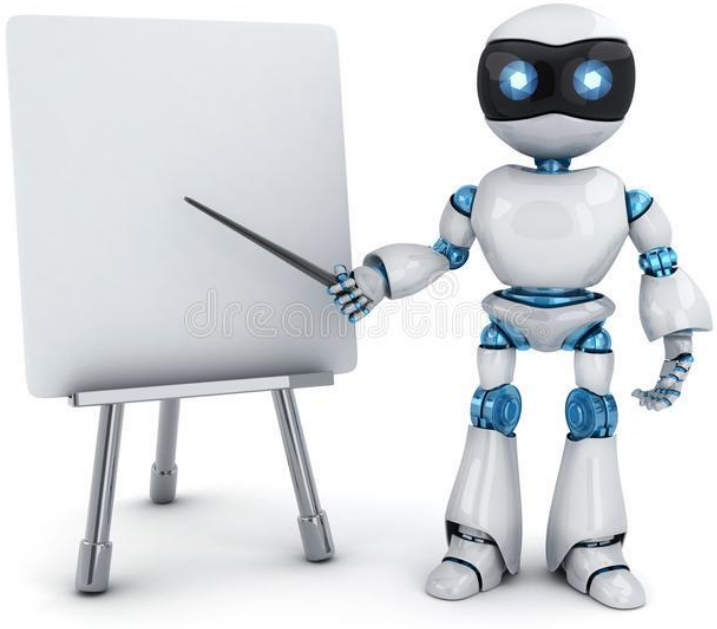
Datamites
Global Institute for Data Science

STATISTICS FOR DATA SCIENCE

Accredited by IABAC™

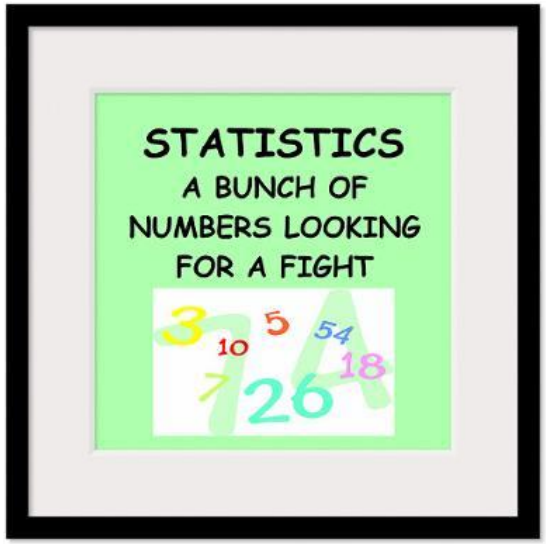


Overview of Statistics

Module - 1

Statistics for Data Science

The science of collecting, describing, and interpreting data is popularly known as Statistical leveraging in Data Science



Two areas of Statistics in Data Science:

Descriptive statistics – Methods of organizing, summarizing, and presenting data in an informative way

Inferential statistics – The methods used to determine something about a population on the basis of a sample

Applications of statistical concepts

Datamites
Global Institute for Data Science

- Finance – correlation and regression, index numbers, time series analysis
- Marketing – hypothesis testing, chi-square tests, nonparametric statistics
- Academic Research – hypothesis testing, chi-square tests, nonparametric tests
- Operating Management – hypothesis testing, estimation, analysis of variance, time series analysis
- Retailing – Sales data, Distribution analysis, Instore promotions, product assortment, new product development

Descriptive Statistics

Descriptive statistics are methods for organizing and summarizing data.

For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.

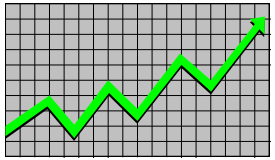
A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

Collect data

e.g., Survey

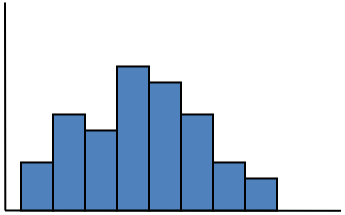
Present data

e.g., Tables and graphs



Summarize data

e.g., Sample mean = $\frac{\sum X_i}{n}$



Inferential Statistics

Datamites
Global Institute for Data Science

- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions about a **population** based on **sample** results

Definition to Basic terms

Datamites
Global Institute for Data Science

Population: A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

Sample: A subset of the population.

Variable: A characteristic about each individual element of a population or sample.

Data (singular): The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

Data (plural): The set of values collected for the variable from each of the elements belonging to the sample.

Experiment: A planned activity whose results yield a set of data.

Parameter: A numerical value summarizing all the data of an entire population.

Statistic: A numerical value summarizing the sample data.

Let's first understand the basic concepts of statistics.

Statistical Population

A collection of all probable observations of a specific characteristic of interest
Example: All learners taking this course

Sample

A subset of population
Example: A group of 20 learners selected for a quiz

Variable

An item of interest that can acquire various numerical values
Example: The number of defective items manufactured in a factory

Parameter

A population characteristic of interest
Example: The average income of a class of people

Cochran's Formula Example

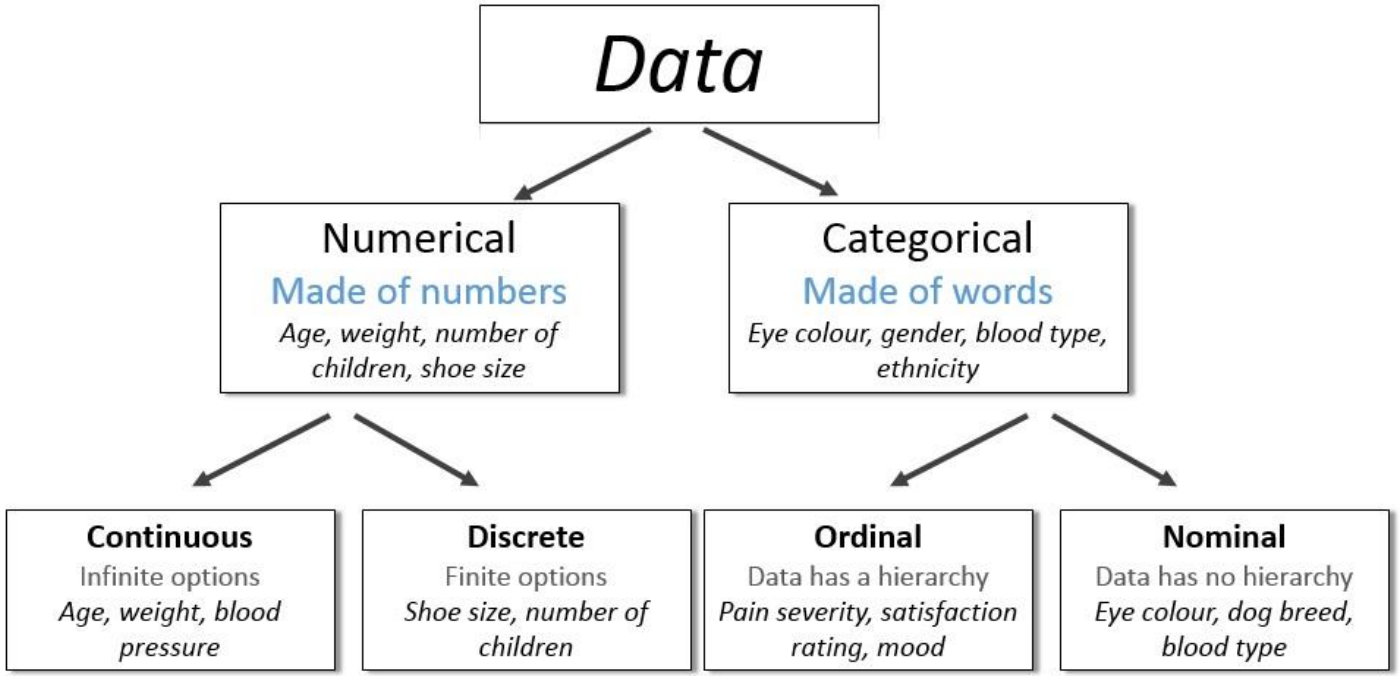
Datamites
Global Institute for Data Science

- Suppose we are doing a study on the inhabitants of a large town, and want to find out how many households serve breakfast in the mornings.
- We don't have much information on the subject to begin with, so assuming 50% households have breakfast: p =0.05
- Now let's say we want 95% confidence, and at least 5 percent—plus or minus—precision.
- A 95 % confidence level gives us Z values of 1.96, per the normal tables, so we get

$$n_0 = \frac{Z^2 pq}{e^2}$$

$((1.96)^2 (0.5) (0.5)) / (0.05)^2 = 385.$

Types of Data



Examples of variables

Datamites
Global Institute for Data Science

Example: Identify each of the following as examples of qualitative or numerical variables:

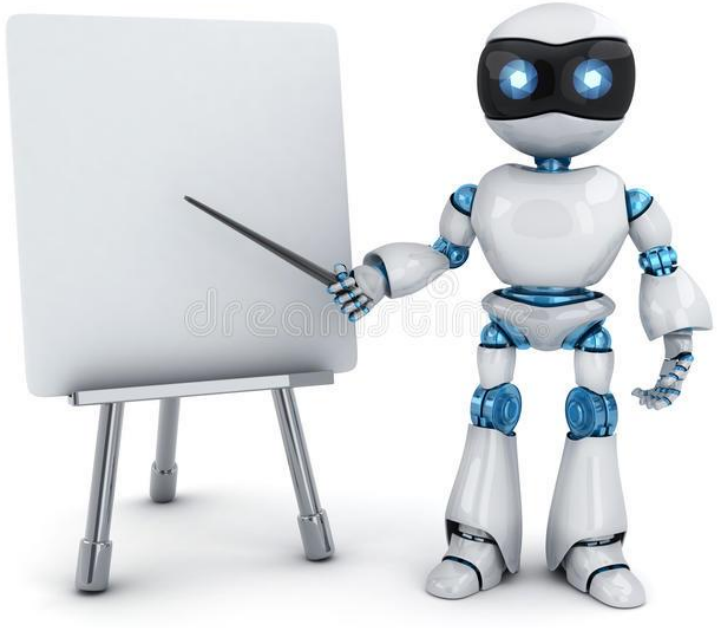
1. The temperature in Barrow, Alaska at 12:00 pm on any given day.
2. The make of automobile driven by each faculty member.
3. Whether or not a 6 volt lantern battery is defective.
4. The weight of a lead pencil.
5. The length of time billed for a long distance telephone call.
6. The brand of cereal children eat for breakfast.
7. The type of book taken out of the library by an adult.

