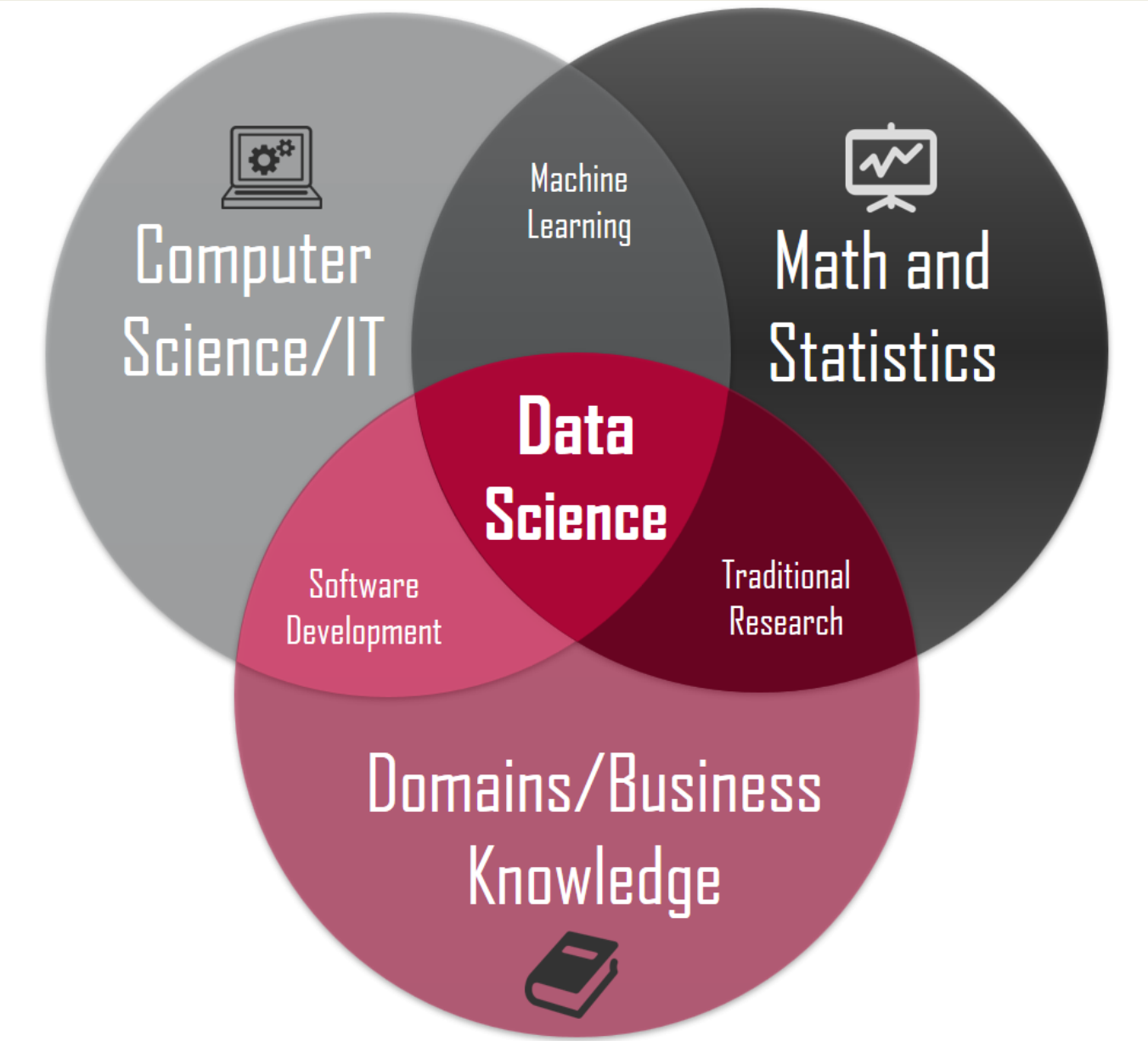
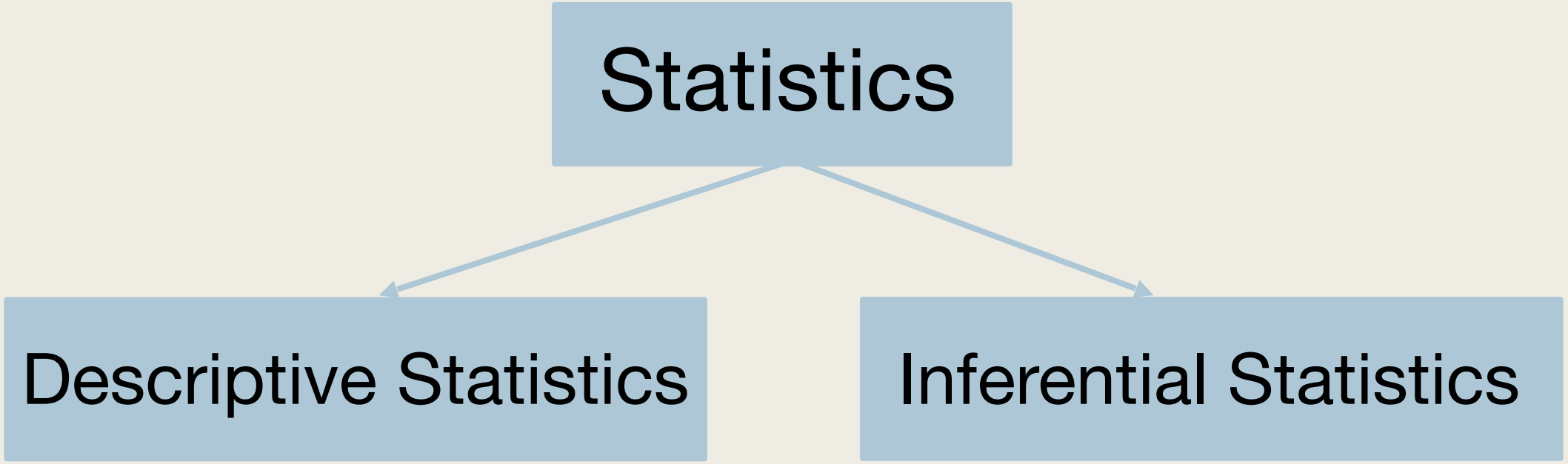
STATISTICS FOR DATA SCIENCE



Source: https://thedatascientist.com/data-science-considered-own-discipline/

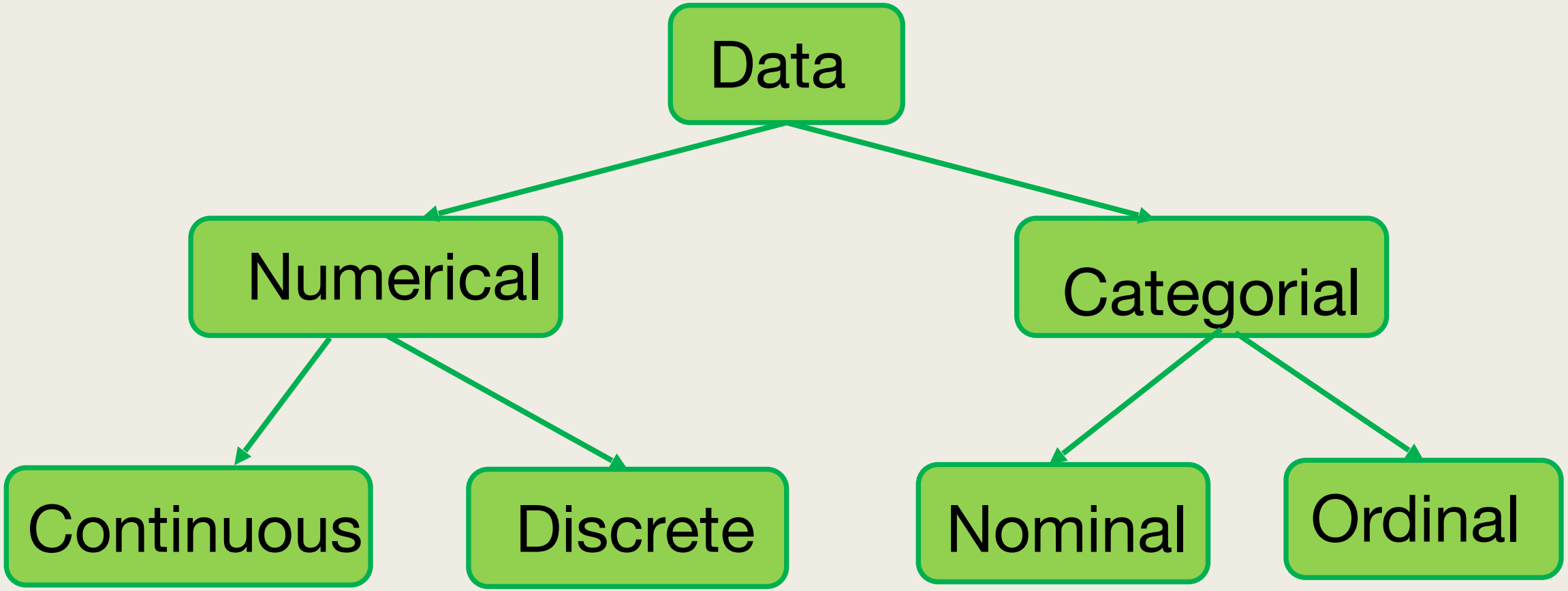
Two types of Statistics



Uses the data to provide description of the population, either through numerical calculations or graphs or tables. For example: maximum, average, minimum.

