**Reference**: Master statistics & machine learning: intuition, math, code

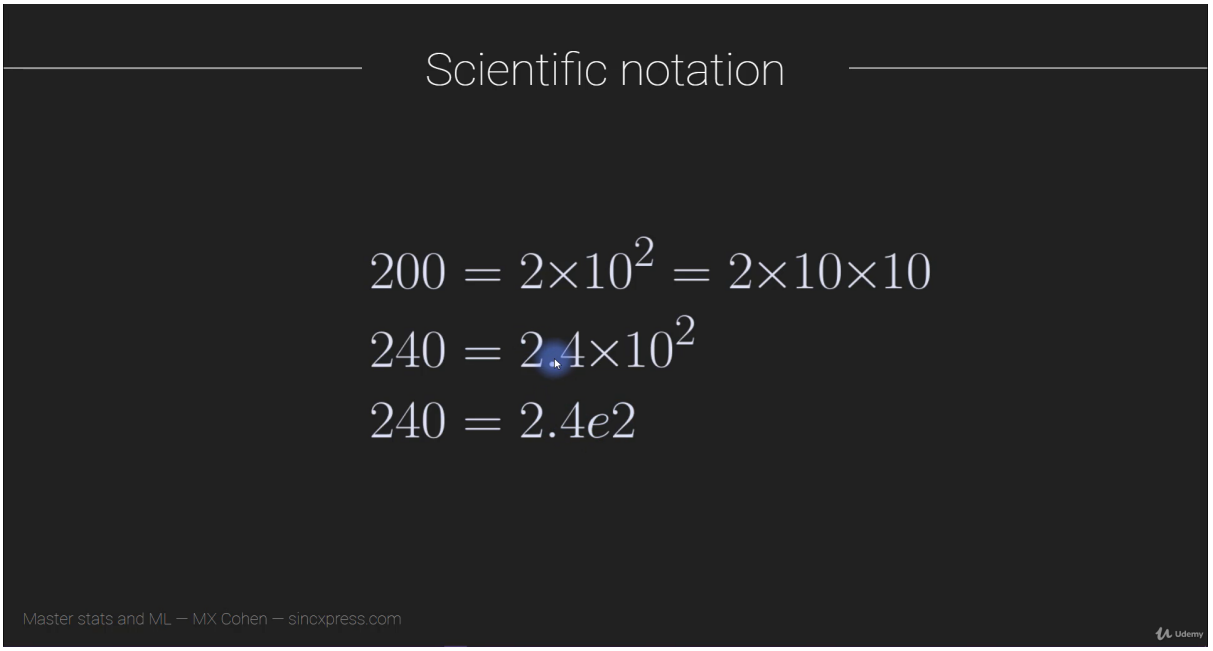
**Table of Content**

# 1. Math prerequisite

**Table of Content**

## Scientific Notation

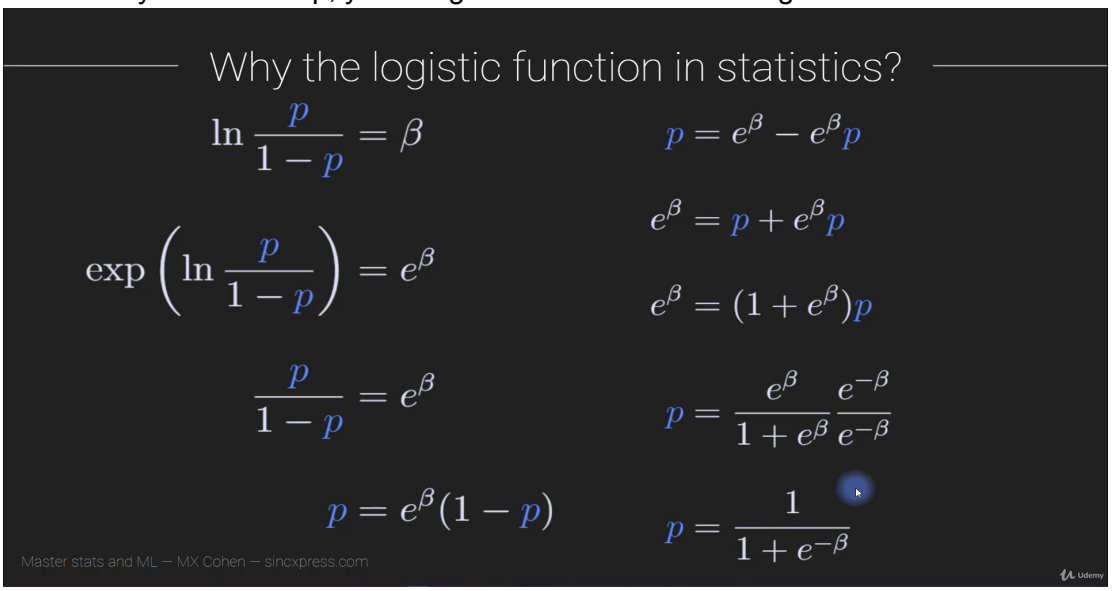
Example




Few other examples:
200 = $2*10^2$ = 2e2
0.02 = $2*10^{-2}$ = 2e-2

## What is Logistic Function

Logistic function → log(p / 1 - p) = beta
And when you solve for p, you will get the solution like the sigmoid formula.



### Rank and tied-Rank

- Say, a = {4.001, 1, 4, -10, 987}
- Rank is basically sort these numbers and their index will be rank.
- Therefore rank of a is ar = {4, 2, 3, 1, 5}
- This is called rank-transformation.
- This is non-linear transformation (directly map the index according to their position on number line)
- But the disadvantage is, consider 4 and 4.001, they are only one index apart, even though their distance is small. Also, consider 4.001 and 987, even though their distance is large enough but still one rank apart.
- This is lossy transformation and also it is non-invertible transformation. You cannot convert this rank transformation into the original set.

### How to deal with ties in rank transformation ?

- say a = {10, -10, 1, 10}
- Now what should be rank transformation ?
- We know it like this = ar = {?, 1, 2, ?}
- Now there is a tie for which 10 should take the rank 3 and which should take 4.
- The solution is, take the average of their possible ranks (i.e 3 and 4) and assign it to both.
- Therefore, ar = {3.5, 1, 2, 3.5}
- This is called tied-rank.

Some Python Basics Code: Github

# 2. What are (is_) data_

**Table of Content**

**Is the word *Data* singular or plural ?**

- Singular: Datun OR data point
- Plural: Data
- Therefore the sentence like: "Data is", is wrong because this means the word Data is singular, instead we should say, "these data are...".

**Type ( most common ) of data:**

- Numeric → Interval → Numeric Scale with meaningful intervals e.g Temp in Celcius. Also, zero degree celsius does not mean absence of temperature.
- Numeric → Ratio → Just like interval but with a meaningful value for zero. e.g height, money. Height of zero means absence of height.
- Numeric → Discrete → No arbitrary precision (integers) , e.g population of country.
- Categorical → Ordinal → We can sort the data e.g education degree (high school, secondary school, bachelors..)
- Categorical → Nominal → Non-sortable e.g movie genres (we cannot say action is greater than romance, ...)

**Why do these types matter ?**

→ Some statistical analyses can be made only to certain data types and also some visualisation methods are valid only for some data types.

**Sample vs Population:**

- **Population:** All the Data e.g Salaries of all employees in a department, all facebook groups (we can gather this data)
- **Sample:** Some amount of data (hopefully randomly sampled from the population, so that it would be sufficient representative of population)
  e.g salaries of all employees in a country (we cannot gather this data.)

**Why should we care about sample vs population ?**

- Many statistical methods are designed either for sample OR population.
- Therefore, applying a procedure to the wrong data type leads to incorrect results and incorrect interpretations!!
- But, in practice, if you have a large sample size, then your results will not get affected much.
- Most of the time, you will work with sample data and you will try to generalise your sample results for the large population.

**When should you create fake data ?**

- Validate analysis methods.
- Learn the pros and cons of analysis methods.
- Controls signals, noise and effect sizes. (you cannot control the real data but you can control fake data)
- Think more carefully and critically about data.
- It allows one to gain great intuition about the statistical methods.

# 3. Visualising Data

**Table of Content**

**Bar plots:**

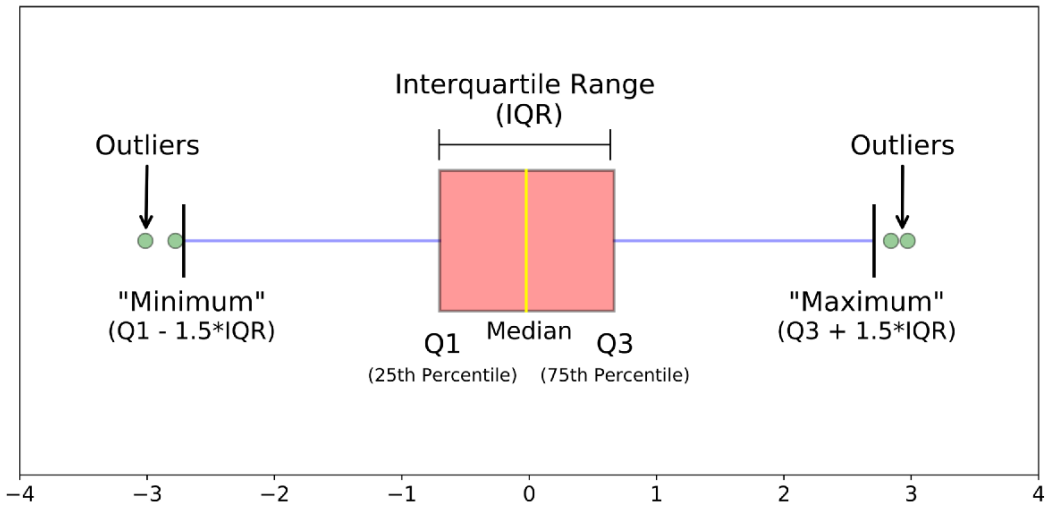
1) What kind of data can be used in the bar plot ?
   - x-axis: Nominal OR Ordinal ||| y-axis -> continuous
   - OR if you want to plot continuous data on x-axis, then you can discretize the data (like create bins in histograms)
   - Also you must take care of the number of bins, and also the width of each bin.(Technically this is histogram)
2) What are error bars in bar plots ?
   - Error bars typically show std-dev, std. error OR confidence intervals around some parameter.
   - We will understand these terms more in-depth later. But whenever you plot error bars in bar plots, you should clearly explain what they represent.

Code:

---

**Box and Whisker Plots:**

Source




Now what you can do is, say you had a colour attribute which contains black and white. You can create one box plot of another variable for colour=black and one for colour=white. And then compare two box plots for more insights.

**Code:**

---

**Histograms:**

1)
- Difference between bar plots and histogram is: Bar plots have categories on x-axis and histograms have binned continuous values on x-axis.
- Another Important question might be, can you swap the locations of the bins on x-axis ?
- If yes, then histogram OR bar plot, anything can be used.
- If not, then because the shape of distribution in histogram is important and the shape of the distribution in bar plot is not.

2) You can have two types of histograms:
- Histogram of counts (where you count on the y-axis)
- Histogram of proportions ( where you normalise the count and use the proportions on y-axis )

3)

| Histogram of counts | Histogram of proportions |
|---|---|
| 1) Can be meaningful to interpret (raw numbers) (we can say how many people have height between 80 and 90.) | 1) Can take extra effort to relate to the raw data. (It will take extra effort to get the count of people between 80 and 90 as we have proportions on y-axis) |
| 2) Can be difficult to compare across datasets. (Say person A went out and stored the height of 125 people and plot histogram and person B thought 125 is not enough, so B stored heights of 400 people and plot histogram. Now it doesn't make sense two compare the heights of two histograms, as they have different sample sizes. ) | 3) Easy to compare across datasets. (But if both person A and B convert their data into proportions (normalised) and plot the histograms, then it becomes easy to compare across datasets) |
| 3) Does not need to sum to 1 OR 100%. | 3) Sums to 1 OR 100%. |
| 4) Usually better for qualitative inspection.(i.e usually better for one dataset) | 4) Better for quantitative analysis.(better for multiple datasets.) |

4) Converting counts to proportions:
Say we had heights of 125 people, and in one bin we had count=17. Now we want to convert this count to proportion

so, proportion = ( count / N ) * 100
= ( 17 / 125 ) * 100 = 13.6%

(Now instead of using count on y-axis we will be using proportion)

**How many bins should we have in histogram ?**

→ There are general guidelines, but they are based on descriptive statistics and therefore we will answer this question later.

**When to use line plots vs histograms ?**

- For the bar plots, the line doesn't make any sense.
- For the histogram, where we have a very large number of bins, the line chart can be used.
- You can have line charts, when you want to compare continuous data for two different groups.
  e.g comparing heights of male and female using line charts, because histograms become hard to interpret.

Code:

---

## PIE CHARTS:

What kinds of data can be used for pie charts ?

- Nominal, Ordinal and discrete.
- If you want to use continuous data for pie charts, make them discrete first use and then use pie charts.

Code

---

## LINEAR SCALING VS LOGARITHMIC SCALING

1) You can scale your axis either linearly or logarithmically.
2) If you do log scale on one axis, then it is called log plot, if both of the axes are log scaled then it is called log log plot.

| Linear scaling | Log scaling |
|---|---|
| 1) Often easier to interpret. | 1) Might need an explanation for the general audience. |
| 2) Easily scaled for big, small or negative numbers. | 2) Might not work with negative numbers. |
| 3) Can obscure trends or comparisons across variables | 3) Often appropriate for physics, finance, etc. for large differences between the numbers. |
| 4) Use unless you have good reason to use log plots. | 4) Use only when required |

Code

# 4. Descriptive Statistics

**Table of Content**

### Difference Between Descriptive Statistics Vs Inferential Statistics

| DESCRIPTIVE STATISTICS | INFERENTIAL STATISTICS |
|---|---|
| 1) Describe the characteristics of the dataset like mean, median, mode, skewness, kurtosis, etc. | 1) Use the features of the dataset to make the claims about the population ike calculating p-values, T/F/chi-square values, Hypothesis testing, confidence intervals, etc. |
| 2) Note: There is no relation to the population OR generalisation to the other datasets OR comparing multiple groups. | 2) Note: The entire purpose is to relate with the population OR generalisation with the other groups. |

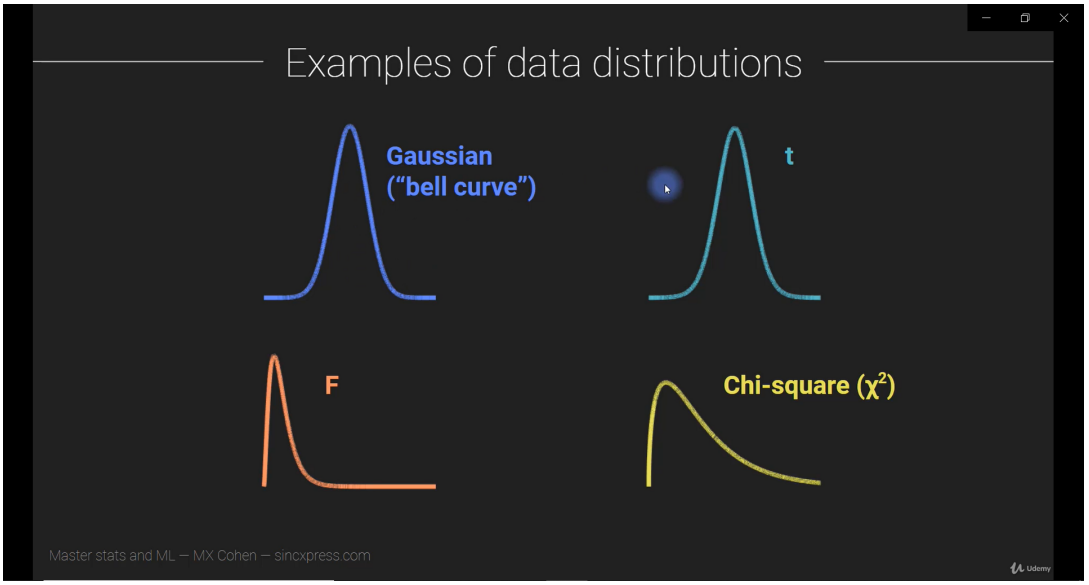
### Difference Between Accuracy, Precision And Resolution:

- **Accuracy**: The relationship between the measurement and the actual truth
- **Precision**: The certainty of each measurement (inversely related to variance)
  e.g say we are measuring the height of a person multiple times, each time we are getting the value which are almost similar, then precision is high.
  If we get the values which are very different, then the precision is low.
- **Resolution**: The number of data points per unit measurement (time, space, ...)
  Say we are measuring some data every one second and each second we are collecting 100 data points. Therefore our resolution will be 100 hz.

### Data Distributions:

- The t-distributions are for evaluating the statistical significance.
- The f-distributions are generally used in general linear models like anova, regression.
- Chi-square distribution is highly used for inferential statistics.
- Note: f-distribution and chi-square distribution has only positive values including zero.

Code



**Who cares about the distributions ?**

- Most statistical procedures are based on the assumptions about the distributions.
- Data distributions provide insights about the nature of the data.

**Beauty and simplicity of gaussian**

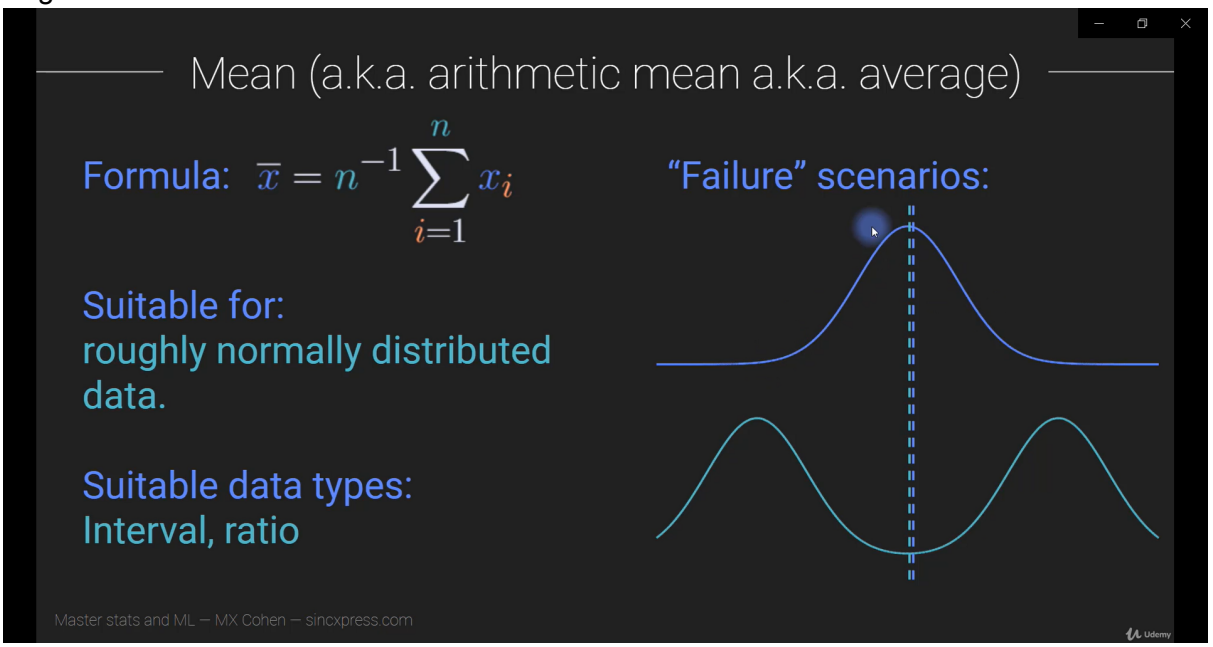
- The formula for pdf of gaussian, in its basic simple form is $e^{-x2}$, and when you plot this, this will be a nice bell curve.
- And this simple expression forms the basis for many of the things in the statistics like Central Limit Distribution.

## Central Tendency

Central tendency is different from the "Expected Value".
Expected Value is the data times its probability of occurrence.

***Mean:***

Mean as valid descriptor in first diagram and mean as inappropriate descriptor in second diagram:




Even though in the second case we can calculate the mean mathematically but it does not make sense

**Is the mean suitable for discrete data types ?**

e.g Average US family has 1.9 children.
Here 1.9 children doesn't really make sense, you have to interpret it correctly when using discrete data.

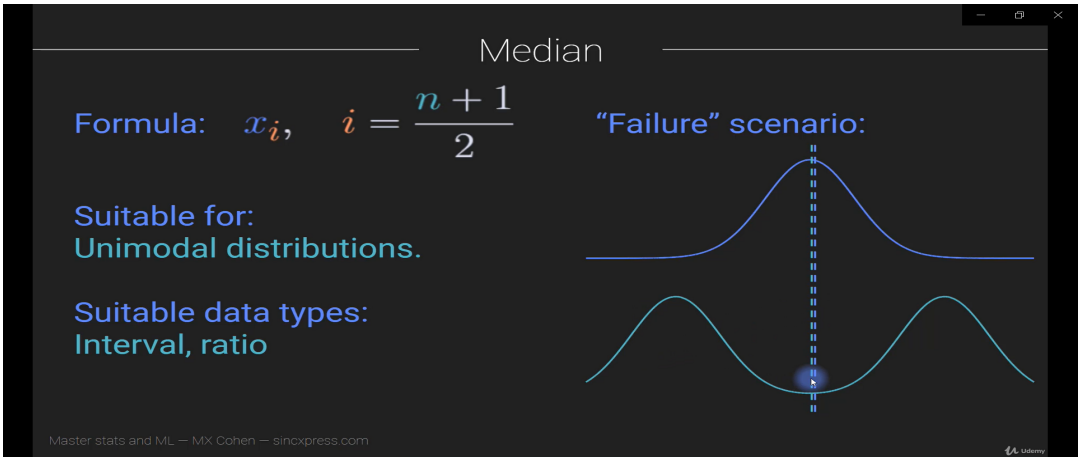
**Is the mean suitable for Ordinal data ?**

e.g Average course rating has 4.3/5
The mean for ordinal data is not suitable because the difference between 3 stars and 4 stars is not equal to the difference between 4 and 5 stars, but still 4.3 rating is informative to make a decision.

**Is the mean suitable for nominal data?**
NO. e.g take the average of red, blue and green colour.
IT DOESN'T MAKES SENSE.

Therefore to conclude:
1) Mean is best applied for interval and ratio based data.
2) Meaning of discrete and ordinal data can be useful, but carefully interpret it.
3) The mean is not appropriate for nominal data.
4) Appropriate for roughly gaussian data.

***Median:***




1) Median is suitable for unimodal distributions and interval and ratio based types.
2) Robust to outliers.

***Mode***
1) Suitable for any distribution.
2) Suitable for any datatype, but continuous data should be converted to discrete first.
3) Mostly used for nominal data.
4) Also, it is possible to have multiple modes

**Summary of Central Tendency**
1) Use a mean with symmetric distribution. (unimodal)
2) Use median with skewed distribution (unimodal)
3) Use mode with multiple modes.

**Code**

## Measures of Dispersion

### Variance

Image Source

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

1) Suitable for any distribution but they are easily interpretable when distribution is unimodal and roughly gaussian.
2) Suitable for numerical and ordinal types (but requires mean)
3) In the variance formula, we have (xi - x_bar)$^2$, this (xi - x_bar) is called mean-centering.
4) **Why mean-centering in variance ?**
   → Because we want the dispersion around the mean.
   e.g d1 = [1,2,2,2,1]
   d2 = [101, 102, 102, 102, 101](basically d1 + 100)

   The variance for both d1 and d2 is the same and that's what we want.
5) **What if we don't square the term ?**
   → The variance will ALWAYS BE ZERO
6) **Why not take an absolute difference ?**
   → Squaring emphasises large values, which is good for better optimization (continuous and differentiable), it is also the second "moment" of the distribution and also squaring has some nice properties.
7) Mean absolute difference is robust to outliers but less commonly used.
8) Mean absolute difference is also closely related to Median Absolute Difference which is used for outliers and cleaning data.
9) Dividing by n-1 is for sample variance and when the denominator is n, then it is population variance.
10) Population mean is theoretical quantity (because it does not change).
    The sample mean is empirical quantity (as it depends on how you sample) (therefore sample mean will not remain constant)
11) Suppose I roll the die 4 times, and then the mean calculated is 3.
    Now my question is, how many data points of dice do you need to know to get the mean of 3 ?
    → You only need to know 3 values, and then you can calculate the 4th observation.
    Say, you get the data from die as 1,2,4,x

$$\frac{1+2+4+x}{4} = 3$$

    Therefore, x=5.
    Hence, you need to know only N-1 values OR there are N-1 free values OR degrees of freedom.
    As the 4th output is dependent on the previous output, therefore it is not free.

**Standard Deviation**
std = sqrt(variance)

**Fano Factor**
Fano factor = variance / mean

**Coefficient of variation**
Coefficient of variation = std / mean

> **Note:** whenever you use
> *np.std(x)*, this basically calculates the population standard deviation and the same for *np.var(x)*
>
> Therefore always pass the parameter ddof (denominator degree of freedom) therefore the correct way for the sample standard deviation is
> *np.std(x, ddof=1)*

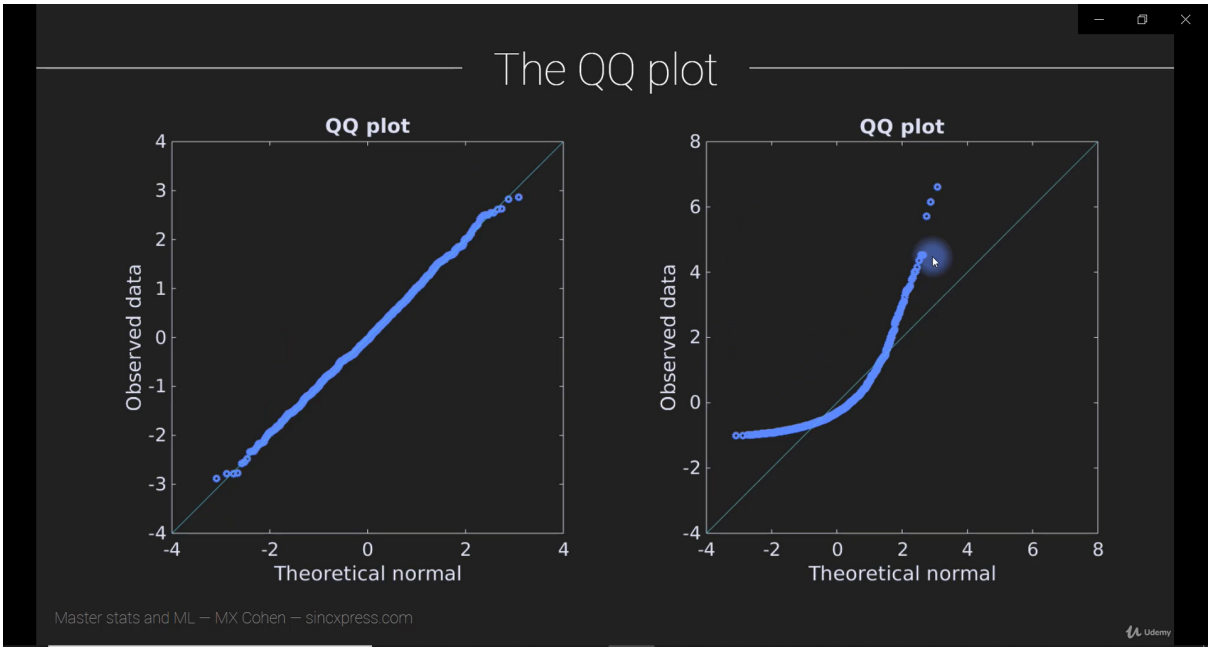
Code

**Interquartile Range**
IQR = Quartile3 - Quartile1
= Q3 - Q1
For detailed IQR: check this out
Range | Interquartile Range (IQR) | Box and whisker plot

Code

## QQ PLOTS (Quartile-Quartile)

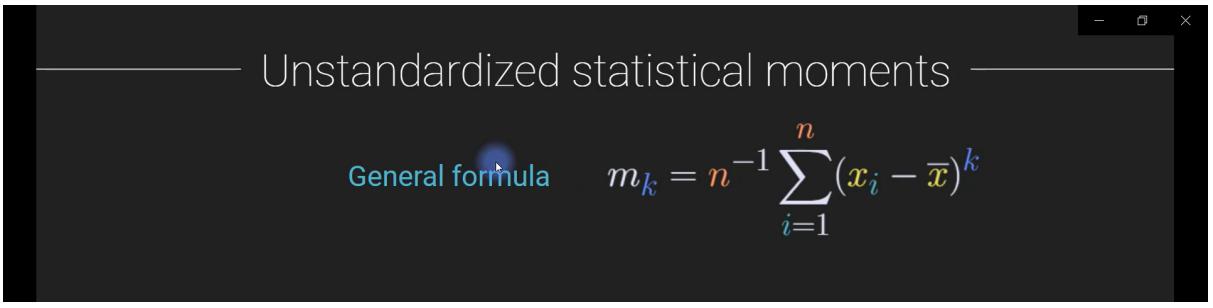
- Suppose you are given some data and I ask you to tell me whether the data came from a gaussian process?
- One way would be to plot the histogram along with the theoretical gaussian to compare.
- Another way is, you plot the observed data on y-axis and theoretical (generated) normal on x-axis and if your data is actually gaussian, then more or less the data points should lie on line (y=x) This is called QQ plot.
- This is useful for qualitative judgement.



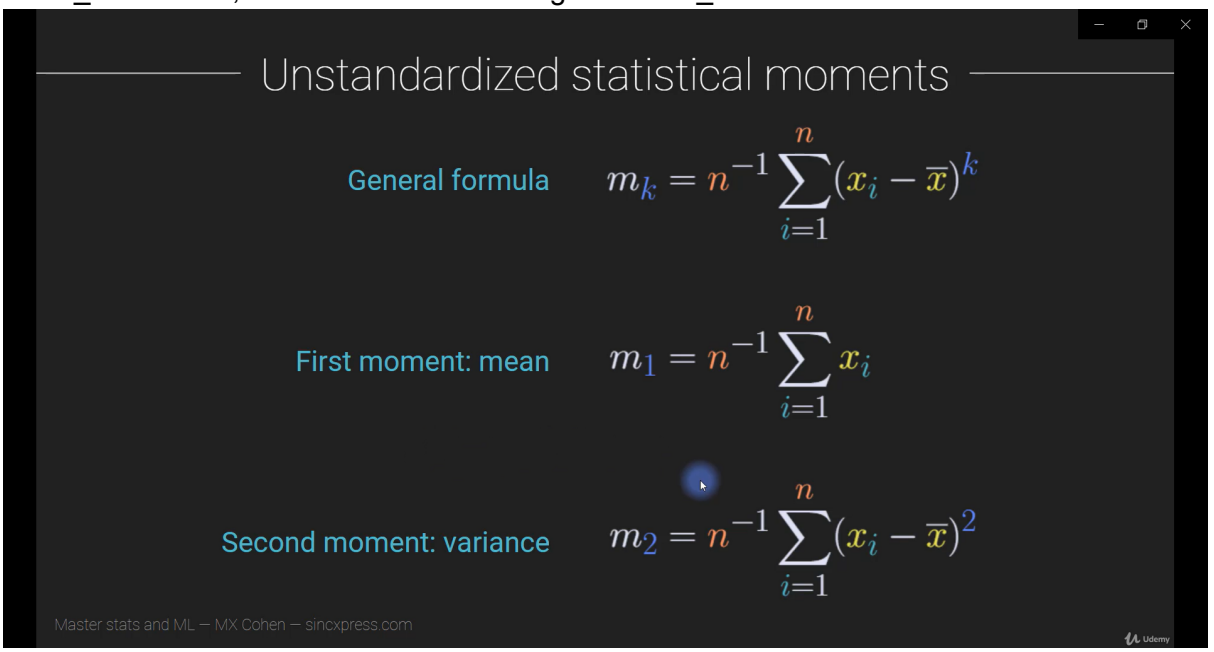

Code

## STATISTICAL MOMENTS:

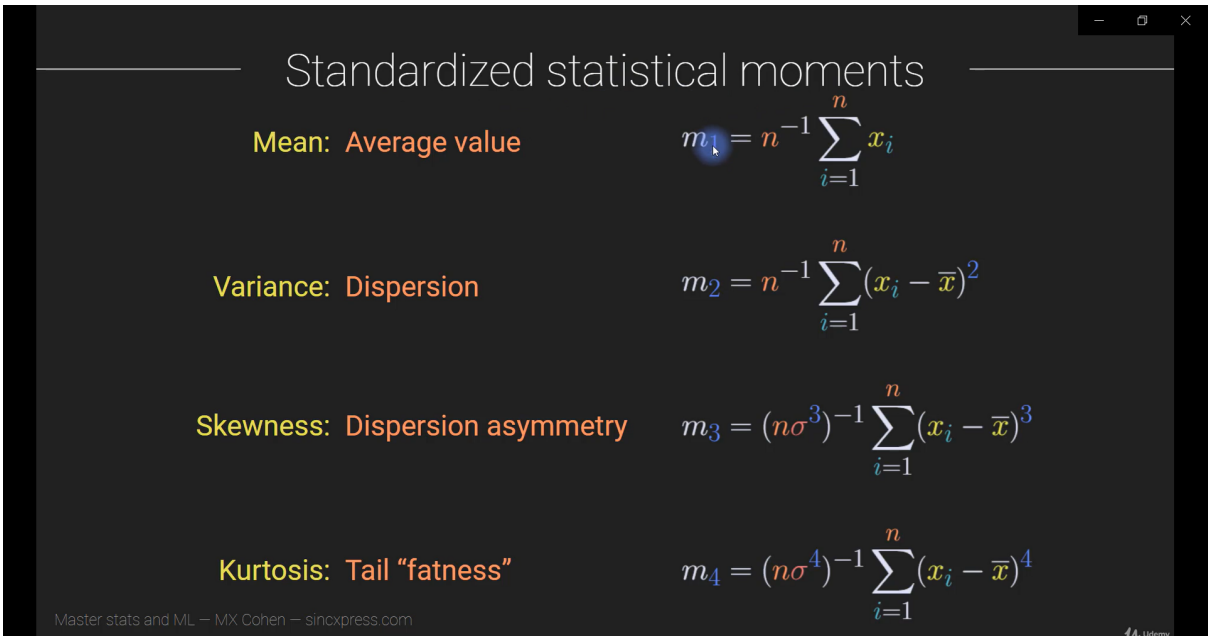
1) Statistical moments bring different kinds of descriptive statistics under a generalised framework.



2) In the first moment, if k=1, then the m1 will always be zero as we are subtracting $x_i$ with x_bar. Hence, in the first moment we ignore the x_bar.



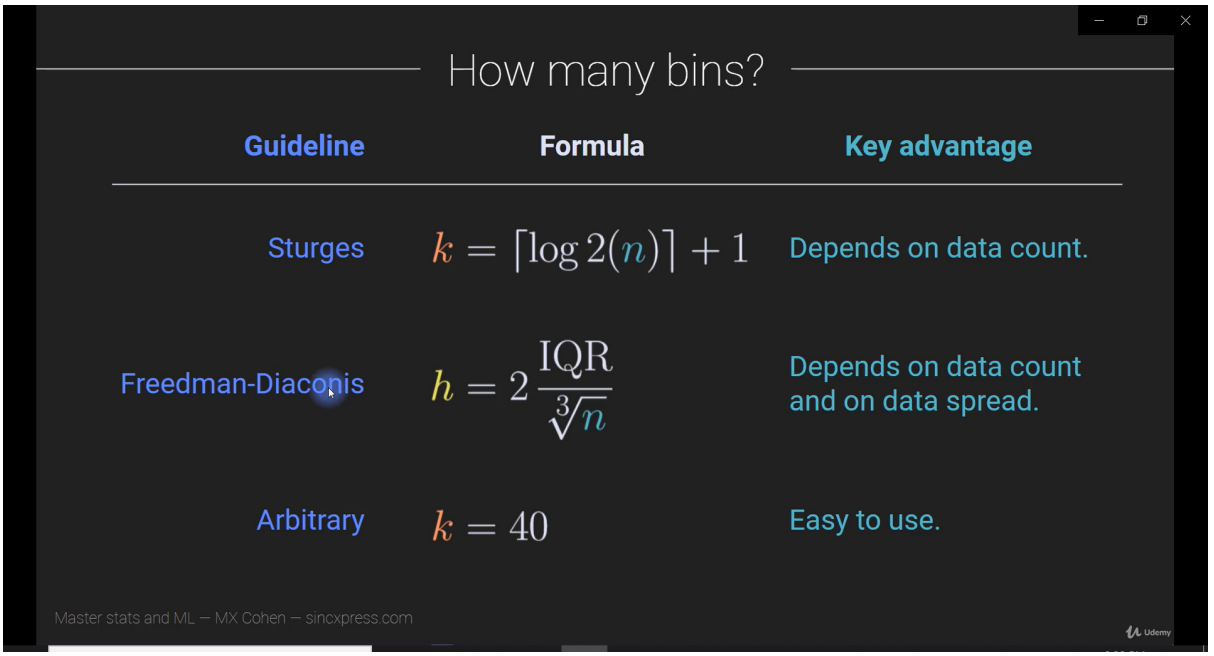

3) The second moment is variance.
4) Skewness is the third moment and fourth moment is called kurtosis.



### How many bins ?

**Remember, in the visualisation section, in the Histograms, we have to decide the number of bins. So here are the few general guidelines for deciding the number of bins.**

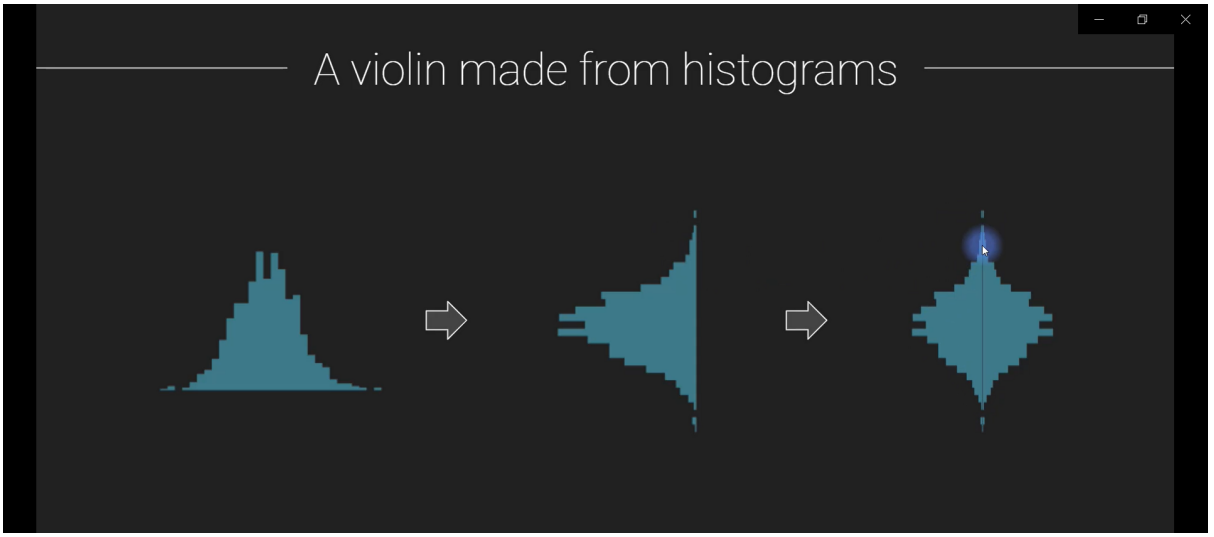
- First way is k = number of bins = *ceil*( (*max*(x) - *min*(x)) / h )
  h = width of the bin you want
- The other ways are mentioned in the image. The Freedman-Diaconis guideline is generally used as it considers data count and the dispersion as well.

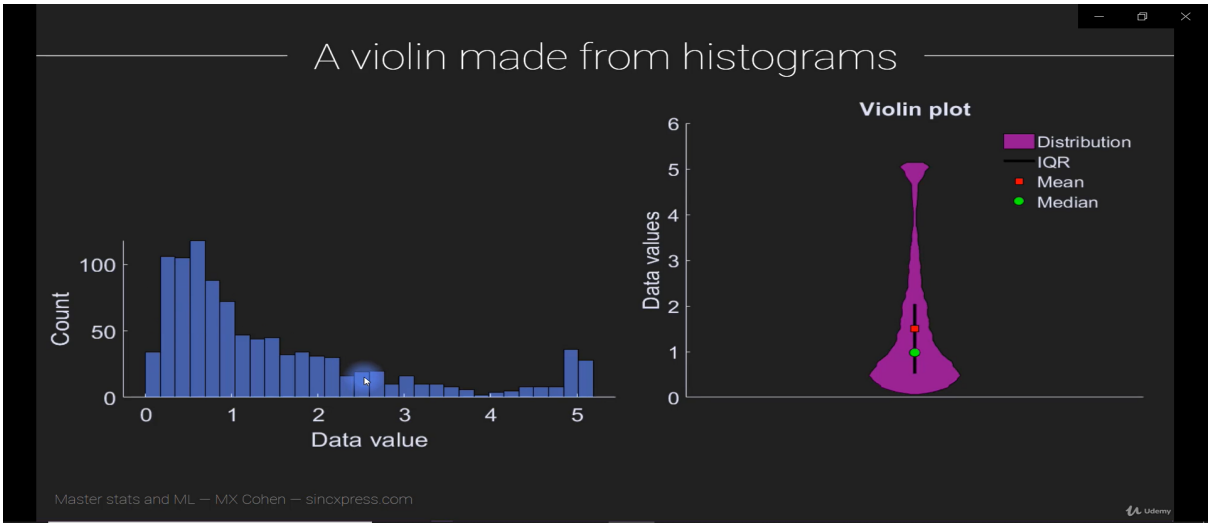



Code

### Violin Plots

- It is one of the beautiful way of illustrating the distribution of data
- Basically you start with the histogram, rotate it and then combine it with the mirror of it.
- And you can show the IQR, mean, median on the plot.






- What other thing you can do is, plot the distribution of one variable on the left side and distribution of another variable on the right and check the violin now. ( Make sure that range of the two variables should be similar otherwise you will get weird shape. )

Code

### Entropy

Check this video for GREAT explanation:
Deep Learning(CS7015): Lec 4.10 Information content, Entropy & cross entropy

- Basically, lower the probability of an event, greater the amount of information will be stored in that event. Therefore, from this we can say that information gain is inversely proportional to the probability.
- Formula for Information gain = $\log_2(1 / p) = -\log_2(p)$
  Where p = probability of an event
- Entropy is the **Expected** amount of information gained.
  entropy = 0
  for i=1 to n:
  entropy += (-p(x) * $\log_2$( p(x) ) )
  p(x) = probability of event x occurring.
- What data types can be used with this formula ?
  → Nominal, Ordinal, Discrete

  Why not interval or ratio ?
  1. The probability of any continuous variable at some specific point is zero, hence entropy is not for these types.
  2. The other option would be to convert continuous variable into discrete (convert it into histogram), then use entropy.
  3. **Note**: It is important that entropy would be affected by bin size and number of bins.
- Interpreting the entropy
  1) Higher the entropy, greater the amount of information in the data ( i.e data has high variance )
  2) Lower the entropy, lower the information in data ( i.e data has lower variance )
- How is entropy different from variance ?
  1) Entropy is non-linear and makes no assumptions about the distribution.
  2) Variance depends on the validity of the means and therefore appropriate for roughly normal data.

Code

# 5. Data Normalisation and Outliers

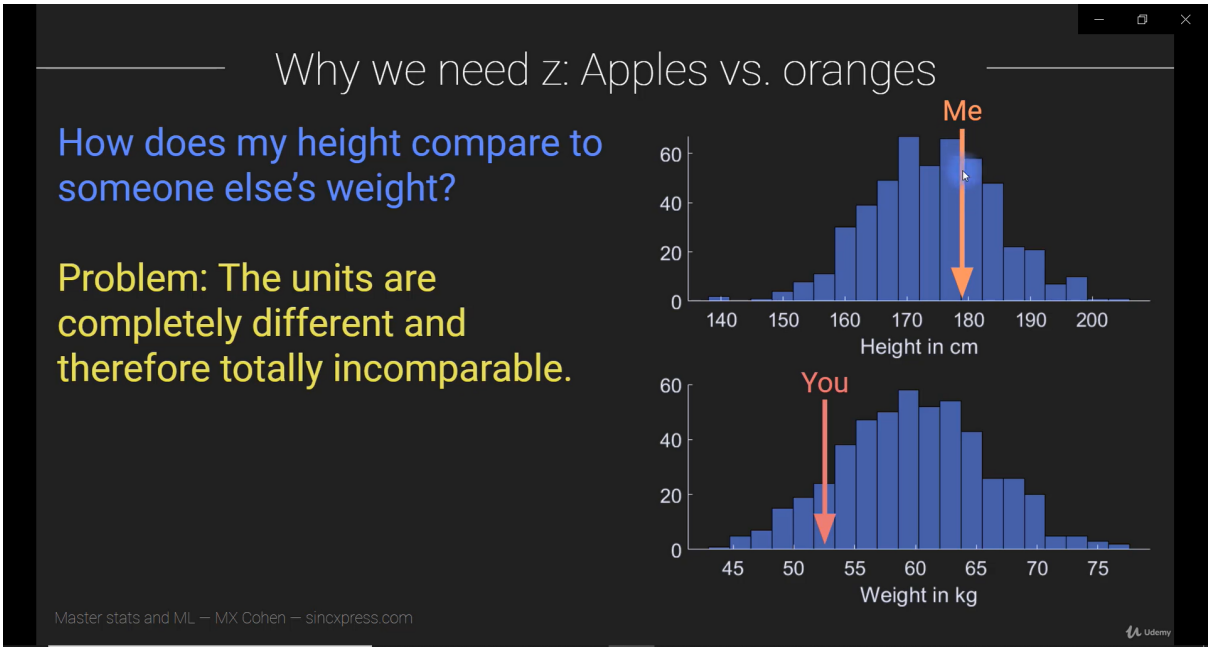
**Table of Content**

## Garbage-In Garbage-out

- This means if you start with crappy data then you get crappy results.
- Also, awesome data doesn't guarantee awesome results. (maybe the hypothesis is not correct)
- So, when possible, always clean data.
- **Note:** While cleaning the data, be patient, critical, use domain knowledge and be unbiased.

## Z-Score Normalisation

- Why do we need a z-score ?
    1. Say we want to compare my height (in cms) with your weight (in kg)
    2. The problem is, the units are completely different and therefore totally incomparable.
    3. Therefore we need a way to compare things independent of data scale.
    4. So, maybe we don't need to care about my height in cms, instead how my height compares with the population of other people's heights.
    5. So now we collect heights of 200 people and find out my height compares with the mean. (say my height is close to mean)
    6. Also we take the weight of the same 200 people and then compare my weight with the mean. (say my weight is one standard deviation less than mean)
    7. There now we can compare these two things based on how far I am from the mean.




- *This is the key insight*: The value on its own is difficult to interpret, but a value relative to its distribution is easy to interpret. Therefore we want to normalise the things into unitless distribution.
- That;s why we need z-normalisation

- Formula

$$z = \frac{x - \mu}{\sigma}$$

$\mu$ = Mean
$\sigma$ = Standard Deviation .

x - mu = mean_centering
division by std: normalisation

- See, the units get cancelled out (say u divide cm / cm ) and therefore z is the number of std. away from the mean of the distribution.
- Z-transformation shifts and stretches, but it keeps the overall shape of the distribution the same
- **What is the key assumption of z-transformation ?**
  Mean and standard deviation are APPROPRIATE descriptors of central tendency and dispersion. Basically, it means the distribution is ROUGHLY normal.
- **Does this mean we cannot use z-transformation when you have a distribution which is not normal ?**
  Well it depends on the specific application and if you are doing z-transform then you must be VERY CAREFUL WHILE INTERPRETING THEM.
- Remember that even if you have data with all +ve values, your z-scores might be -ve, therefore whether or not interpretation of -ve values is awkward, it depends on specific application.

