Statistics

2020

By Ayub Odhiambo
http://www.a4ayub.me/

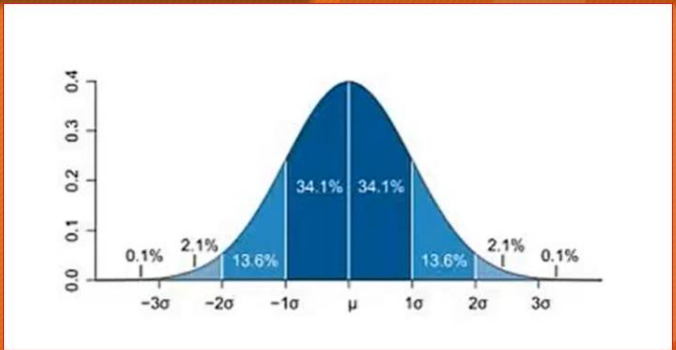
Objectives

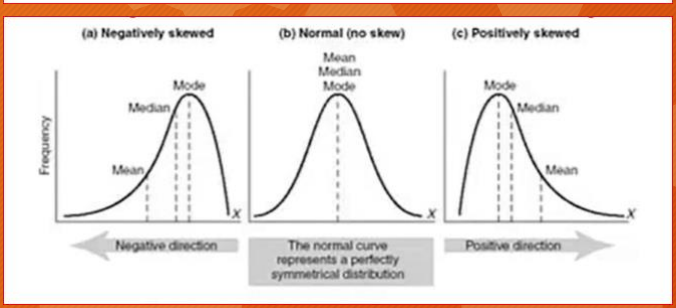
- Review of previous class on learnings
- Intro to Statistics
- What is hypothesis Testing
- Null Hypothesis Vs Alternative Hypothesis
- Writing hypothesis tests
- Important statistical definitions
- The process of hypothesis testing
- Statistical tests
- P-Value and it’s use in Hypothesis Testing
- Hands-On Lab on Jupyter
  - Calculation of Test Statistic
  - Visualization of supermarket data using Matplotlib

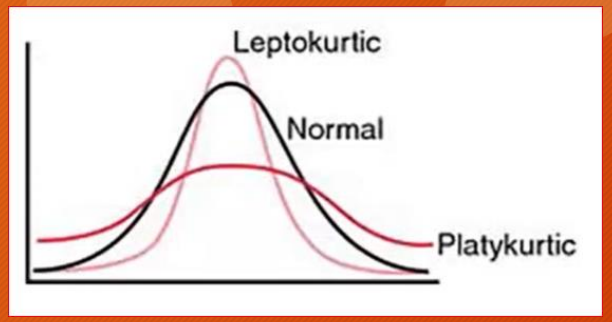
Introduction to Statistics

- Process of collecting information/data and presenting this in different ways and drawing conclusions in some ways.
- This is also the actual **numeric descriptions** of sample data e.g. mean, standard deviations e.t.c
- Branches of Statistics:
  1. Descriptive Statistics – Gather, Sort and Summarize data using numbers and graphs (Bar graphs, Histograms, Pie Charts, Shape of graph and Skewness). Measures of Central Tendency (Mean, Media, Mode). Measures of Variability (Range, Variance, Standard Deviation)
  2. Inferential Statistics – Run different tests and use of probability to determine how confident we can be that the conclusion we make are accurate (Confident Intervals & Margin of errors) e.t.c

Descriptive Statistics

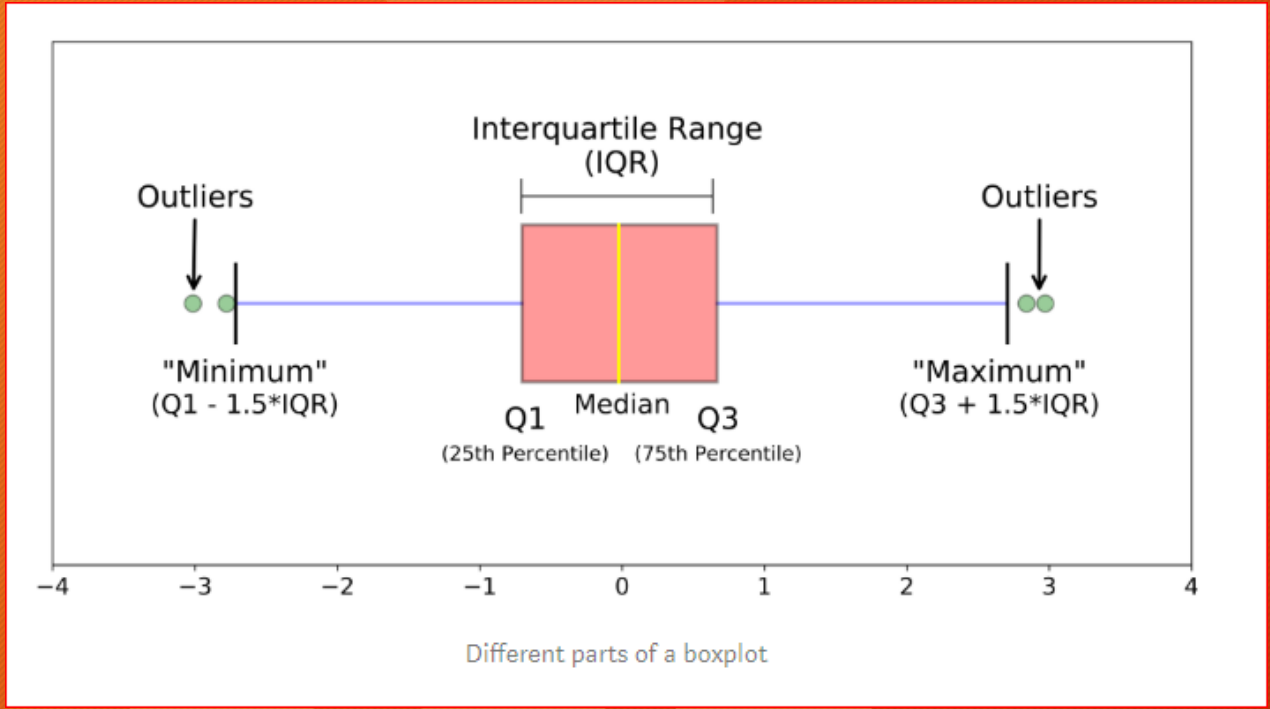






- Distributions – This is a list of scores taken on a particular variable e.g. student scores, age e.t.c
  - Single Mode
  - Bi Modal
  - Multi modal
- Mean – This is the average
- Median – This is the mid-point (50th Percentile)
- Mode – Most occurring
- Standard Deviation – How much variation exists from the average/mean. Best used when data is single mode.
- Variance – This is the squared standard deviation
- Skewness – How symmetric is the distribution of the variables
- Kurtosis – provides a visual estimate of variance in a sample. A kurtosis > 2 is leptokurtic, has thicker tails with majority of the values concentrated at the center meaning higher probability for extreme values and little variance. IN a platykurtic distribution the probability of extreme values is low. 0 kurtosis is mesokurtic
- Inter-percentile Measures – describes a score relative to others in a distribution
- Range – Difference between the largest and smallest