Examples of variables…

Example: Identify each of the following as examples of
(1) nominal, (2) ordinal, (3) discrete, or (4) continuous variables:

- The length of time until a pain reliever begins to work.
- The number of chocolate chips in a cookie.
- The number of colors used in a statistics textbook.
- The brand of refrigerator in a home.
- The overall satisfaction rating of a new car.
- The number of files on a computer's hard disk.
- The pH level of the water in a swimming pool.
- The number of staples in a stapler.

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔



Harnessing Data

Module - 2

Making sense of the data

Datamites
Global Institute for Data Science

A sample should have the same characteristics as the population it is representing.

Sampling can be:

- **with replacement**: a member of the population may be chosen more than once
- **without replacement**: a member of the population may be chosen only once (lottery ticket)

Sampling methods

Datamites
Global Institute for Data Science

Sampling methods can be:

- **random** (each member of the population has an equal chance of being selected**)**
- **nonrandom**

The actual process of sampling causes **sampling errors**.

For example, the sample may not be large enough or representative of the population. Factors not related to the sampling process cause non sampling errors. A defective counting device can cause a Non sampling error.

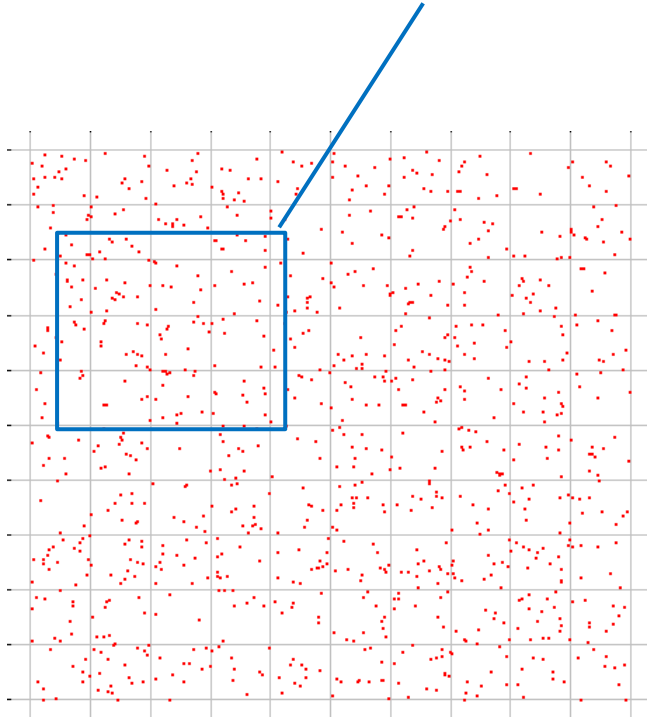
Measurement on Samples

- Many datasets are **samples** from an **infinite population**.
- We are most interested in **measures on the population**, but we have access only to a **sample** of it.

A sample measurement is called a **"statistic".** Examples:

- Sample min, max, mean, std. deviation

That makes measurement hard:
Sample measurements are **"noisy,"** i.e. vary from one sample to the next
Sample measurements may be **biased**, i.e. systematically be different from
the measurement on the population.
Sample measurements have **variance**: variation between samples
Sample measurements have **bias**, systematic variation from the population value.



Sampling Example

Datamites
Global Institute for Data Science

Example: An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded.

There are 2712 employees.

Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

Random sampling methods

Datamites
Global Institute for Data Science

- **simple random sample** (each sample of the same size has an equal chance of being selected)
- **stratified sample** (divide the population into groups called strata and then take a sample from each stratum)
- **cluster sample** (divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.)
- **systematic sample** (randomly select a starting point and take every n-th piece of data from a listing of the population)

Sampling methods…

Datamites
Global Institute for Data Science

Biased Sampling Method: A sampling method that produces data which systematically differs from the sampled population. An **unbiased sampling method** is one that is not biased.

Sampling methods that often result in biased samples:

1. **Convenience sample**: sample selected from elements of a population that are easily accessible.
2. **Volunteer sample**: sample collected from those elements of the population which chose to contribute the needed information on their own initiative.

Process of data collection

Datamites
Global Institute for Data Science

Define the objectives of the survey or experiment.

Example: Estimate the average life of an electronic component.

2. Define the variable and population of interest.

   Example: Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate descriptive or inferential data-analysis techniques.

Sampling Error

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics

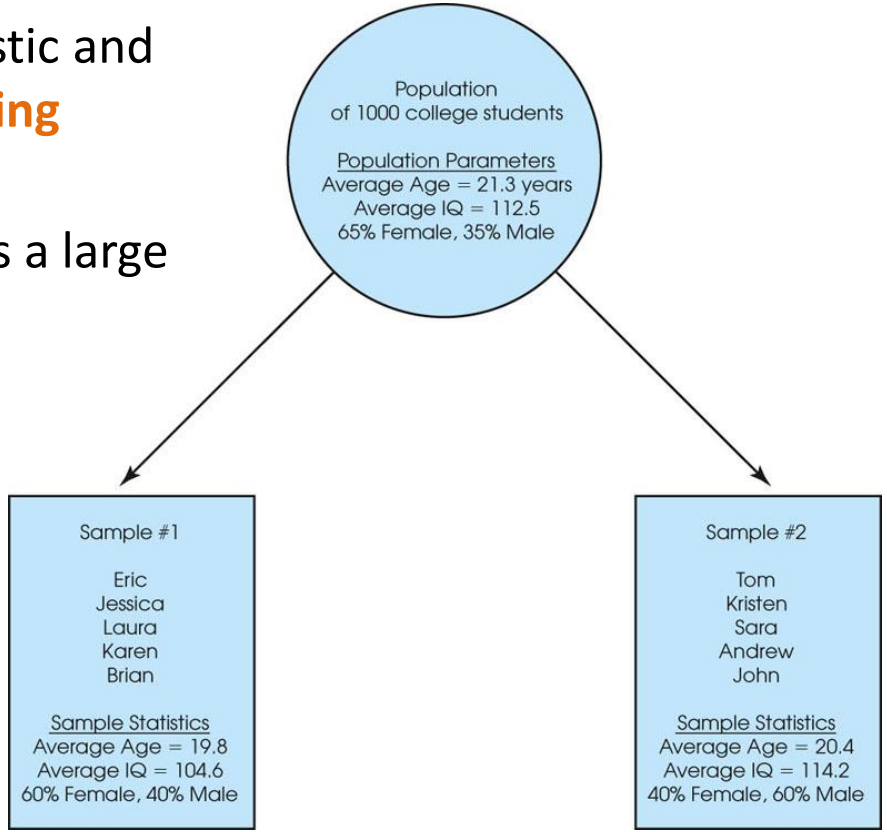


Figure 1.2
A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called sampling error.

Methods used to collect data

Datamites
Global Institute for Data Science

Experiment: The investigator controls or modifies the environment and observes the effect on the variable under study.

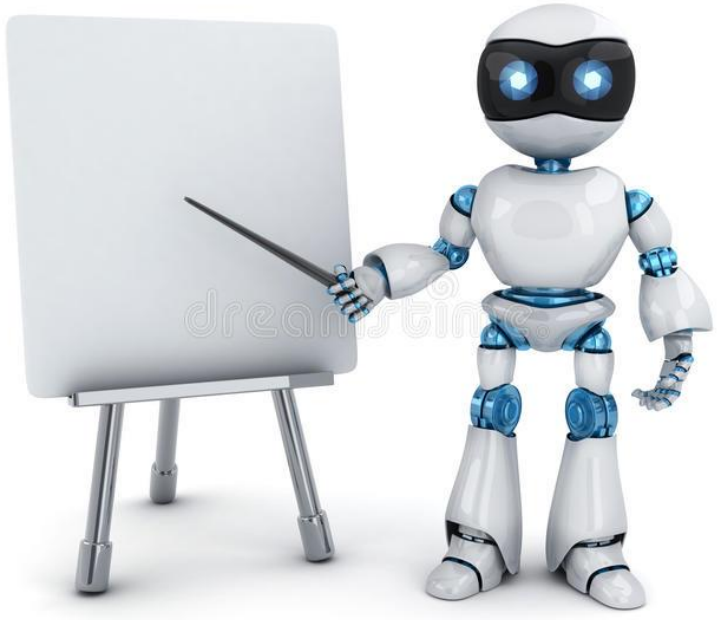
Survey: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

Census: A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

Judgment Samples: Samples that are selected on the basis of being "typical."

