

Learn Something Every Day



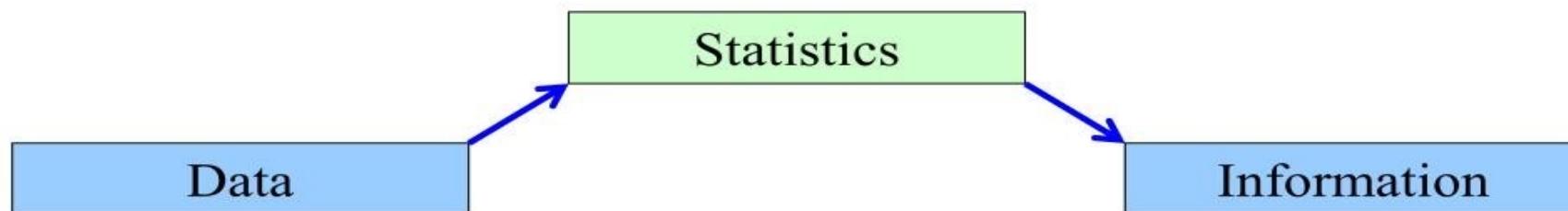
Basic of Statistics

Statistics



What is Statistics?

“Statistics is a way to get information from data”



Data: Facts, especially numerical facts, collected together for reference or information.

Information: Knowledge communicated concerning some particular fact.

Statistics is a **tool** for creating ***new understanding*** from a set of numbers.

