# Sample and Population

## Population vs Sample

### Population:

* Collection of all items of interest (N)
* The numbers obtained when using the population are called parameters
* Hard to define & hard to observe in real life

### Sample:

* Subset of the population (n)
* The numbers obtained when using the sample are called statistics
* Less time consuming
* Less costly (cheaper)

You will almost always be working with sample data

## Samples have two defining features:

### Randomness:

* A random sample is collected when each member of the sample is chosen from the population strictly by chance

### Representativeness:

* A representative sample is a subset of the population that accurately reflects the members of the entire population

# Descriptive Statistics

## Data Types

### Categorical

* Describes categories or groups

### Numerical

* Represents either discrete or continuous numbers
* Discrete: number of children you have
* Continuous: the temperature outside

## Measurement Levels

### Qualitative

#### Nominal

* Arent numbers and cannot be ordered
* The seasons, car types etc..

#### Ordinal

* Arent numbers but can be ordered
* Rating food: disgusting, unappetizing, neutral, tasty, delicious

### Quantitative

#### Interval

* Don't have a true 0
* Farenheit, Celcius
* Can't say 40 degrees C is double as hot as 20 degrees C because 0C is not a true 0

#### Ratio

* Have a true 0
* Number of objects, distance, time
* Kelvin
* We can say 40K is double as hot as 20K because 0K is a true 0

# Statistical Measures

## Mean, Median, Mode

### Mean

* Add all the components and divide by n
* Easily affected by outliers

### Median

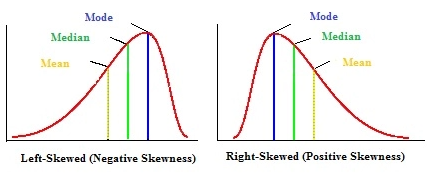
* The middlemost option

### Mode

* The value that occurs most often
* No mode if all options occur once (or you can have multiple modes)

## Skewness

* Indicates whether the data concentrated on one side



### Right Skew

* Mean > Median
* The outliers are to the right (tail trails off to the right)

### Zero Skew

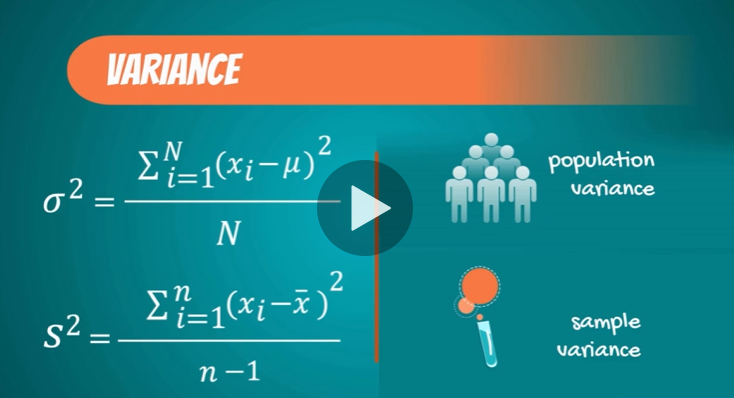
* Also called symmetrical
* Mean = Median = Mode

### Left Skew

* Mean < Median
* The outliers are to the left (tail trails off to the left)

## Variance

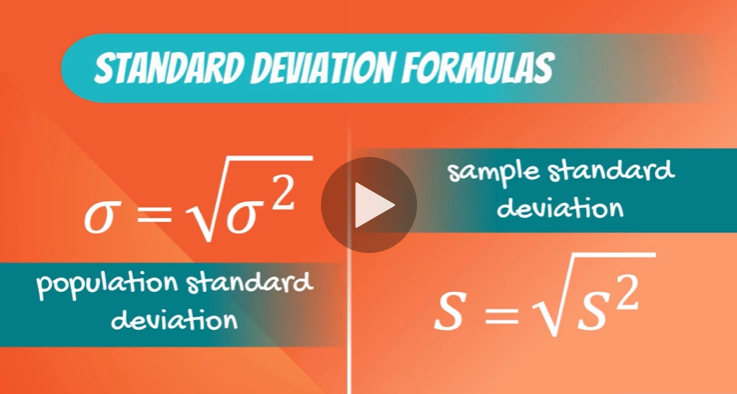
* A sample statistic is an approximation of the population parameter
* If you sample the data 10 separate times you will get 10 different measures
* Statisticians solved this problem by defining equations differently for sample and population purposes
* Variance measures the dispersion of a set of data points around their mean



* Sample variance is higher due to the fact that there’s a higher potential variability when considering the whole population (there could be some outlier points that the sample simply didn’t capture and we want to include this uncertainty in our sample calculation)

## Standard Deviation

* The most common measure of variability for a single dataset

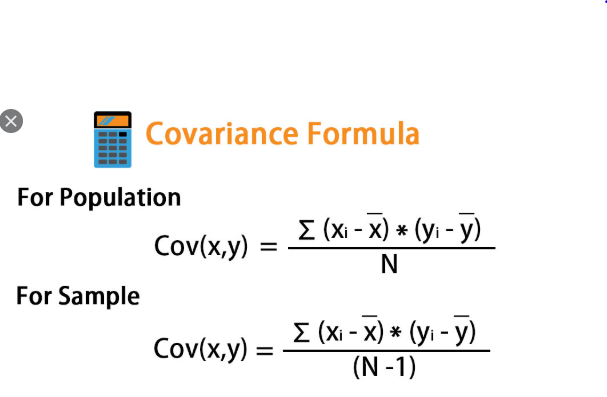


## Coefficient of Variation (CV)

* Standard deviation divided by the mean
* Or relative standard deviation
* We need this so we can compare the standard deviation of two different datasets – we need to normalize by it’s own mean
* It is unit-less