IQR : Explanation using Box Plots

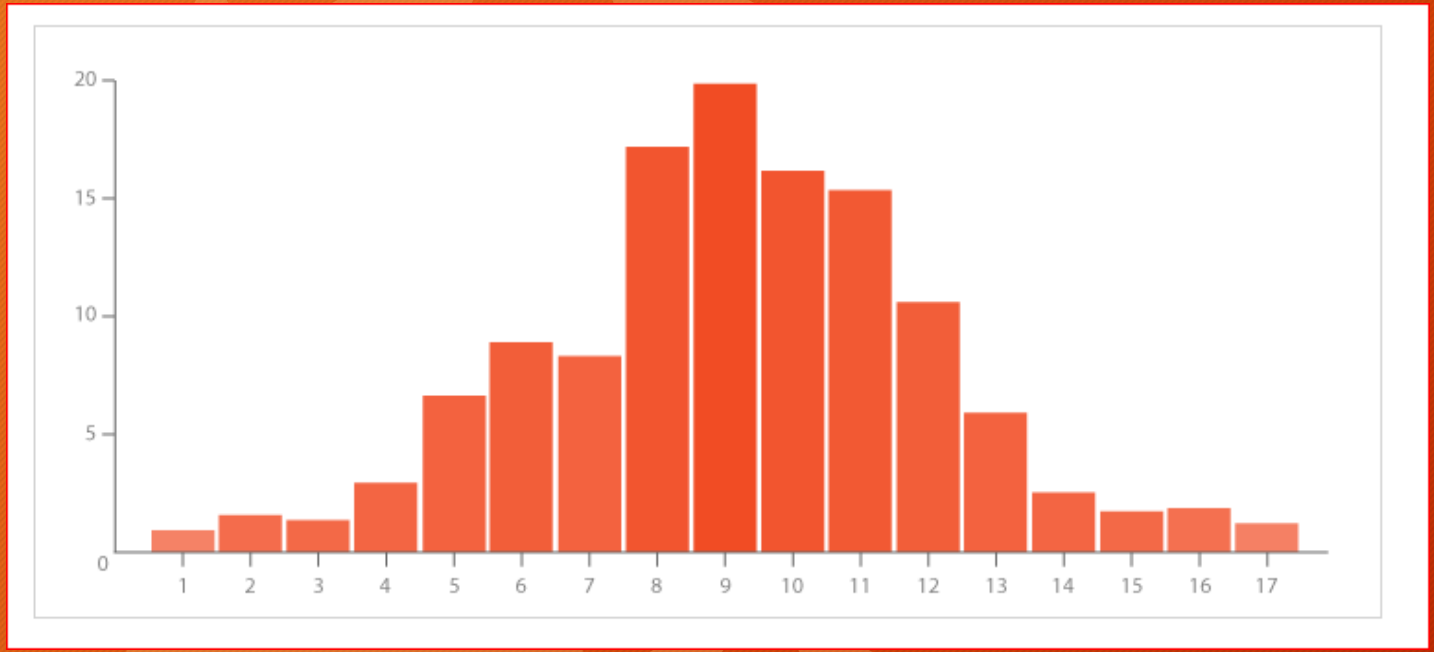


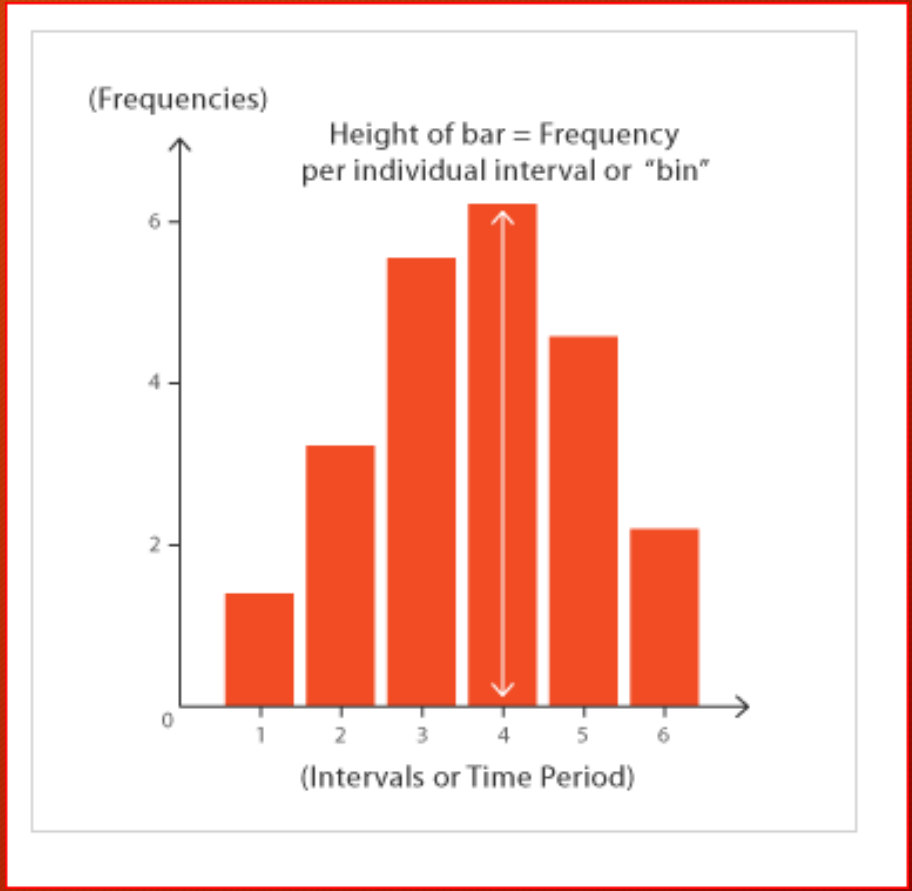
Different parts of a boxplot

Shows the distribution of your data-set

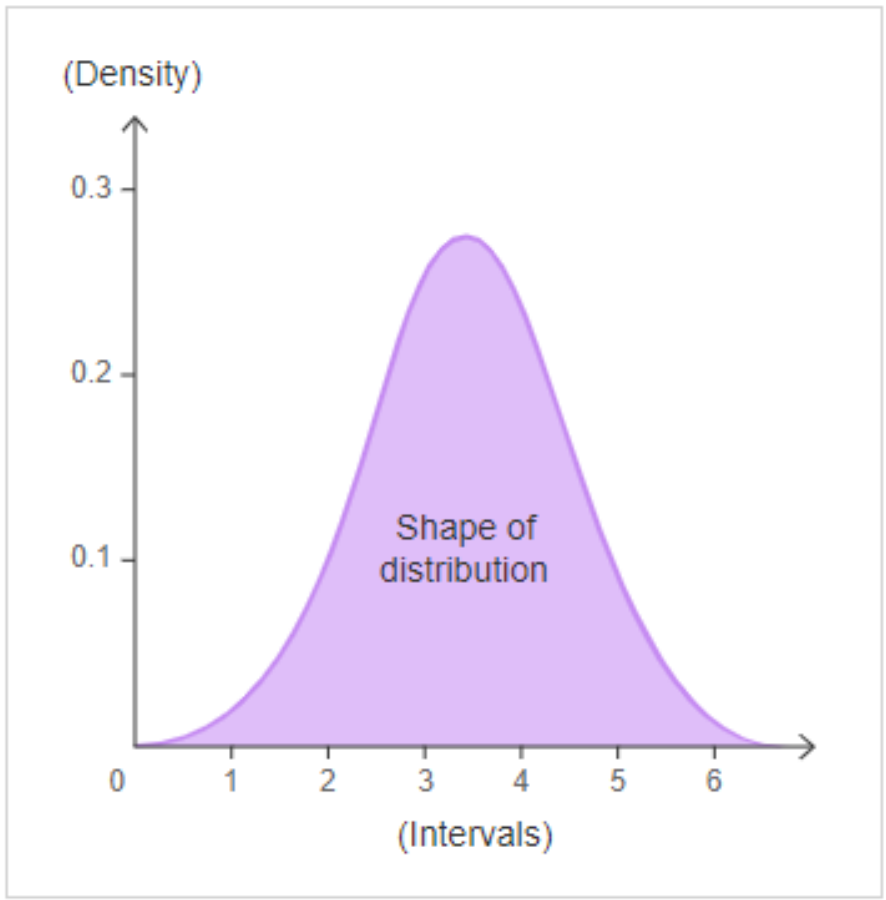
- **Outlier Detection** - Tells you about your outliers and their values
- **Skewness** - Can tell if your data is symmetrical i.e. by what degree does the distribution fall away from its mean
- **Variability** - It explains the variations of the data amongst the data-set or around its average
- **Types :** Outliers can be Suspected/Mild or Stronger by deriving inner fences and outer fences
- Others :
  - Histogram
  - Density Plots