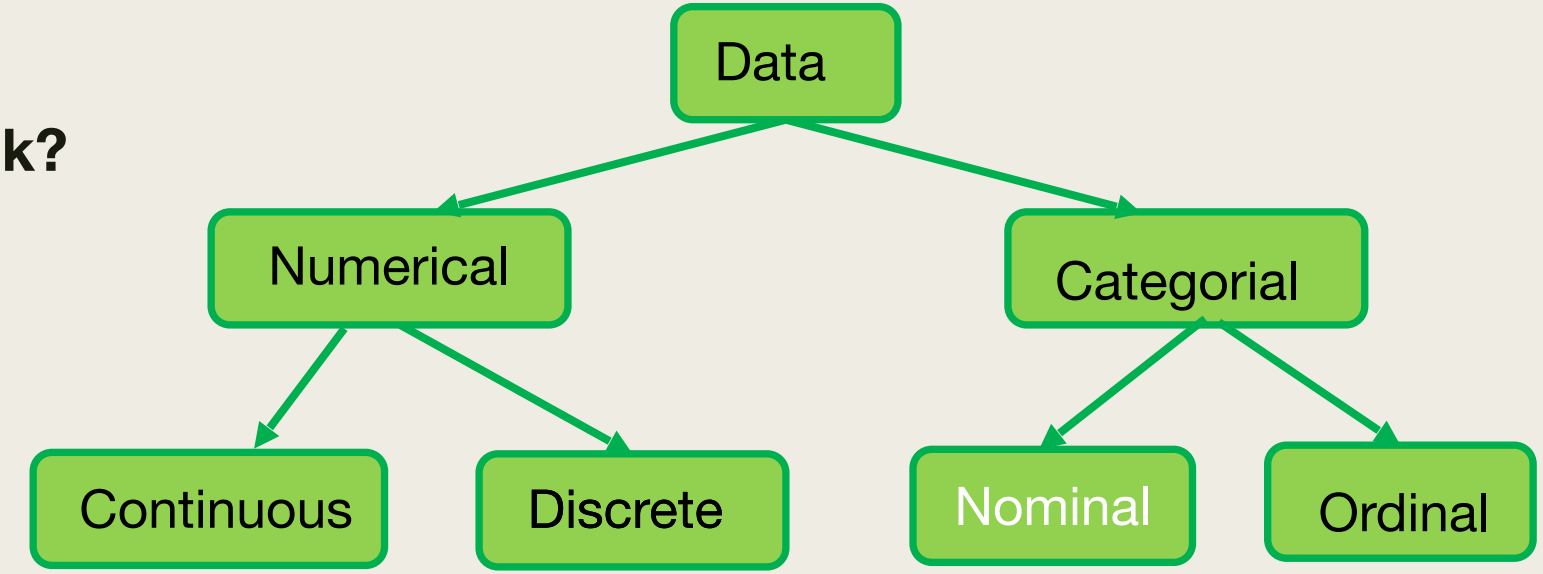
Makes inference or predications about a population based on a sample of data taken from the population in question.

Types of Data - Levels of Measurement



Categorical: Nominal Data

- **Data with no inherent order or ranking such as gender or race, such kind of data called Nominal data.**
- **Gender?**
- Female
- Male
- **languages you speak?**
- English
- French
- Spanish



Categorical: Ordinal Data

- **Data with an ordered series, such as shown in the table, such kind of data is called ordinal data.**

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

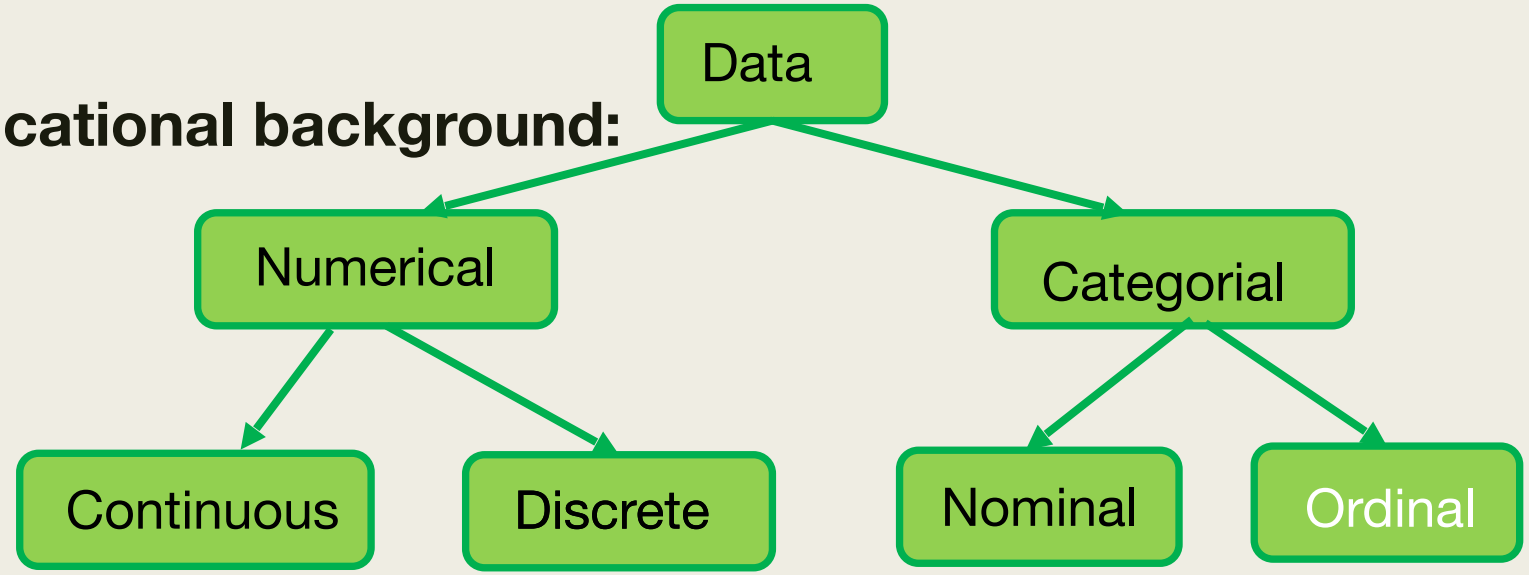
- **Another example, educational background:**