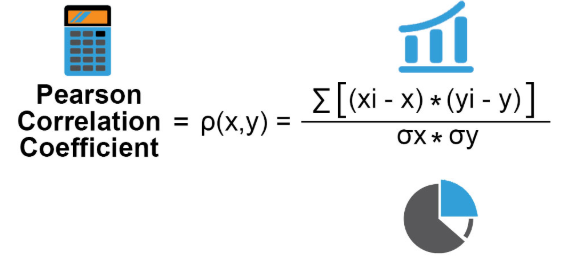
## Covariance

* The relationship between the movement of two variables



## Correlation Coefficient

* The covariance of two variables divided by the product of the standard deviations of both variables



### Perfect Positive Correlation

* The entire variability of one variable is explained by the other
* Correlation coefficient of 1

### Independent Correlation

* Correlation of 0

### Perfect Negative Correlation

* The entire variability of one variable is explained by the other (inversely)
* Correlation coefficient of -1

### Causality

* It is important to understand the direction of causal relationships
* In the housing business – the size of the house causes the price to be high, and not vice-versa
* Correlation does not imply causation

# Distributions

* A probability distribution
* Examples: Normal, Binomial, Uniform
* A distribution is a function that shows the possible values for a variable and how often they occur
* Rolling a dice: the probability of all outcomes is 1/6 - all else is 0
  + This is called a discrete outcome distribution
* A distribution is not a graph itself. The graph is just a visual representation

### Example: Rolling 2 Dice

* Possibilities?
  + (1,1) (2,1) (3,1) …(5,6) (6,6)
* What’s the probability of getting a sum of 1? It’s zero (impossible)
* What’s the probability of getting a sum of 2?
  + Possible only when both dice = 1
  + This happens one time out of the 36 (6 ^ 2) possible outcomes
  + 1 / 36
* What’s the probability of getting a sum of 3?
  + Possible for (1,2) and (2,1)
  + 2 / 36

## Inferential Statistics

* Methods that rely on probability theory and distributions to predict population values based on sample data

## The Normal Distribution

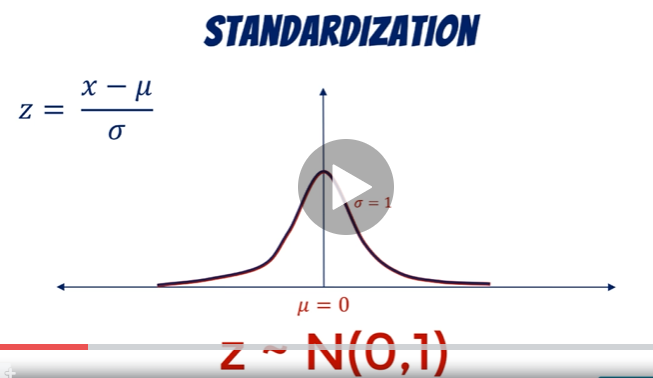
* Also known as bell curve or Gaussian distribution
* Mean = median = mode

## Standardization

* The process of transforming the variable to a mean of 0 and a standard deviation of 1



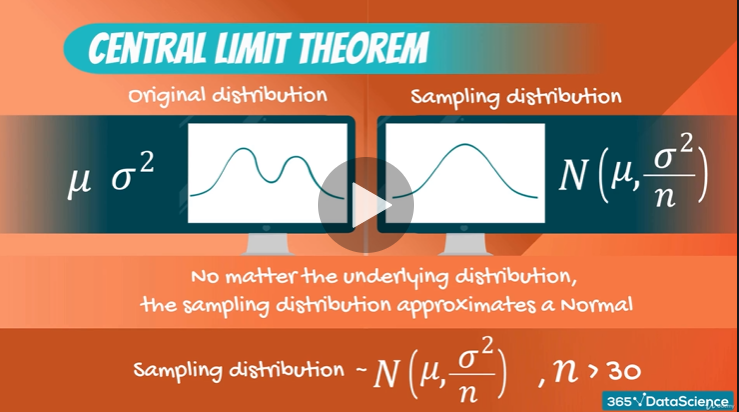
* When we standardize a normal distribution – the result is a standard normal distribution



* The standardized variable is called the Z-Score

## Central Limit Theorem

* No matter the underlying distribution , the sampling distribution approximates a normal distribution
* For CLT to apply we need a sample of at least 30 observations
* CLT allows us to perform tests, solve problems, and make inferences using Normal Distribution even when the population is not normally distributed



### Sampling Distribution

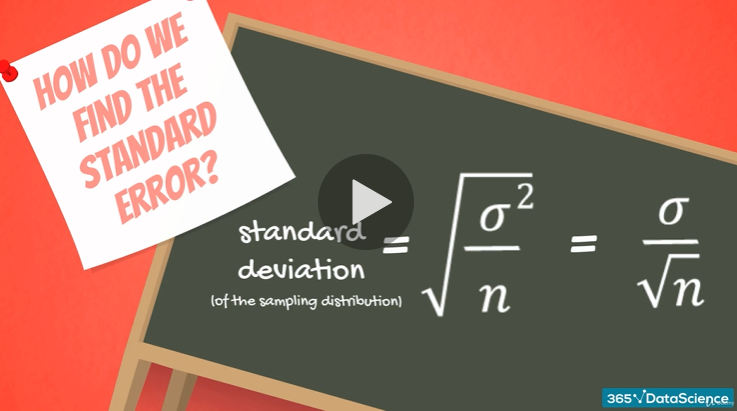
* A distribution formed by different samples measuring some quantity of the samples picked
* If we analyze these results they’re different – but they concentrate around some value
  + This value is (approximately) the population mean
* The variance of the resultant sampling distribution is the population variance divided by the sample size

