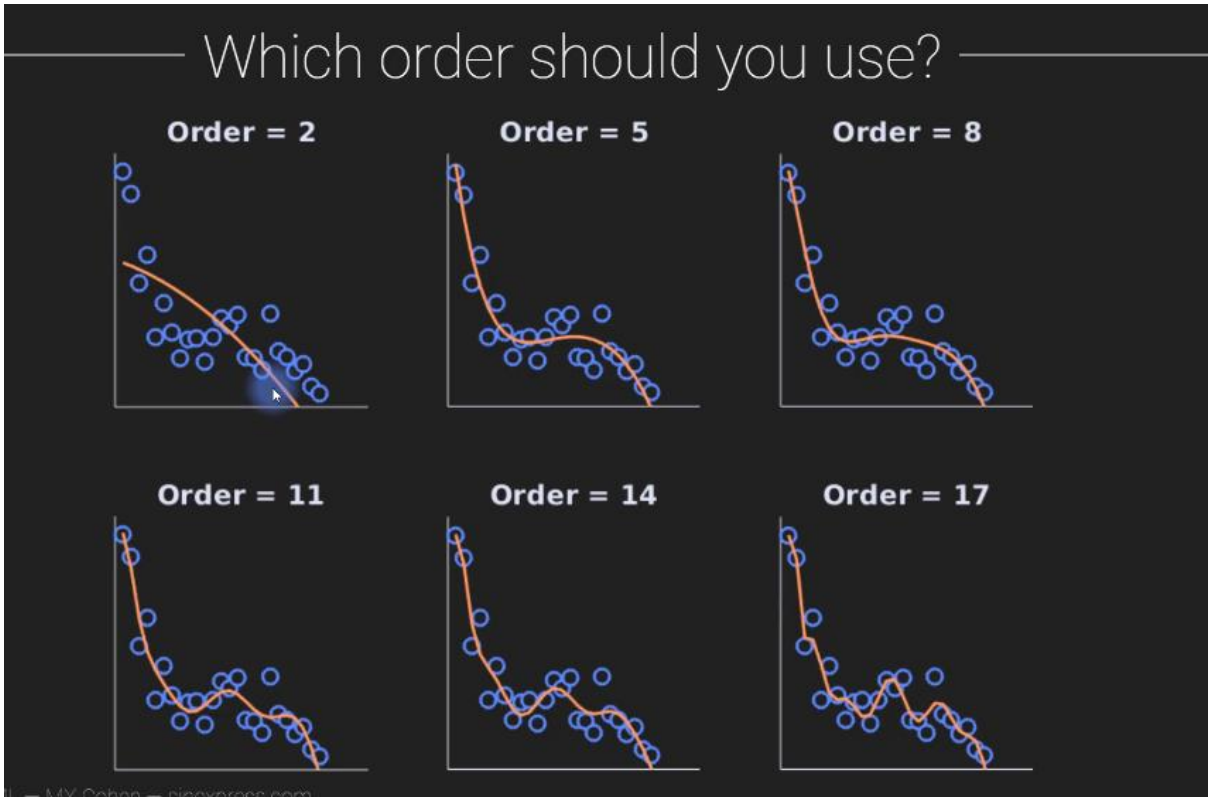
- **Basic Idea:** Normalize β so that its variance is 1.
- **Method 1:** z-normalize DV and IVs before the regression. All β's will be in the units of the data, which are already standard deviation units.
- **Method 2:** Scale the unstandardized β by the standard deviations of the IV and corresponding DV.

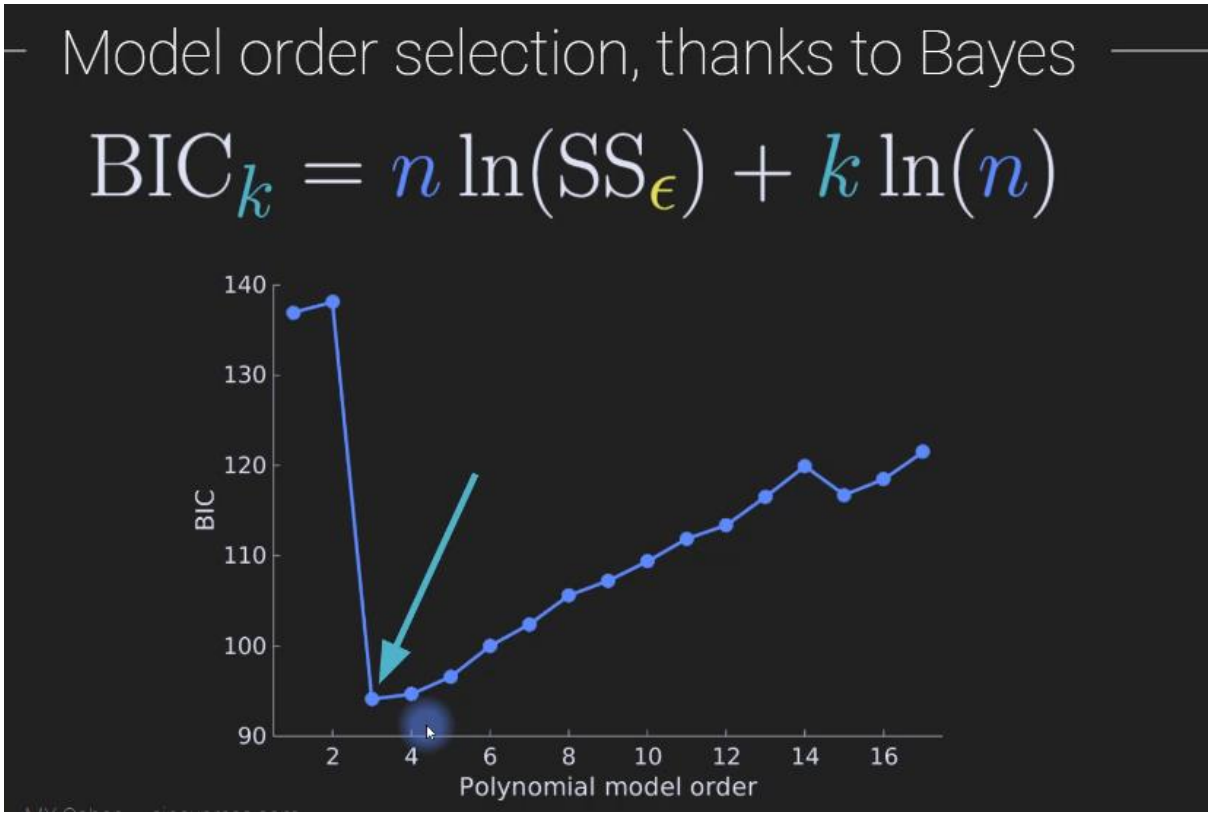
$$\boldsymbol{b_k} = \beta_k \frac{\boldsymbol{s_{x_k}}}{\boldsymbol{s_y}}$$



Polynomial Regression Models:







Logistic Regression:

What is a logistic regression?

- A logistic regression (a.k.a. binary logistic regression) has a binary DV ("logical"). Examples: true/false, male/female, has tumor/ doesn't have tumor, win/lose
- Can be extended to any number of categorical outcomes (a.k.a. multinomial logistic regression). Cat/penguin/truck.

What is the outcome of a logistic regression?

- A logistic regression does not classify; it returns probabilities of category membership.
- Classification can be implemented using a threshold, e.g., r > 0.5

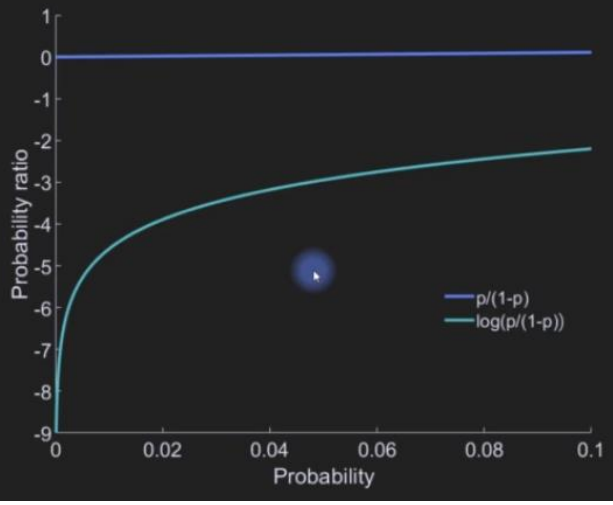
Setting up the logistic regression equation

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}$$

Why take the log of probabilities?



Main point: The log of small values has a larger dynamic range and is easier to work with in optimization problems.

How to find the best regression coefficients?

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}$$

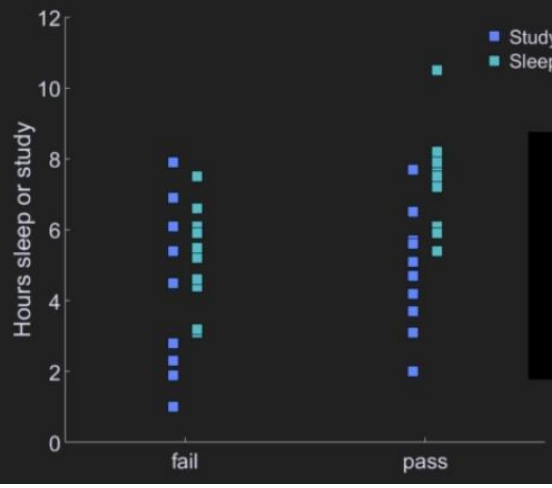
The nonlinearities in the coefficients prevent the left-inverse from being a viable solution.

Instead, iterative methods such as gradient descent are applied to find the set of parameters that make the probabilities best match the DV.

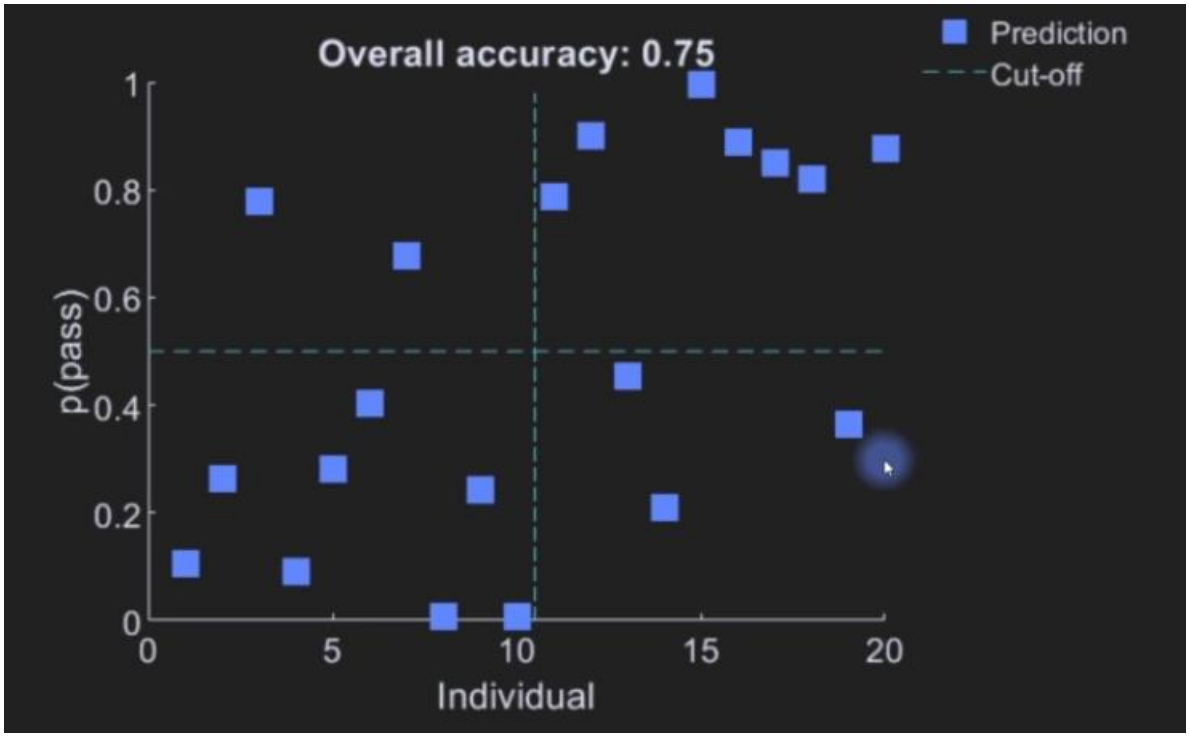
Example logistic regression:

- **Research question:** Does the amount of sleep and number of study hours predict passing an exam?
  Experiment: Ask students (N=20) to report the number of hours they slept, and the number of hours they studied. After the exam, hack into the university secure network, find those students' records, and steal their grades. In the interest of privacy, label each grade as pass or fail.