1 - Elementary

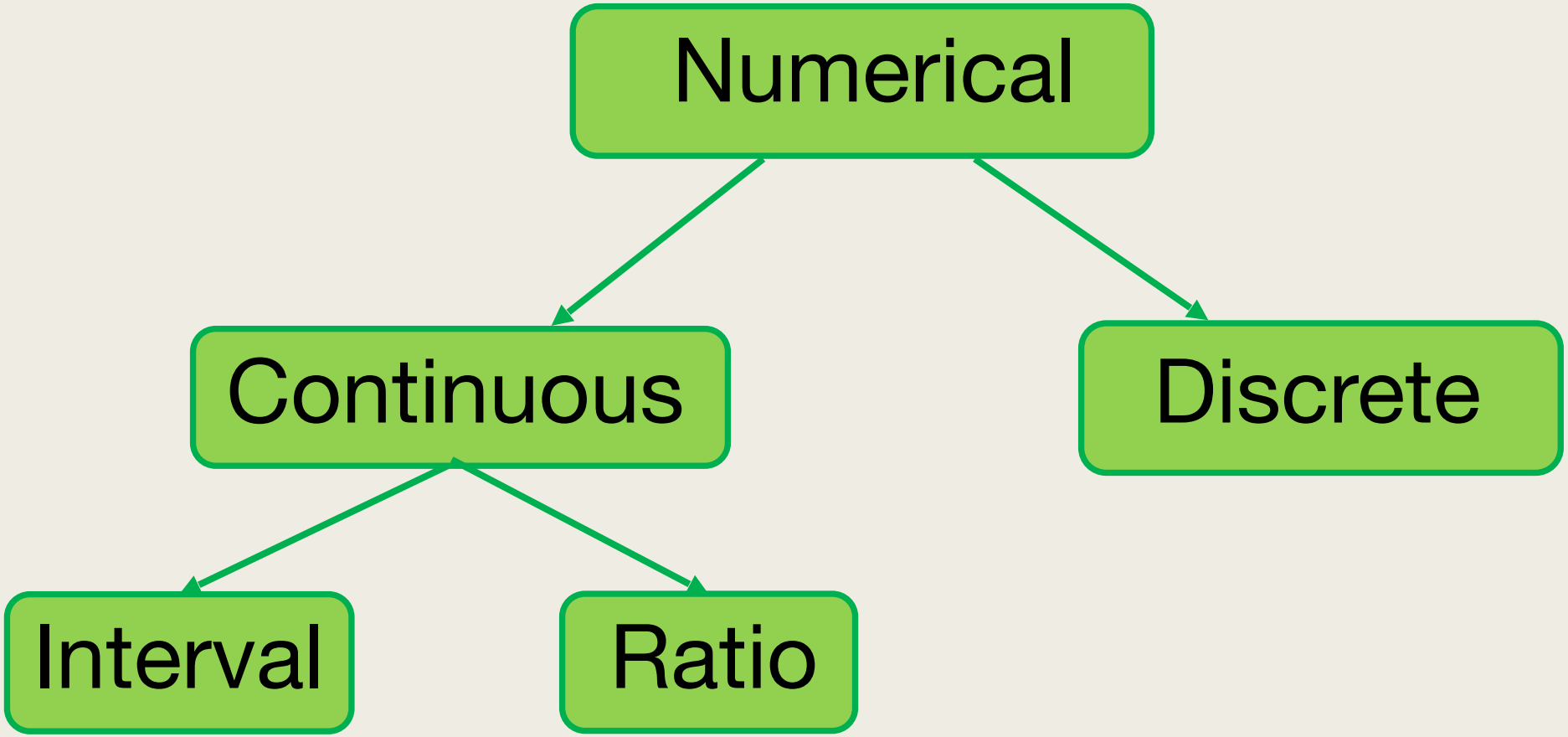
2 - High School

3 - Undergraduate

4 - Graduate

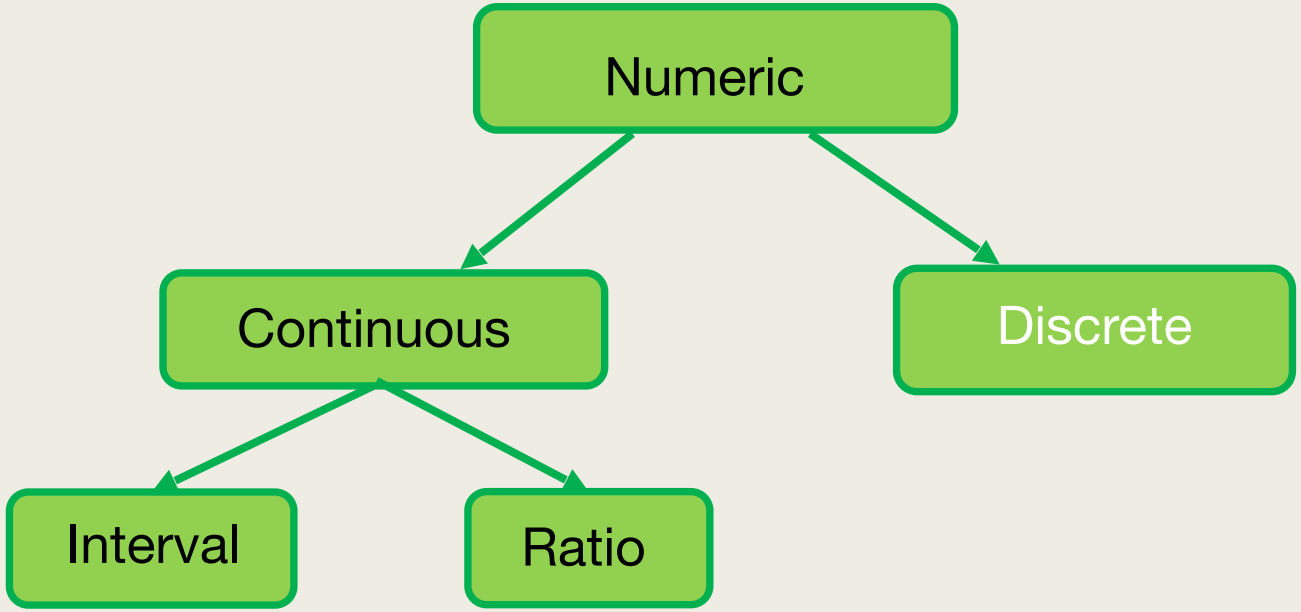


Types of data



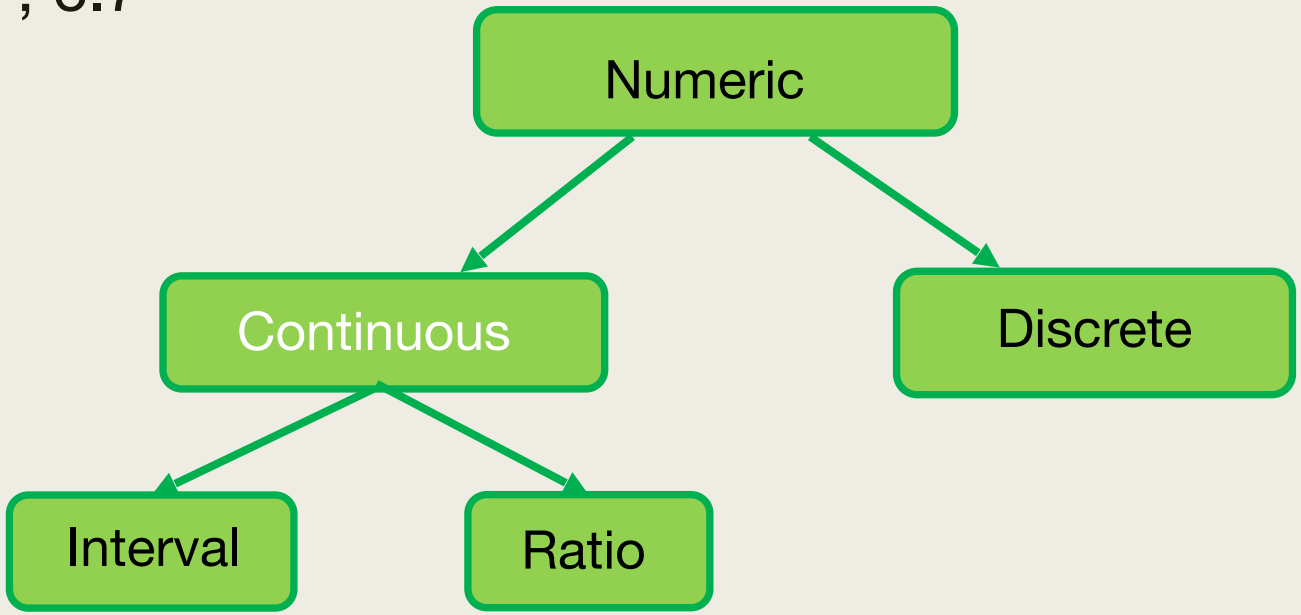
Numerical: Discrete

- Discrete data take on countable and distinct values.
- For example: Number of students in a class.
- Integers



Numerical: Continuous

- Data that can hold infinite number of possible values that can be measured.
- For example a person's weight, distance, speed
- Fractional numbers 2.4, 5.7, 6.7

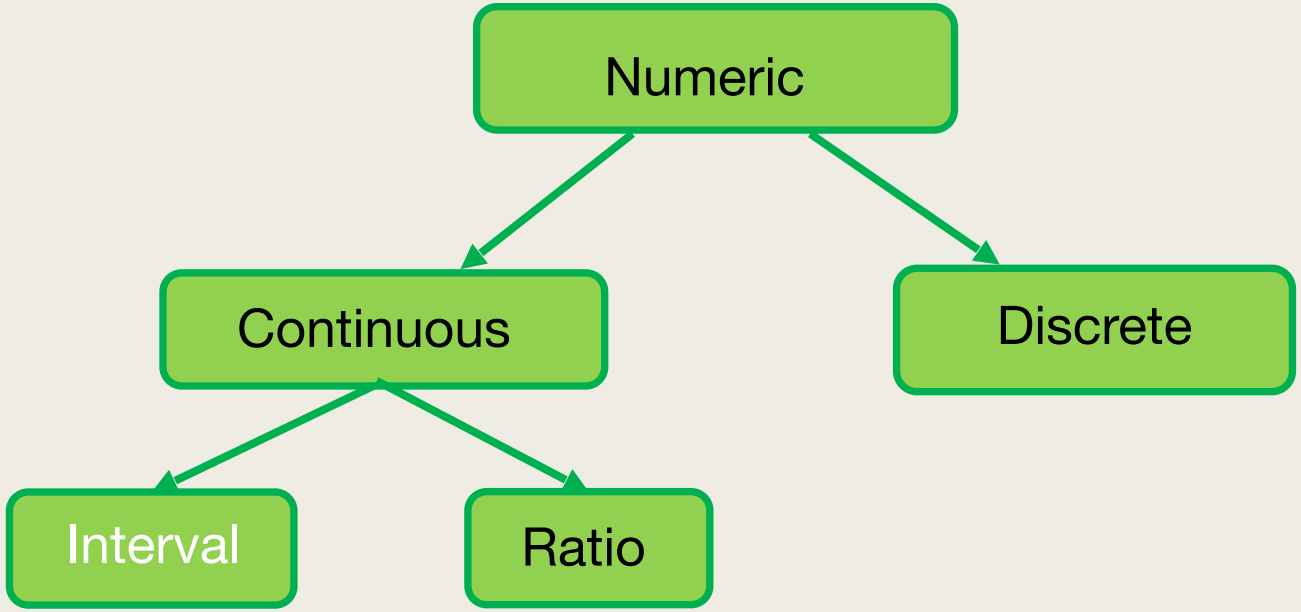


Continuous: Interval data

- Ordered units having the same difference
- No true zero

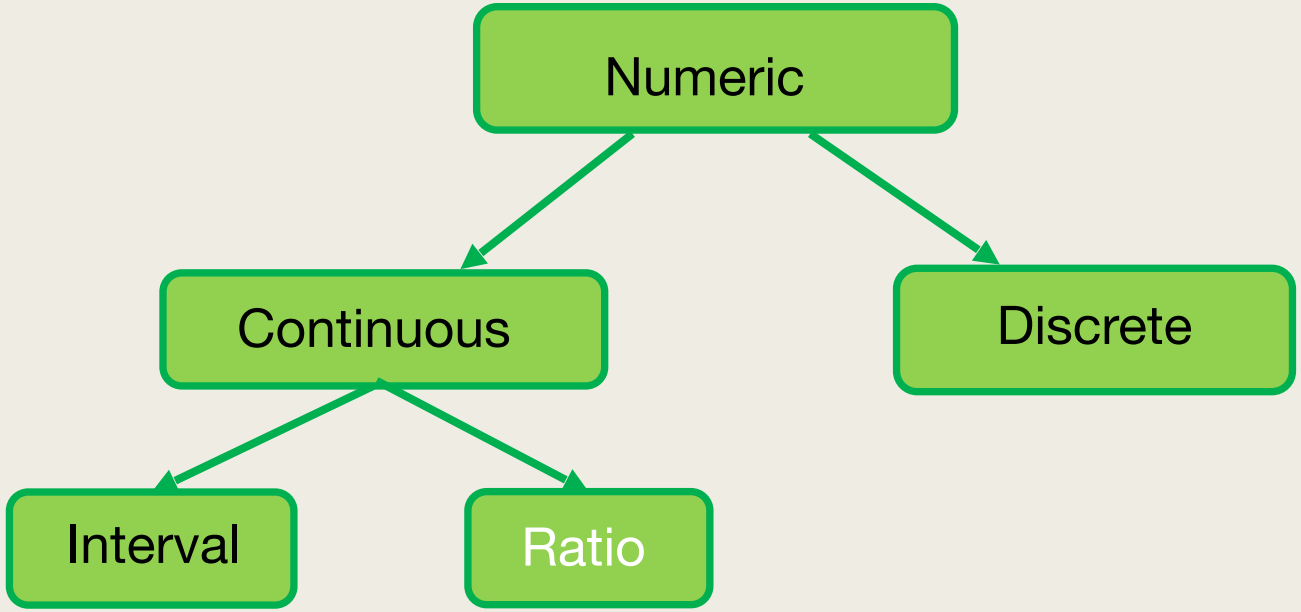
Temperature?

