Statistics

SAMPLE AND POPULATION

It is very hard to understand and analyze the population so we work on the samples drawn from the population. Sample statistics is used to estimate the parameters of the population.

Understanding the sample is less time consuming and less costly. The sample should be random and representative where each sample point is strictly chosen by chance.

CLASSIFICATION OF DATA

Data can be classified into two ways

1. Type of data
2. Measurement level

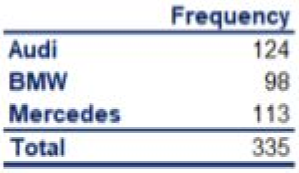
TYPE OF DATA

1. Categorical
2. Numerical
    a. Discrete
    b. Continuous

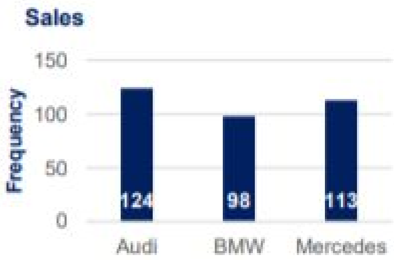
MEASUREMENT LEVEL

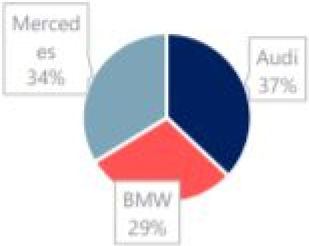
1. Qualitative
    a. Nominal - Categories where order doesn't matter.
    b. Ordinal - Follows order. For example rating (1-5)
2. Quantitative
    a. Interval - Uncommon, Does not have a true 0. For example temperature
    b. Ratio - Has a true 0. Represents most of the things. Example Age of person

VISUALIZATION - Categorical Data

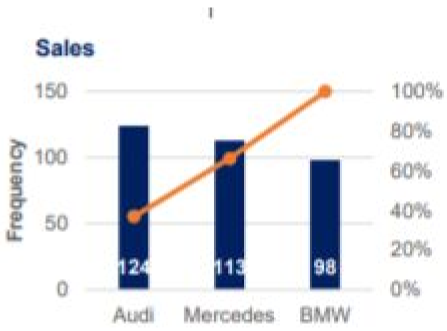
1. 1.Frequency distribution table
    a. Group by the categories and find the total frequency of occurring each category
    b. 

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

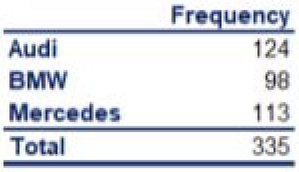
2. Bar charts
    a. Constructed from the frequency distribution table
    b. 

3. Pie charts
    a. Calculates the relative frequency of all the categories in the frequency table
    b. Pie chart not just compares the items from each other but also gives the share of the total
    c. Market share is mostly represented by Pie charts
    d. 

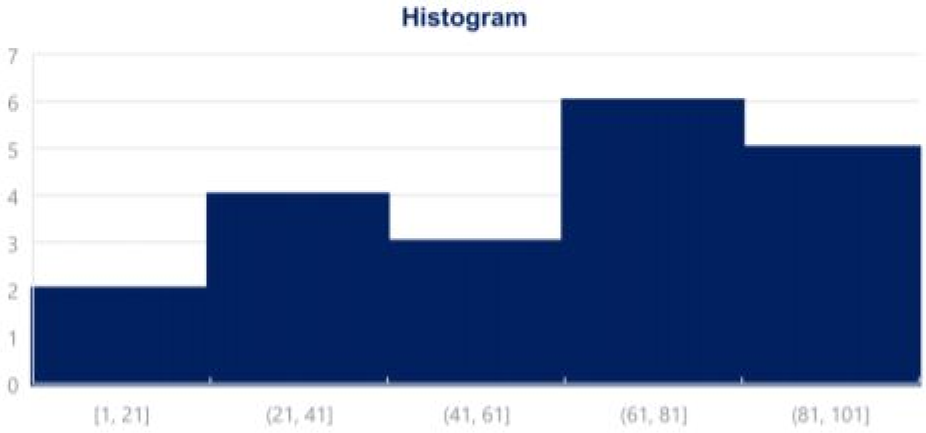
4. Pareto diagrams
    a. It is a special type of bar charts where categories are shown in the descending order of frequency.
    b. There is a curve also on the same graph showing the cumulative frequency
    c. It can be used to showcase the Top N categories.

d. 