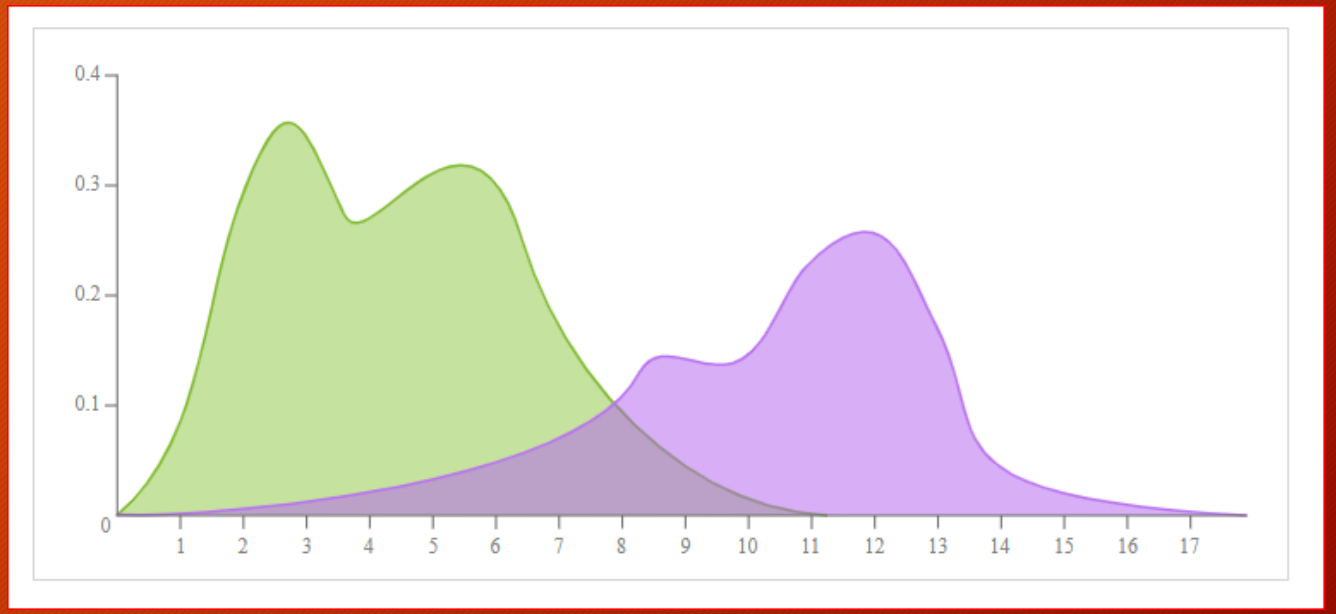
Histogram : Frequency of Data in Bins



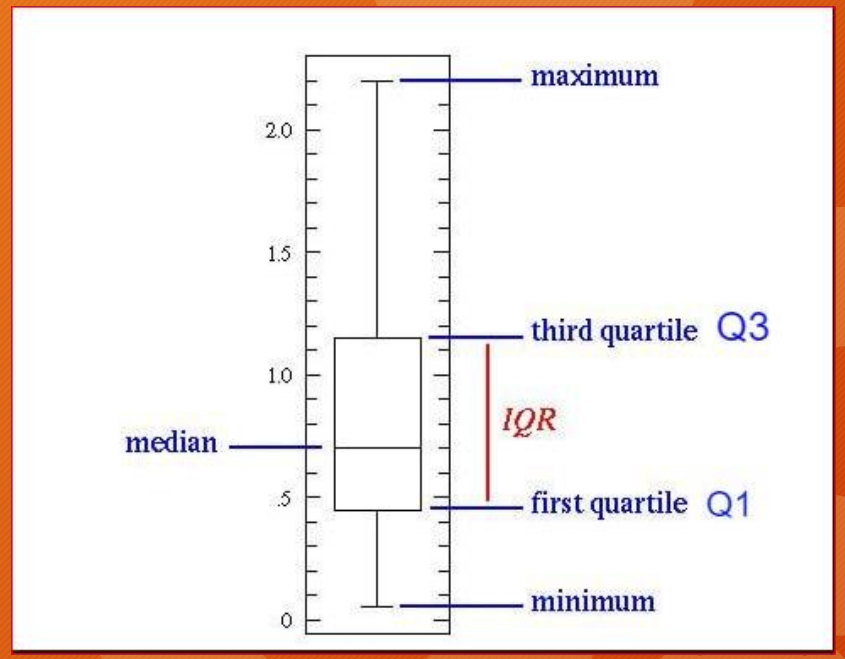


Density Plots : Removal of Histogram Noise





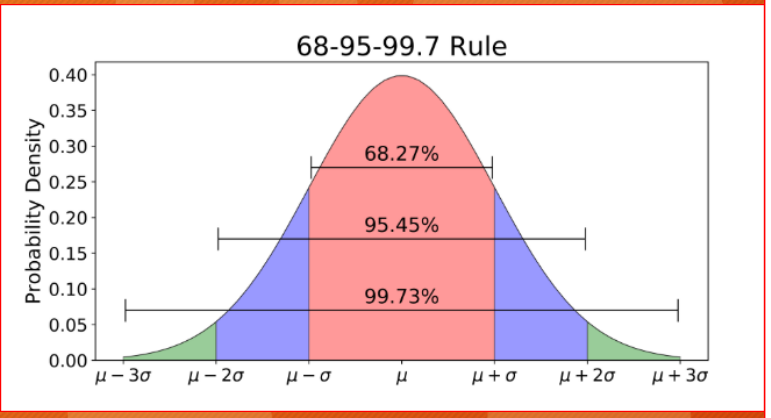
IQR : The Rules

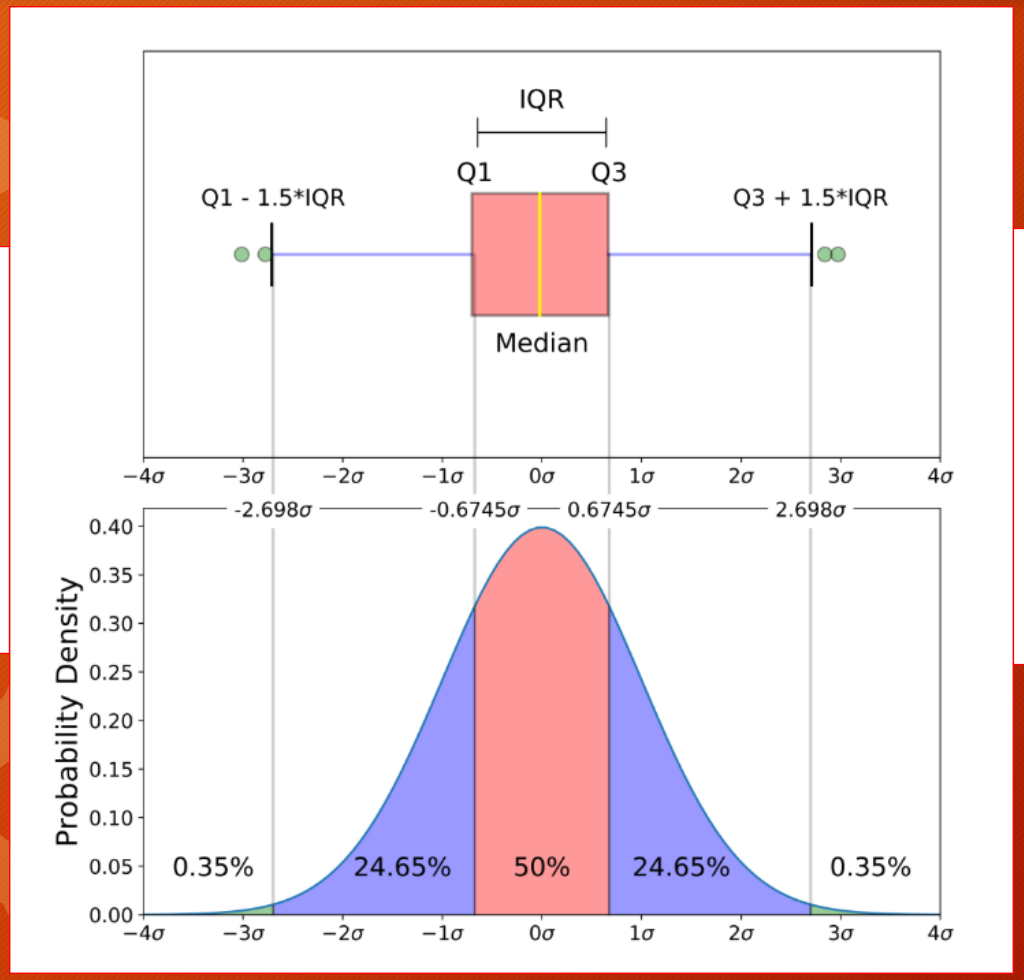


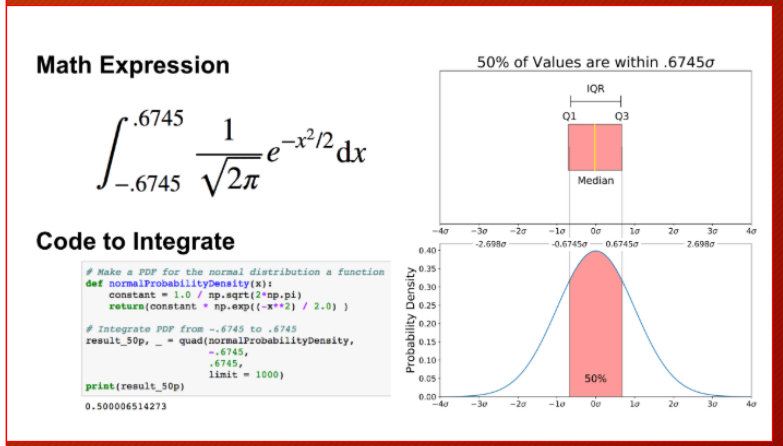
Five Number Summary

1. Calculate the Inter-Quartile Range for the data-set
2. Multiply the inter-quartile range for the data with the number 1.5 (Inner Fence) or by 3 (Outer Fence)
3. Add 1.5 * IQR to the third Quartile. Any number greater than this is a suspected/mild outlier
4. Subtract 1.5 * IQR to the first Quartile. Any number less than this is a suspected/mild outlier
5. Add 3 * IQR to the third Quartile. Any number greater than this is a strong outlier.
6. Subtract 3 * IQR to the first Quartile. Any number less than this is a strong outlier

Boxplot Vs Probability Density Function







Outlier Detection Class Illustration

Problem Statement : Is 17 an Outlier?