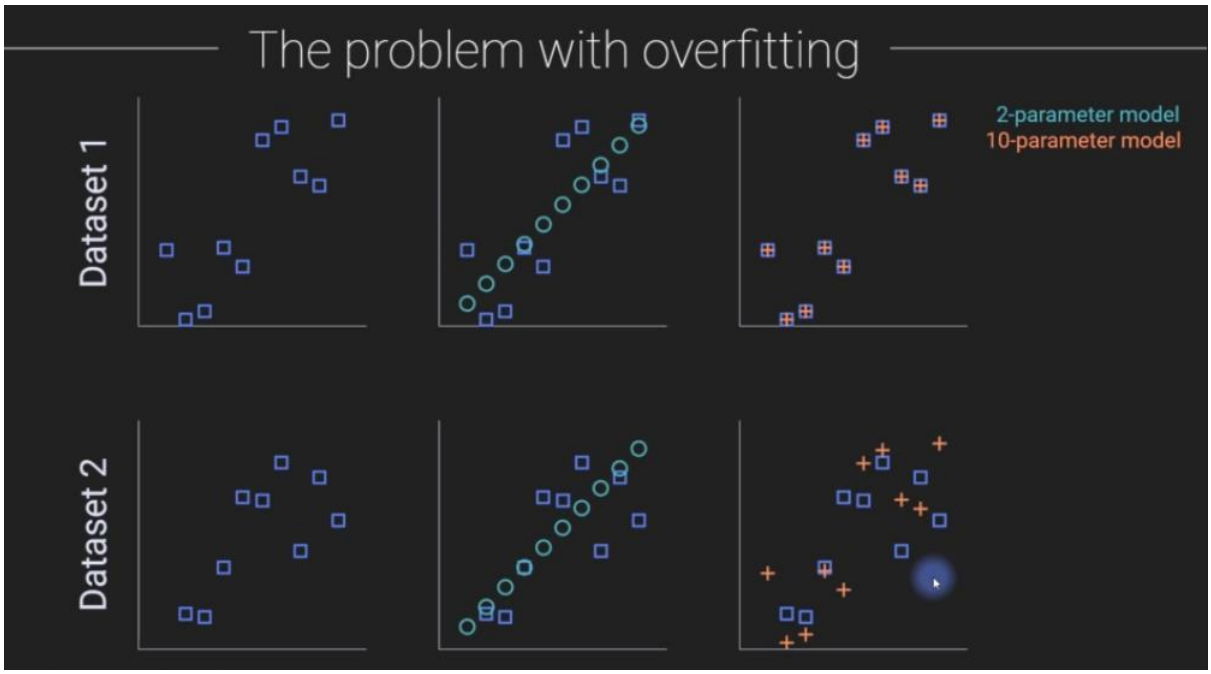
Example logistic regression

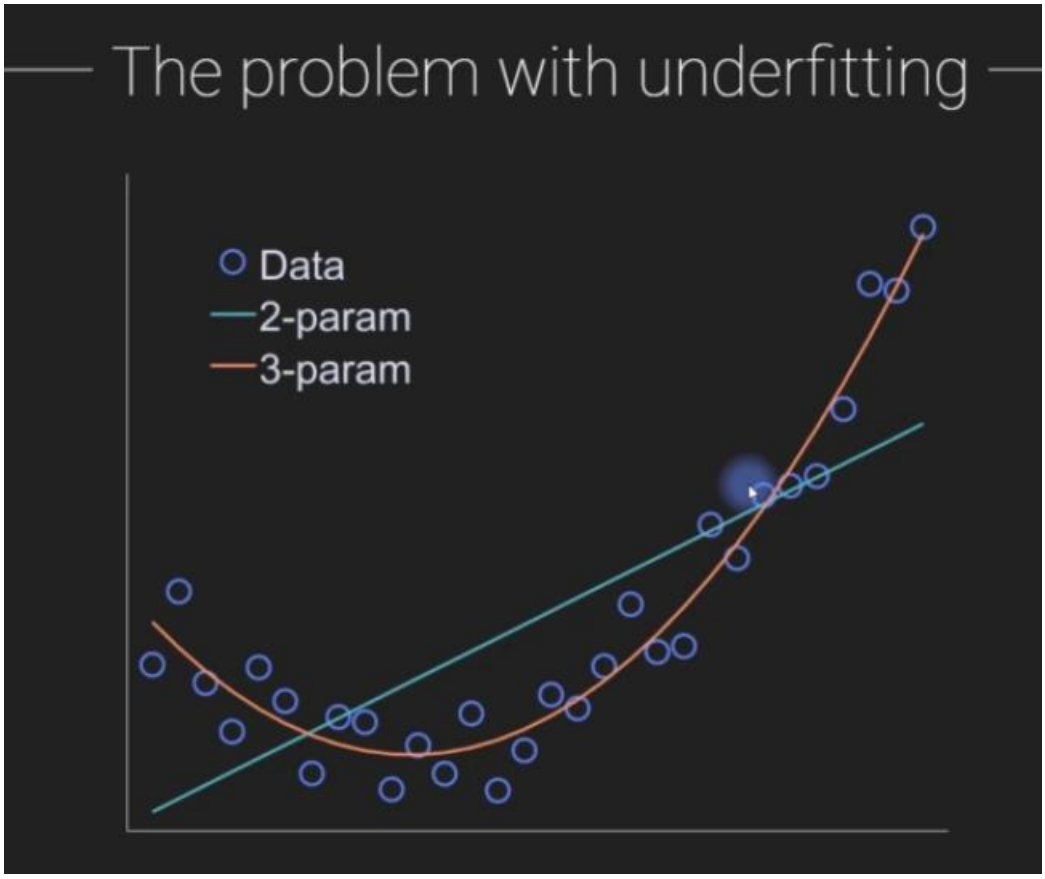


Source of variance	β coeff.	Std. error	t-value (df=??)	p-value
Intercept	-9.72	4.97	-1.9	.050
Study hours	.18	.35	.52	.603
Sleep hours	1.39	.63	2.21	.026



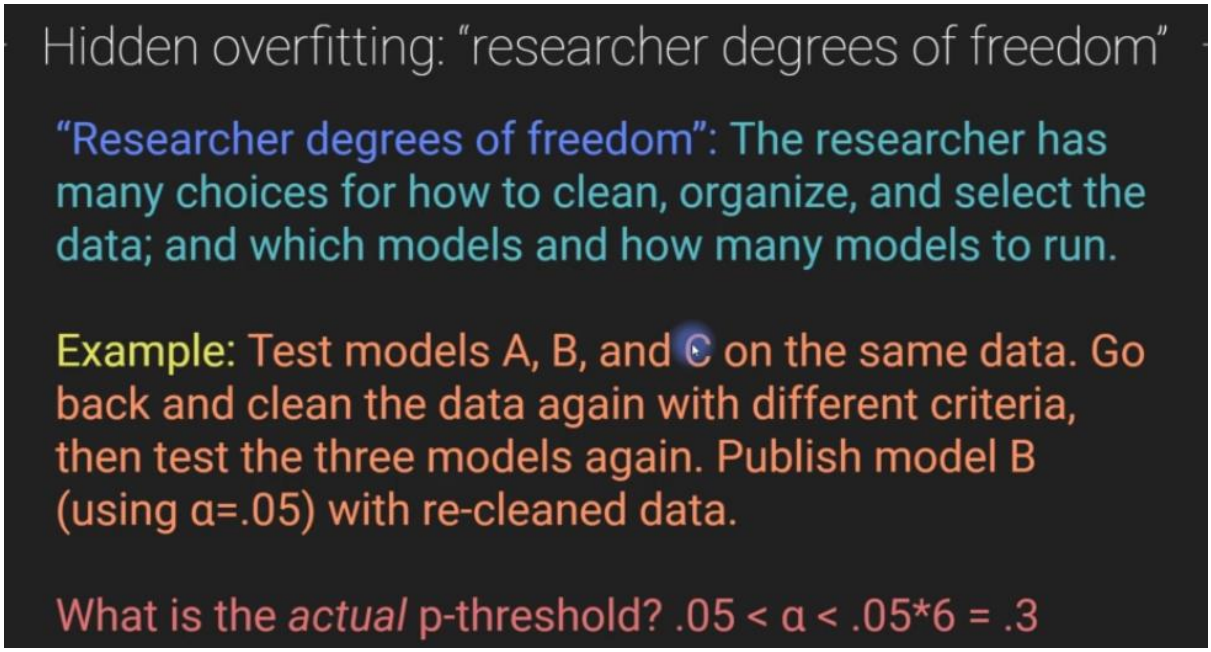
Under and Over Fitting:





How to know the correct number of parameters?

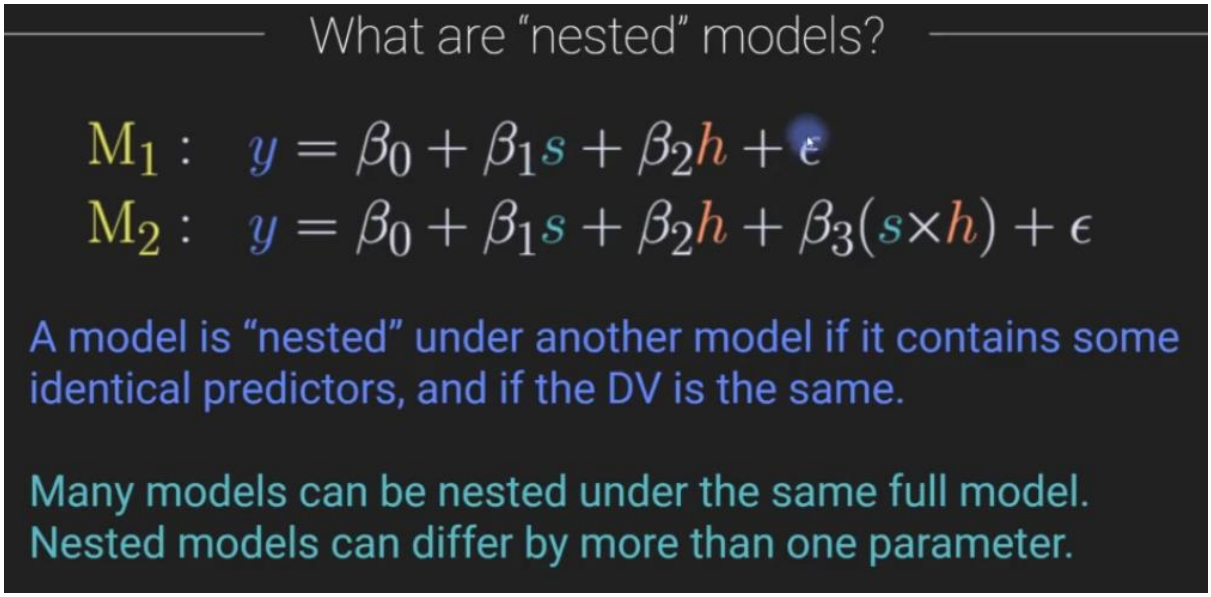
- ⇨ **With 1-2 dimensions:** Visualize the data and make an informed decision.
- ⇨ **With 3+ dimensions:** Formal model comparision (and also the polynomial regression lecture).



How to avoid researcher overfitting?