Code

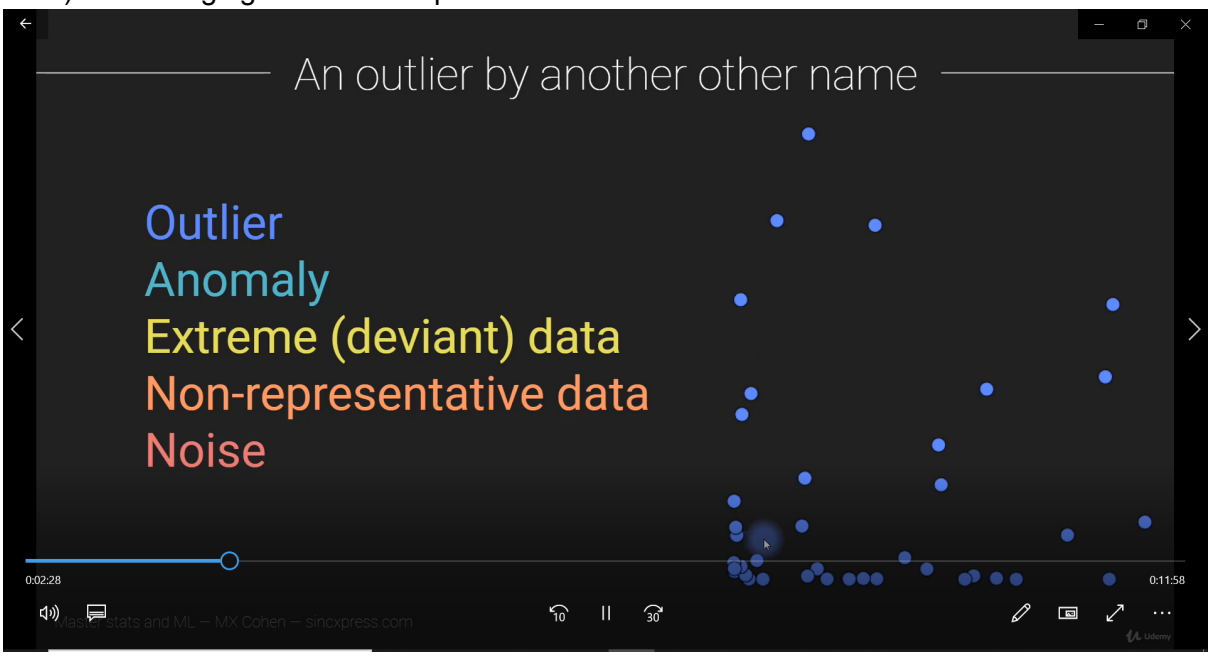
### Min-Max Scaling

- Scale the data into any arbitrary range, most commonly used range [0,1] and scaling the data in [0,1] range is called unity-normed data scale.
- The shape of the distribution is preserved.
- This is not lossy transformation, this means from the transformed data we can get back to the original value.
- Formula: xi_transformed = (xi - x_min) / (x_max - x_min), this transforms the data into [0,1] range
- Now suppose we want to scale between any arbitrary range [a,b]
  then, this would be 2 step
  First use the above formula to scale between [0,1]
  Then, scale to range between a and b →
  x_betweenAandB = a + x_transformed*(b-a)
- This is a useful transformation, especially when you are working on some analysis which requires the data into a certain range.
- You can also combine the two step formula in one like this
  x_betweenAandB = a + ( (xi - x_min) / (x_max - x_min) ) * (b - a)

Code

## Outliers

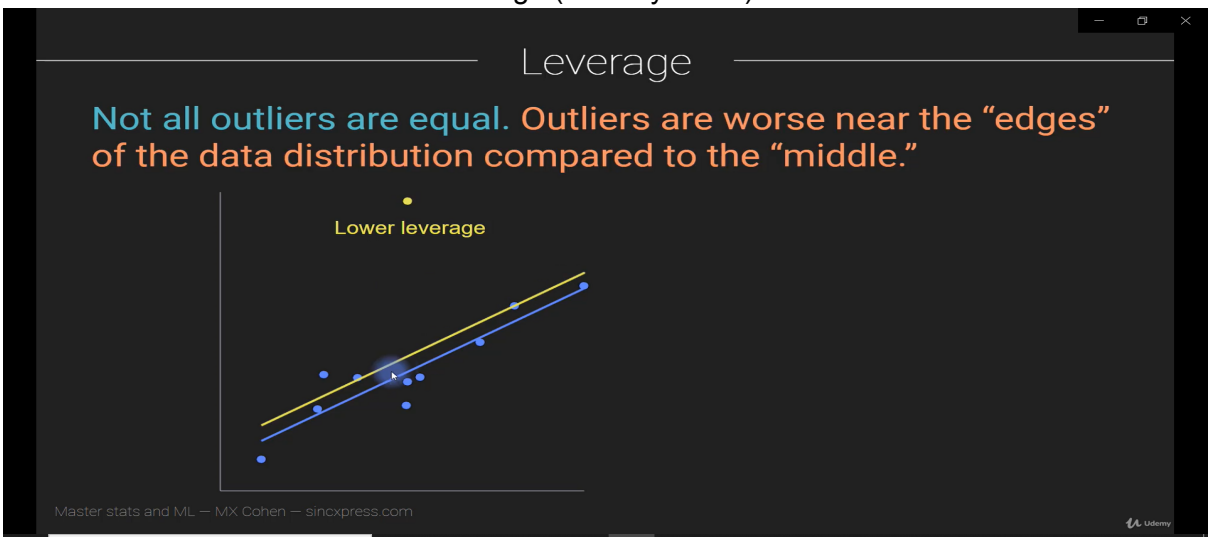
1) The other names for the outlier are: Anomaly, Extreme (deviant) data, Non-representative data, Noise
2) There are cases when outliers are easy to identify and there are cases when outliers are not easy to identify. ( check this image )
3) The things gets more complicated when we have multi-dimensional data



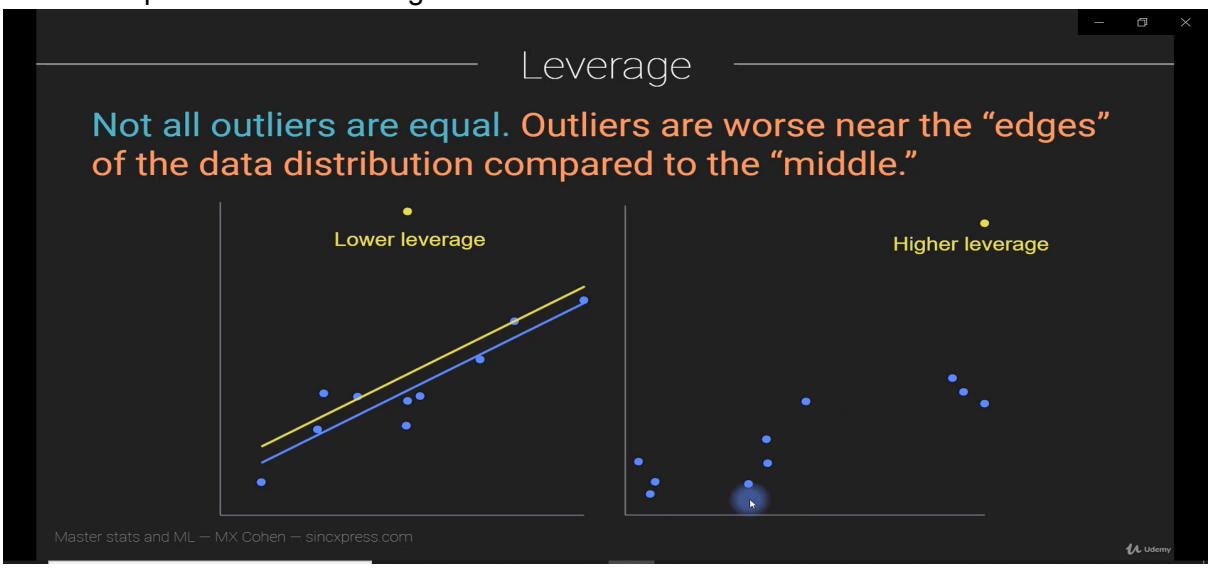

4) Where do the outliers come from ?
   → Noisy data, faulty equipment, human error, people filling the forms in a very weird way :D, OR it is natural variation in data.
5) Why are outliers bad ?
   - Most statistical analysis uses squared terms like (variance, anova, polynomials, GLM, correlation,etc.), so when you square them, the outliers become HUGE.
   - Therefore, the results of your analysis may become incorrect, OR become skewed just because of outliers.
   - Outliers can have more impact on analysis when the sample size is small.
   - Therefore, you should always worry about outliers even if you have large data and if you have small n, then you should be more worried.

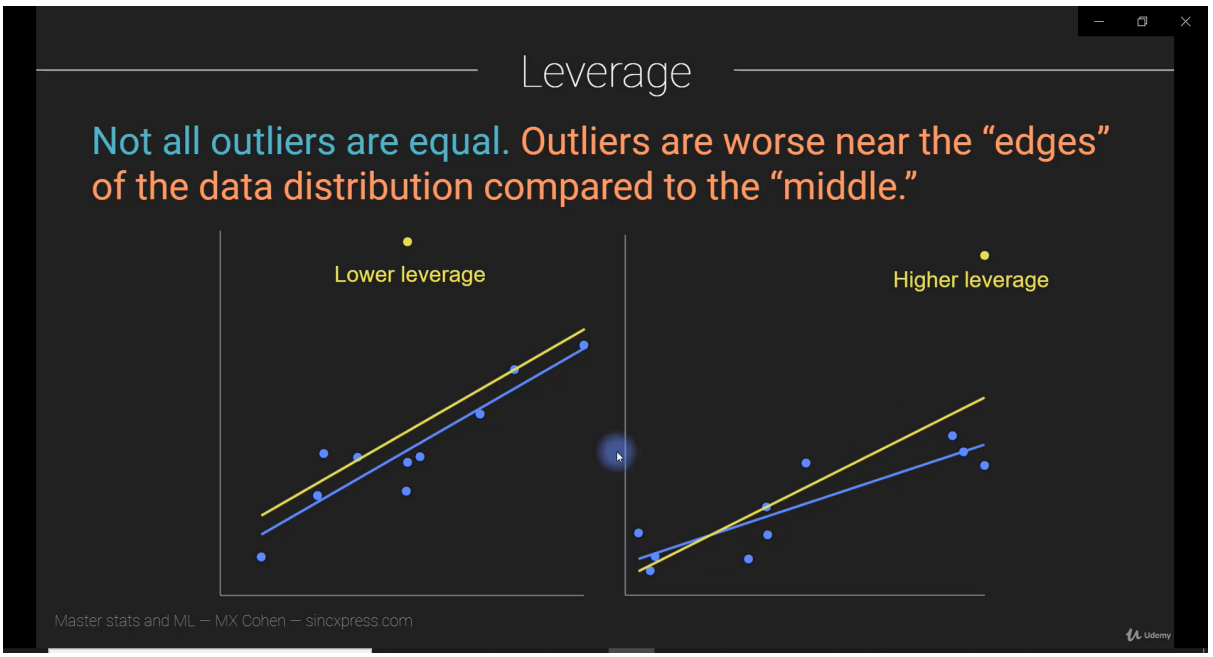
6) What is leverage ?
   - Not all outliers are equal. Outliers are worse "near the edges (or ends)" of the data distribution and outliers near the middle of distribution are less worse.

The outlier in the image does not really have much impact on the slope of best fit, just it gets shifted. Therefore the slope of the line remains almost same with or without that outlier, therefore that outlier has "Lower Leverage (can say effect)"

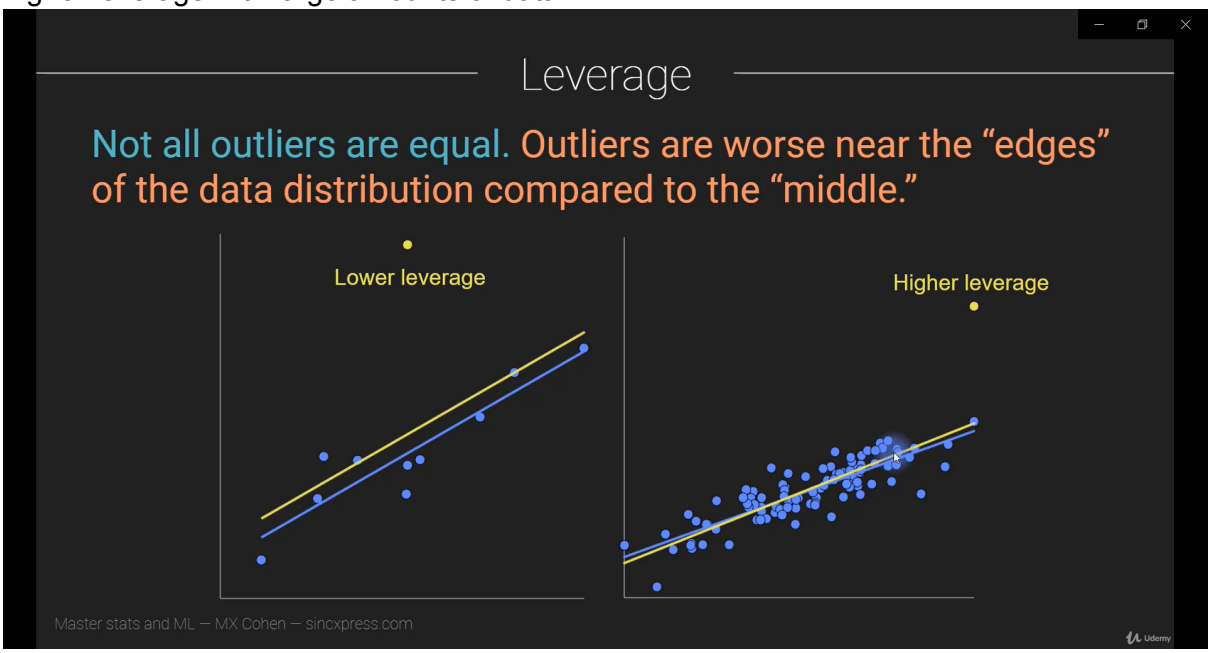



Now check the effect of the outlier with higher leverage. Also see, when the n was small, the impact on the result due to the outlier was quite high, and when we have large amount of data the impact decreases to a great extent.






Higher leverage with large amounts of data.




7) How to deal with outliers ?
   → There are two general strategies:
   1) Identify the outliers and remove them before running any analysis.
      But you are making the assumption that your outliers are just noise or otherwise invalid.
   2) You keep the outliers in the data (even though you know that can negatively impact your results) and use the robust methods for the analysis that minimises the effect of outliers on the results.

      Assumption: Outliers are unusual but valid data.

Robust methods like: non-parametric t-tests, permutation testing, spearman correlation, robust regression and iteratively weighted regression, etc

3) Therefore, you should always investigate the outliers and evaluate them.

8) **Never remove outliers without any thought**, think about why you are removing outliers and whether it is justified removing those outliers. **Outliers can be informative based on the experiment that you are running and they might be a crucial part of your research.**

### Removing Outliers using Z-Scores

1) As said before, the z-score is appropriate if your distribution is roughly gaussian, if your data is strictly non-normal then you can use a modified z-score (which we will learn later).
2) Algorithm:
    - Convert your data to z-scores.
    - z-scores with "x" standard deviation greater are considered as outliers (typically x = 3, but it can change based on use-case)
    - We can use z-scores because we are not changing the relationship between data points, we are not changing distribution shape, all we are doing is z-transformation.
3) Also, why typically x=3, and not 2.17 or 2.9 or 4, what is done in many cases, you plot the z-scores and inspect what could be the appropriate threshold for your data.
4) This algorithm can be extended with iterative z-score based outlier removal.
5) Algorithm:
    - Convert your data to z-scores
    - z-scores with "x" standard deviation greater are considered as outliers.
    - Repeat steps 1 and 2 until there are no more outliers.
6) But I am not a huge fan of this iterative method, because there can be few data points which are at the edge of distributions which may not be considered as outliers but after several iterations, they might get removed, even though it was a potential inlier.

## Code

### Modified z-transformation for non-normal distributions

1) Formula:

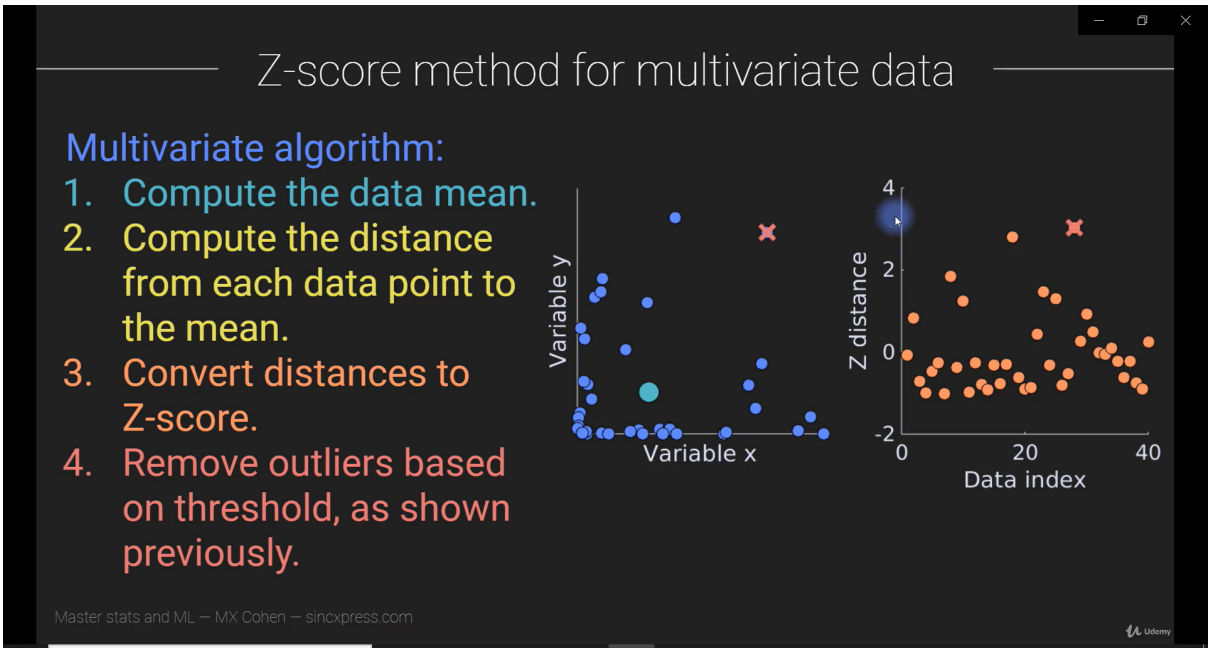
$$\text{modifiedZScore}_i = \frac{0.6745 * (x_i - \text{median}(x))}{\text{MedianAbsoluteDifference}}$$

MedianAbsoluteDifference(MAD) is basically, you take each $x_i$, subtract it with the median(x), take the absolute difference, and you do this for all xi's then,
MAD = median( | xi - median(x) | )

It is like the mean absolute difference, just instead of mean, we are taking the median.

2) This 0.6745 acts as a normalizer and it also helps us to treat this modifiedZScores as the normal zScores, and we can use this method for outlier removal, also an iterative method.
3) 0.6745 is the value on the x-axis for normal distribution when the area under the distribution is 0.75

### Multivariate Outlier Detection using z-scores

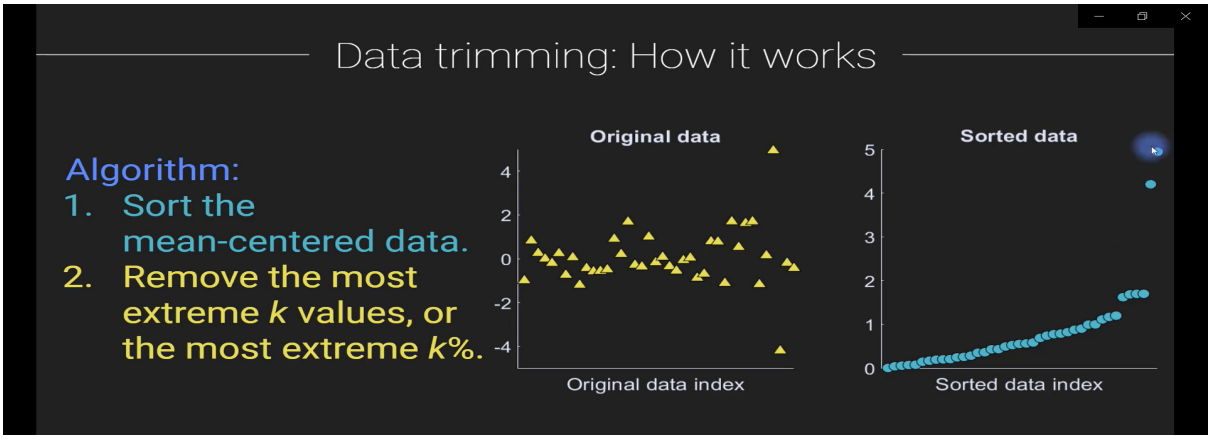



## Code

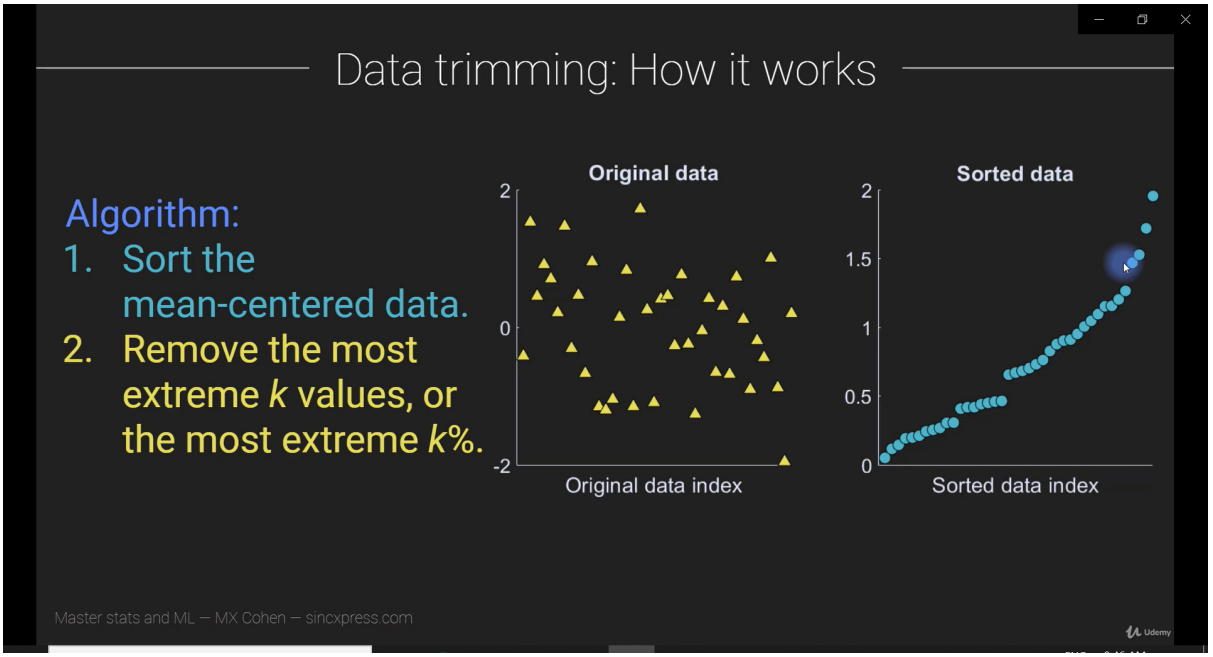
### Removing Outliers based on Data Trimming

Algorithm:

1. Mean centre the data
2. Sort the mean centred data.
3. Remove the extreme k data points OR remove the extreme k% of data points. (where k is the threshold you decide)




Check this image where it will also remove some of the non-outliers as well.




Code

**Non Parametric Solutions to the Outliers**

1. Remember that we discussed two general strategies of dealing with outliers, in the first we assumed that outliers are invalid data and therefore we removed them prior to any analysis. So till now whatever solutions that we discussed z-scores, trimming, modified z-scores follow strategy 1.
2. But there are some cases where we have to follow strategy 2, which assumes that outliers are unusual data but valid data. Therefore we have to use some robust methods to decrease the effect of unusual data on the analysis.
3. Why are these non-parametric methods less sensitive to outliers ?
   → Because they are based on median OR ranks, which are insensitive to outliers.

# 6. Probability Theory

From now on, we will cover the topics from inferential statistics, which is based on probabilities.

**Table of Content**

## Basics

**What is probability ?**

- It is a numerical description of how likely an event is to occur. It is a number between 0 and 1, where 0 means impossibility and 1 means certainty.

**Now I will give you a statement and its three interpretations, tell which ones are right or wrong**

Statement: There is a 20% probability of rain today.
Interpretation 1: It will rain for 20% of the day i.e 0.2*24 = 4.8 hours
Interpretation 2: There is a 1 in 5 chance that it will rain today.
Interpretation 3: We can be 20% confident that it will rain today.

- Now interpretation 1 is incorrect, interpretation 2 is correct, tell me why interpretation 3 is incorrect ?
- Because 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%]
- Why is interpretation 2 correct ?
  If we could repeat the exact same conditions today a very large number of times (approaching infinity), then it would rain on 20% of those days.

**When do we need probability ?**

- We need probability when there is uncertainty about the outcome of the event.
- Say, for example I say, tomorrow the sun will rise. Now in this statement, we know that an event WITH 100% CERTAINTY is going to occur. So, we don't need probability here.
- Say, some medical test stated that you have cancer, then you want to know what are the chances that you have cancer GIVEN THAT the test is positive. Here we will need probability.

Code

## Proportion vs Probability

1) Probability: Likelihood of an event to occur.
   Proportion: The fraction of a whole.
2) I will give one statement and three phrases, tell which statements are correct or incorrect.
   **Statement**: I spend a total of 5.1 minutes each day brushing my teeth out of a total of 17 hours (1020 minutes)(when I am not sleeping).
   **Phrase 1**: The proportion of my waking day, spent brushing my teeth is 5.1/1020 = 5%
   **Phrase 2:** The probability that a randomly selected minute of my day involves teeth brushing = 5%
   **Phrase 3**: The probability that I will brush my teeth during the day is 1.

   All the three phrases are CORRECT.
   *Therefore the probability and proportion might be different depending on how the question OR how the statement is phrased*.

3) **Statement**: Say I toss the fair coin 10 times, and we get 6 heads and 4 tails.
   **What is the probability of getting heads ?**

→ The probability = 0.5, because the probability of getting heads will always remain the same irrespective of the number of heads we get.

**What is the proportion of getting heads?**
→ Proportion = 0.6, because remember the proportion is basically the fraction as a whole.

**Say I used the same statement, then what is probability, out of 10 flips, I select any random flip and its outcome is heads ?**
→ 0.6

### Valid Data types for probability

- Remember that we have 5 types of data: interval, ratio, discrete, nominal, ordinal
- Now the probability is valid for discrete, nominal, ordinal and not for interval and ratio based data.
- What you have to do is convert them to discrete then calculate probability.
- e.g, what is probability of height of randomly selected men = 122.64982913874323
  You cannot calculate the height of men with such precision and chances = 0
  But we can calculate p(height>121 and height<123) and you can vary these intervals.

**Also the probability is valid if the data have mutually exclusive outcomes.**
For e.g when you toss a coin, you will get either heads or tails but not both at the same time.
When you roll a die, you will get a number from 1 to 6, and not 2 different numbers at the same time.

### Probability and Odds

Image source
Odds is basically the ratio of events not occuring divided by event occurring.
*It is commonly called the odds ratio.*

$$\text{Odds} = \frac{\text{probability of the event}}{\text{1 - probability of event}} = \frac{P}{1-P}$$

OR

$$P = \frac{\text{Odds}}{\text{1+ Odds}}$$

e.g what are the odds of drawing the king from the randomly shuffled deck of cards.

$$r = \frac{4 / 52}{48 / 52} = 1:12$$

i.e the odds of drawing a king from a deck of cards is 1:12

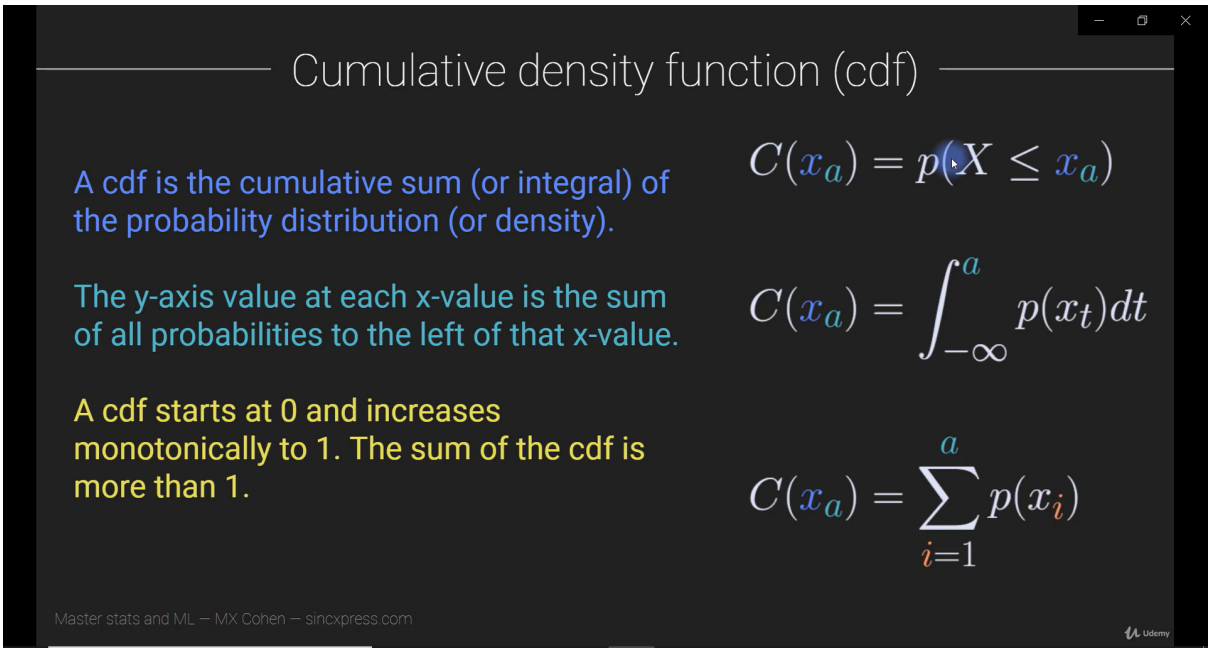
### Probability mass and density functions:

- **PMF**: For discrete events, generally bar plots or histograms are used for visualisation.
- **PDF**: for continuous events, line plot, with values on x-axis and their probabilities will be area under the curve.

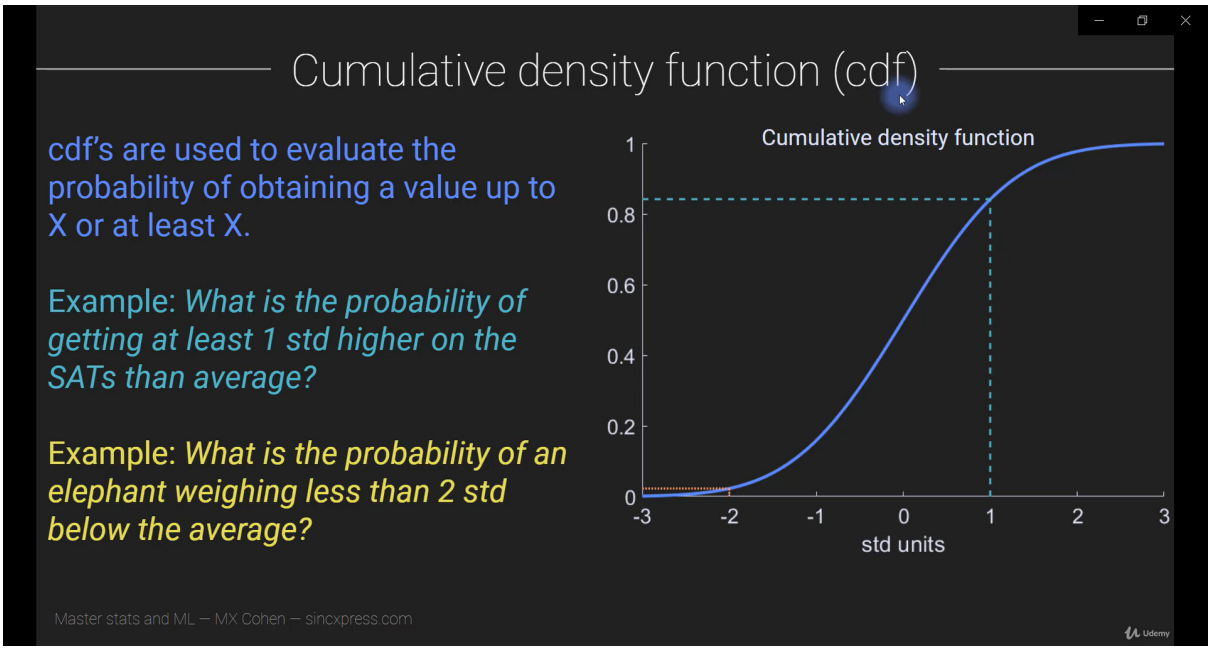
- In statistics, applied statistics, ML, medicine, etc. we generally try to discretize things and hence we generally use pmf.
- Because, in computers you really cannot represent the analog signal with high precision, hence we try to estimate the true probability density function using pmf.

Code

**CDF**




**Cumulative Distribution Function** : used for pmf
**Cumulative Density Function** : used for pdf



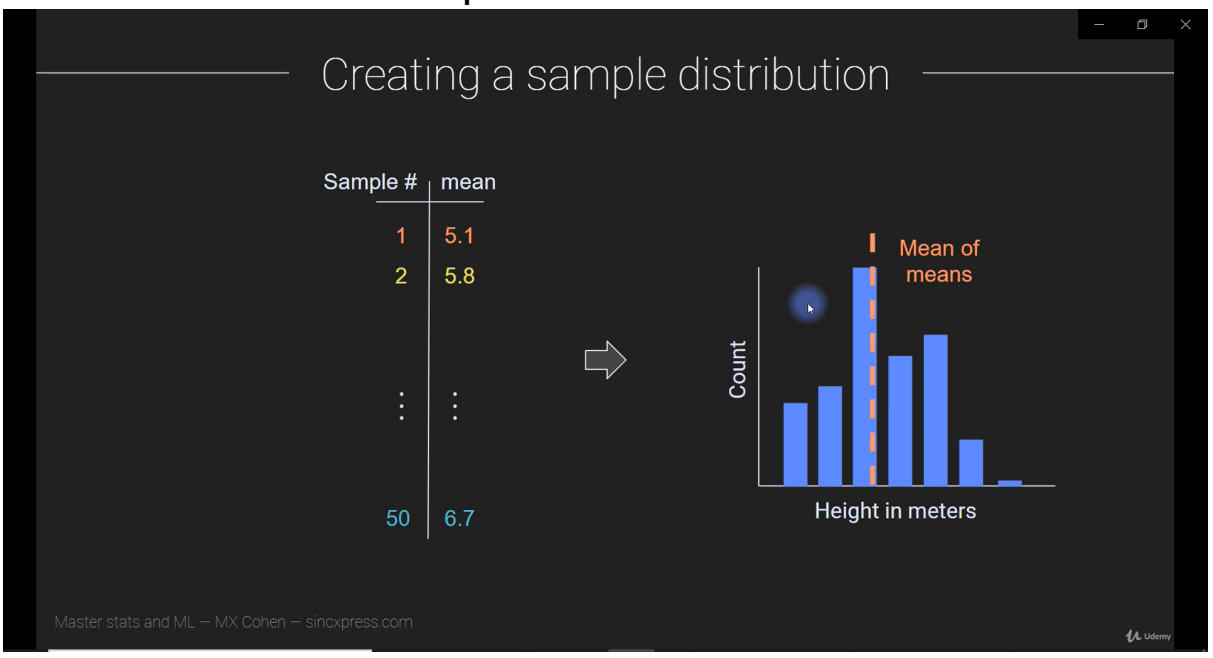

The answer for first question:
0.15: p(X >= 1) = 1 - p(X < 1) = 1 - 0.85

The answer for second question:
p(x <= -2) = 0.03

Code

### Creating Sample Estimate Distributions ( V.V. Imp )

- It is very important for inferential statistics, for generalising from sample to population, which we often want to do.
- Let's start by asking one question
  **How tall are the giraffes ?**
  The statistical way of asking this question is
  What is the population parameter of giraffe height ?
  IT IS IMPOSSIBLE TO ANSWER THIS QUESTION!!!
- WHY?
  → Because we cannot measure the height of ALL the giraffes.
- Instead what we can do is, we can measure the heights of samples of giraffes.
  1. So we start by randomly sampling giraffes from the whole population of giraffes and start measuring their heights.
  2. Say we collected the heights of 100 giraffes and plotted their histogram along with the vertical line at mean.
  3. IMPORTANT: This data distribution (HISTOGRAM) produces one sample estimate (mean)
  4. Now we again sample the giraffes height (sampling with replacement i.e giraffe in the first sampling might be present in the second sampling also)
  5. Then again plot the histogram and calculate the mean. (This mean might be different from the first one because of the sampling variability and so on)
  6. We repeat this experiment N times, and calculate and N means.
  7. Now plot these N means and it is called **Sample Distribution** and the mean of **Sample Distribution** is called **Mean of the Means.**




**NOTE**: "mean" is just one example for parameter estimate. (it could be var, std, ..)

Now the basics are clear, let's go further and ask a new question.
**Are giraffes taller than cats ?**

- Remember that when we are quite certain about the event, then there is no need for probability. Here we dont need probability, but let's say we need
- probability.
- Now, we sample giraffes and sample cats ( say sample size is 40 ), say the sample mean of giraffe is gi and of cats is ci, we calculate their difference and store it.

```
meanDiffs = []
for i = 1 to N:
    gi = sample(giraffe).mean()
    ci = sample(cats).mean()
    meanDiffs.append(gi - ci)
```

  **Note:** you can do ci - gi, but this will generate -ve difference, which can lead to harder interpretation OR you can simply use absolute value, since you are interested only in magnitude and answering that question.
- Even though it is possible that in the real world there exists some difference in heights of giraffes and cats, due to sampling variability ,and other stuff, the difference might be zero.

*In the first question, we were trying to estimate the population parameter but in the second question, we are trying to ask a question involving two groups of population. We will continue this discussion later, but the basics are clear.*

### Monte Carlo

- **Monte Carlo methods:** Solve really hard problems by sampling the solution space instead of actually solving the complete problem.
  i.e we are determining what the possible solution could be based on the sampling of solution space.
- e.g suppose there is a very very hard integral to calculate, so instead of solving it, we use monte carlo methods.
- It is often used in physics, statistics, deep learning, even in pure maths.
- There are a whole bunch of monte carlo methods.
- "**Monte Carlo Sampling**" is basically the same thing as randomly sampling from the population space to estimate an unknown population parameter.
- "**Markov Chain Monte Carlo**" It is variant of Monte Carlo Sampling, basically the way you pick sample depends on the previous sample
  Check this more in-depth article:
  https://machinelearningmastery.com/markov-chain-monte-carlo-for-probability/

### Sampling Variability, Noise And Annoyances:

- How tall is the Indian (the person who lives in India) ?
  Don't worry about the actual answer.
  The actual question is: How do we go about finding the answer?
- What we do is, the first person we meet, we measure their height and claim it as the average height of an Indian (160 cm), but we met another person whose height is 192cm, that's way above average we thought, and we keep measuring heights.
- But the internet says 183.2cm (population average). The answer on the internet must be true, but how can we have people with different heights ? You know the answer, it is due to variability. Not everyone has the same height.

- Therefore "**Sampling Variability**" means Different samples from the same population can have the different values of the same measurement (parameter like mean).
- **Implication of the sampling variability**: A single measurement may be an unreliable estimate of a population parameter.

**Where does this variability come from ?**

- "**Natural variation**": Often seen in biology (e.g weight, height) and physics (e.g earthquake magnitude, no. of stars, etc..)
- Also, sometimes the variability might be less (there is very less variation when you calculate the size of a proton multiple times)
- Therefore, sometimes variation might be large or small depending on the use case.
- "**Measurement noise**": Imperfect instruments, less precise instruments (say you are trying to measure weight in micrograms, but your system only measures in grams)
- "**Complex System**": Say I am measuring heights ignoring the age of a person, there might be very high variability, but if we keep the age as fixed, then your variability will be less.
- "**Stochasticity**"(Randomness): The universe is wild and unpredictable place (e.g photons hitting the lens of the camera)
- Some of the sources you can control like the quality of instruments, precision, etc.

**What to do about sampling variability ?**

- Taking many samples !!! Large number of samples will lead to better approximation of population parameters.

**The main outcome is, the mean of the sample means approaches closer to the population mean and it leads to two important concepts in statistics:**
**Law of Large Numbers, Central Limit Theorem.**

## Code

### Expected Value

- Average = sum(x) / n
- ExpectedValue E[X] = sum(xi*pi), where pi = probability of value xi occuring.
- When the average and Expected value are equal ?
  → When all the pi's = 1/n, i.e each and every data point is equally likely to occur.
  ALSO, when drawing large and repeated REPRESENTATIVE random samples from the population.
- Average is an empirical sample estimate based on the finite data.
- Expectation value is the expected average in the population OR from a very very large number of samples(approaching infinity).

## Expected value vs. average: example

"Illegal" die

Side 1: $p$=¼
Side 2: $p$=¼
Side 3: $p$=⅛
Side 4: $p$=⅛
Side 5: $p$=⅛
Side 6: $p$=⅛

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

$$= 1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6$$

$$= \frac{1}{4} + \frac{2}{4} + \frac{3}{8} + \frac{4}{8} + \frac{5}{8} + \frac{6}{8}$$

$$= 3$$



## Expected value vs. average: example

"Illegal" die

Side 1: $p$=¼
Side 2: $p$=¼
Side 3: $p$=⅛
Side 4: $p$=⅛
Side 5: $p$=⅛
Side 6: $p$=⅛

8 random rolls of the loaded die:

1 3 4 4 4 3 2 5

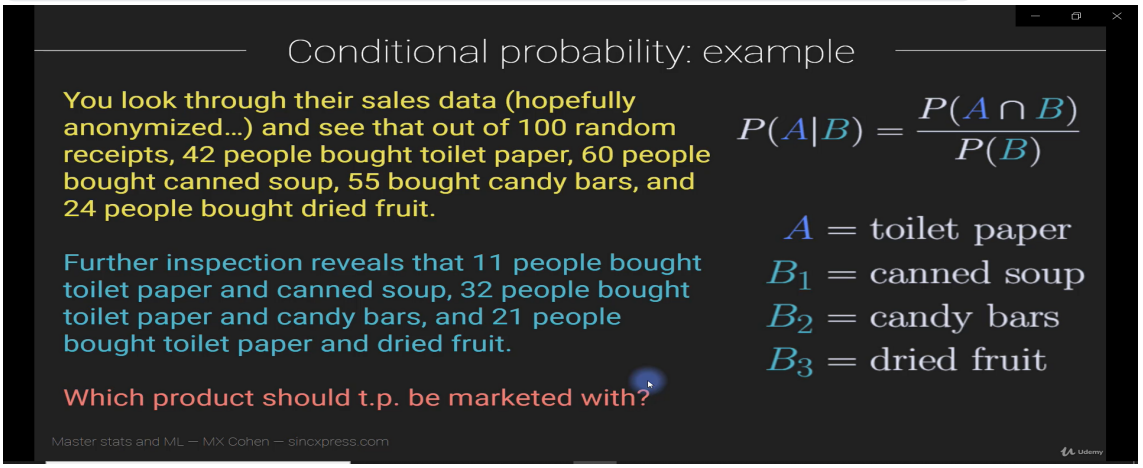
Average: 3.25
Expected value: 3

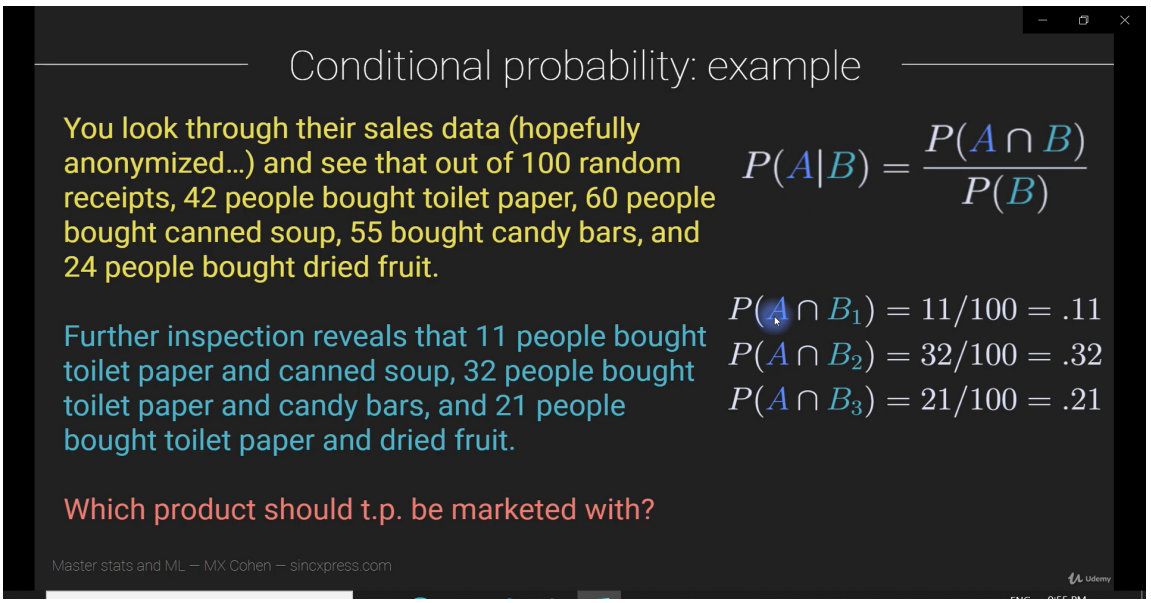
Master stats and ML — MX Cohen — sincxpress.com

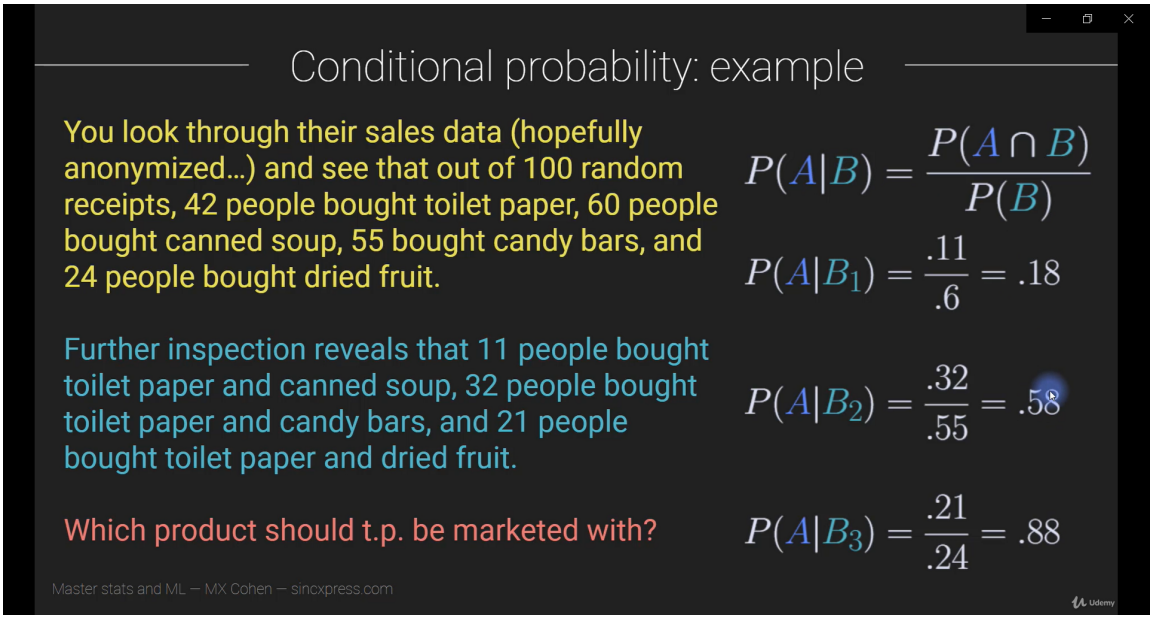
## Conditional Probability

For quick explanation:

Deep Learning Part - II (CS7015): Lec 16.0 Recap of Probability Theory Example



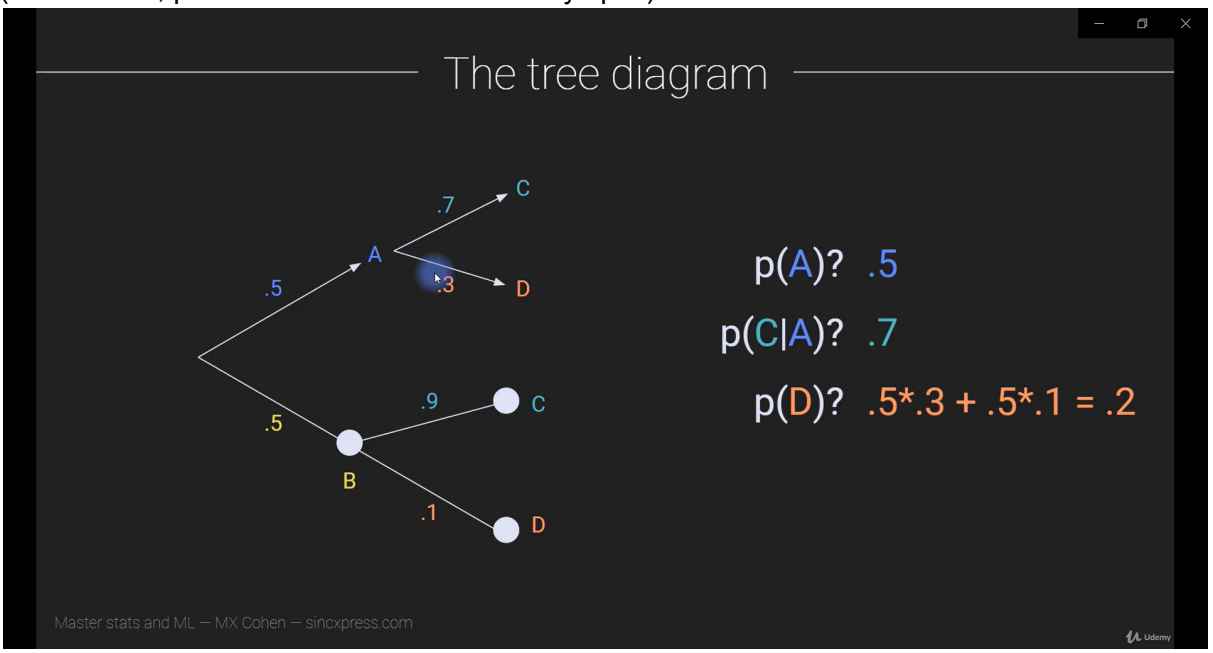



- P(A| B1) = 0.18 means probability of people buying toilet paper given that they have bought canned soup is 18%
- P(A| B2) = 0.58 means probability of people buying toilet paper given that they have bought candy bars is 58%
- P(A| B3) = 0.88 means probability of people buying toilet paper given that they have bought dried fruits is 88%

**Be cautious, not to confuse the correlation with causation, therefore, it should NOT be interpreted like buying dried fruit causes people to buy toilet paper, what does p(A | B3) means is, if someone buys dried fruits then the probability of that person buying toilet paper is 88%**

- These types of questions can get a little bit tricky and it depends on how you phrase the question and how you set up the problem.
- For example, the supermarket should be advertising the candy bars with toilet paper because it should be clearly seen, out of 100, 32 people bought candy bars and toilet paper and only 21 people buy toilet paper and dried fruits.
- But if the question is, which product will maximise return on investment, then we can clearly say that people who buy dried fruits are 88% likely to also buy toilet paper.

**Tree Diagram for Visualising Conditional Probabilities**
( Notice that, probabilities sums to 1 at every split )




Code

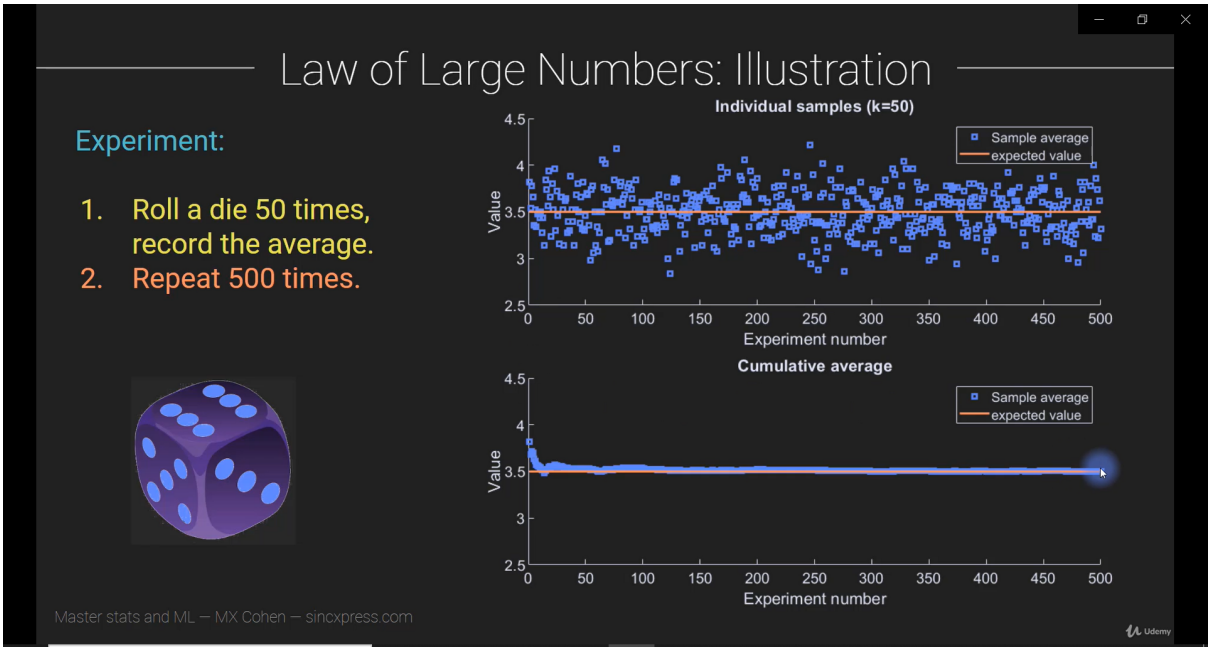
### Law of Large Numbers

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

  lim (n → inf) p( |x_bar - mu| > epsilon ) = 0
  n = number of times the experiment is repeated.
  x_bar = average of the n samples means.
  mu = population mean
  epsilon = very small number

We are making an absolute difference, as we are interested in how far we are and not whether we are underestimating or overestimating.