- -11
- -5
- 0
- +5
- +17
- +20

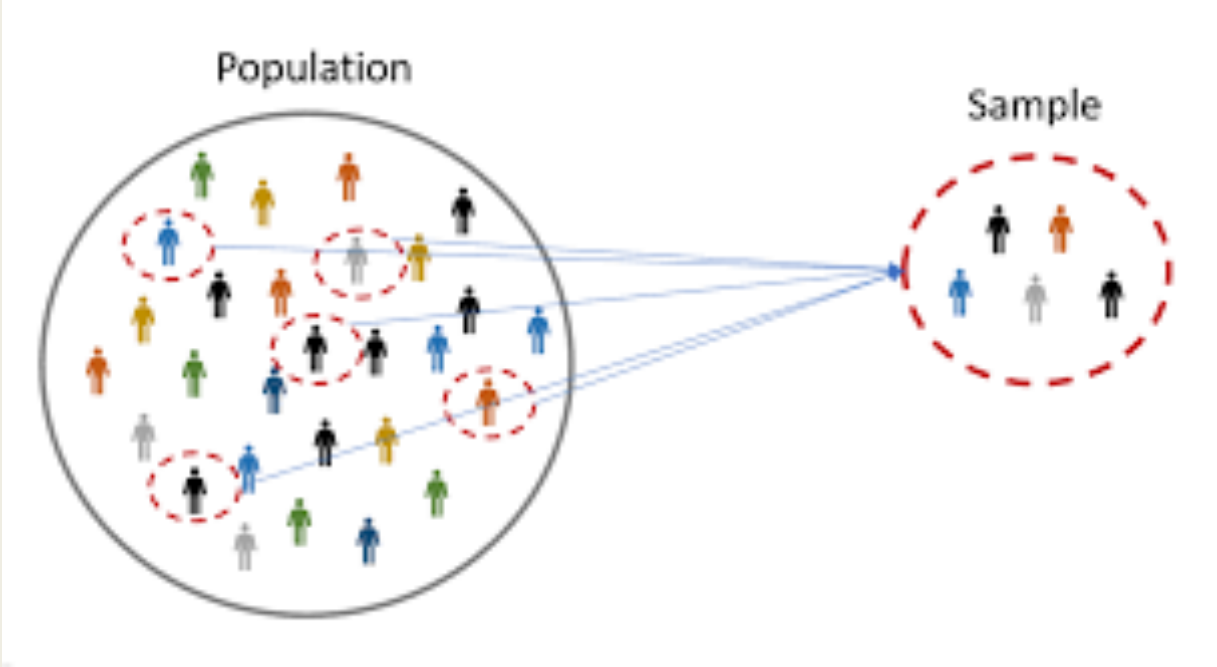


Continuous: Ratio data

- Same as interval data, but…
- Has a true zero
- Length(inch):
- 0
- 5
- 12
- 16



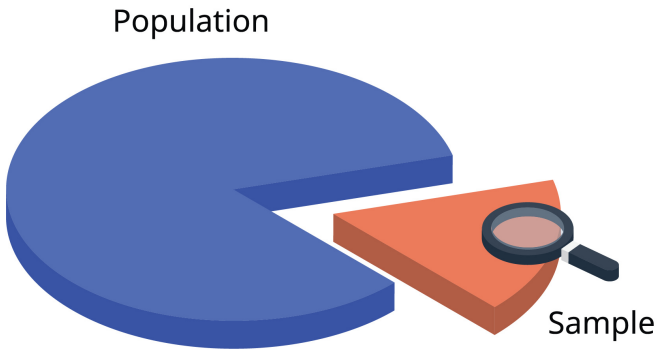
Population and Sample:



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

Sample: A subset of population is called ‘sample’. A well chosen sample will contain most of the information about a particular population parameter.

VS



Descriptive Statistics

Inferential Statistics

Measure of Centrality

mean, median, mode

outliers

- An extreme value in the dataset compared to all other values.
- Exam scores of 5 students
- Can you identify the outlier?

[40, 87, 88, 90, 95]

Mean

- Average of the data
- Formula:

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} = \frac{x_1 + x_2 \ldots + x_n}{n}$$

- Greatly affected by ***outliers***

Median

- **Middle value of a sorted dataset**
- **Formulas:**

If *n* is odd the median is the value at position p where

$$p = \frac{n+1}{2}$$

If *n* is even the median is the average of the values at positions *p* and *p + 1 where*

$$p = \frac{n}{2}$$

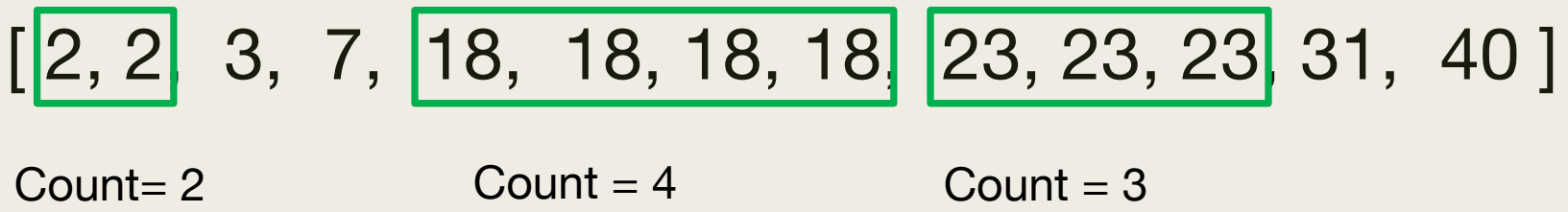
$$\tilde{x} = \frac{x_p + x_{p+1}}{2}$$

***n* is the size of the ordered data set**

- **Median performs well in the presence of outliers**

Mode

- Value that occurs most often in a dataset



Mode is **18**

- Not affected by outliers
- A dataset can have two or more modes, and can also have no mode

Example

With outliers, odd size of dataset

[40, 87, 88, 90, 95]
Outlier: 40

Mean:	Median:	Mode:
$\frac{40+87+88+90+95}{5} = \mathbf{80}$	40, 87, 88, 90, 95	There is no mode.

Without outliers, even size of data set

[87, 88, 90, 95]

Mean:	Median:	Mode:
$\frac{87+88+90+95}{4} = \mathbf{90}$	87, 88, $\frac{88+90}{2}$, 95 = **88.5**	There is no mode.

summary

- Mean is a good measure for a dataset **without** outliers
- Median is a good measure for a dataset **with** outliers
- Mode is a good measure to use when we have categorical data

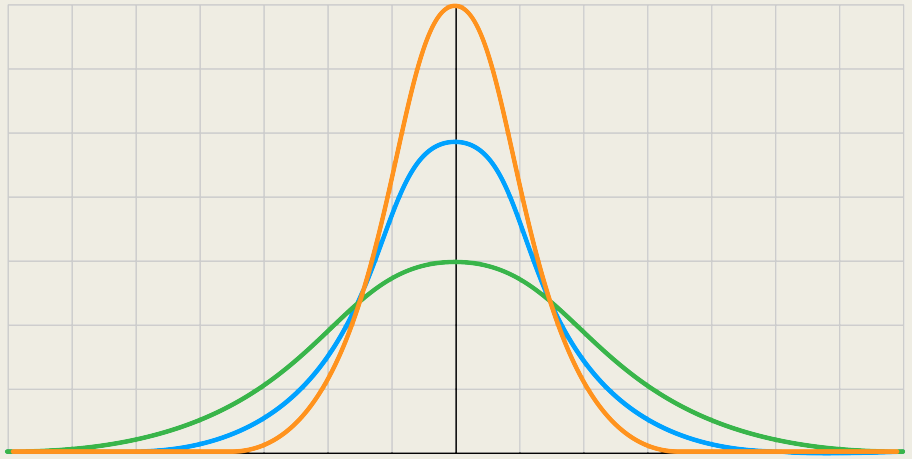
Measure of Dispersion

Variance, Standard deviation, Covariance

Variance

- **Measure the spread of data around the mean**
- **Variance formula:**

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance	s^2 = sample variance
x_i = value of i^{th} element	x_i = value of i^{th} element
μ = population mean	$\bar{x}$ = sample mean
N = population size	n = sample size



Sample variance vs Population Variance

- Data Scientists mostly deal with sample data rather than the whole population data
- The variance of the sample data is usually greater than the variance of the population data
- Dividing by n-1 in the variance sample formula compensates for the lack of information about the population data

Standard deviation

- Variance is too large for any visualization/ comparison
- that's why we need standard deviation
- Standard Deviation Formula:

$$s = \sqrt{\frac{\sum_{i=0}^{n} (X_i - \bar{x})^2}{n - 1}}$$

s = sample standard deviation
n = the size of the sample
x_i = each value from the sample
$\bar{x}$ = the sample mean

Coefficient of Variation (CV)

- Dispersion of data relative to its mean

$$CV = \frac{Standard\ Deviation\ (s)}{Sample\ Mean\ (\bar{x})}$$

- Helpful when comparing two datasets
- In most fields, **lower values for the coefficient of variation are considered better** because it means there is less variability around the mean. The lower the value of the coefficient of variation, **the more precise the estimate**.