Data-Set :[1,3,4,6,7,7,8,8,10,12,17]

Five Number Summary:

- Minimum : 1
- Q1 : 4
- Median : 7
- Q3 : 10
- Maximum : 17

Solution :

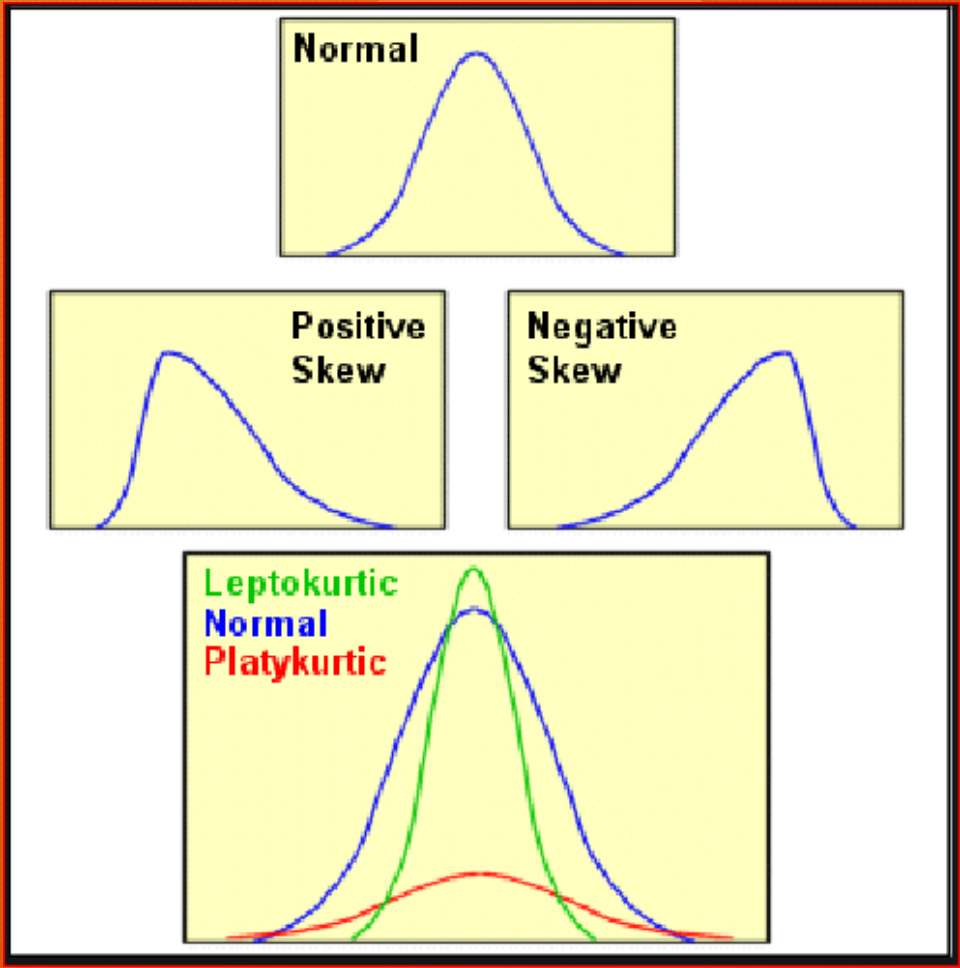
IQR : 6

Inner Fence Values [-5 , 19]

Outer Fence Values [-14 , 28]

Answer : 17 is neither a suspect outlier or a strong outlier.

Skewness Vs Kurtosis (Measures of Shape)



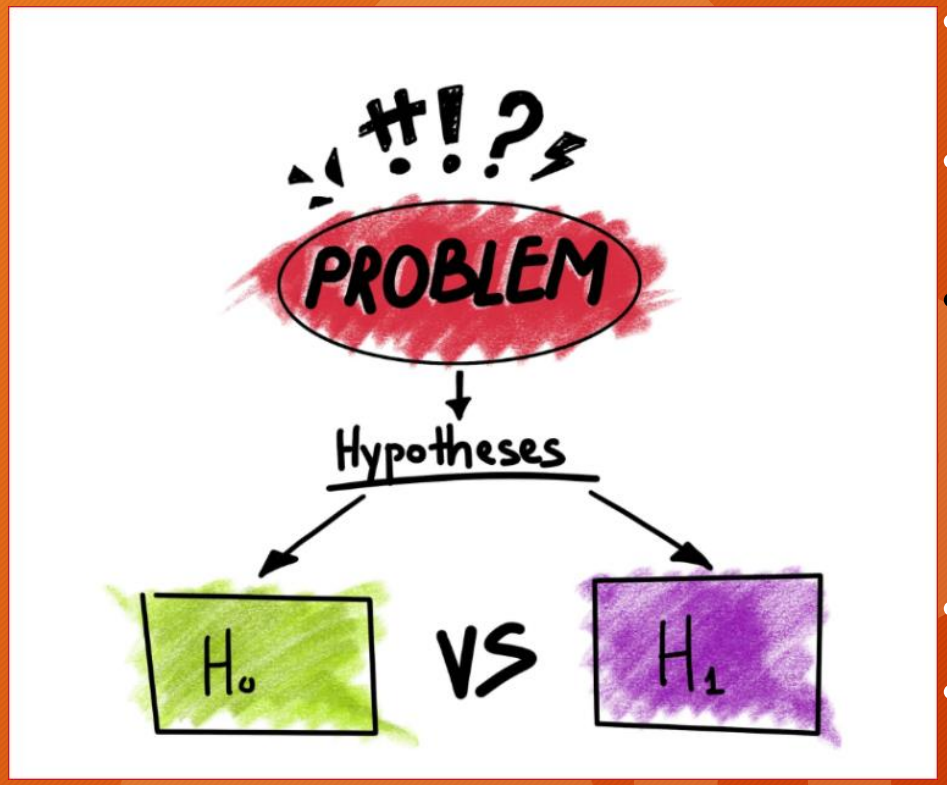
Skewness – This is a measure of symmetry in a distribution. A symmetrical distribution will have a skewness of zero. This is relatively a measure of the relative size of the two tails

- If skewness is < 0 this is a negative skewness and the tail is longer with the hump tending towards the right
- If the skewness is > 0 is a positive skewness and the tail is longer with the hump towards the left

Kurtosis – This is a measure of the combined weights of the tails relative to the rest of the distribution.