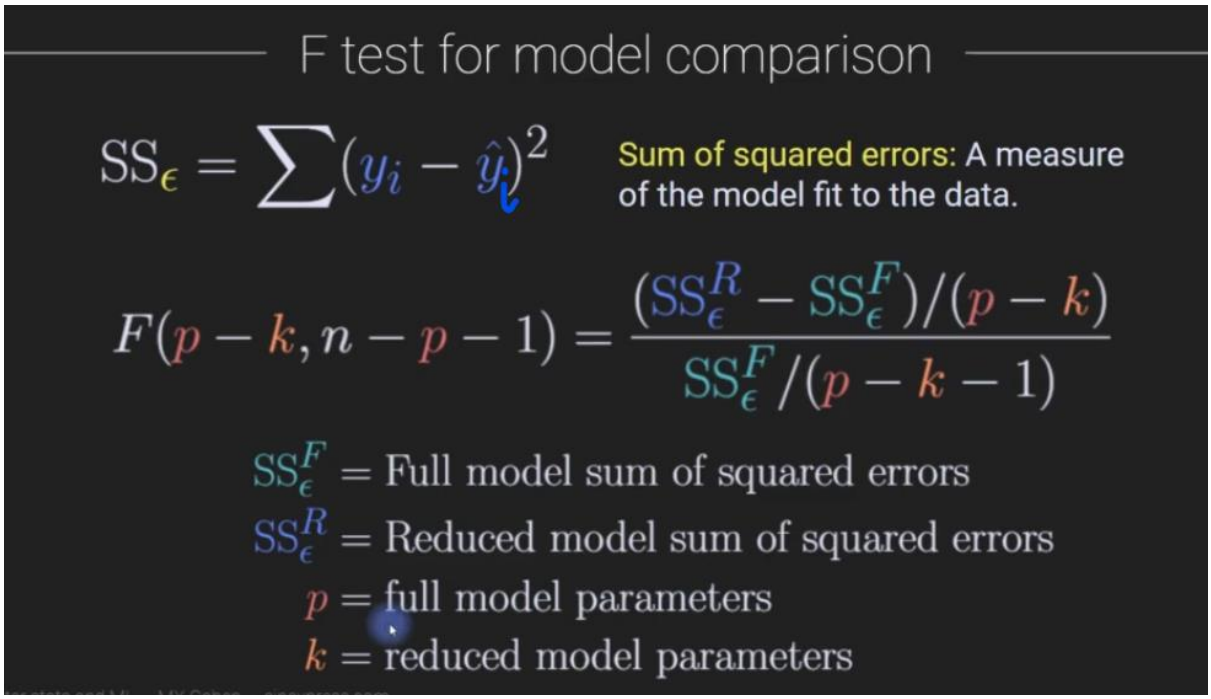
- ⇨ Decide your complete analysis pipeline in advance (before data collection). Any deviations from this plan must be clearly stated (e.g., pre-registered report)
- ⇨ Explore/develop an analysis pipeline in a small sample of the data, and then apply that pipeline to the rest of the data. (informal training/testing sets)

Comparing "nested" models:



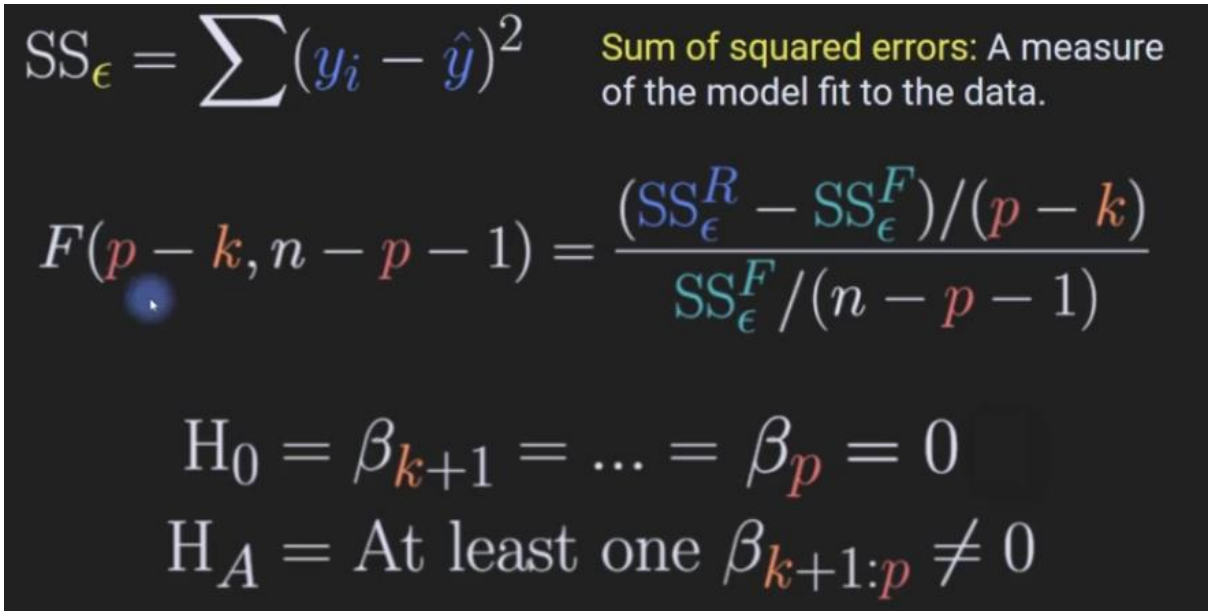
Important observation: The full model will always fit the data better than the reduced model. ALWAYS!

But models should be penalized for having more parameters. Thus, more parameters (more complicated model) is justified only when the model significantly improves the fir to the data.



F is statistically significant: More parameters improve the model. Prefer the more complicated model.

F is nonsignificant: The model with fewer parameters fits as well as the model with more parameters. Prefer the simpler model.

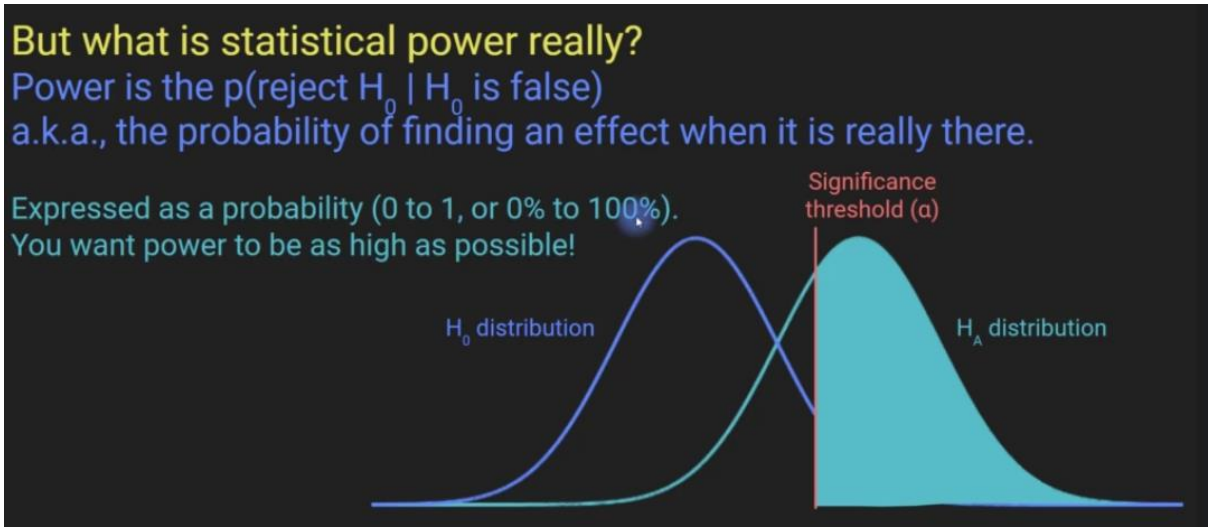


What to do About Missing Data?

- **Option 1:** Complete removal that rows.
  - Used for:
    - Paired data
    - Analysing changes
- **Option 2:** Selective removal.
  - Used for:
    - Unpaired data
    - Within-variable analyses
- **Option 3:** Replacement (Replace missing data with column mean)
  - Used for:
    - Small datasets
- **Option 4:** Prediction
  - Used for:
    - Enough data and other columns (features) to generate a useful predictive model.

Statistical Power and Sample Size

What is Statistical Power and Why is it important?



How to maximize statistical power:

- Sample Size
- Effect size
- Lower α (e.g., $p < 0.1$)

Statistical power decreases with:

- Variability
- Higher α (e.g., p < 0.01)

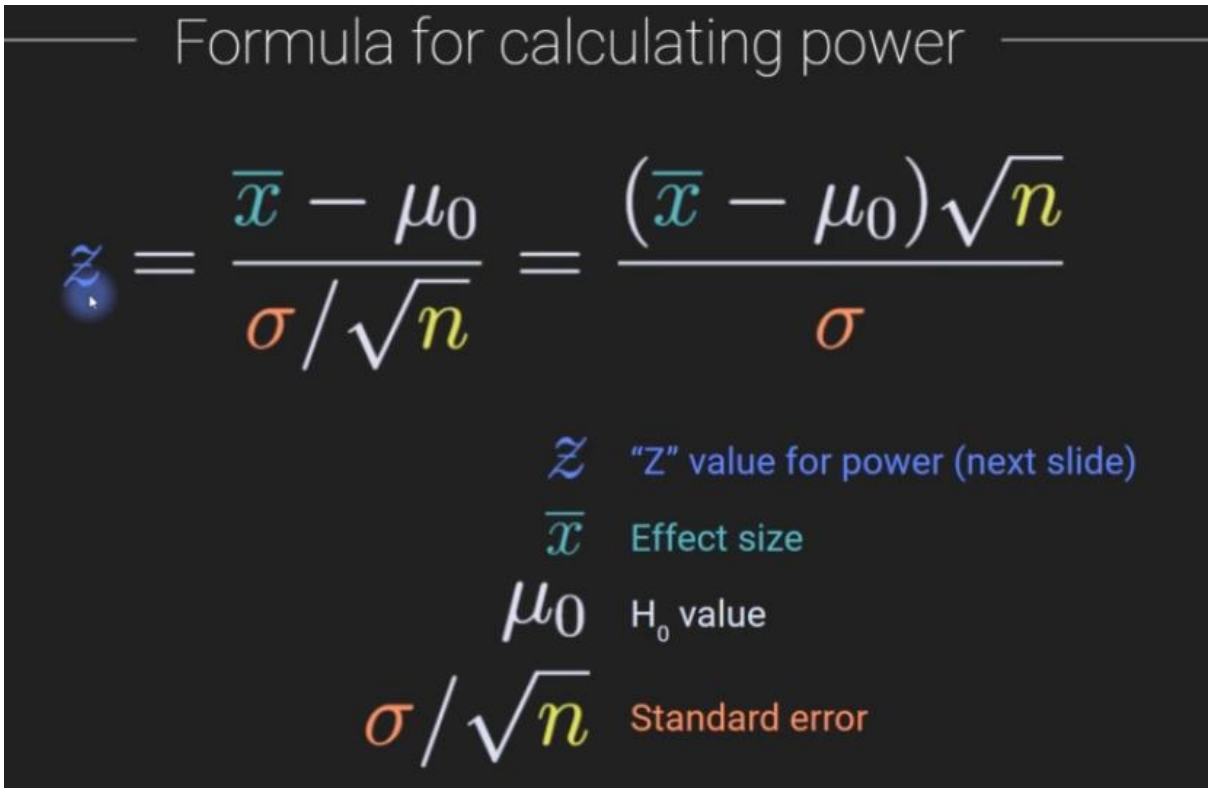
The problems with power:

- Increasing alpha increases power but also increases p (Type I error).
- Some factors that influence power you can control; others you cannot (sample size, effect size, variability).
- Power can differ for different analyses in the same experiment.
- Computing "true power" requires knowing the true effect size.
- Power calculations from published studies are unreliable due to "publication bias" (non-significant findings are not reported).

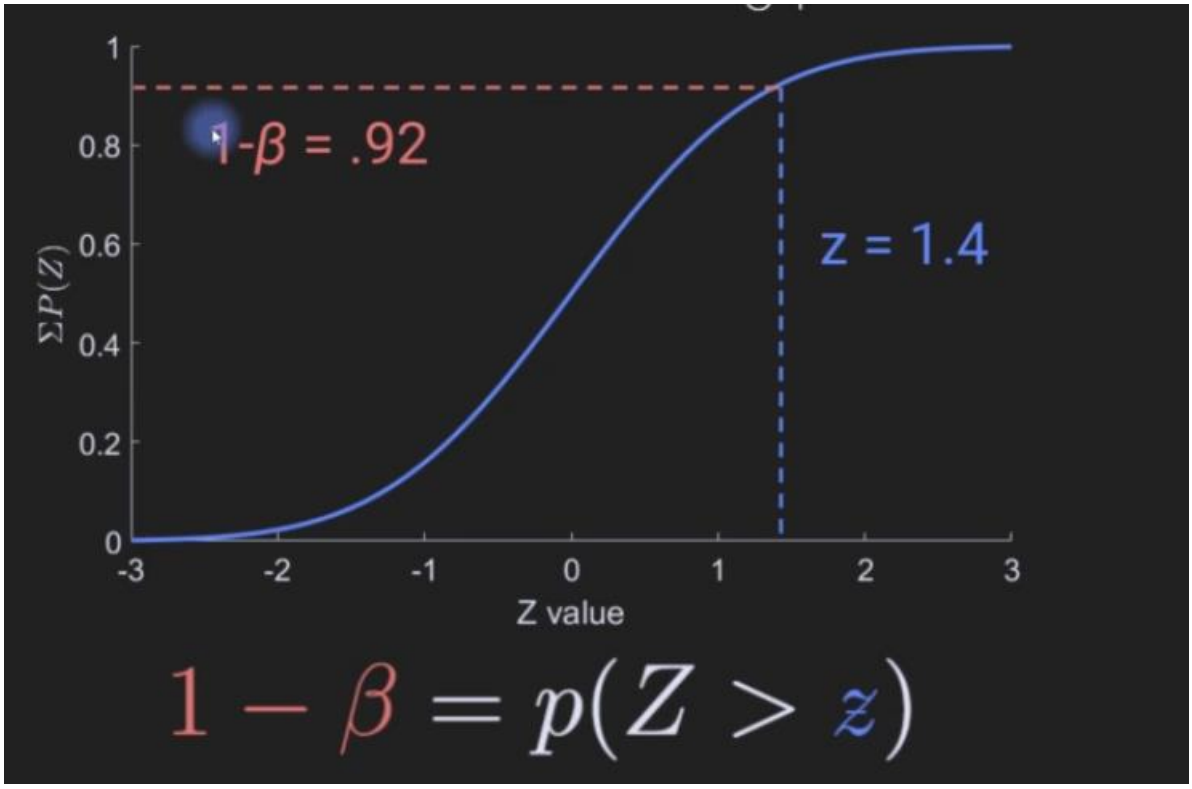
Conclusion:

Statistical power is more of a useful guideline than a precise and trustworthy numerical value.