Items are selected that are representative of the population. The validity of the results from a judgment sample reflects the soundness of the collector's judgment.

Probability Samples: Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.



Exploratory Analysis

Module - 3

Measures of Central Tendencies

Measures of Central Tendency

- Mean
- Median
- Mode

Mean

The mean is the average of all numbers and is sometimes called the arithmetic mean.

Median

The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median

Mode

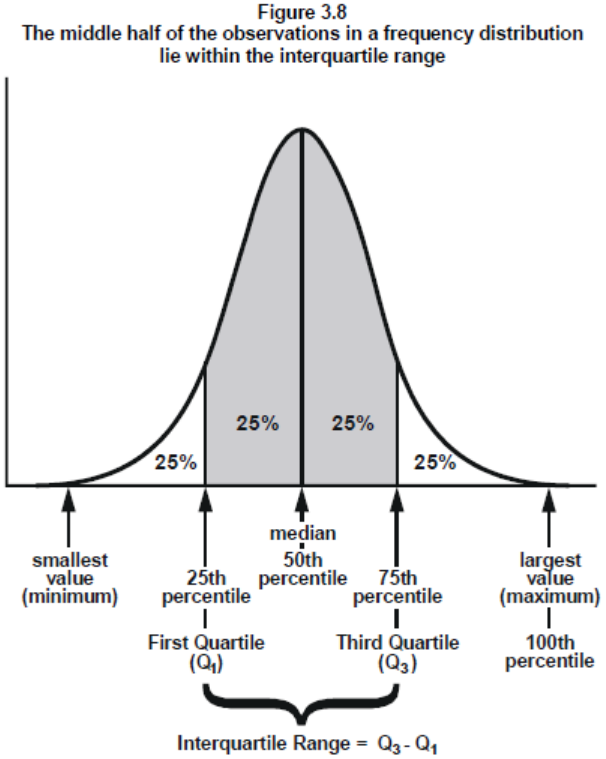
The mode is the number that occurs most often within a set of numbers.

Data Variability

Range

Datamites
Global Institute for Data Science

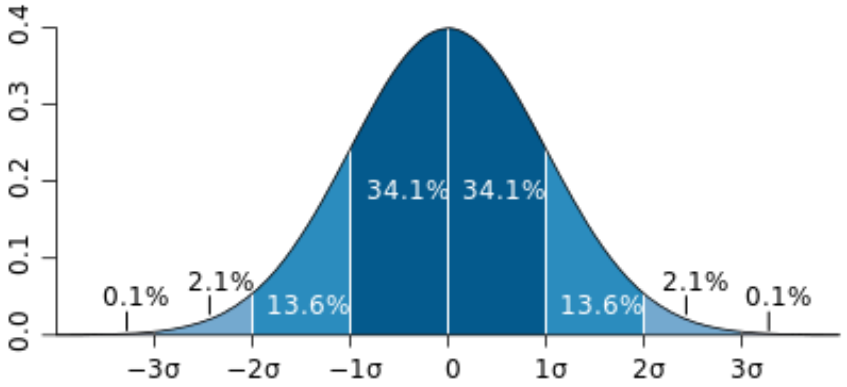
The range is the difference between the highest and lowest values within a set of numbers.



Standard Deviation (**σ**)

Datamites
Global Institute for Data Science

Standard Deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.



Calculating Standard Deviation

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Example:

$\bar{x}$ = mean = ⑧

n = sample size = ④

sample size minus 1 = ⓝ − 1 = 3

$$S = \sqrt{\frac{1}{3}\sum_{i=1}^{4}(x_i - 8)^2}$$

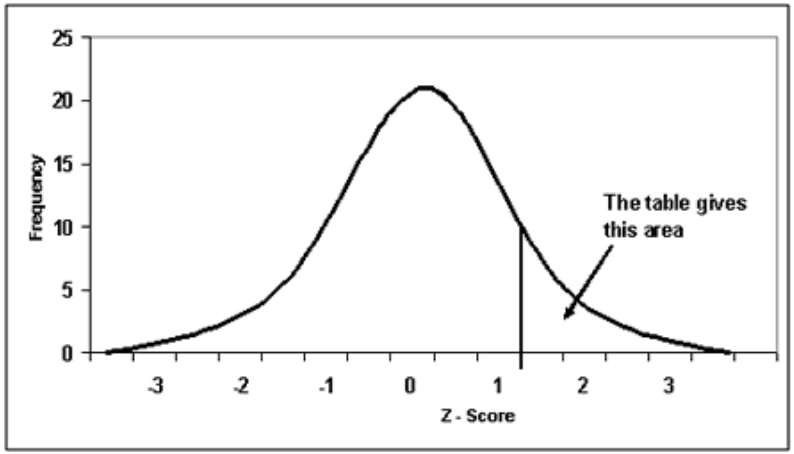
$$S = \sqrt{\frac{(2-8)^2 + (8-8)^2 + (10-8)^2 + (12-8)^2}{3}}$$

$$S = \sqrt{\frac{36 + 0 + 4 + 16}{3}} = \sqrt{18.7} = 4.32$$

Z- Score / Z-Value/ Standard Score

- A z-score (aka, a standard score) indicates how many standard deviations an element is from the mean. A z-score can be calculated from the following formula.

- $z = (X - \mu) / \sigma$



Calculating Z-Score

Datamites
Global Institute for Data Science

Formula:

$$z\text{-}score = \frac{x_i - \bar{x}}{s}$$

---

x_i = data point

$\bar{x}$ = mean

s = standard deviation

Example:

$$z = \frac{231 - 130.1}{47.85} = 2.11$$

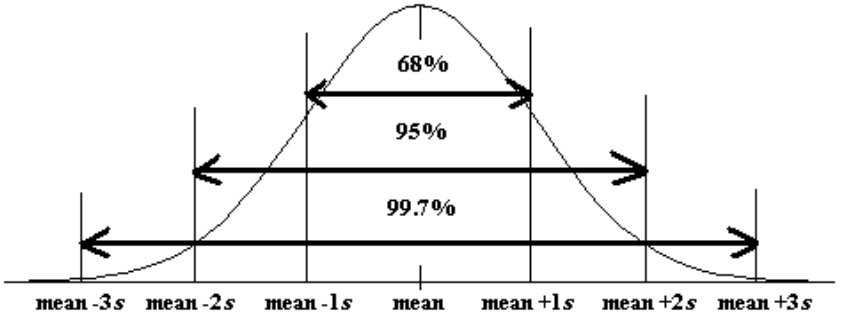
$$z = \frac{50 - 130.1}{47.85} = -1.67$$

Emphirical Rule

Datamites
Global Institute for Data Science

- The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

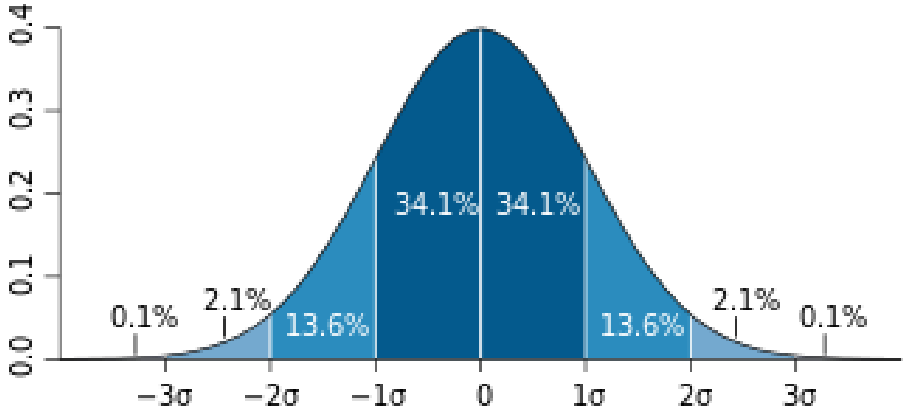
- 68% of data falls within the first standard deviation from t
- 95% fall within two standard deviations. (2 Sigma)
- 99.7% fall within three standard deviations. (3 Sigma)
- 99.99966% → (6 Sigma)



Calculating Percentiles

Datamites
Global Institute for Data Science

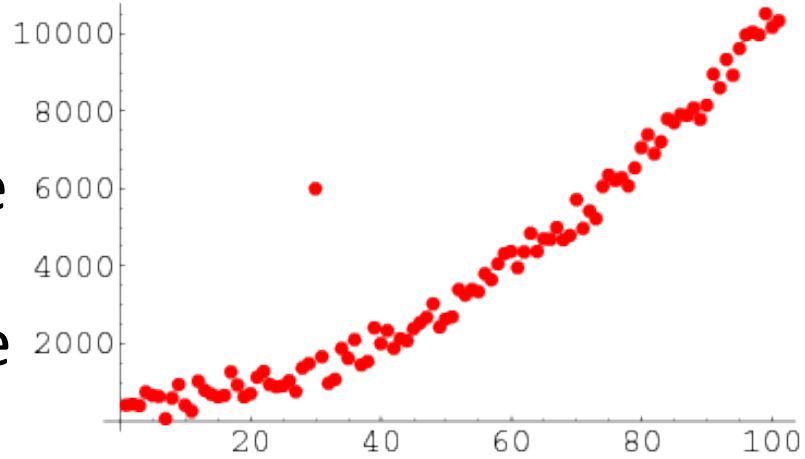
- A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.