Example of Coefficient of Variation

- **Exam scores for two different classes**

Class1 scores	Class2 scores
70	98
79	55
50	68
86	98
94	66
100	83
62	50

Class1:

Standard deviation: 16.49

Mean: 77.28

CV: (16.49/77.28) *100 = **21.34%**

Class2:

Standard deviation: 18.007

Mean: 74

CV: 18.007/74 = 0.2433 * 100 = **24.33%**

Another example:

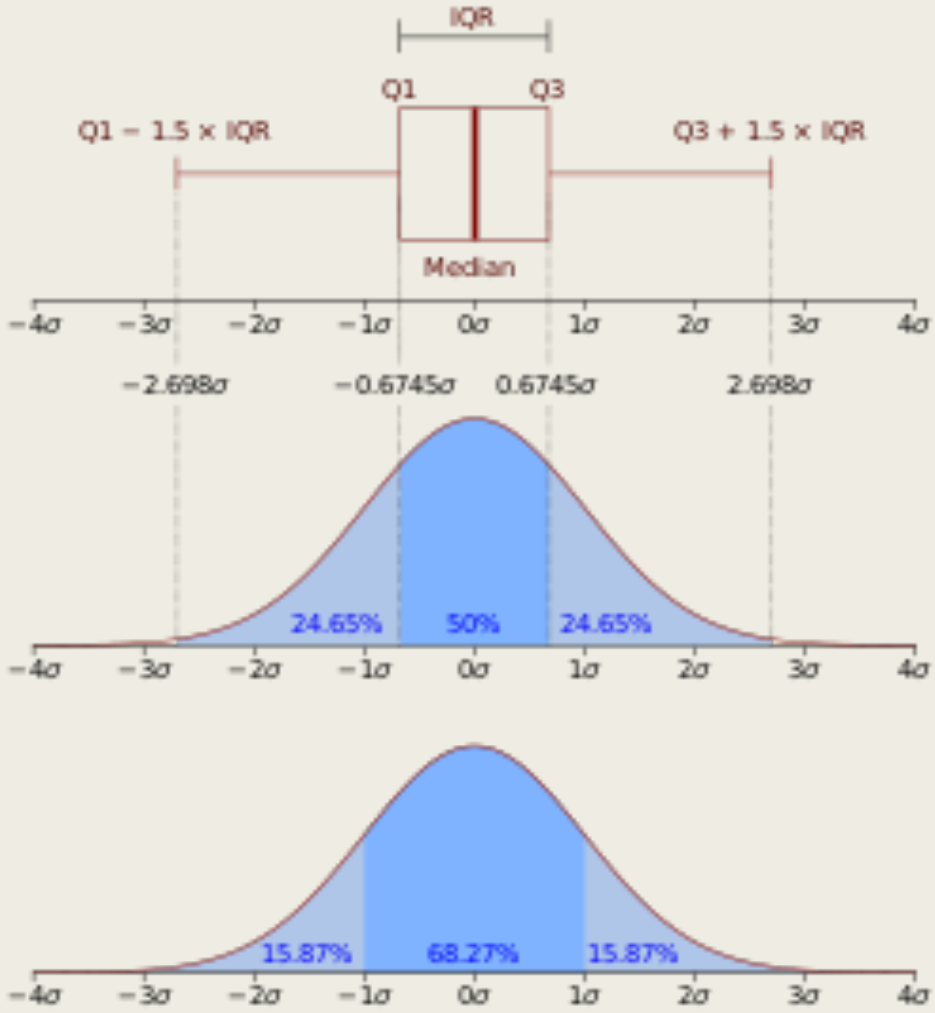
- Consider the following mean weekly sales and standard deviation of weekly sales for two different companies:
  - Company A: Mean Weekly Sales = $4,000, Standard Deviation = $1,500
  - Company B: Mean Weekly Sales = $8,000, Standard Deviation = $2,000
- We can calculate the coefficient of variation for each store:
  - CV for Company A: $1,500 / $4,000 = 0.375
  - CV for Company B: $2,000 / $8,000 = 0.25

Since Company B has a lower CV, it has lower volatility in weekly sales relative to the mean compared to company A. This means Company B can likely predict their weekly sales with more certainty than Company A.

Range , Quartiles & Interquartile Range

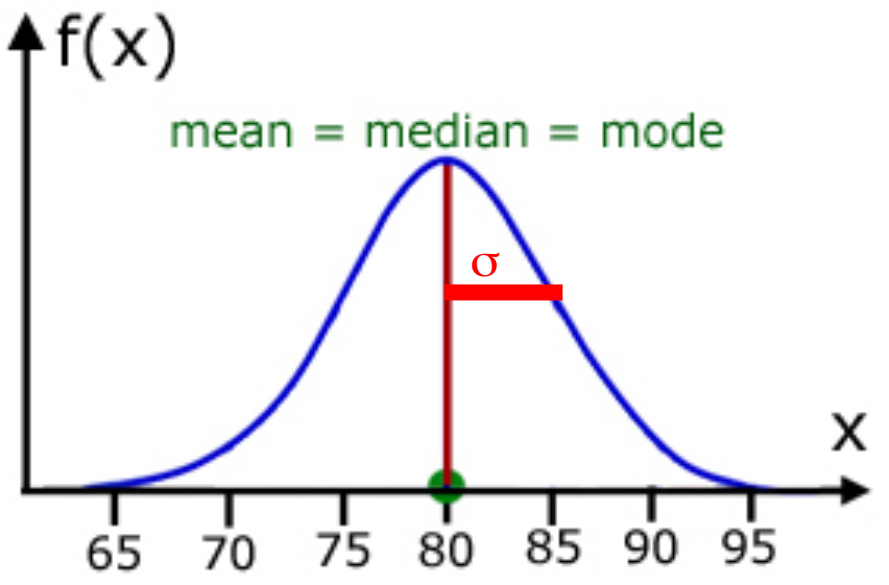
- **Range** - the distance between the maximum and minimum values
- **Quartiles** - the values that separate each quarter.
- **Interquartile Range** - the difference between the 1st and 3rd quartiles.

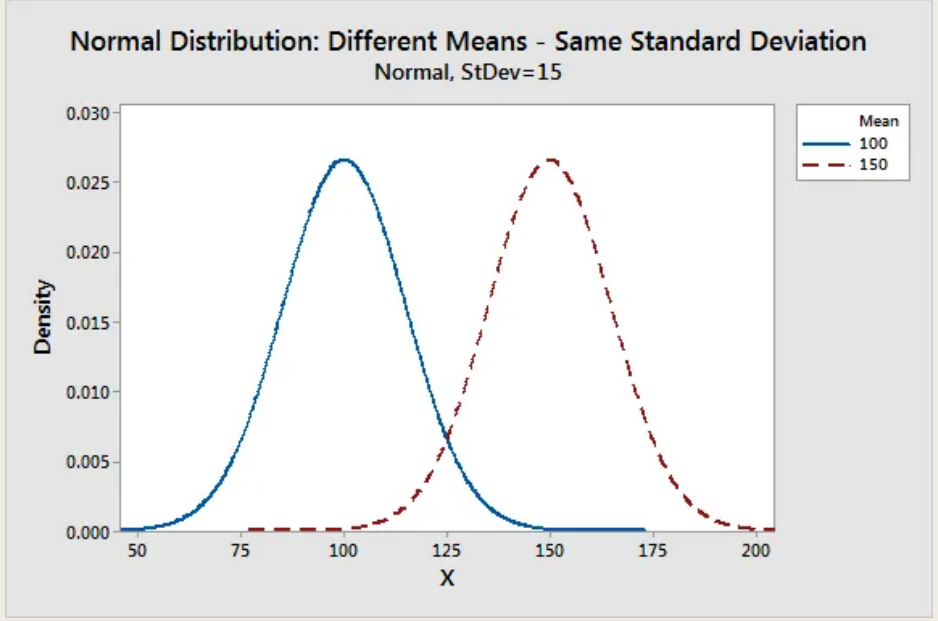
Source: https://en.wikipedia.org/wiki/Interquartile_range