### Reasons to use the Normal Distribution

* They approximate a wide variety of random variables
* Distributions of sample means with large enough sample sizes could be approximated to normal
* All computable statistics are elegant
* Decisions based on normal distribution insights have a good track record

## Standard Error

* The standard deviation of the distribution formed by the sample means (the standard deviation of the sampling distribution



* Like standard deviation – the standard error shows variability
* It shows the variability of the means of the different samples we extracted
* It’s used in most statistical test because it shows how well you approximated the true mean
* It decreases as the sample size increases (bigger samples give a better approximation of the population

# Estimators & Estimates

* Estimates are specific value approximations depending solely on sample information
* Estimators are a type of statistic

## Types of Estimates

### Point Estimates

* A single number
* Located exactly in the middle of the confidence interval
* For example:
  + the sample mean is an **estimate** of population mean
  + The sample variance is an estimate of population variance

### Confidence Intervals

* An interval

### Bias

#### Unbiased Estimator

* Expected value = population parameter

#### Biased Estimator

* Say you’re told that the estimate for the average height of someone in the USA is the average of the sample plus 1 foot
  + We would say that this estimator has a bias of 1 foot

### Efficiency

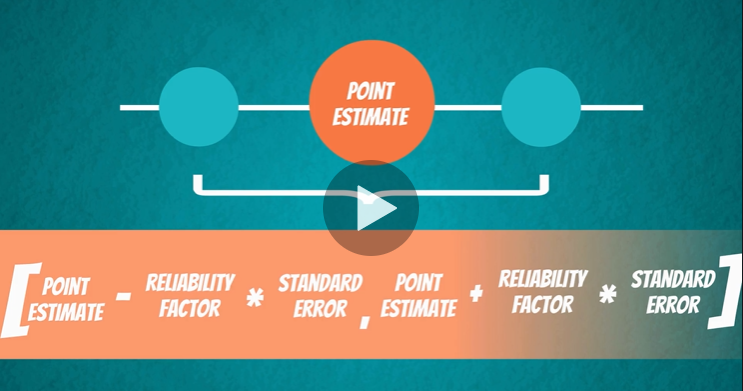
* The most efficient estimator is the unbiased estimator with the smallest variance

## Confidence Intervals

* A more accurate representation of reality
* There is still uncertainty (which we measure as a % confidence)
* You can never be 100% confident unless you count the entire population
* 95% CI means there is only a 5% chance that the population parameter is outside the range
* A confidence interval is the range within which you expect the population parameter to be
* 100% confidence is completely useless as all possibilities must be within the range

### Confidence Level

* 1 – alpha
* If we want to be 95% confident – alpha is 5%
* 95% is the accepted norm in the real world



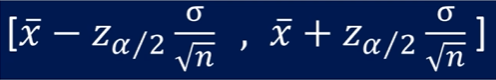


### Known Population Variance

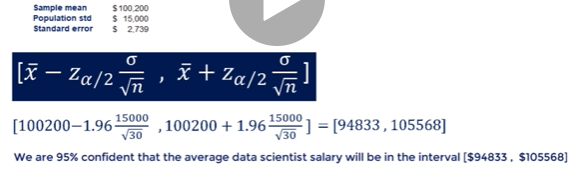
* There are two different cases for creating a confidence interval: known and unknown population variance
* Even if you have a lot of data it doesn’t necessarily mean you have population data
  + For example: even Google doesn’t have population data for all people who browse the internet
* If Google wants to use statistical methods to target users with Google Ads – they will basically be using sample data with an unknown population variance to guess their preferences

#### Example: Estimate Data Scientist Earnings

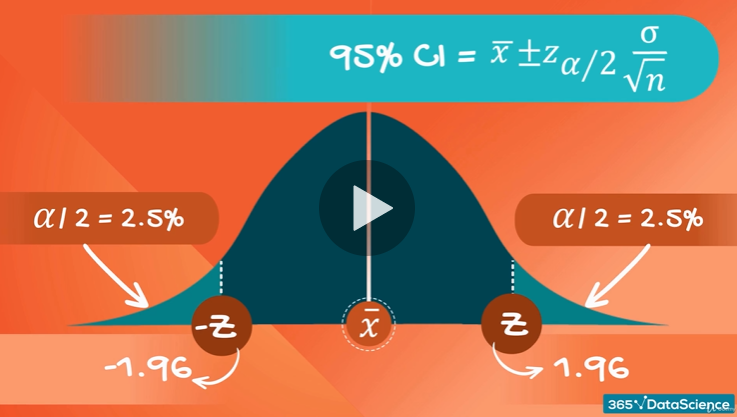
* Let’s say the population standard deviation is $15,000 annually
* Sample mean is $100,200
* The sample you’re using consists of 30 salaries