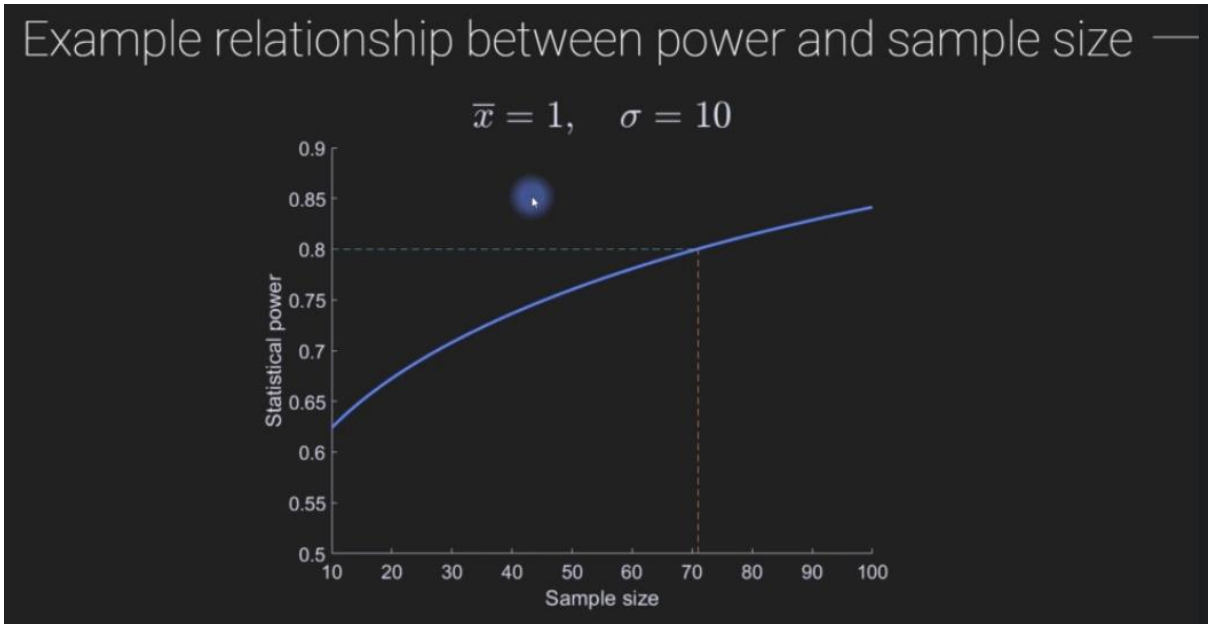
Estimate Statistical Power & Sample Sizes:



$$1 - \beta = p(Z > z)$$



$$n = \left(\frac{z\alpha}{\bar{x} - \mu_0}\right)^2$$



Is it all so simple?

- The formulas for more complicated statistical models (multiple regression, AVOVAs, etc) are complicated, although the principal is the same.
- In practice, best to use an online statistical power calculator.

A priori power vs post-hoc power:

- **A priori power**
  - Get effect size and standard deviation from published studies or from pilot data.
  - Used to compute sample size prospectively.
  - Widely used and accepted but wrought with difficulties and uncertainties.
- **Post-hoc power**
  - Compute the power based on your completed study.
  - Proportional to the p-value, thus provides no new info.
  - Sometimes asked for by well-meaning people…

Compute Power & Sample Size Using G*Power:

https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

Clustering & Dimension Reduction:

K-means Clustering:

Objective:

- The goal of k-means clustering is to group multidimensional data into k groups.
- Group membership defined as distances.
- Hence, k-means clustering attempts to minimize with-group distances and maximize between-group distances.

Basic k-means Algorithm:

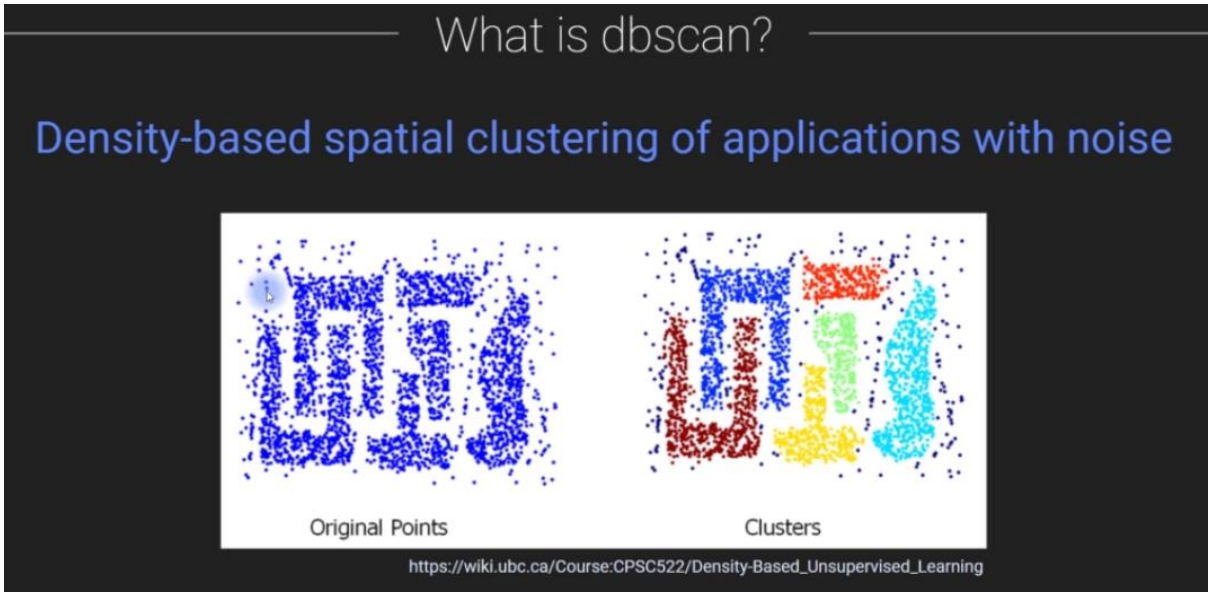
1. Select k.
2. Create k centroids at random locations inside the dataset.
3. Compute sum of squared distances (errors) from all data points to all centroids.
4. Assign each data point to its closest centroid.
5. Create new centroid at average of all data points.
6. Repeat steps 3-5 until convergence.

Difficulties with k-means:

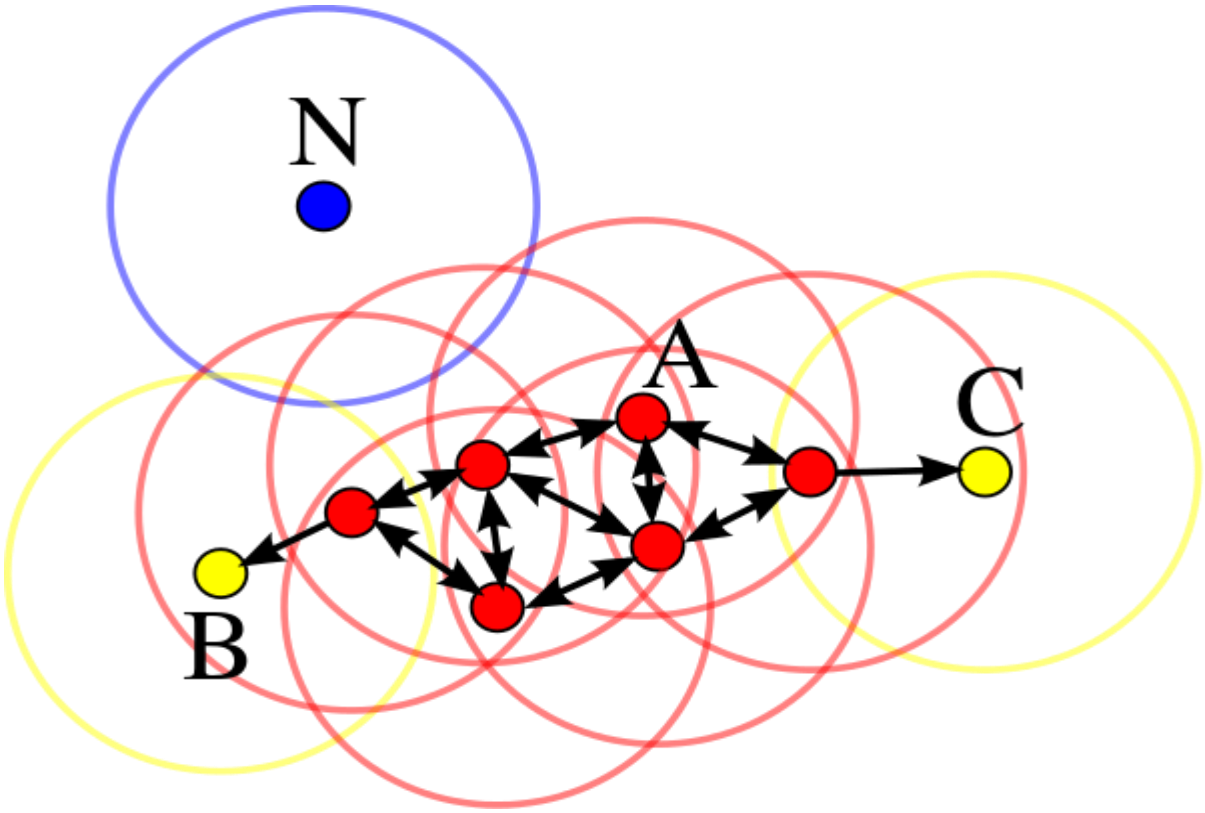
1. Proper k can be difficult to know a priori.
2. Evaluating the proper k is also difficult (except in simple examples).
3. Multidimensional clustering is difficult or impossible to visualize.
4. Repeating the algorithm can give different results.
5. Can be suboptimal when cluster sizes differ.
6. Not all clusters are distance-based.