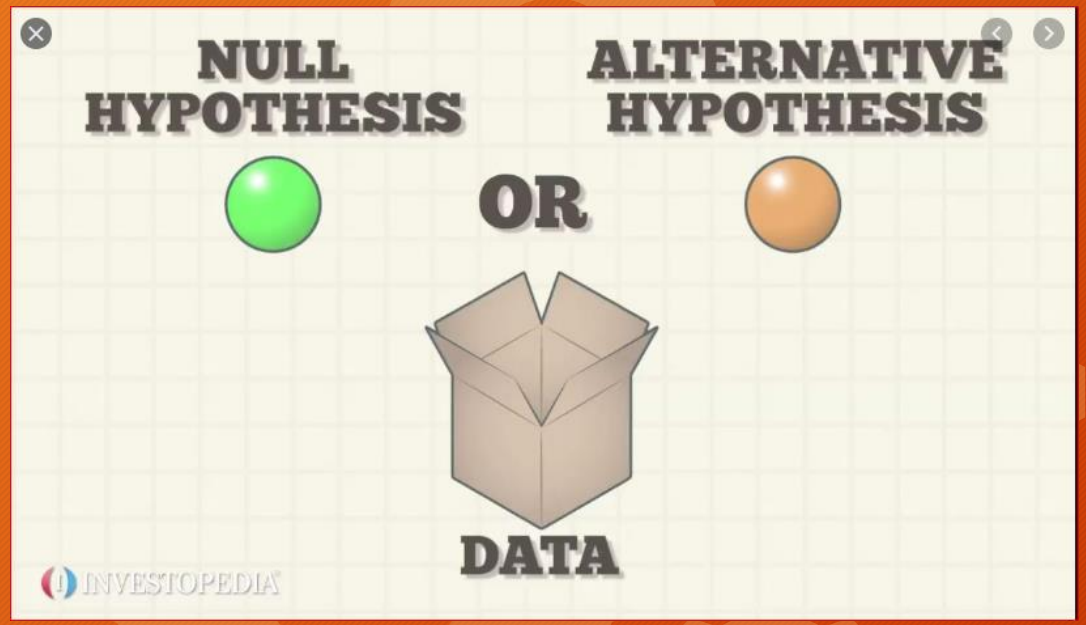
- If Kurtosis is > 0 it means the tail is thinner and it has positive kurtosis (leptokurtic). Large outliers
- If Kurtosis is < 0 it means the tail is thinner and it has negative kurtosis (platykurtic)
- If kurtosis is = 0 this is mesokurtic and it is same as a normal distribution
- Excess Kurtosis = Kurtosis - 3

Inferential Statistics



- A fact is a simple statement that everyone believes. It is innocent, unless found guilty.
- A hypothesis is a novice suggestion that no one wants to believe. It is guilty until found effective.
- **For a question to become a hypothesis, it must be provable** e.g. you can prove that changing the title in an add will increase conversion by 20% but you won't be able to check the question, will changing the title help increase conversion.
- The hypothesis should be specific and not vague.
- The process of hypothesis testing consists of forming questions about the data based on the gathered information and testing them using statistical methods

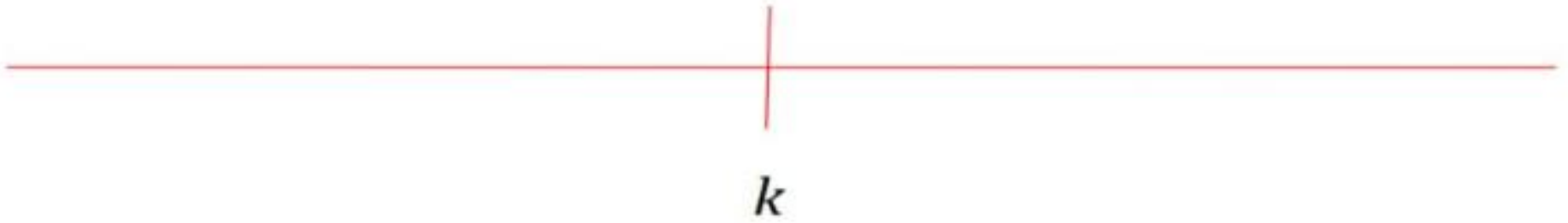
Null Hypothesis Vs Alternate Hypothesis



- **Null Hypothesis** : This is the claim about a population, it is the currently accepted value of a parameter. For our case this would be assumed to be **women spend double the time when shopping as compared to men.**
- **Alternative Hypothesis** : This is an assumption under test that is not always strictly opposite to the null hypothesis. Also called research hyposis. In our case it would be **women do not spend double the time shopping when compared to their male counterparts**

Writing Hypothesis Tests

Two-tailed test	Right-tailed test	Left-tailed test
$H_0: \mu = k$ $H_1: \mu \neq k$	$H_0: \mu = k$ $H_1: \mu > k$	$H_0: \mu = k$ $H_1: \mu < k$



State the H_0 and H_1

The average basket value of a daily shopper is KShs 2,000/=

State the H0 and H1

61% of Supermarkets customers buy less than 3 items

https://bankelele.co.ke/2017/11/nairobi-supermarket-trends.html

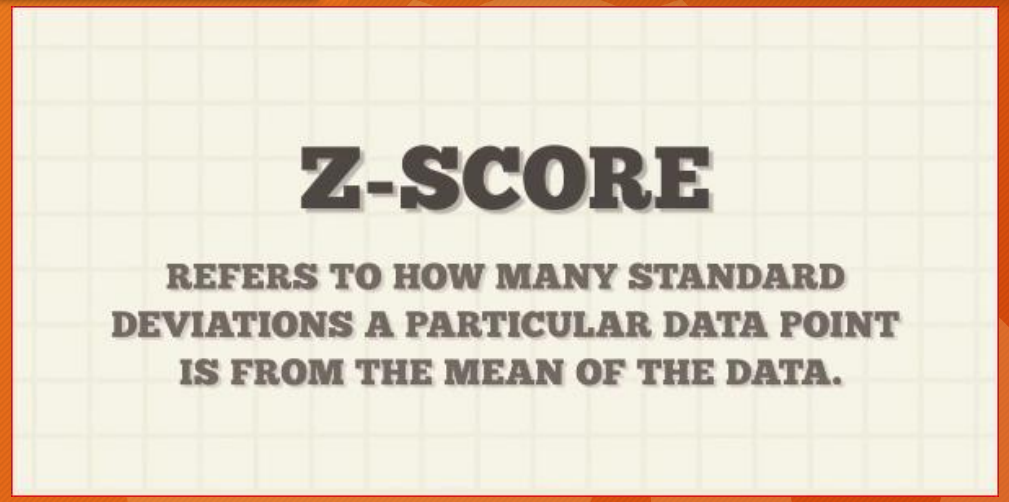
Hypothesis Testing : Possible Outcomes

- **Reject** the Null Hypothesis.
- **Fail to Reject** the Null Hypothesis

Important Statistical Definitions

Terms	Definition
Test Statistic	Uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected
Statistically Significant	Where do we draw the line to help us decide whether to reject the null hypothesis or not. Some conclusions are subjective, statistics provides a concrete way to decide when to reject the hypothesis and when to fail to reject it through an hypothesis test. An hypothesis test collects the data, puts it in an equation, you get a number back and that number is used as boundaries that helps decide when to reject or fail to reject a null hypothesis
Confidence Level	e.g. 95% or 99% i.e. How confident are we in our decision
Significance Level	Also called alpha and it is normally 1 – Confidence Level
Z Score	This is a single value representing sample data
P-Value	This is the probability of obtaining a sample “more extreme” than the ones observed in your data, assuming the Null hypothesis is true. Area under the curve beyond the z score