* The formula contains alpha / 2 because the total alpha has to be split between the left and right quadrents of the distributions tails
* Z (the reliability factor) is the standardized variable with a normal distribution
  + Comes from the standard normal distribution table
* Alpha is defined based on confidence level (95% confidence alpha is 5%)

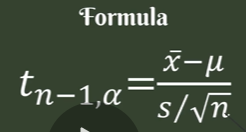






### Student’s T-Distribution

* “Student” was the penname of William Sealy Gosset while he worked for Guinness and couldn’t publish under his real name (company policy)
* Allows inference through small samples
* Unknown population variance
* Huge real-life applications
* Looks like a Normal Distribution except with fatter tails
* Higher dispersion of variables and more uncertainty
* In the same way the Z-statistic is used for the Normal distribution – the t-statistic is used for Student’s T



* There are degrees of freedom (n-1 in the formula above)
  + If the sample size is n, the degrees of freedom is n-1
* For samples bigger than 50 just use the Z-table instead



#### Example: Estimate Data Scientist Earnings

* This time we don’t have the population standard deviation
* Sample of 9 compensations
  + Mean: 92533
  + Sample Standard Deviation: 13932
  + Standard Error: 4644
* Since we don’t know the population variance we’ll use Student’s T
* Get the t-statistic from the t-table using alpha and dof (n-1)
  + In this case for 95% CI: t8,0.025 = 2.31

### Margin of Error

* Margin of error is half the span of the confidence interval

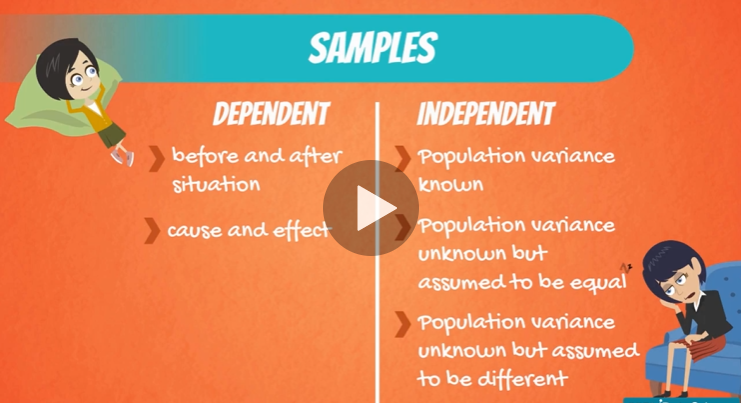


## Advanced Confidence Intervals

* Sometimes we’re looking into confidence intervals of two populations
* More important – wide range of real-world applications