Code: Part 1 - K-Means Clustering

Clustering Via dbscan:



Overview of dbscan algorithm:



Algorithm:

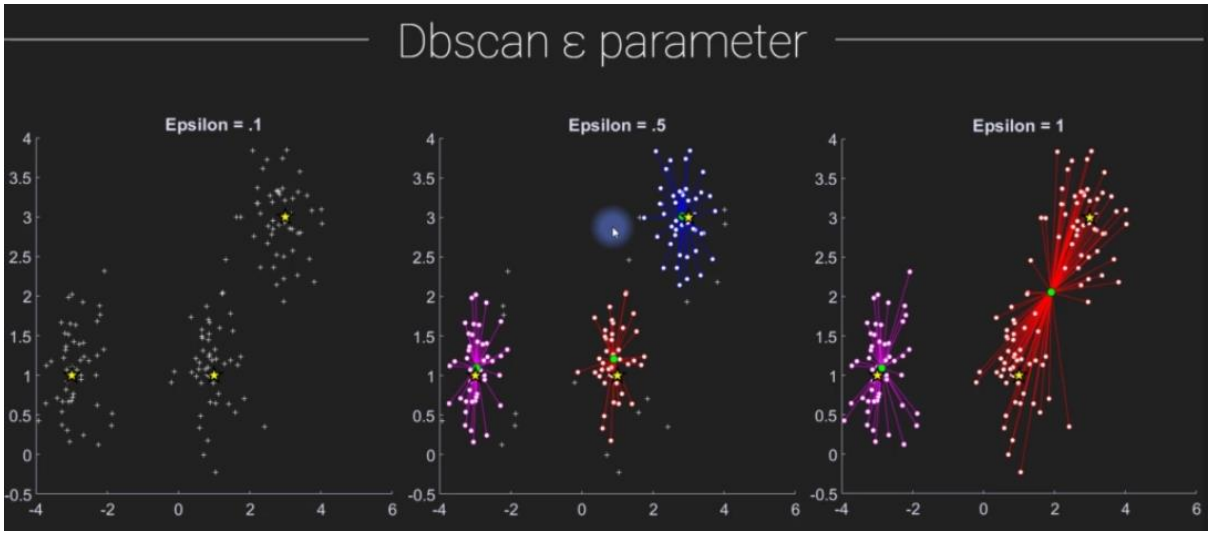
- https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html
- https://towardsdatascience.com/dbscan-algorithm-complete-guide-and-application-with-python-scikit-learn-d690cbae4c5d

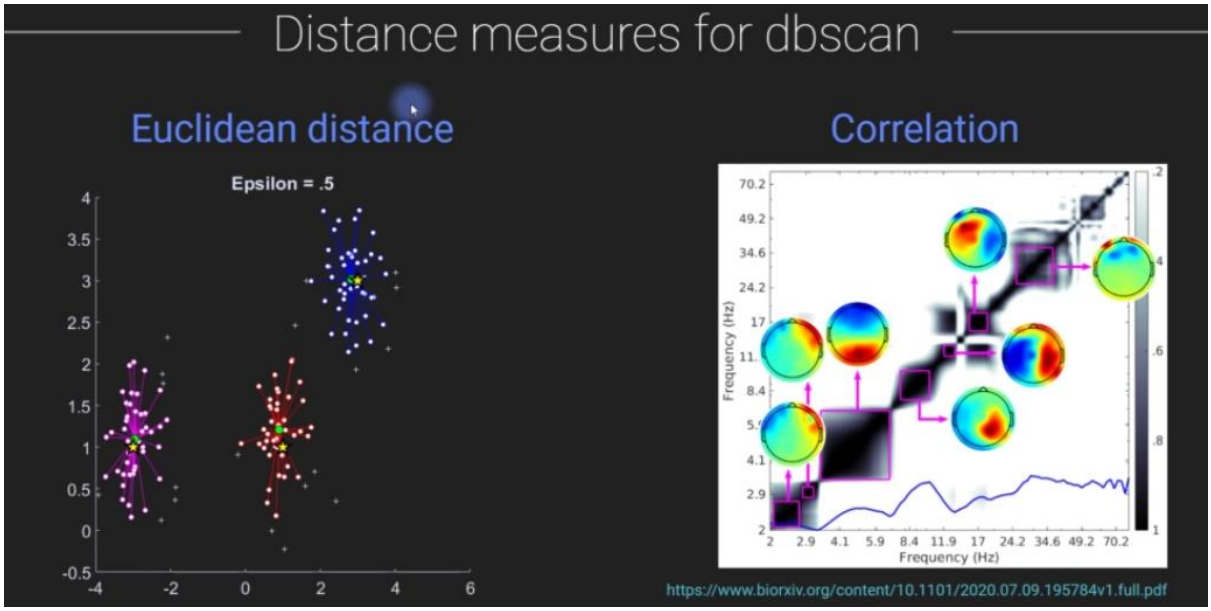
- https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering

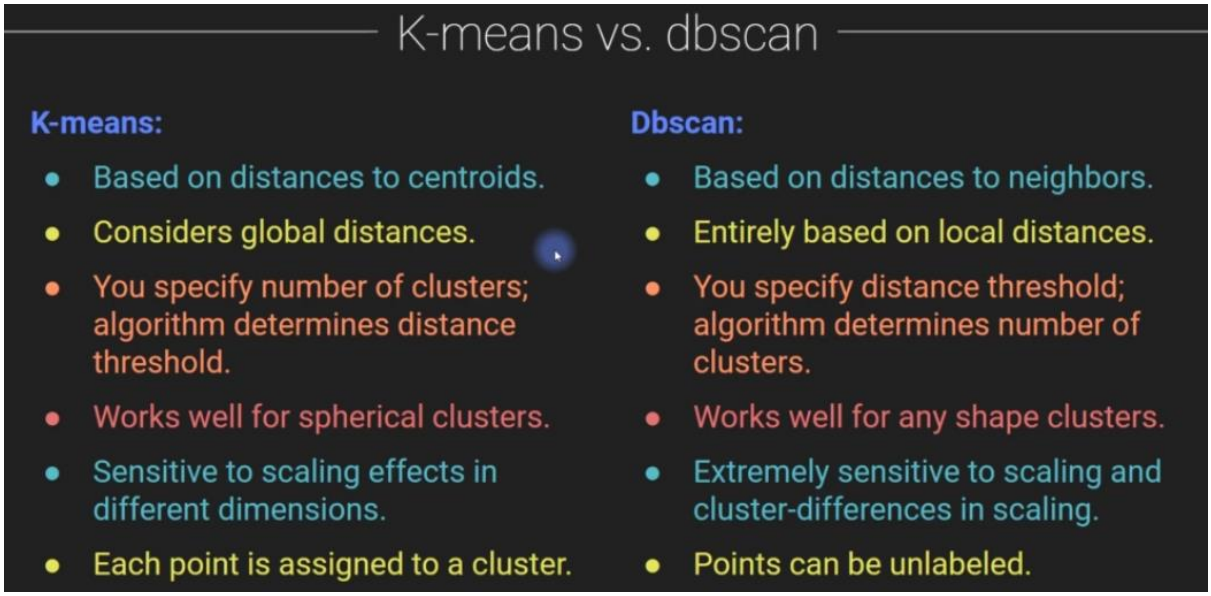
Dbscan parameters:

- **Epsilon(ε):**
  - Step size for finding clusters.
  - Too small -> clusters broken up into many clusters.
  - Too large -> separate clusters combined into one

- **Minimum points:**
  - Minimum number of points to be a "cluster"
  - Too small -> many small clusters
  - Too large -> true small clusters ignored.





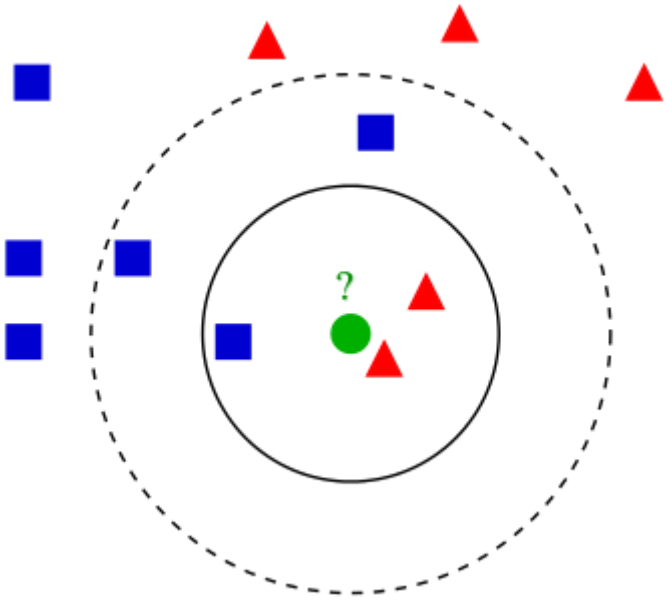


Code: Part 2 – DBScan Clustering

Hierarchical vs DBScan Clustering:

https://ryanwingate.com/intro-to-machine-learning/unsupervised/hierarchical-and-density-based-clustering

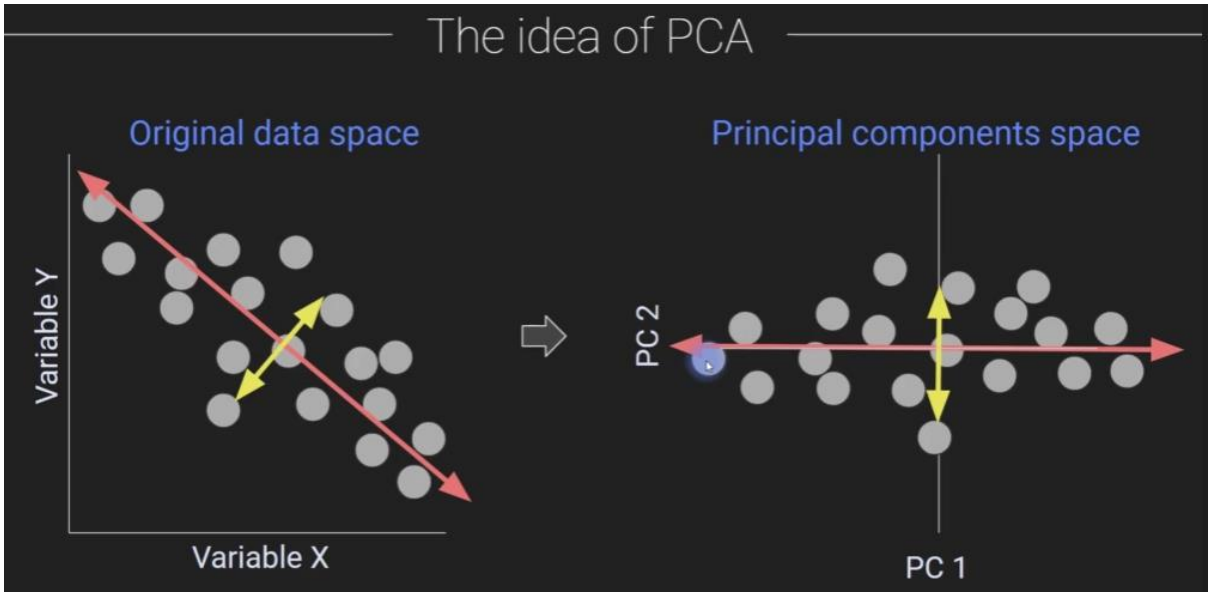
K-Nearest Neighbors Classification (KNN):