Basics of Statistics

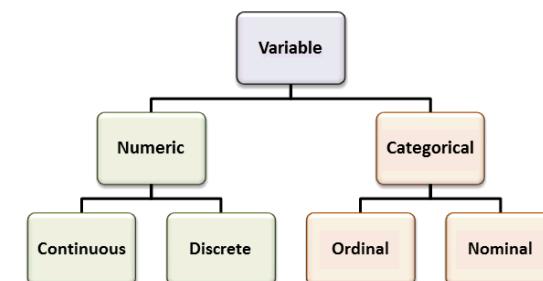
Population



Samples



Variables

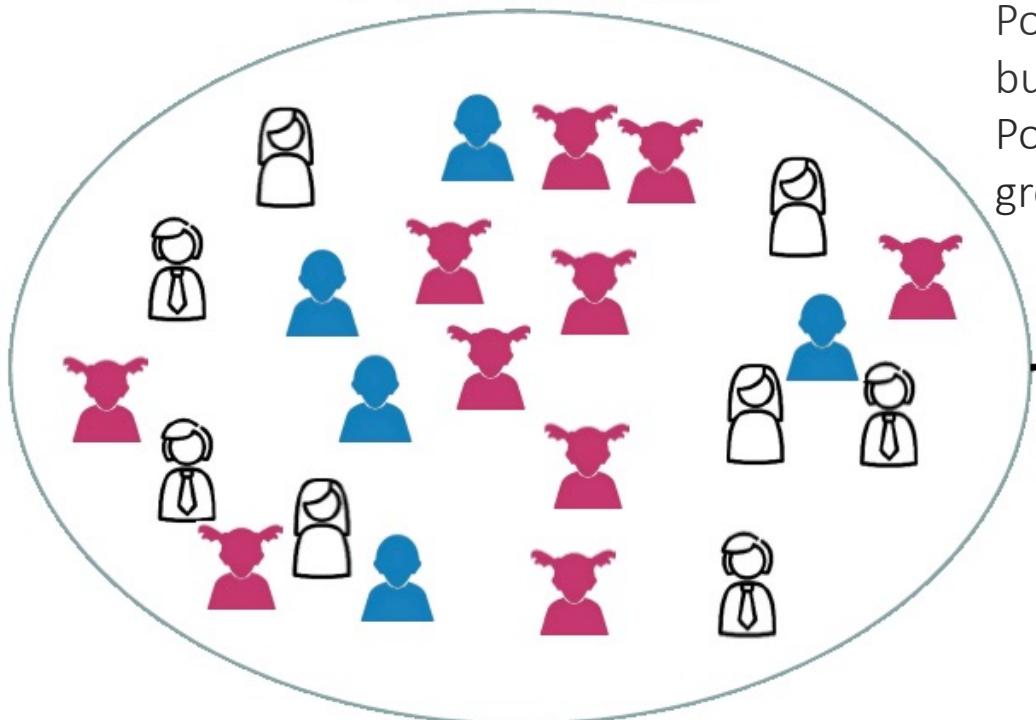


Statistical Model

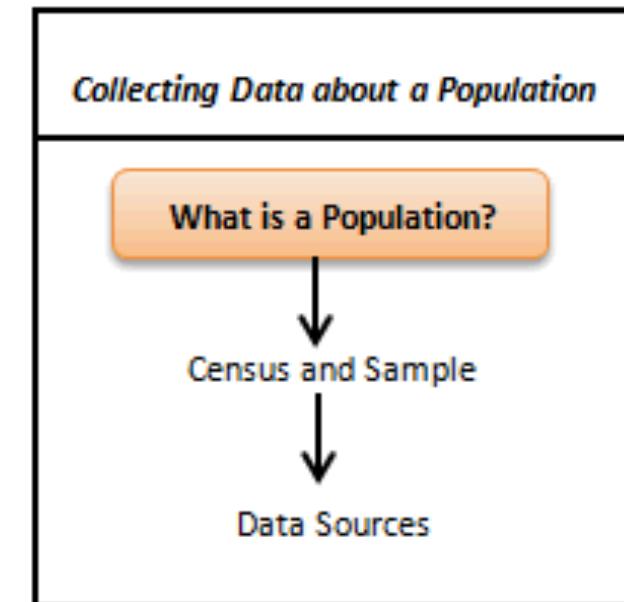


Population

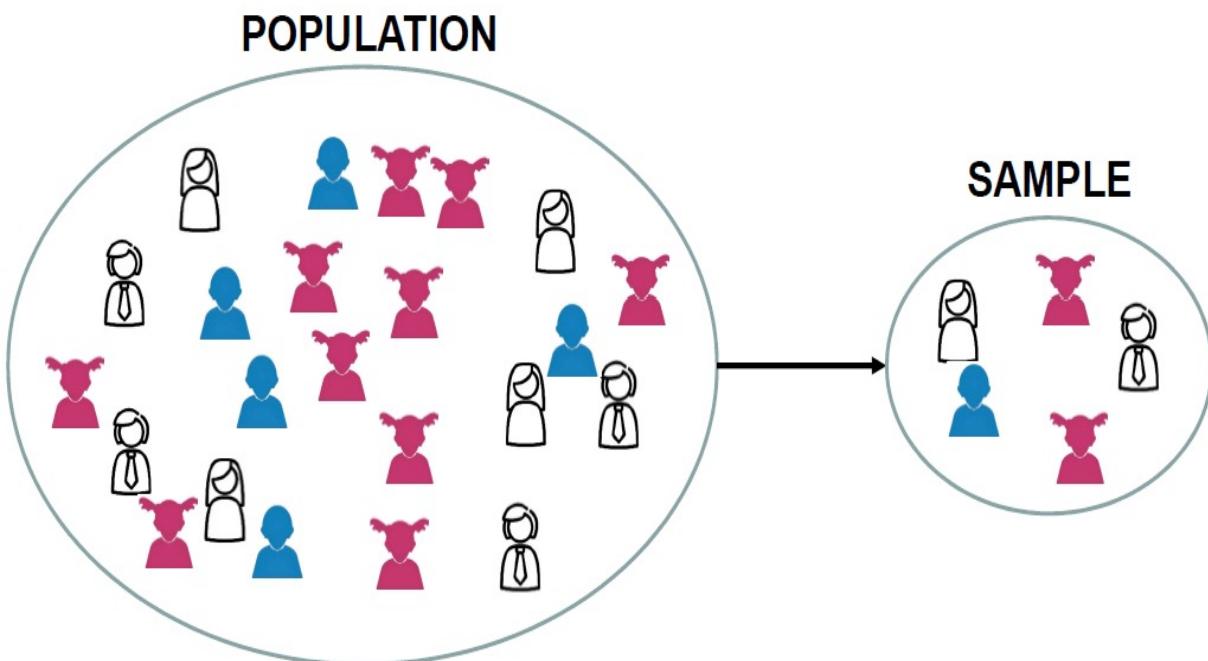
POPULATION



A population is any complete group with at least one characteristic in common. Populations are not just people. Populations may consist of, but are not limited to, people, animals, businesses, buildings, motor vehicles, farms, objects or events. Population contains all the data points from a set of data. It is a group from where we collect the data.



Sample



A sample consists of some observations selected from the population. The sample from the population should be selected such that it has all the characteristics that a population has. Population's measurable characteristics such as mean, standard deviation etc. are called as parameters while Sample's measurable characteristic is known as a statistic.

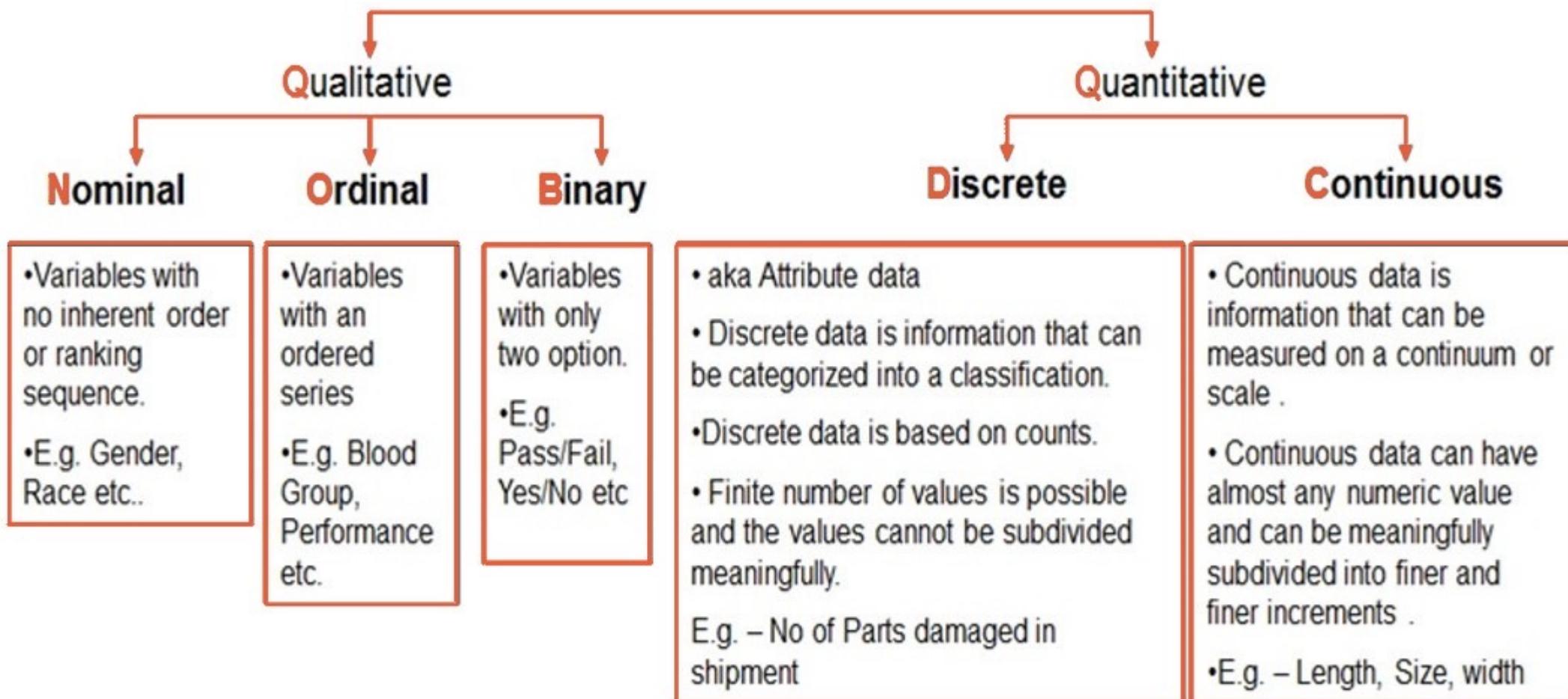
**Cluster
Sampling**

**Stratified
Sampling**

**Random
Sampling**

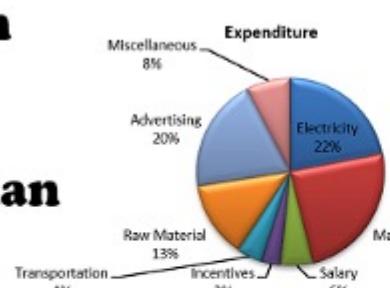
**Systematical
Sampling**

Variables in statistics

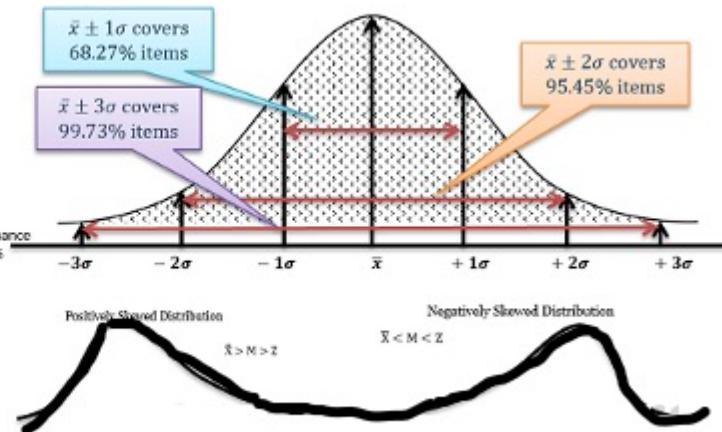


Statistical Model

Mean



Median



Mode

$$Std. Dev. \sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

Statistical modelling is a method of mathematically approximating the world. Statistical models contain variables that can be used to explain relationships between other variables. We use hypothesis testing, confidence intervals etc to make inferences and validate our hypothesis.

A statistical model will have sampling, probability spaces, assumptions and diagnostics etc, to make inferences.

We use statistical models to find insights given a particular set of data. We can conduct modelling on a relatively small set of data just to try and understand the underlying nature of the data.