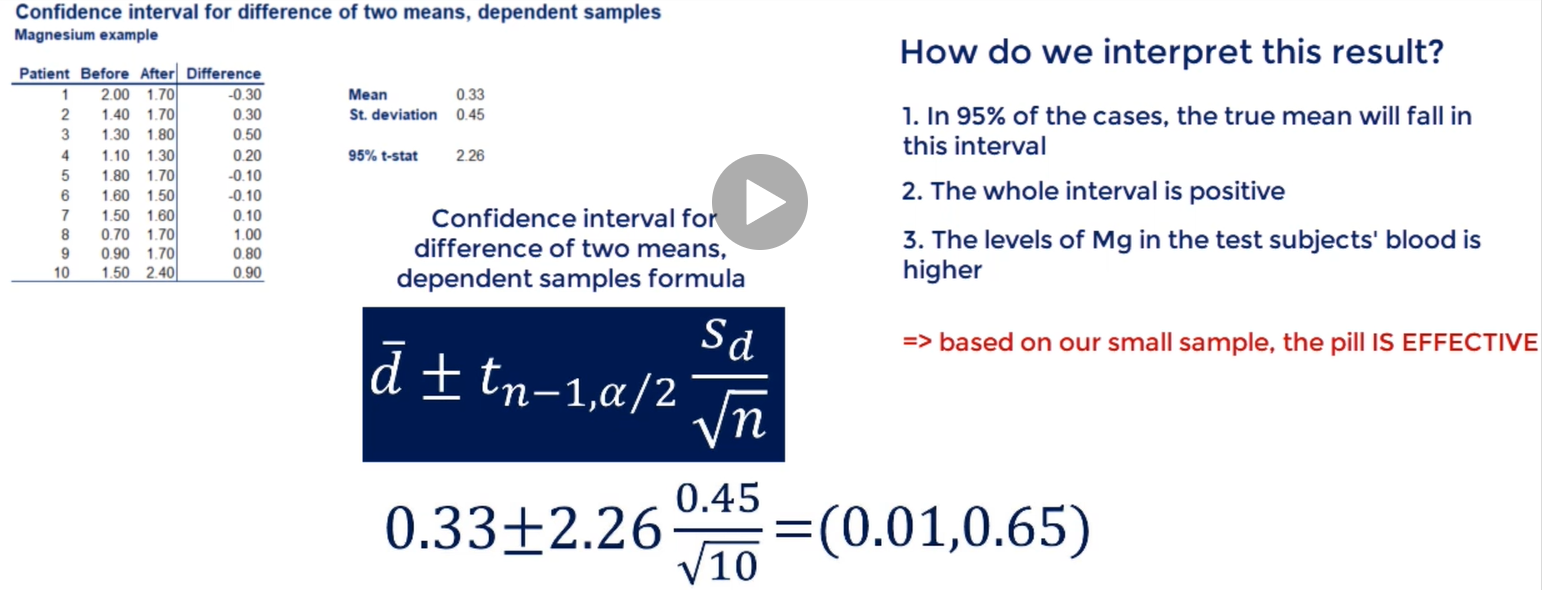
### Samples

* Can be dependent or independent



#### Dependent Samples

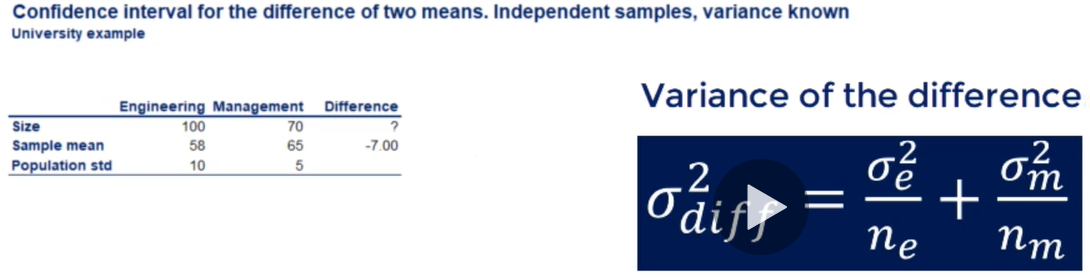
* Often used in medicine
  1. Lets say you have a pill that increases the % of magnesium in the blood
  2. Take a sample of 10 people and test their magnesium levels before and after taking the pill
  3. The people are the same so it’s clear the samples are dependent (before sample and after sample)

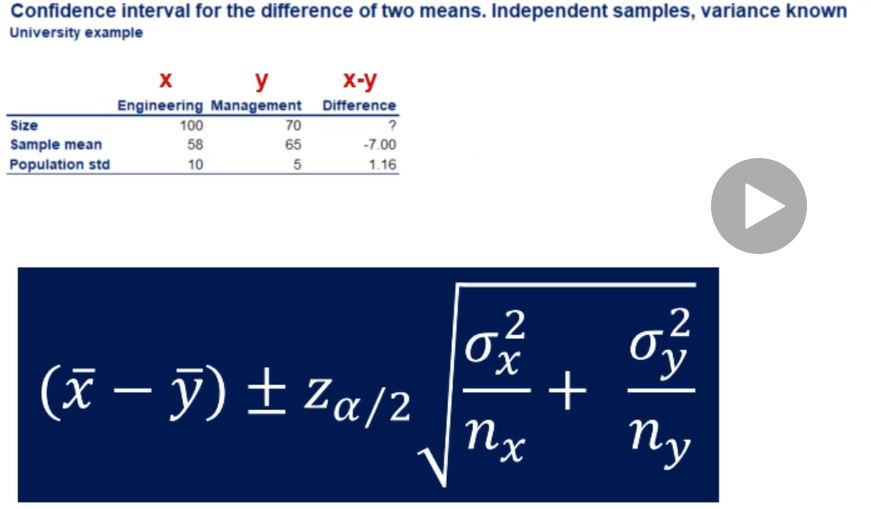


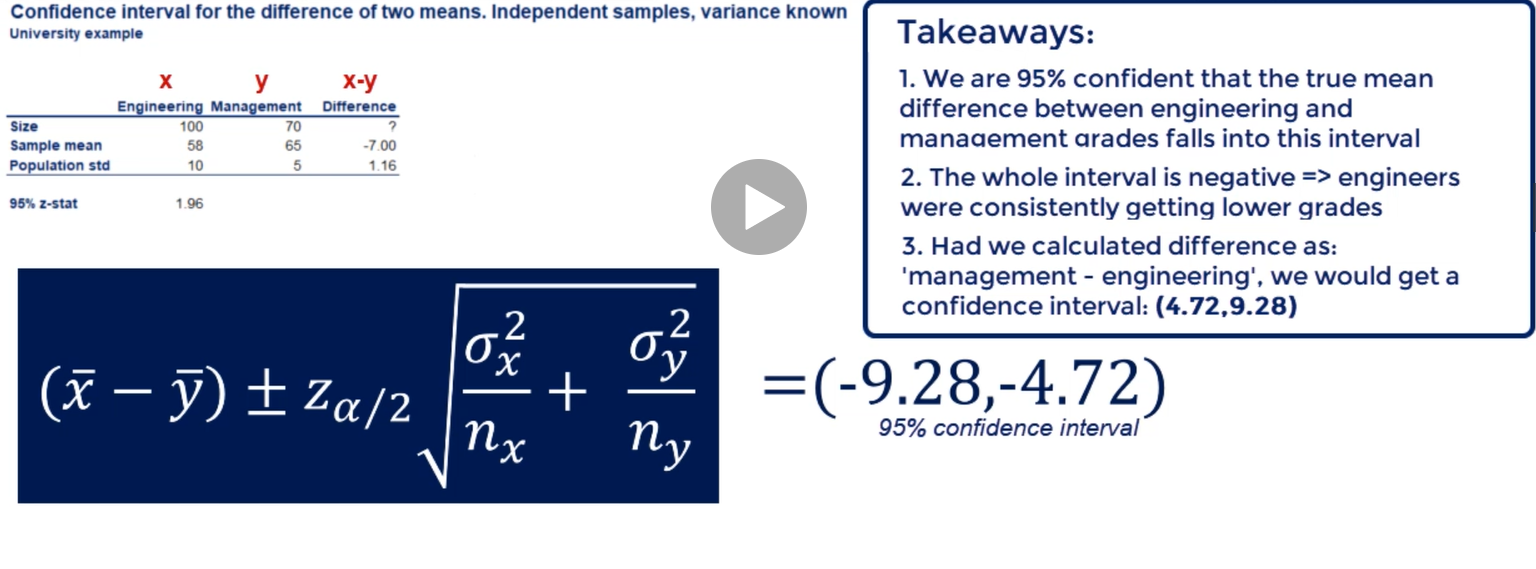
#### Independent Samples

They can come in 3 varieties:

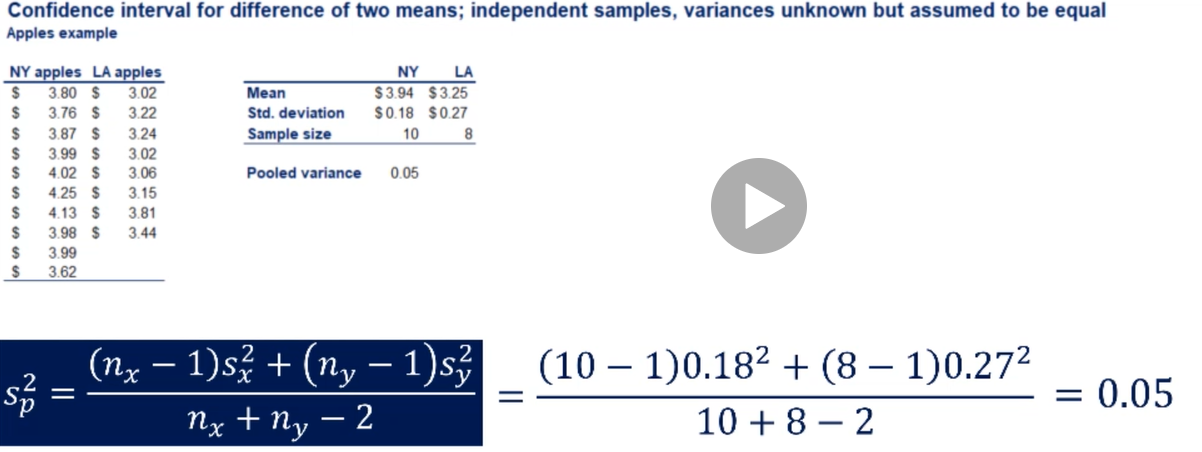
1. Known population variances

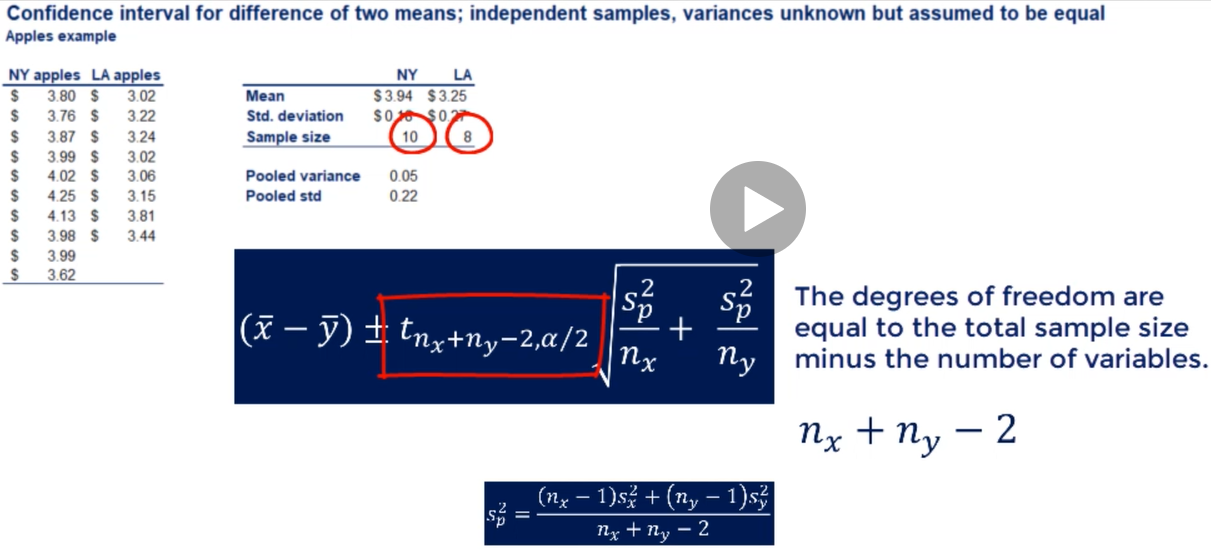


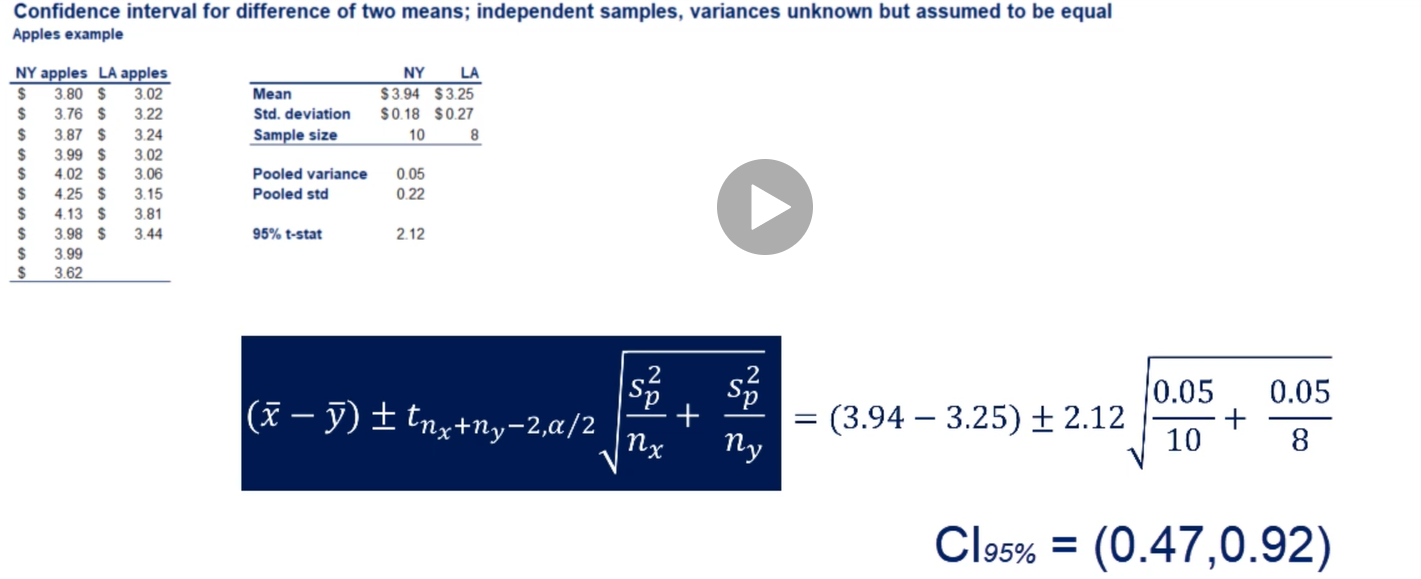




1. Unknown population variances (but assumed to be equal)

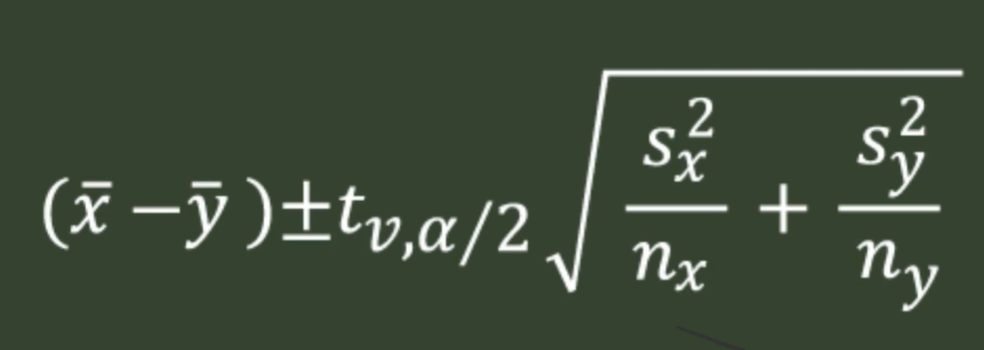


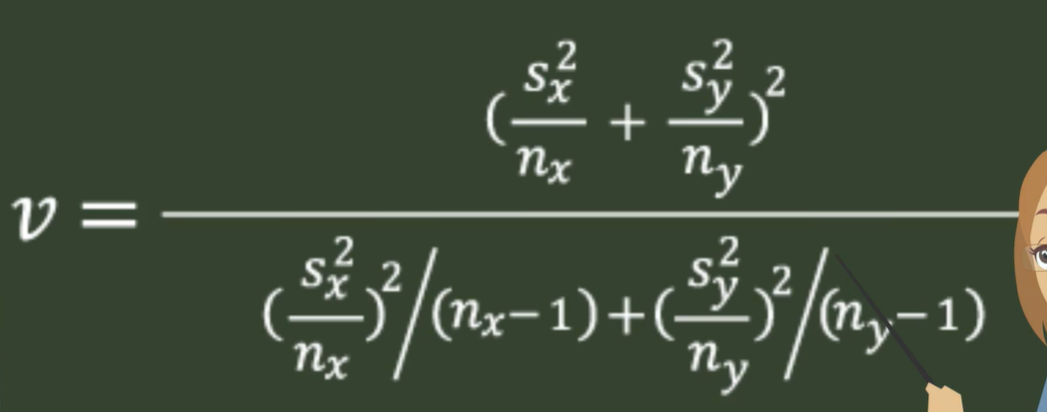




* Interpretation of above: We are 95% confident that the actual difference between the two prices of apples in LA and NY is between $0.47 and $0.92 (NY apples being more expensive)

1. Unknown population variances (but assumed to be different)
   1. Example: comparing apples and oranges
   2. See equation below (the issue is coming up with the dof denoted “v”)