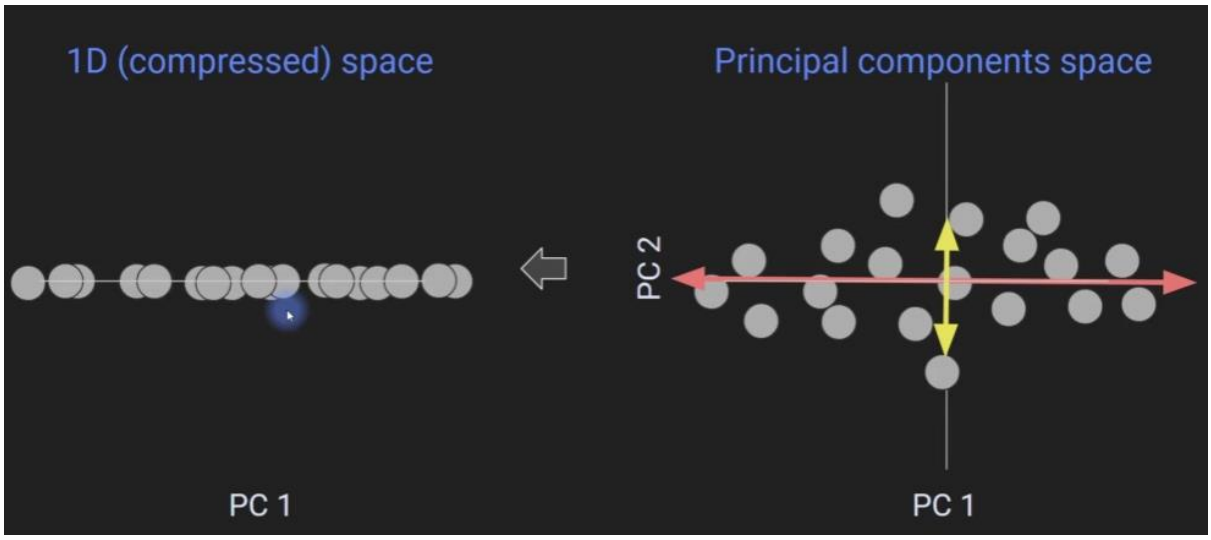


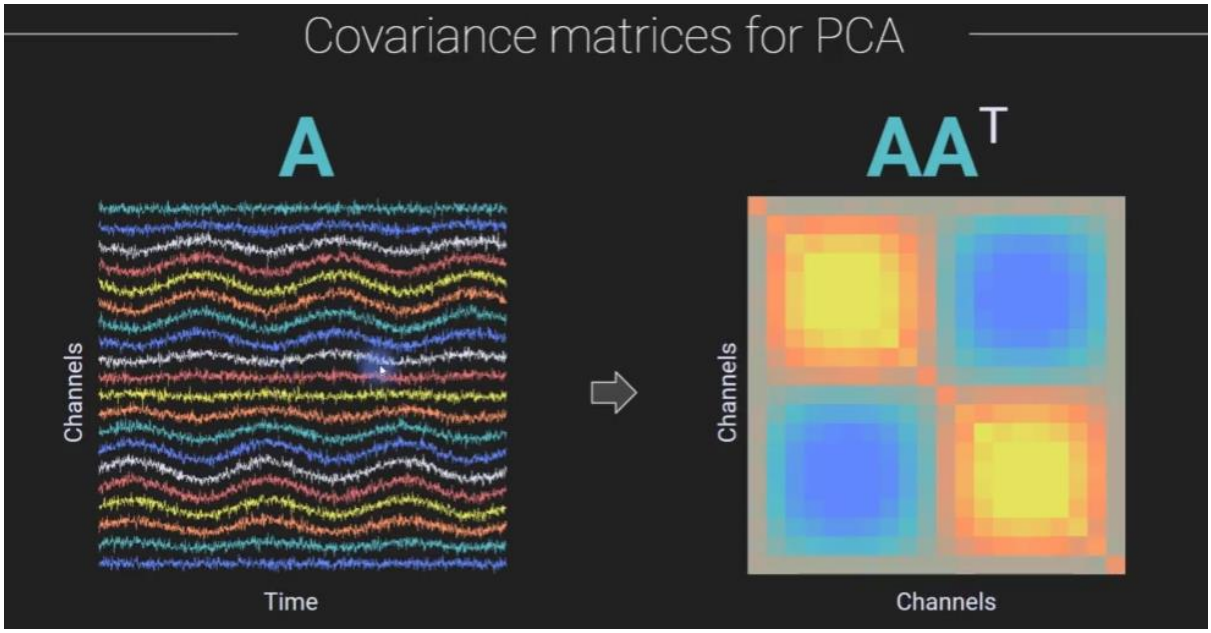
Wikipedia Page: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Code: Part 3 – KNN Classifier

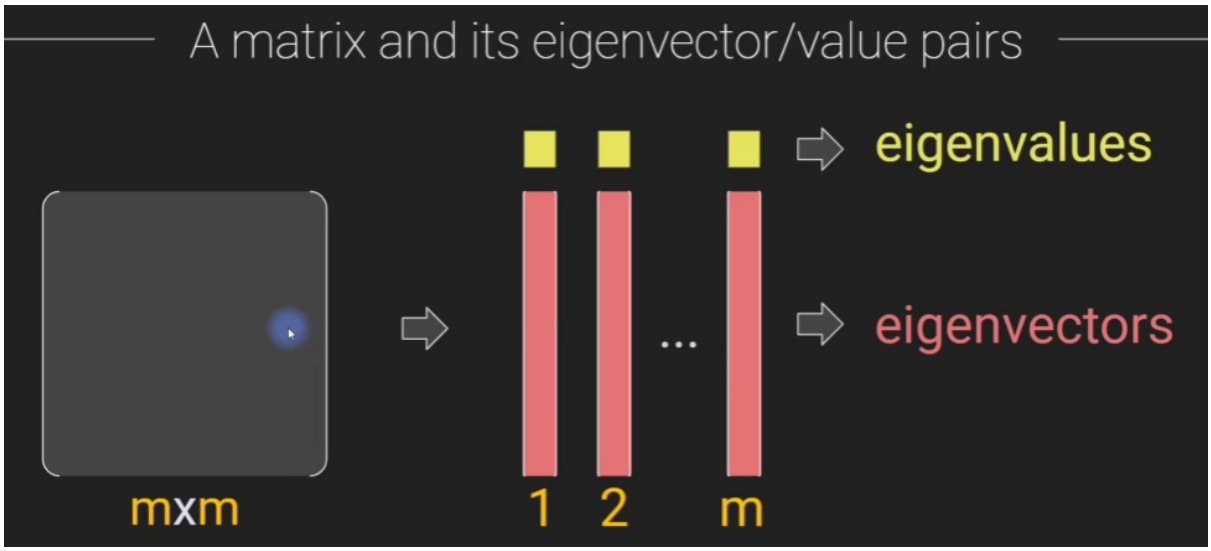
Principal Components Analysis (PCA):



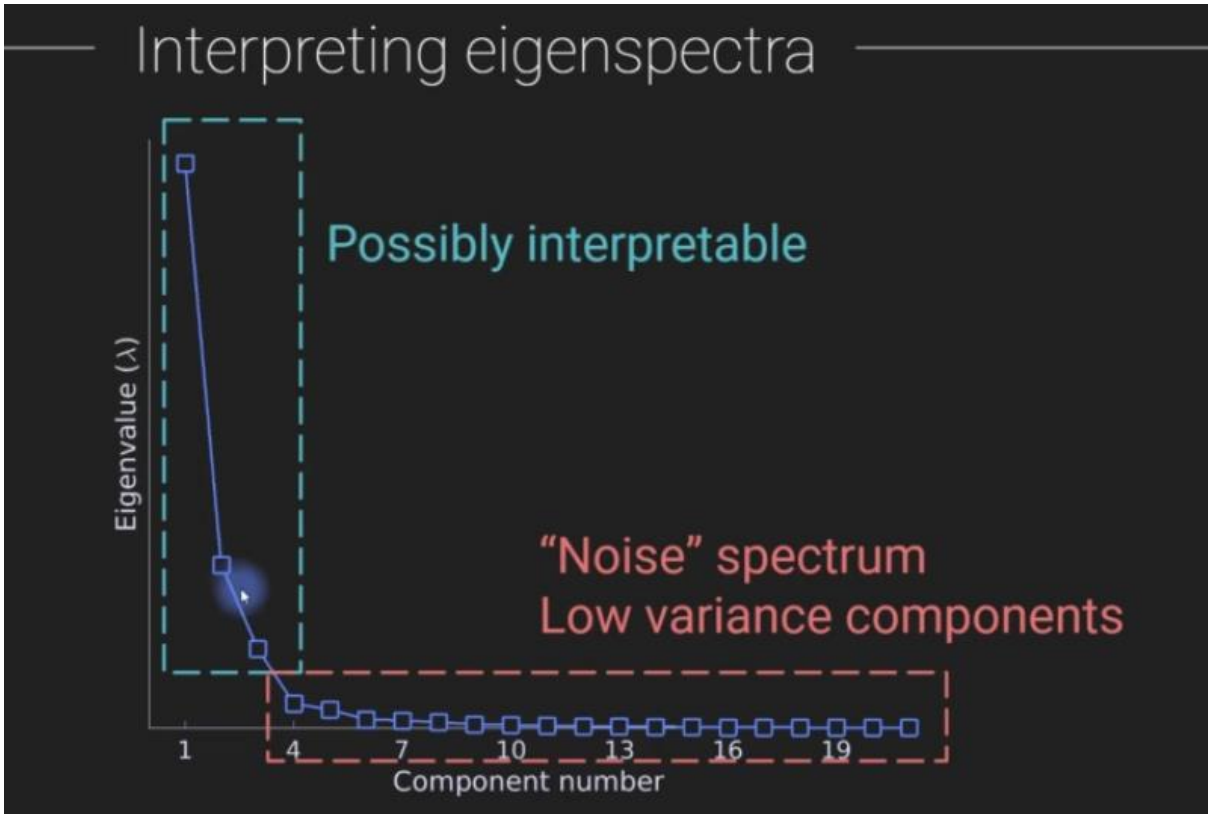




Right side matrix is the covariance matrix

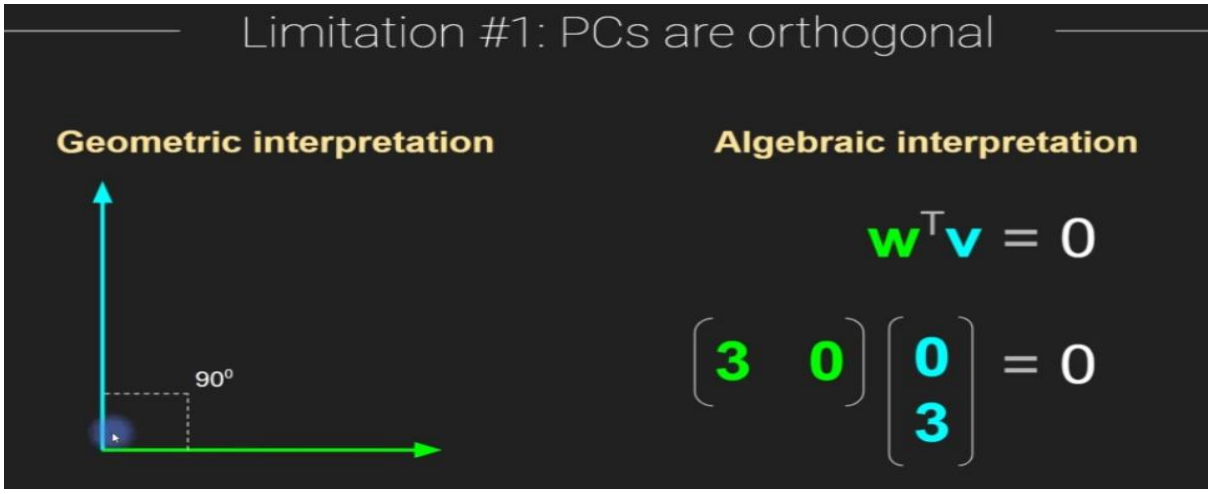


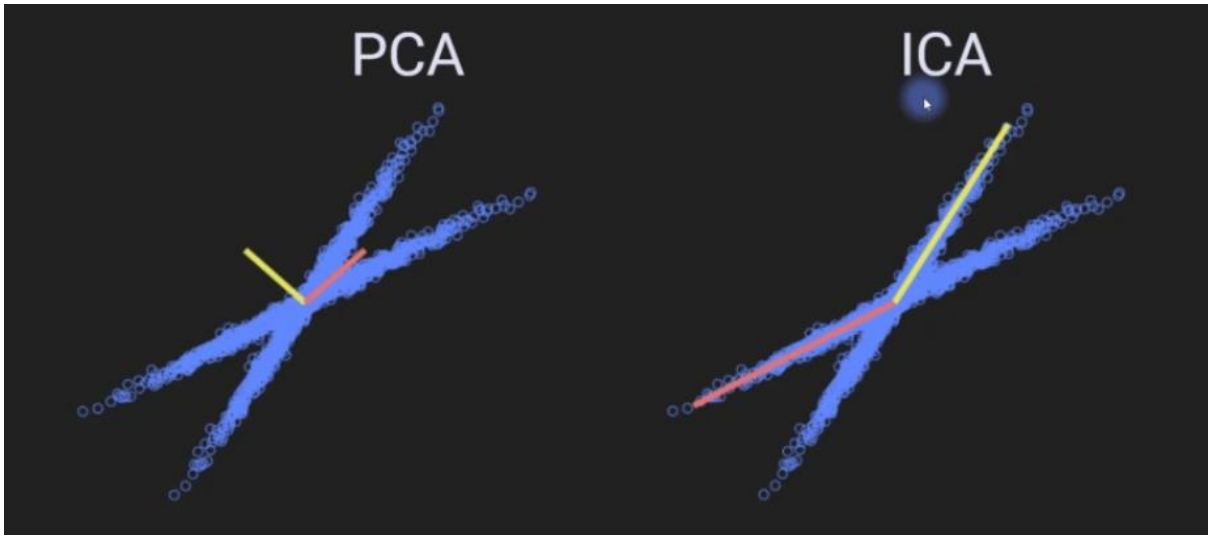
Here, mxm is the covariance matrix. Each eigenvectors has same number of elements as number of elements in a column of the covariance matrix. So, every eigenvectors have m elements.



Two limitations of PCA:

These are not necessary disadvantages.





Conclusion of PCA:

- PCA is great for data compression (dimensionally reduction) and visual exploration.
- PCA may be suboptimal or misleading when interpreting PCs as "factors" or unique sources of variance in the data.