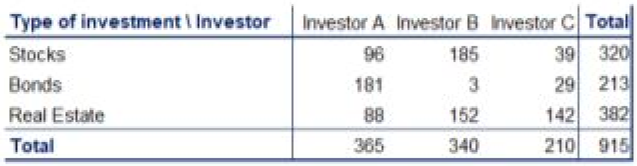
VISUALIZATION - Numerical Data

1. Frequency distribution table
    a. We group the data into intervals and then find the corresponding frequencies
    b. We can make from 5-20 intervals of our data
    c. We can also calculate the relative frequencies of the grouped data
    d. 

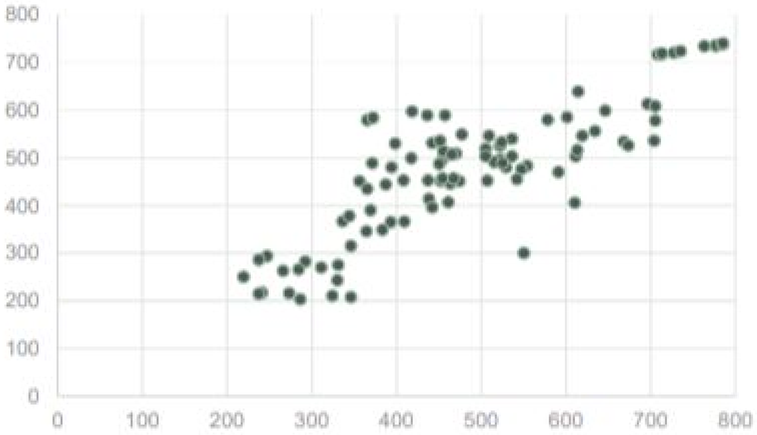
	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

2. Histograms
    a. Looks similar to Bar chart but conveys a different meaning
    b. X-axis represents the intervals of the data and the Y-axis represents the frequencies
    c. Histograms can also be created from unequal intervals. For example Age group 18-25, 25-35, 35-50, 50+
    d. 

VISUALIZATION - Multiple variables

1. Cross tables
    a. 

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

2. Scatter plots
    a. Used to represent two numerical variables
    b. X and Y axes represent each of the variable and each point on the graph shows 1 record
    c. Used to see if there is any trend in the data
    d. 

MEASUREMENT OF CENTRAL TENDENCY - Mean, Median and Mode

1. Mean is the average of all the observations
2. Mean is affected by the outliers
3. Median is the central item and is unlikely to get affected by the outliers
4. Mode is the most common observed value
5. None of them best describes the data. One should look at all the three statistics to describe the central tendency of the data

MEASURING SKEWNESS

1. If mean > median, it is positive/right skewed. It means the tail is leading to the right which means outliers are towards the right
2. If mean < median, it is negative/left skewed. It means the tail is leading to the left which means the outliers are towards the left

MEASUREMENT OF THE VARIABILITY

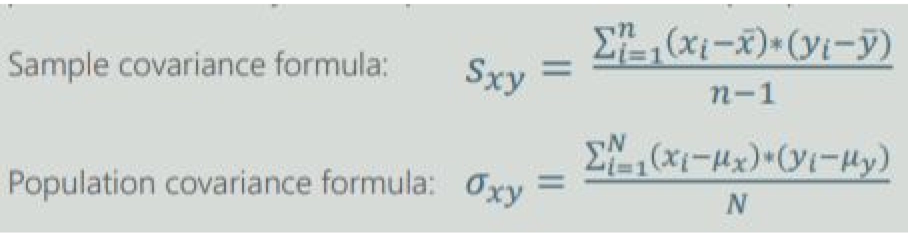
Variability of the data is mainly represented by the 3 statistics