Outliers

Datamites
Global Institute for Data Science

an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.



Distributions & Central Limit Theorem

Normal Distributions, Mean, Variance

Datamites
Global Institute for Data Science

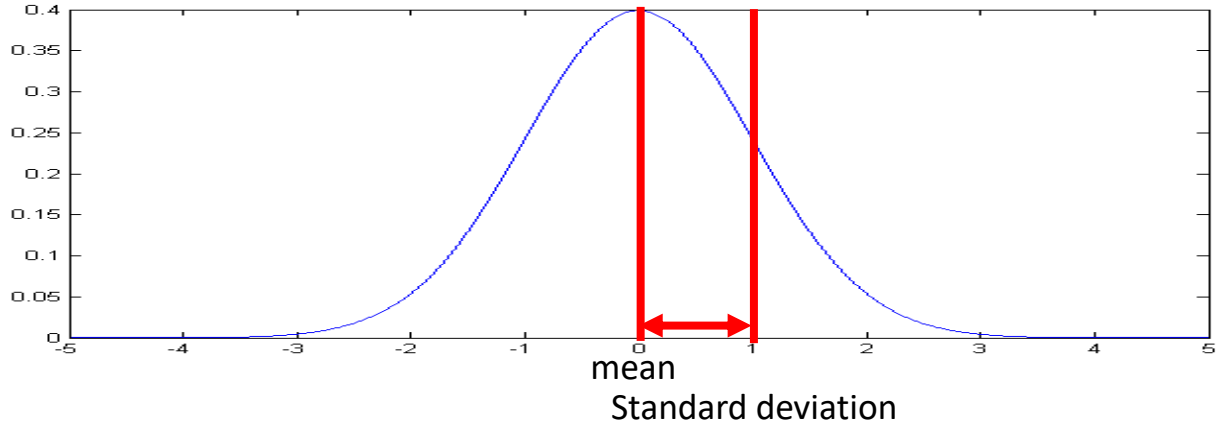
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of points from the mean:

The standard deviation is the square root of variance.

The normal distribution is completed characterized by mean and variance.

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$



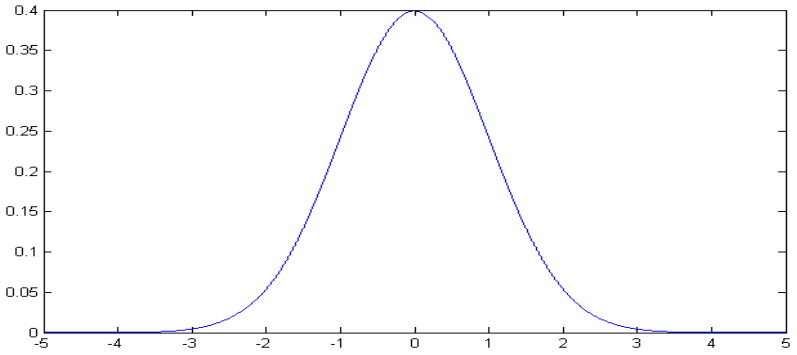
Central Limit Theorem

Datamites
Global Institute for Data Science

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as $n \rightarrow \infty$.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

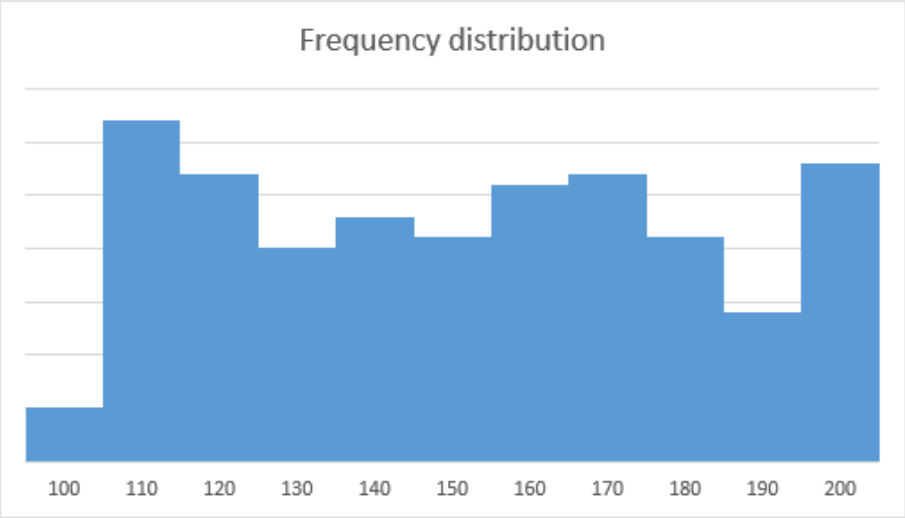
Def: The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are identical in size), regardless of population distribution shape..



Central Limit Theorem - Explanation

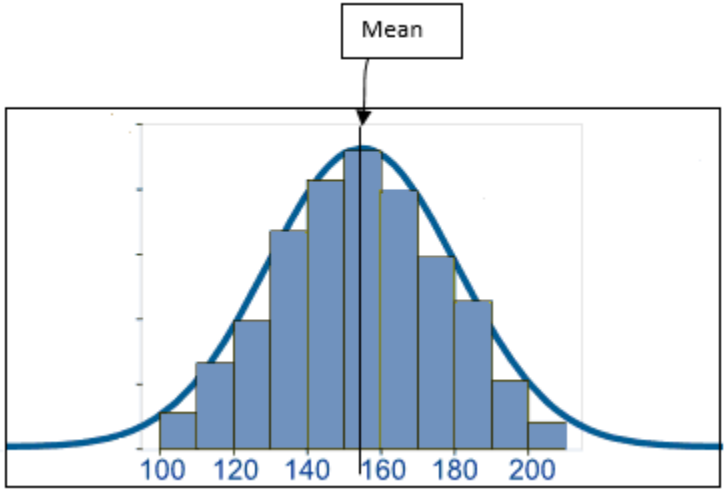
Datamites
Global Institute for Data Science

- Let's say we have the cholesterol levels of all the people in India, we can look at the mean, median and mode of the data. Maybe plot a histogram with sensible ranges and look at the data. Let's assume this is how the data looks like.



Central Limit Theorem - Explanation

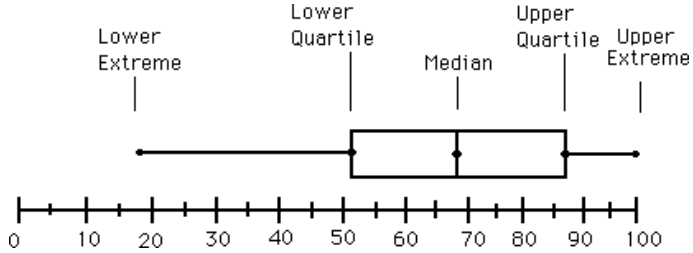
- we take the data of some 50 people and calculate their mean.We again take a sample of some 50 people and calculate the mean and we keep doing that for quite a number of times. We now plot the means of these samples.



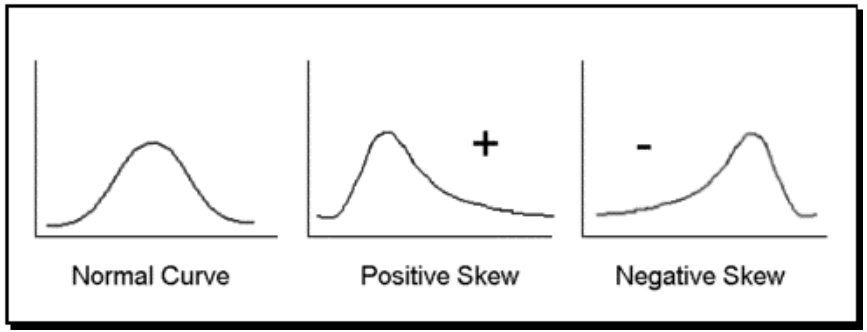
Correcting distributions

Many statistical tools, including mean and variance, t-test, ANOVA etc. **assume data are normally distributed**.