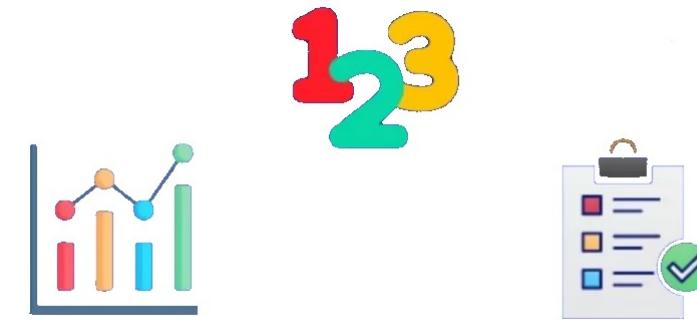
Different Types of Analysis

Different Types of Analysis



Qualitative

Qualitative Analysis



Quantitative

Quantitative Analysis

Different Types of Analysis

Qualitative Data

(Categorical)

- Gender
- Religion
- Marital status
- Native language
- Social class
- Qualifications
- Type of instruction
- Method of treatment
- Type of teaching approach
- Problem-solving strategy used

Quantitative Data

(Numerical)

- Age
- Height
- Weight
- Income
- University size
- Group size
- Self-efficacy test score
- Percent of lecture attended
- Clinical skills performed
- Number of errors

Qualitative Analysis

Content Analysis

- It is a method for subjective interpretation of content of text through the systematic classification process of coding and identifying themes or patterns

Narrative Analysis

- These approaches typically focus on the lives of individuals as told through their own stories.

Discourse Analysis

- It involves analyzing a naturally occurring language use and types of written texts.

Grounded Theory

- is an approach for theory construction through the analysis of data. it is usually inductive in nature.

Thematic Analysis

- It is a method for identifying and analyzing patterns (themes) in the data by means of thematic codes.

Quantitative Analysis

Survey Research

Survey methodology studies the in-depth sampling of individual units from a population and administering data collection techniques on that sample.

Correlation Research

Correlational research is a type of nonexperimental research in which the researcher measures two variables and assesses the statistical relationship between them with little or no effort to control extraneous variables.

Causal-Comparative Research

In causal-comparative research, the researcher investigates the effect of an independent variable on a dependent variable by comparing two or more groups of individuals.

Experimental Research

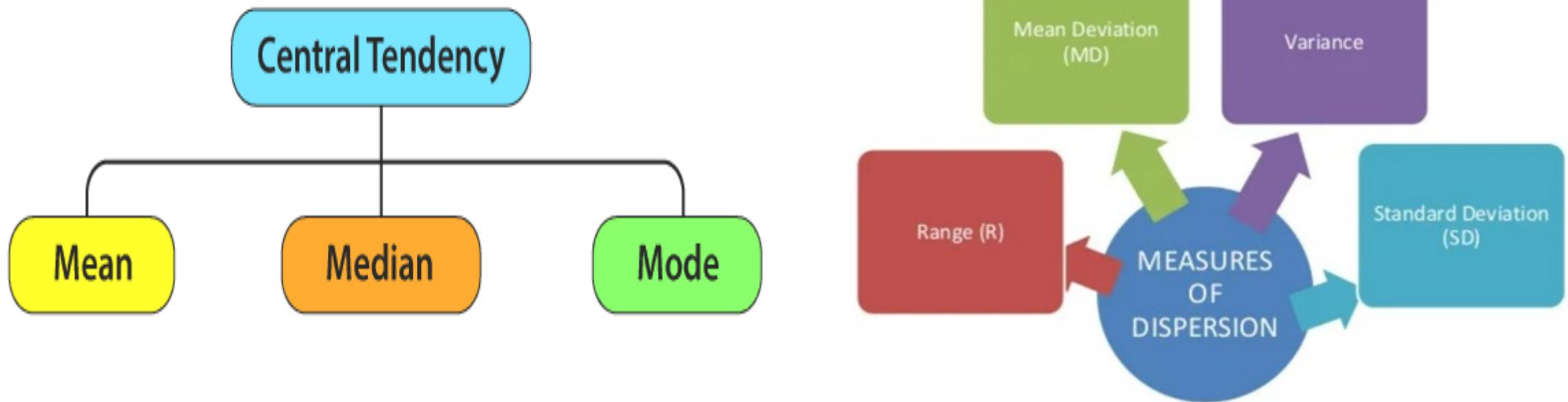
Experimental research is research conducted with a scientific approach using two sets of variables. The first set acts as a constant, which you use to measure the differences of the second set.

Types of in **Statistic Analysis**

Statistic Analysis

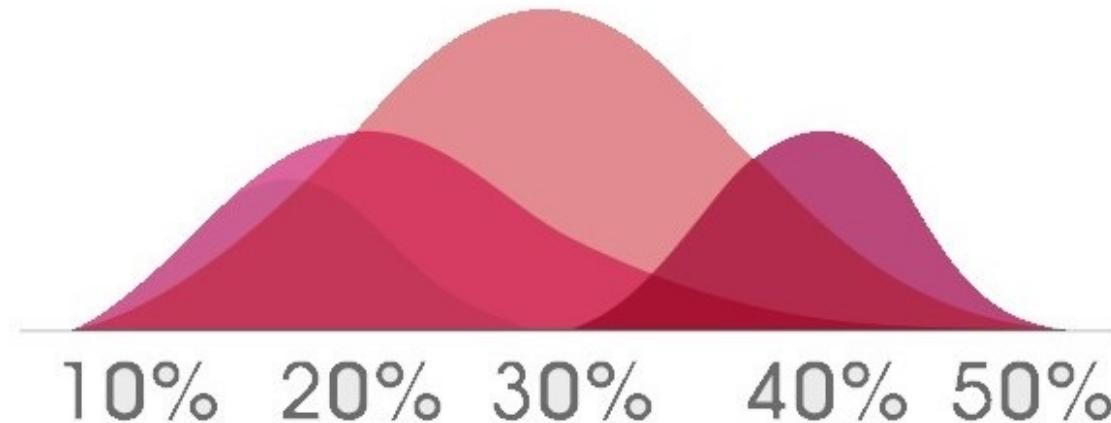
Descriptive Statistics		Inferential Statistics	
Measures of Central Tendency	Measures of Dispersion	Hypothesis Testing	Regression Analysis
Mean	Range	Z test	
Median	Standard Deviation	F test	Linear Regression
Mode	Variance Absolute Deviation	T test	