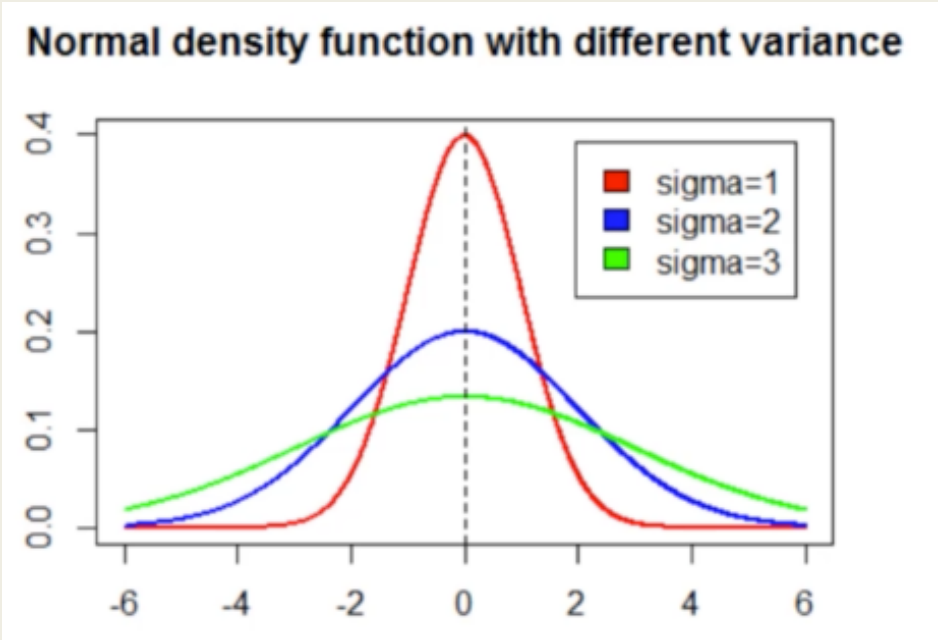


Normal Distribution

- Also known as Gaussian Distribution
- Bell Curve
- Mean = median = mode





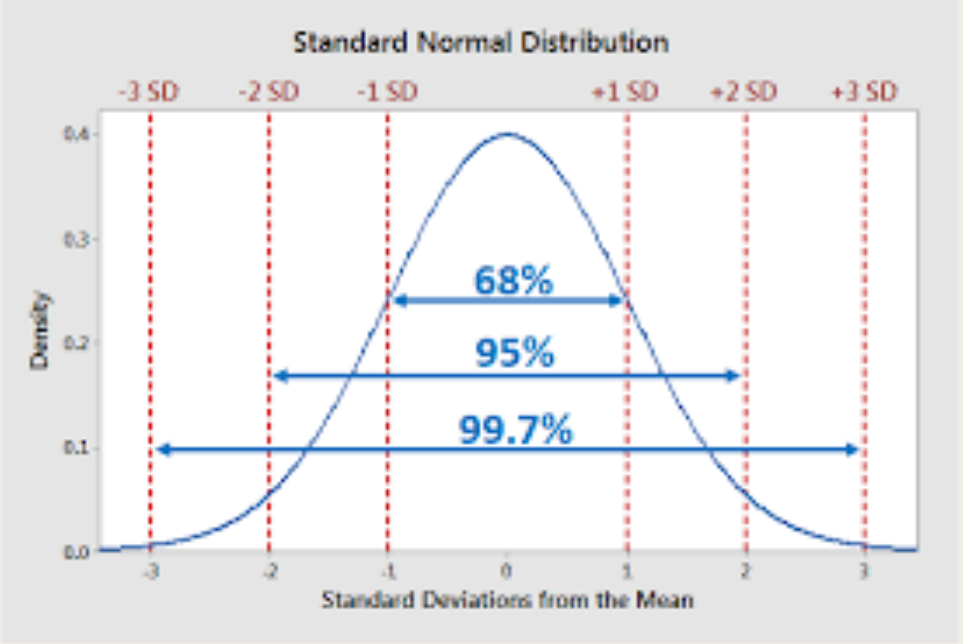


Standard Normal Distribution

- Mean = 0
- Standard deviation = 1

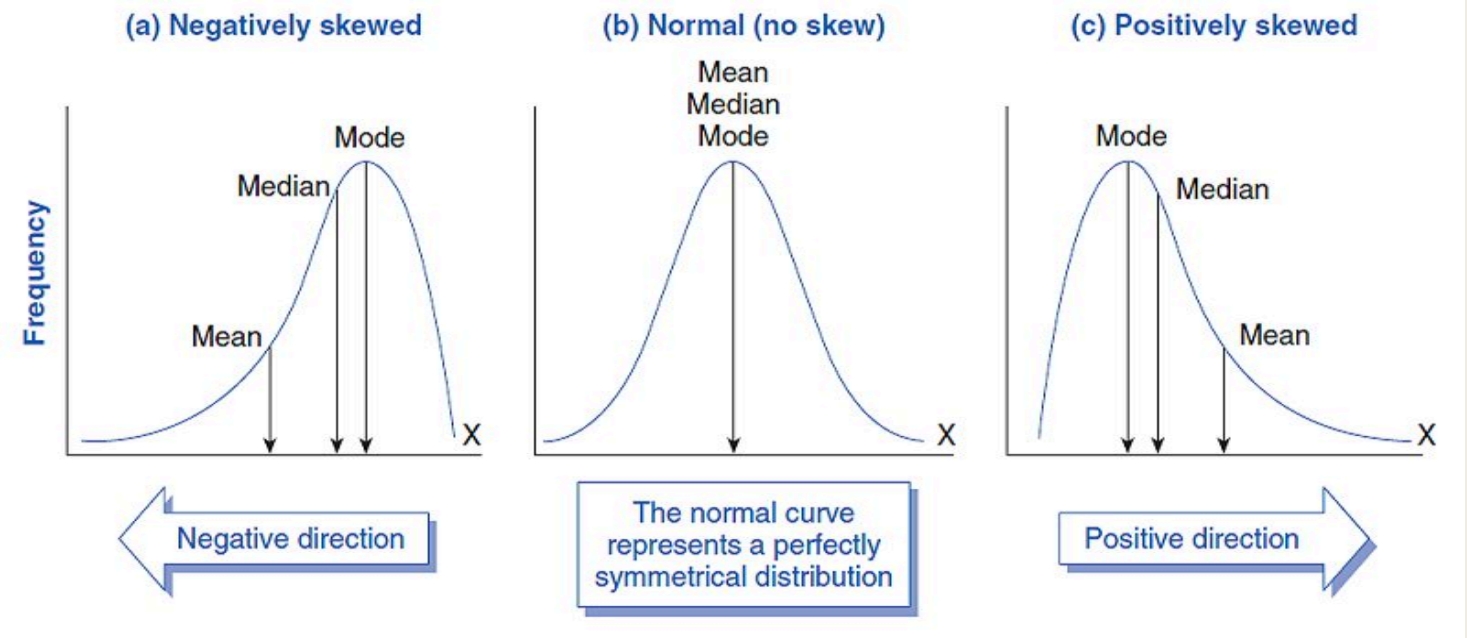
$$z = \frac{x - \mu}{\sigma}$$

x = raw score
μ = mean
σ = standard deviation (std)



Measures of Asymmetry

Skewness: measures asymmetry (imbalance) from the mean of a data distribution



- Negative/left Skewed: In negative skewness, the extreme data values are smaller, which decreases the mean value of the dataset or the negative skew distribution is the distribution having the tail on the left side. In Negative Skewness: Mean < median < mode
- Positive/Right Skewed: In positive skewness, the extreme data values are larger, which in turn increase the mean value of the data set, or in the simple term in positive skew distribution is the distribution having the tail on the right side. In Positive Skewness: Mean > Median > Mode

Measures of Asymmetry

Kurtosis: measure of whether the data is heavy tailed or light tailed relative to a normal distribution.

- A **platykurtic** distribution shows a negative kurtosis. The flat tails indicate the **small outliers** in a distribution.
- Data that follows a **normal** distribution shows an excess kurtosis of zero or close to zero.
- **Leptokurtic** indicates a positive kurtosis. The distribution shows heavy tails on either side, indicating **large outliers**.

https://corporatefinanceinstitute.com/resources/data-science/kurtosis/

Other Distributions

- **t-distribution** - similar to a normal distribution but with fatter tails.
- **Bernoulli distribution** - a discrete probability distribution of the outcomes of a single experiment with a yes-no question. Where *p* is the probability of yes (or 1) and 1 - *p* is the probability of no (or 0).
- **Binomial distribution** - a discrete probability distribution of the outcomes of a n independent experiments with a yes-no question.

Central Limit Theorem

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

- Suppose that you draw a random sample from a population and calculate a statistic for the sample, such as the mean.
- Now you draw another random sample of the same size, and again calculate the mean.
- You repeat this process many times, and end up with a large number of means, one for each sample.
- The distribution of the sample means is an example of a **sampling distribution**.

*The central limit theorem says that the sampling distribution of the mean will always be **normally distributed**, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.*

The mean of the sampling distribution is the mean of the population.

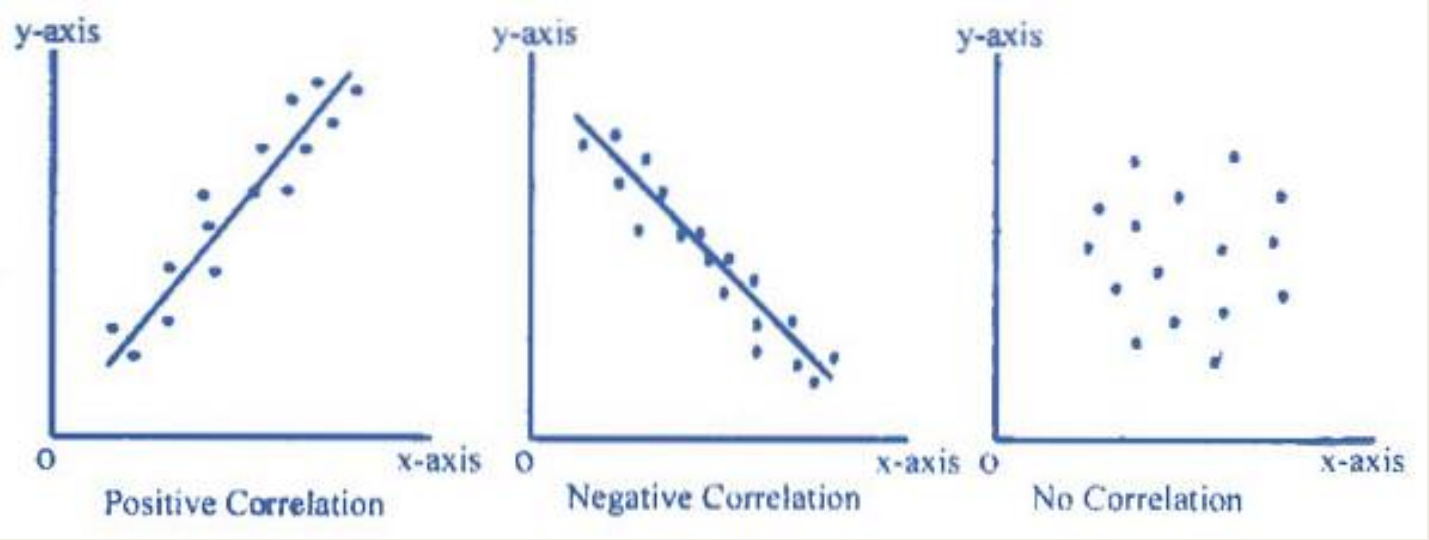
$$\mu_{\bar{x}} = \mu$$

Statistics How To

Measures of Variable Relationship

covariance

- **Measures the directional relationship between two variables**
- **Three types:**
  - **Zero COV:** no relation between the variables
  - **Positive COV:** if one variable increases, the other variable increases too
  - **Negative COV:** if one variable increases, the other variable decreases