Very often this is not true. The box-and-whisker plot is a good clue

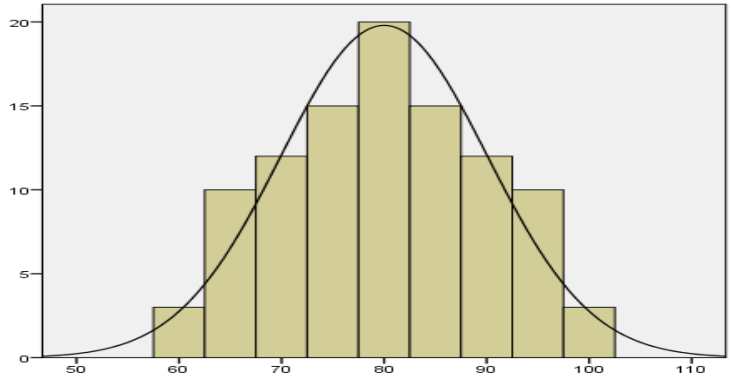


Whenever its asymmetric, the data cannot be normal. The histogram gives even more information



Histogram Normalization

Its not difficult to turn histogram normalization into an algorithm:



- Draw a normal distribution, and compute its histogram into k bins.
- Normalize (scale) the areas of the bars to add up to 1.
- If the left bar has area 0.04, assign the top 0.04-largest values to it, and reassign them a value “60”.
- If the next bar has area 0.10, assign the next 0.10-largest values to it, and reassign them a value “65” etc.

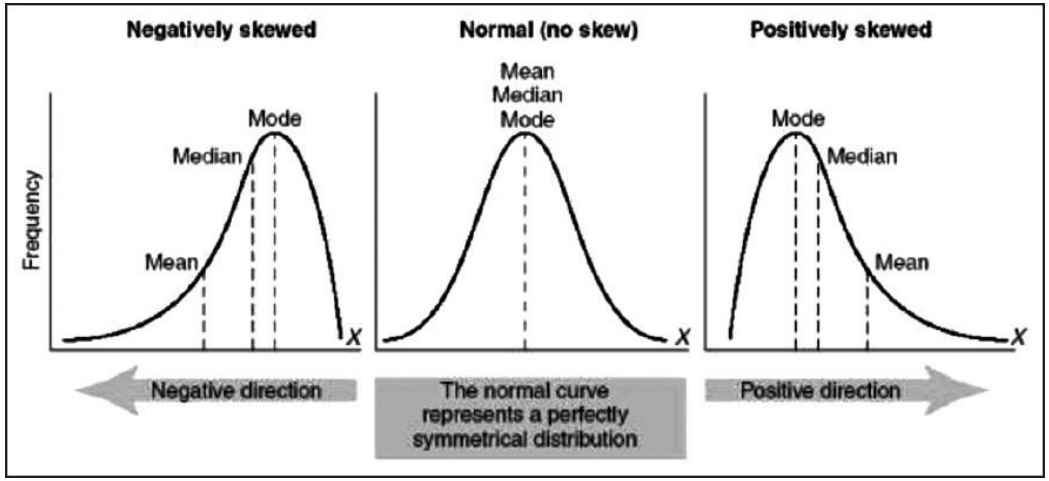
Normality Testing

Datamites
Global Institute for Data Science

- Skewness
- Kurtosis

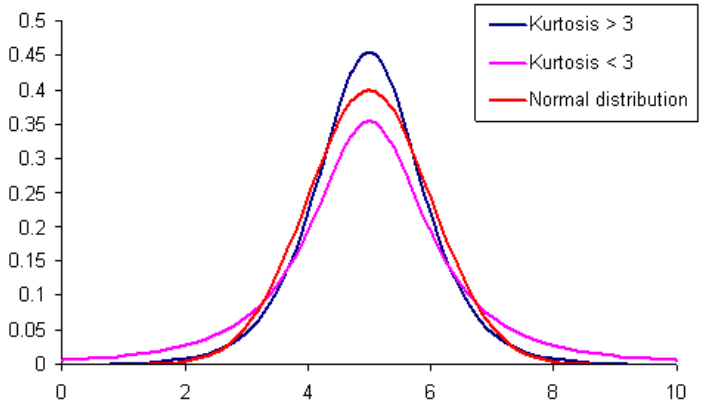
Skewness

- In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined.



Kurtosis

- In probability theory and statistics, kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.



How much Skewness and Kurtosis

Datamites
Global Institute for Data Science

- when |S| > 1.96 the skewness is significantly (alpha=5%) different from zero.
- Same for the kurtosis |K| > 1.96, from 3

Measure of Distance

Measure of distance

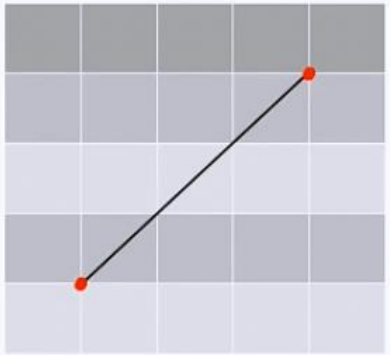
Datamites
Global Institute for Data Science

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance

Euclidean Distance

It is a classical method to calculate the distance between two objects X and Y in the Euclidean space (1- or 2- or n- dimension space). This distance can be calculated by traveling along the line, connecting the points.

Euclidean Distance



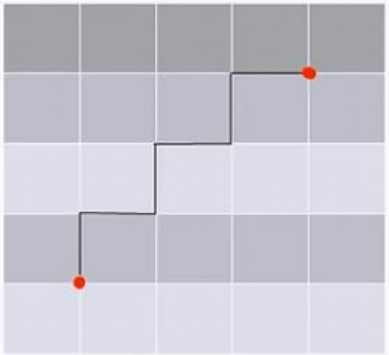
You can use the Pythagorean Theorem to compute this distance:

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Manhattan Distance

It is similar to Euclidean Distance, but the distance (for example, two points, separated by building blocks in a city) is calculated by traversing vertical and horizontal lines in the grid-based system.

Manhattan Distance



You can use the following formula to compute this distance:

$$d_t = |x_2 - x_1| + |y_2 - y_1|$$

Minkowski Distance

It is a metric on the Euclidean space and can be considered as a generalization of both the Euclidean and Manhattan distances.

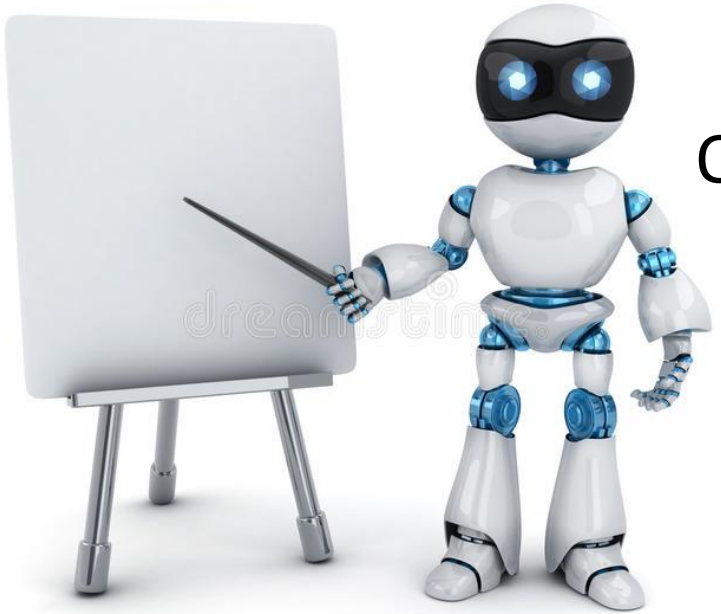
You can use the following formula to compute this distance:

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

When r =1; it computes the Manhattan distance.

When r =2; it computes the Euclidean distance.

When r =∞; it computes Supremum.



Hypothesis Testing &

Other computational Techniques

Module - 4

Hypothesis Testing

Datamites
Global Institute for Data Science

We want to prove a hypothesis H_A but its hard so we try to **disprove a null hypothesis H_0**

A **test statistic** is some measurement we can make on the data which is likely to be **big under H_A** but **small under H_0**.

Example:

- We suspect that a particular coin isn't fair.
- We toss it 10 times, it comes up heads every time
- We conclude it's not fair, why?
- How sure are we?

Now we toss a coin 4 times, and it comes up heads every time.

- What do we conclude?