Variance, Standard Deviation and coefficient of variance

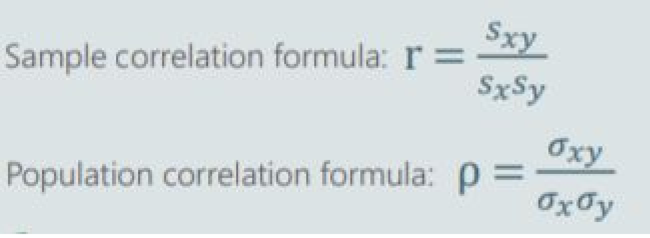
1. Variance measures the dispersion of the data points around their mean
2. Variance is equal to the sum of the squared differences between their observed value and their mean and then divided by the total number of observations
3. The closer the numbers are from mean, smaller the number we would obtain
4. The denominator in the formula of mean of population is N whereas in case of sample, it is (n-1)
5. Variance is also called as the second central moment
6. The unit of measurement of variance is squared so it is hard to relate it to the underlying data so that derives the need of the Standard Deviation
7. Standard Deviation is the square root of the variance and is expressed in the same unit as the unit of the data
8. Coefficient of variance is the relative standard deviation to the its mean and is equal to the standard deviation divided by the mean. Comparing the standard deviation of two different datasets is meaning less but comparing the coefficient is not

MEASURES OF RELATIONSHIP BETWEEN VARIABLES

Covariance and correlation

1. Covariance is the measure of the joint variability of the two variables. It gives a sense of the direction in which the two variables are moving. Its outcome is positive if both the variables move in the same direction otherwise negative. If the movement of the variables is independent then the covariance will be 0
2. 

   Sample covariance formula: $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})*(y_i-\bar{y})}{n-1}$

   Population covariance formula: $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)*(y_i-\mu_y)}{N}$
3. Covariance(similar to variance) is very hard to interpret because it can take even too small or too large values so this derives the need of the correlation
4. Correlation scale is between -1 and 1. The correlation of 1 means that both the variables are moving in the same direction and are perfectly correlated. Correlation of -1 means that both the variables are moving in exact opposite direction but are perfectly correlated. 0 means no correlation
5. 

   Sample correlation formula: $\mathrm{r} = \frac{s_{xy}}{s_x s_y}$

   Population correlation formula: $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
6. As a general rule, any correlation between -0.2 to 0.2 can be disregarded
7. Correlation does not imply causation. In the housing example, the size of the house causes the price of the house not the vice versa

INFERENTIAL STATISTICS

1. This relies on the probability theory and the distribution of the data
2. It uses the sample statistics to predict the population parameters
3. [TODO] Mention all the types of probability distributions here

CENTRAL LIMIT THEOREM

1. When we are referring to a distribution formed by the samples then we use the term called as sampling distribution.

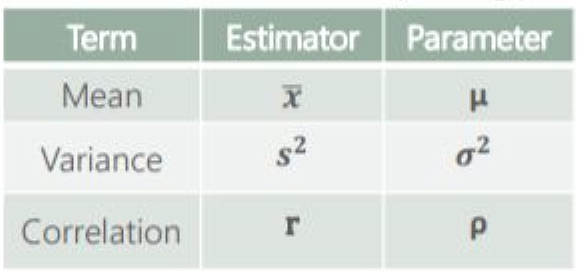
2. When we are dealing with a sample distribution of the means of the samples drawn from the population then the mean of the sampling distribution approximates the population mean
3. No matter the distribution of the population, the sampling distribution of the mean will approximate a normal distribution
4. Usually a CLT to apply we need a sample size of at least 30 observations.

STANDARD ERROR

1. Standard error is the standard deviation of the distribution formed by the sample means
2. It is defined as the standard deviation divide by the square root of the n
3. It is the variability of the means of the different samples we extracted
4. As the sample size increases, standard error decreases