- As any one sample (or one experiment) is sensitive to sampling variability, noise and other sources of variation.
- But if we repeat the experiment many many times, our estimate will be really better. (independent replications of experiment)
- **Important question:** How large is large ? what does n (number of times we perform experiment) need to be to get a reasonable estimate.
  → Unfortunately, it depends on many factors like effect size, noise, system complexity, etc
- We will talk more about this in the statistical power and sample sizes section.
- If you also plot the sample means which you collected after n experiments, then you will find that it will be roughly gaussian.( Continued in Central Limit Theorem )




Code

### Central Limit Theorem

- The DISTRIBUTION of the sample means (you can also take sample variance, entropy, or any other statistic) approaches the gaussian distribution, irrespective of the shape of the population distribution.
- Note: In law of large numbers, we calculate the average of sample means, but here we are talking about the distribution of the sample statistic.
- Why is the CLT so important ?
  1) Because many statistical analyses rely on the assumption of normality (or roughly gaussian).
  2) Gaussian distributions are easy to parametrize (mean, variance, skew, kurtosis) and compute confidence intervals.
  3) **Second Interpretation of CLT:** If we take Random samples from independent variables, their mixture will tend towards a normal distribution, even if the variables are non-normally distributed.
  4) **Independent Component Analysis** (ICA) is based on the assumption that the multivariate signals which are non-gaussian distributed, while random mixtures of signals are gaussian. (This is based on second interpretation of CLT)

Code

# 7. Hypothesis Testing

**Table of Content**

**Stats Lingo**

- DV: Dependent variable, the variable which you are trying to explain.
  IV: Independent Variables, the explanatory variables, the variables that you HOPE will explain the variance in Dependent Variable.
- You generally manipulate the IVs and have control over them in the experiment and sometimes you don't have control over IV, you observe the IV from the real world and you HOPE that it will be able to explain the DV.
- Examples (Identify IV and DV):
  1) Effects of soil moisture on plant growth: Quite easy, IV = soil moisture, DV = plant growth
  2) Effects of time spent on facebook on irritability: This makes us think, IV = time spent on facebook, DV = irritability
  Because we can take the other way as well, like the more you are irritated, the more time you spend on facebook. In this case, the IV = amount of irritation, DV = time spent on facebook.

  | |
  |---|
  | • Whenever you are thinking about which is IV and which is DV, you should think about, "What is the likely flow of causality in the world".<br>• In statistics, you have to be very careful when discussing casualty, just because we find an effect, it doesn't mean there is a casual relationship between them like in the second example. |

  3) Relationship between people who spend more money on clothes and go to bars frequently.
  The *first case* can be: the people who buy new clothes tend to go to bars more frequently to show off their clothes.
  i.e IV = money spent on new clothes
  DV = frequency of going in bars.

  The *other case* can be people who frequently go to bars, want to look more presentable so they end up spending more money on clothes.
  i.e IV = frequency of going into bars
  DV = money spent on clothes.

  Therefore, sometimes it can go either way, your thoughts, use case, assumptions and how you want to interpret the data.
  But sometimes it is clear, which is IV and DV.

## Model-Fitting In Statistics

**1) What is modelling ?**

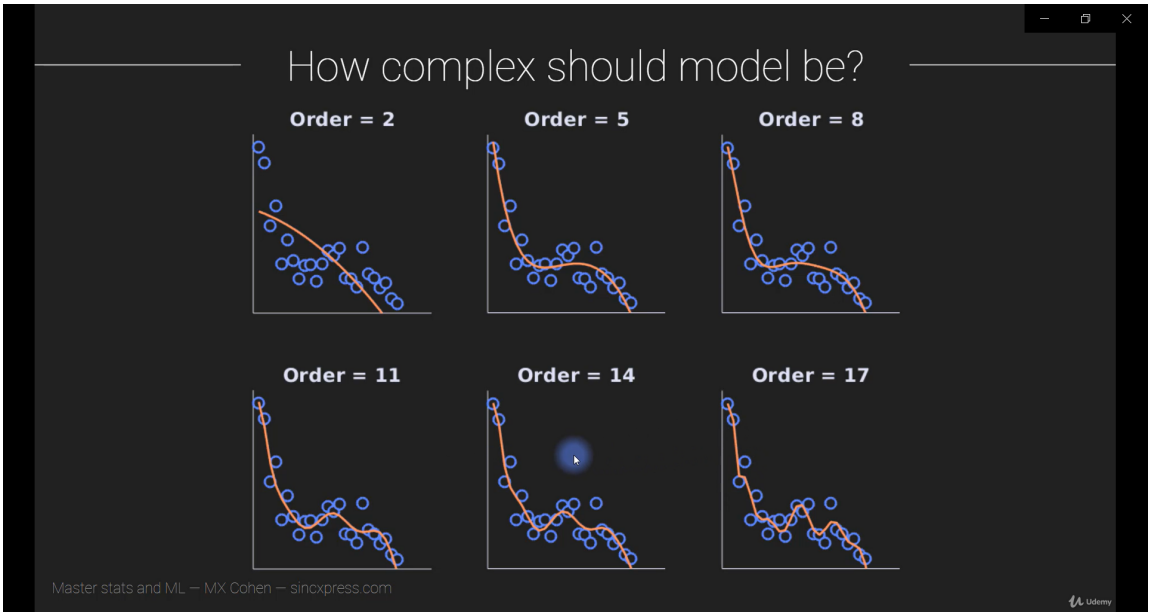
- A model is an equation(or set of equations) that explains some features in the dataset.
- For example, we want to know why there is a large variety in the height of the adults ?
- This is a really complicated question, there are many many factors that determine the height of the adult.
- Now the simplified version of this complicated question can be
  h = a1*x1 + a2*x2 + a3*x3 + epsilon
  h = height
  x1 = gender
  x2 = parents height
  x3 = childhood nutrition

*Now we know there are many more variables that explain someone's height, but we are not trying to explain 100% of every feature about height, we want to capture the most important features in our simplified equation. Of course, these three variables are not going to explain every single pico-metre of height, therefore we add the term residual, all the differences which are not explained by our equation will go into residual.*

**IMPORTANT POINT IS**:

- The residual should be small, but the model should be simple. (Easy to interpret)
- As the complicated models are too hard to interpret and also they can be unstable and also less generalizable.
- Therefore how you find balance between residual and simplicity is a bit hard.

Example: Order 2 is way too simple and order 17 is quite complicated and less generalizable.



## Hypothesis and Models

- All the hypotheses are model comparisons.
- Hypothesis is about which model is better.
- When we say, model1 is better fit to the data than model2, that is an hypothesis and models are basically some equations that describe some real data.

**What are hypotheses and how do you specify one**

**1) What is an hypothesis ?**

- A falsifiable claim that requires verification, typically from experimental or observational data, and that allows us to predict future observations.

**2) Why are hypotheses important ?**

- Hypotheses improve the experiment design, critical thinking and very very important, data analysis.
- It allows you to transform your loose thoughts or ideas or assumptions into concrete and specific claims.

3) A **strong hypothesis** is
- Clear
- specific
- Falsifiable (can be proved wrong)
- Based on the prior data OR theory
- leads to a statistical test (can be tested with some statistical test)
- provides the prediction about the direction of the effec.
- relevant for the unobserved data OR phenomenon
- relevant for understanding the nature of data OR phenomenon.
- Statement and not a question.

4) Some examples of Not an hypothesis:

- Medical research is important for curing disease. (This is true and we 100% agree by this, but this is NOT an hypothesis)
- Will students pass this course (This is question and not the statement, we can think of generating hypothesis from this, but this thing on its own is not an hypothesis)

5) Some examples of Weak hypothesis:

- The medication has an effect.
  (It is not a good hypothesis, but technically it is a hypothesis and we can try to falsify this statement, but it;s not very clear and specific. And it's not relevant to understand the medication or disease. Say for example, you take medication for sleep at night, but it turns your fingers black. That would be an awful thing. The medication clearly has an effect, but the effect has nothing to do what the medication is for and it is awful effect)
- Studying improves grades (Obviously it is  true, but it is not specific, and it is hard to contextualise this.)

6) Some examples of Strong hypothesis:

- The medication reduces the symptom X in the dose-response fashion.
  (dose-response means the more the dose you take, the more symptom reduces)
  (So this is clear hypothesis, it is specific, it is falsifiable, it leads to particular statistical test, and we can make prediction about unobserved data)
- A combination of self study and group study will improve grades by at least 10%.
  (it is clear, specific, falsifiable, ..)

7) Now I will be giving a few examples and think about which is not, week or strong hypothesis.

1) There are other universes with different physical rules.
2) Wearing purple underwear improves mood.
3) Plant grows differently in sugar-water
4) Mike's courses are awesome.
5) Washing hands for 20 seconds reduces the disease spread.
6) An apple a day keeps the doctor away.
7) Are people more creative after watching stand-up comedy ?

→

1. Not an hypothesis, because it is not falsifiable, we can't do experiments, we cannot collect the data or provide support or go against this statement.
2. Strong hypothesis, because it is clear, falsifiable, can be used for predicting future observations, a bit of a silly hypothesis but still it is strong.
3. Weak hypothesis: because it can be falsifiable, we can collect data and get the evidence, but it is not specific, because what does differently means, better or worse.
4. Not hypothesis
5. Strong: it is clear, specific, we can test various things, like is it really 20 seconds or 15 seconds and so on.
6. Weak hypothesis
7. Not an hypothesis: because it is a question and not the hypothesis, you can come up with a strong hypothesis using this question, but this question on its own is not an hypothesis.
   Example of strong hypothesis from this question: people score 15% more on creative problem solving after watching stand up comedy.

**Null Hypothesis**

- Null hypothesis is the hypothesis that nothing interesting happened in the data (i.e no effect on process you did on data)
- In the research, you specify the alternative hypothesis.
- ***In statistical analysis, you always test null hypothesis and not alternate hypothesis.***
- Then the question raises, that do your data support the null hypothesis, the hope is, it do not support null hypothesis and we reject it in favour of alternate hypothesis
- Example:
  **Alternate hypothesis H1**: People will buy more product X after seeing advertisement 1 as compared to the advertisement2.

**Null Hypothesis H0**: The advertisement type has no effect on purchases.

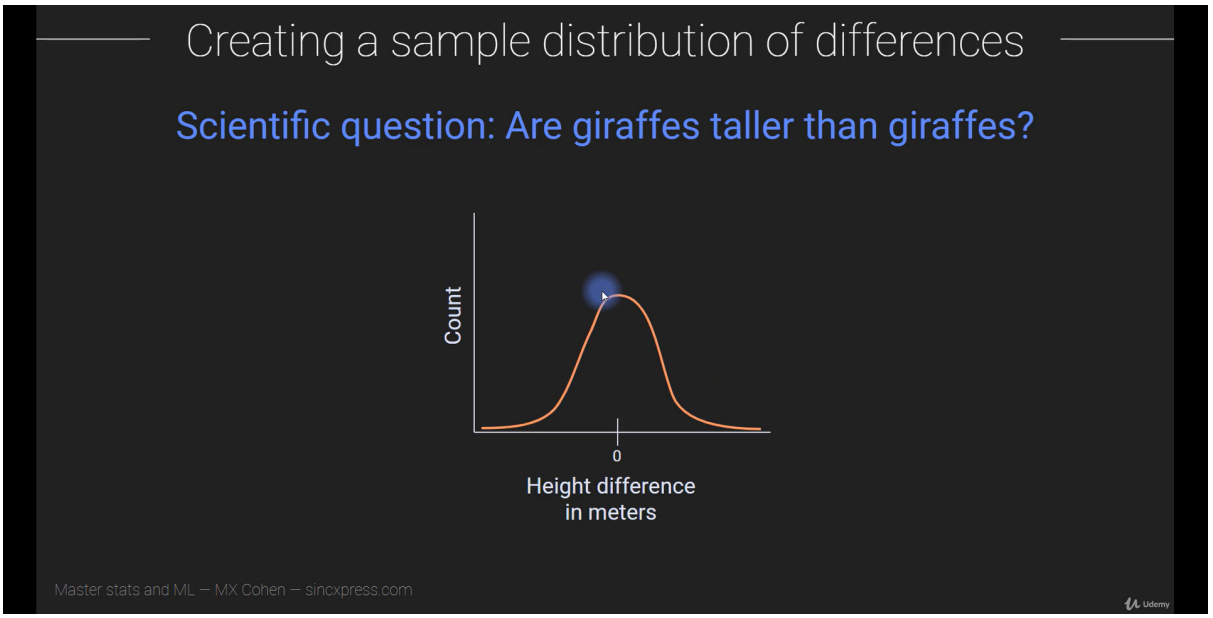
- Research is all about the alternative hypothesis, but in statistical analysis we try to quantify the evidence of the weak hypothesis using the data, if evidence is weak, then we reject it.

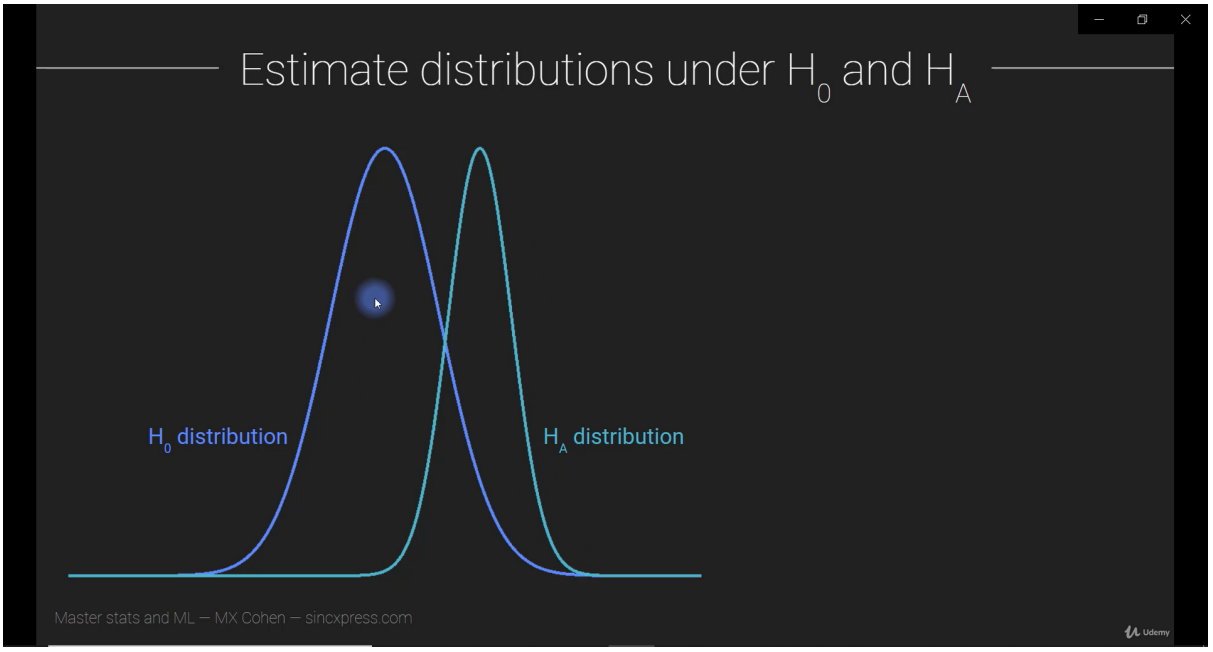
### Hypothesis proving ?

1) It is NOT possible to PROVE the hypothesis.
2) Hypothesis can be rejected OR failed to be rejected (interpreted as tentatively supported until a better hypothesis comes.)

### Sample Distributions under Null and Alternative hypothesis:

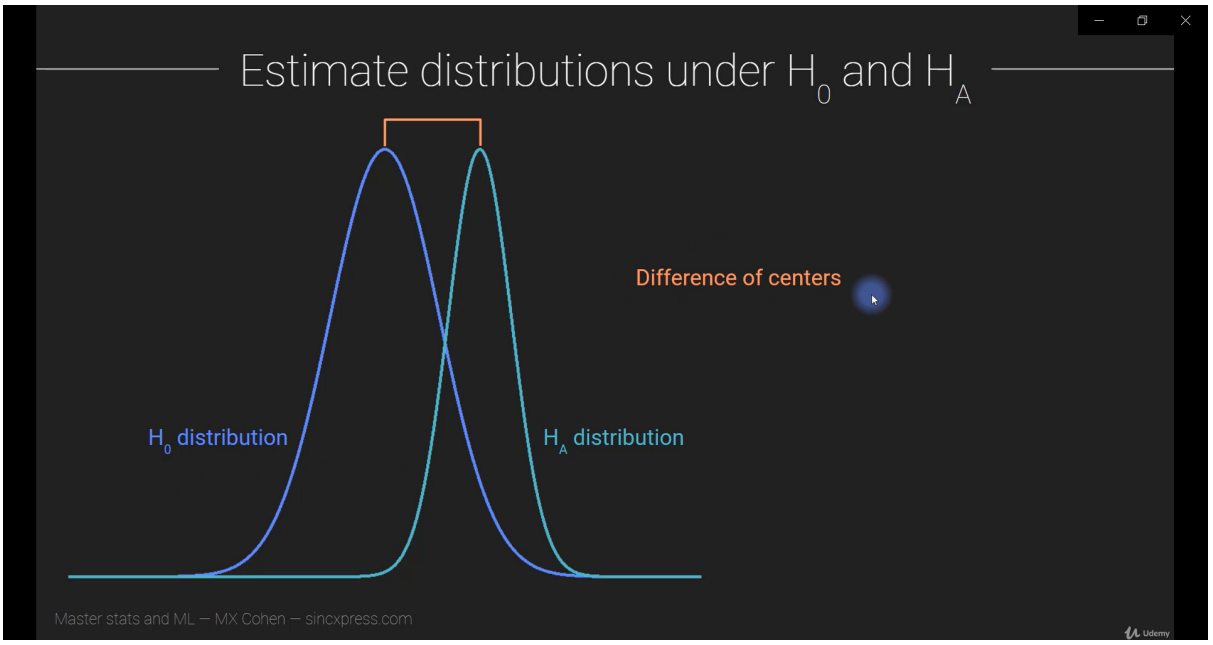
- Reminder: The distribution of sample means is gaussian (CLT)
- Lets create null hypothesis from the scientific question:
  Are giraffes taller than giraffes ? Obviously the answer is no.
  This is the null hypothesis question that we are asking here.
- The reason why we are setting up this question is to illustrate how to create the distribution (of sample parameter estimates) under the null hypothesis.
- Say we sampled the mean of heights for n1 times and we get the distribution. The mean of the sample means was = 6 metres.
- Again we sampled mean of heights for n2 times (n1 == n2 and different groups of giraffes), then the mean of sample means = 6.2 metres.
- Then, is it true that giraffes are 0.2 metres taller than giraffes?
  NO. It means that on an average, giraffes(of group 2) are 0.2 taller than giraffes(of group 1).
  The difference is because of sampling variability, noise, ...
- The point is, we expect the difference to be zero, but actually the difference is non-zero because of all causes of sampling variability.
- **Therefore, we are willing to tolerate some difference around zero.**



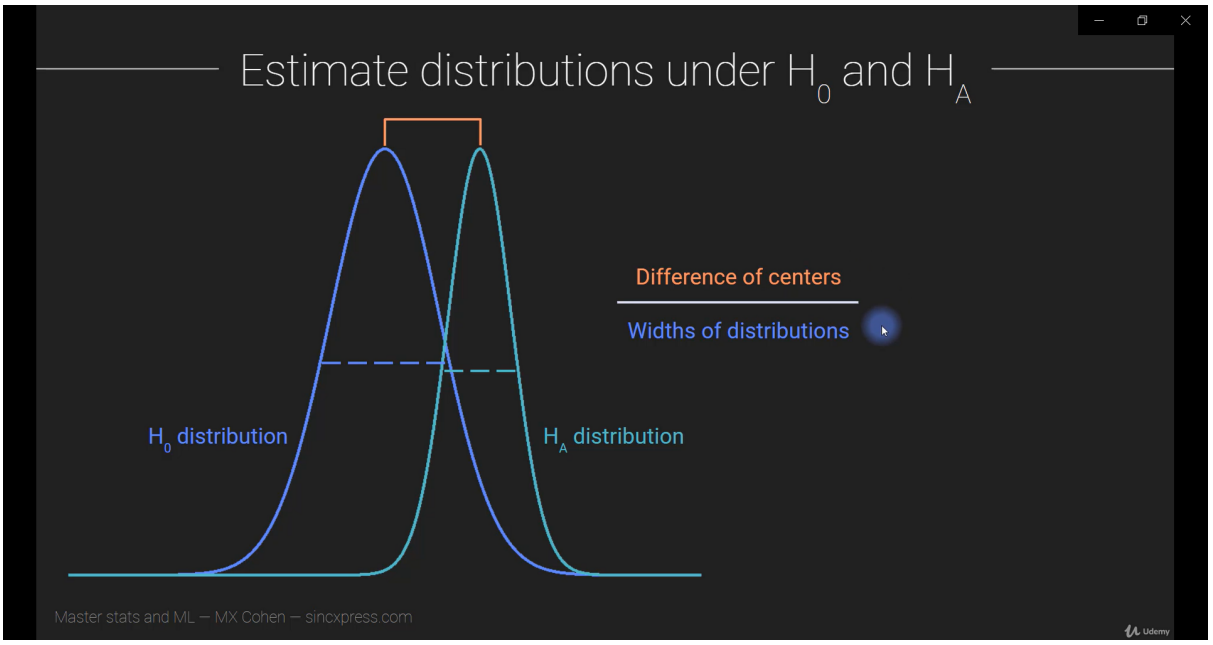



**We want some probability that null hypothesis distribution and alternate hypothesis distribution are different from each other.**

So one way we can say they are different based on comparing the difference in centres of distribution (means), but it will be scale dependent (metres), and we want a normalised way and also we will not be considering the width of the distribution.

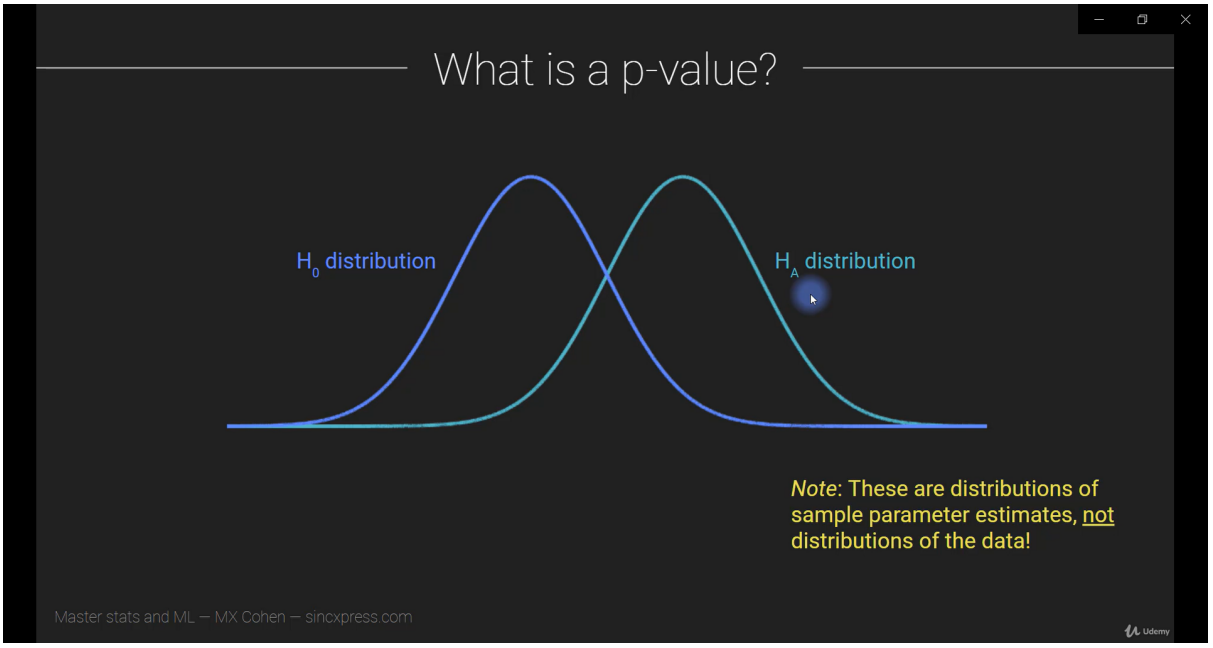


The other way could be, **difference_of_centers / width_of_distribution** OR **signal / noise**
*Many statistical tests and basically the entire inferential statistics are based on this formula, these tests basically adapt this framework and add things, have different definitions of widths, based on some assumptions, but in general follow this framework.*

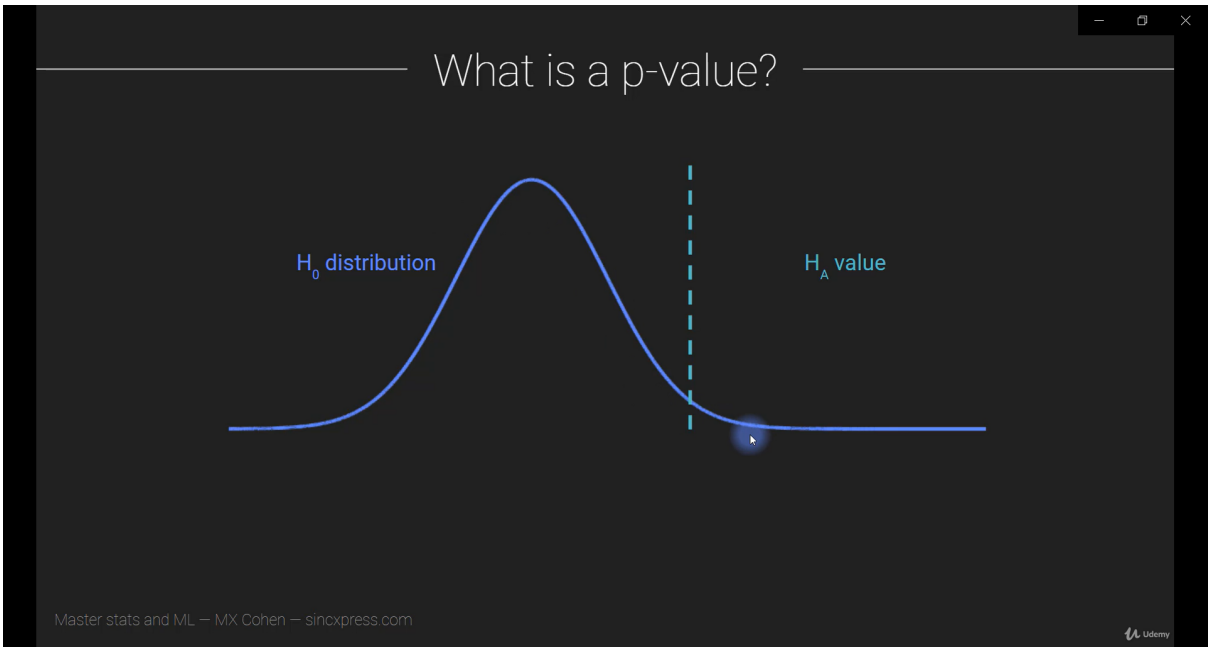



### p-values, tails and misinterpretation

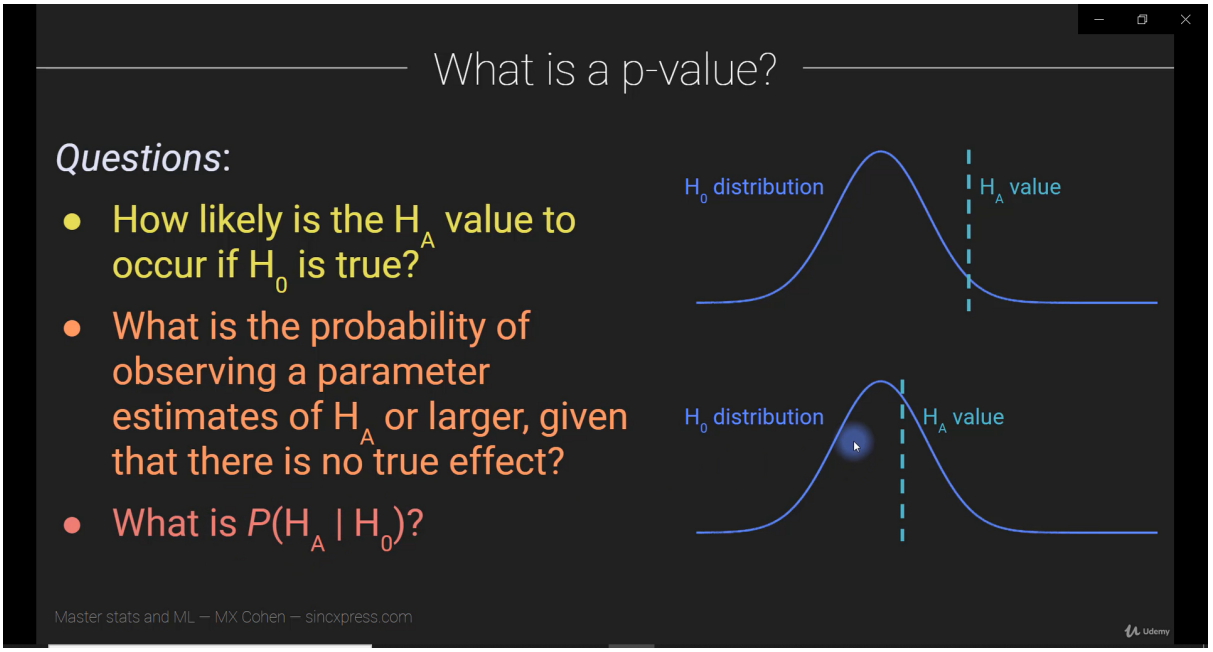
1) Null hypothesis distributions are based on some assumption, some formula, but we really don't know about the distribution of alternate hypotheses.



Since we don't know the distribution under alternative hypothesis

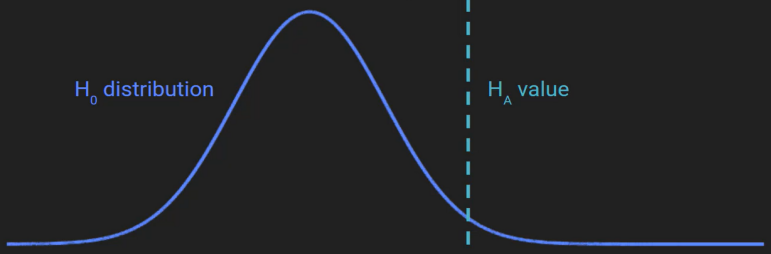



Answer these questions:




All three questions are the same.
p-value is the probability of observing the effect GIVEN the null hypothesis is true.

## What is a p-value?

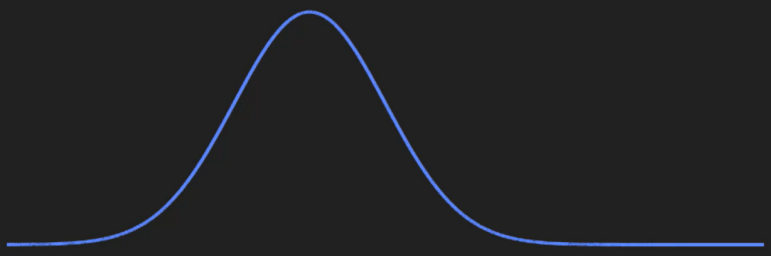



Important concept:

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.



## What is a p-value?



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

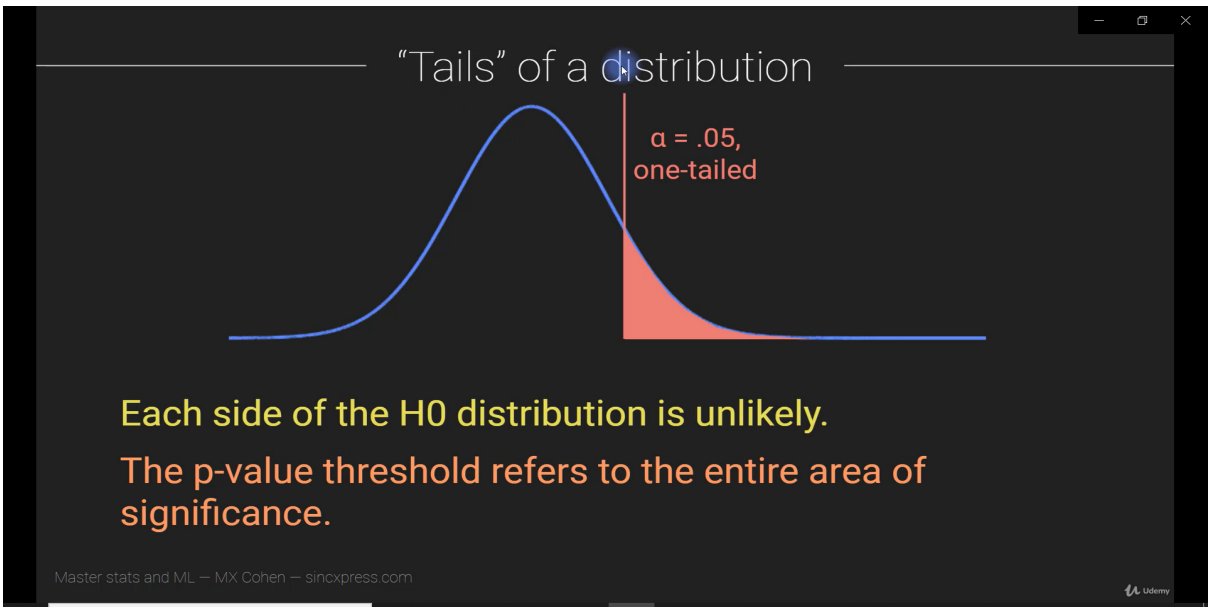
Values closer to zero indicate low probability of $H_A \mid H_0$, and values closer to one indicate high probability of $H_A \mid H_0$.

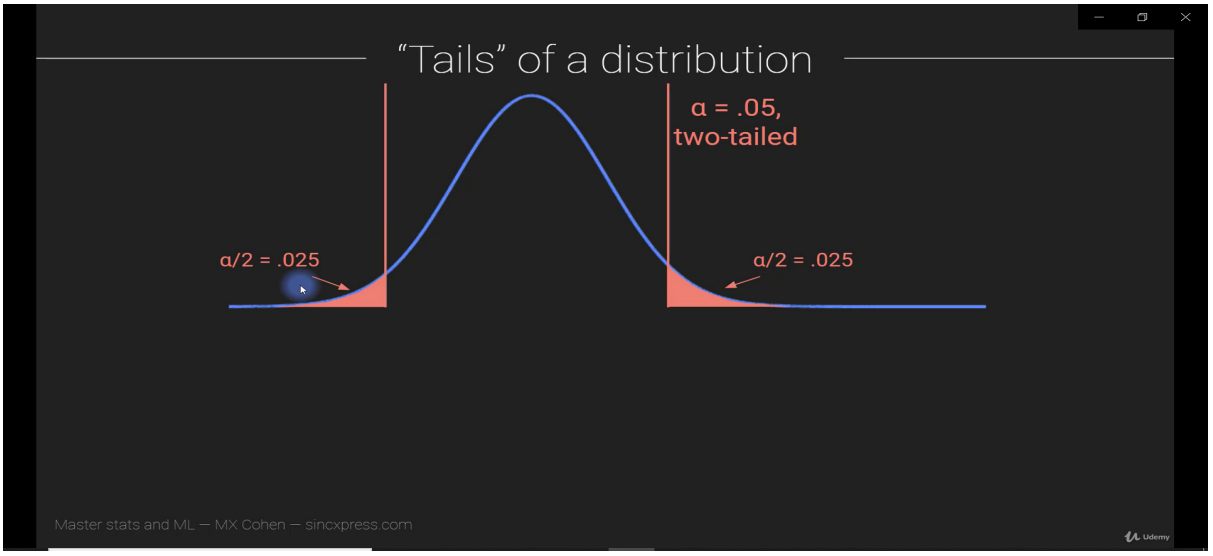


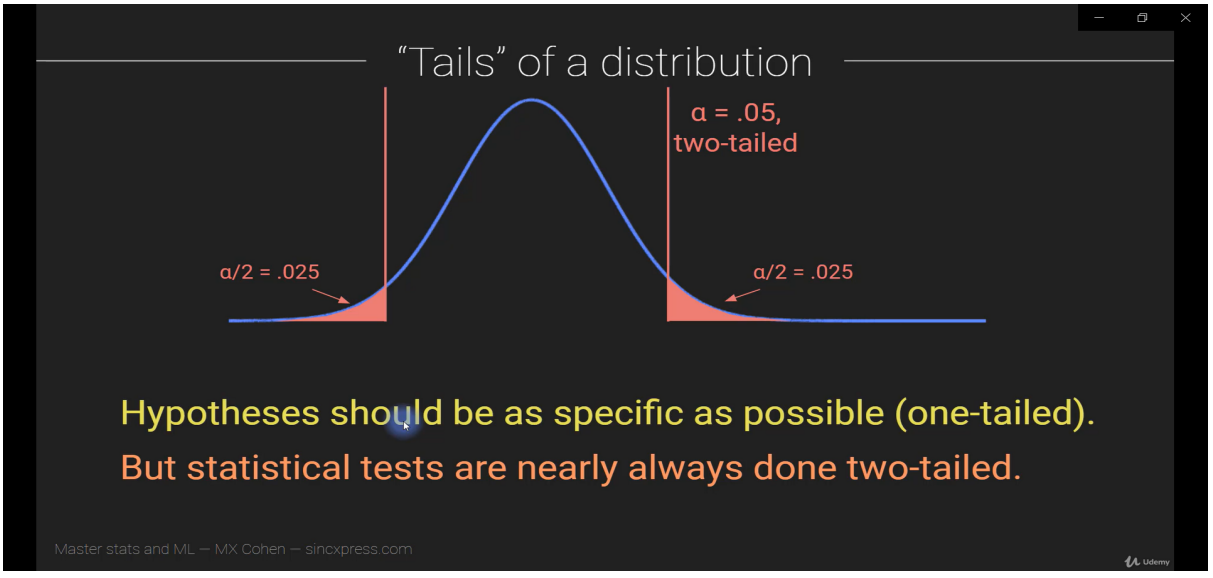
## When is a finding "statistically significant"?


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

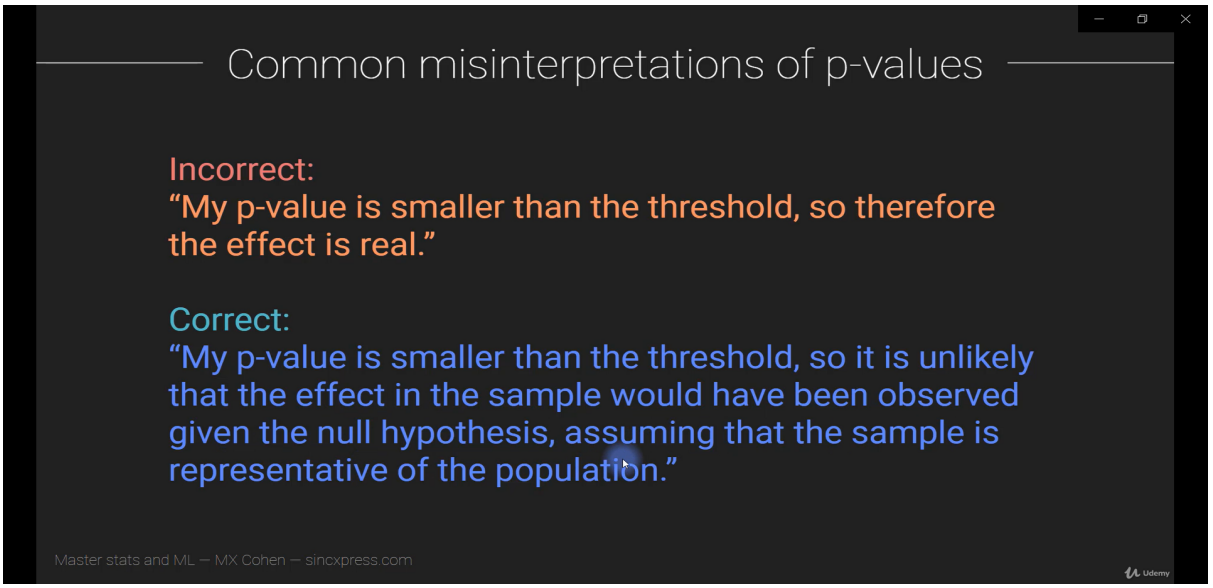
Threshold is arbitrary; common values are $p<.05$ or $p<.01$.

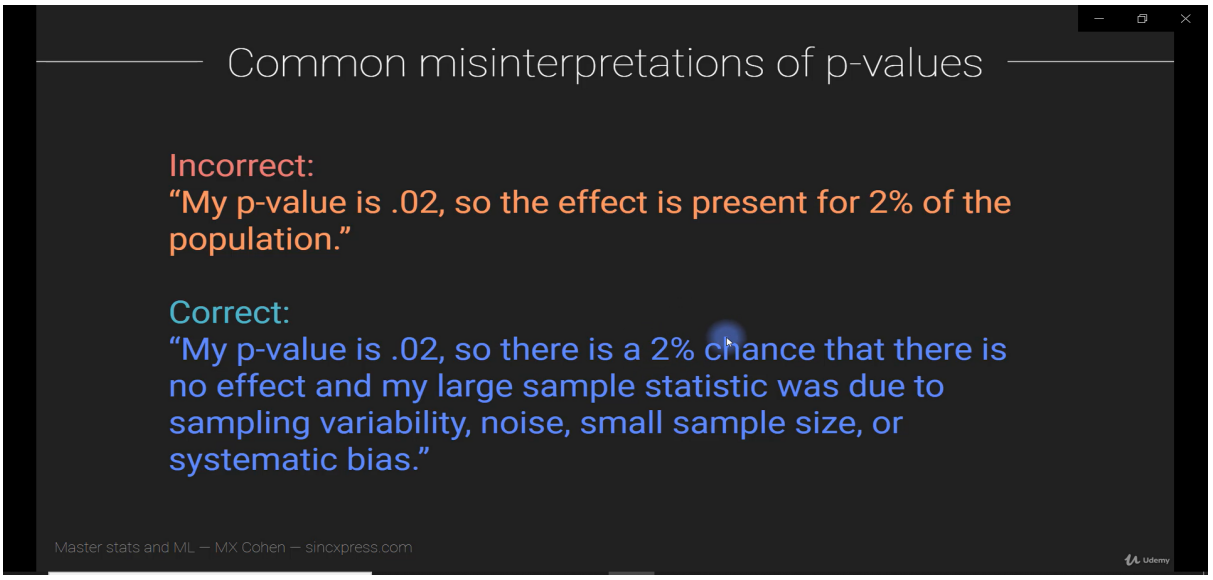

"Tails" of a distribution
α = .05,
one-tailed
Each side of the H0 distribution is unlikely.
The p-value threshold refers to the entire area of significance.
Master stats and ML — MX Cohen — sincxpress.com

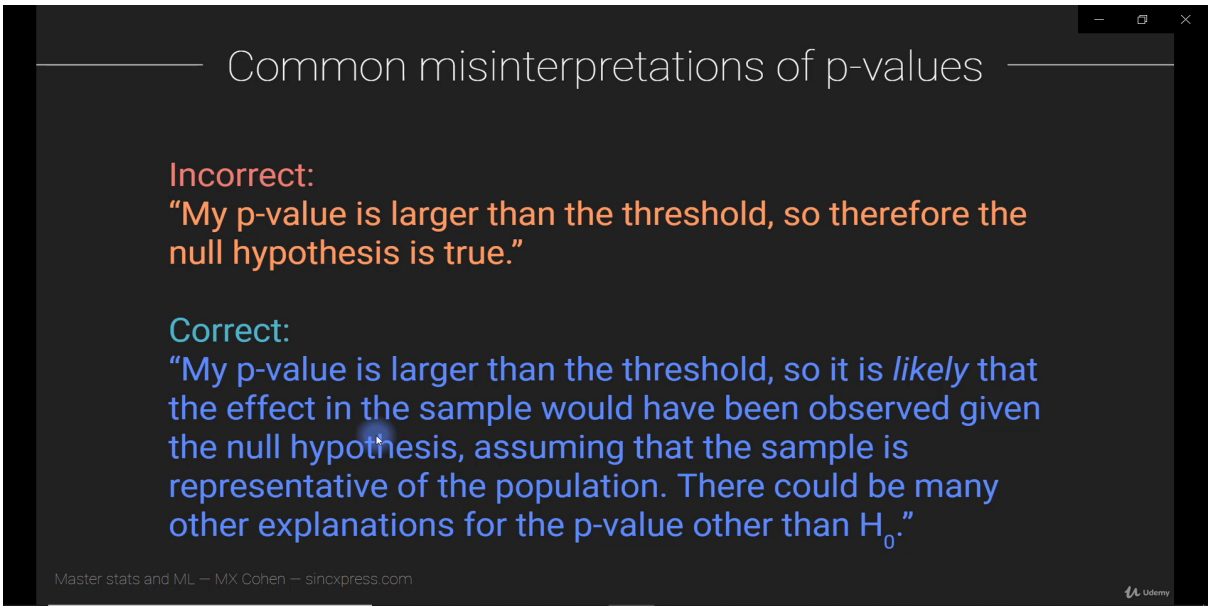

"Tails" of a distribution
α = .05,
two-tailed
α/2 = .025
α/2 = .025
Master stats and ML — MX Cohen — sincxpress.com

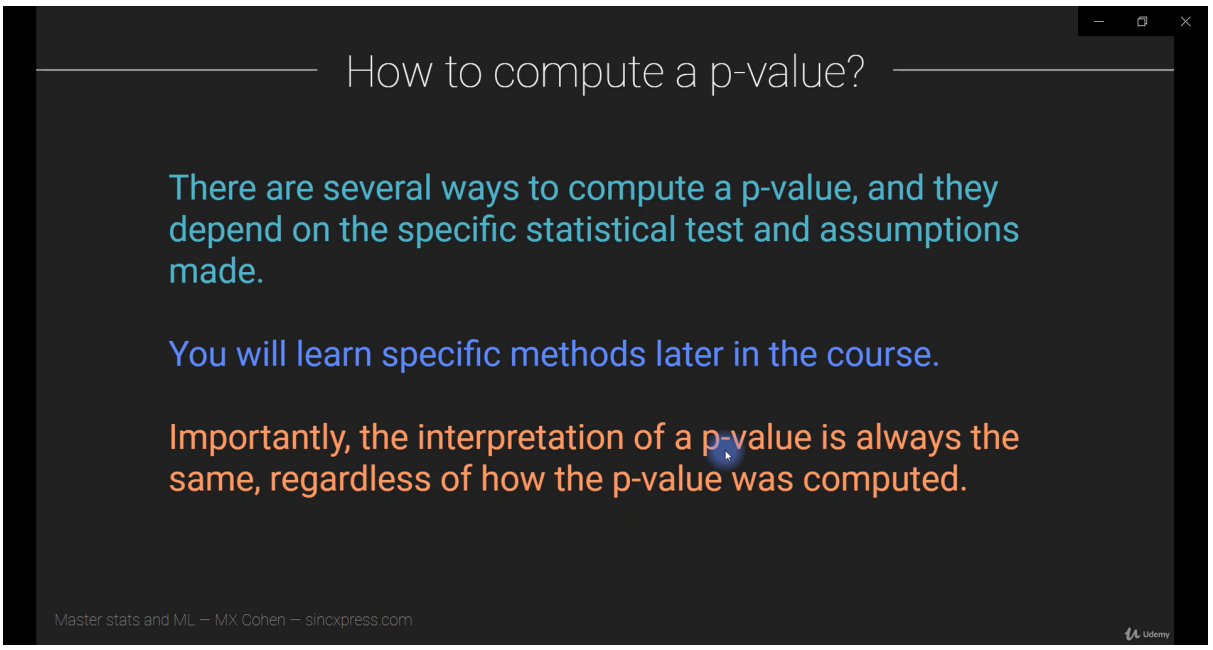

"Tails" of a distribution
α = .05,
two-tailed
α/2 = .025
α/2 = .025
Hypotheses should be as specific as possible (one-tailed).
But statistical tests are nearly always done two-tailed.
Master stats and ML — MX Cohen — sincxpress.com

## Common Misinterpretations of p-values



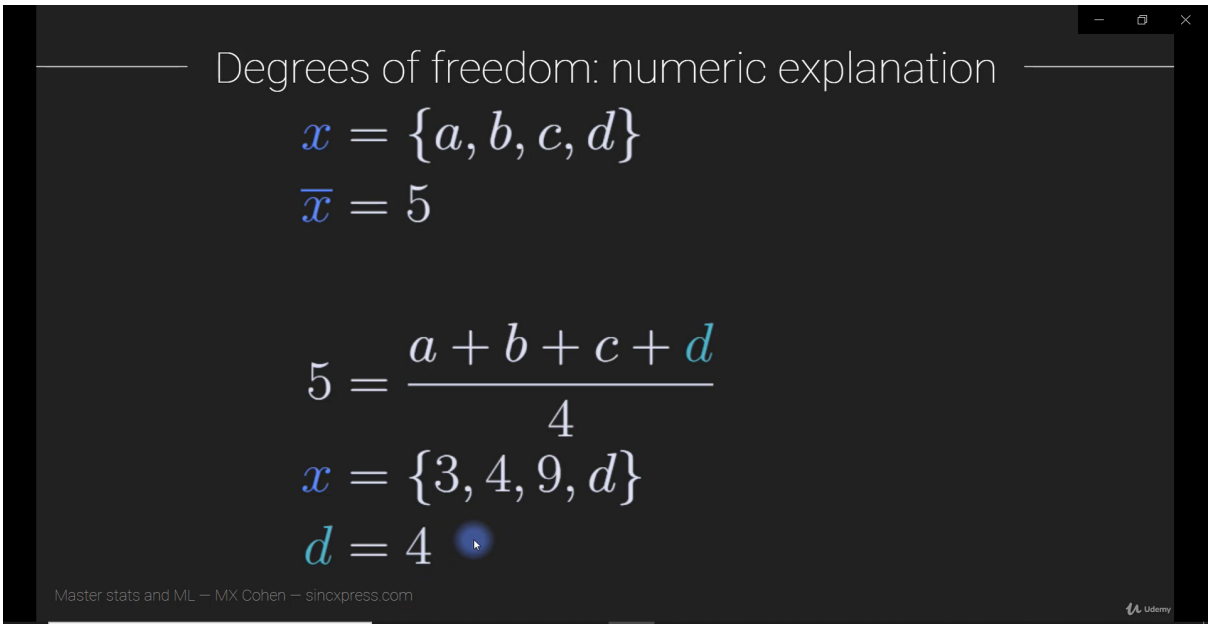



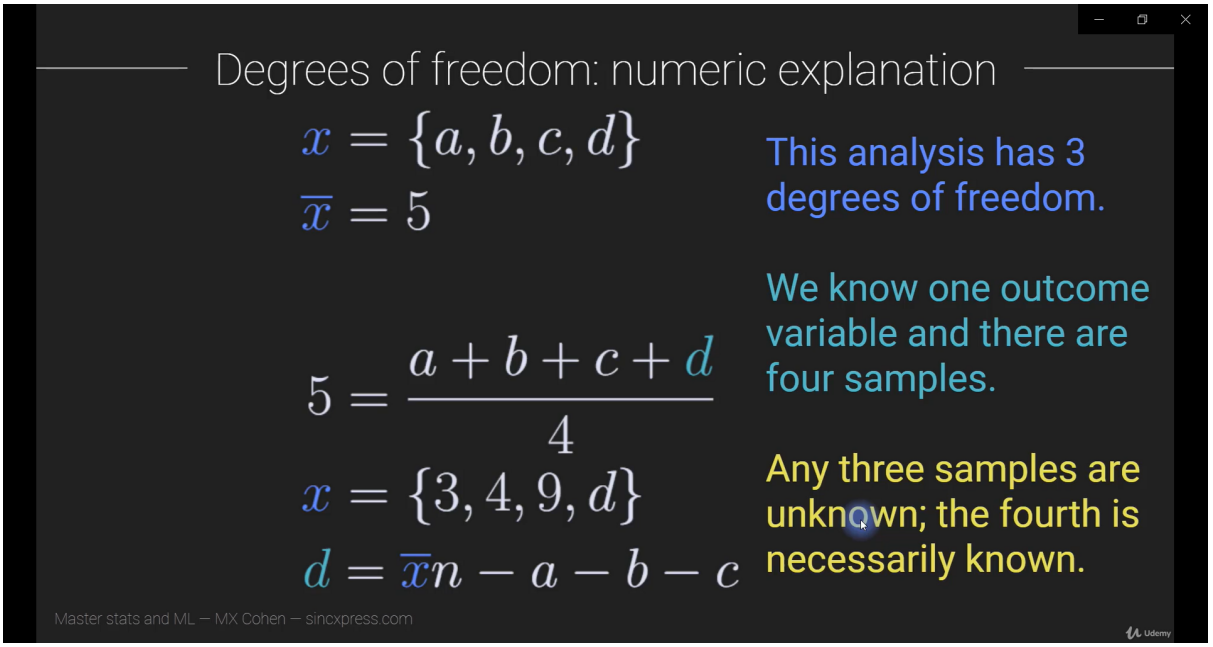



**Degrees of Freedom**

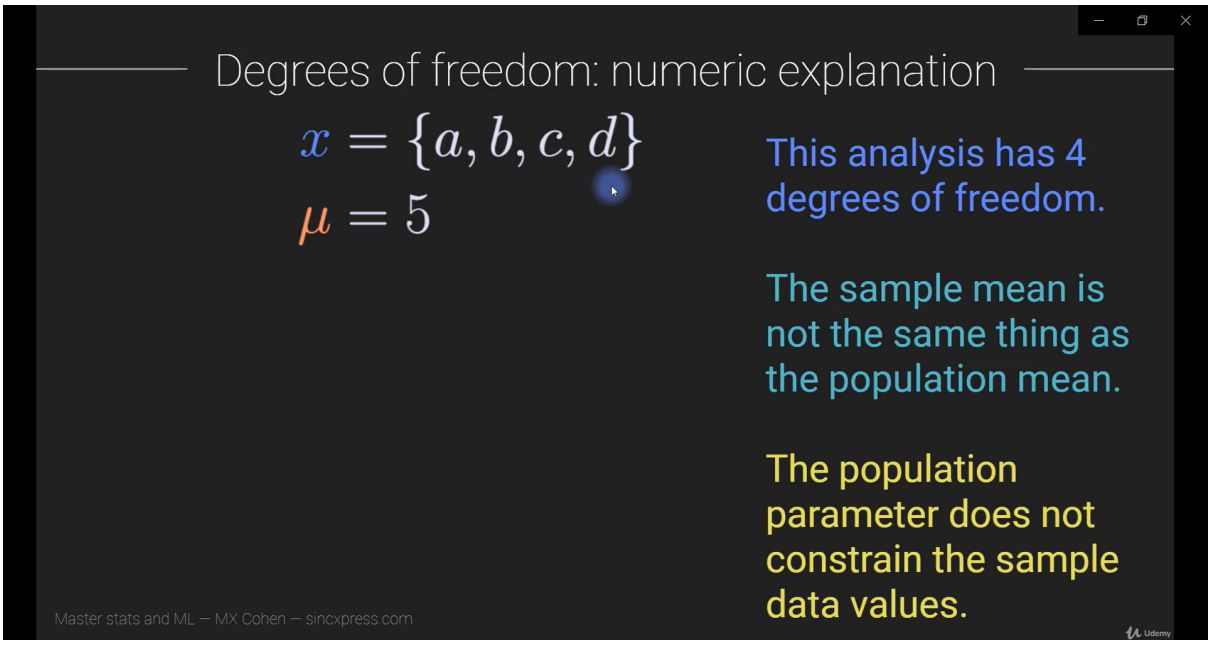
- Say you had four numbers and their mean, you know only the first three values and the fourth value is unknown, then you can calculate it easily.
- That is, the first three values can take on any values and the fourth value depends on the first three values, i.e it has 3 free values to take i.e it has 3 degrees of freedom.