Null Hypothesis

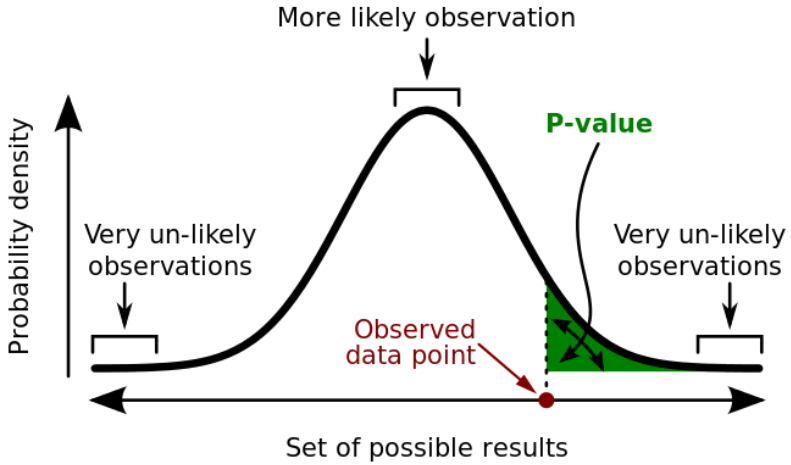
Datamites
Global Institute for Data Science

- Null Hypothesis (H_0)-> No difference between groups (or to a hypothesized mean value)
- Alternate Hypothesis (H_a)-> There is a difference
- The alternative hypotheses are typically of the form <, > or ≠
- You usually only reject H_0 or H_a Can't be proven TRUE

p –Value

- In statistical hypothesis testing, the p-value or probability value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or of greater magnitude than the actual observed results

- Threshold value of p, called the significance level of the test, traditionally 5% or 1%
- Low p -> reject Null
- High P -> fail to reject Null

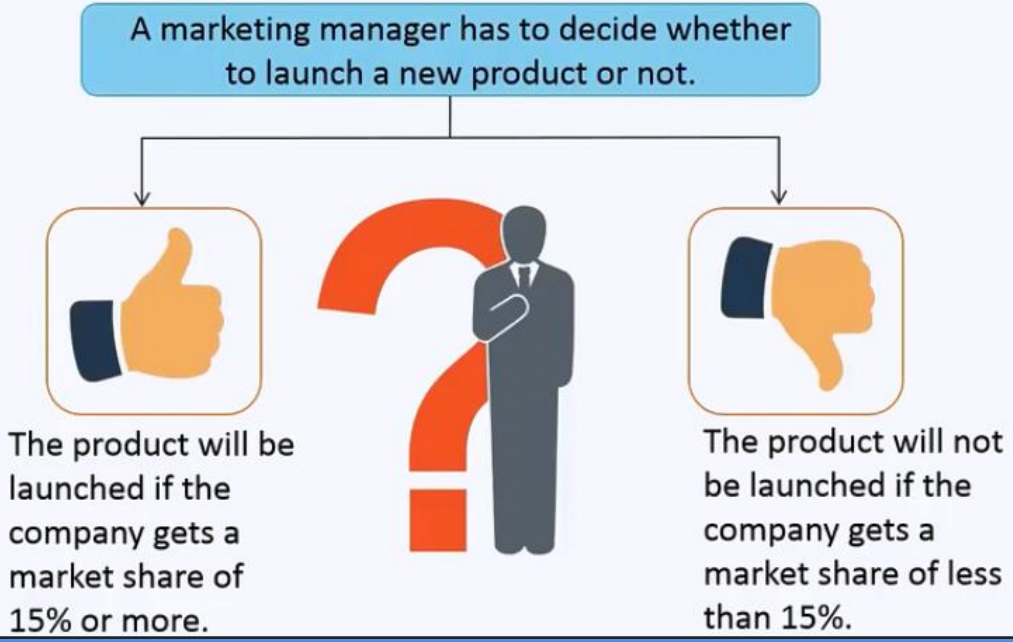


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Need for Hypothesis Testing in Business

Datamites
Global Institute for Data Science

Businesses collect and analyze data to help managers make optimal decisions that maximize profit at minimum risk. This depends on the acceptance or rejection of a hypothesis. For example, suitable hypothesis formulation and testing can help in the situation described below:



Types of Hypothesis Tests

Datamites
Global Institute for Data Science

- Parametric Tests
- Non – Parametric Tests

Hypothesis Tests (Parametric)

Datamites
Global Institute for Data Science

- One Sample T
- Two Sample T
- Paired t
- Correlation/Regression Analysis
- One Way ANOVA

Non Parametric

Datamites
Global Institute for Data Science

- Chi Square Test
- Mann-Whitney Test
- Wilcoxon Signed-Rank Test
- Kruskal-Wallis Test
- Friedman’s ANOVA

Hypothysis Testing – Case Discussion

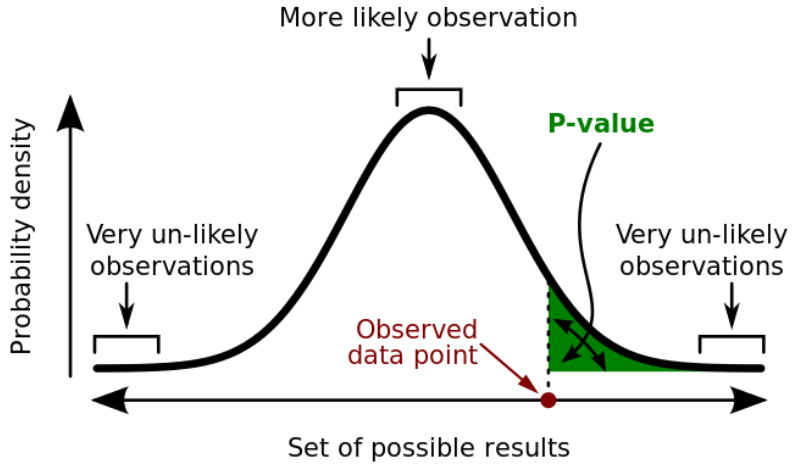
Datamites
Global Institute for Data Science

- Consider a court of law; the null hypothesis is that the defendant is innocent
- We require evidence to reject the null hypothesis (convict)
- If we require little evidence, then we would increase the percentage of innocent people convicted (type I errors); however we would also increase the percentage of guilty people convicted (correctly rejecting the null)
- If we require a lot of evidence, then we increase the the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors)

p –Value

- In statistical hypothesis testing, the p-value or probability value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or of greater magnitude than the actual observed results

- Threshold value of p, called the significance level of the test, traditionally 5% or 1%
- Low p -> reject Null
- High P -> fail to reject Null



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

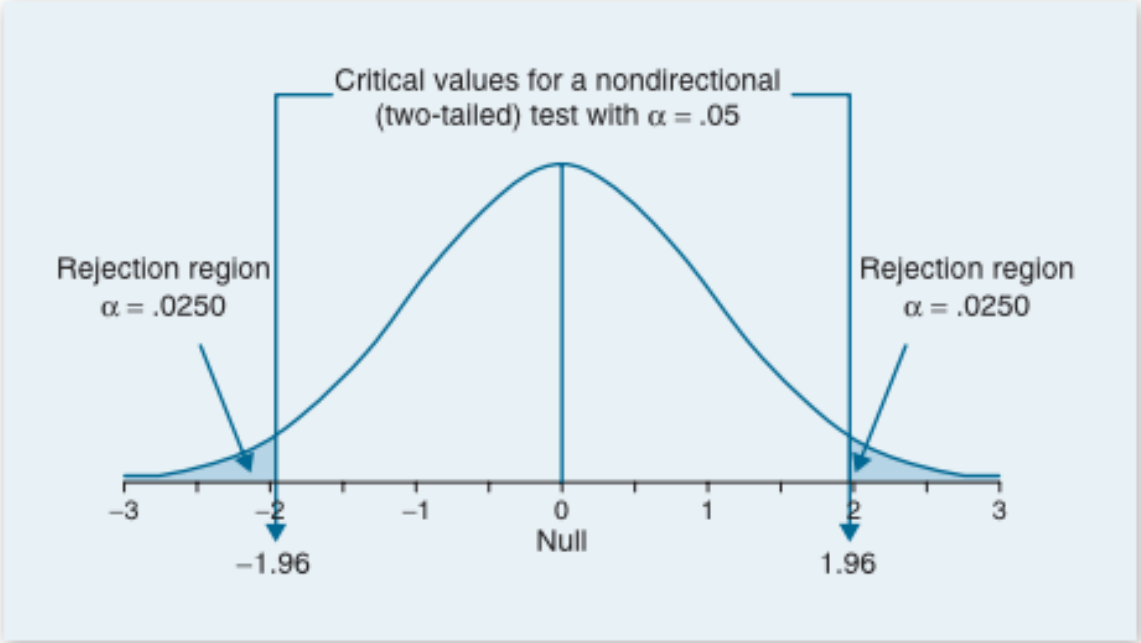
Type 1 & Type 2 Error

Datamites
Global Institute for Data Science

		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT 1 - α	TYPE II ERROR α
	False	TYPE II ERROR β	CORRECT 1 - β

Important Parametric tests

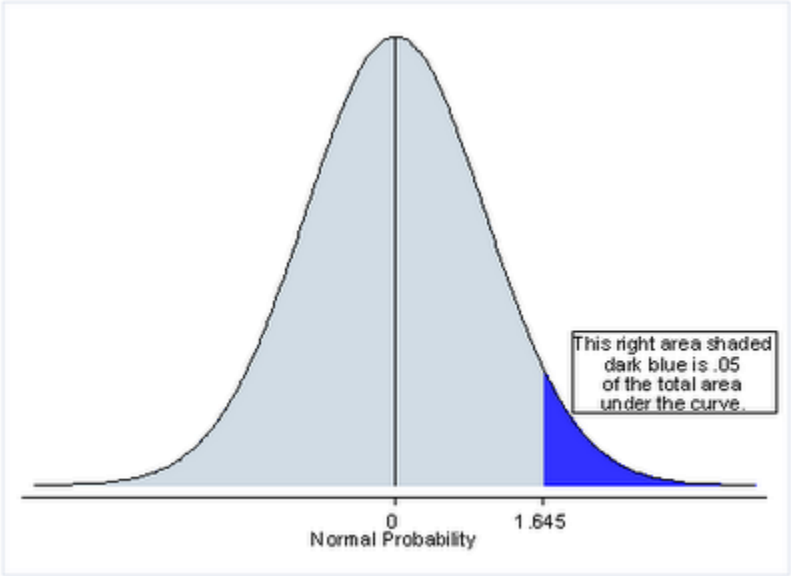
- **T-test:** compare two groups, or two interventions on one group.
- **CHI-squared and Fisher's test.** Compare the counts in a "contingency table".
- **ANOVA:** compare outcomes under several discrete interventions.