Z-Score : What it is and Calculating it



Steps:

1. **Get the variance (σ^2)** - This is a measure of the spread between numbers in a dataset. i.e. It measures how far each number in the set is from the mean and therefore from every other number in the set.
2. **Get the Standard Deviation (σ)** - This is the measure of dispersion/variability of a dataset relative to its mean and is calculated as the square-root of the mean (Population or Sample)
3. **Calculate the Z-Score** - This is the number of standard deviations a particular data point is from the mean.

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

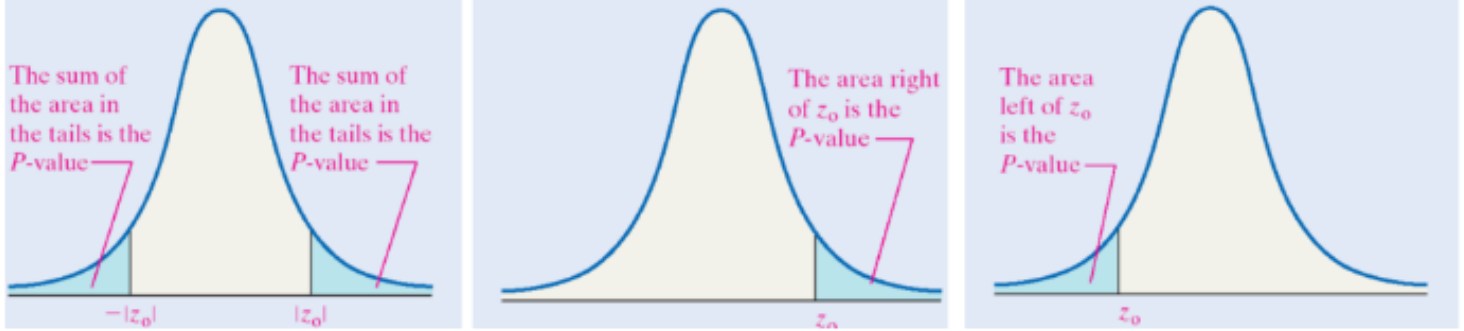
μ = Mean

σ = Standard Deviation

P-Value : Types of Tests

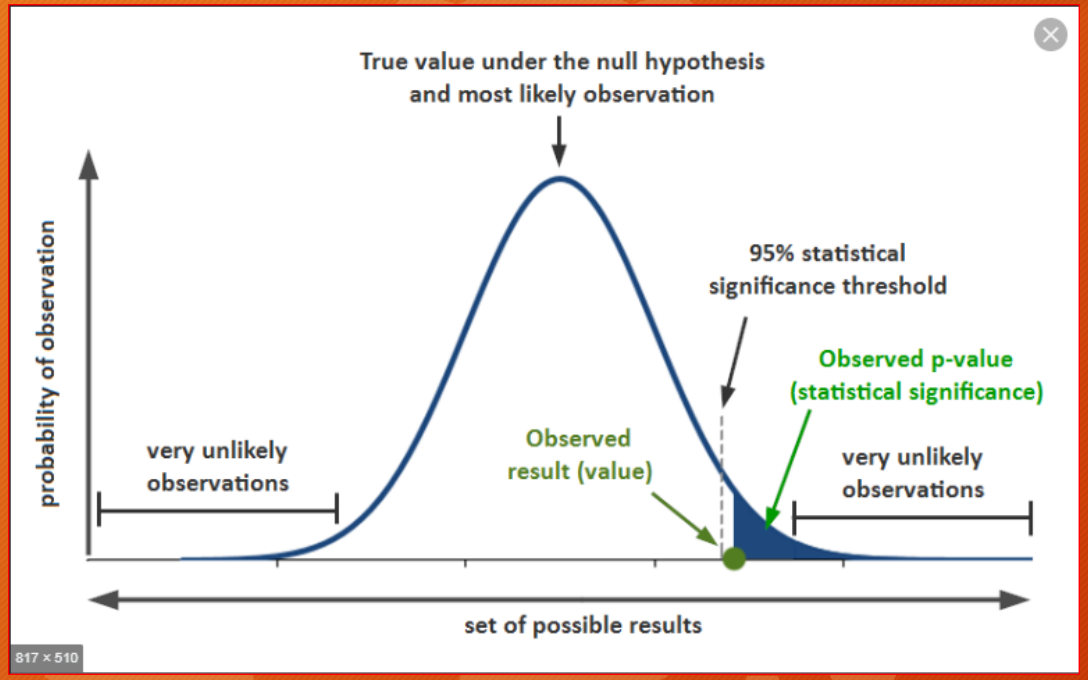
Step 1: Compute the test statistic $z_0 = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