**Note:** *In this case, we were given the sample mean and that's why the degrees of freedom is 3*




Now, say we were given the **Population Mean** and the four numbers, then the degree of freedom is 4.



## Why are degrees of freedom important?

Degrees of freedom (df) determine the shape of H0 distributions.

Higher df generally indicates more power to reject the null hypothesis (related to statistical power).



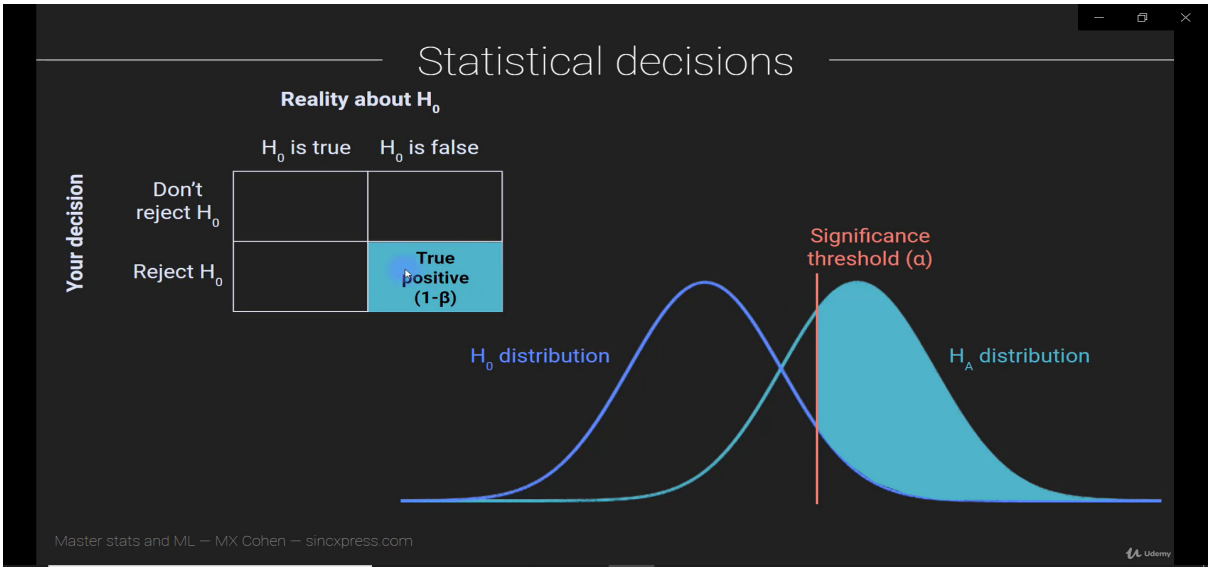
## Degrees of freedom definition

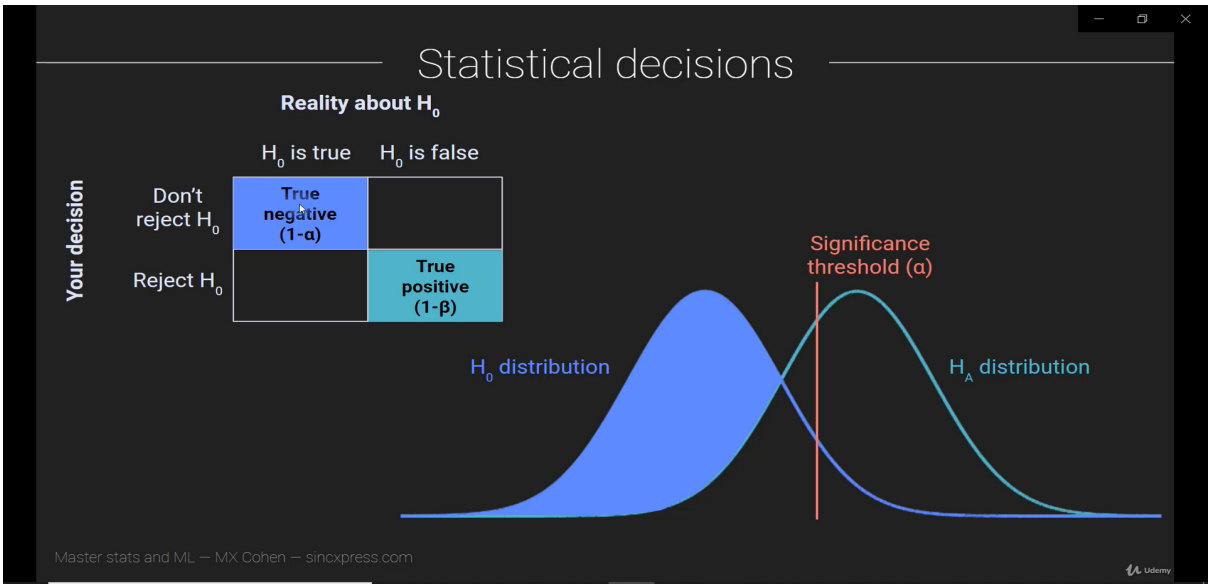
Degrees of freedom:

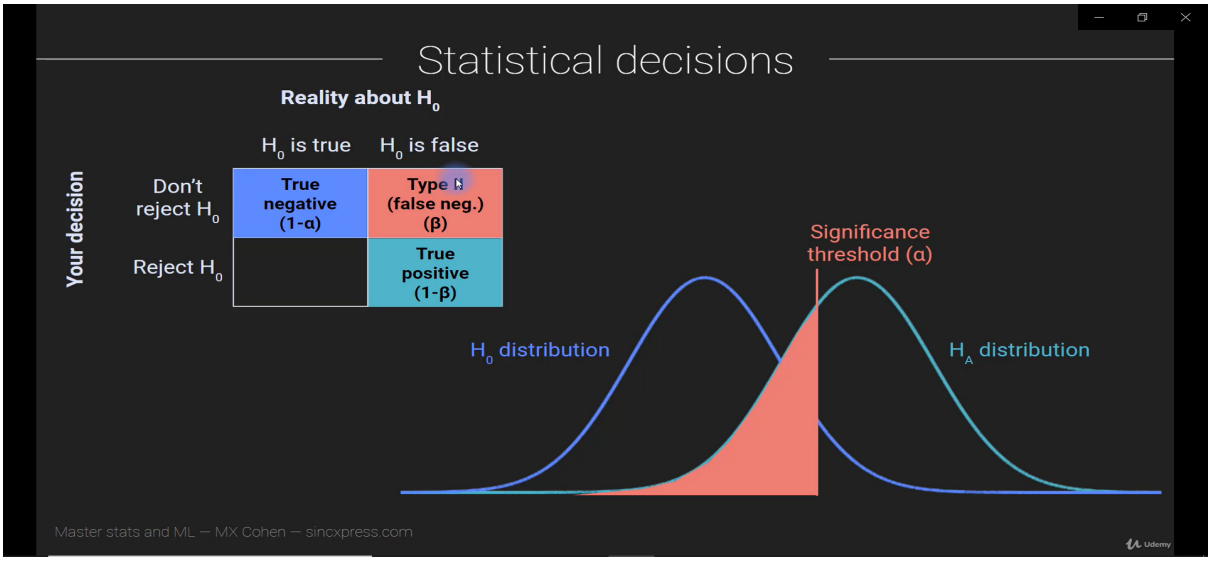
- The number of independent sample values.
- The number of sample data points that can vary.
- The number of data values that are unconstrained by the rest of the data values.

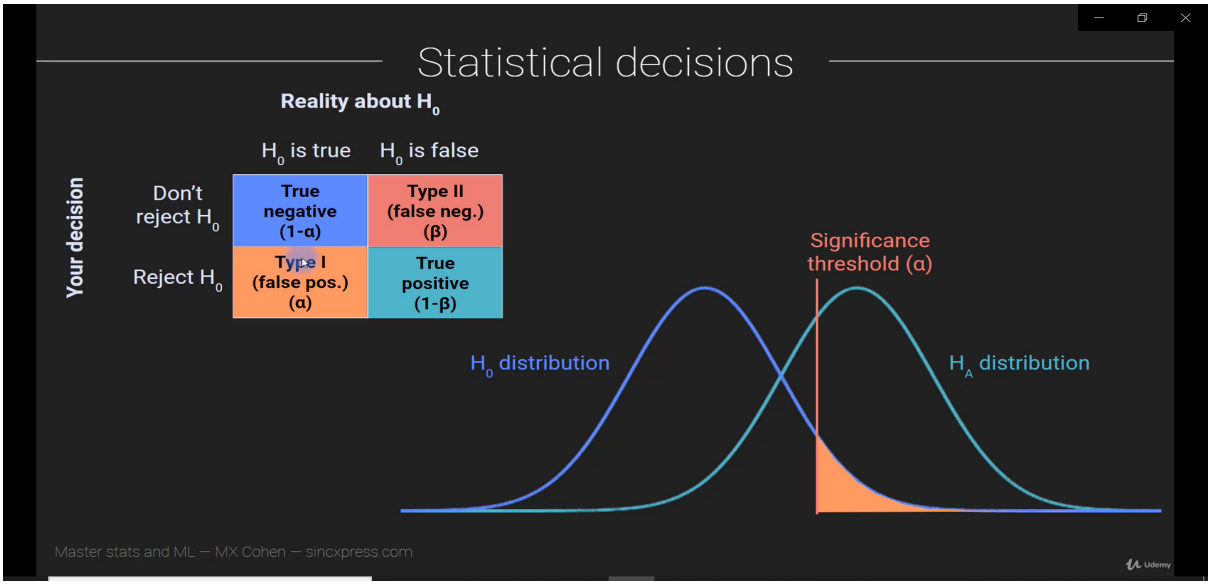
Generally: df = N-k (N data points, k parameters)

## Type I and Type II Errors

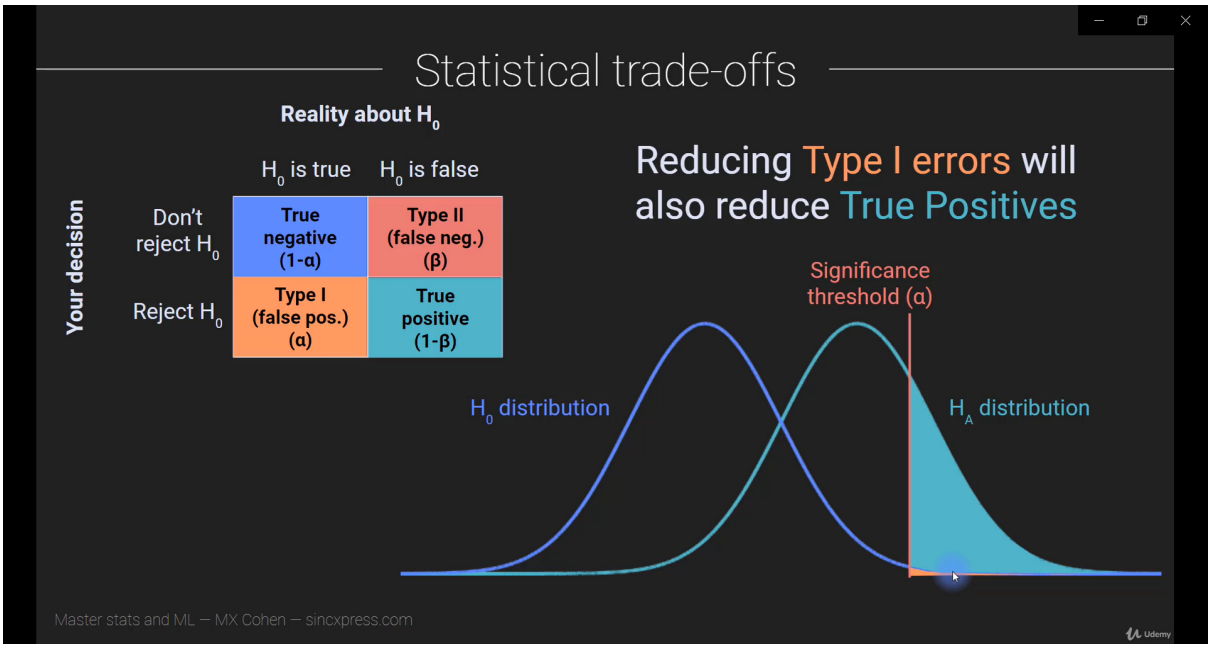


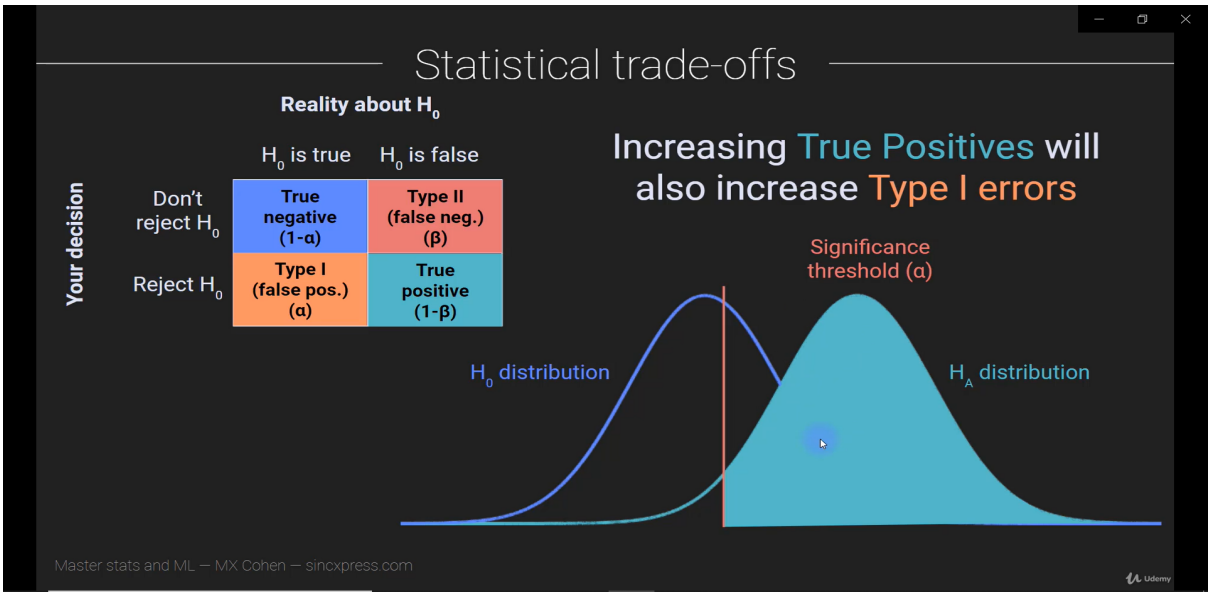



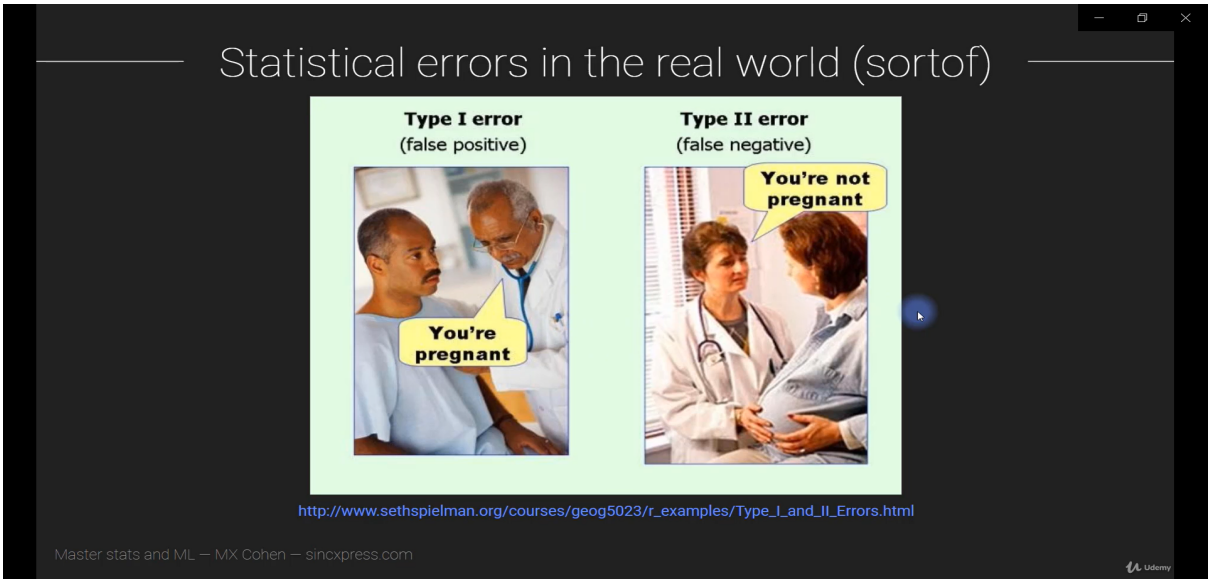

Statistical decisions
Reality about $H_0$
$H_0$ is true
$H_0$ is false
Your decision
Don't reject $H_0$
Reject $H_0$
True negative (1-α)
Type II (false neg.) (β)
True positive (1-β)
Significance threshold (α)
$H_0$ distribution
$H_A$ distribution
Master stats and ML — MX Cohen — sincxpress.com
Udemy


Statistical decisions
Reality about $H_0$
$H_0$ is true
$H_0$ is false
Your decision
Don't reject $H_0$
Reject $H_0$
True negative (1-α)
Type II (false neg.) (β)
Type I (false pos.) (α)
True positive (1-β)
Significance threshold (α)
$H_0$ distribution
$H_A$ distribution
Master stats and ML — MX Cohen — sincxpress.com
Udemy

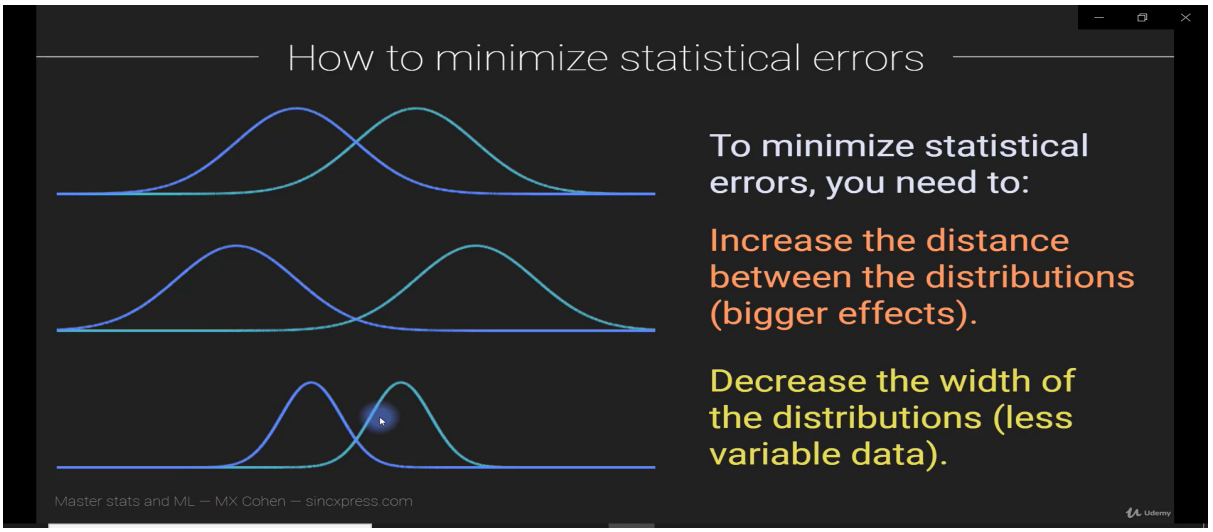
## Statistical Trade-offs



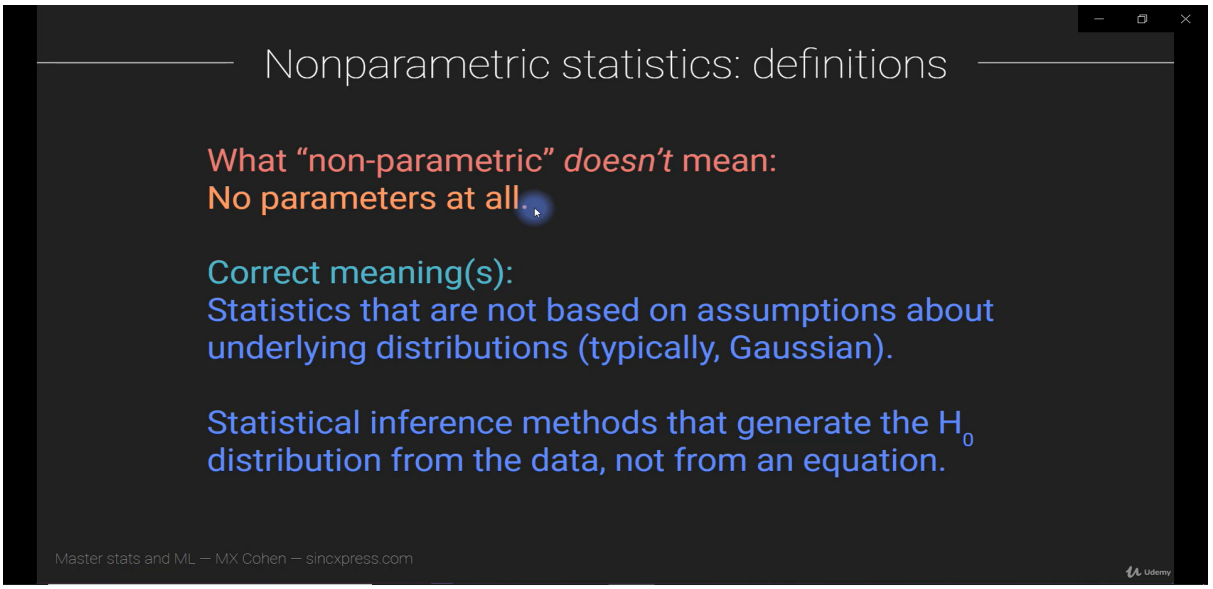



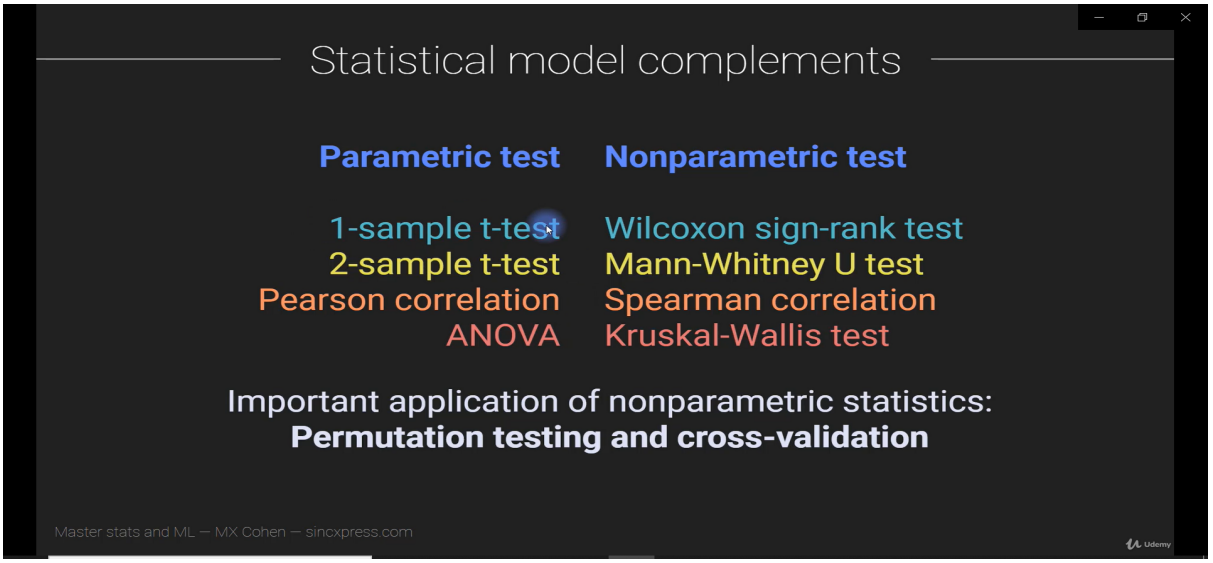


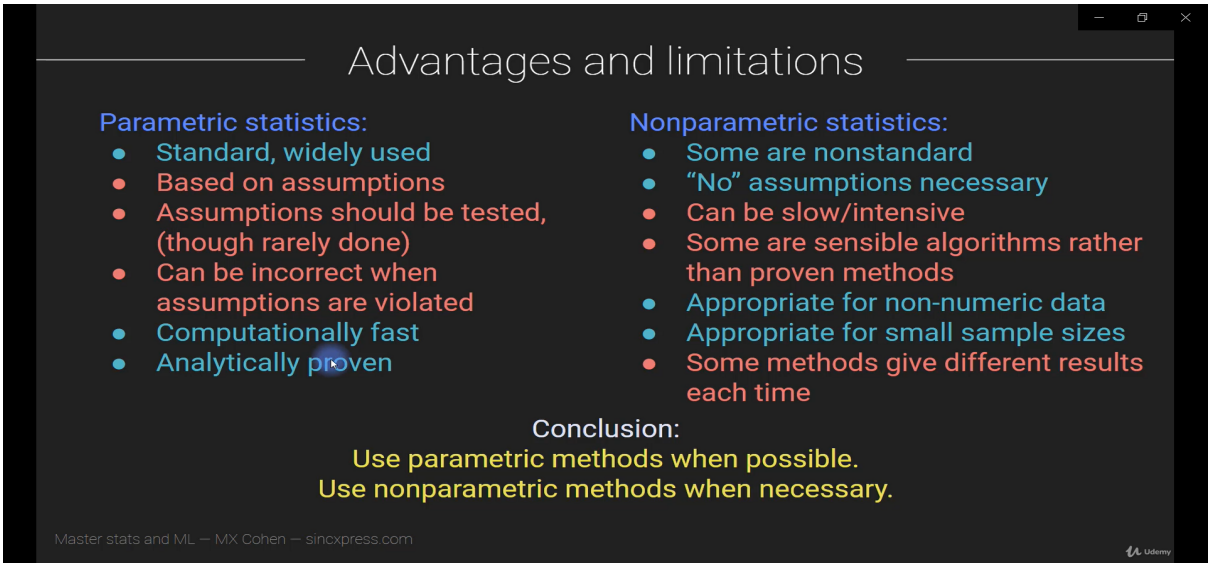

## How to reduce Statistical Errors



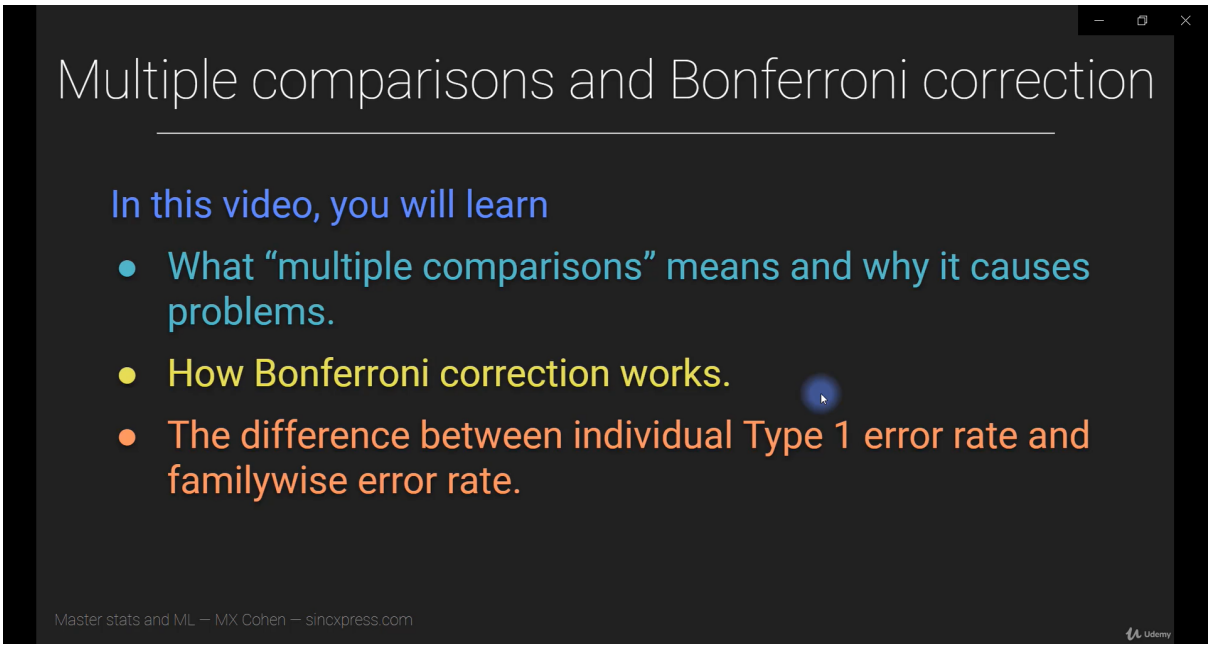
## Parametric and Non-Parametric Statistics



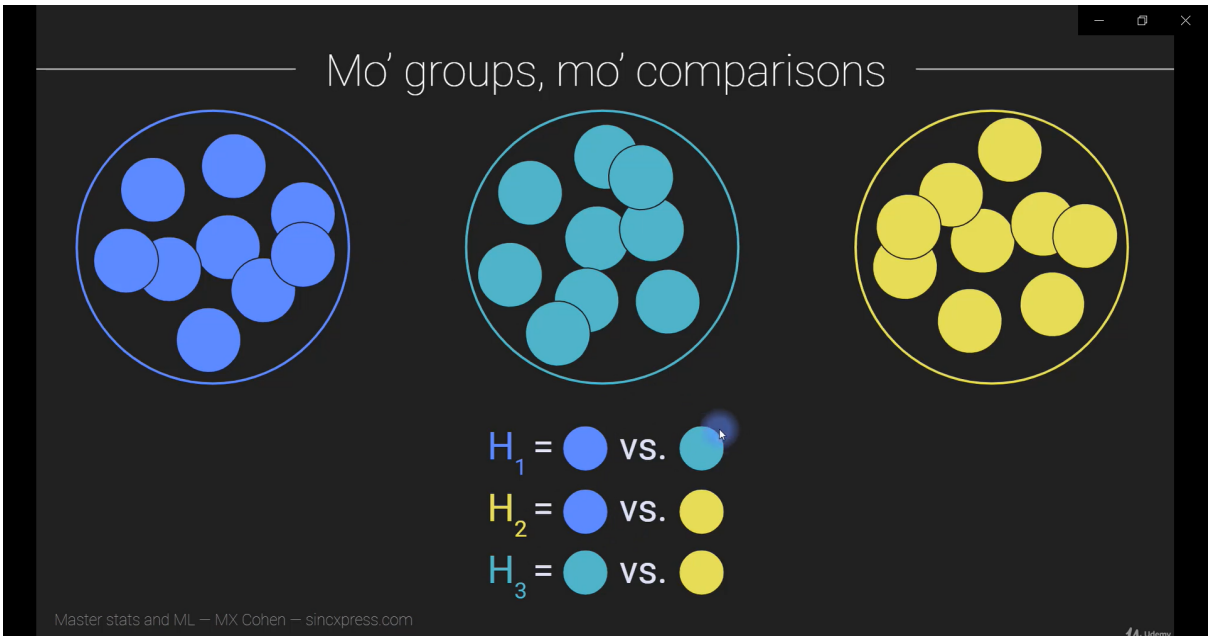



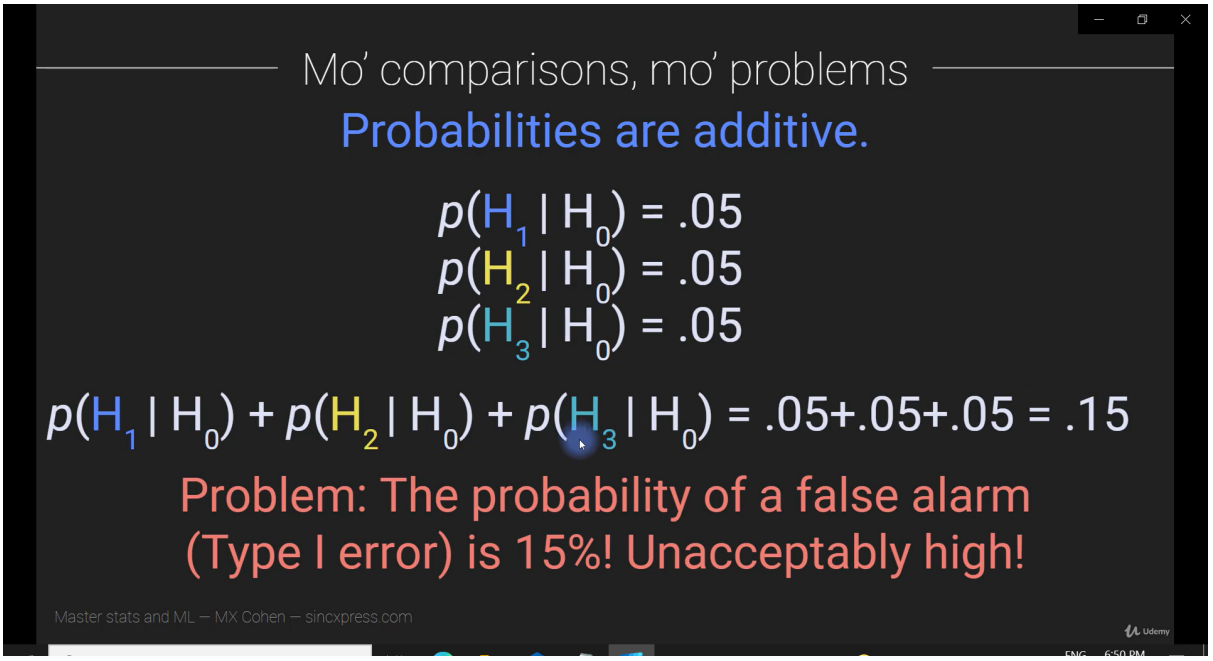
## Multiple Comparisons and Bonferroni Correction



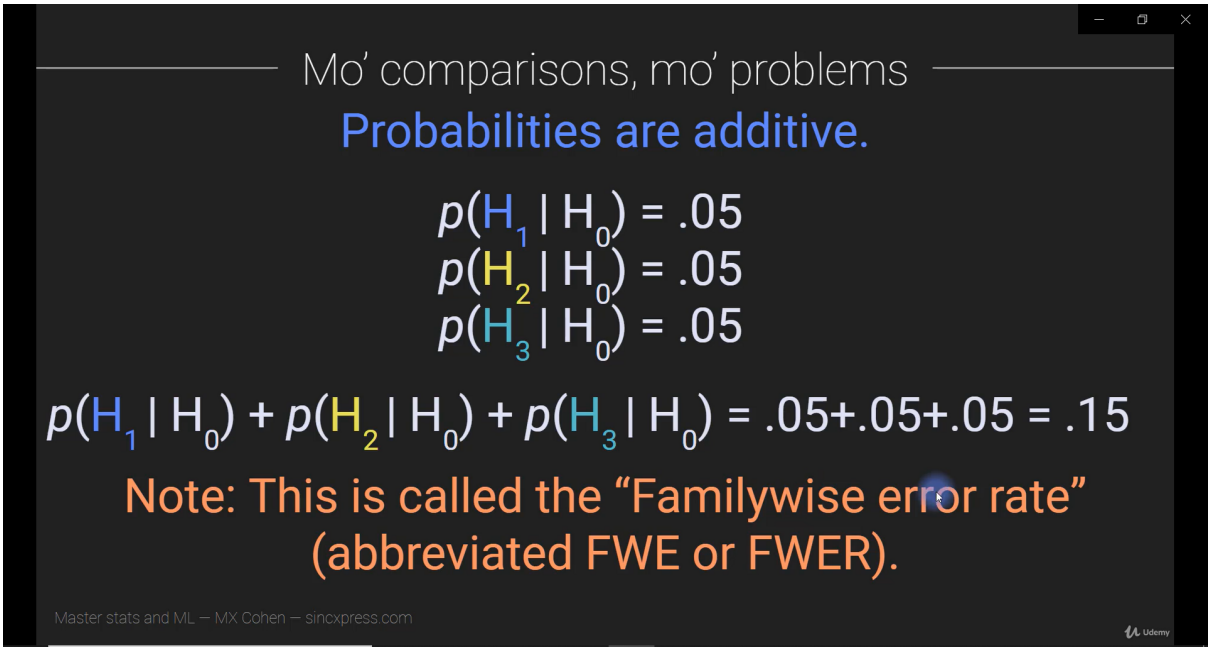

Say we have three groups (blue, green, yellow) and we want to compare the means of these three groups.
In how many ways can we compare their means ?
→ 3 ( $H_1$ = (blue,green), $H_2$ = (green,yellow), $H_3$ = (yellow,blue) )
(assuming that comparison is asymmetric, i.e comparison between blue and green is same as green and blue)



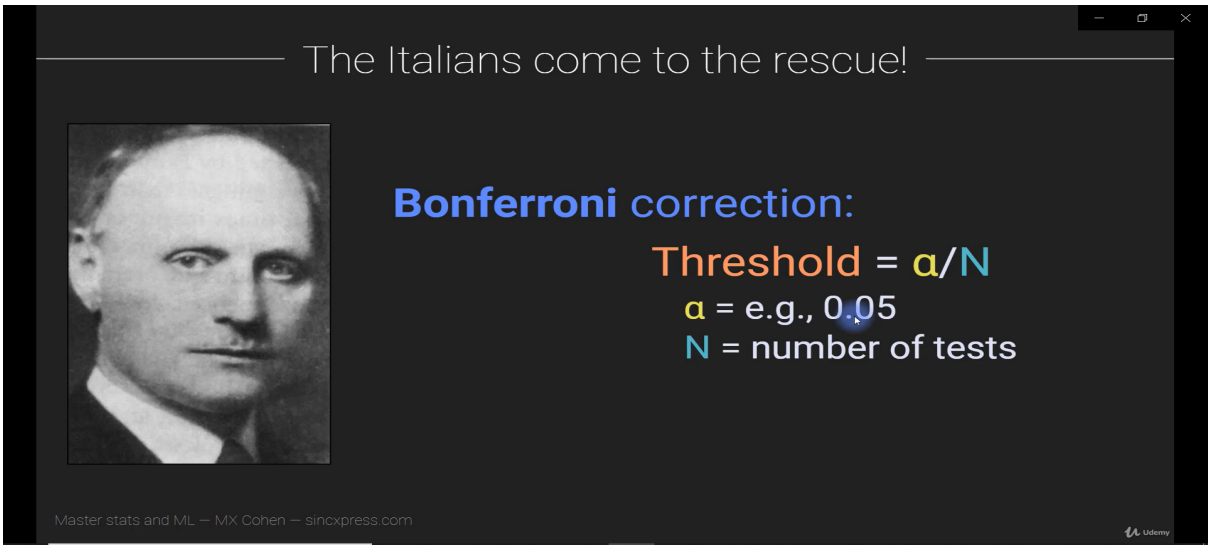
- So there are three hypotheses which we want to perform with the threshold of 0.05.
- The problem is that probabilities are additive. The first comparison is independent of the second and so on.
- Therefore, even though for individual hypothesis we have 0.05 threshold (e.g P(H1 | H0) = 0.05 ), the total test will have the threshold of 0.15

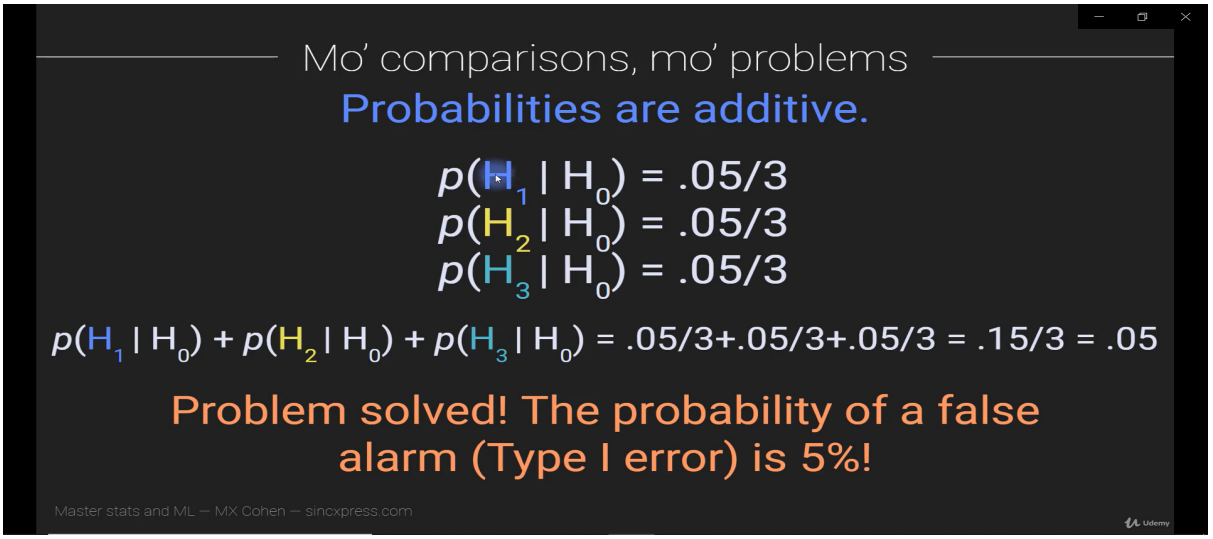



- Because they are all based on the same dataset, when we put all the tests in the set, the overall threshold becomes 0.15.
- Imagine, when we have many many comparisons to do, the overall p-value will be really high.

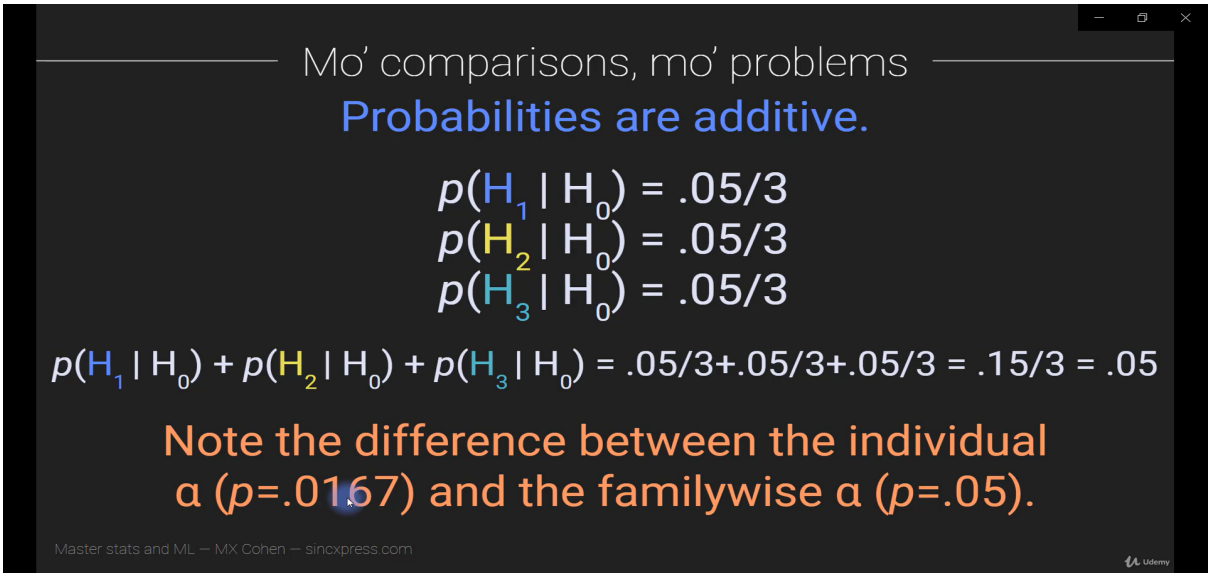


So by the Bonferroni correction, the threshold for the individual test will become alpha / n



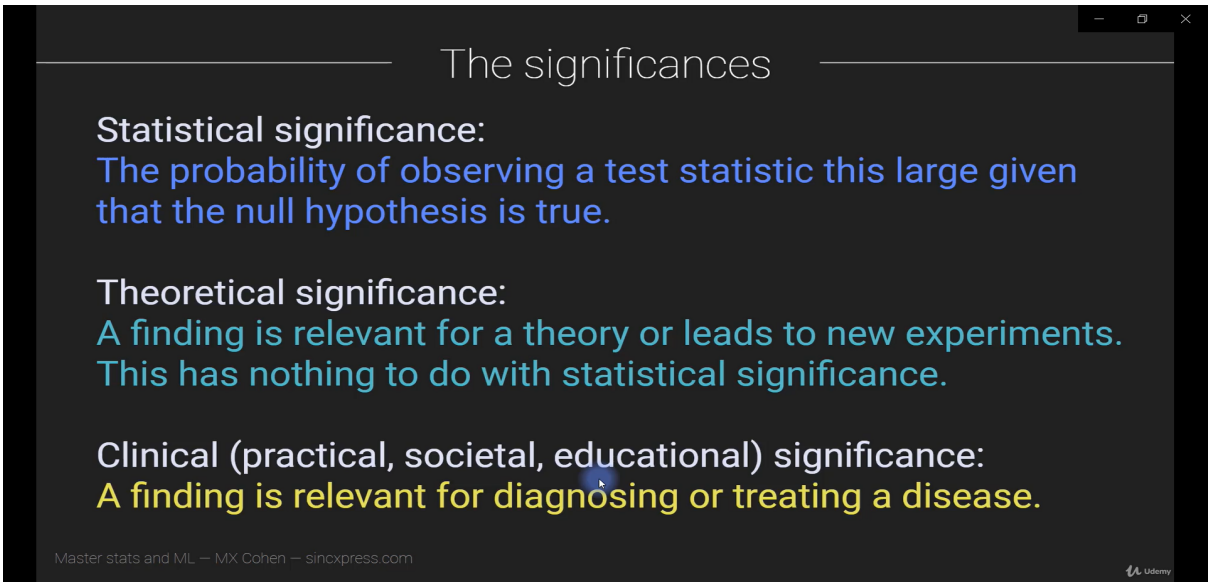



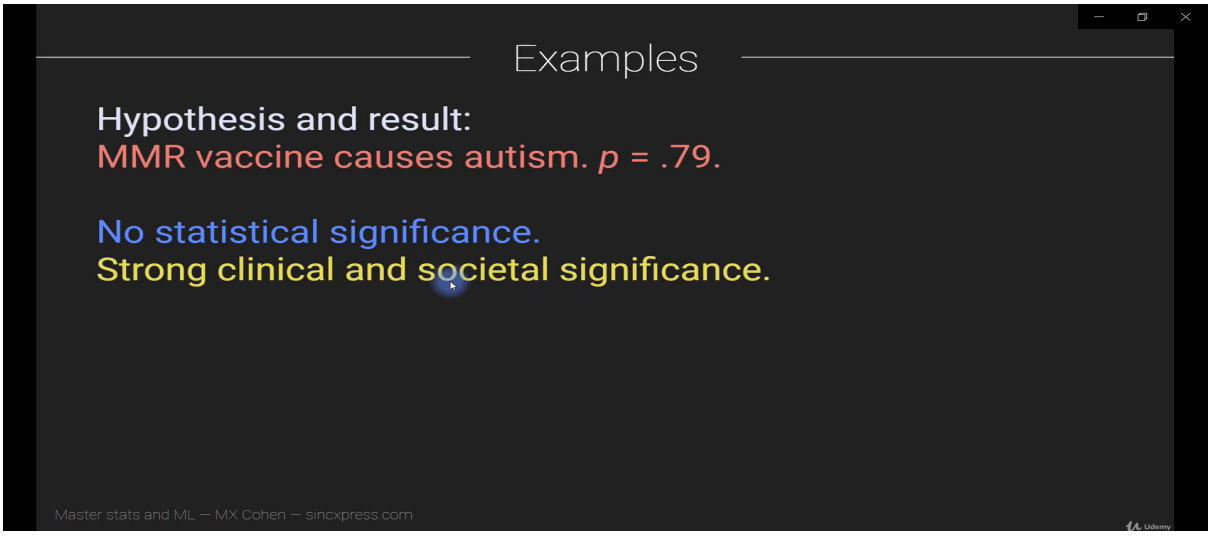

Now individual test should have more evidence in-order to reject the $H_0$ in favour of $H_1$

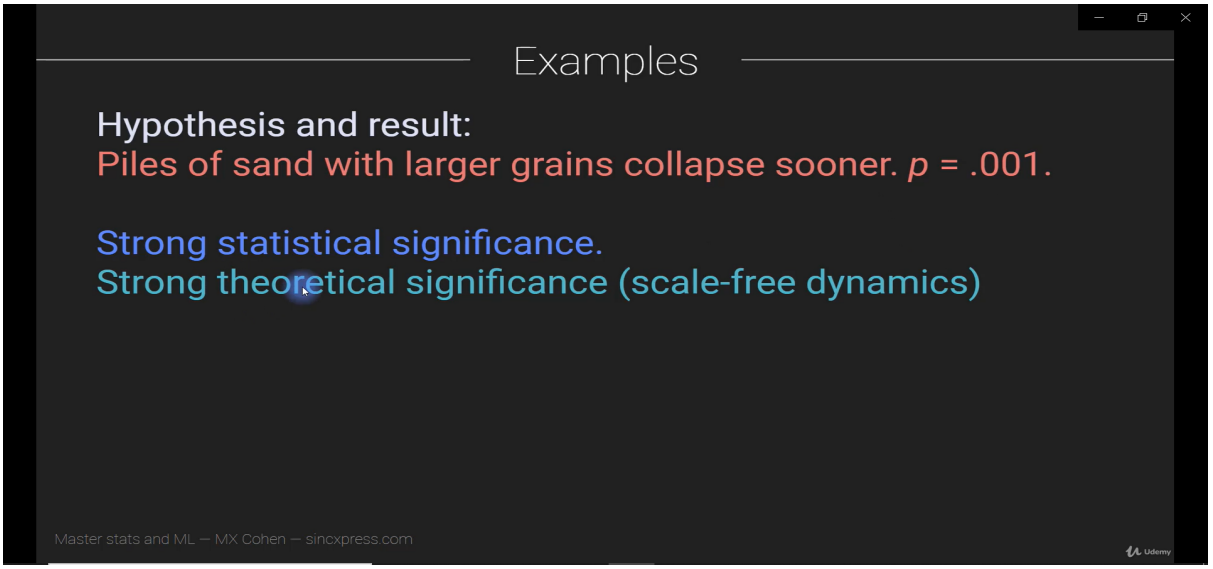


- Also note that Bonferroni correction is great and works well (when you have multiple tests and they are independent of each other).
- But there are many other correction methods which are more suitable for other specific situation like FDR (False discovery rate), cluster based correction (if you have spatially or temporally correlated data),...