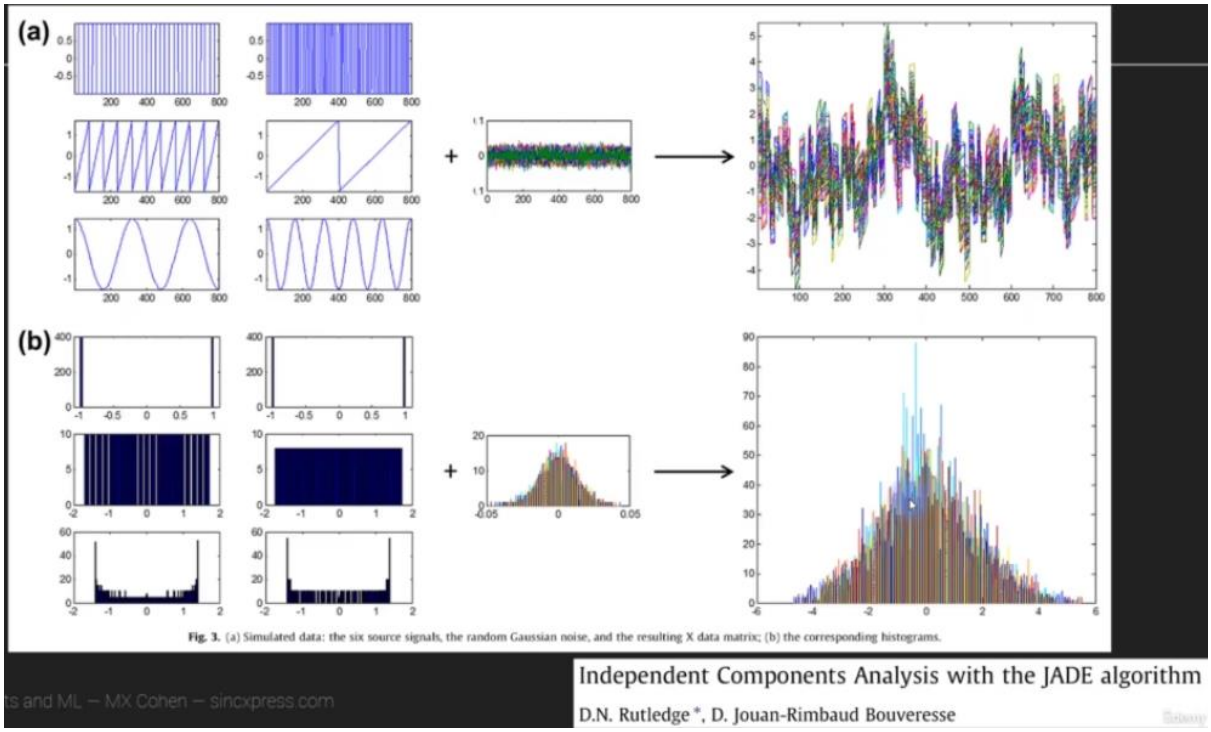
Code: Part 4 - PCA (Principle Component Analysis)

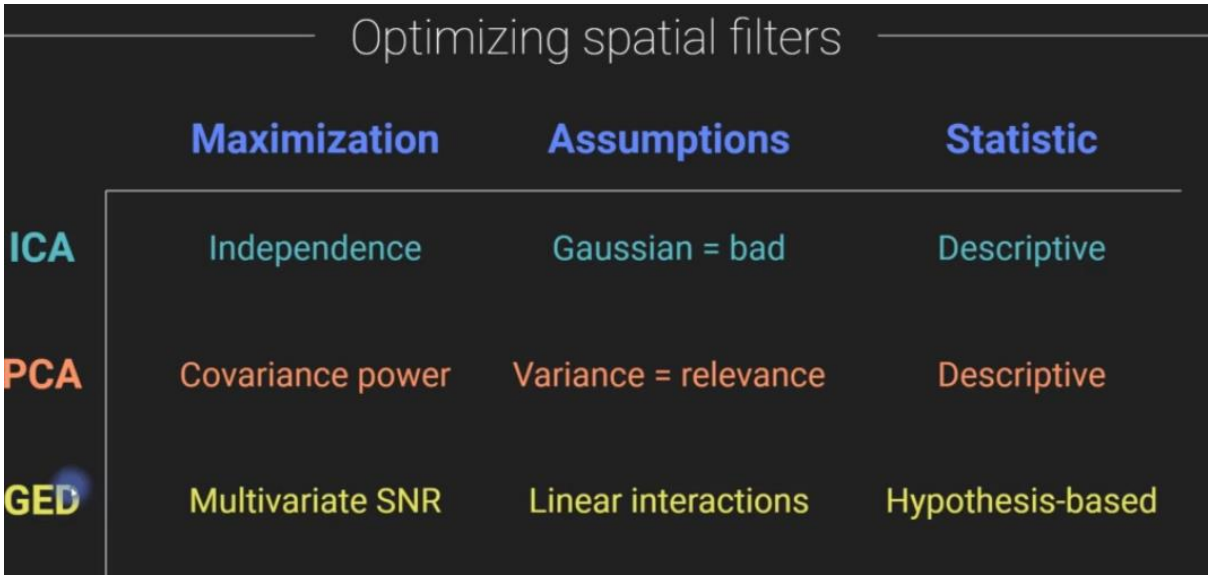
Independent Components Analysis (ICA):

Blogs:

- https://towardsdatascience.com/independent-component-analysis-ica-a3eba0ccec35
- https://www.geeksforgeeks.org/ml-independent-component-analysis

Central Measure Theorem says that ICA: Gaussian = BAD

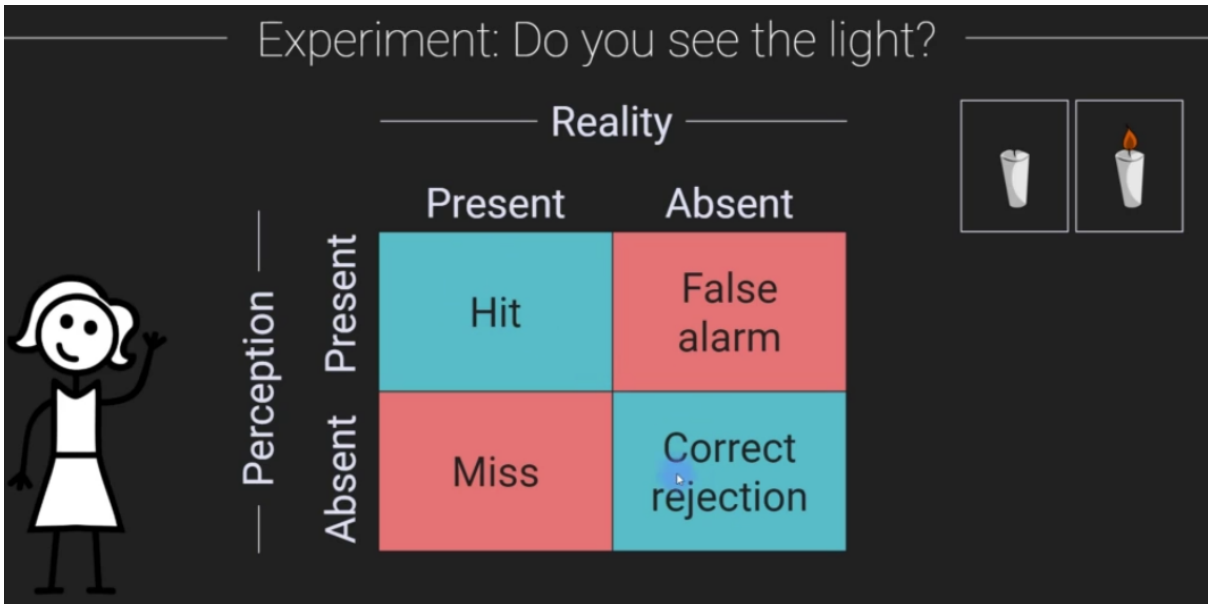




Code: Part 5 - ICA (Independent Component Analysis)

Signal Detection Theory

The Two Perspectives of the World:

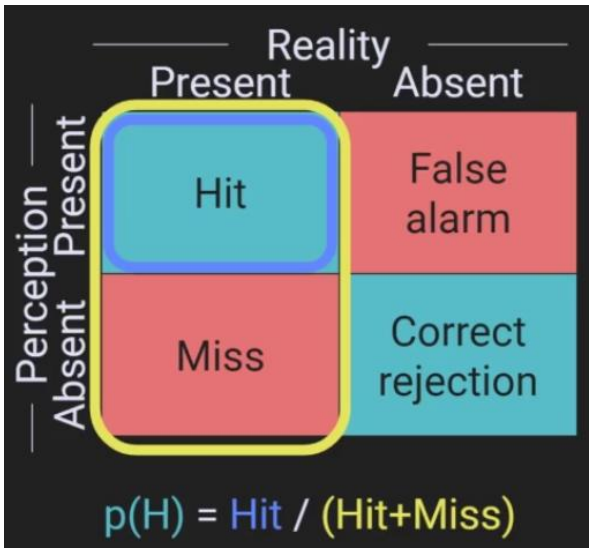


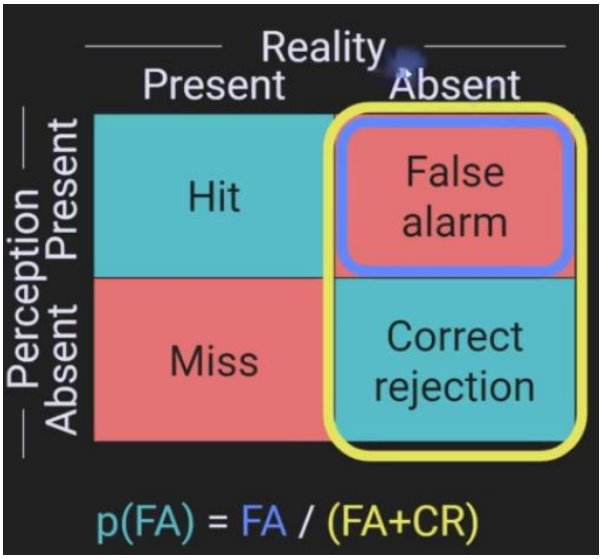
d-prime (d'):

- d-prime is a measure of discrimination.
- It is necessary to distinguish between all yes responses (hits + false alarms) and correct yes responses (hit only).
- d' is therefore a more accurate measure of performance.

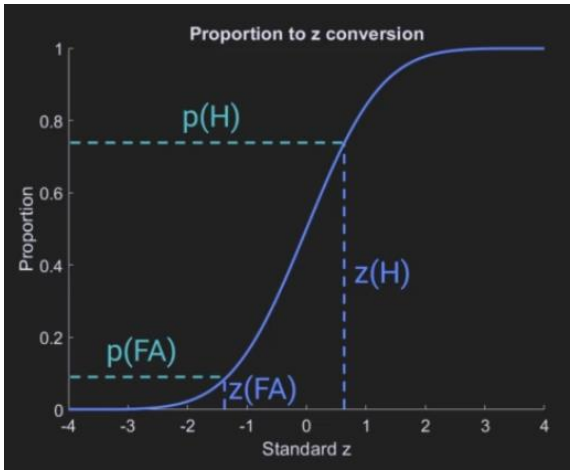
Algorithm to compute d':

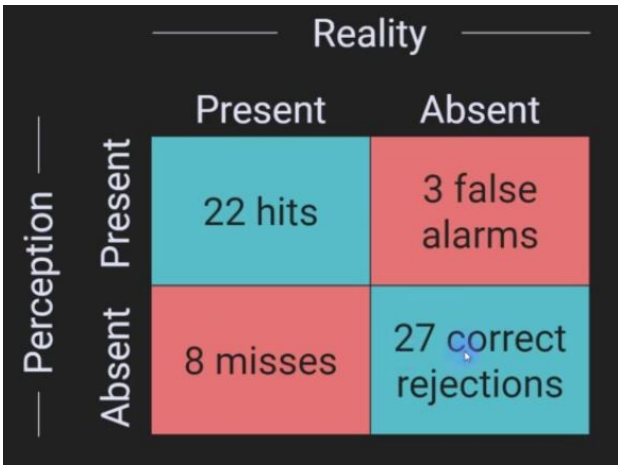
1. Convert hits and false alarms to proportion relative to total number of veridical events.