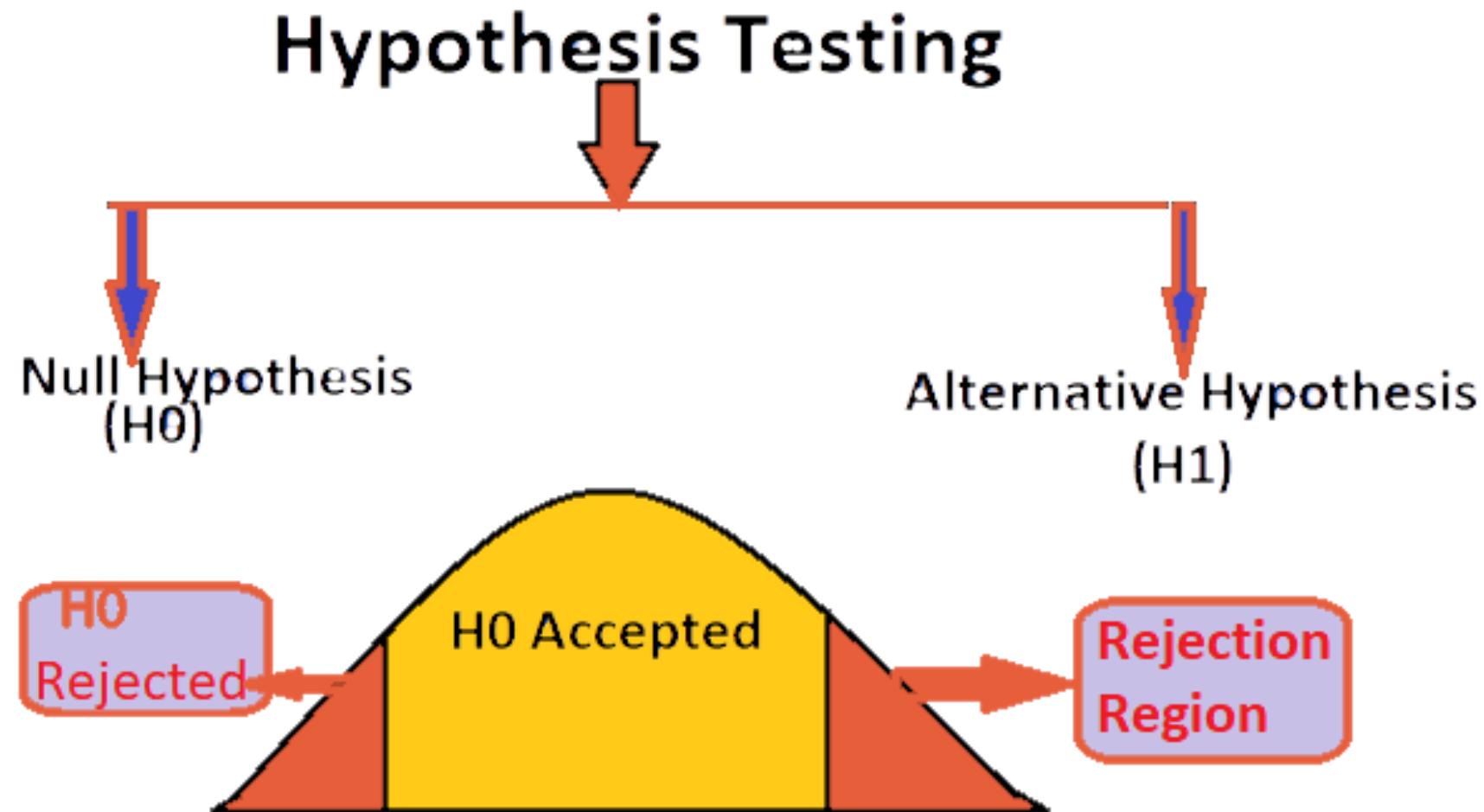
Measures of central tendency and Dispersion



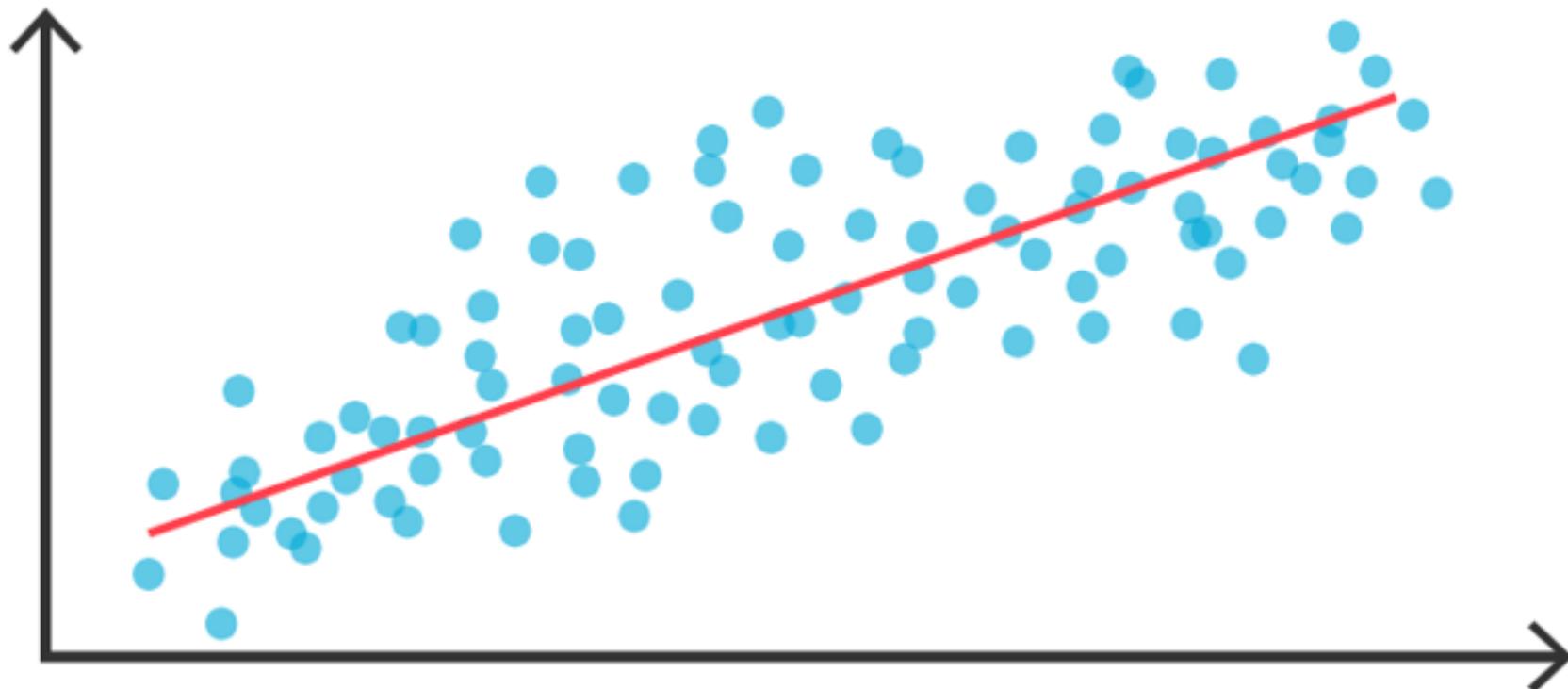
Inferential statistics

studies a sample
of the same data.