### Type of Significances

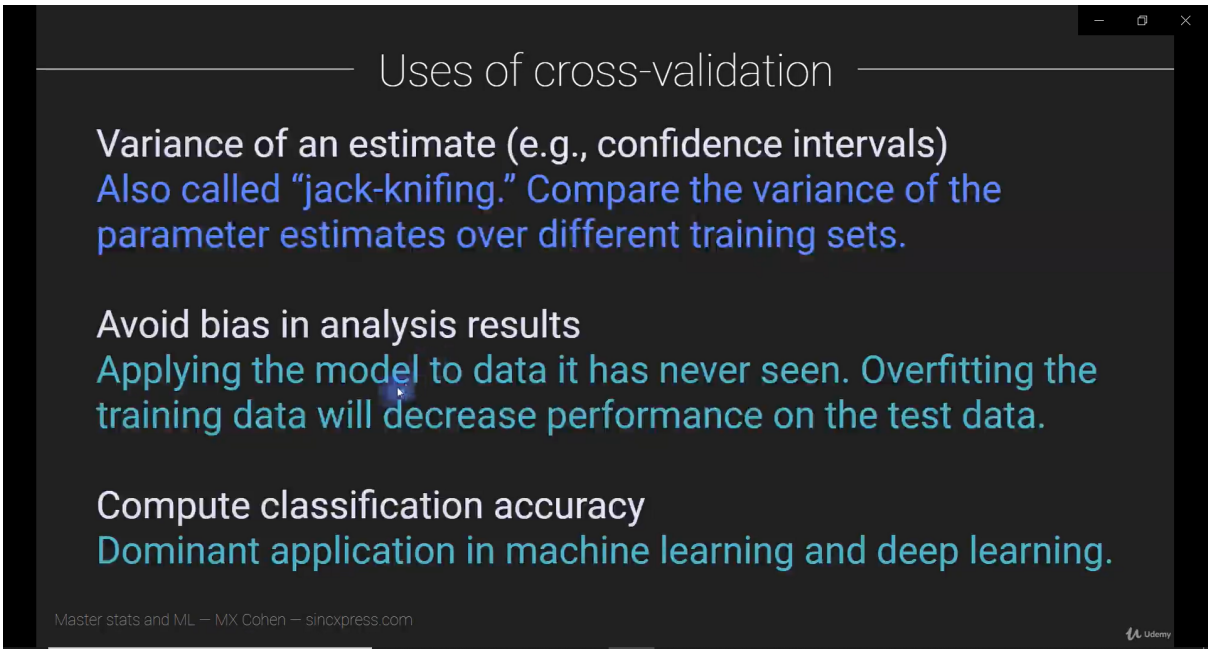


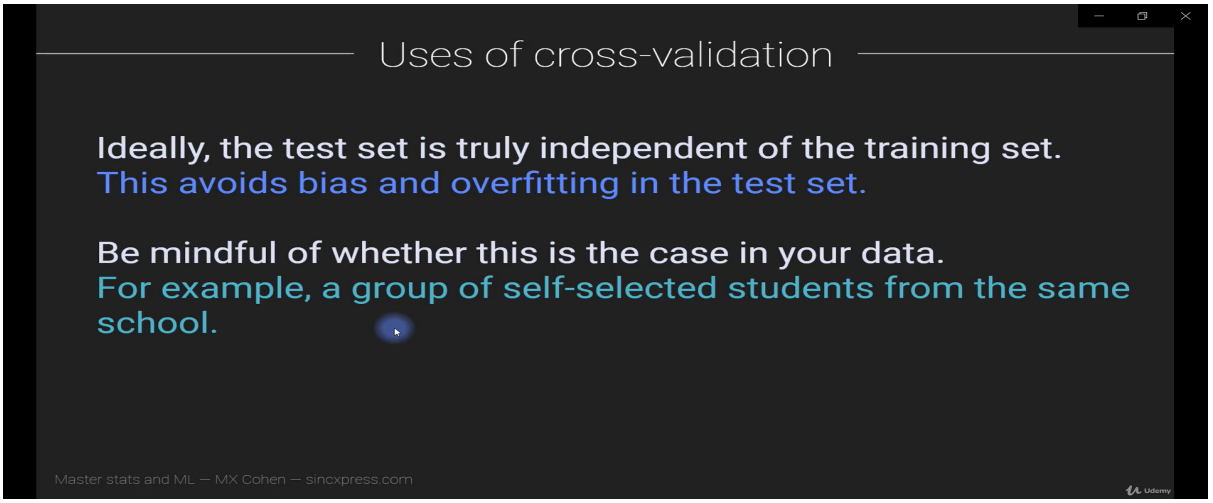






### Cross Validation

Nice Explanation: Machine Learning Fundamentals: Cross Validation




Uses of cross-validation
Ideally, the test set is truly independent of the training set.
This avoids bias and overfitting in the test set.
Be mindful of whether this is the case in your data.
For example, a group of self-selected students from the same school.
Master stats and ML — MX Cohen — sincxpress.com
Udemy

# 8. The t-test Family

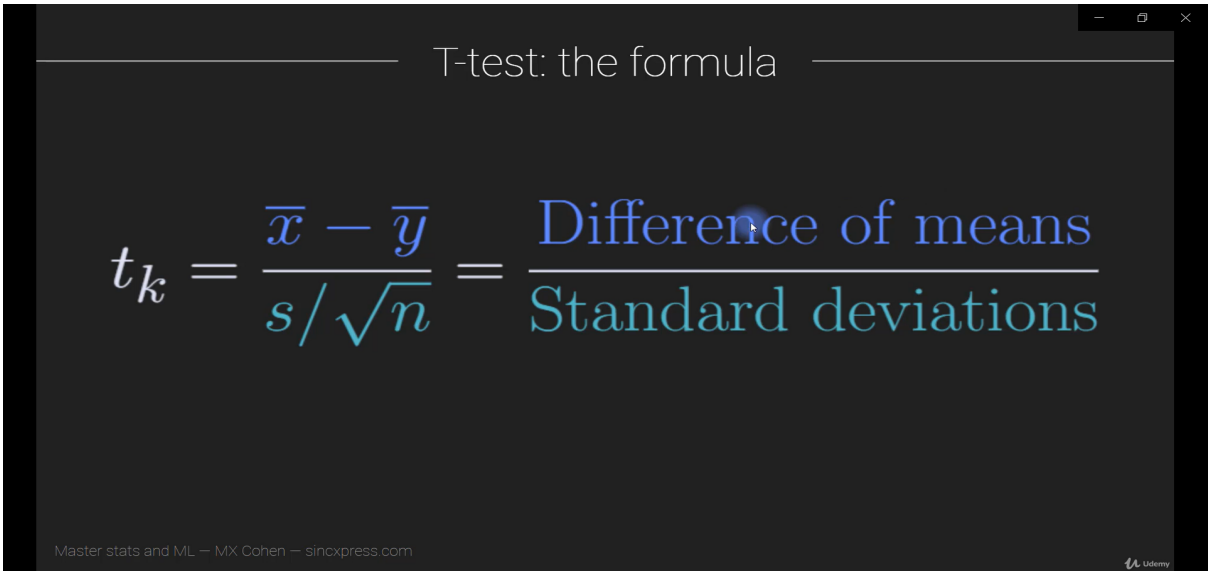
**Table of Content**

### One Sample T-test:

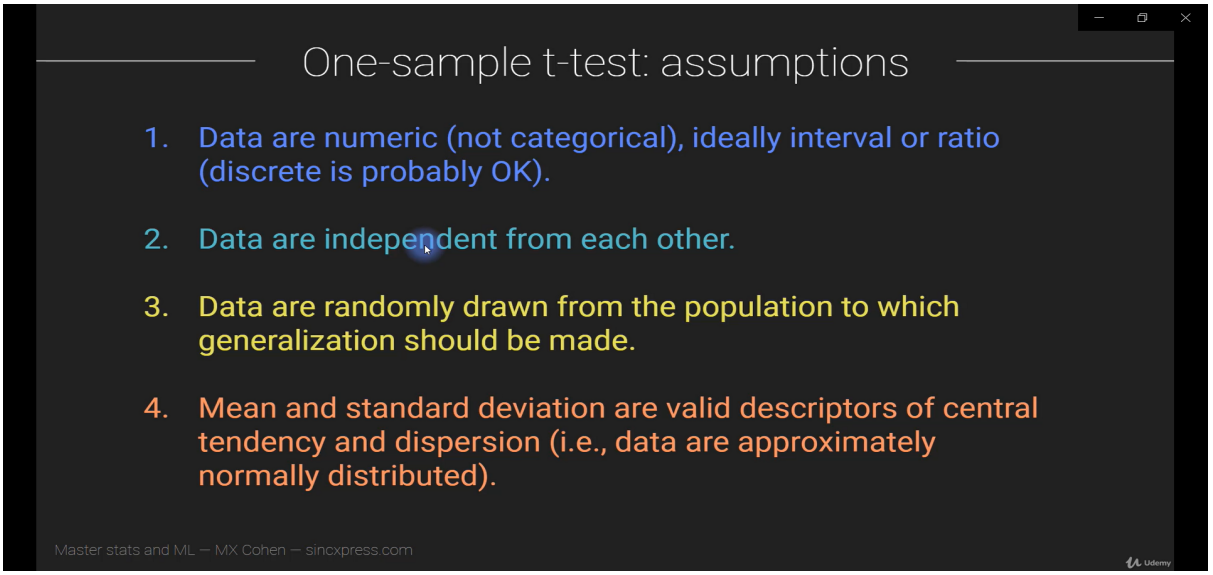
- One sample T-Test tests if the given sample could have been generated from a population with a specified mean.
- If it is found from the test that the means are statistically different, we infer that the sample is unlikely to have come from the population.
- **Example**: If you want to test a car manufacturer's claim that their cars give a highway mileage of 20 kmpl on an average. You sample 10 cars from the dealership, measure their mileage and use the one sample T-test to determine if the manufacturer's claim is true.
- **Purpose of One Sample T Test**
  The purpose of the One Sample T Test is to determine if a sample observation could have come from a process that follows a specific parameter (like the mean). It is typically implemented on small samples.
- **How did we determine One sample T-test is the right test for this?**
  Because, there is only one sample involved and you want to compare the mean of this sample against a particular (hypothesised) value.
- Null hypothesis will be that the given sample has been drawn from the specified population parameter, i.e there is no difference between sample statistic and population parameter.
- Depending on the how the problem is stated, the alternate hypothesis can be one of the following 3 cases:

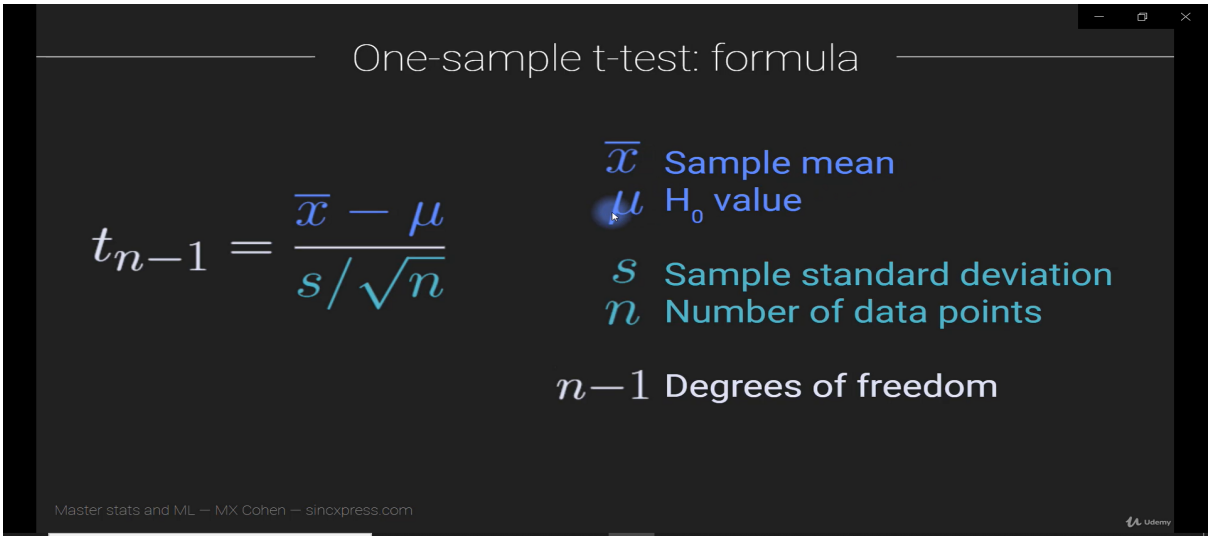
- **Case 1**: $H_1 : \bar{x} != \mu$. Used when the true sample mean is not equal to the comparison mean. Use Two Tailed T-Tests.
- **Case 2**: $H_1 : \bar{x} > \mu$. Used when the true sample mean is greater than the comparison mean. Use the Upper Tailed T-Tests.
- **Case 3**: $H_1 : \bar{x} < \mu$. Used when the true sample mean is lesser than the comparison mean. Use Lower Tailed T Test.

**The BASIC formula for the t-test is which gets modified based on different variations of t-test.**



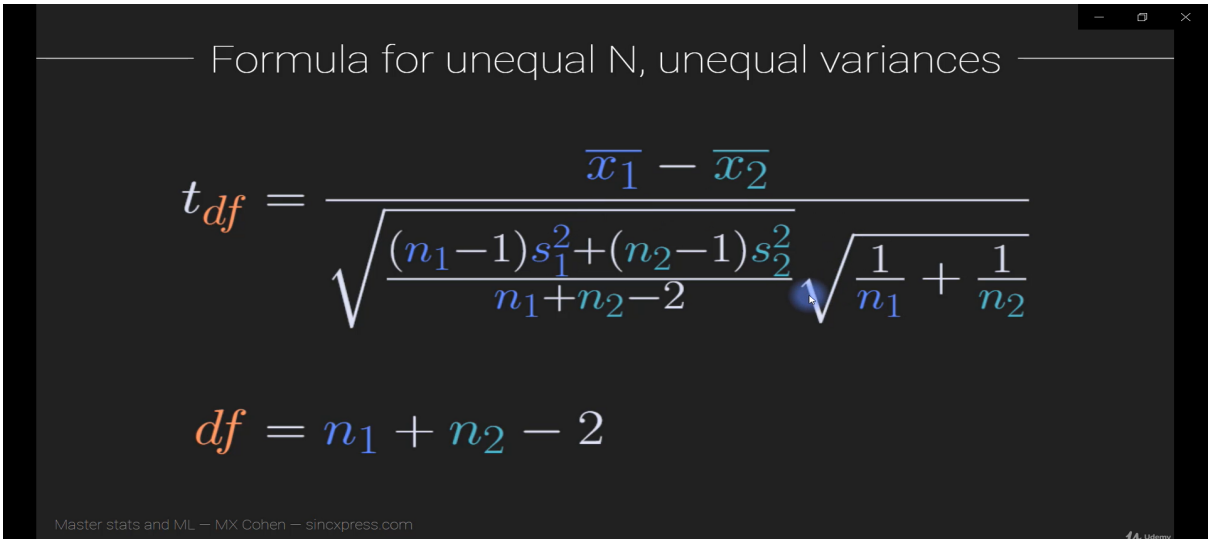
**The assumptions of one sample t-test**



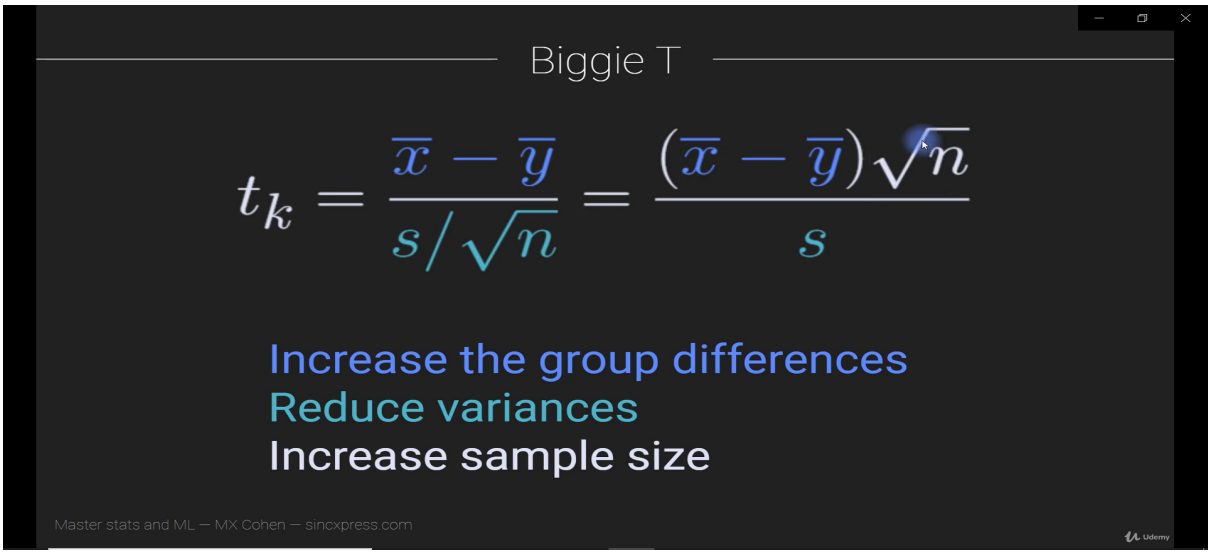




Code

Since in one sample t-test we have only one sample, so substitute n1 = n2 and s1 = s2 and check if the formula matches with one sample t-test.




### How to get Statistically Significant Results




### Two Sample t-test

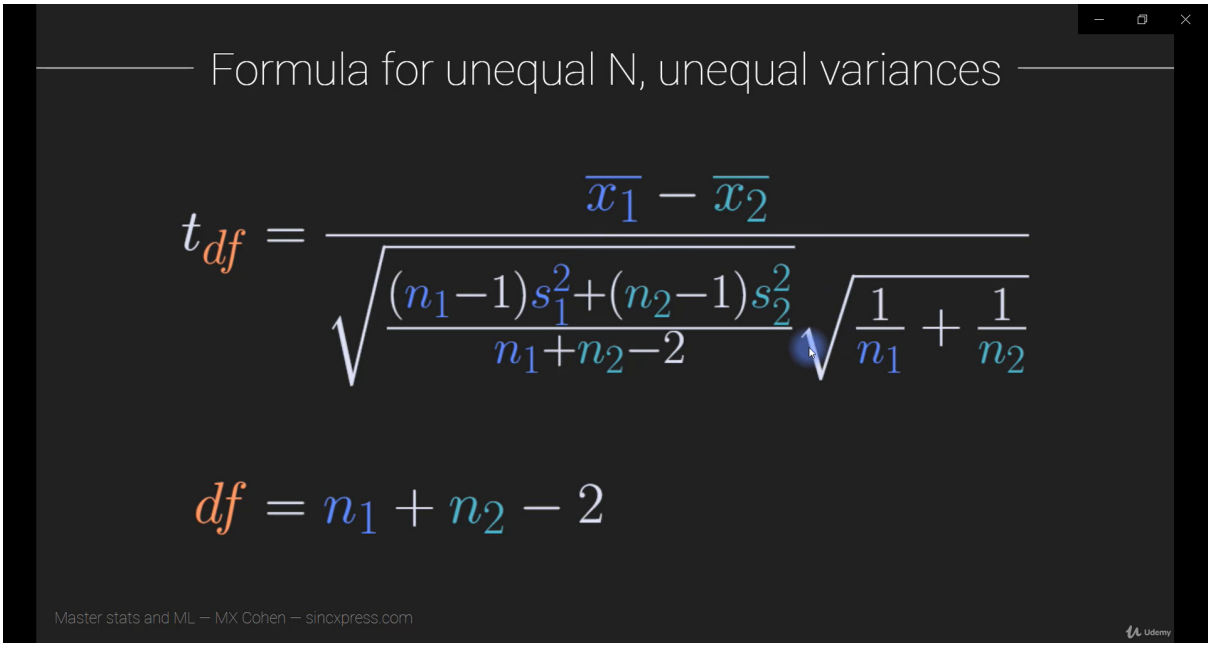
- It is a simple extension of the one-sample t-test.
- The numerator is pretty similar with the one sample t-test and the denominator depends on the characteristics of the data groups.
- Purpose: Test whether the two groups could have been drawn from the same distribution.
- i.e It is applied to compare whether the average (or any parameter) difference between two groups is really significant or if it is due instead to random chance.
- e.g It helps to answer questions like whether the average success rate is higher after implementing a new sales tool than before or whether the test results of patients who received a drug are better than test results of those who received a placebo.

i.e you collect data of group1 and group2 and compare them OR you take data before some effect (like we had drug) and then you again collect data after that effect. Now you want to compare them.

- There are several formulas for two sample t-tests. The numerator is always the same.
- *The denominator depends on whether the groups are paired or unpaired, groups have (ROUGHLY)same or different variances, groups have (ROUGHLY)same or different sample sizes.*
  (unequal sample sizes apply only to the unpaired groups.)

**Paired and unpaired groups:**

- **Paired groups**: We compare two groups data collected from the same individual .(like data of health before taking drug and data of health after drug on same individual)
- **Unpaired groups**: We compare the two groups data collected from different individuals. data of effect of drug on men's and data of effect of drug on females.

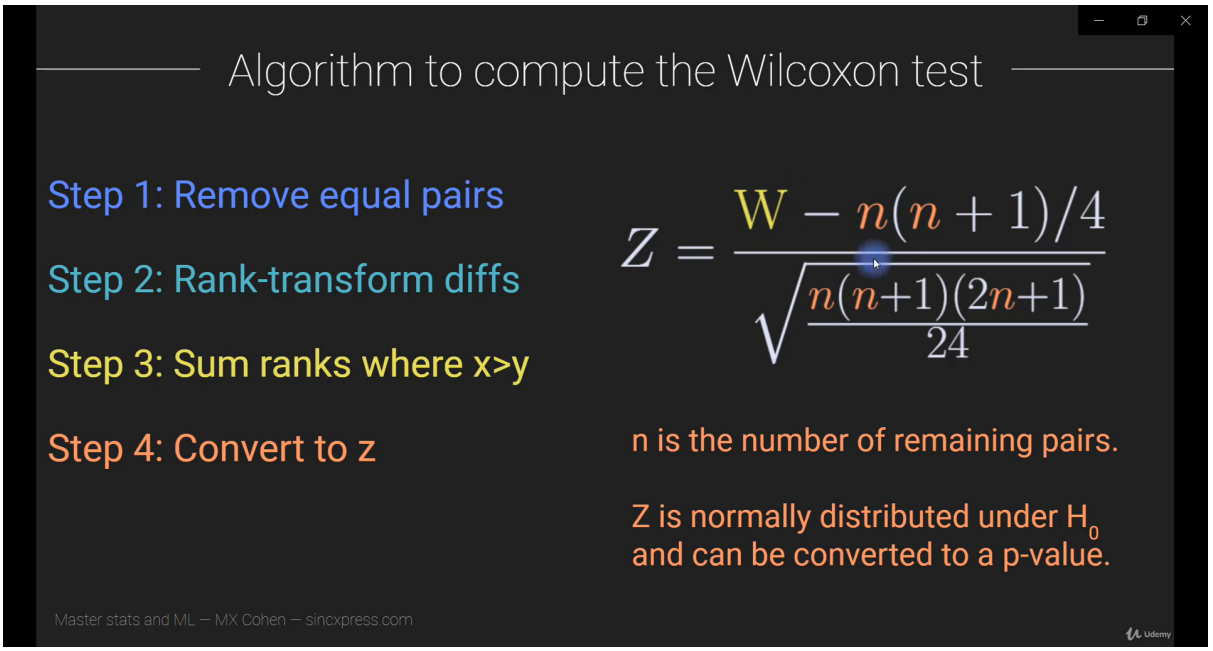



- df = n1 + n2 - 2
  Because for the first group, df = n1 - 1 and for second group, df = n2 - 1, therefore overall df = n1 + n2 - 2
- Assumptions are the same as one Sample T-test.

Code

**Wilcoxon Signed Rank Test (also called Signed Rank Test) (Non parametric t-test):**

- It is a non-parametric alternative to the one sample and two sample PAIRED ONLY t-test.
- Mainly used when data is not close to gaussian AND For ordered (ranked) categorical variables without a numerical scale.
- Tests for the differences in the medians (instead of differences in the mean)




**Algorithm:**

- **Remove the equal pairs**
    1. If we are doing two sample paired test, then remove pairs which are equal to Ho value (null hypothesis value)
    2. If we are doing one sample test, then remove data points which are equal to Ho value (null hypothesis value)
    3. Reasoning: Equal pairs do not contribute to the test either way.
- **Rank Transform diffs**
    1. r = rank( |x - y| )
    2. If it is two sample test, then x and y are pairs
    3. If we are doing one sample test, then y is replaced by Ho (null hypothesis value)
- **Sum the rank where x > y,**

    sumR = 0
    for i=1 to n:
        sumR += ri, if xi > yi

    W = sumR
- **Convert to Z**

    (Z is normally distributed under Ho and can be used to find out p-values)

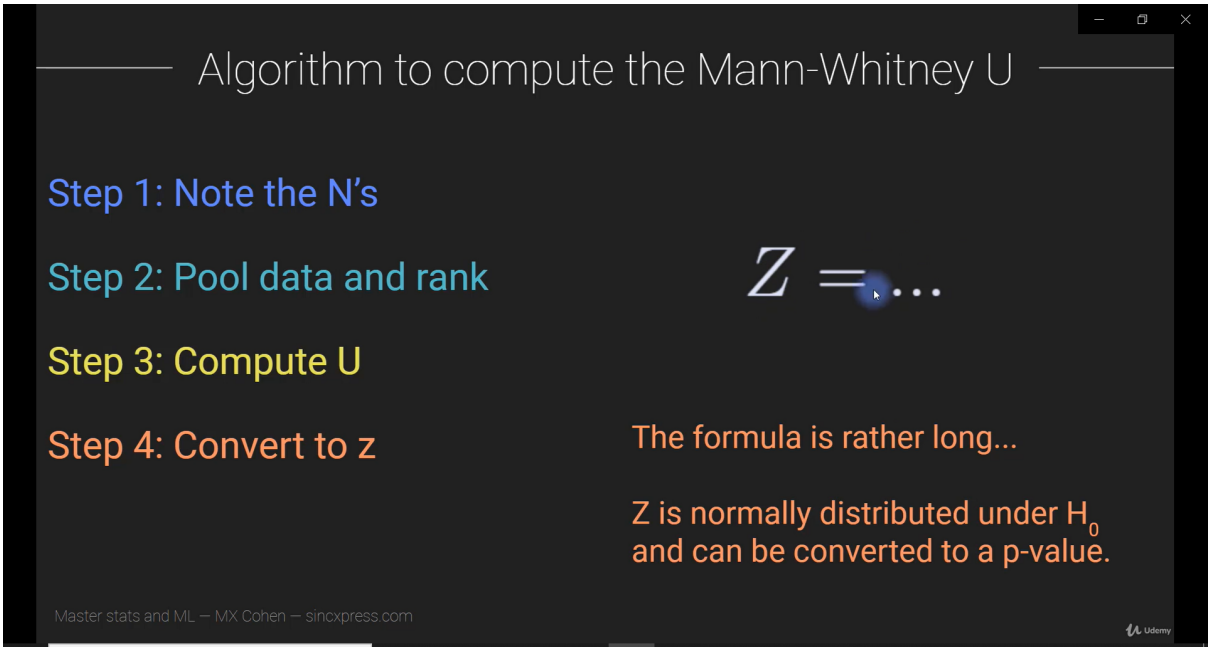
,**Null hypothesis:** Both the data comes from the distribution (The null hypothesis for this test is that the medians of two samples are equal.)
**Alternative hypothesis:**
two tailed ( m1 != m2 )
one-tailed ( m1 < m2 OR m1 > m2 )

For paired non-parametric t-test example:
https://www.statisticshowto.com/wilcoxon-signed-rank-test/

Code

**Mann-Whitney U Test (Also called Mann-Whitney-Wilcoxon t-test OR Wilcoxon Rank-Sum test) (Non parametric t-test):**

- It is a non-parametric alternative to the two sample UNPAIRED t-test. (i.e two independent groups)
- Used when data is not normal and the dependent variable is either ordinal or continuous, but not normally distributed.
- Tests for the differences in the medians (instead of differences in the mean).
- It can have different sample sizes.




**Algorithm:**

- Do the assignments
  xf = dataset with fewer data points
  xm = dataset with more data points
  nf = number of data points in xf
  nm = number of data points in xm.
- Concatenate the xf and xm (note: xm should get concatenated after xf). After concatenation, compute rank
  r = rank( xf.append(xm) )
- Compute U (sum ranks upto nf)
  U ← sum( r[:nf] )
- Compute Z (Z is normally distributed under Ho and can be used to find out p-values),

**Null hypothesis**: Both the data comes from the same distribution (The null hypothesis for this test is that the medians of two samples are equal.)
**Alternative hypothesis:**
two tailed ( m1 != m2 )
one-tailed ( m1 < m2 OR m1 > m2 )

Code

**Permutation Testing for t-test:**

- The key difference between parametric and permutation based t-test significance is, in general parametric approach is, you evaluate the position of your test statistic relative to the analytic distribution *(based on your assumptions and this is not something that you compute by running some analysis i.e you are not generating the null hypothesis distribution on your own)* under the null hypothesis. That is standard parametric approach (like t-tests, correlation, etc)
- In contrast, in permutation testing, I don't want to make any assumptions like where the null hypothesis distribution comes from OR what it looks like. Instead, I want to compute the **null hypothesis distribution based on the data I have** and then interpret the value of the observed test statistic relative to the empirical distribution.

**Q. How do we create empirical hypothesis distribution ?**

- Say we had 7 data points, 4 from group1 and 3 from group2.
  group1 = {1,2,3,4}
  group2 = {5,6,7}

- **Step1** → Mix the data together so that there are no labels on them
  pooledData = {1,2,3,4,5,6,7}

- **Step 2** → Randomly assign labels (group1 OR group2) to these data points, but the ratio of those labels should be the same (i.e there should be 4 points for group1 and 3 points for group2). Note: We are not modifying any value of the data points, just randomly relabeling them
  group1 = {1,4,5,6}
  group2 = {2,3,7}

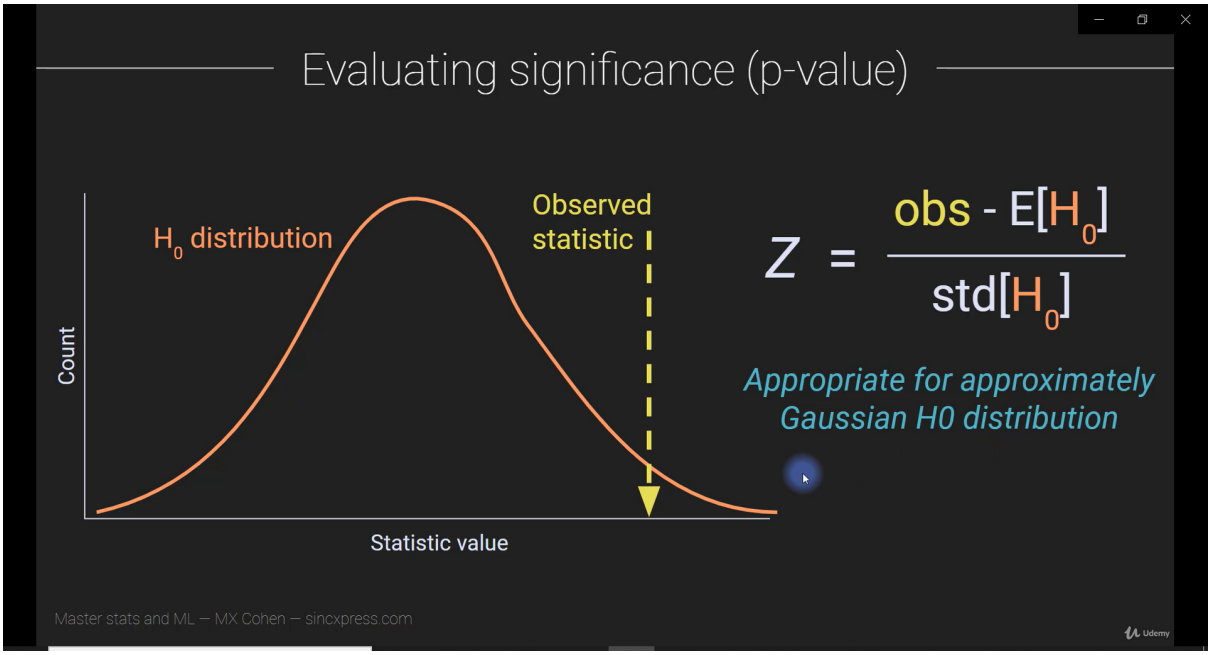
- **Step 3** → calculate t value from these groups (t-test for two unpaired samples) and store it.

- Now repeat the steps 2 and 3, say N number of times.

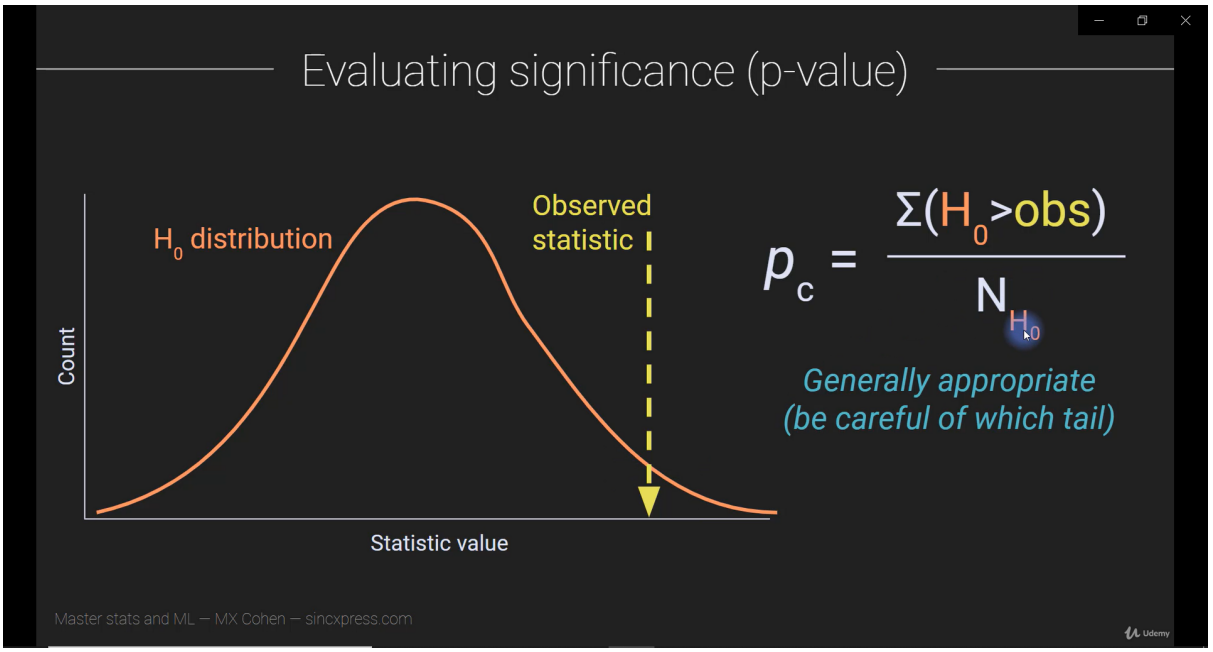
- **Step 4** → You will get N t-values. You plot those t-values and this is your null hypothesis distribution.
- **Step 5** → Now, you go back to your original data and perform t-test between your two groups and evaluate your test statistic with respect to the null hypothesis distribution you calculated.

Now for evaluating the p-values, there are two ways,

This way is appropriate if you distribution is approximately gaussian.




Second Way:
pc = probability counts, basically you take all the possible Ho values which are greater than observed test statistic and divided by total possible values of Ho



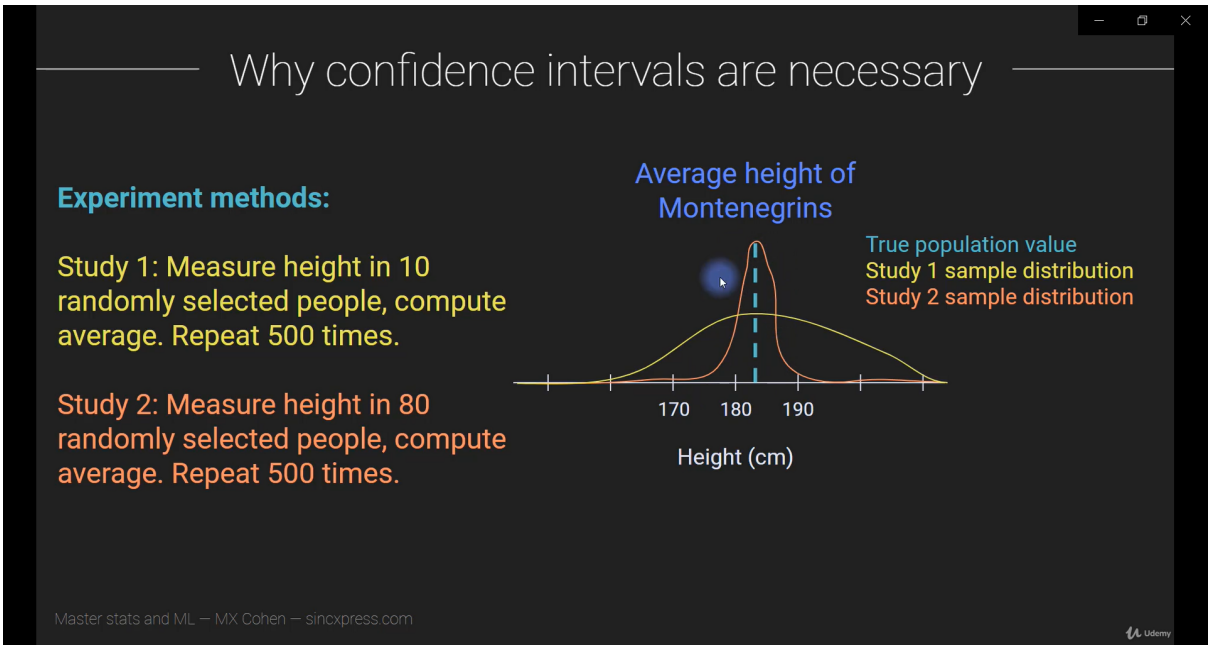

Code

# 9. Confidence Intervals on Parameters

**Table of Content**

**Confidence Intervals and why do we need them:**

- say the population mean of heights of Montenegrins is 182.
- We perform 2 experiments now:
    1. Measure the height of 10 randomly selected Montenegrins and store the average. Repeat this experiment 500 times.
    2. Measure the height of 80 randomly selected Montenegrins and store the average. Repeat this experiment 500 times.
- plot the distribution of means for both the experiments.




→ **Which distribution from the above two experiments will be more confident?**
obviously (qualitatively visually) distribution 2, the distribution of means from experiment 2 because it will be narrower as compared to distribution from experiment 1 because we measured more people in experiment 2.
( see their mean of means is same but the spread is different )

This is the base idea of confidence intervals, in experiment 2, we measured more people data and got a narrow distribution and in experiment 1, we got a wider distribution (that's why we are less confident).

**Naive Definition:** The probability of unknown population parameter falls within the range of values after repeating the experiment multiple times.

**NOTE:**

1. *The confidence interval does not guarantee that the true mean (population parameter) will lie between sample estimate and confidence intervals.*
2. *In fact, it just tells our confidence on our estimation of our sample estimate, if we repeat this same experiment many times in future, about 95% of times, the true mean will lie between sample estimate and confidence intervals.*

**Mathematically Confidence Interval:**
P(L < mu < U) = c
L = Lower bound
U = Upper Bound
c = Confidence
mu = population parameter
P = probability

Typical confidence intervals probabilities: 95%, 99%, 90%.

### Factors influencing the confidence intervals:

1) **Sample size:** as sample size increases, the confidence intervals decreases
2) **Variance:** as variance decreases, the confidence interval also decreases.

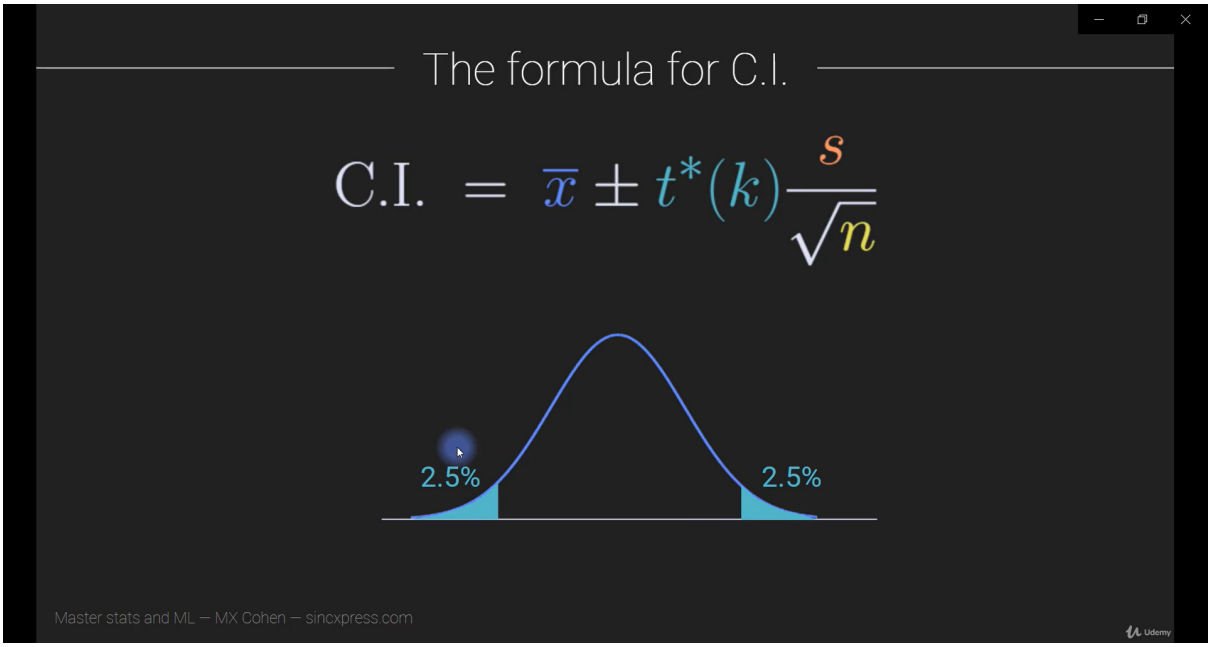
### There are two methods to compute the confidence intervals:

1) By applying formula
2) By using an empirical bootstrapping or resampling method.

### Confidence Intervals using formula:

- Since, typically we are not able to repeat the experiments multiple times due to some reasons and we just have one sample. In that case, the formula method is used typically.
- Formula:
  Confidence Interval (CI) = x_bar +- $t^*_k$ * ( s / sqrt(n) )
  x_bar = sample mean
  s = sample std
  n = sample size
  $t^*_k$ = t-value with k-degrees of freedom (we will talk more about this)

  This t* is not the same kind of value that we have been learning in the previous section, it does come from t-distribution, but this is not the test statistic.




t* = t-value associated with the one tail of the confidence interval = (1-c) / 2
t* = t_inverse( (1-c) / 2, n-1 )
t_inverse is basically some hypothetical function which takes the area under the curve (probability) and the degrees of freedom (n-1) and returns the corresponding t-value in the t-distribution.

e.g c=95%, n=20
t* = t_inverse( (1-0.95) / 2, 20-1 ) = 2.093
The reason we are dividing by 2 is because we want the confidence intervals to be two tails. (both the sides).

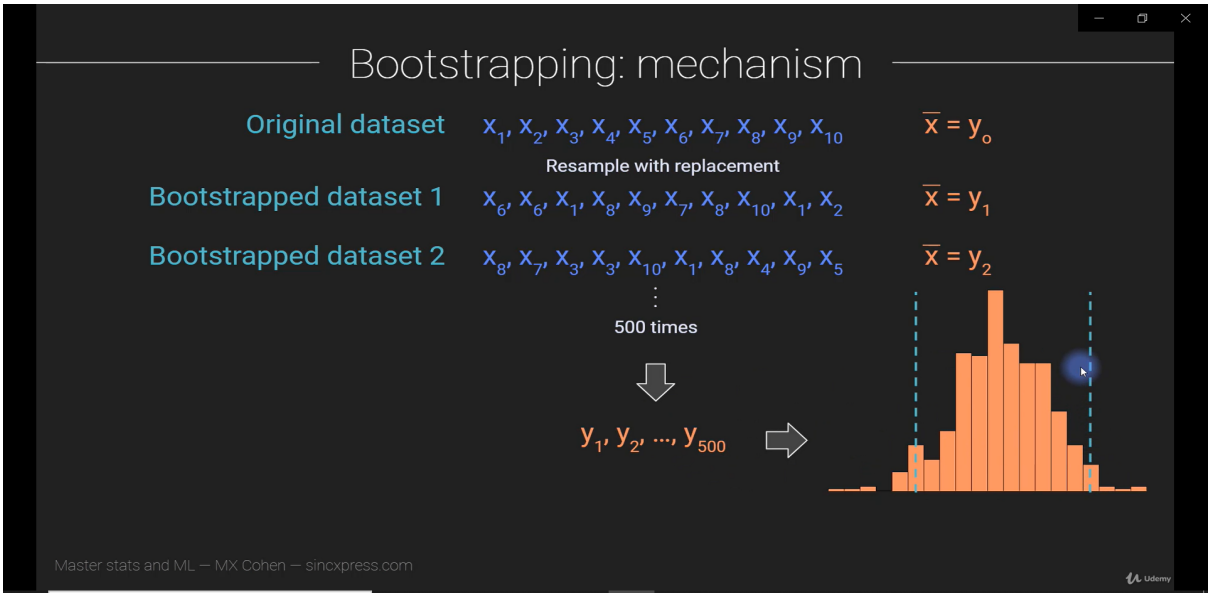
**Assumptions of this Formula:**
1) s is an appropriate measure of dispersion in data. **That's why it is parametric way of computing confidence intervals.**

**Code**

### Confidence Intervals using Bootstrapping:

- Non parametric method for computing confidence intervals.
- Instead of using the formula for computing confidence intervals, we can compute them directly by using data.
- This is done by repeatedly randomly sampling the dataset.
- Thus, pretend that your sample is the population and your resampling is the sample.
- Mechanism:
  say you had 10 points with mean (y0)

  You randomly sample with replacement data of size 10 (the size can be smaller than actual n OR equal to n) and compute the mean. say y1.
  You repeat this experiment 500 times and you will get 500 means (y1,y2,y3,...,y500)
  You plot those means and compare the mean of this distribution with y0.
- Now you can also find out 2.5 percentile and 97.5 percentile of distribution, which will give you the empirical estimation of confidence interval.




**Pros and Cons of Bootstrapping:**
**Pros**:
1) Works for any kind of parameter (mean, variance, correlation, median, etc.)
2) Do not assume any normality.
3) Useful for limited data.
**Cons**:
1) Gives (slightly) different results each time.
2) Can be time consuming for large datasets (even though you can use bootstrap sample size < N(size of dataset) )
3) Sample must be a good representative of the population.

Code

# 10. Correlation

**Table of Content**

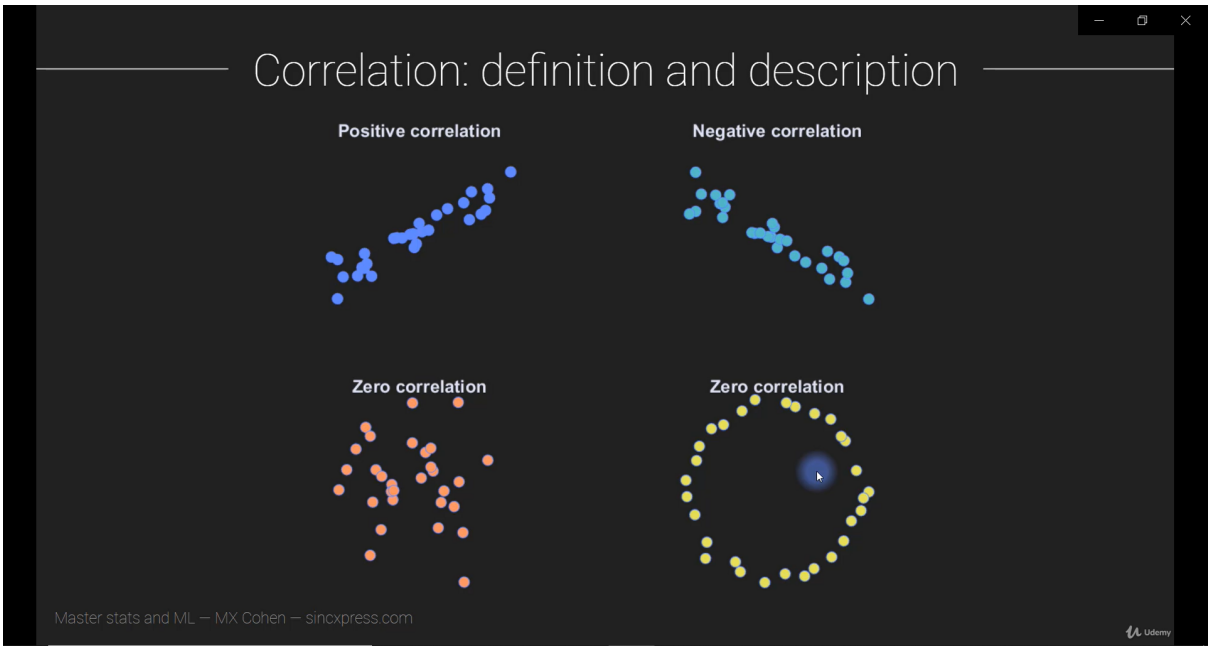


**Correlation**:

**Motivation and Description of correlation.**

- In this,  we basically check how the two groups are related to each other.
- Correlation analysis computes the correlation coefficient (r)
- It is standardised between -1 and 1. -1 indicating perfect inverse relationship, 0 indicating no relationship and 1 indicating perfect relationship.
- The r is a single number that shows the LINEAR RELATIONSHIP between two groups.
- WE SHOULD NOT JUST COMPUTE THE CORRELATION COEFFICIENT ONLY. WE SHOULD ALWAYS COMPUTE p-value TO INTERPRET ITS STATISTICAL SIGNIFICANCE.