2. Convert proportion to standard z.
3. d' = z(H) – z(FA)





P(H) = 22/30 = 0.73

P(FA) = 3/30 = 0.1

z(H) = 0.622

z(FA) = -1.28

d' = 1.90

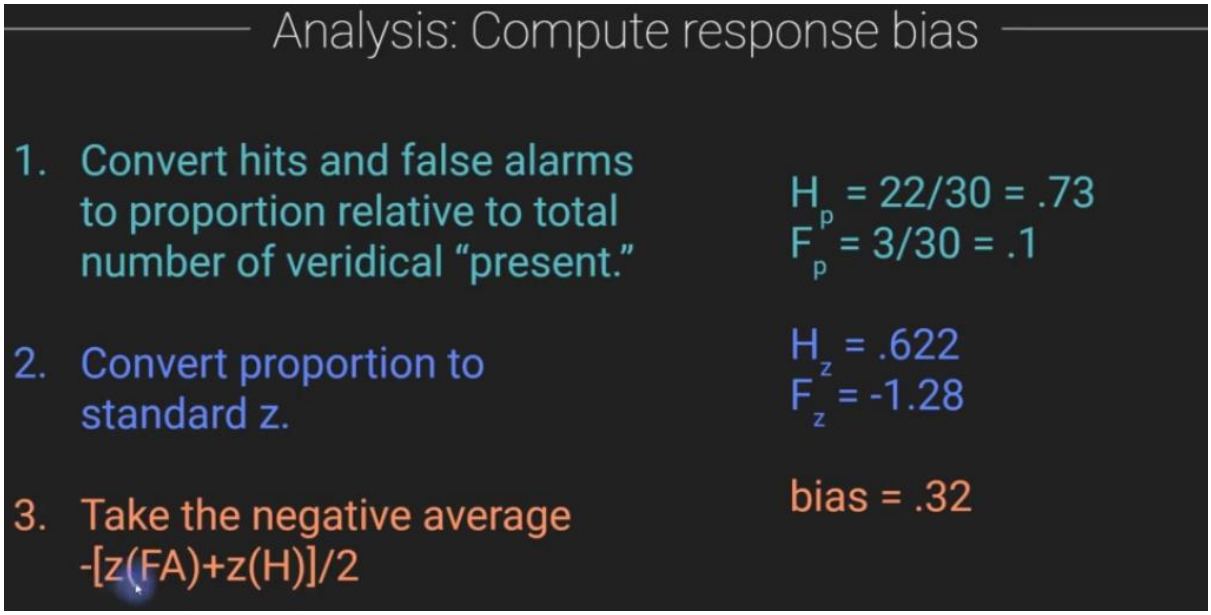
How to interpret d'?

Issues with d':

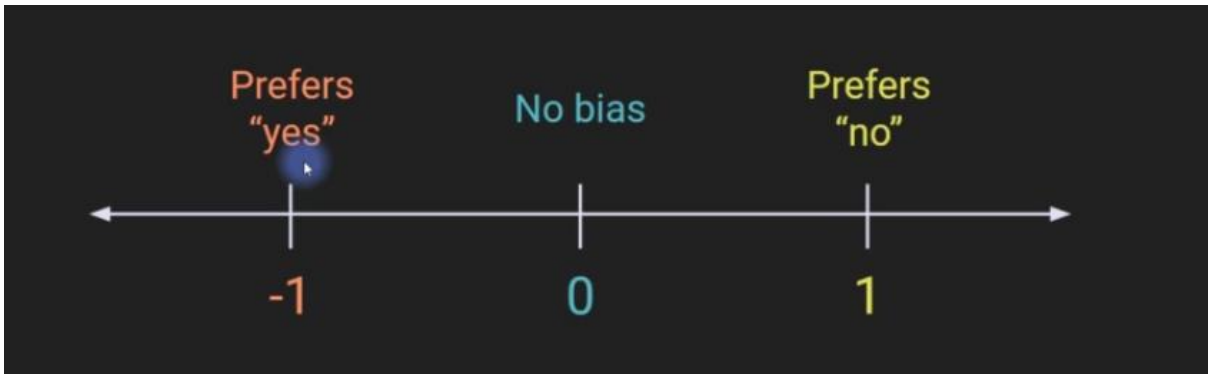
- Doesn't work when hits or false alarms are 0% or 100%
- Insensitive to response bias.

Response Bias:

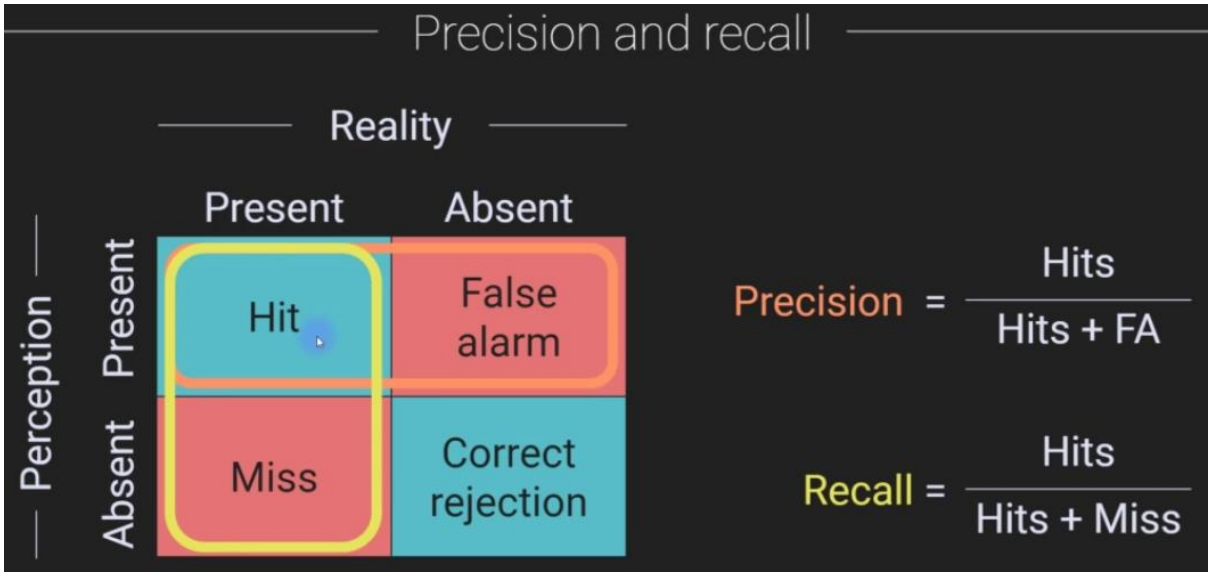
- The tendency to respond "Present" more often vs "Absent" more often.
- Response bias determines whether someone tends to response "yes" or "no" more often.
- Response bias is orthogonal to (unrelated to) d', because very different d-primes can be associated with the same bias.



How to interpret response bias?



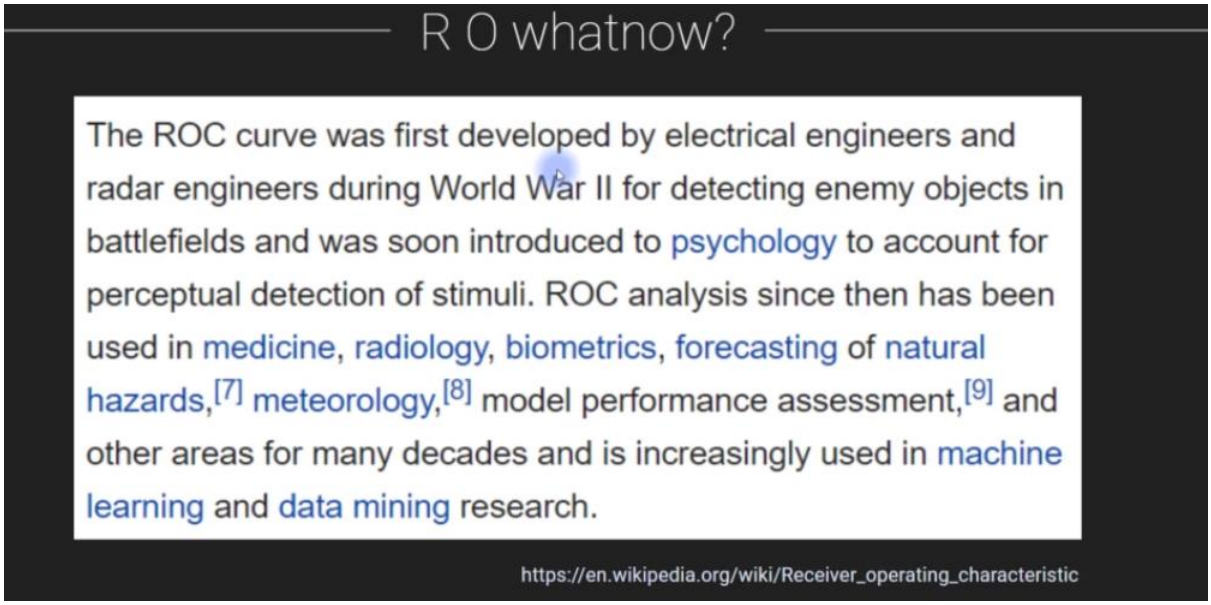
F-Score:

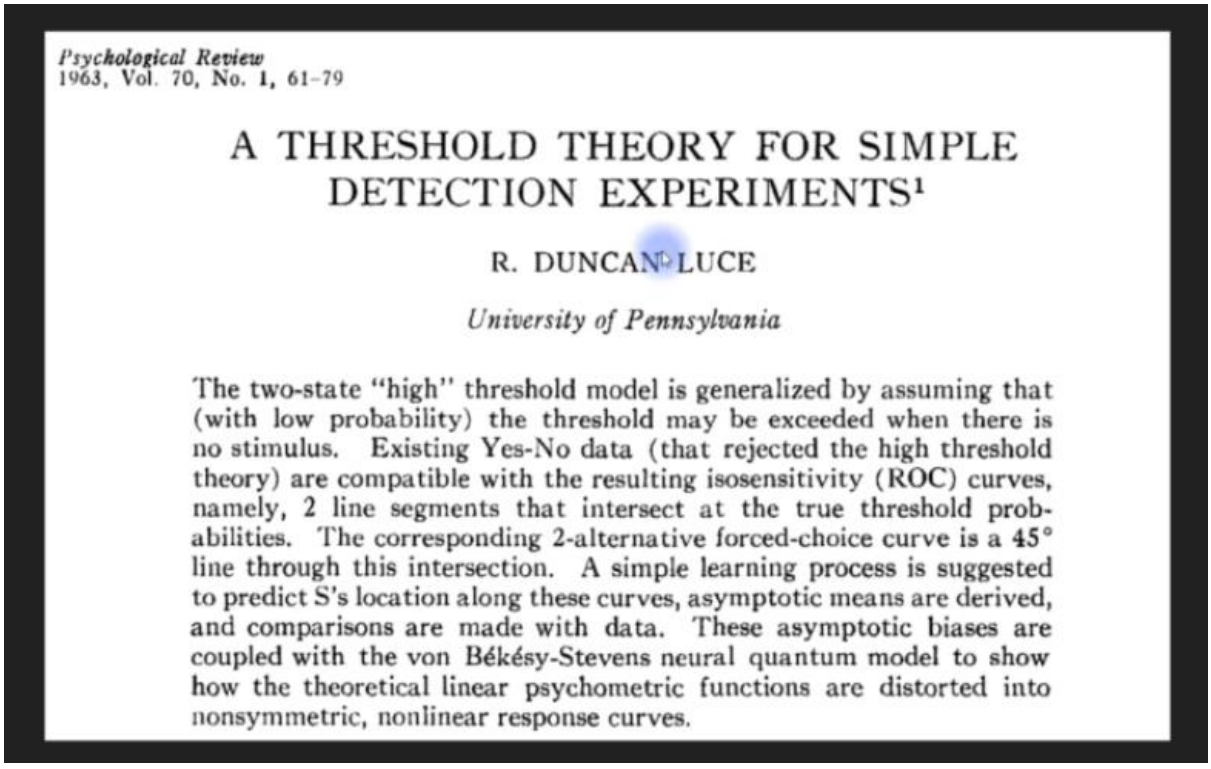


$$F_1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_1 \text{ score} = \frac{\text{Hits}}{\text{Hits} + (\text{FA}+\text{Miss})/2}$$

Receiver Operating Characteristics (ROC):





Psychological Review
1963, Vol. 70, No. 1, 61–79

A THRESHOLD THEORY FOR SIMPLE DETECTION EXPERIMENTS[1]

R. DUNCAN LUCE

University of Pennsylvania

The two-state "high" threshold model is generalized by assuming that (with low probability) the threshold may be exceeded when there is no stimulus. Existing Yes-No data (that rejected the high threshold theory) are compatible with the resulting isosensitivity (ROC) curves, namely, 2 line segments that intersect at the true threshold probabilities. The corresponding 2-alternative forced-choice curve is a 45° line through this intersection. A simple learning process is suggested to predict S's location along these curves, asymptotic means are derived, and comparisons are made with data. These asymptotic biases are coupled with the von Békésy-Stevens neural quantum model to show how the theoretical linear psychometric functions are distorted into nonsymmetric, nonlinear response curves.

Isosensitivity curves in "yes" space

Motivation, part 1: Identical d' values can be obtained from a variety of specific p(H) and p(FA).

Motivation, part 2: Hits and false alarms are separate events.

Therefore: We can generate a 2D space of hits by false alarms.