When the p value is less than 5% ($p < .05$),
we reject the null hypothesis

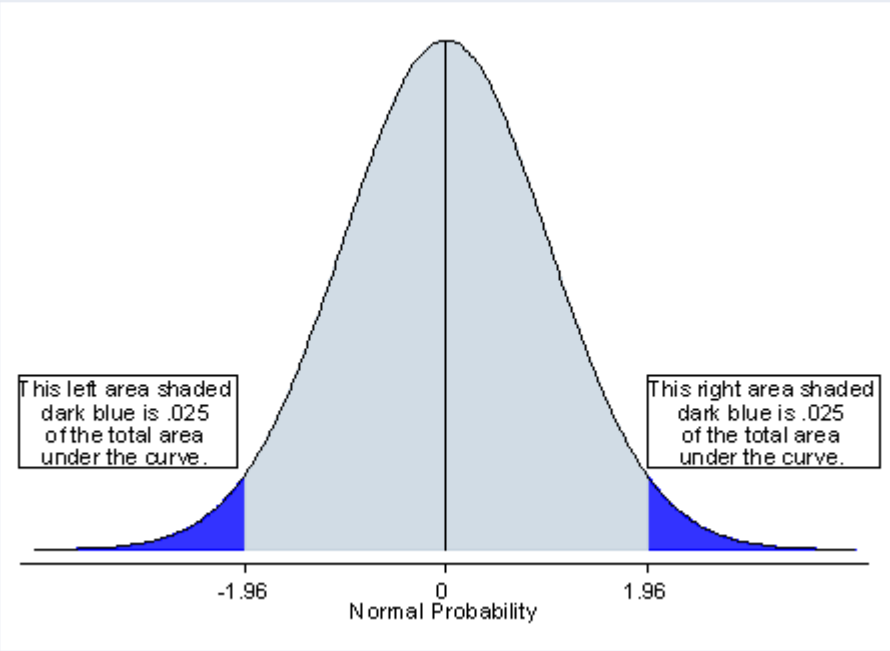
1-tailed Test

- The critical value is either + or -, but not both.
- In this case, you would have statistical significance ($p < .05$) if t ≥ 1.645.

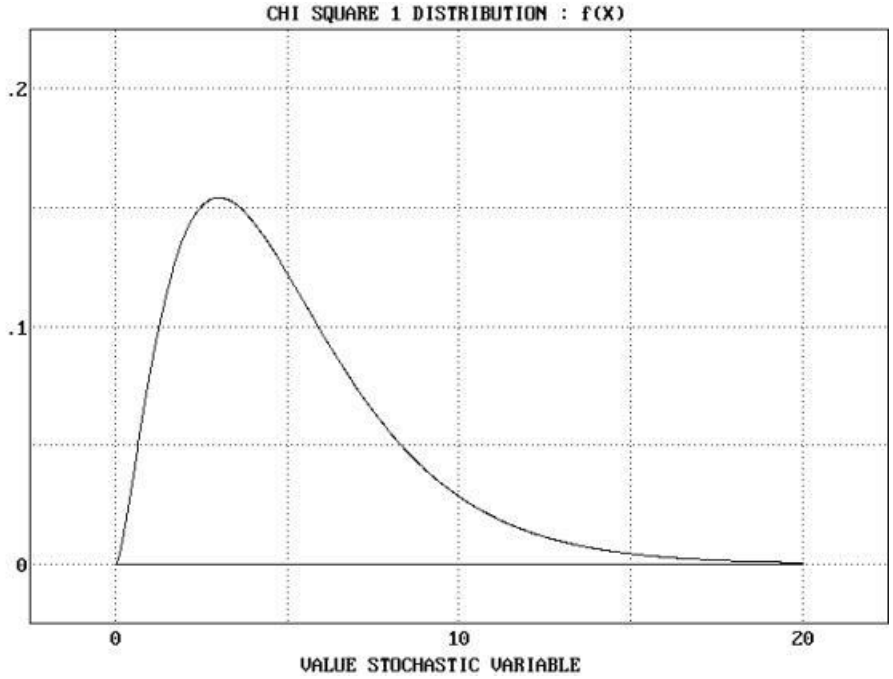


2-tailed Test

- The critical value is the number that separates the "blue zone" from the middle (± 1.96 this example)
- In a *t*-test, in order to be statistically significant the *t* score needs to be in the "blue zone"
- If α = .05, then 2.5% of the area is in each tail



Chi-Square (χ^2)



- Any number squared is a positive number
- Therefore, area under the curve starts at 0 and goes to infinity
- To be statistically significant, needs to be in the upper 5% ($\alpha = .05$)
- Compares observed frequency to what we expected

T-test

- A t-test is an analysis of two populations means through the use of statistical examination; a t-test with two samples is commonly used with small sample sizes, testing the difference between the samples when the variances of two normal distributions are not known

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}}$$

X -> Mean of sample set
μ -> Mean of Population
S -> Standard deviation of sample
N -> Sample size

One Sample T-test

- 29
- 26
- 30
- 28
- 26
- 99
- 31
- 31

Two sample Ind T-test

Datamites
Global Institute for Data Science

- 29
- 26
- 30
- 28
- 26
- 29
- 31
- 31

- 30
- 29
- 34
- 25
- 24
- 26
- 30
- 34

Paired T-test

Datamites
Global Institute for Data Science

Before Red Bull	After Red Bull
36	18
32	12
28	10
23	12
27	11
23	13

One-Way ANOVA

ANOVA (ANalysis Of VAriance) allows testing of multiple differences in a single test. Suppose our experiment design has an independent variable Y with four levels:

Y

Primary School	High School	College	Grad degree
4.1	4.5	4.2	3.8

- The table shows the mean values of a response variable (e.g. avg number of Facebook posts per day) in each group.
- We would like to know in a single test whether the response variable depends on Y, at some particular significance such as 0.05.

ANOVA

Datamites
Global Institute for Data Science

In ANOVA we compute a single statistic (an F-statistic) that compares variance between groups with variance within each group.

$$F = \frac{VAR_{between}}{VAR_{within}}$$

- The higher the F-value is, the less probable is the null hypothesis that the samples all come from the same population.
- We can look up the F-statistic value in a cumulative F-distribution (similar to the other statistics) to get the p-value.
- ANOVA tests can be much more complicated, with multiple dependent variables, hierarchies of variables, correlated measurements etc.

Lemon Water	Coffee	Red Bull
36	23	18
32	26	12
28	28	10
23	20	12
27	18	11
23	21	13

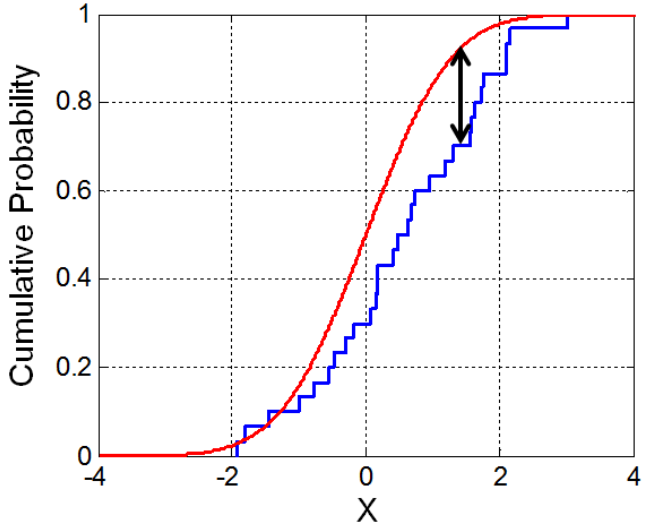
Lemon Water	Coffee	Red Bull
36	33	31
12	10	9
46	38	39
10	12	13
15	13	10
20	19	15

Important Non-Parametric tests

Datamites
Global Institute for Data Science

The Non-Parametric tests make no assumption about the distribution of the input data, and can be used on very general datasets:

- K-S test
- Permutation tests
- Bootstrap confidence intervals



The K-S (Kolmogorov-Smirnov) test is a very useful test for checking whether two (continuous or discrete) distributions are the same.