Examples:




**Why does the fourth figure have zero correlation even though the figure seems to have a relation between x and y ?**
→ There is zero correlation because *correlation measures the linear relationship*. There is a non linear relationship.

- Null hypothesis: there is zero correlation.
  Alternative Hypothesis: There is non-zero correlation.

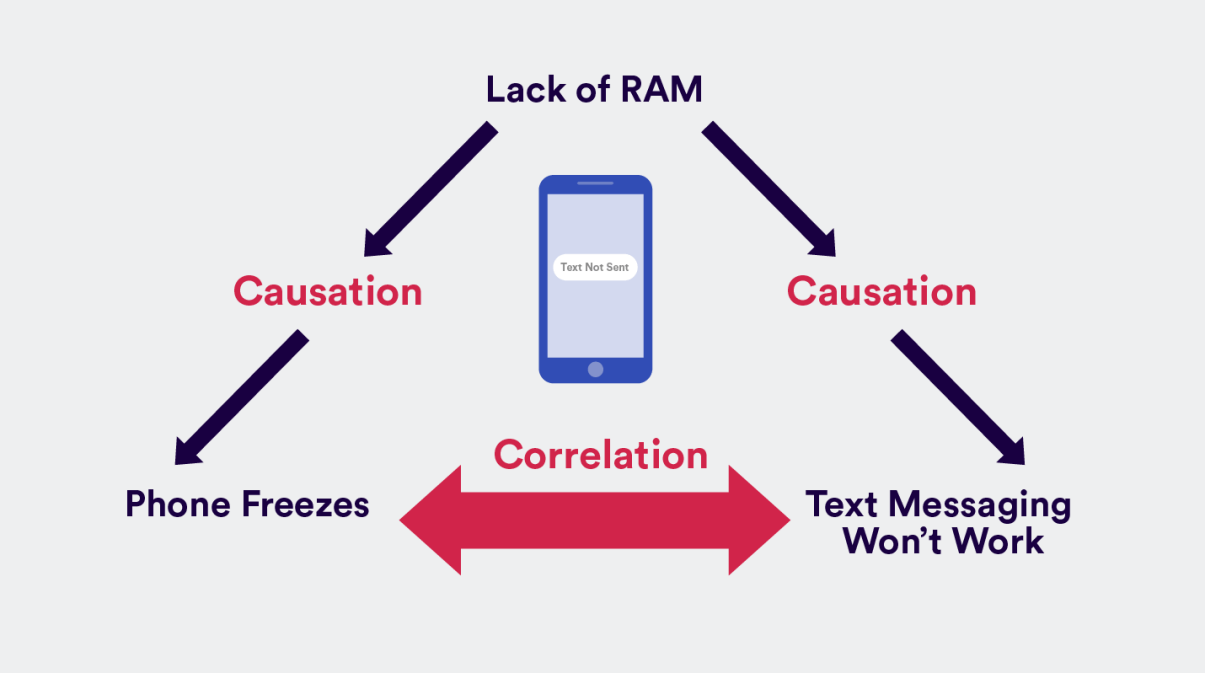
### What is Causation?

- Causation is implying that A and B have a cause-and-effect relationship with one another. You're saying A causes B. Causation is also known as causality.
- Firstly, causation means that two events appear at the same time or one after the other.
- And secondly, it means these two variables not only appear together, the existence of one causes the other to manifest.

### Correlation vs Causation:

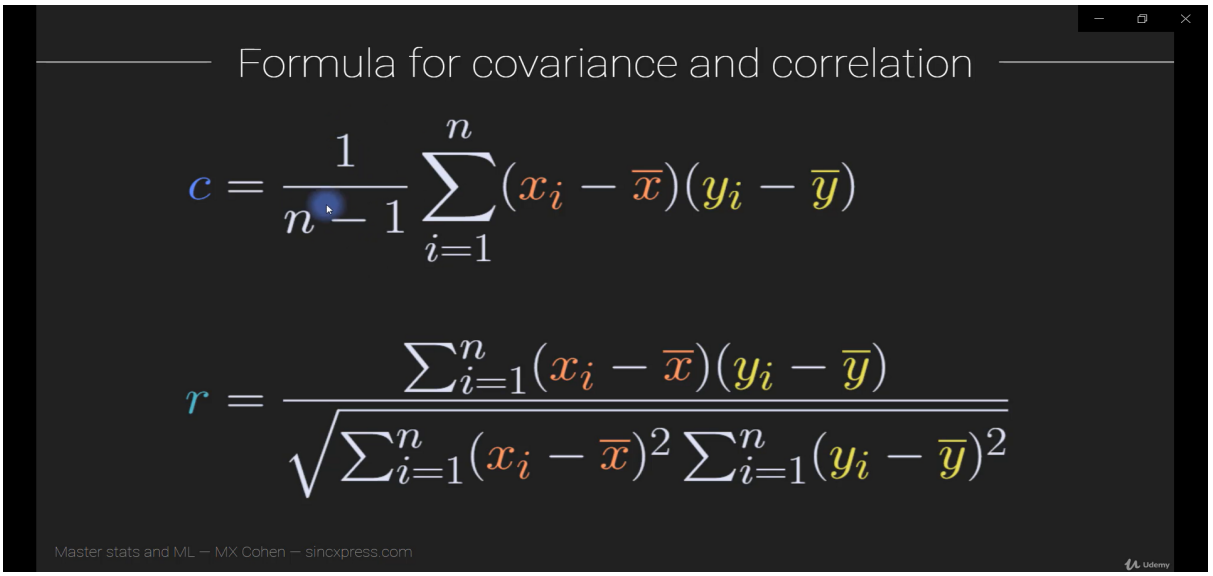
- Correlation just measures the relationships.
- It does not reveal or imply causation.
- Causality can be demonstrated by experimental manipulations.
- Example
  Ref: https://clevertap.com/blog/correlation-vs-causation/

  My mother-in-law recently complained to me: "Whenever I try to text message, my phone freezes." A quick look at her smartphone confirmed my suspicion:
  she had five game apps open at the same time plus Facebook and YouTube. The act of trying to send a text message wasn't causing the freeze,the lack of RAM was. But she immediately connected it with the last action she was doing before the freeze. She was implying a causation where there was only a correlation.

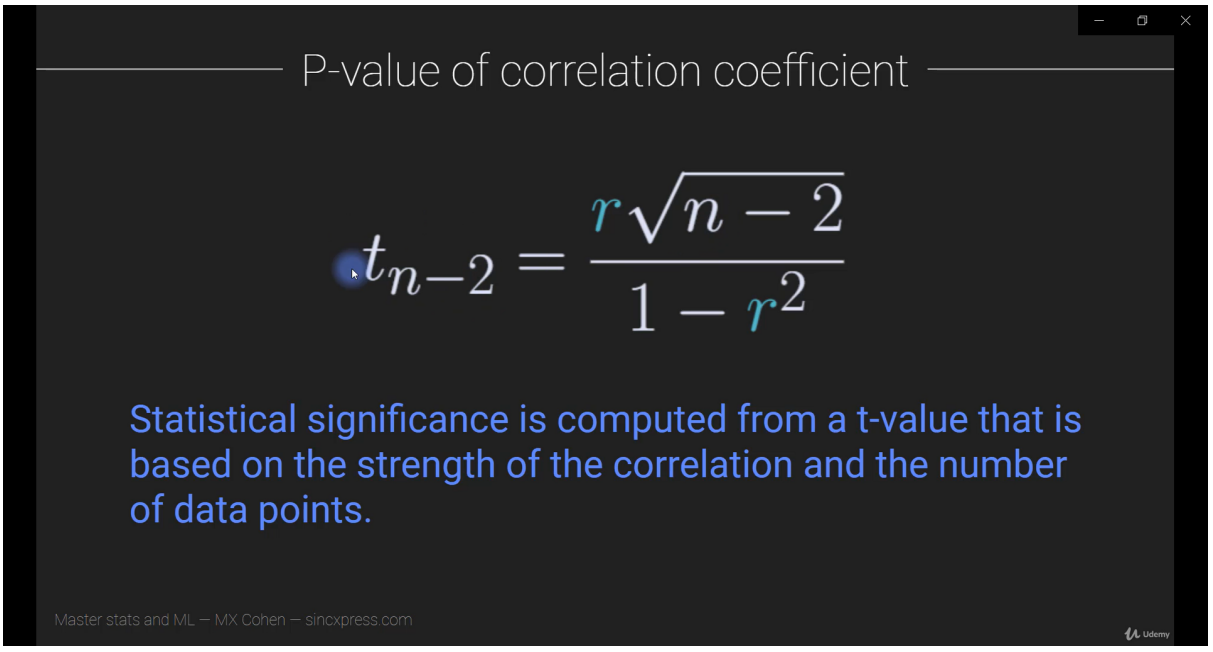


**Covariance and Correlation:**

- Covariance is a single number that measures the linear relationship between the two variables.
- Correlation is just normalised covariance. (from -1 to 1)
- Division by n-1 is not as critical, it depends on what you are doing with covariance.



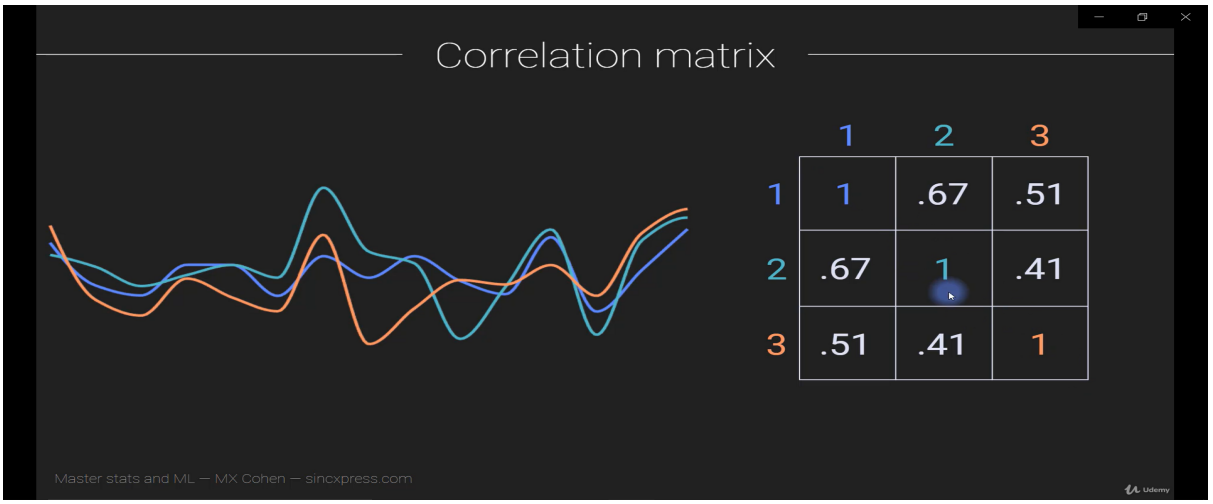

- But remember, we should not blindly trust the correlation value, we should check the p-value, is it really statistically significant ?

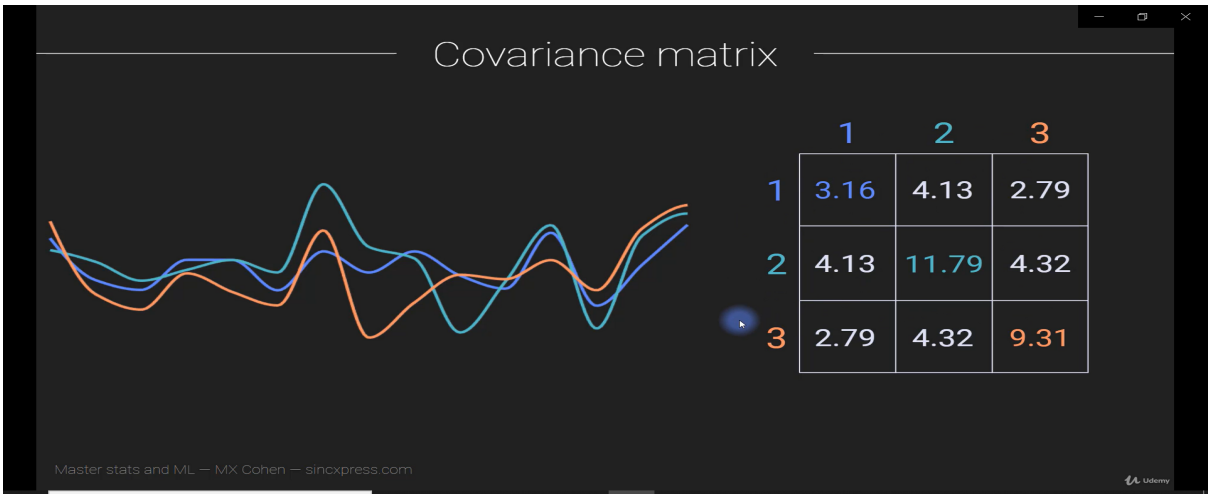



- Say, we are increasing the sample size, but assume that r will always be constant, then see the t-value will also increase i.e being more significant.
- Just an observation, when the mean of both the variables is zero and variance is 1, then the covariance formula is the same as correlation formula.

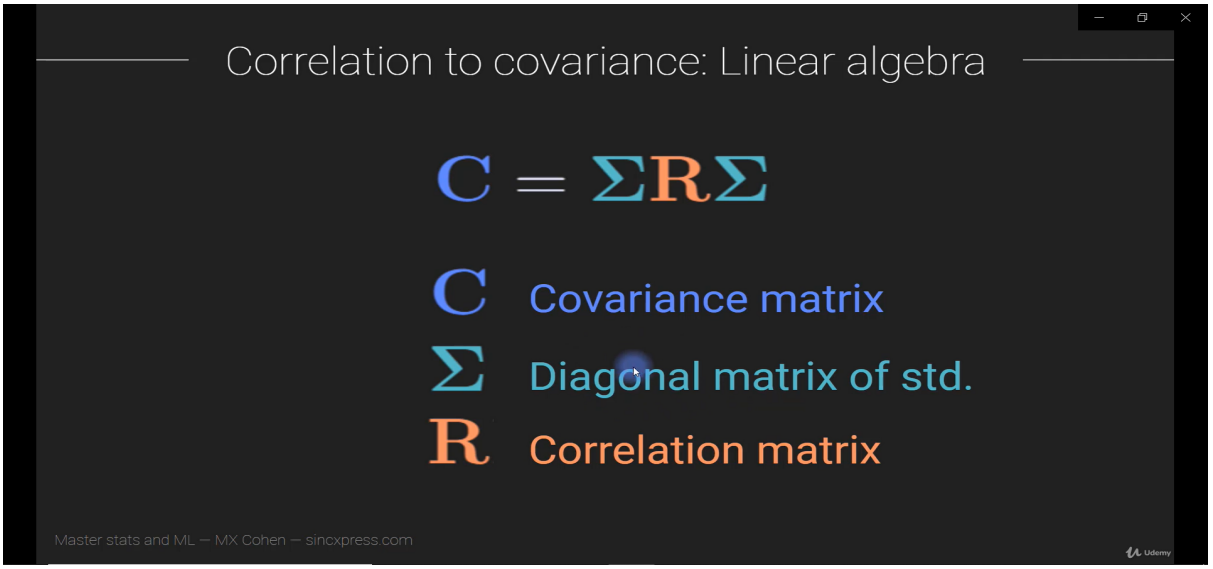
Code

**Correlation Matrix:**



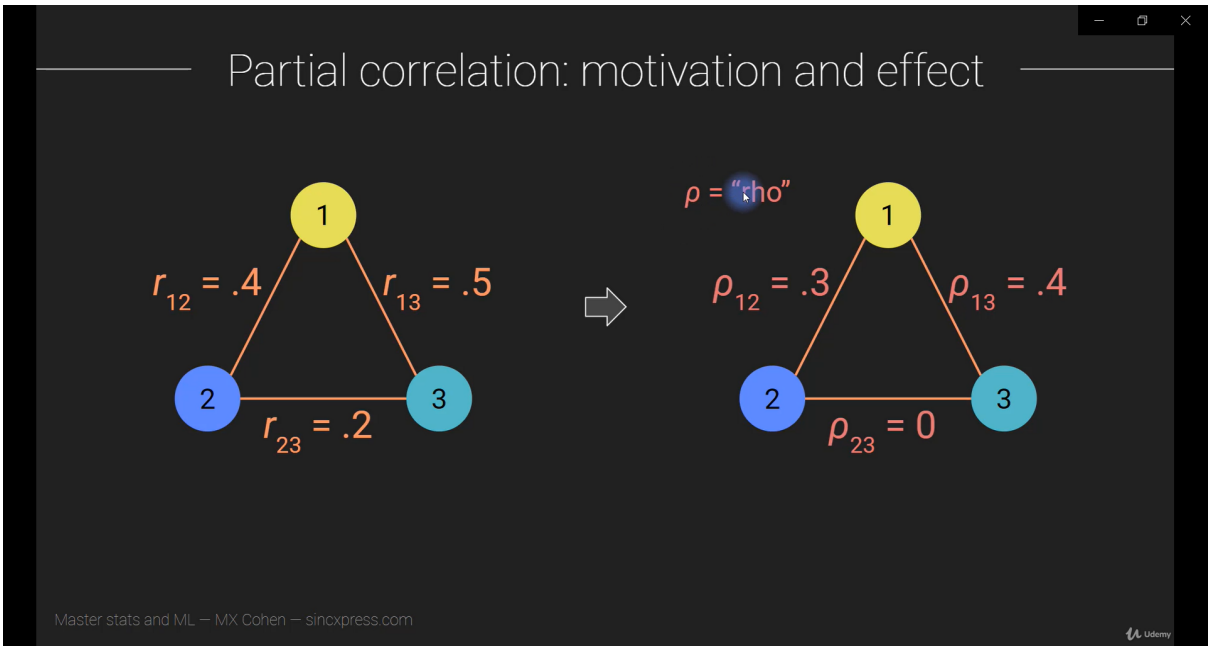

**Expressing Covariance Matrix in terms of Correlation Matrix**




**Code**

**Partial Correlation:**

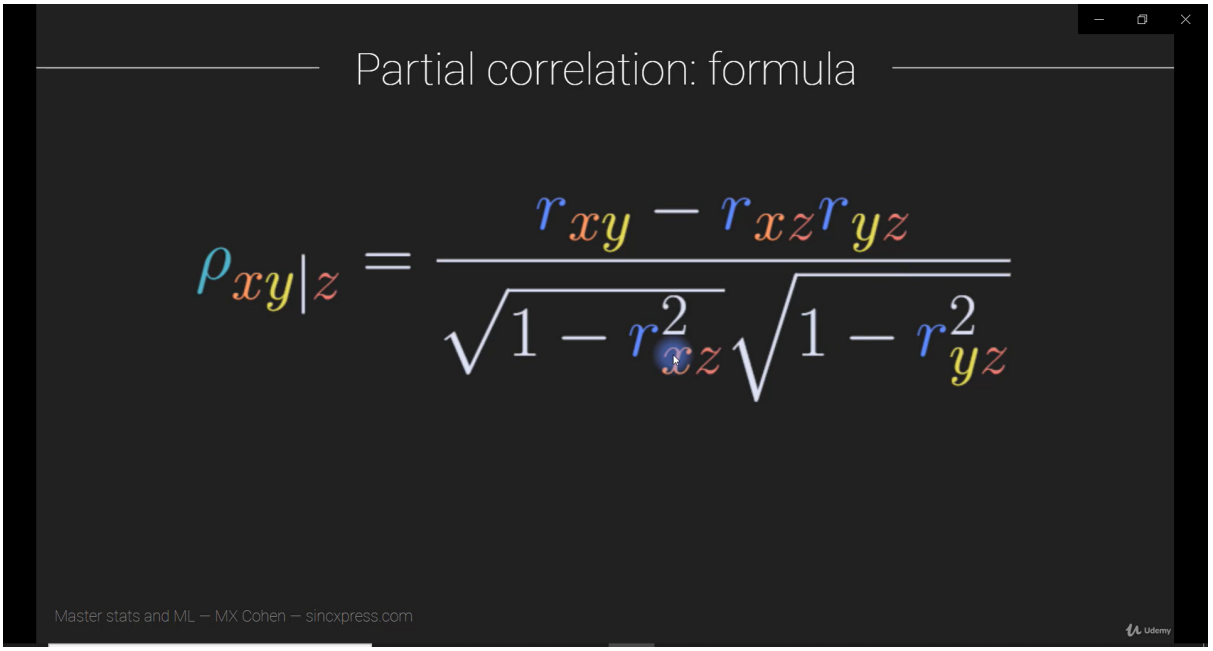
- Motivation: Say you had three variables, weather, ice cream consumption and shark attacks.
- From the data, it has been found that, as the ice cream consumption increases the shark attacks also increases. i.e there is some correlation between them. This sounds weird right.
- From the data ( hypothetical ) :
- The correlation between:
    1. weather and ice cream consumption = 0.4
    2. ice cream consumption and shark attacks = 0.2
    3. weather and shark attacks = 0.5




(check the above image, where node 1 = weather, 2 = ice cream consumption, 3 = shark attacks)

- *But actually, what is happening is, as the weather is hot, people consume more ice cream and go to the beach and as people go to the beach, more shark attacks happen.*
- *Therefore, the weather caused the increase in ice cream consumption which led to an increase in shark attacks.*

Now the **idea behind partial correlation** is, what is the correlation between variable A and variable B if I keep aside the effect of variable C.
Therefore, partial correlation between:
1) weather and ice cream consumption = 0.3
2) ice cream consumption and shark attacks = 0
3) weather and shark attacks = 0.4




- Intuitively, the formula for partial correlation makes sense because,
- say we wanted p(xy|z) i.e we are interested in the correlation between x and y by removing the shared variance of z,
- In the numerator of formula, it is $r_{xy}$ - $r_{xz}.r_{yz}$
  $r_{xy}$ = correlation between x and y, in which we are interested.
  $r_{xz}$ = correlation between x and z.
  $r_{yz}$ = correlation between y and z.
  i.e removing(subtracting) the shared variance of z on x and y.
- And denominator is just some normalisation factor.

In the formula screenshot, we just partialled out the single variable, but it is also possible to waive out the many variables as well.
say we wanted, p(xy | z,w)
i.e we want correlation between x and y, by removing the shared variance of z and w, then

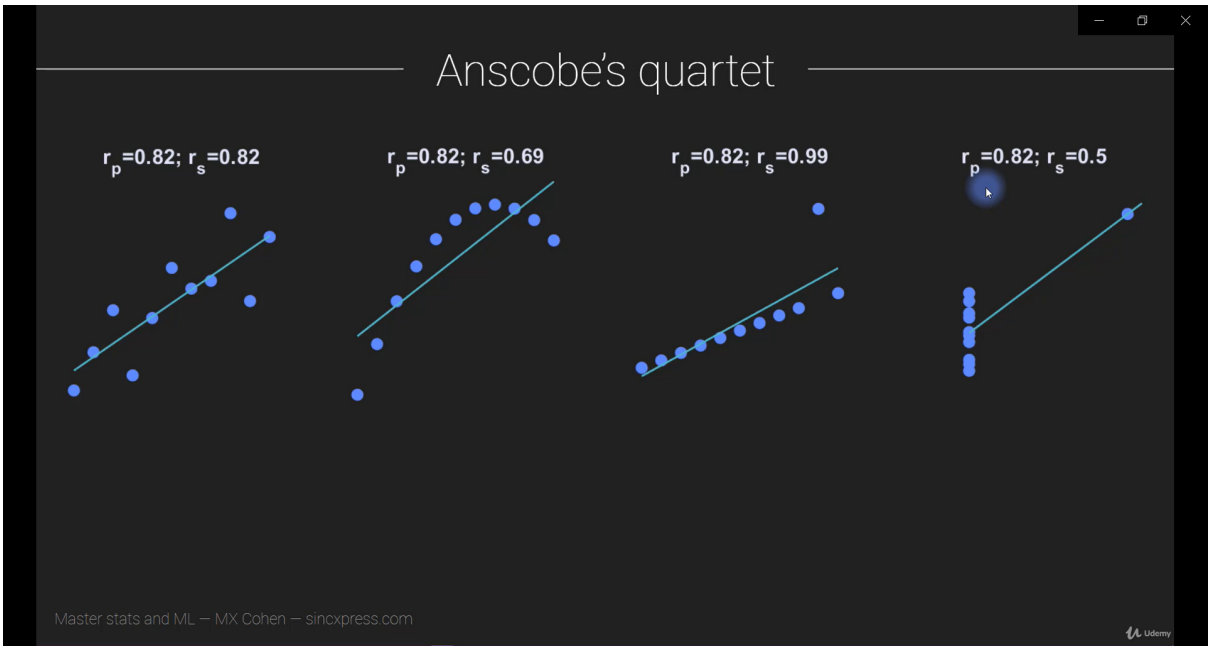
$$p(xy \mid z,w) = \frac{r_{xy} - r_{xz}.r_{yz} - r_{xw}.r_{yw}}{\sqrt{1 - r_{xz}^2}.\sqrt{1 - r_{yz}^2}.\sqrt{1 - r_{xw}^2}.\sqrt{1 - r_{yw}^2}}$$

and you can extend this to many variables.

Code

### Limitations of Pearsosn's Correlation Coefficient

rp = pearson correlation, rs = spearman correlation
**Notice**: The pearson correlation is same in all the four diagrams but the nature of data is different in all 4 cases.




- In the first diagram, the data seems to have a linear relationship.
- In the second diagram, the data has non-linear relation, but Pearson correlation is trying to capture its linear component.
- In the third diagram, the data seems to have a perfect linear relationship but with one outlier, hence pearson correlation gets strongly affected by it.
- In the fourth diagram, the data is almost vertical with one outlier due to which the rp = 0.82. Also spearman correlation also gets affected by that outlier but has much lower value than rp.

**Conclusion**:

- The Pearson correlation can over-represent or under-represent relationships if the data is non-linear or contains outliers.
- Therefore, the Pearson coefficient is appropriate if x and y are normally distributed and their relationship is linear without outliers.

| Note:<br>The other correlation measures which we will talk about later, also finds linear relationships in the data just as they are differently sensitive to outliers. |
|---|

| Way to generate data with the correlation coefficient we want: Code |
|---|

### Non-Parametric Correlation: Spearman Rank

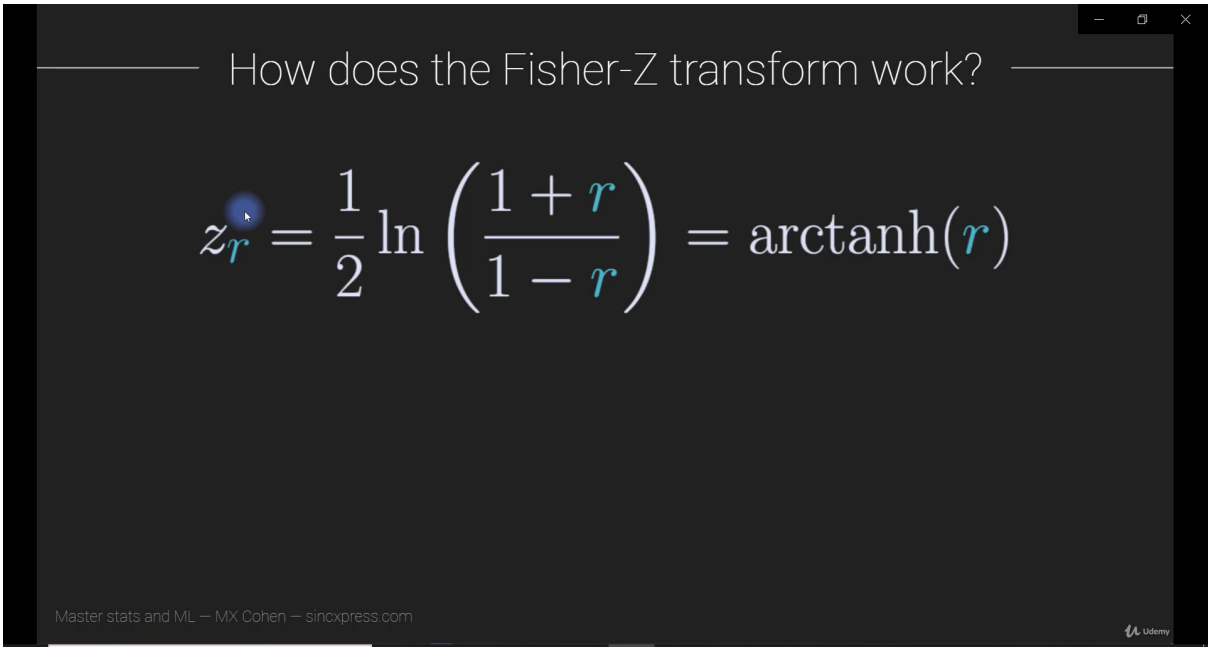
- Pearson and spearman correlation converge when the x and y are normally distributed.
- The main benefit of spearman correlation is that they are robust to outliers.
- The spearman correlation tests for the MONOTONIC RELATIONSHIP, regardless of whether the relationship is linear or nonlinear.
- **Algorithm**:
  1) Transform both x and y into ranks.
     e.g reminder of rank: (874387438, -40, 1, 0) --> (4,1,3,2)
  2) Compute the Pearson correlation on ranks.
  3) p-value calculated the same as in Pearson correlation.

- See how spearman correlation deals with the outliers, instead of considering the distance of data points with mean, it simply transforms the variable into ranks which totally reduces the effect of outliers.
- see the above example, how the outlier 874387438 is converted into rank and this will not affect the correlation process.

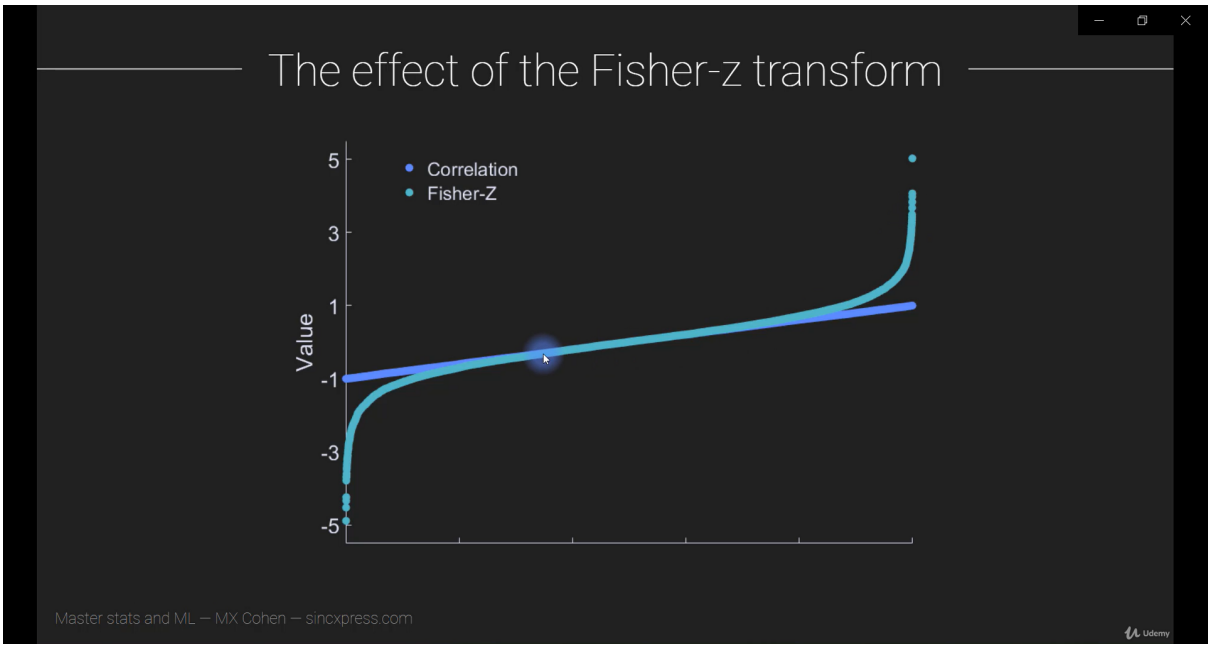
Code

### Fisher Z-Transformation for Correlation:

- Let's say you have two populations x and y. You sample from x (say sx) and sample from y (say sy), you compute the correlation and store it.
- You repeat the above step N(large) times, you will get a correlation list and when you plot it, say it will be a uniform distribution bounded between -1 and 1.
- Now, many analysis methods use normality as the assumption and say you want to do some analysis on that correlation list and that analysis requires that correlation should be normally distributed, so you need to do some transformation on it to make it normal.
- Fisher-z transformation is used to transform that correlation list into (roughly) normally distributed values.



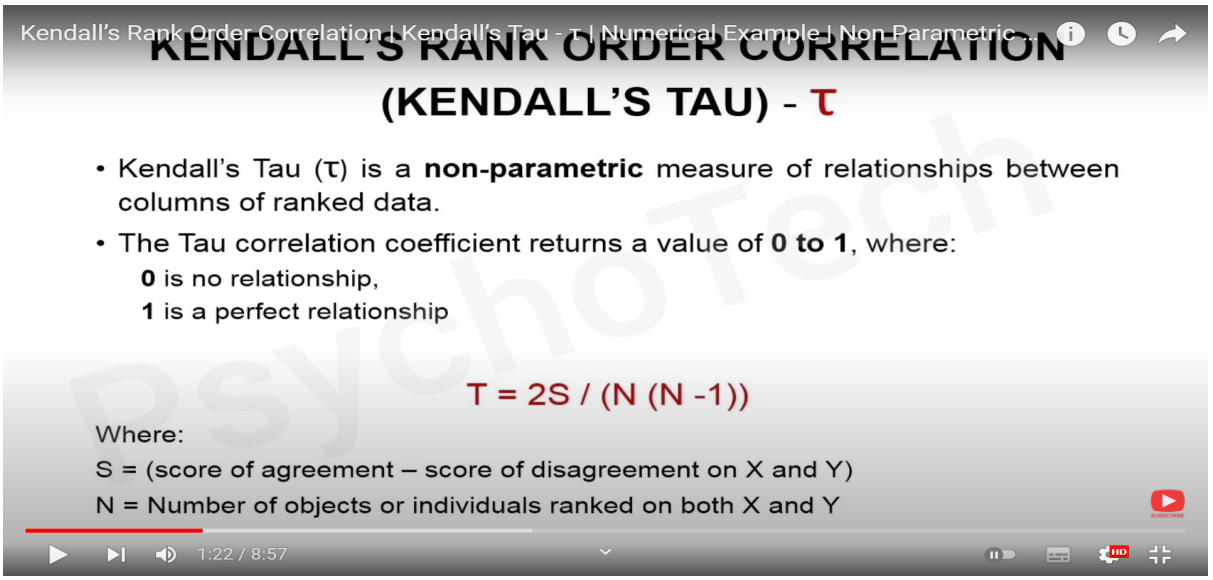

arctanh means inverse hyperbolic tangent of correlation coefficient.
See the correlations are bounded between -1 and 1.
But the transformed z-values differs, specially at the tails, but in the middle the transformed values are quite similar to actual correlation coefficient,

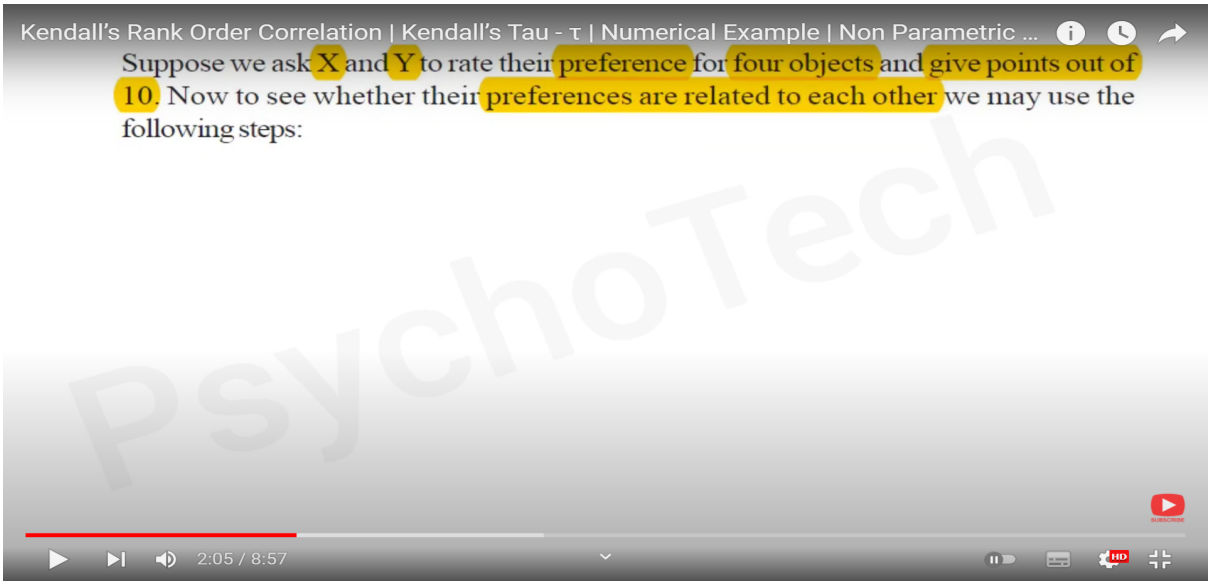


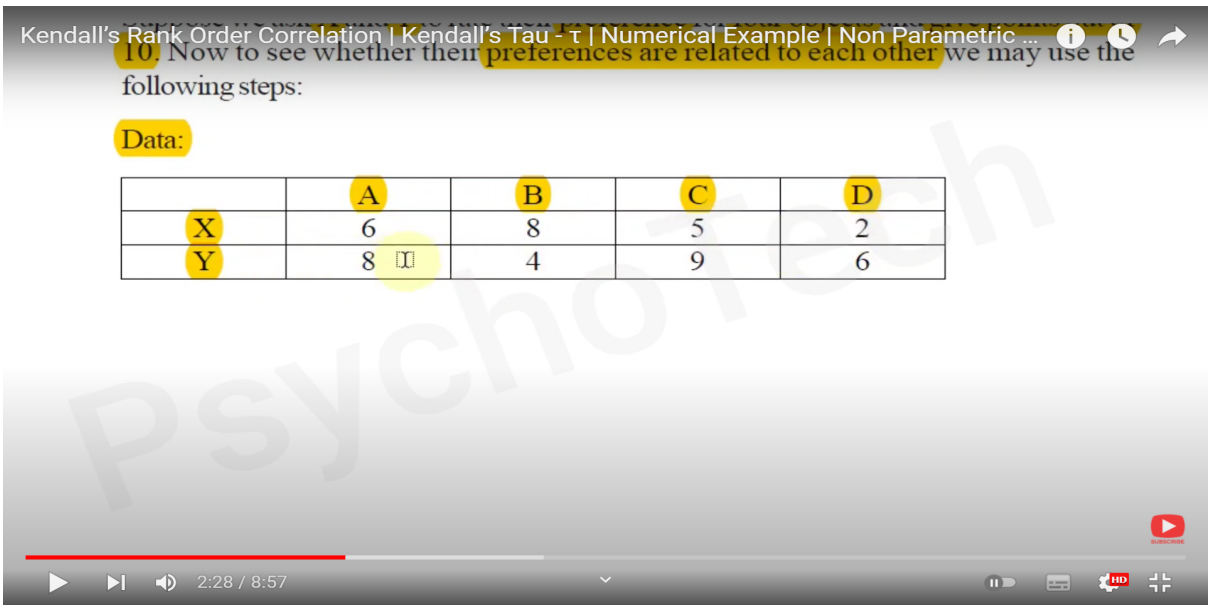

- If you want to know the relationship between two samples, then you don't need to worry about fisher-z transform ,but if you have many many correlation values and you want to do some analysis, then you might use this.

**Kendall's Correlation for Ordinal Data:**

- Used for the ordinal data (i.e the order in the data is important and some numeric values can be assigned)
  e.g size: small, medium, high -> 0,1,2 (1.5 doesn't make sense in this case).
- How it works:






Data:

| | A | B | C | D |
|---|---|---|---|---|
| X | 6 | 8 | 5 | 2 |
| Y | 8 | 4 | 9 | 6 |


| | A | B | C | D |
|---|---|---|---|---|
| X | 6 | 8 | 5 | 2 |
| Y | 8 | 4 | 9 | 6 |


**Step 1:** Ranking the data of X and Y

| | A | B | C | D |
|---|---|---|---|---|
| X | 3 | 4 | 2 | 1 |
| Y | 3 | 1 | 4 | 2 |

**Step 2:** Rearrange the data of X in order of 1 to N (4 in this case)

| | D | C | A | B |
|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 |
| | | | | |

**Step 2:** Rearrange the data of X in order of 1 to N (4 in this case)

| | D | C | A | B |
|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 |
| | | | | |

**Step 3:** Put the corresponding score of Y in order of X and Determine number of agreements and disagreements

| | D | C | A | B |
|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 |
| Y | 2 | 4 | 3 | 1 |

To calculate S we need number of agreements and disagreements. This can be calculated by

Using the Y scores, starting from left and counting the number of ranks to its right that are larger, these are agreements in order. We subtract from this the number of ranks to its right that are smaller- these are the disagreements in order. If we do this for all the ranks and then sum the results we obtain S:

Using the Y scores, starting from left and counting the number of ranks to its right that are larger, these are agreements in order. We subtract from this the number of ranks to its right that are smaller- these are the disagreements in order. If we do this for all the ranks and then sum the results we obtain S:

| Y | 2 | 4 | 3 | 1 | Total |
|---|---|---|---|---|---|
| | 2 | + | + | - | +1 |
| | | 4 | - | - | -2 |
| | | | 3 | - | -1 |
| | | | | 1 | 0 |
| | | | | Grand Total= S | - 2 |

**Step 4:** Calculate T

$T = 2S / (N (N -1))$

$T = 2 (-2) / (4 (4 - 1))$

$T = -4 / 12$

$T = -0.33$

Thus, $T = -0.33$ is a measure of the agreement between the preferences of X and Y.

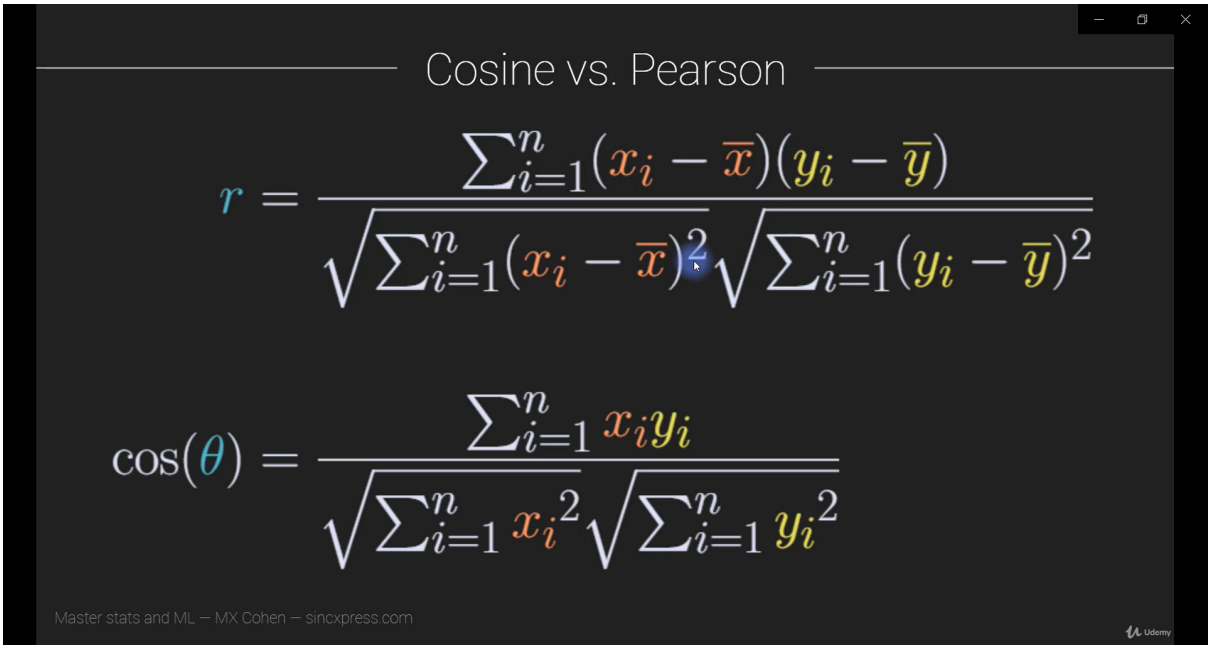
- Kendall tau-b has an adjustment for the ties and it is most often used for ordinal data.
- kendall-tau-a do not have adjustment for ties.
- Interpretation is identical to pearson and spearman correlation (i.e bounded by -1 and 1)
- **The assumptions for Kendall's Tau include:**
    1. Continuous or ordinal
    2. Monotonicity: Your two variables should have a monotonic relationship. This means that the direction of the relationship between the variables is consistent. For instance, when one variable goes up, the other goes up (in general). The relationship would also be monotonic if when one variable goes up, the other goes down (in general).

Code

**What is the difference between Spearman's Rho and Kendall's Tau?**
→ Spearman's Rho and Kendall's Tau are very similar tests and are used in similar scenarios.
We recommend using Kendall's Tau first and Spearman's Rho as a backup.

### Cosine Similarity:

1) Very very similar to Pearson correlation.




- where x is a point and y is point in n-dimensional space.(and there is line between origin and x, origin and y)
- The Pearson and cosine similarity are identical when in cosine similarity, the data is already mean-centred.
- The interpretation of cosine similarity and Pearson correlation is the same.

Code

### Point Biserial Correlation

- Point-biserial correlation is used to understand the strength of the relationship between two variables. Your variables of interest should include one continuous and one binary variable.
- Assumptions for Point-Biserial correlation
    1. Continuous and Binary
    2. Normally Distributed: Only use Point-Biserial Correlation on your data if the continuous variable is normally distributed.
    3. No Outliers: The continuous variables that must not contain outliers.
    4. Equal Variances (variance in continuous variable for category1 should be similar to variance in continuous variable for category 2)

# 11. ANOVA

**Table of Content**

**ANOVA**:

- It stands for ANalysis Of VAriance.
- The goal of an ANOVA is to determine the effects of discrete independent variables (categorical) on a continuous dependent variable.
- Setting up steps for ANOVA:
  Step 1) Review the experiment design and make sure ANOVA is an appropriate method.
  Step 2) Identify the independent and dependent variables.
  Step 3) Create a table of factors and levels (when possible) (more on this later)
  Step 4) Compute the model and interpret the results.

Let's start with each steps:

**Step 1)**
**Research Goal**: Test whether the new covid-19 medications are effective for different groups of people.
**Experiment**: Randomly assign patients to receive medication A, B and placebo. Measure the disease severity after 10 days. Separate older people(>50 years) from younger people(<= 50)

**Step 2:**
Dependent Variable: The variable for which you are trying to explain variance.
Independent variables: The variables which you HOPE will explain the Dependent variable.

Therefore, for the above the experiment:
IVs: Medication and age group ( Both are categorical )
DV: Severity after 10 days.

**Step 3:**
Create a table of factors and levels
**Factors**: The dimensions in the IVs
**Levels**: The specific groups or manipulations within each dimension ( basically possible values in each dimension )

In the above experiment,
**Factors**: medication, age
**Levels**: A, B, placebo (for medication)
older, younger (for age)

Now create table like this

| | Medication A | Medication B | Placebo |
|---|---|---|---|
| Younger people | | | |
| Older People | | | |

**We have written WHEN POSSIBLE why?**

- Say if we had a third categorical variable as well, then we had to create a 3d table and not 2d table like above, and if we have n variables, then we cannot visualise this table of factors and levels.
- We can compute ANOVA values with n variables, and it is okay if we cannot create tables.

**Step 4** ( Very Important )
We will get into the maths later, but for now there are 3 main interpretations for ANOVA:

- **Main effect:**
  How the one factor influences the DV even when ignoring all the other factors.
  e.g The younger people's symptoms improve faster as compared to the older people, regardless of medication type.
  (i.e main effect of age on severity independent of medication type)
- **Interactions**:
  The effect of one factor on DV depends on the levels of other factors.
  e.g The medication A works better for older people and medication B works better for younger people.
  (i.e effect of medication type depends on levels of age)
- **Intercept**:
  The average value of DV is different from zero. The intercept of ANOVA is usually ignored.
  e.g the symptoms improve for almost everyone after 10 days.

We are mainly concerned with main effects and interactions between the factors.

### Some Important Terminologies in ANOVA:

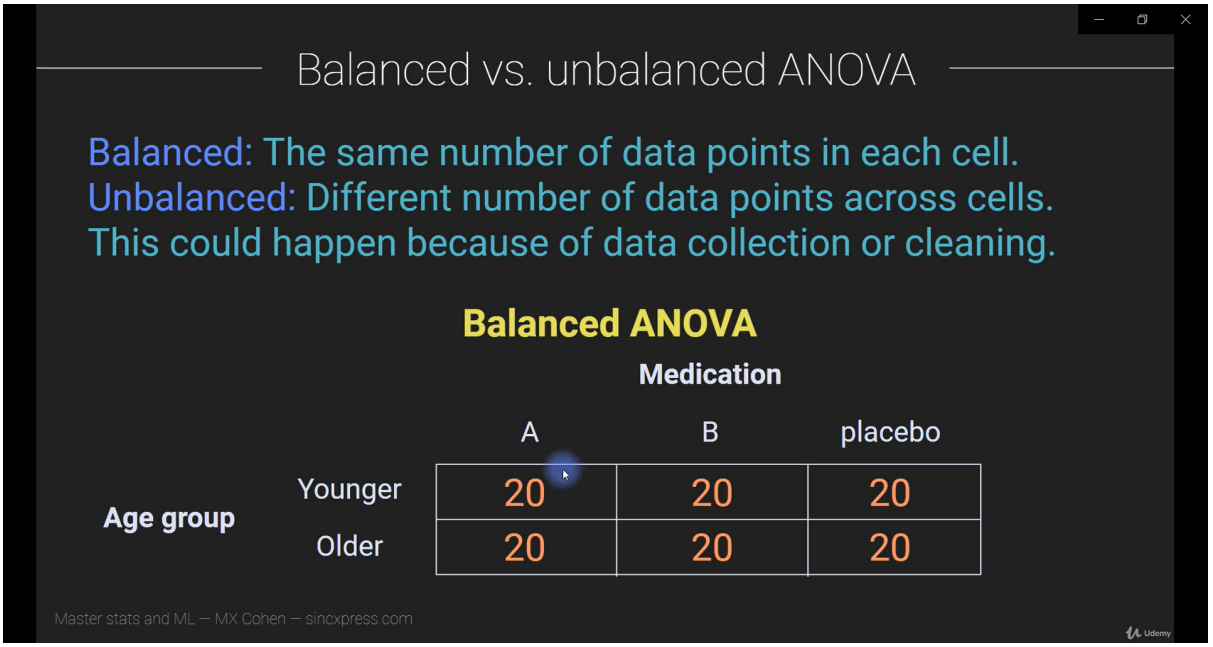
- <x>-way ANOVA, where x = number of factors
- e.g one-way ANOVA: i.e number of factors considered = 1
  (it can have any number of levels.)
  e.g Determine the influence of day-of-week on the purchase of iphones.
  There is only one factor = day-of-week and number of levels = 7
- Two-way ANOVA: number of factors considered = 2
  e.g Determine the influence of day-of-week and gender on iphone purchases.