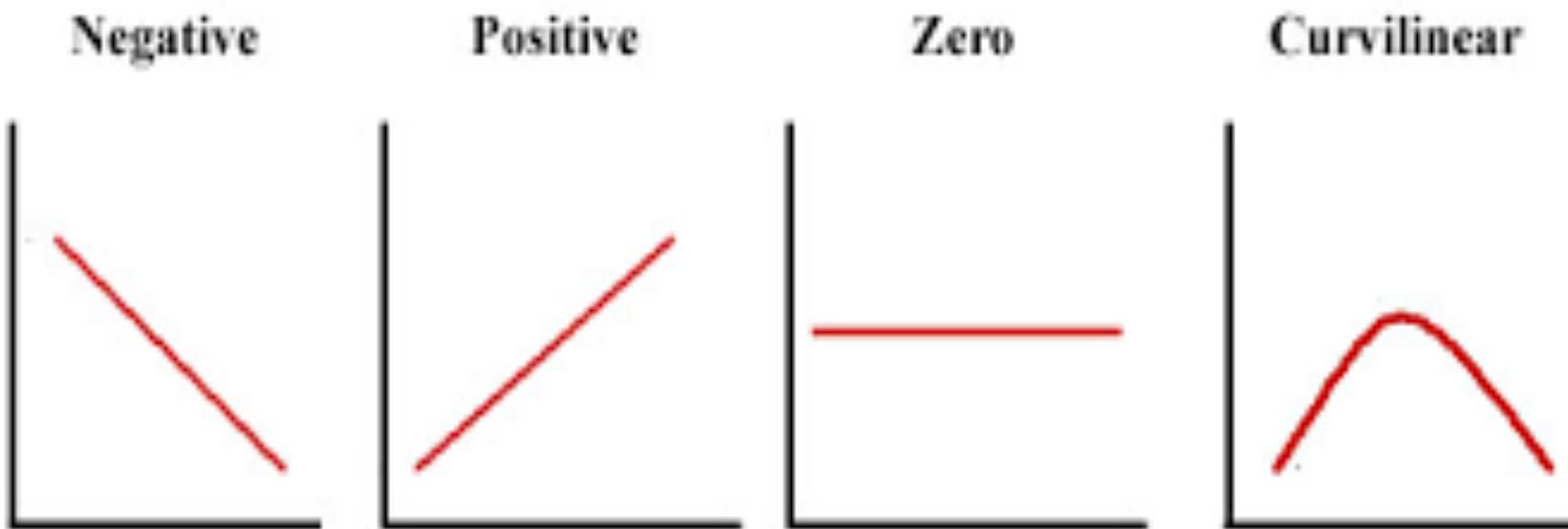




Regression Analysis



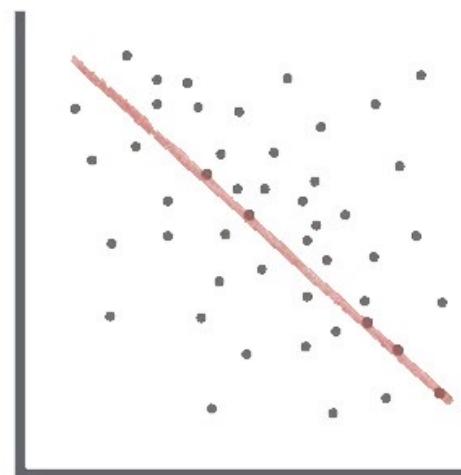
Relationship between variables



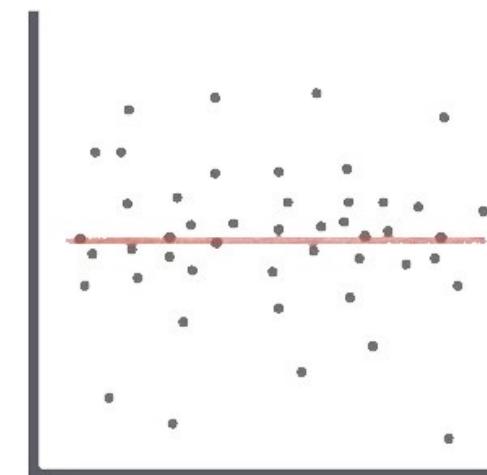
Relationship between variables



Positive Correlation



Negative Correlation

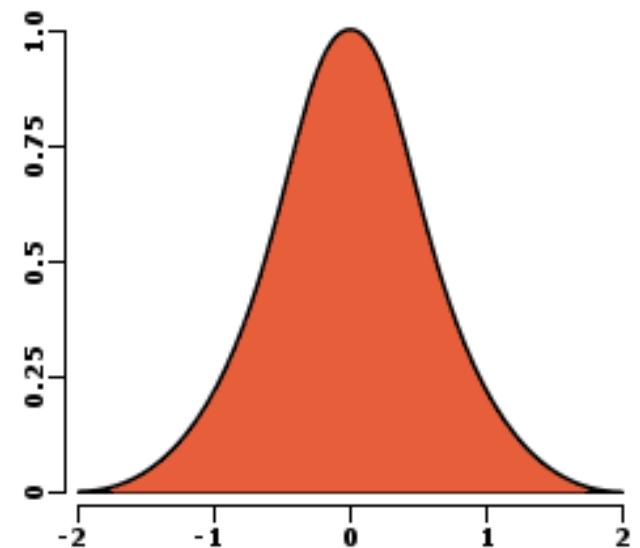
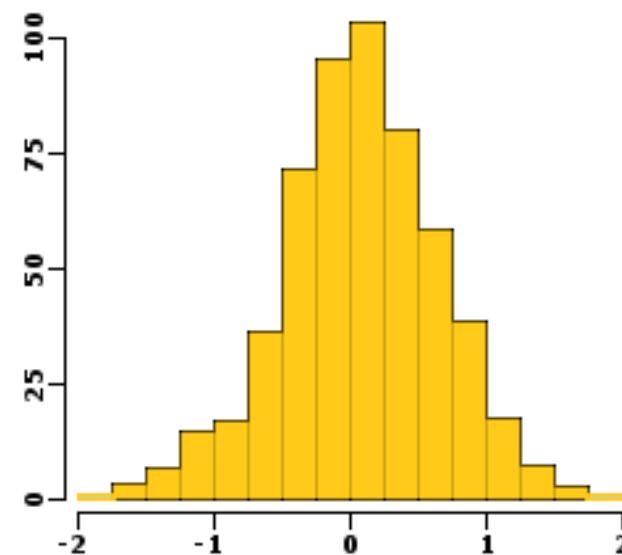


No Correlation

Probability Distribution

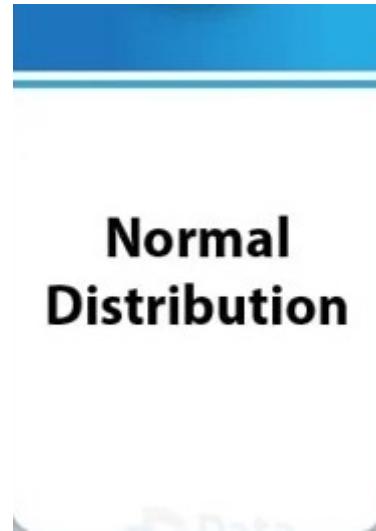
Probability Distribution

Distribution function is a mathematical expression that describes the probability that a system will take on a specific value or set of values.