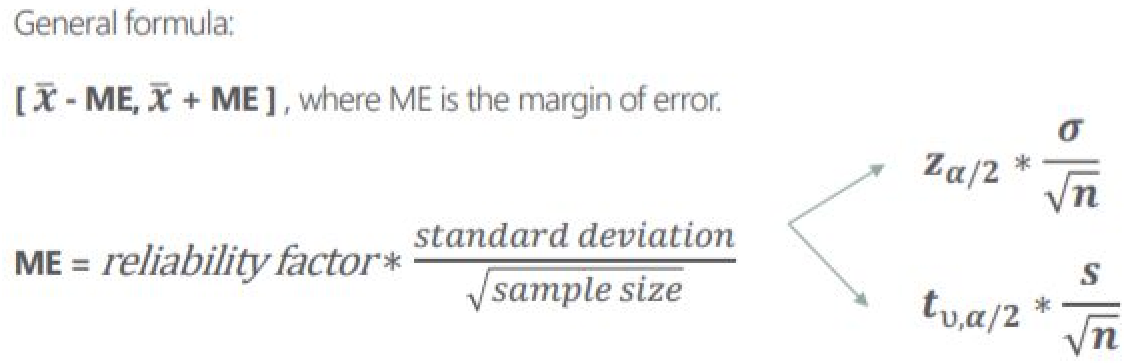
ESTIMATORS AND ESTIMATES

Point estimates and confidence interval estimates

1. A point estimate is a single number where as the confidence interval is an interval with an associated confidence level
2. Point estimate is located exactly in the middle of the confidence interval
3. For example, the sample mean is the estimate of the population mean and similarly the sample variance is the estimate of the population variance
4. There is efficiency and bias associated with each estimators. The most efficient estimator is the unbiased estimator with smallest variance
5. 

Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	r	ρ

CONFIDENCE INTERVAL



General formula:

[$\bar{x}$ - ME, $\bar{x}$ + ME] , where ME is the margin of error.

$$\textbf{ME} = \textit{reliability factor} * \frac{\textit{standard deviation}}{\sqrt{\textit{sample size}}}$$

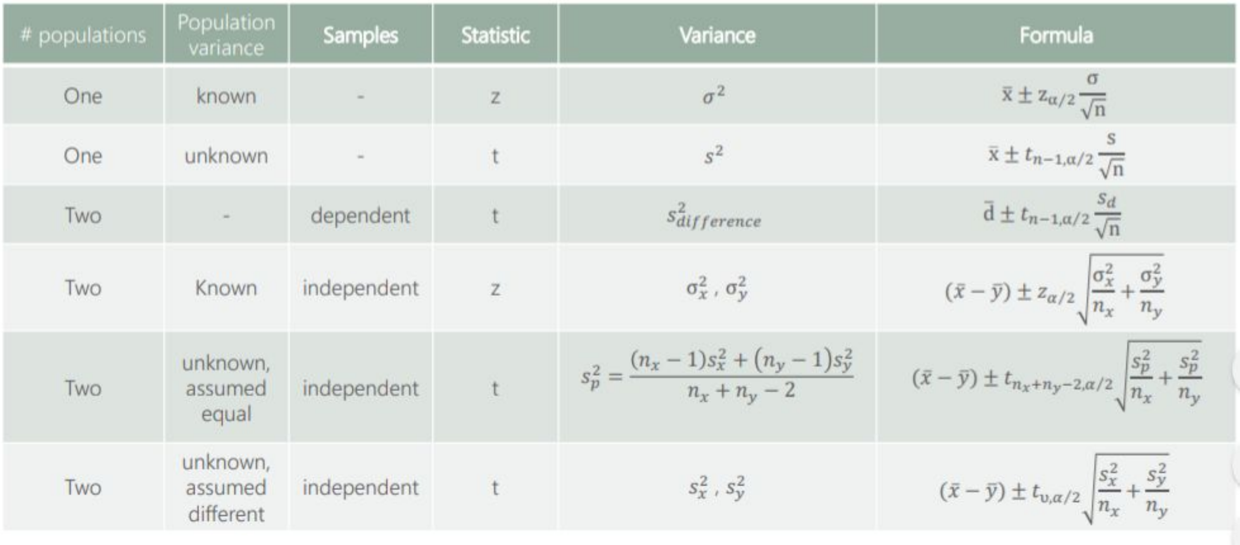
$$z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$t_{\upsilon,\alpha/2} * \frac{s}{\sqrt{n}}$$