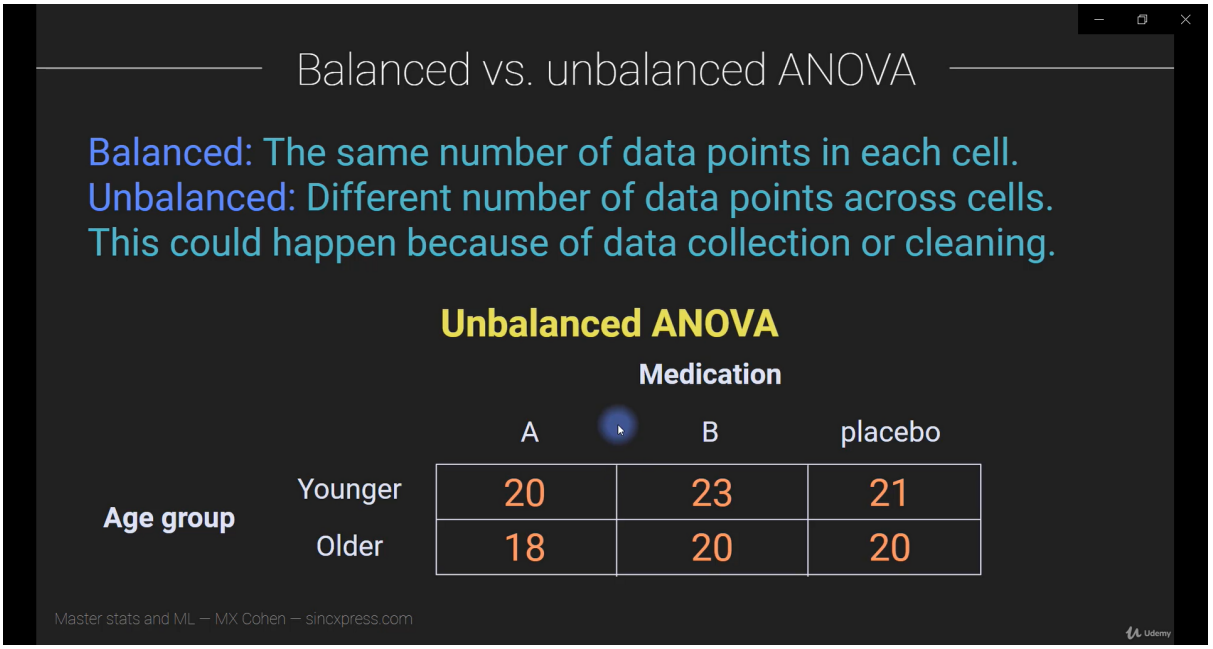
**Repeated-measures ANOVA**:

- If atleast one factor involves multiple measurements from the same individual.
- e.g
  **Research question**: Determine the effect of snack-type on mood.
  **Experiment:** Volunteers eat chocolate for 2 days, potato chips for 2 days and icecream for 2 days. (randomised order)

  There is only one factor called snack-type and dependent variable mood.
  This is repeated-measures ANOVA, because the SAME INDIVIDUAL will eat chocolate for 2 days, potato chips for 2 days and icecream for 2 days and their mood will be noted after every 2 days.

**Balanced vs unbalanced ANOVA:**




**ANOVA vs MANOVA:**

- anova: only one dependent variable (you can have many independent variables)
- manova: multivariate dependent variable (as many independent variables as appropriate)
  e.g effect of medication type and age on disease severity and medical expenses.

**Fixed and Random Effects ANOVA:**

- **Fixed effects**: The number of levels of a factor is always fixed. (e.g operating systems: They are always three: mac, windows, linux)
- **Random effects:** The number of levels of a factor are not fixed. (e.g say age is factor, which is not fixed, so we can discretize into specific number of
- levels.)
- **Mixed effects:** Some of the factors are fixed effects and some factors are random effects.

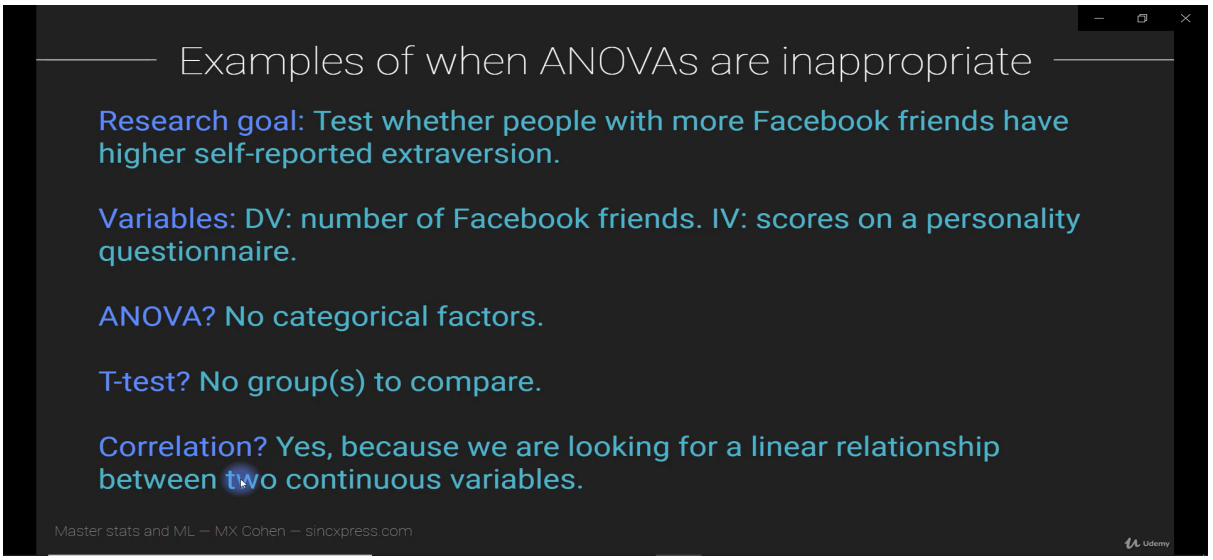
### Assumptions of ANOVA:

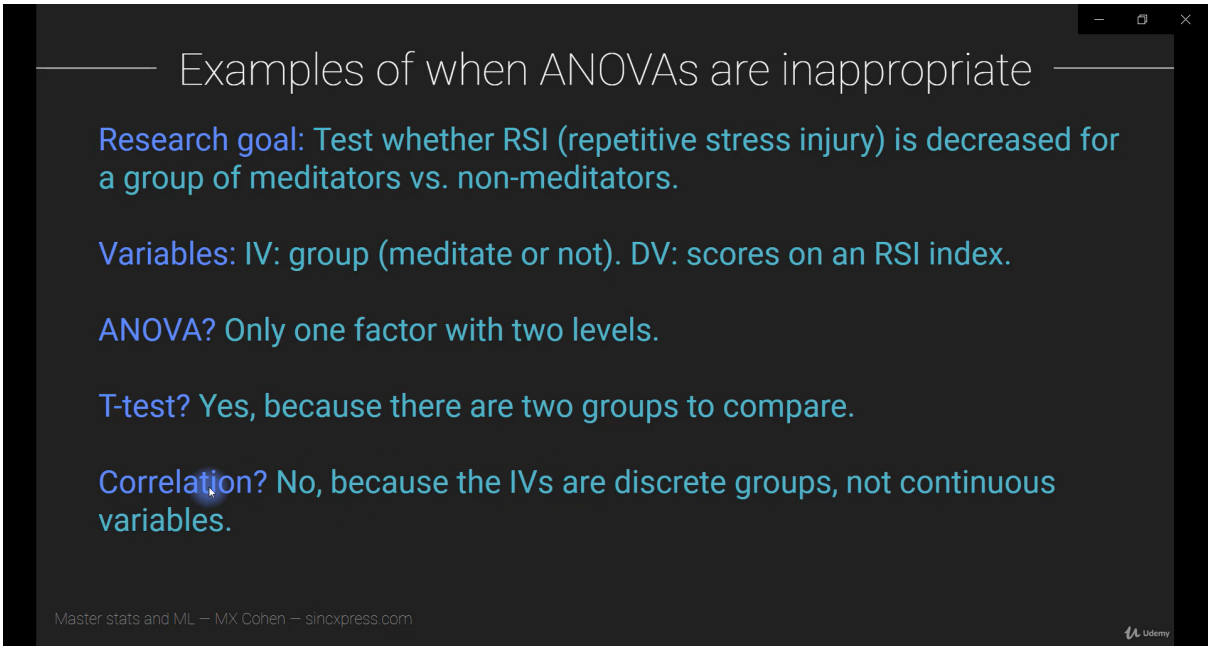
- Independence: The data points should be sampled independently from the population.
- Normality: The residuals (unexplained variance after fitting the model) are normally distributed.
- Homogeneity of variance: The variance in the factor-levels table should be roughly the same, otherwise there might be some bias.

Note:
If these assumptions are strongly NOT met, then you can use Non-parametric ANOVA alternatives:
Kruskal-Wallis test (KW-ANOVA): Alternative to the one-way ANOVA on the rank-transformed data.

Note:
The anova are generally robust to violations of the assumptions. non-parametric anovas are rarely used and may cause more of a headache than they are worth.

**When ANOVA is not appropriate:**



**Now mathematics**

( Just building some blocks )

- Remember we had a table of factors vs levels, let ni be the number of data points in each cell and $mu_i$ be the mean of each cell.
- Null hypothesis: all the cells have the same mean i.e mu1 = mu2 = ... = muk i.e means all the groups are the same.
- Alternative hypothesis: the mean of at least two cells differ from each other. This is a very generic alternative hypothesis. So to get exactly which two cells have different means, we have to do some more investigation. i.e mean of at least one group differs from at least one other group.

- Anova is based on the sum of squares i.e
  sum( $(x_i$ - x_bar$)^2$ )
  This formula is very similar to variance.
- But in anova, why don't we have a denominator of n-1 ?
  → Because in anova we calculate multiple sum of squares, we are not actually interested in sum of squares values, we calculate their ratios, therefore when we do ratios, the denominators cancel out.
- The total variation in the dataset is expressed as the sum of variations across different groups and the variation within each group.

We are basically going to calculate the f-statistic and it is ratio of two things
f = "Explained" variance (due to factors or levels in each factor) (sum of squares between groups)
—----------

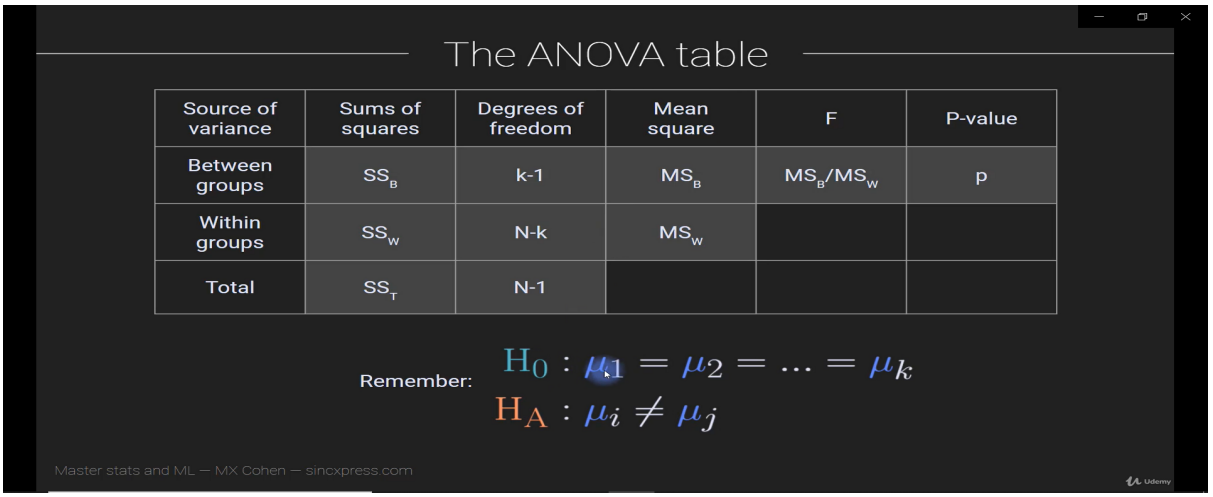
"Unexplained" variance" (due to natural variation) (sum of squares within groups)

***Omnibus f-test*** *means: your p-value associated with the f-statistic is significant but you don't know exactly the means of which two groups differ from each other, so that f-value is called omnibus. So you have to do post-hoc analysis i.e comparison of multiple groups.*

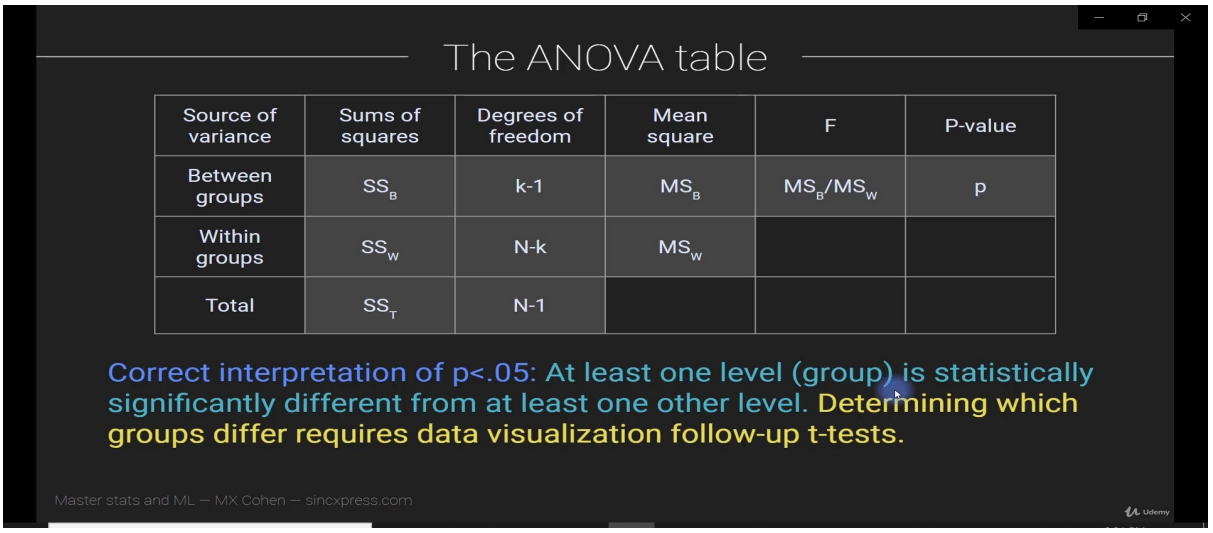
Now check this article

https://medium.com/data-science-in-your-pocket/anova-mathematics-explained-c85a36172ac1

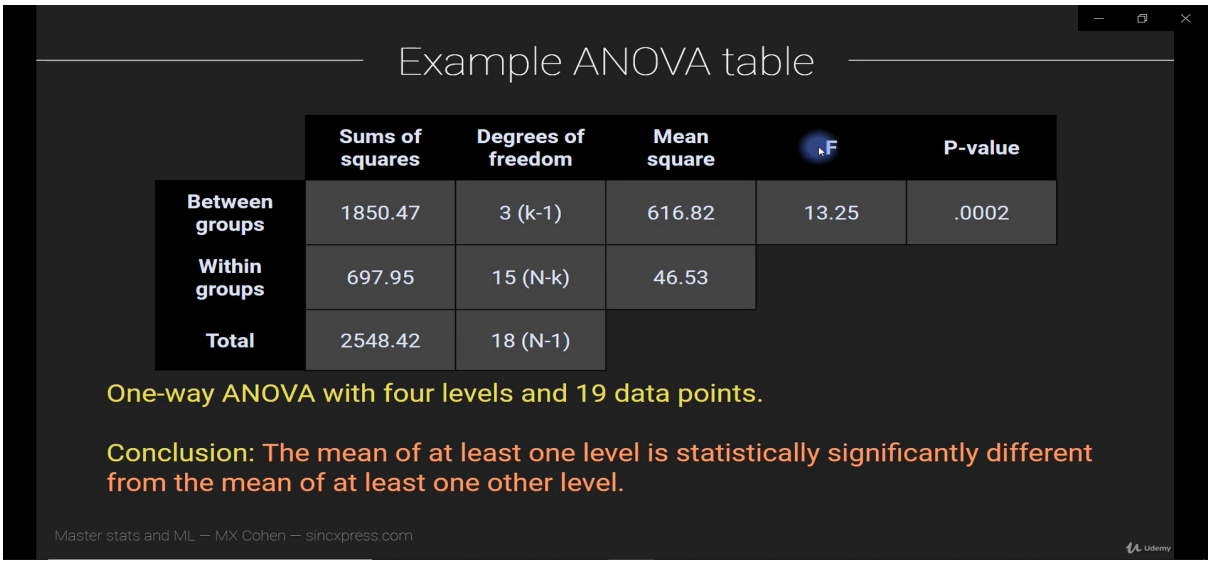
This is called the ANOVA table ( k = number of levels in a factor )

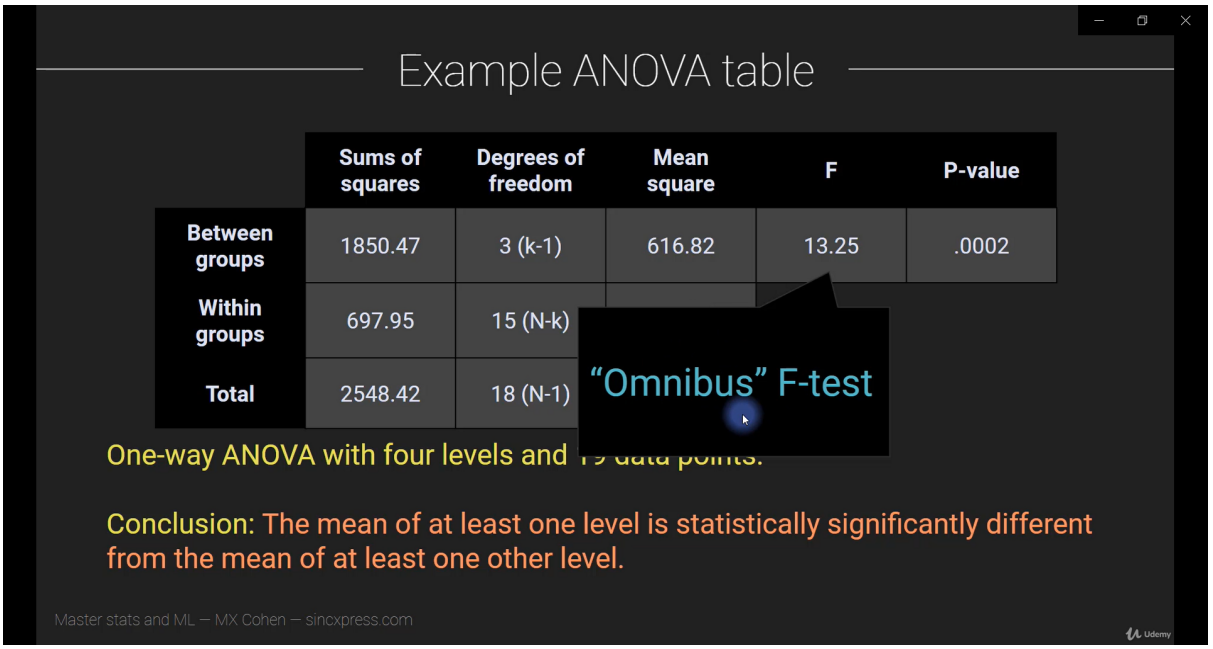



**Important: Correct interpretation of p-value:**



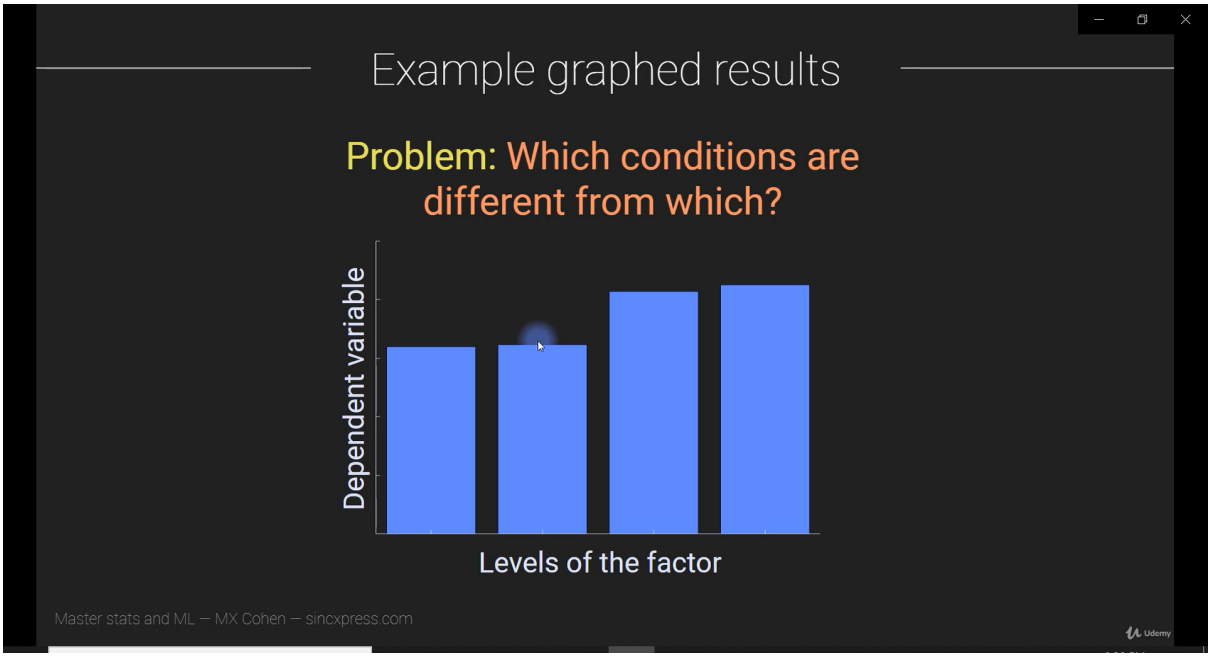

Example for correct interpretation of p-value:



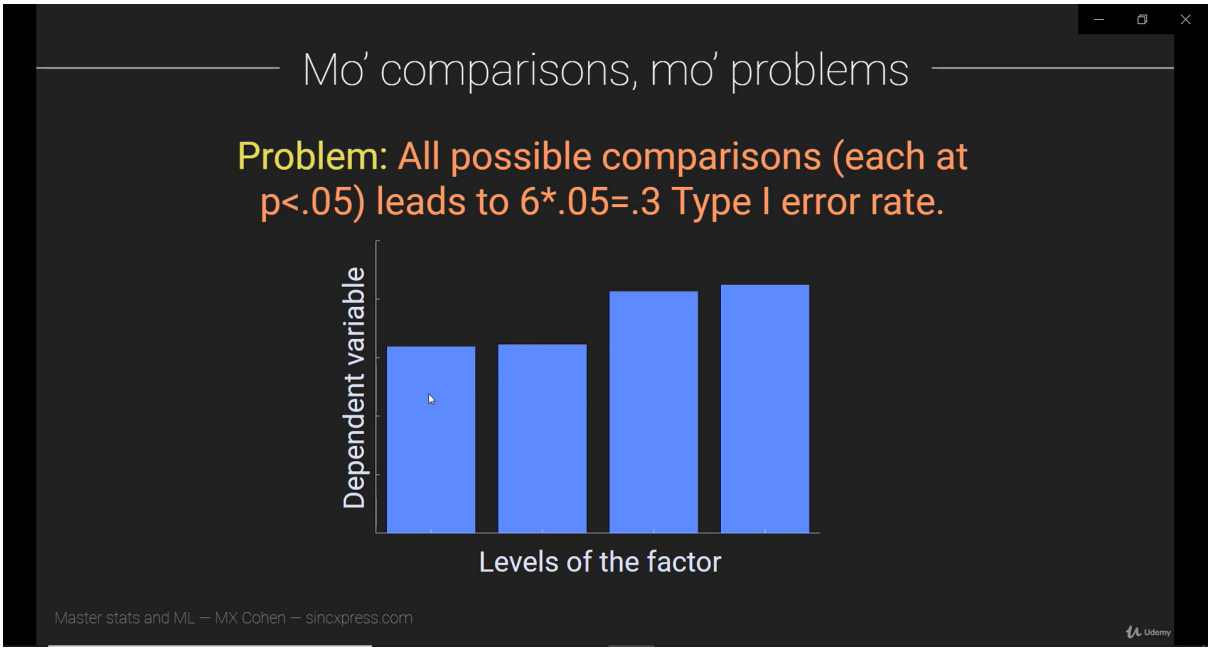


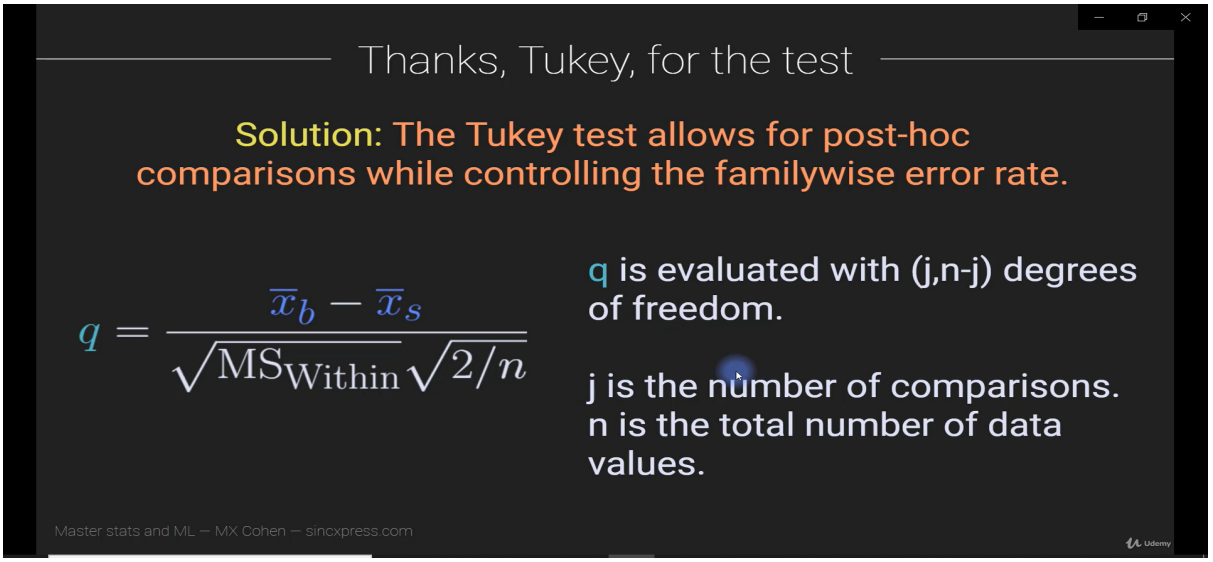

**omnibus f-test** means: your p-value associated with the f-statistic is significant but you don't know exactly the means of which two groups differ from each other, so that f-value is called omnibus. So you have to do post-hoc analysis i.e comparison of multiple groups

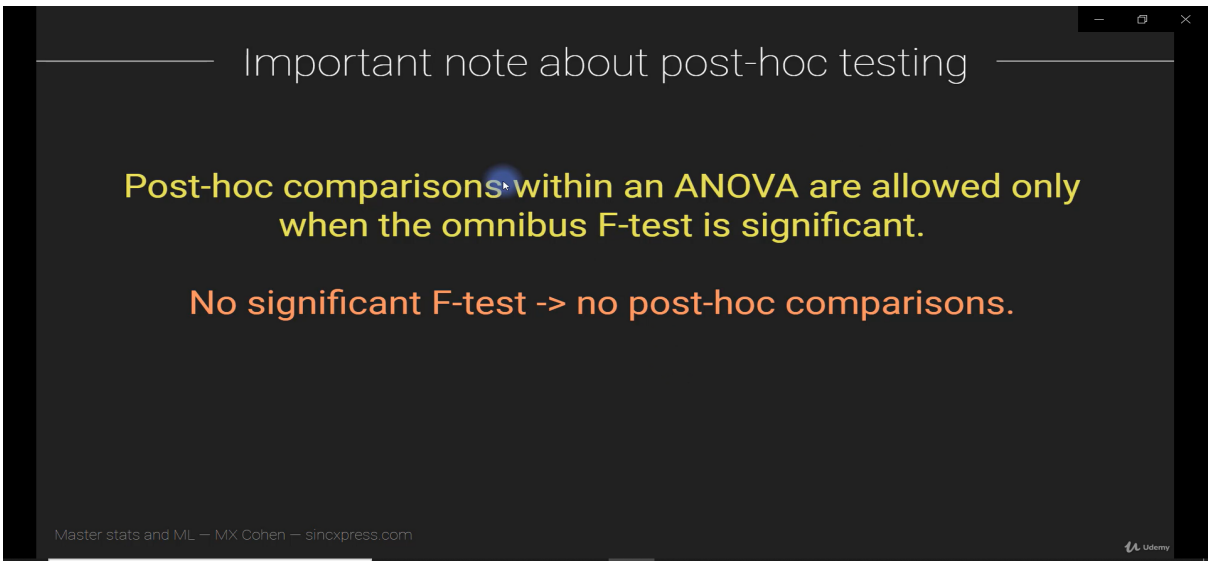
Say for the independent category there are 4 levels, then we can have 6 possible comparisons with 4 levels.



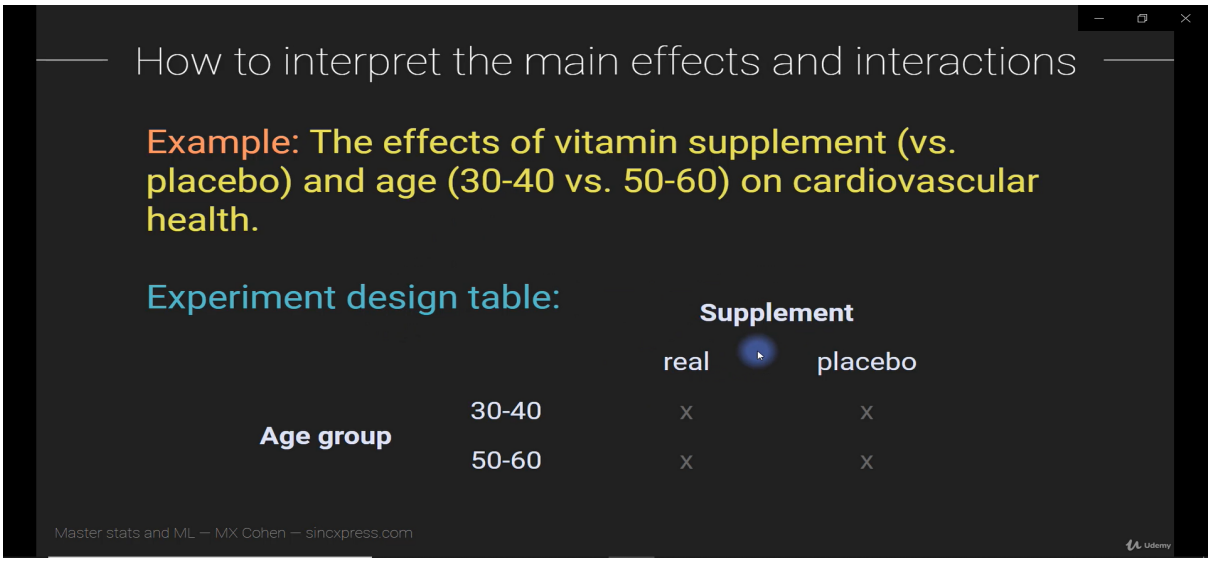
All 6 possible comparisons with p <= 0.05 will lead to such high combined type-1 error rate.



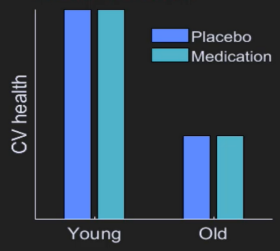



**Interpretation of Main Effects and Interactions:**

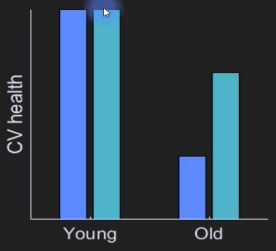


# How to interpret the main effects and interactions

Main effect of age
No effect of medication
No interaction



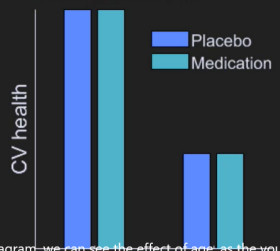

Main effect of age
Effect of medication
Interaction



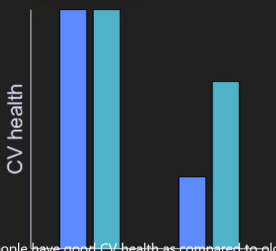



# How to interpret the main effects and interactions

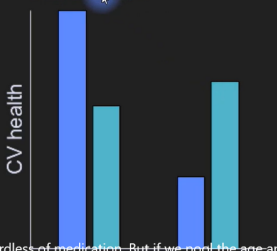
Main effect of age
No effect of medication
No interaction




Main effect of age
Effect of medication
Interaction




Main effect of age
No effect of medication
Interaction

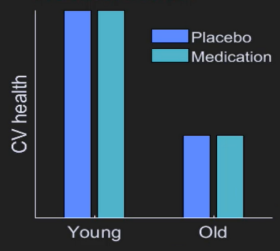



In first diagram, we can see the effect of age, as the young people have good CV health as compared to old regardless of medication. But if we pool the age and average the effect of medication and placebo they are same, hence there is no effect of medication on CV health regardless of age.
In the second diagram, again we can see the CV health of younger people is good as compared to older people (that is effect of age) regardless of medication. Now if we pool the effects of medication and placebo, the average of medication is greater than average of placebo. Therefore there is effect of medication regardless of age. Now in case of older people, the medication improves CV health (therefore there is interaction term).
In the third diagram, again the average CV health of age is better as compared to older people (effect of age) irrespective of medication. Now if we pool the effect of medication on CV health, the average of medication and placebo remains the same (therefore there is no effect of medication irrespective of age). Now in case of older people, the CV health increases after medication (interaction term), also for younger people the CV health decreasea after medication.

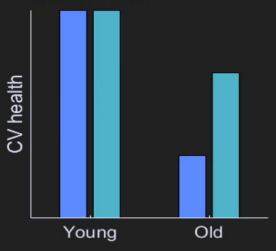


# How to interpret the main effects and interactions

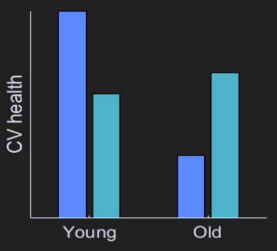
Main effect of age
No effect of medication
No interaction



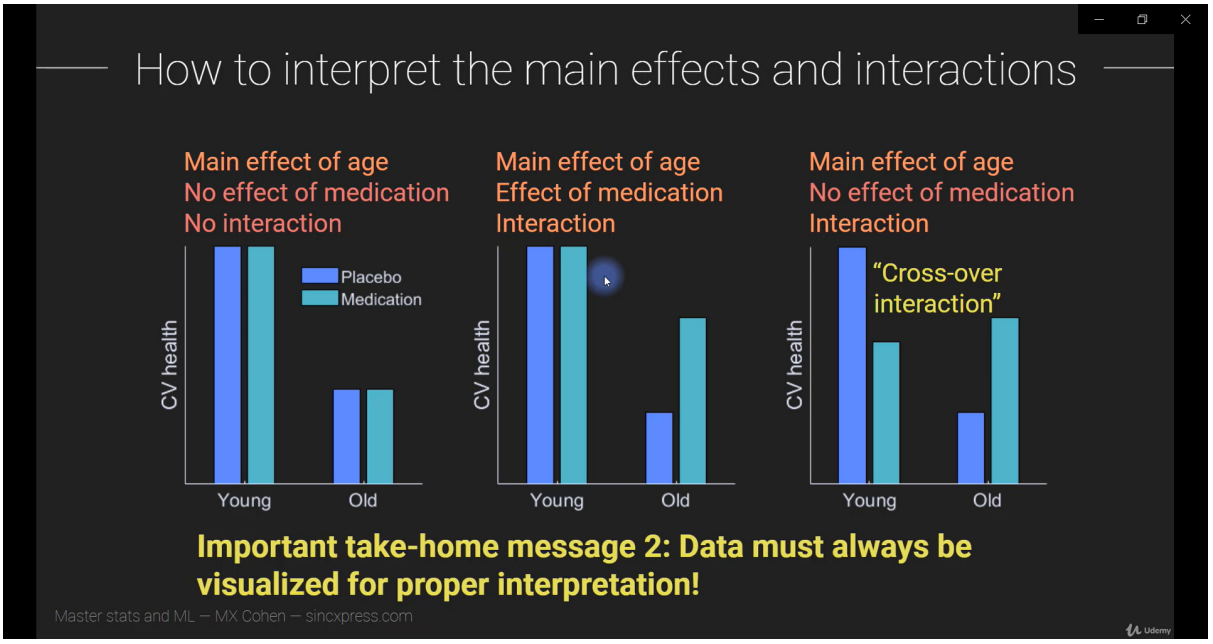

Main effect of age
Effect of medication
Interaction




Main effect of age
No effect of medication
Interaction




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




**One-Way ANOVA Example:**

( This is unbalanced-ANOVA )

## The data

| | 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 |

* Note: Made-up data!



After running the ANOVA on the above data, we get the following ANOVA table

## The ANOVA table

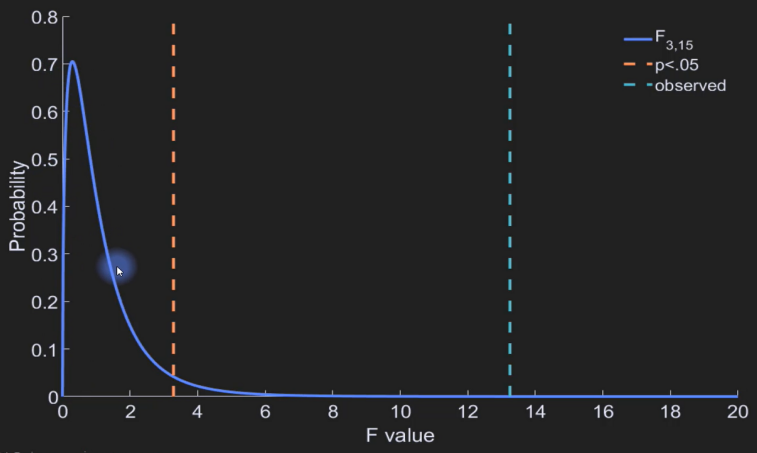
| | Sums of squares | Degrees of freedom | Mean square | F | P-value |
|---|---|---|---|---|---|
| Between groups | 1850.47 | 3 (k-1) | 616.82 | 13.25 | .0002 |
| Within groups | 697.95 | 15 (N-k) | 46.53 | | |
| Total | 2548.42 | 18 (N-1) | | | |

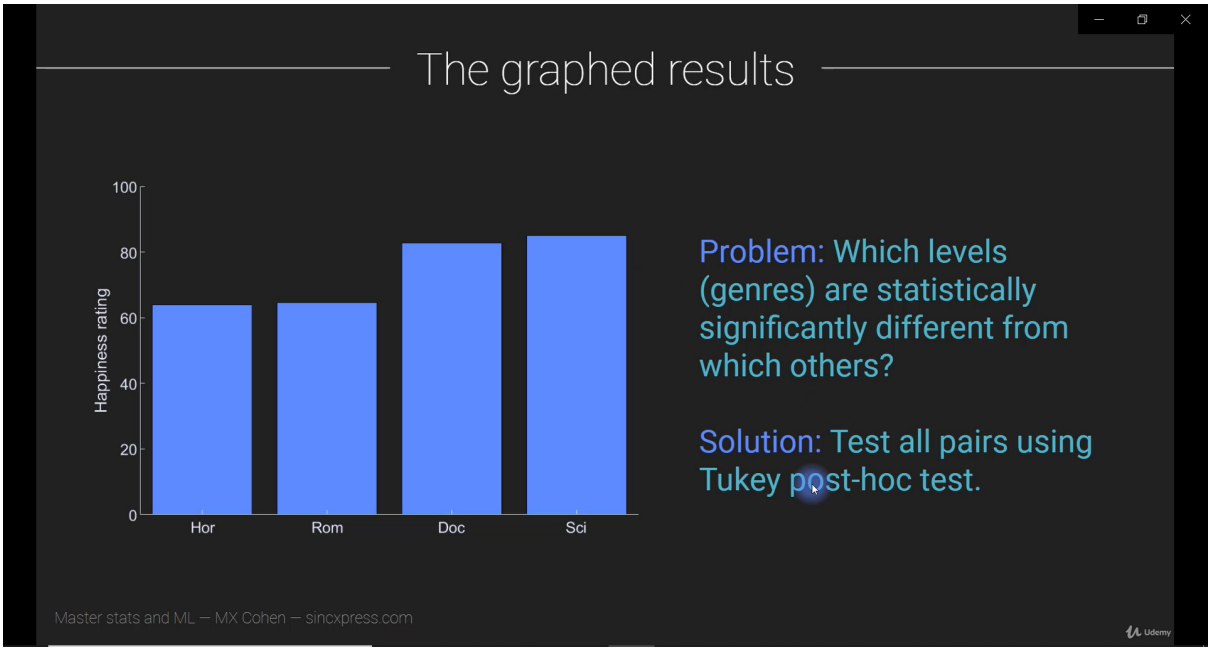
Conclusion: The mean of at least one level is statistically significantly different from the mean of at least one other level.

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

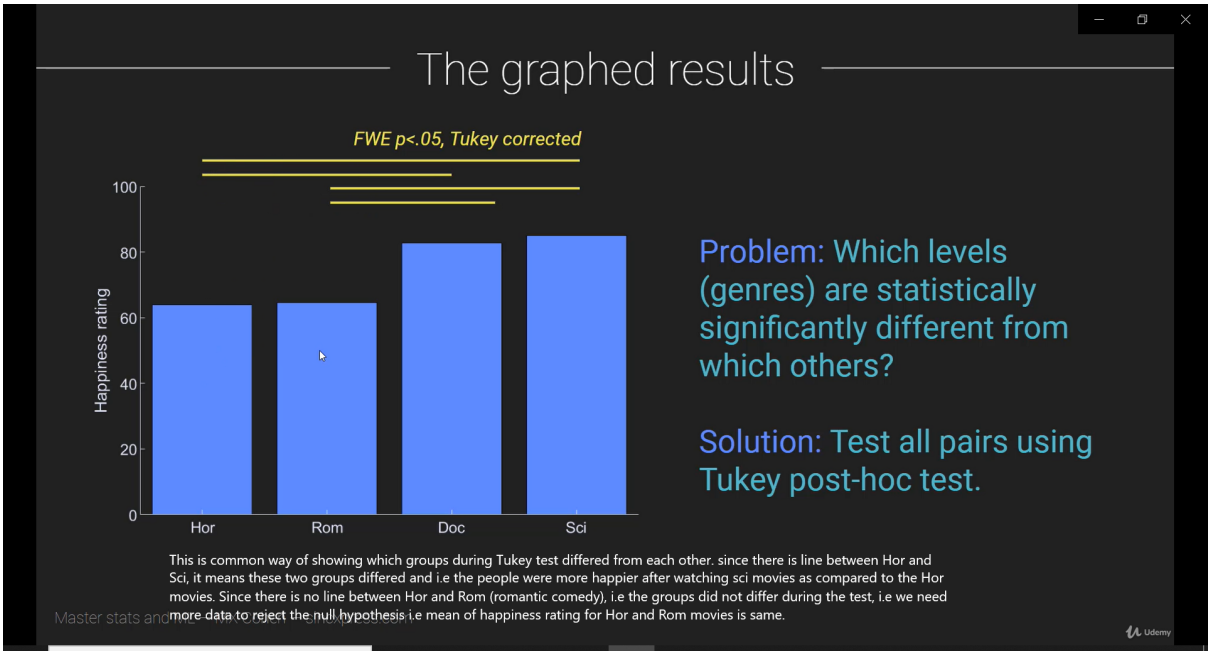


## The $H_0$ and observed F values





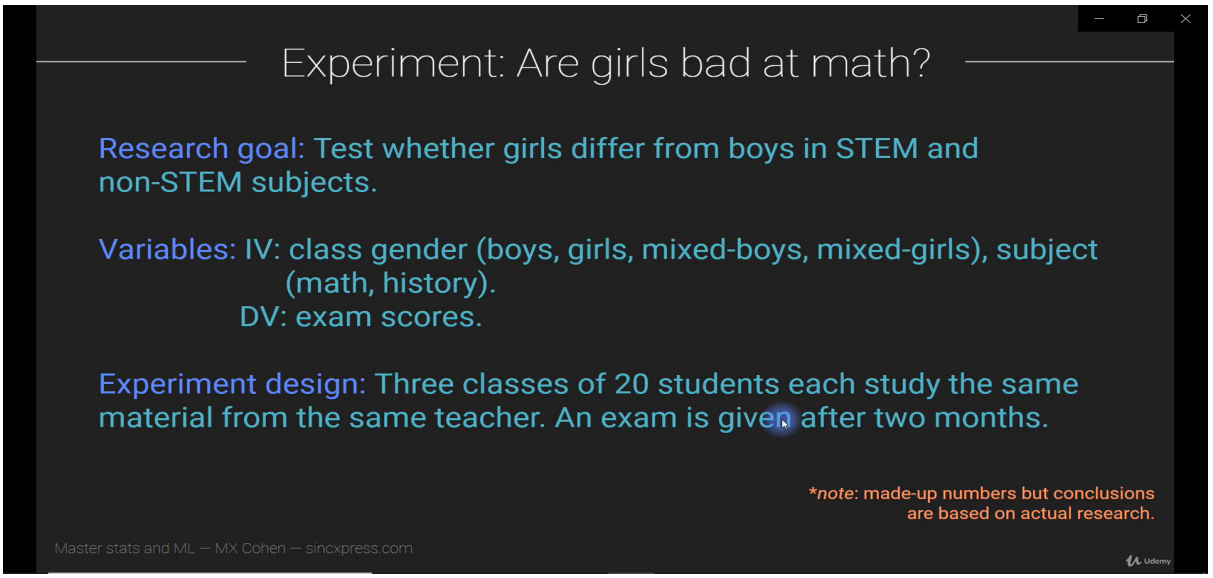

( FWE: Family-Wise Error rate )

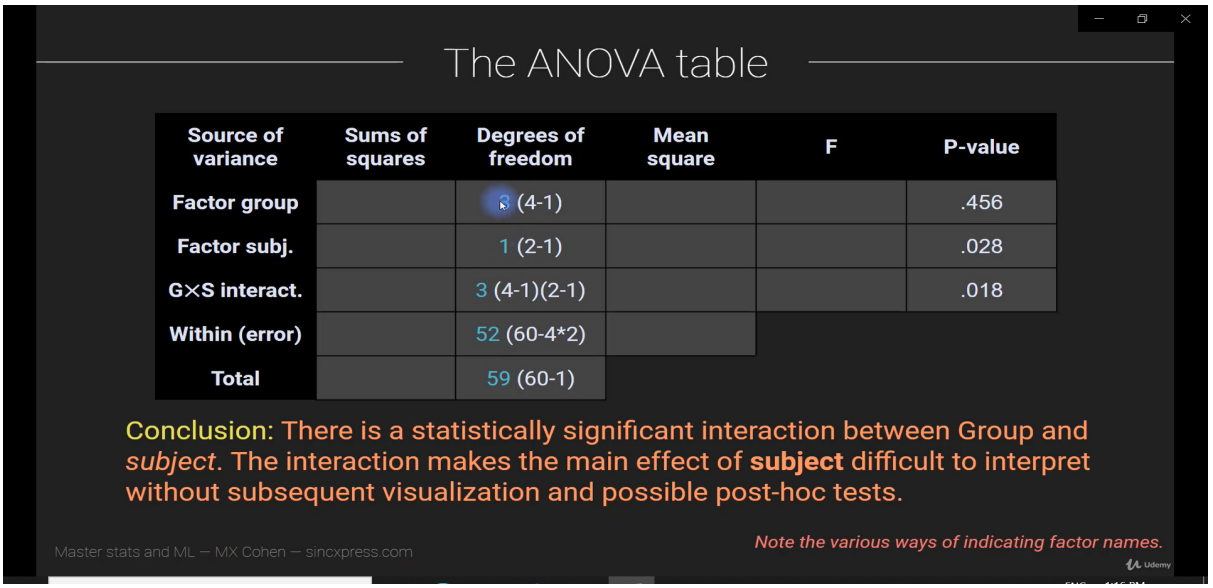


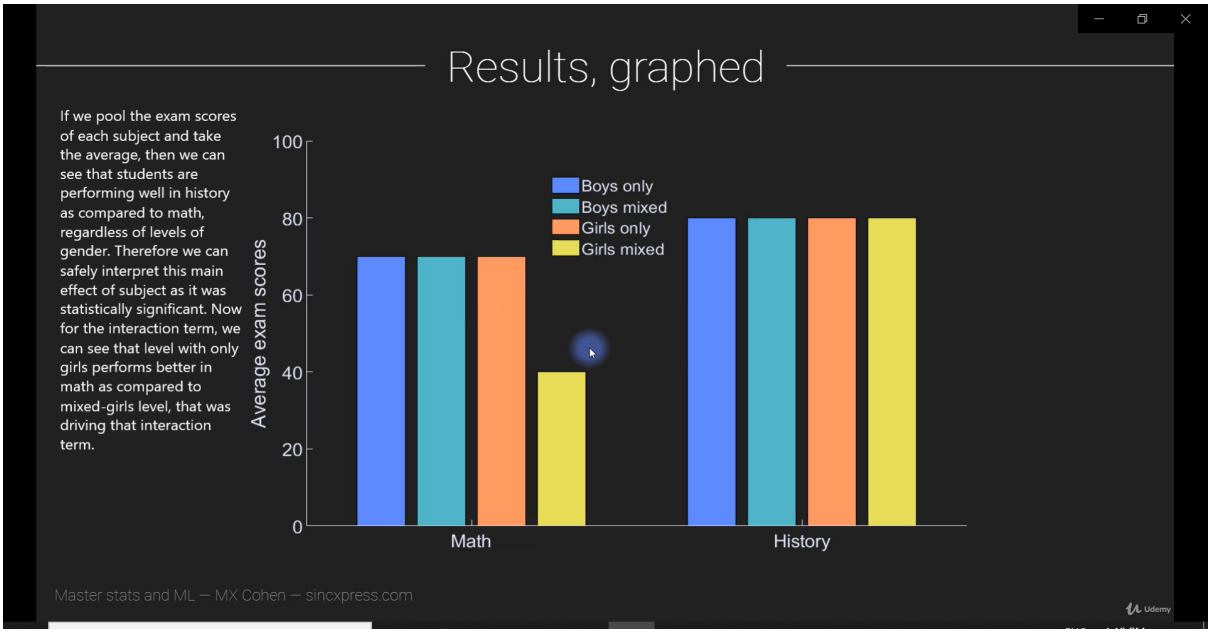

**Code**

**One-way ANOVA with repeated measures: Code**

**Two-Way ANOVA Example:**






Two-way mixed effects ANOVA: Code

# 12. Regression

**Table of Content**

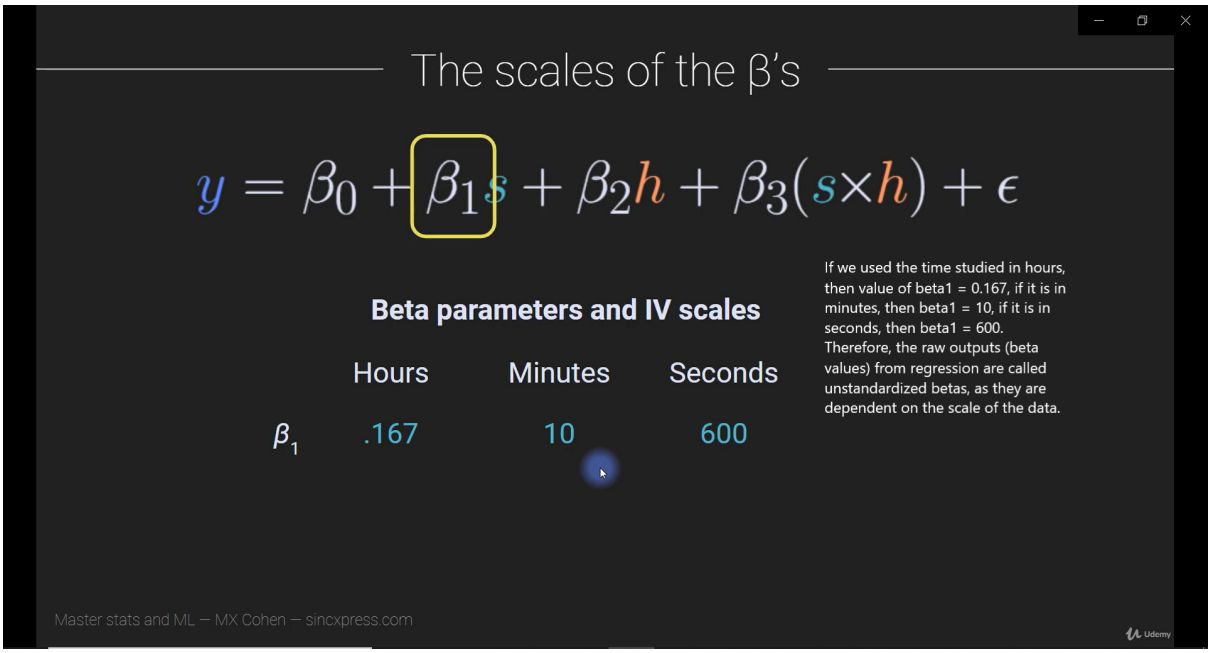
**For detailed lectures on Regression**

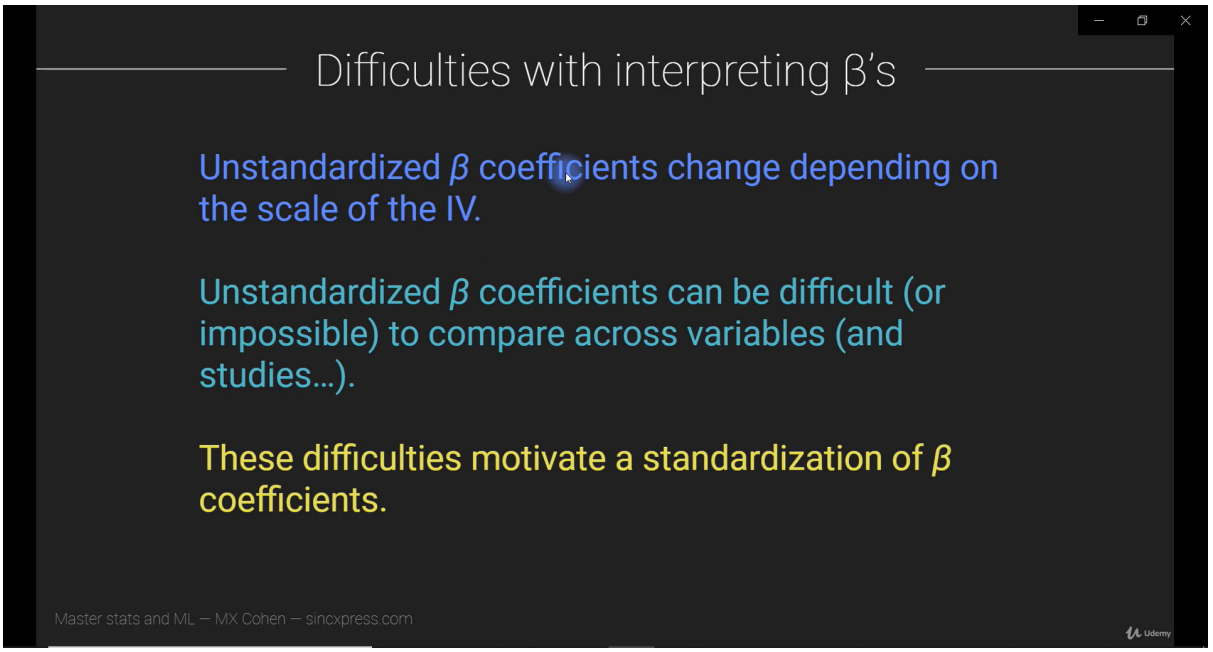
check out this playlist
https://www.youtube.com/playlist?list=PLTNMv857s9WUI1Nz4SssXDKAELESXz-bi

**Evaluating Regression Model using f-test:**

F-Test in regression analysis (Hypothesis test using F-Statistic)

**Standardising Regression Coefficients**



## The scales of the β's

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

**Standardized beta parameters**

| | Hours | Minutes | Seconds | Calories |
|---|---|---|---|---|
| $\beta_1$ | .6 | .6 | .6 | |
| $\beta_2$ | say s=time you sleep and h=calories you intake before exam. Now after standardization of betas, see the value of beta1 is same irrespective of what the unit of s is. Also, from this we can interpret that the effect of calories you intake has more effect on exam score as compared to amount of time you sleep before exam | | | .8 |



## The scales of the β's

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

Standardized $\beta$ 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!