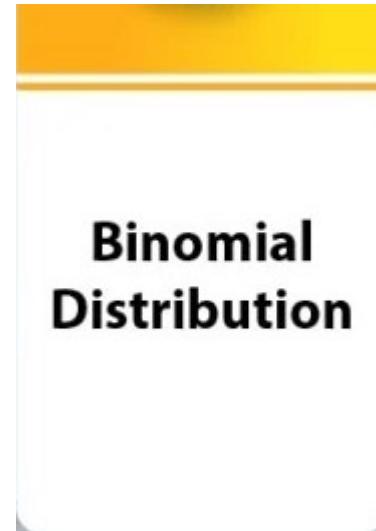


Probability Distribution

**Normal
Distribution**



**Binomial
Distribution**



**Poisson
Distribution**



Probability Distribution

Normal Distribution

Normal Distribution:

The normal distribution is a symmetric probability distribution centered on the mean, indicating that data around the mean occur more frequently than data far from it. The normal distribution is also called Gaussian distribution. The normal distribution curve resembles a bell curve.

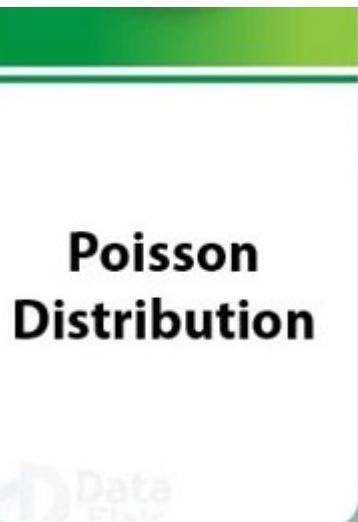
Probability Distribution

Binomial Distribution

Binomial Distribution

Under a given set of factors or assumptions, the binomial distribution expresses the likelihood that a variable will take one of two outcomes or independent values.

Probability Distribution



Poisson Distribution:

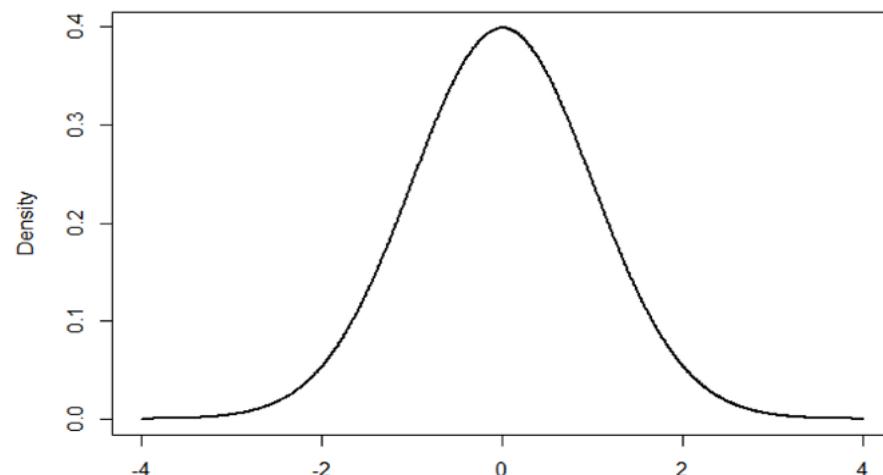
A Poisson distribution is a kind of probability distribution used in statistics to illustrate how many times an event is expected to happen over a certain amount of time. It's also called count distribution.

Probability density Function

&

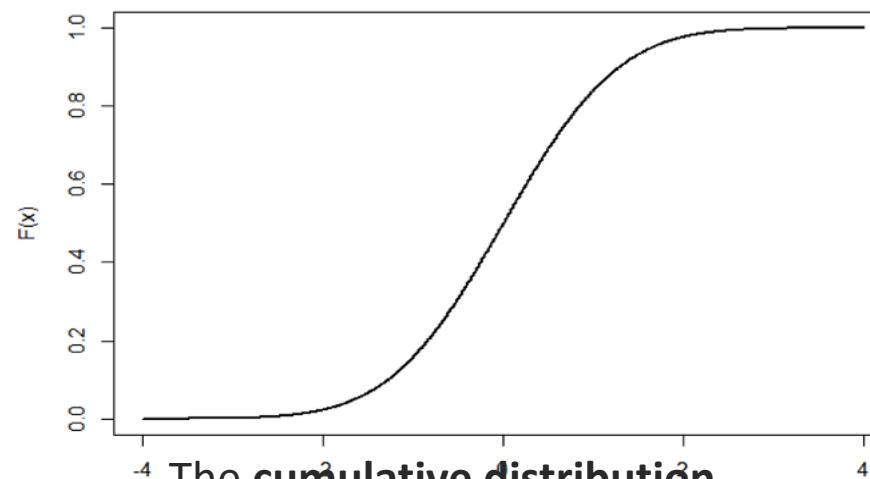
Cumulative Distribution Function

Probability density function (PDF)



*PDF is a statistical term that describes the probability distribution of the **continues** random variable*

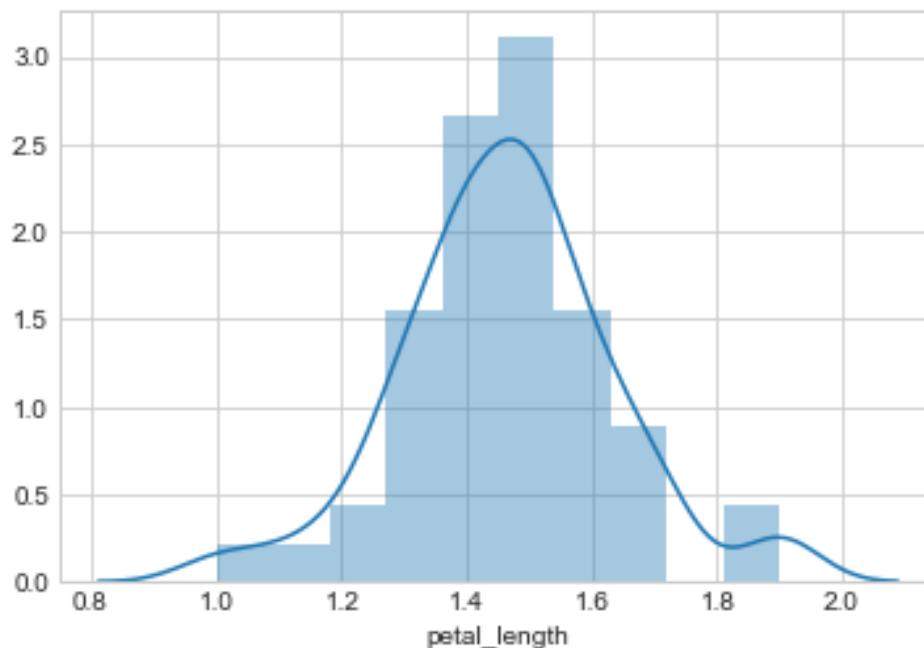
Cumulative distribution function (CDF)



The cumulative distribution function is applicable for describing the distribution of random variables either it is continuous or discrete