1.
2. A confidence interval is an interval associated with a confidence level.
3. The most popular confidence levels are 99, 95 and 90% depending upon the case to case
4. Alpha(significance level) = 1 - confidence level
5. Alpha or the level of significance is the probability of rejecting the null hypothesis given that the null hypothesis is true. It is also called the Type 1 error
6. [Point estimate - critical value * standard error, point estimate + critical value * standard error]
7. (critical value * standard error) is also called the Margin of Error (ME). It is inversely proportional to the size of the sample
8. [Point estimate - ME, point estimate + ME]. Smaller ME means that the confidence interval will be narrower
9. The broader the confidence interval is, more confidence we get
10. To find the critical value, we either use Z statistic or T statistic. If the population variance is known then we use Z statistic otherwise T statistic. It also depends on the sample size, if the sample size is less than 30 we use T statistic otherwise Z statistic
11. To find a value in Z table, we find the value of (1 - alpha/2) in the table. The corresponding Z comes from the sum of the row and column headers associated with the cell

STUDENT T DISTRIBUTION

1. Visually it looks much like a normal distribution but with fatter tails
2. Degrees of freedom = Sample size - 1
3. In T table, the rows indicates the degrees of freedom (d.f) and the columns are the alpha values. The corresponding cell value is the critical value
4. After the 30th row, the numbers don't very much and the T statistic table becomes almost the same as Z statistic table



# populations	Population variance	Samples	Statistic	Variance	Formula
One	known	-	z	σ^2	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
One	unknown	-	t	s^2	$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$
Two	-	dependent	t	$s^2_{difference}$	$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$
Two	Known	independent	z	σ_x^2 , σ_y^2	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
Two	unknown, assumed different	independent	t	s_x^2 , s_y^2	$(\bar{x} - \bar{y}) \pm t_{\upsilon,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

CONFIDENCE INTERVAL FOR TWO MEANS WITH DEPENDENT SAMPLE (2 POPULATIONS)

1. When we are researching the same subject over time for eg - weight loss/blood samples, we are looking at the same person before and after the test
2. This test is mainly used when developing medicines. The patients are observed before and after taking the pill. Here we have two means and the dependent sample

CONFIDENCE INTERVAL FOR TWO MEANS WITH INDEPENDENT SAMPLE (2 POPULATIONS)

1. When we compare the grades of the students from two different departments (Engineering and Management) then we will two means and the samples are

independent

2. Variance of the difference of the two means is equal to the (variance of first population/sample size1 + variance of second population/sample size2)
3. [(mean1 - mean2) - critical value * variance of the difference, (mean1 - mean2) + critical value * variance of the difference]

HYPOTHESIS TESTING

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”

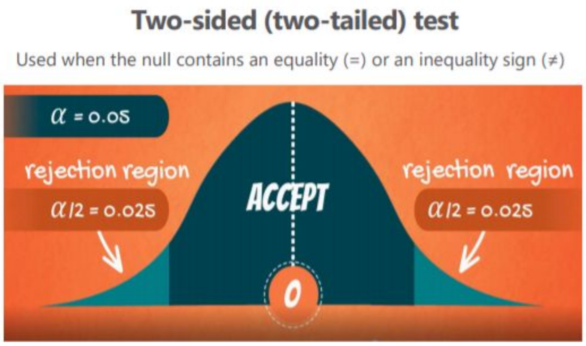
There are two hypotheses that are made

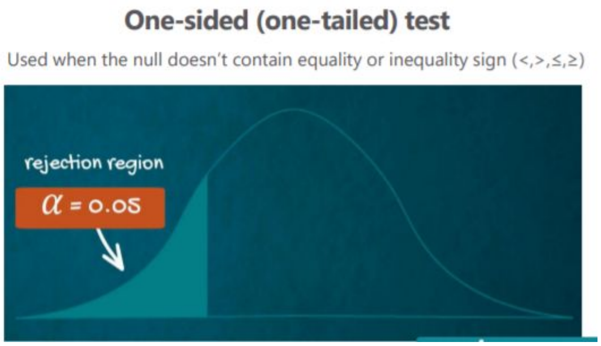
1. Null hypothesis - H0 - This is to be tested
2. Alternative hypothesis - H1 - This is everything else than H0

Generally we formulate the hypothesis in such a way that we try to reject the null hypothesis and the alternative hypothesis is the one which challenges the null hypothesis. In a more formal way, The null hypothesis is the present state of affairs while the alternative hypothesis is our personal opinion