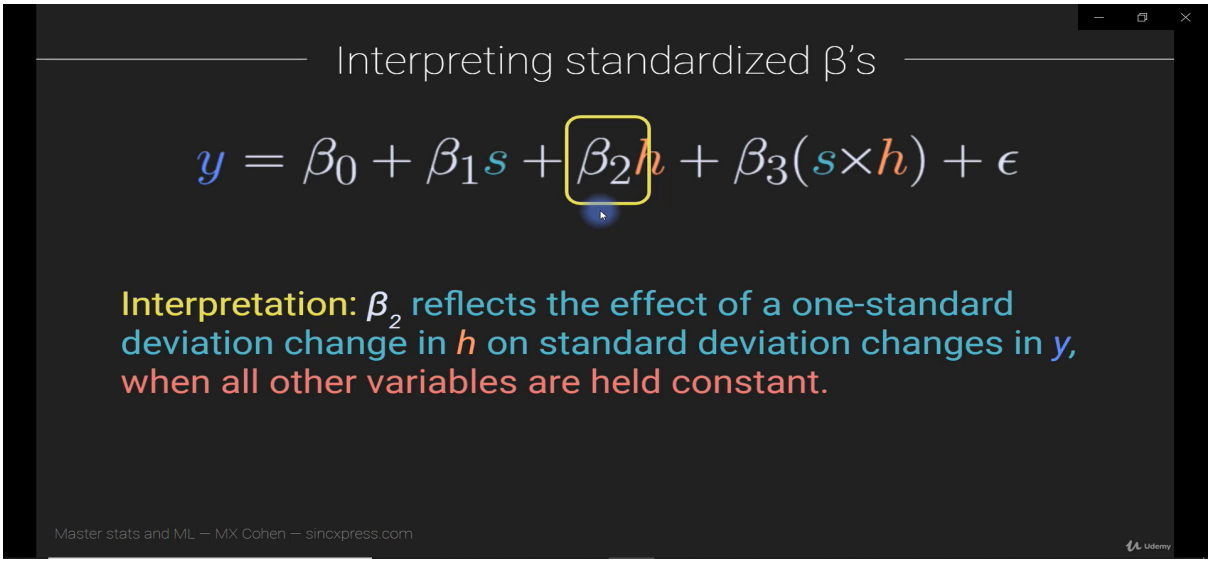
## How to standardize regressors

Basic idea: normalize $\beta$ so that its variance is 1.

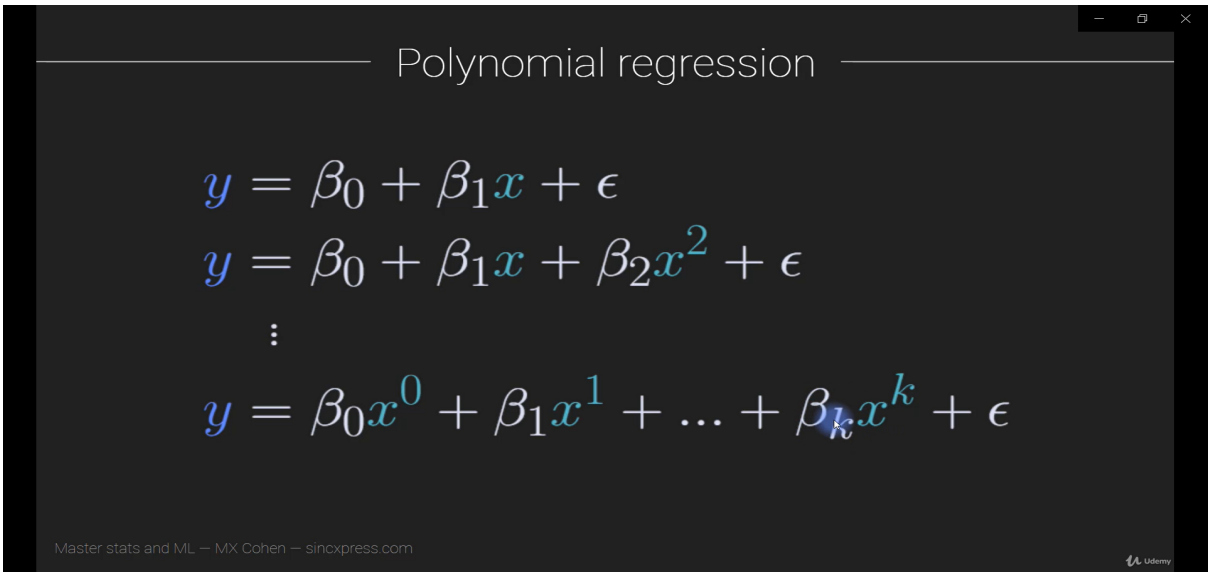
Method 1: z-normalize DV and IVs before the regression. All $\beta$'s will be in the units of the data, which are already standard deviation units.

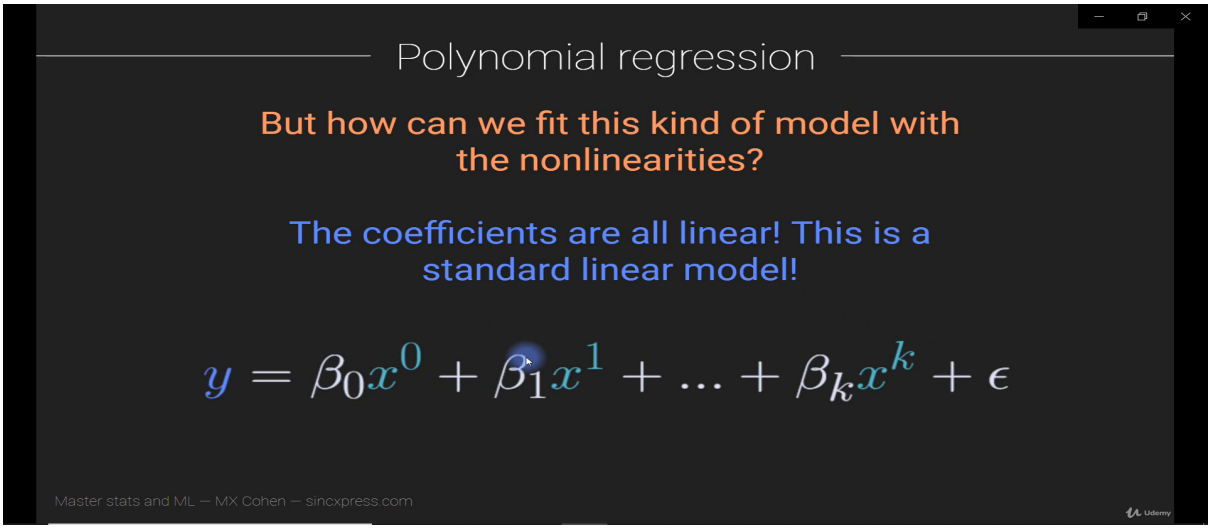
Method 2: Scale the unstandardized $\beta$ by the standard deviations of the IV and corresponding DV.

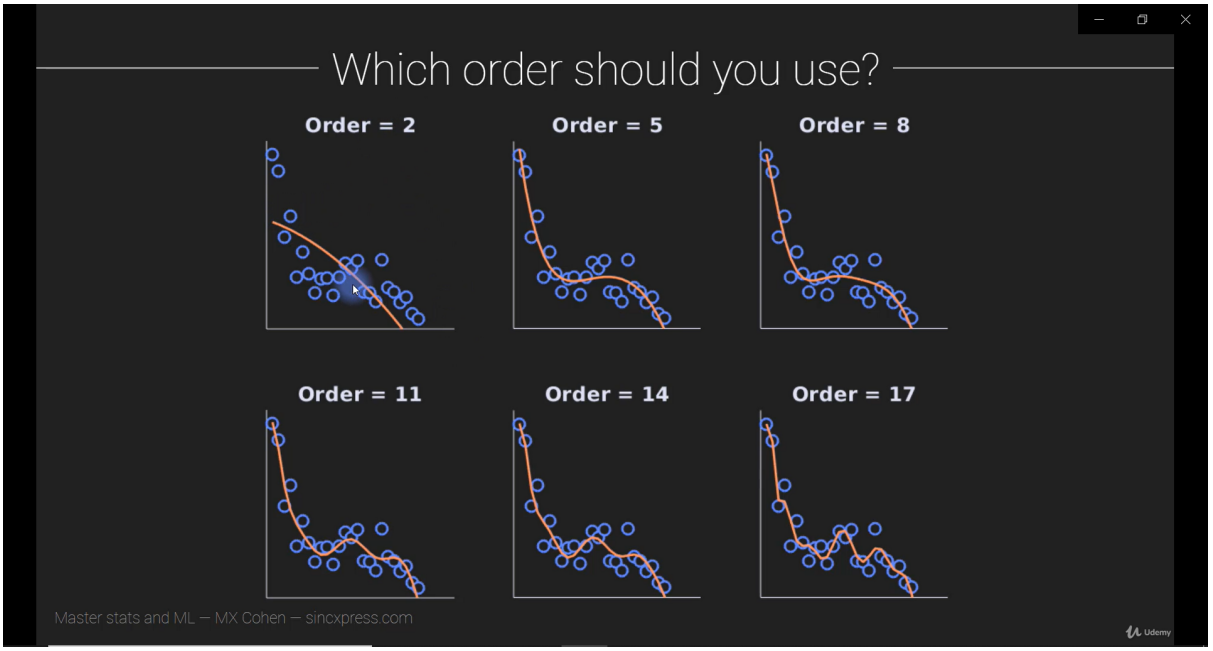
$$b_k = \beta_k \frac{s_{x_k}}{s_y}$$



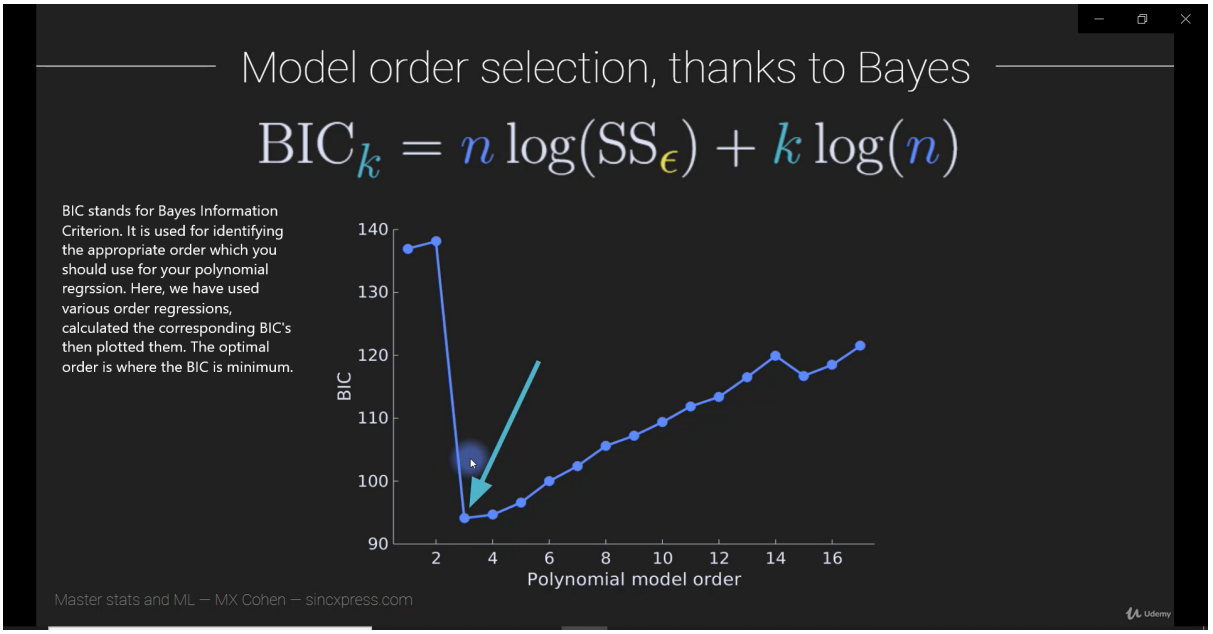

**Polynomial Regression**





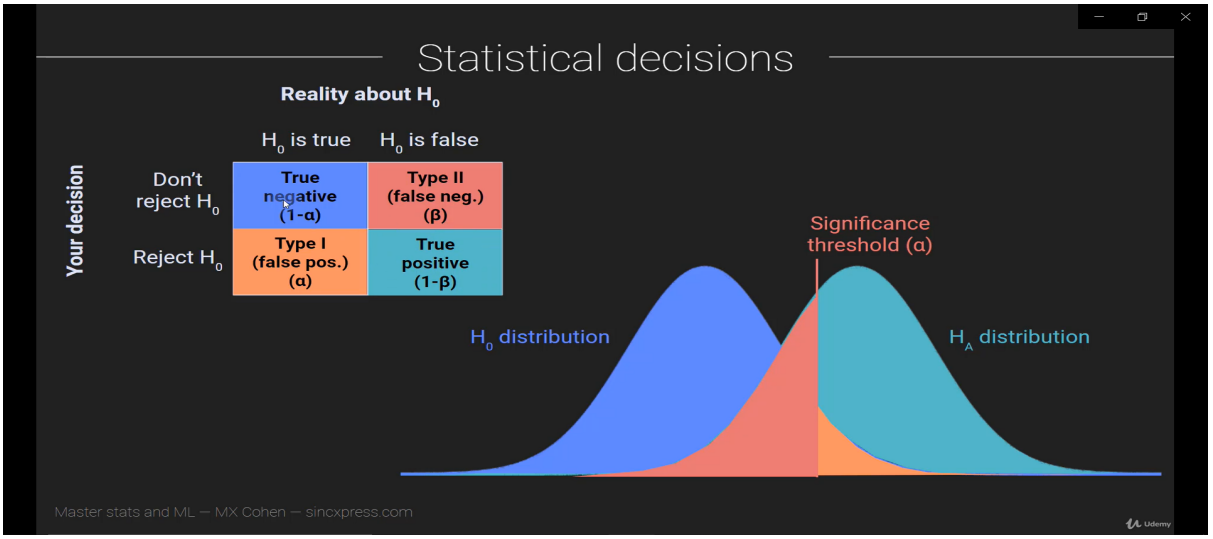

## Bayes Information Criterion

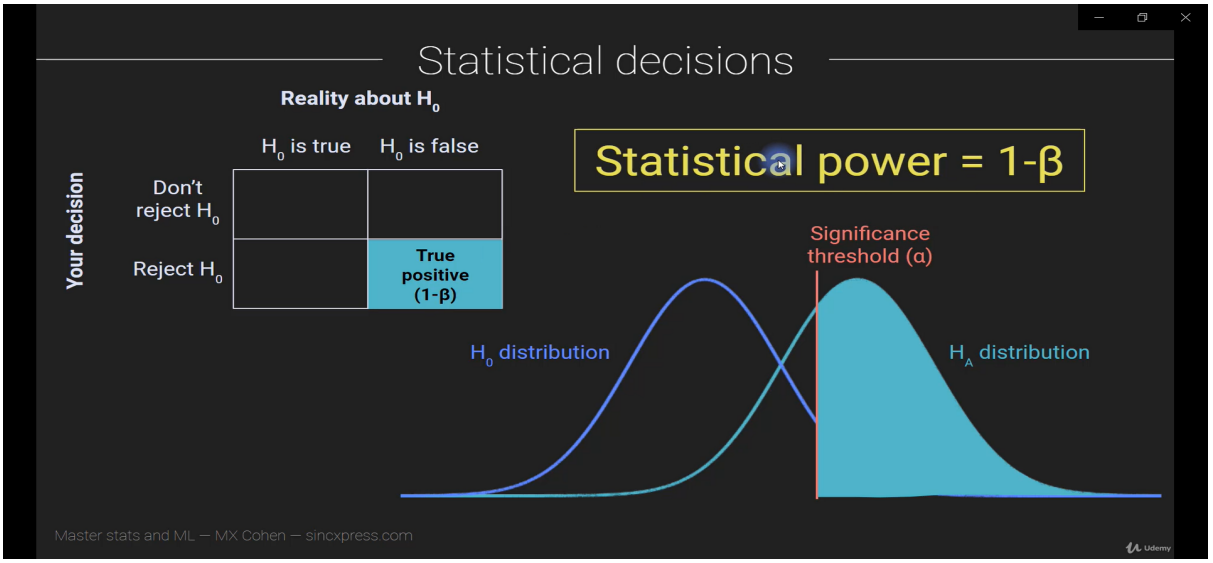


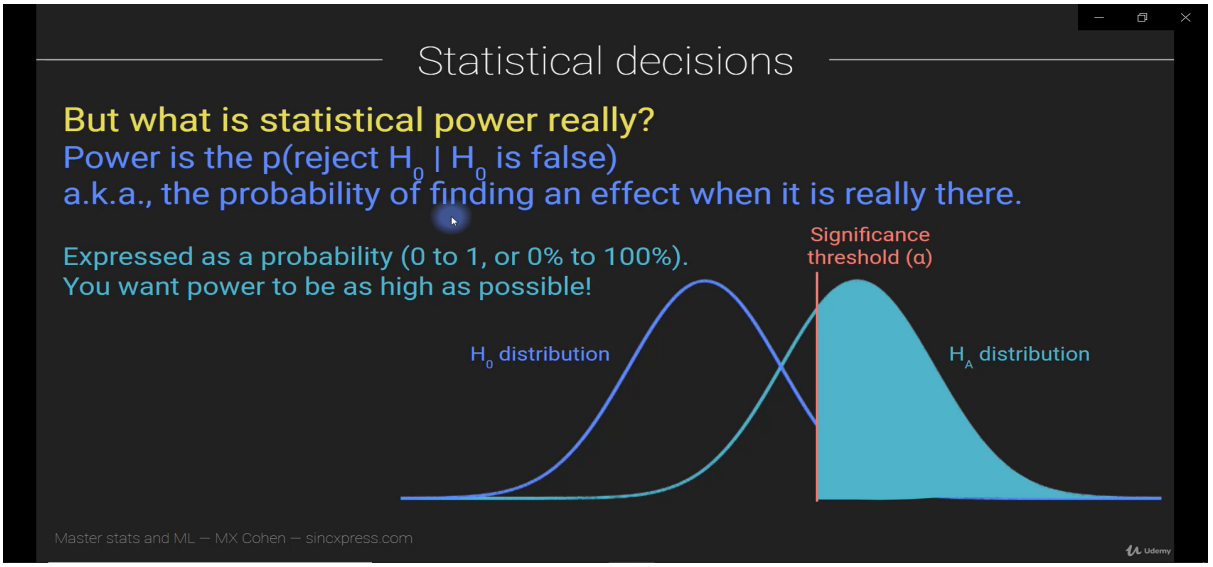
# 13. Statistical Power and Sample Sizes

**Table of Content**

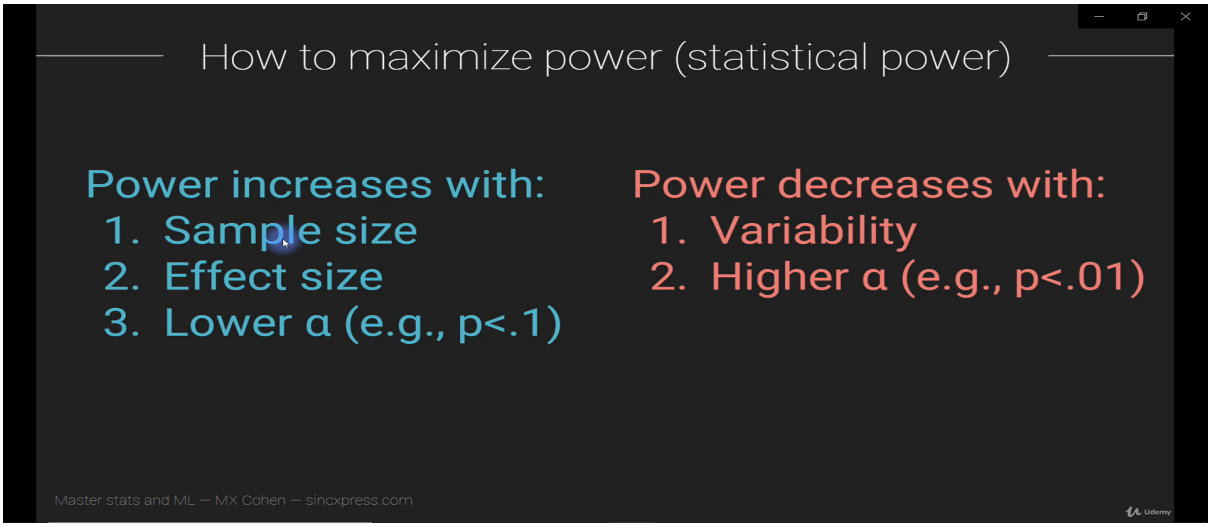
## What is Statistical Power ?

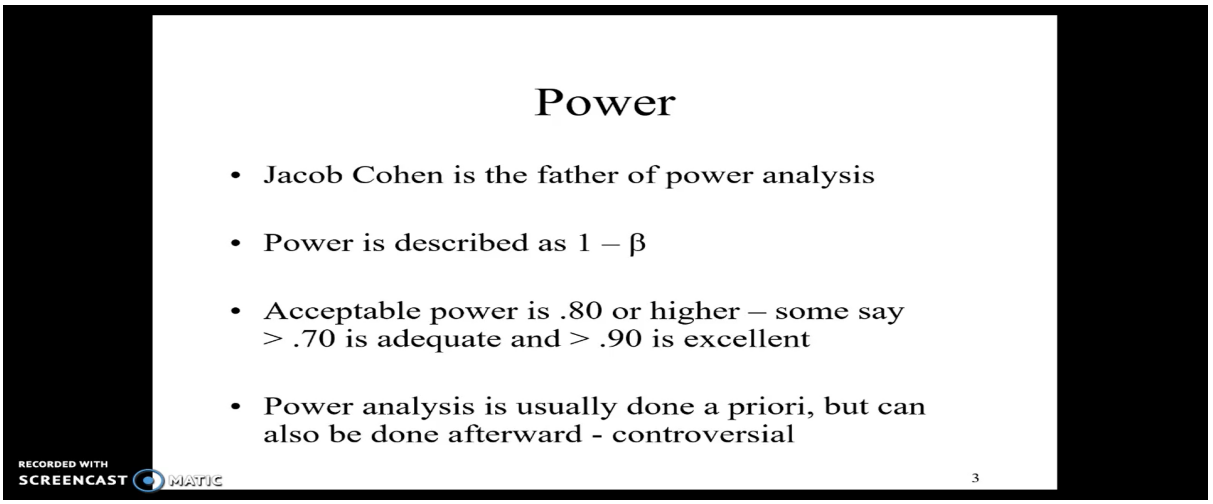


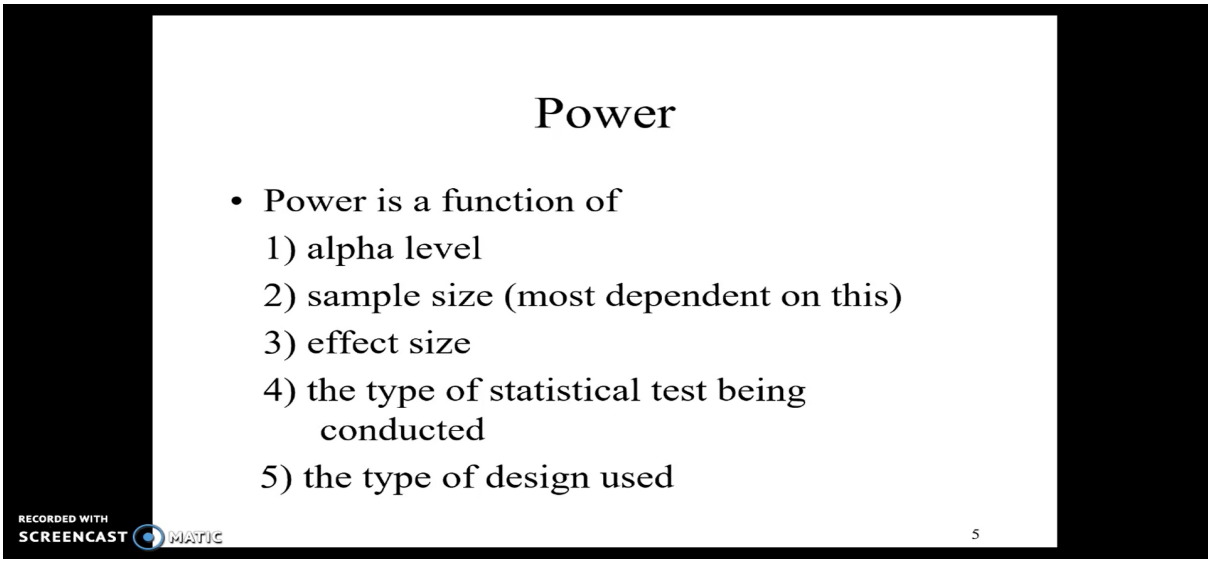



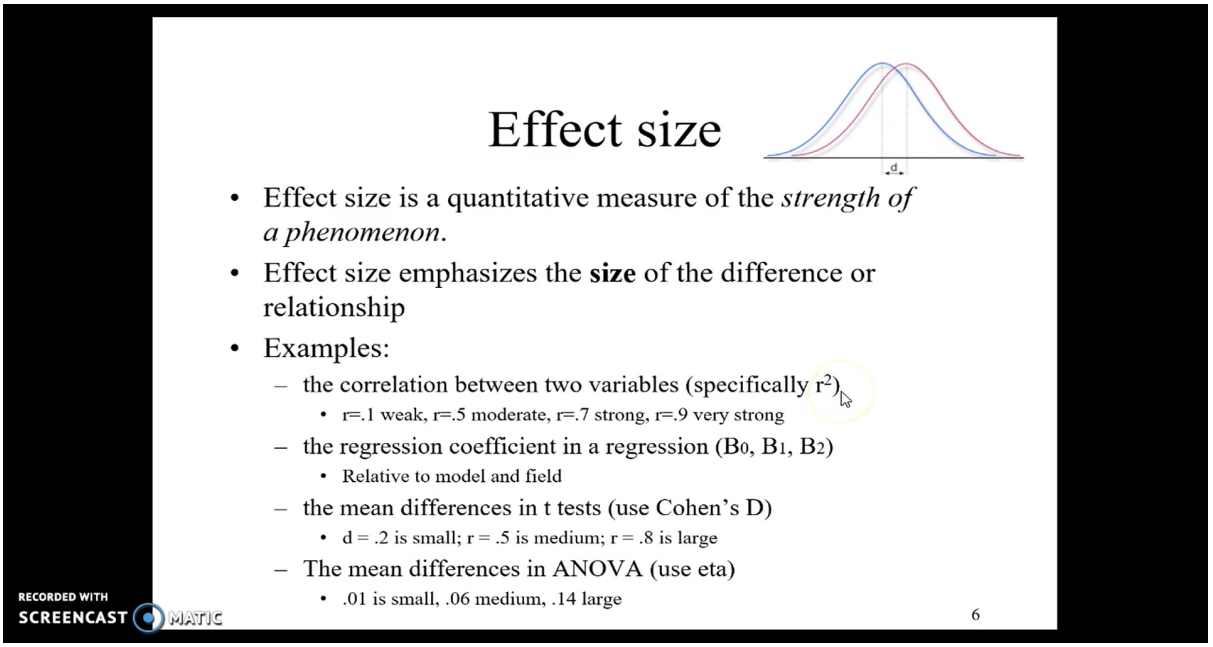
## How to maximise power ?



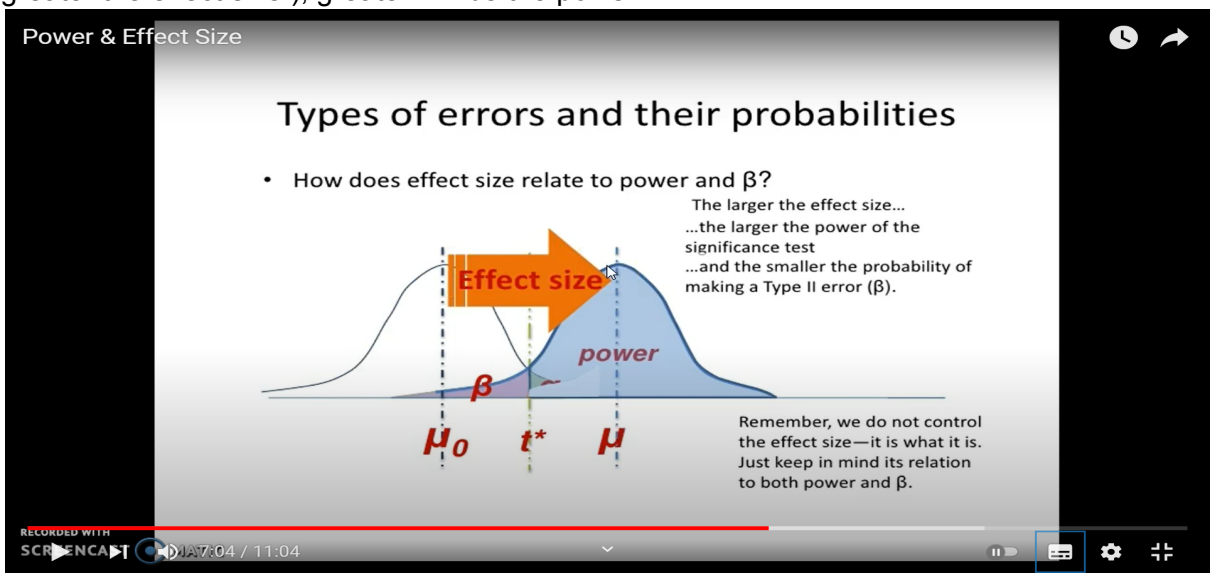



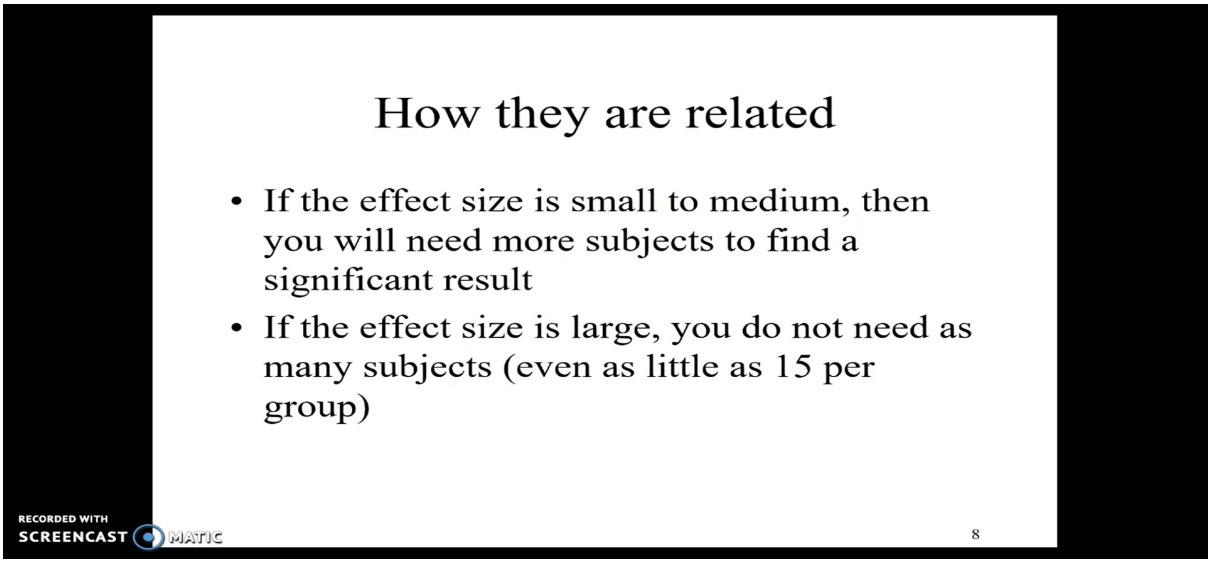
## What is Effect Size ?



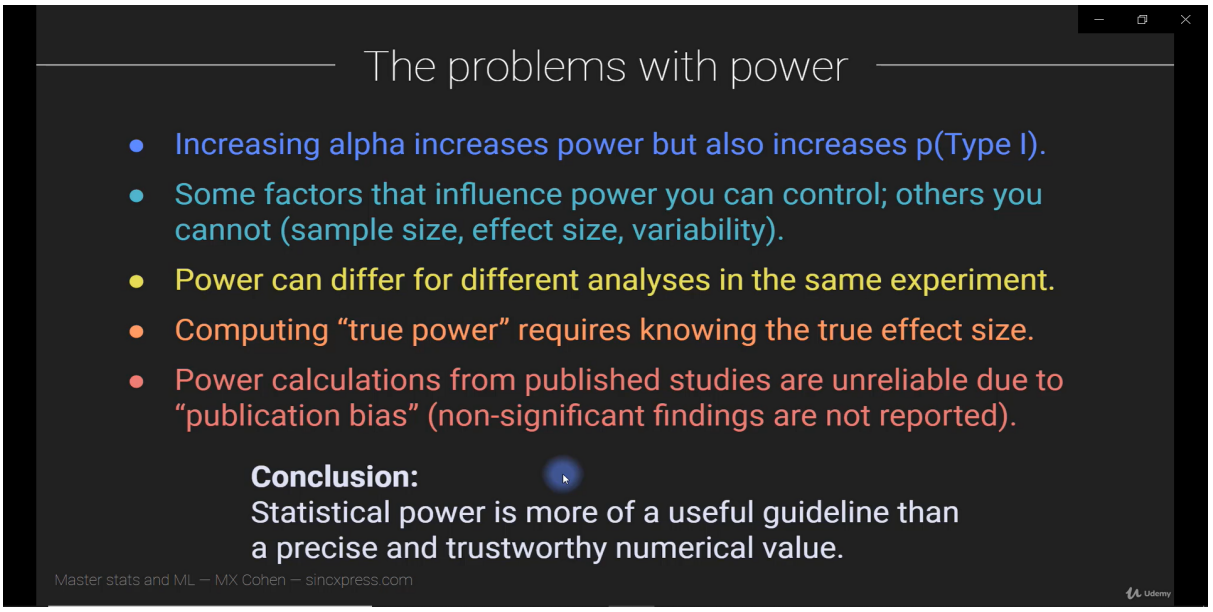

## Effect of effect size on Power:

Greater the effect size ( in the below example, greater the difference between the means, greater the effect size ), greater will be the power.

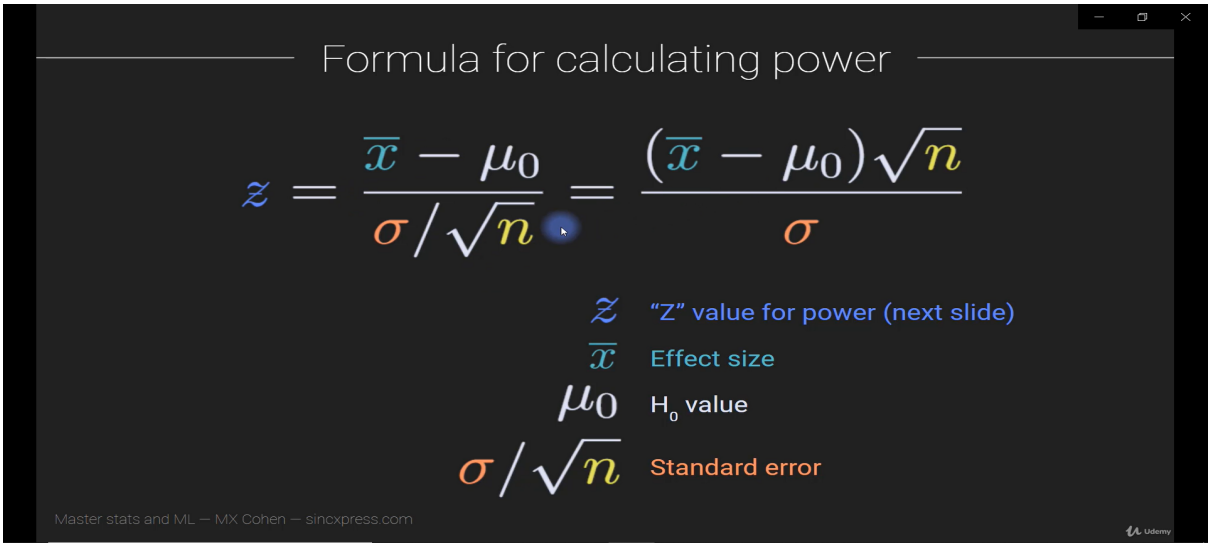


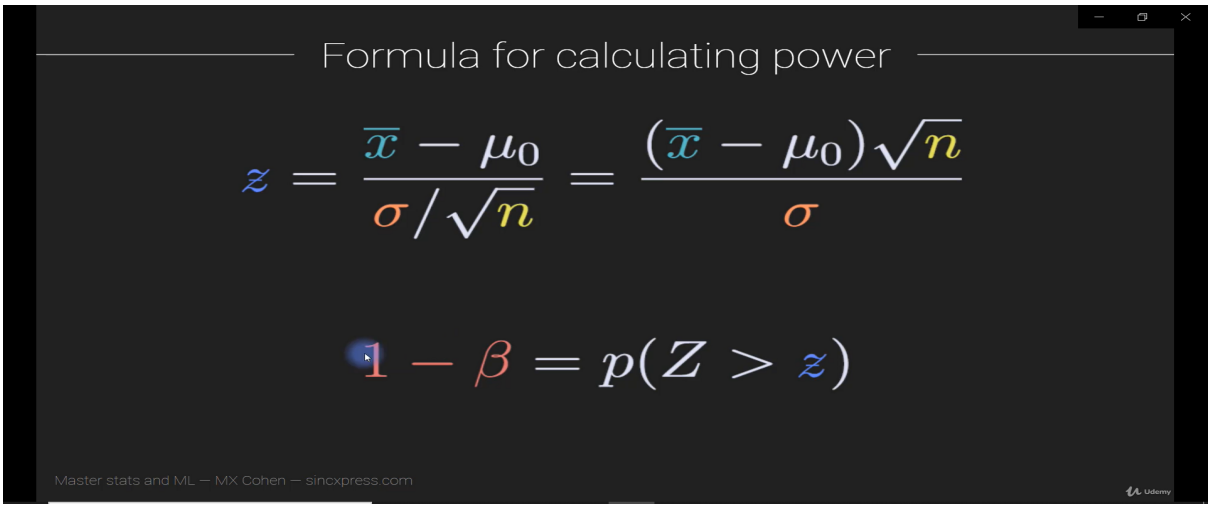



## Problems with Power

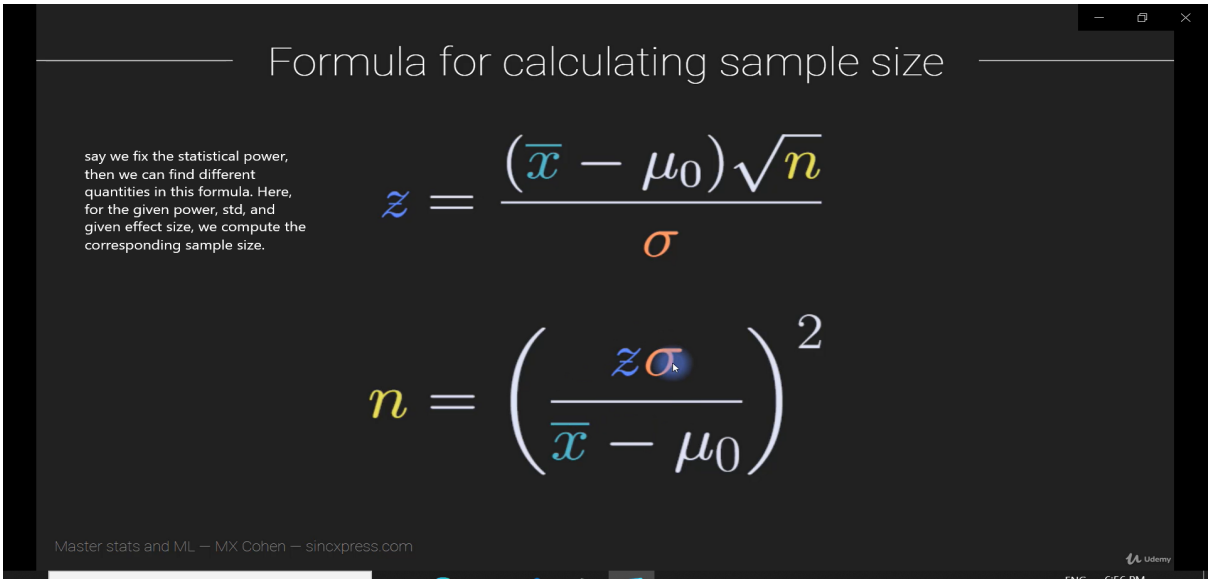


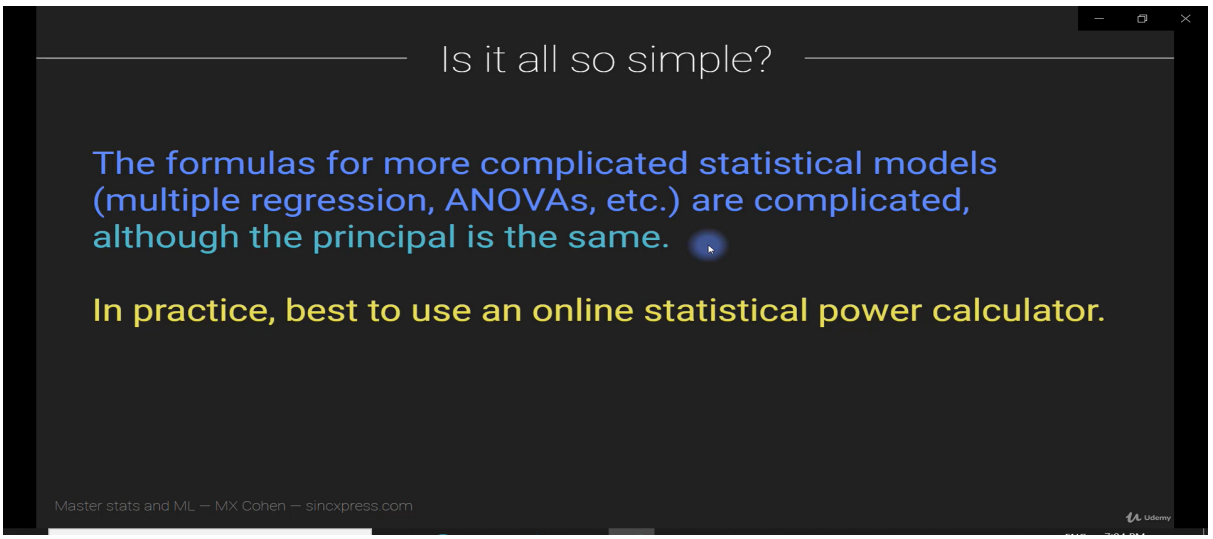
## Estimating Power and Sample Sizes



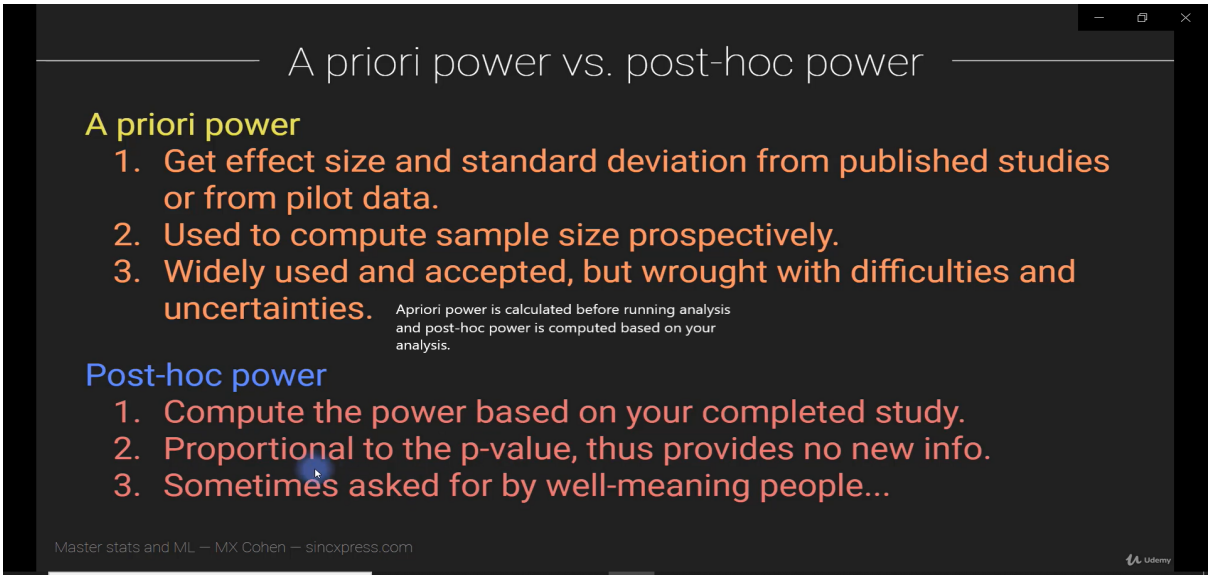




## Calculating sample size






## A Priori Power vs post-hoc power




## Great Online Tool for Calculating Power