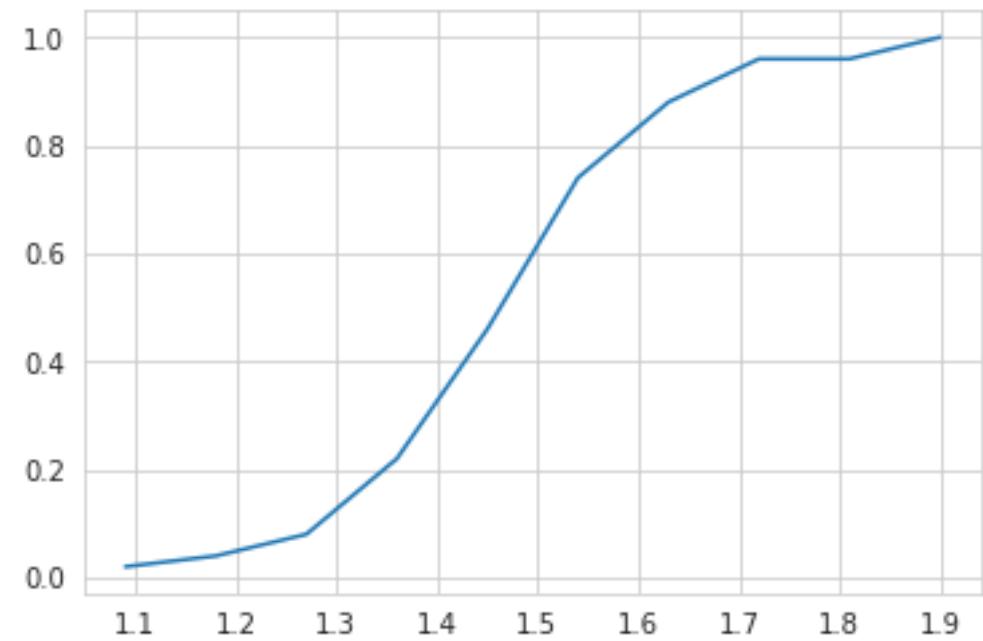
PDF & CDF

PDF On Iris



PDF for ['species']== 'setosa' on petal length

CDF on Iris

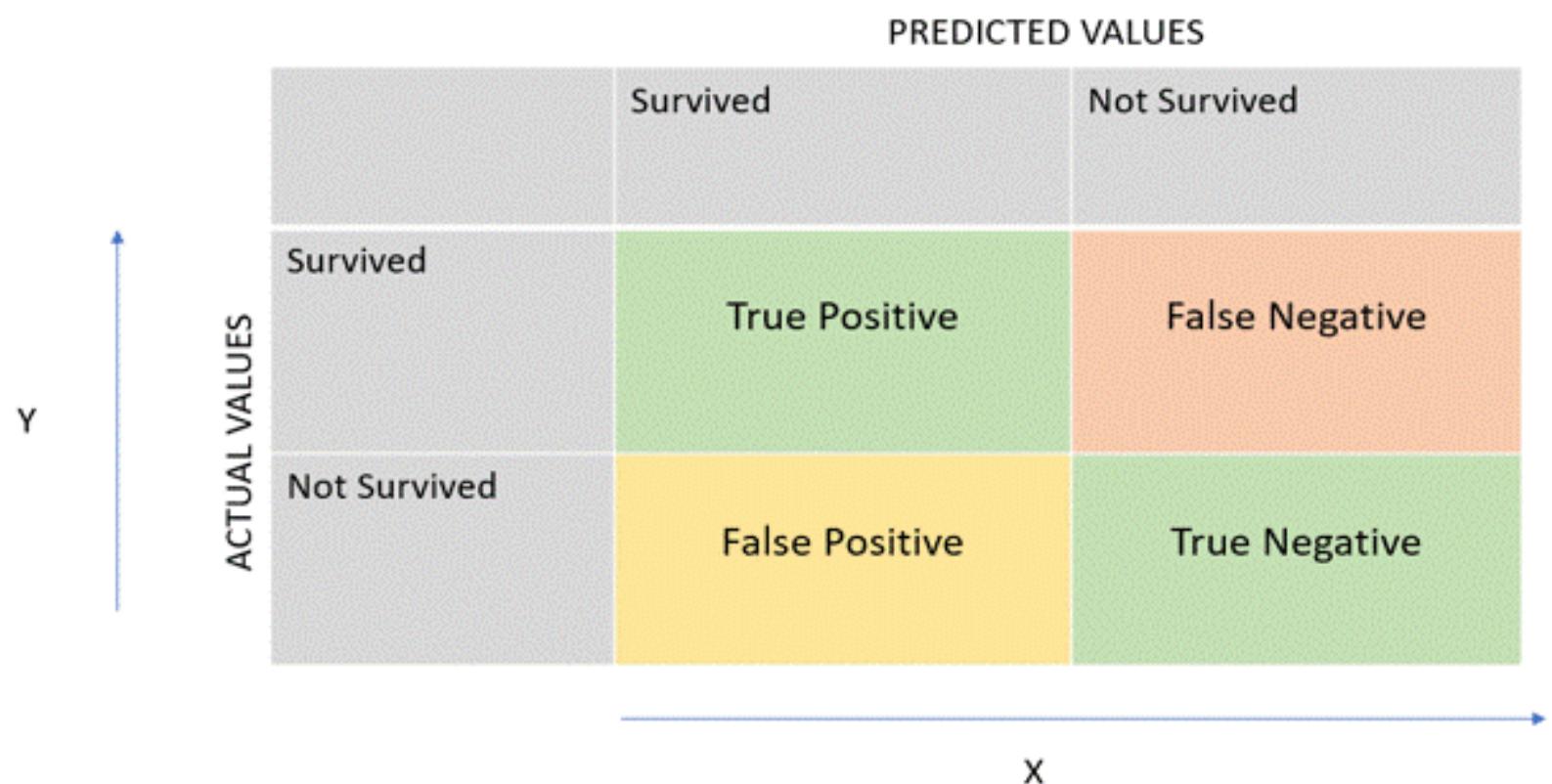


CDF of iris_setosa using petal length

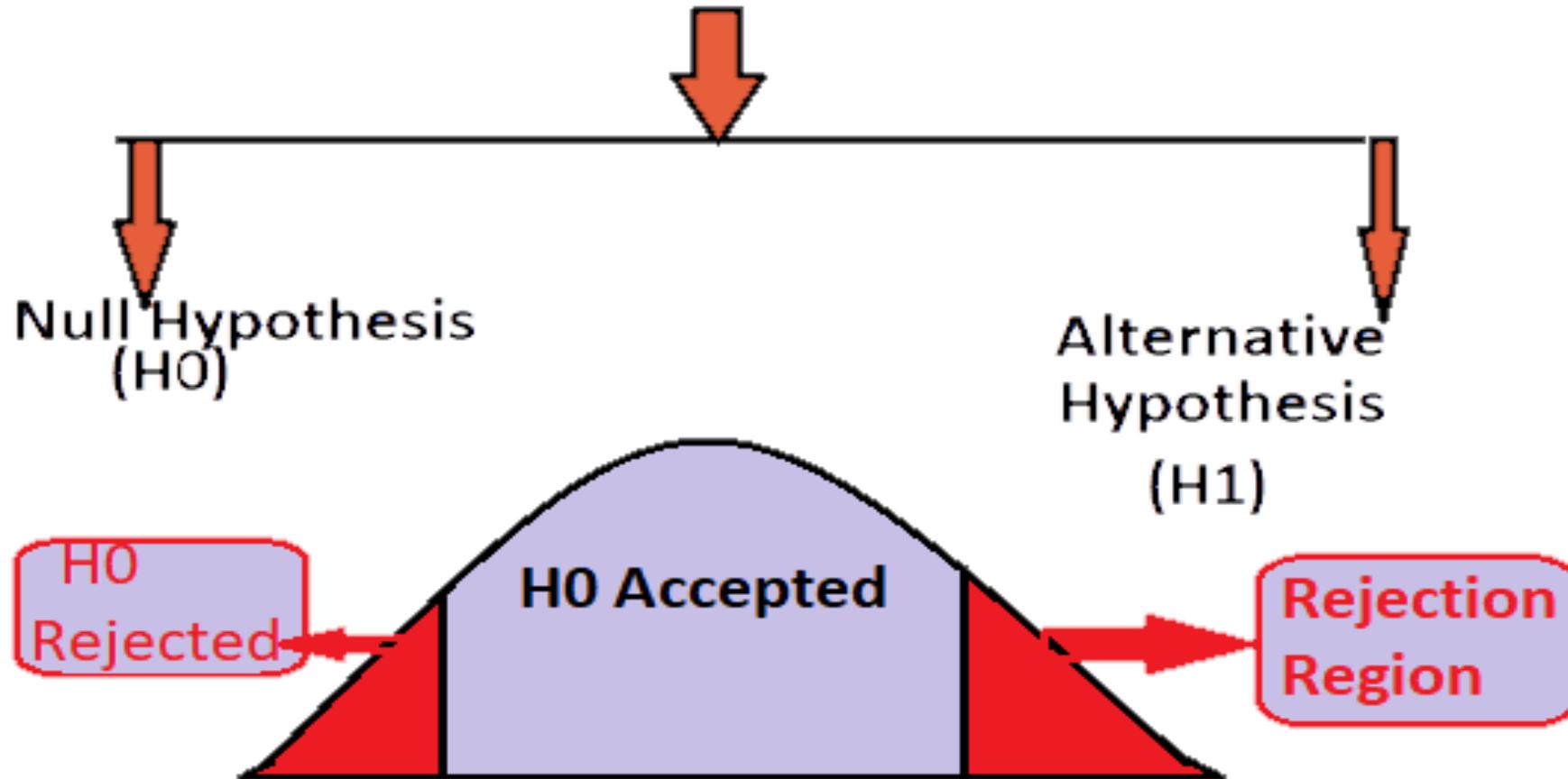
Accuracy

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

CONFUSING!



Hypothesis Testing



Hypothesis Testing

Null Hypothesis

$$H_0$$

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include $=$, \leq , or \geq sign.



Alternative Hypothesis

$$H_a$$

A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include \neq , $>$, or $<$ sign.

Hypothesis Testing

The two types of hypothesis testing are null hypothesis and alternate hypothesis.

[Null hypothesis](#) is the initial assumption about an event (also referred to as the ground truth).

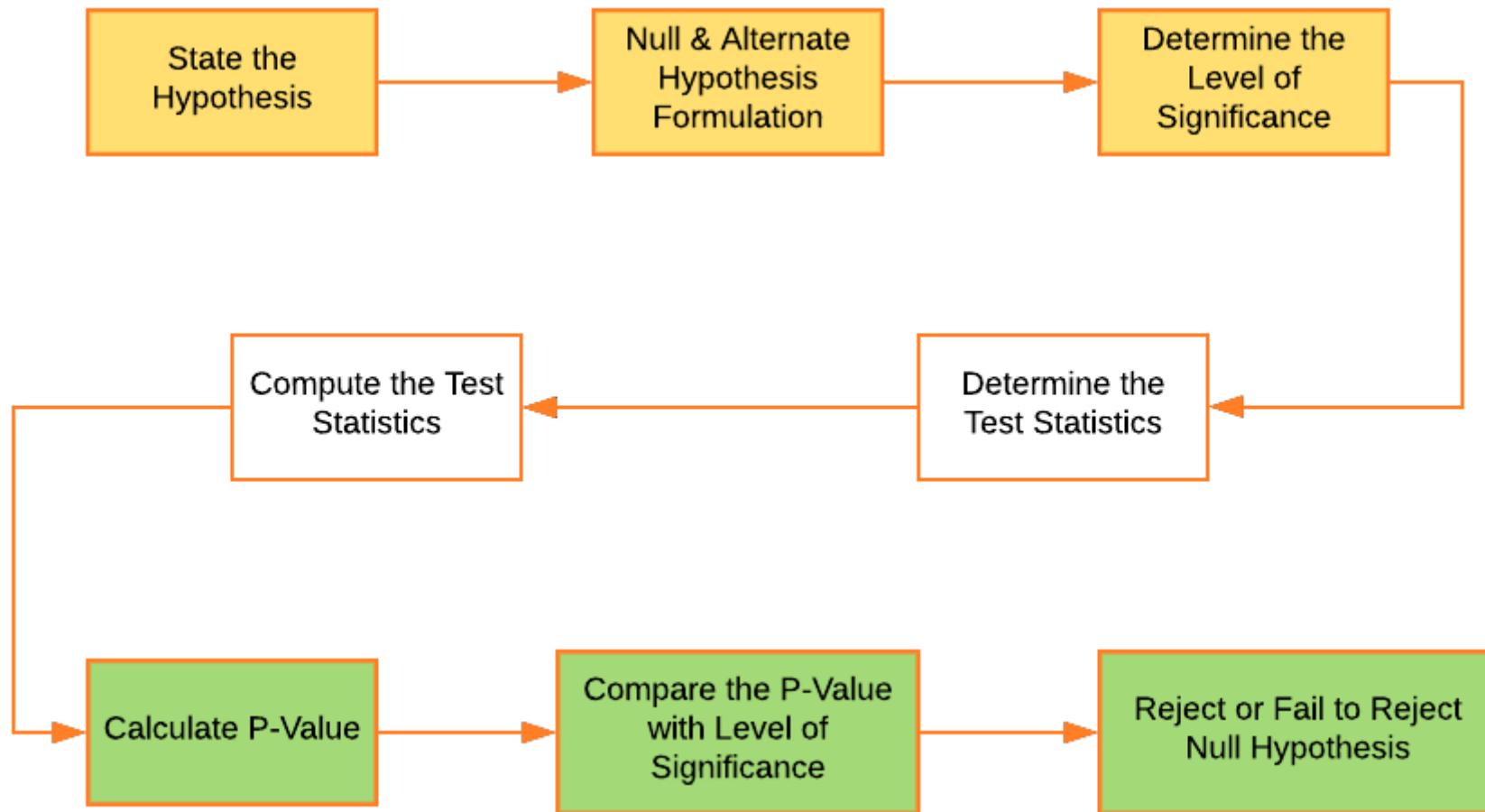
[Alternate hypothesis](#) is an assumption that counters the initial assumption.

Example of Hypothesis Testing With T test

A **t-test** is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

- The null hypothesis (H_0) is that the true difference between these group means is zero.
 - The alternate hypothesis (H_a) is that the true difference is different from zero.
-
- A p-value is a statistical measurement used to validate a hypothesis against observed data.
 - A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
 - The lower the p-value, the greater the statistical significance of the observed difference.
 - A p-value of 0.05 or lower is generally considered statistically significant.
 - P-value can serve as an alternative to or in addition to preselected confidence levels for hypothesis testing.

Hypothesis Testing Workflow



THANK YOU!