Significance level is the probability of rejecting the null hypothesis, if it is true

We always set the value of significance level and then test our hypothesis against it. 0.05 is the most common value but it varies from case to case


Two-sided (two-tailed) test
Used when the null contains an equality (=) or an inequality sign (≠)
α = 0.05
rejection region
α/2 = 0.025
ACCEPT
rejection region
α/2 = 0.025
0


One-sided (one-tailed) test
Used when the null doesn't contain equality or inequality sign (<,>,≤,≥)
rejection region
α = 0.05

<u>Example of forming a hypothesis</u>

Assume that you are carrying out an analysis on how the students are performing on average. The university dean told you that the mean population grade is 70%. Being a data scientist, you decided to test it

H0 : Population mean grade = 70%

H1 : Population mean grade != 70%

Perform the Z test and is given as

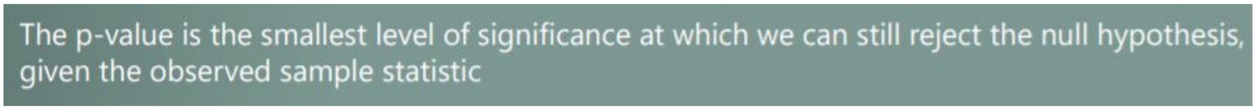
1. Z = (sample mean - hypothesized mean) / (SD of sample/sqrt(n))
2. If sample mean to close to the H0 mean then the Z will be close to 0
3. Find the Z-critical value from Z table and compare it with the Z score
4. If Z-calculated > Z-critical, we reject the null hypothesis otherwise not

TYPE 1 VS TYPE 2 ERROR

1. Type 1 error is when you reject the true null hypothesis and also called the False Positives and its probability is defined as alpha (the level of significance)
2. Type 2 error is when you accept the null hypothesis when it is actually false and also called the False Negative.
3.

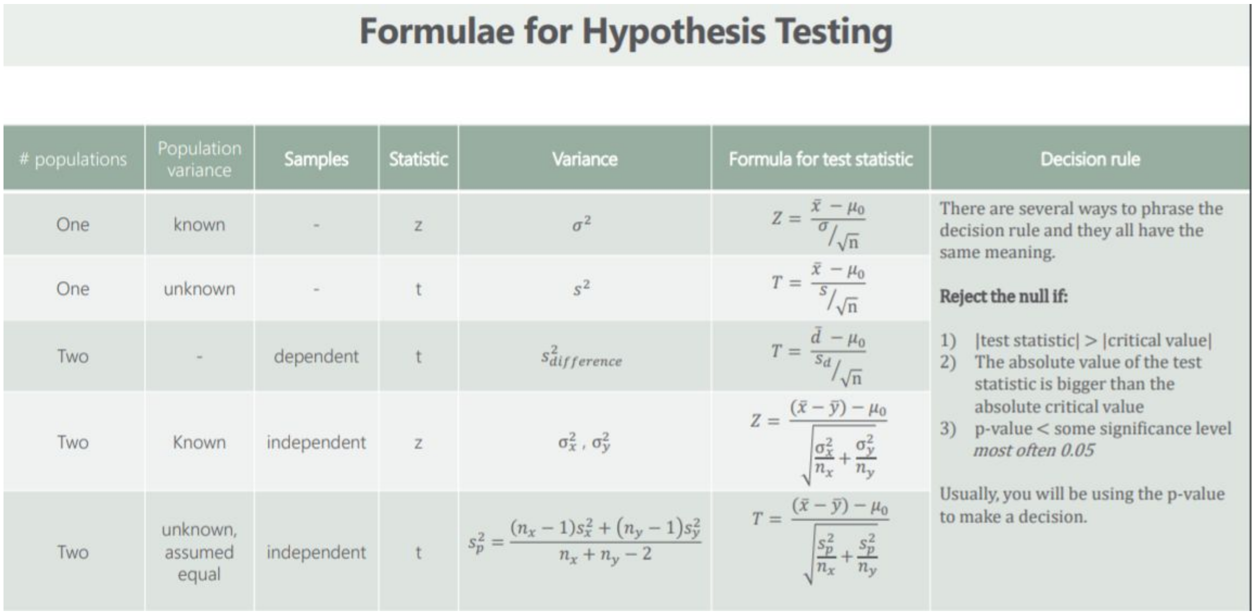
	H0 is true	H0 is false
Accept	Correct	Type 2 error
Reject	Type 1 error(False positive)	Correct (False negative)