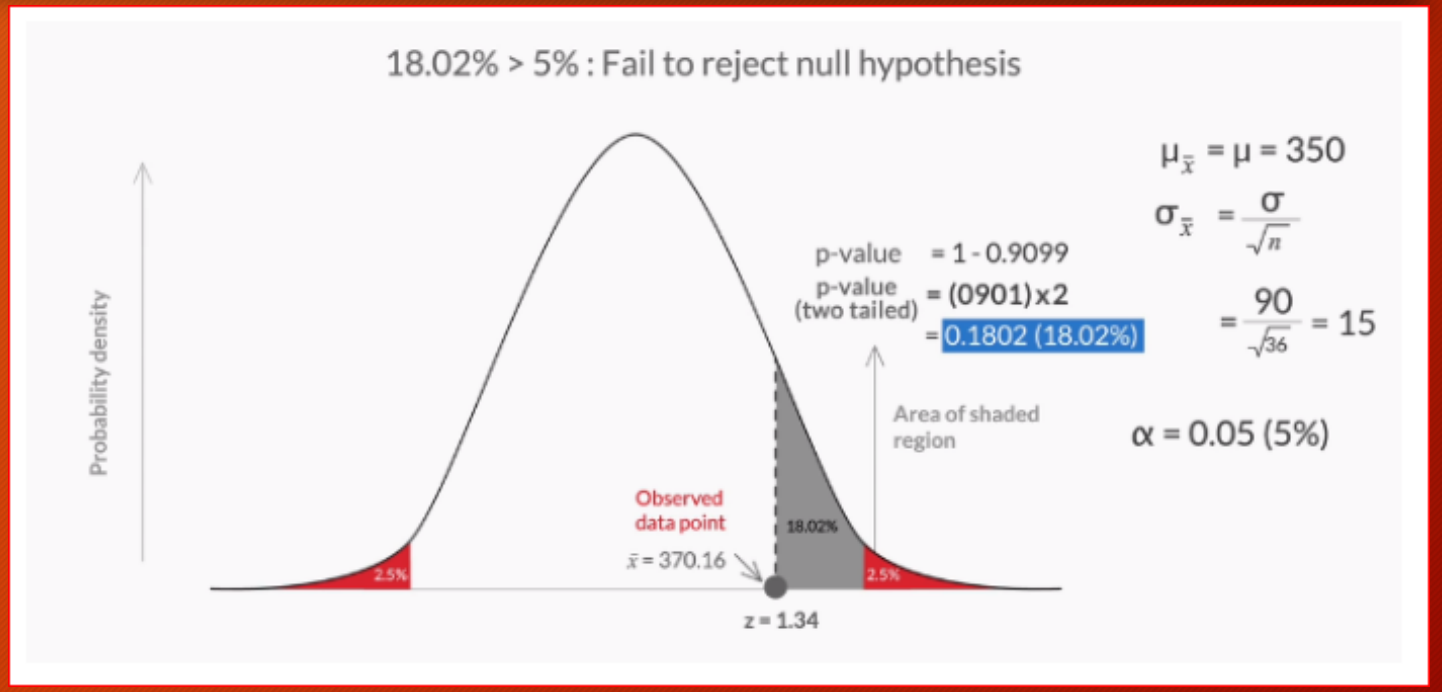
Two-tail	Right Tail	Left Tail
Two-Tailed	**Right-Tailed**	**Left-Tailed**
$P\text{-value} = P(Z < -\lvert z_0\rvert \text{ or } Z > \lvert z_0\rvert)$ $= 2P(Z > \lvert z_0\rvert)$	$P\text{-value} = P(Z > z_0)$	$P\text{-value} = P(Z < z_0)$



The shaded area are known as the “rejection regions” and these corresponds to the z-score values (Area = 0.05 or 5% for a single tailed test and Area = 0.025 or 2.5% for a two tailed test). So any value in these regions means the Null Hypothesis needs to be rejected.

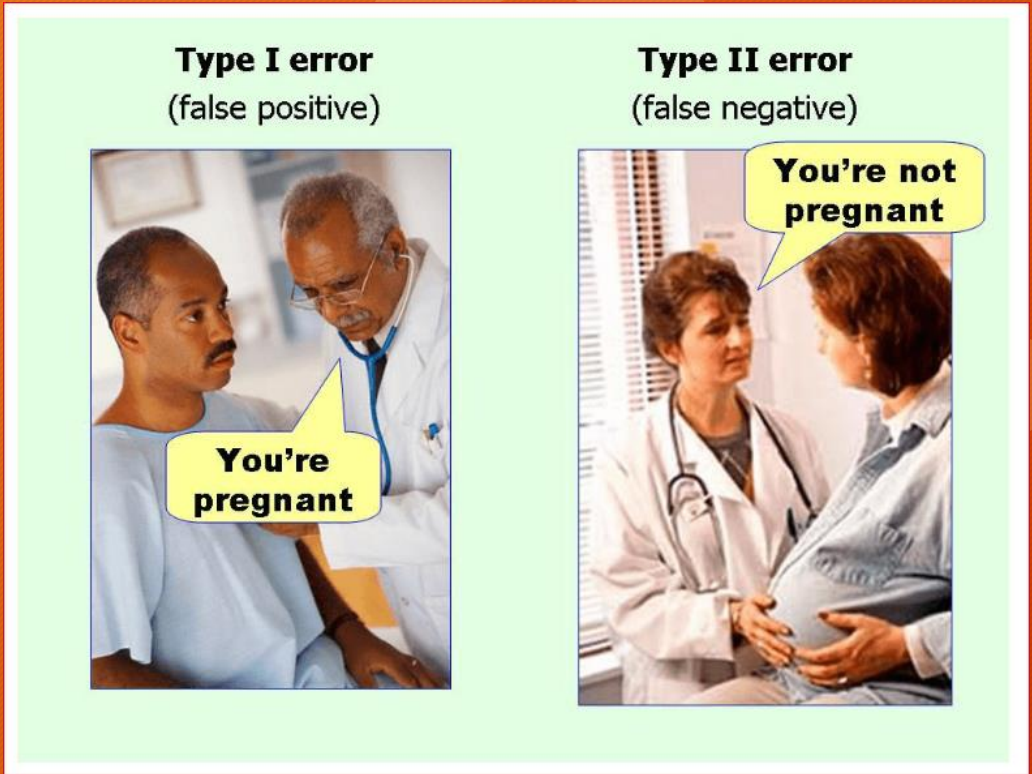
P-Value : Conclusions of p-test





If the P value is less than the level of significance then you reject the null hypothesis

Statistical Errors



While testing statistical hypotheses, it is possible to make a mistake by accepting or refusing the correct hypothesis.

The level of significance is the probability of making a Type I error.

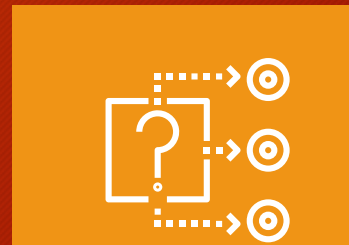
- **Type I Error**: The null hypothesis is true but it is rejected. The null hypothesis in this case is that all men can never be pregnant.

- **Type II Error**: Not rejecting the null hypothesis when it is false. The null hypothesis in this case is all pregnant women have protruding stomachs

Assignment : Week #1

Assignment	Description
Descriptive Statistics	1. Identify the measures of central tendencies for your target variable 2. Plot a pie chart showing the proportions of your target variable 3. Plot a histogram showing the distribution of your target variable 4. Describe the skewness and kurtosis of your data
Inferential Statistics	Perform and Inferential experiment on your project by: 1. Stating your null hypothesis 2. State the alternate hypothesis 3. Perform the T Statistics 4. Calculate the Z-Score 5. Identify the p-value 6. Conclude on whether to reject or fail to reject the null hypothesis

Statistical Jupyter Notebooks



Hands-On practical sessions on Key Concepts.