The K-S statistic is just the **max distance between the CDFs** of the two distributions. While the statistic is simple, its distribution is not! But it is available in most stat packages

Questions to Choose Test

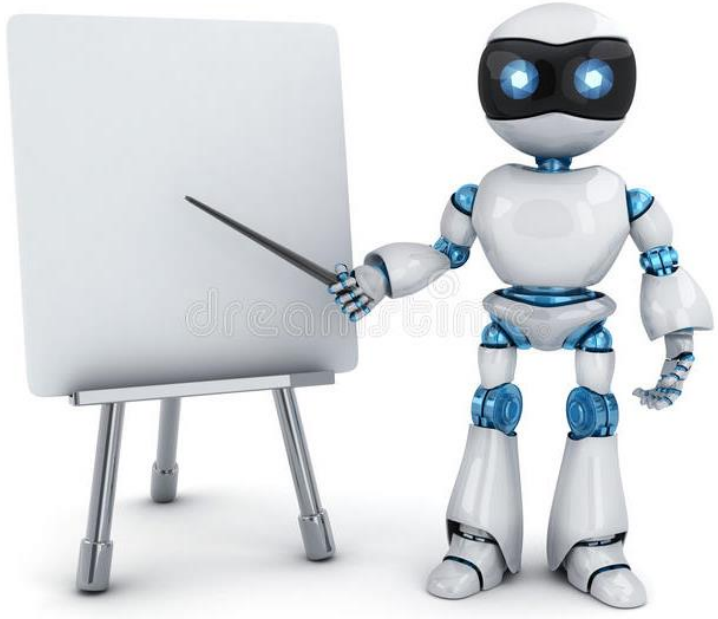
Datamites
Global Institute for Data Science

- What type of Data? (Categorical/Quantitative)
- How many different Groups?
- Comparing Data / Seeking relationship?

Which Test?

Datamites
Global Institute for Data Science

	Compare the data		Seek Relationships
	Categorical Data	Quantitative Data	
One Sample	1 sample proportion	1 sample t	
Two Sample	2 sample proportion	2 sample t	
Two Sample Special		2 sample t Paired t	Correlation/Regression
Three or more Samples		One-Way ANOVA	



Correlation & Regression

Module - 5

Introduction to Regression

It is a technique used to:

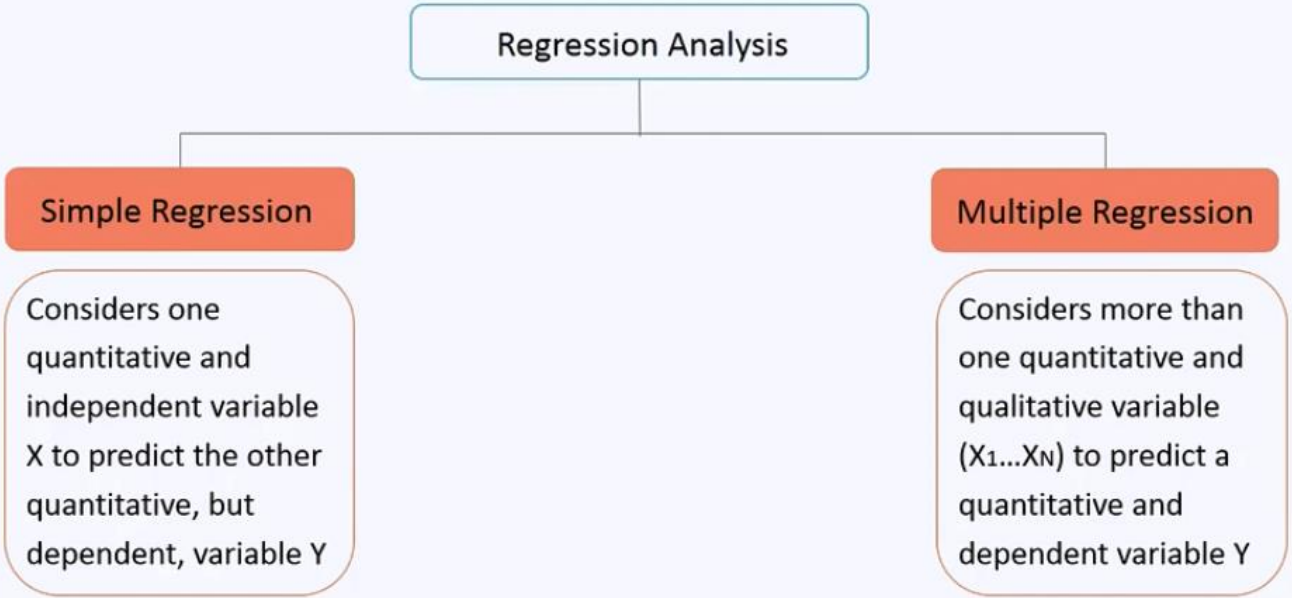
- Estimate a relationship between variables
- Predict the value of one variable (dependent variable) on the basis of other variables (independent variables)

Example:

$Y = \beta_0 + \beta_1 x + \varepsilon$

Here, Y is a dependent variable, whereas β_0, β_1, x, and ε are independent variables.

Types of Regression

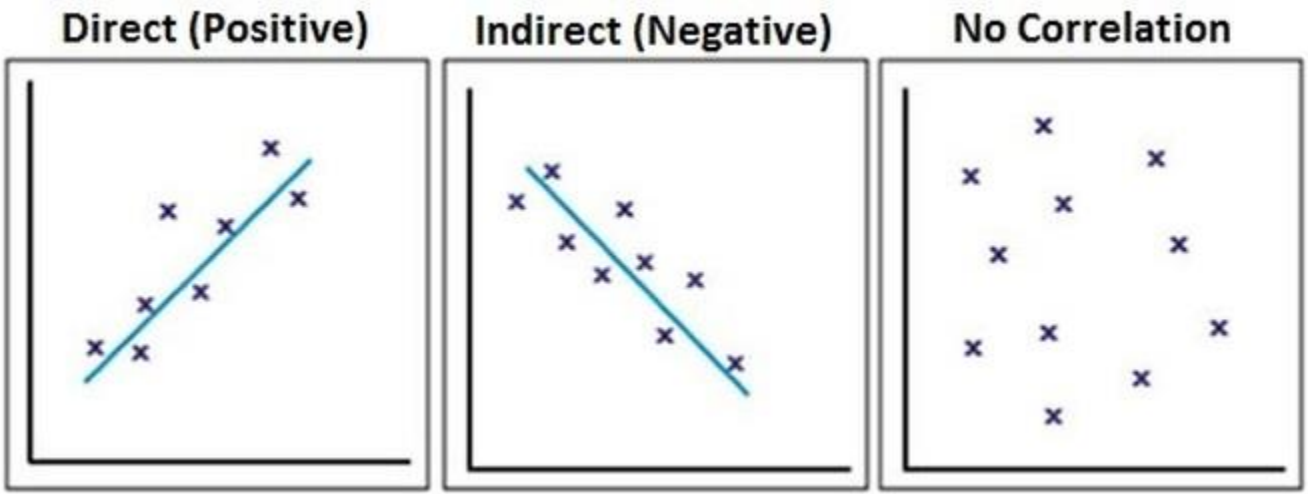


Correlation

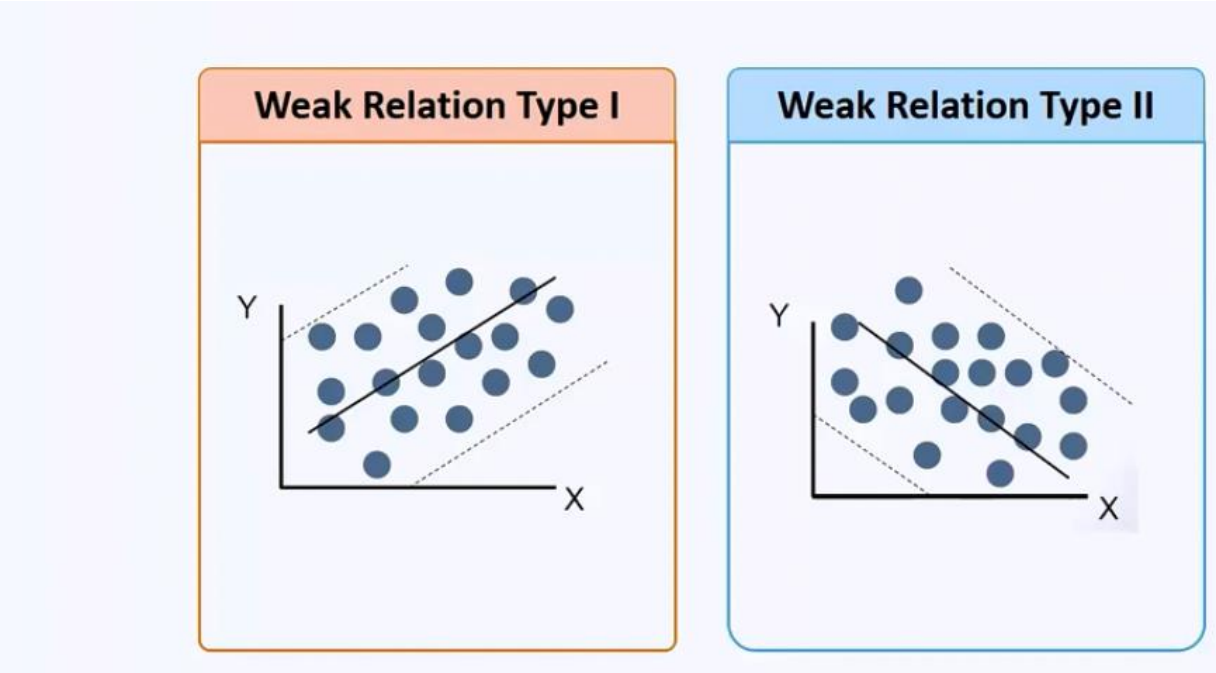
What is Correlation ?

Datamites
Global Institute for Data Science

- **Correlation** is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people.



Weak Correlation



Questions?

Q&A Session

Thank You.