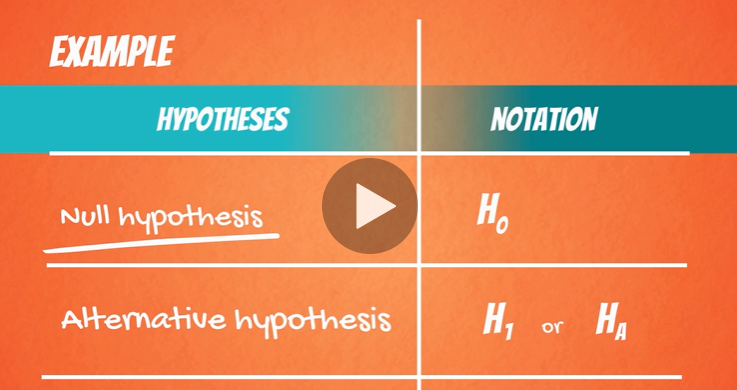




# Hypothesis Testing

## Introduction

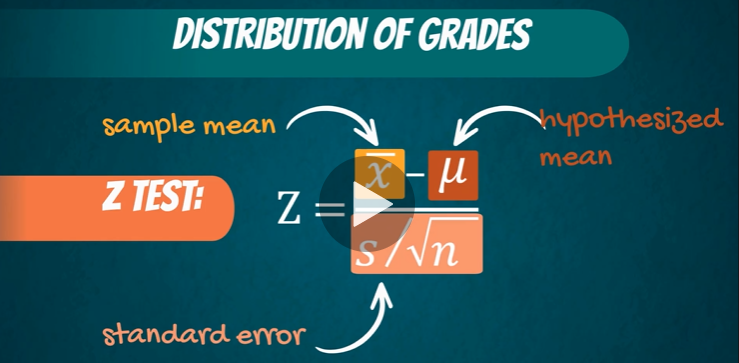
* There are 4 steps in data-driven decision making:
  1. Formulate a hypothesis
  2. Find the right test
  3. Execute the test
  4. Make a decision
* A hypothesis is an idea that can be tested
* The null hypothesis: the hypothesis to be tested (Ho)
* Alternative hypothesis: anything else (H1 or HA)



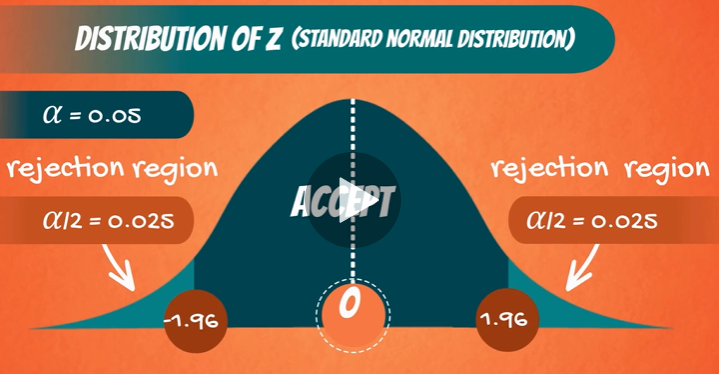
* The concept of the null hypothesis is “innocent until proven guilty” – assume it is true and try to prove otherwise to reject the null
* Outcomes of tests refer to the population parameter rather than the sample statistic
* Generally the researcher is trying to reject the null hypothesis
  + Think of the null hypothesis as the status quo and the alternative is the “change” or “innovation” that the statistician is trying to uncover
* The null hypothesis is the statement we are trying to reject, therefore the null is the present state of affairs while the alternative is our personal opinion

## Significance Level

* Denoted by alpha, and is the probability of rejecting the null hypothesis, if it is true
* Typical values of alpha are 0.01, 0.05, and 0.1 (value selected based on the % certainty you need)
  + 0.05 is commonly the significance level used (corresponding to 95% confidence)
* Example of challenging a hypothesis:

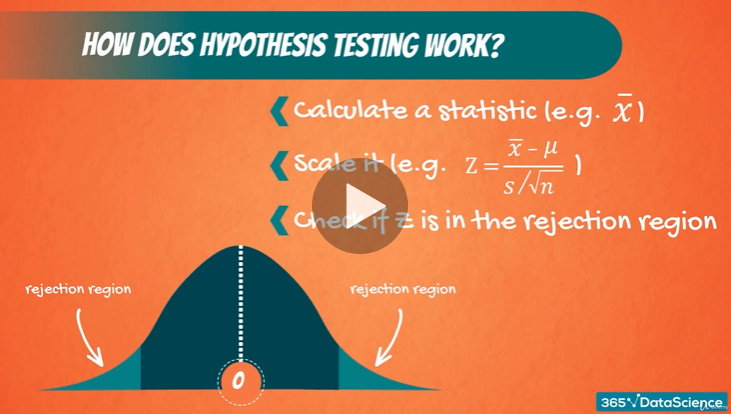


* The question above is: how big does Z need to be to reject the null that the mean population is some value?
  + Since we’re conducting a two-tailed test (the population average can either be higher or lower) the logic resembles the following diagram:



* The area of these regions depends on the significance level used in our test
* If our Z-score is higher or lower than 1.96 and -1.96 respectively, we can reject the null hypothesis that the population mean is the same as assumed

## Two-Sided Test



## One-Sided Test

* For example: the null hypothesis that data scientists make more than $125,000 USD
  + Ho: µo is bigger or equal to $125,000
  + H1: µo is less than $125,000



## Testing Errors

### Type 1 Error

* When you reject a true null hypothesis
* Also called a false positive
* Propability denoted as α
* The responsibility for making this error falls on the statistician ONLY

### Type 2 Error

* When you accept a false null hypothesis
* Also called a false negative
* Probability of making this error denoted as β
* Β depends on sample size and the magnitude of the effect
  + It is more likely to make this type of error
* The probability of rejecting a false null hypothesis is 1-β
  + This is also known as the power of the test
  + Can increase the power of the test by increasing the sample size