P VALUE



The p-value is the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic

1. It is the probability of finding the observed or more extreme value when the null hypothesis is true. It is a number between 0 and 1
2. If p <= 0.5 , then this is a strong evidence to reject the null hypothesis
3. If p> 0.5, then this is a week evidence against the null hypothesis
4. The lesser the value of P, the strong is the evidence for rejecting the null hypothesis



Formulae for Hypothesis Testing

# populations	Population variance	Samples	Statistic	Variance	Formula for test statistic	Decision rule
One	known	-	z	σ^2	$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	There are several ways to phrase the decision rule and they all have the same meaning. **Reject the null if:** 1) |test statistic| > |critical value| 2) The absolute value of the test statistic is bigger than the absolute critical value 3) p-value < some significance level *most often 0.05* Usually, you will be using the p-value to make a decision.
One	unknown	-	t	s^2	$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	
Two	-	dependent	t	$s^2_{difference}$	$T = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}}$	
Two	Known	independent	z	σ_x^2 , σ_y^2	$Z = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$	
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$T = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$	

REGRESSION ANALYSIS

1. The most common method of prediction. It is used whenever we have a causal relationship between the variables. For example, the amount of money you spend depends on the amount of money you make
2. Before performing the regression analysis, first check the correlation between the two variables either by a scatter or most preferably the coefficient of correlation
3. Always remember that the correlation does not imply causation

LINEAR REGRESSION

1. Linear regression is a linear approximation of a causal relationship between two or more variables
2. Get the data => Design a model that works for that sample => Make predictions for the whole population
3. The simplest linear regression is the simple linear regression model
   y = B0 + B1x1 + e
4. The independent variable x has a causal relationship with the dependent variable y. For example, the income you receive is related to the years of education but its not true vice versa
5. The linear regression analysis is known for the best fitting line that goes through the data points and minimizes the distance between them. On the other hand, the correlation is a single point which only tells the direction where the two variables are moving
6. Whenever you find the data that looks regressable, don't dive straight into regression analysis. Always look for the causality
7. The distance between the observed value and the regression line is called the residual

SST, SSR, SSE(RSS) > SST = SSR + SSE

1. SST or sum of squares total is the squared distance between the observed dependent variable and its mean. Think of this as the dispersion of the observed variables around the mean and it is a measure of the total variability of the dataset
2. SSR or sum of squares regression is the sum of the squared distance between the predicted value and its mean. Think of it as a measure of how well your line fits the data
3. SSE or RSS or sum of squares errors is the sum of the squared distance between the observed and the predicted values. We always try to minimize it to get the least errors in our model

R-SQUARE

1. R^2 = SSR/SST
2. It is equal to the variability explained by the regression divided by the total variability.
3. It takes the values from 0 to 1
4. The larger value of R^2 means the regression model is a better fit for the data
5. If R^2 = 1, means that the model explains the entire variability of the data
6. R^2 = 0 means the model failed to explain any variability of the data

OLS

1. OLS or Ordinary Least Squares is the most common method to estimate the linear regression equations
2. Least square stands for the minimum squares error (SSE) so this method aims to find the line which minimizes the sum of the squared errors
3. There may be many lines possible to do the regression on data but the OLS finds the one with the smallest error
4. In case of multiple regression, we increase the explanatory power by 0 or more. We cannot lower it

ADJUSTED R-SQUARE

1. Adjusted R^2 < R^2
2. It penalizes the excessive use of the variables so it is used to find the optimal number of variables required

F STATISTIC

1. It is similar as the Z statistic or T statistic and is used for testing the overall significance of the model
2. The lower the F statistic, the closer to a non-significant model

OLS ASSUMPTIONS

1. Linearity
2. No endogeneity
3. Normality and homoscedasticity
4. No autocorrelation
5. No multicolliearity