Covariance Formula:

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$COV(X,Y)$ = *covariance between variable X and Y*

X_i = *data value of X*

Y_i = *data value of Y*

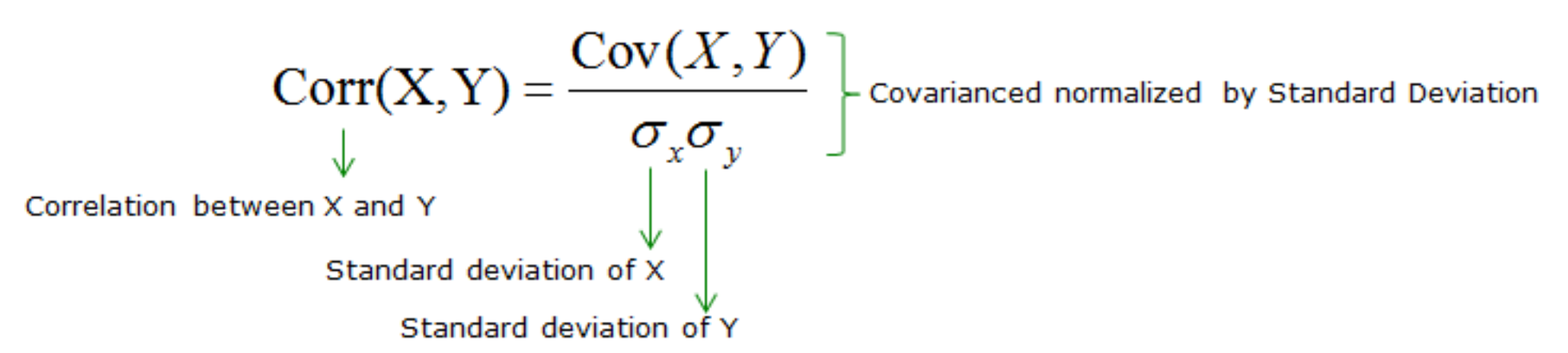
$\bar{X}$ = *mean of X*

$\bar{Y}$ = *mean of Y*

n = *number of data values*

Correlation

- **Covariance** explains the type of relationship between two variables**.** How strong this relationship is can be explained by **correlation.**
- The correlation coefficient is measured on a scale that varies from + 1 through 0 to – 1. for example, 0.8 indicate a strong positive correlation, where -0.9 indicates a strong negative correlation. While a coefficient of -0.3 or lower indicates a very weak correlation.
- A weak positive correlation indicates that, although both variables tend to go up in response to one another, the relationship is not very strong. A strong negative correlation, on the other hand, indicates a strong connection between the two variables, but that one goes up whenever the other one goes down.
- **Formula:**



Example

- Consider the following two variables

Temperature(X)	No. Of Ice-cream sold(Y)
98	20
85	16
87	15
64	7
66	8
73	11

- Mean of X = **78.83**. std= 13.3479. CV= 0.1693
- Mean of Y = **12.833**. std= 5.036. CV= 0.39245

(x-Xmean)	(y – Ymean)	product(x-Xmean)(y – Ymean)
98-78.83= 19.17	20-12.83= 7.17	137.4486
85-78.83= 6.17	16-12.83= 3.17	19.5589
87-78.83= 8.17	15-12.83= 2.17	17.7289
64-78.83= -14.83	7-12.83= -5.83	86.4589
66-78.83= -12.83	8-12.83= -4.83	61.9689
73-78.83= -5.83	11-12.83= -1.83	10.6689
	Sum:	333.7889

COV(X,Y)= sum/n-1 = 333.7889/5 = **66.75**

(x-Xmean)	(y – Ymean)
367.48	51.40
38.06	10.04
66.74	4.708
219.92	33.98
164.60	23.32
33.98	3.348
Sum: 890.78	126.796

$\sigma x = \sqrt{890.78/5} =$ 13.34

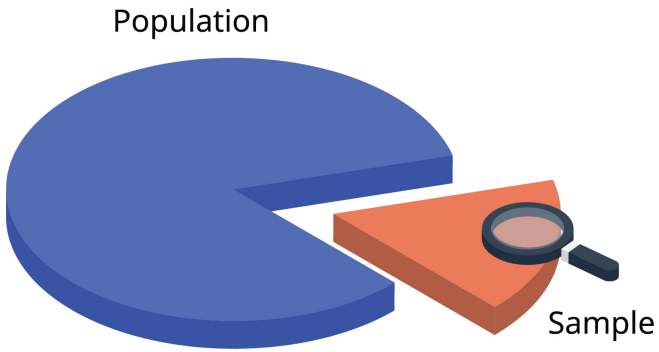
$\sigma y = \sqrt{126.796/5} =$ 5.03

Strong correlation

Correlation = COV(X,Y)/ σx σy = 66.75/(13.34x 5.03) = 0.9947

Jupyter Notebook Descriptive Statistics

VS



Descriptive Statistics

Inferential Statistics

Steps to Make Statements about a Population

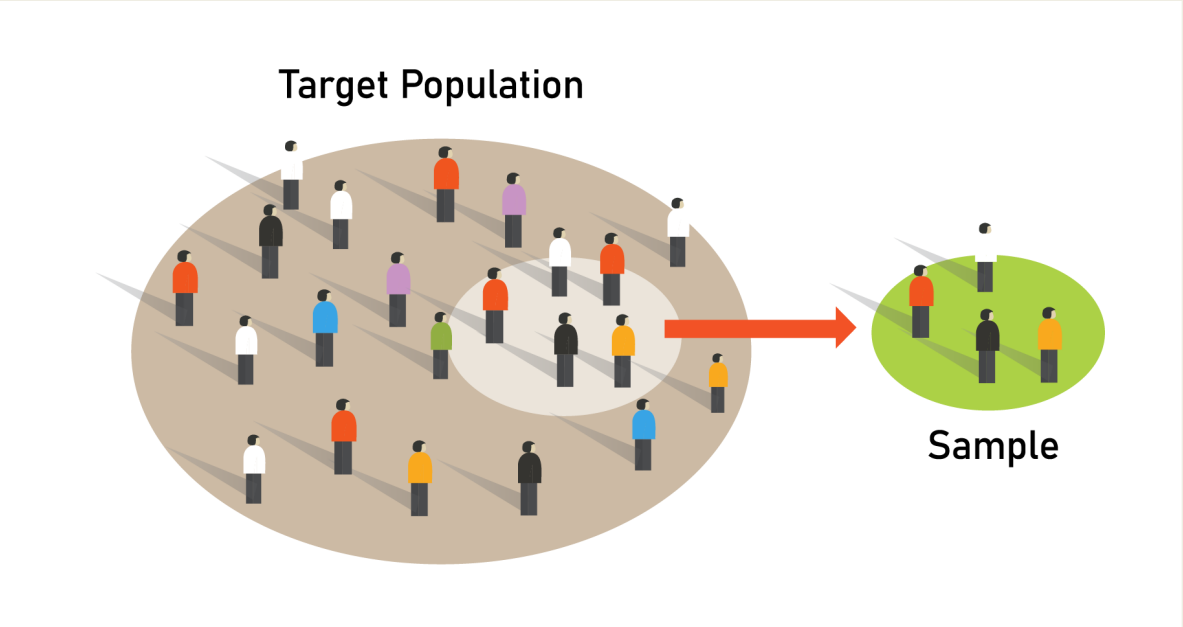
1. **Hypothesis**
2. **Population and Sample**
3. **Hypothesis Testing**
4. **p-Value**
5. **Errors**

1. Hypothesis

- Alternative hypothesis (H_1) - the hypothesis we are trying to find evidence for. Cannot be tested directly with a hypothesis test.
  - e.g. a certain drug has an effect on blood pressure
- Assume the Null (Opposite) hypothesis (H_0)
  - e.g. the drug has no effect on blood pressure

2. Population and Sample

- Two ways to collect data:
- Collect all the data
- Collect a sample from the data
- Sampling techniques: using probability, and non-probability.
- Non-probability: convenience, consecutive, judgement, quota
  - Not useful for inferential statistics



Probability sampling techniques:

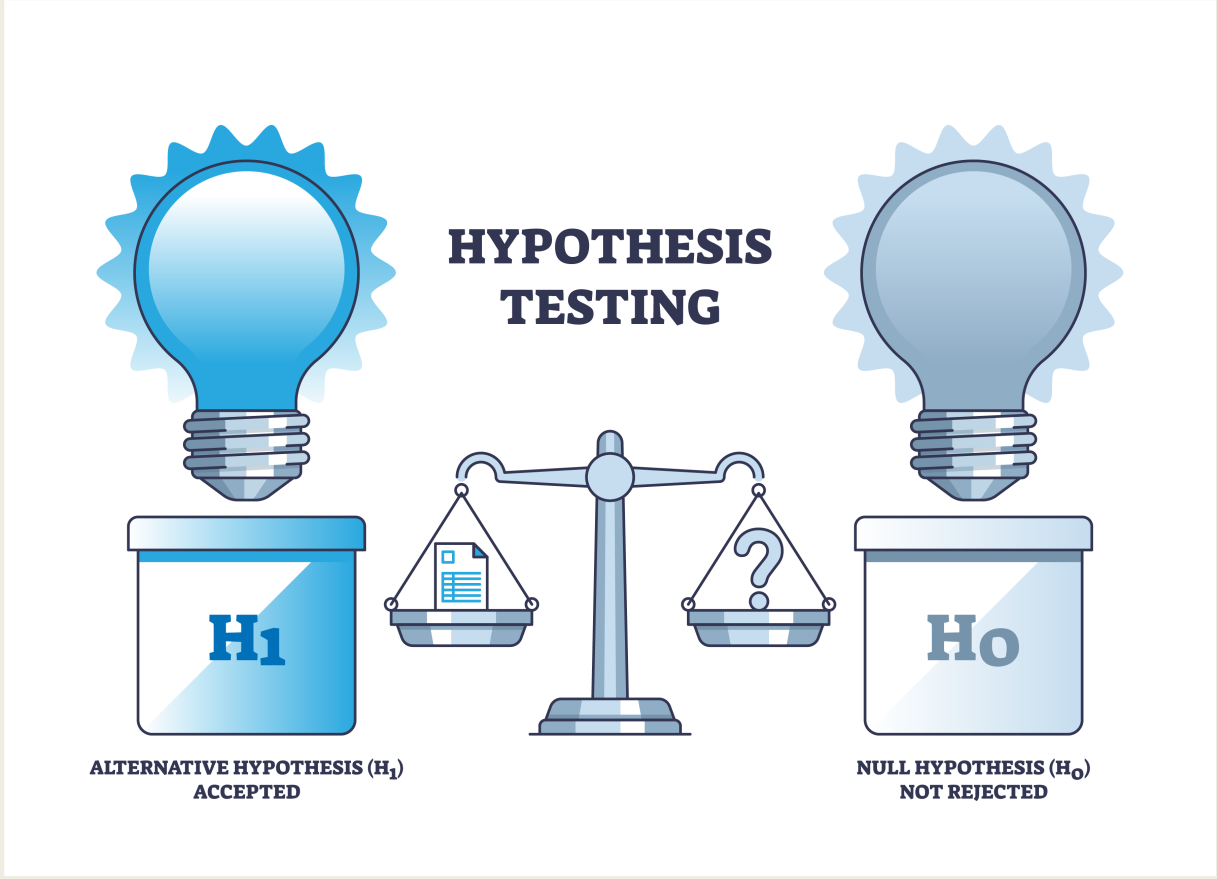
- **Random sampling:** each member of the population has equal chance of being selected.

- **Systematic sampling:** every nth record is chosen from the population to be a part of the sample. For example, every second record is chosen.

- **Stratified sampling:** a stratum is a subset of the population that shares at least one common characteristic, such as gender for example. Random sampling is then used to select a sufficient number of subjects from each stratum.

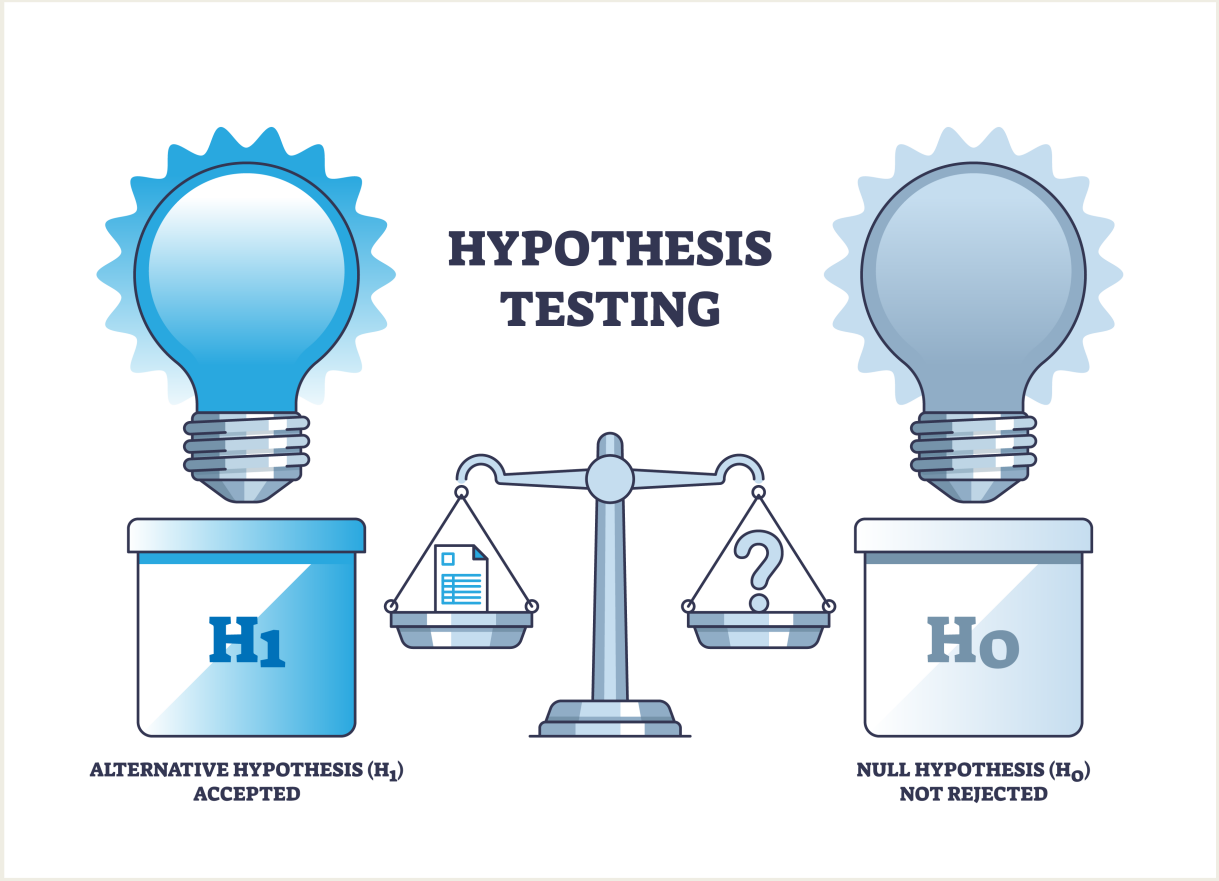
3. Hypothesis Testing

- Measure how much the sample deviates from the null hypothesis.
- Assuming the null hypothesis, how likely is it to draw a sample with such a deviation?
- Common parametric hypothesis tests:
  - t-Test
  - Chi²-Test
  - ANOVA



4. p-Value

- Indicates the probability of drawing a sample that deviates as much or more as our sample.
- Assuming the null hypothesis is true



5. Significance

- At what point is the probability (p-Value) small enough that we can reject the null hypothesis?
- If the p-Value is less than a *pre-determined threshold*, the result is considered statistically significant.
  - The result is unlikely to have occurred by chance alone.
  - We can reject the null hypothesis

$$??? < 0.05$$

5. Error

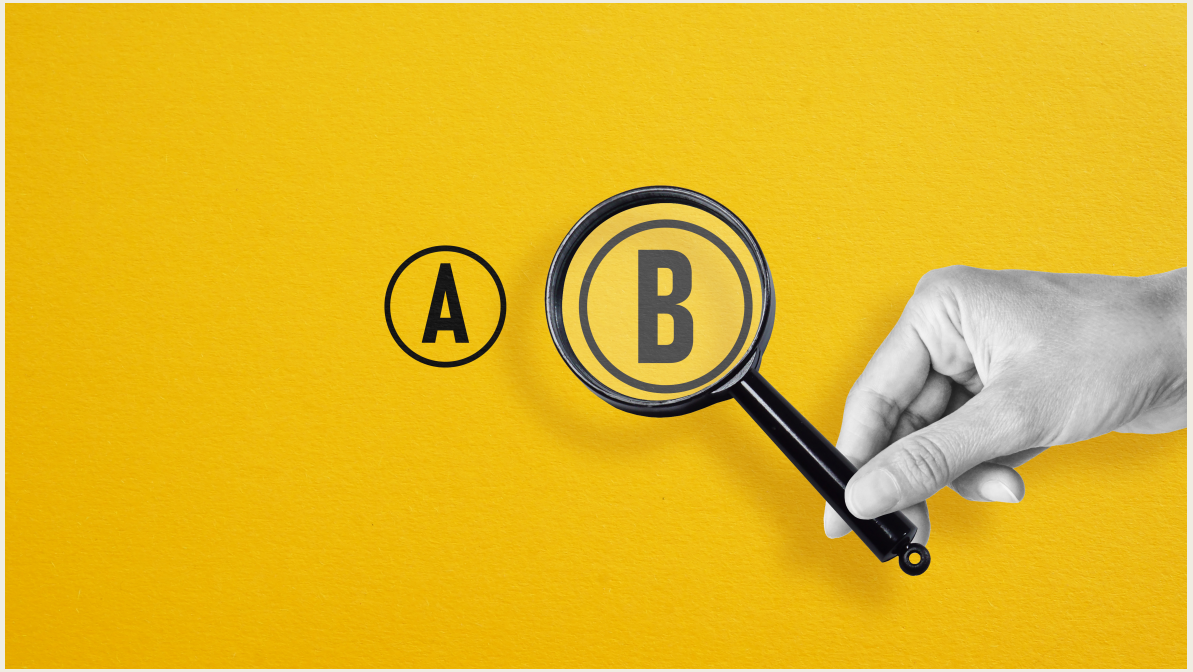
- A p-value below the threshold does not prove that the alternative hypothesis is true.
- Only that it is unlikely to get such a result if the null hypothesis is true.
- Likewise, a p-value greater than the threshold does not prove that the null hypothesis is true.

		In **reality** Null hypothesis is:	
		TRUE	**FALSE**
Decision about null hypothesis	**Accepted**	Correct	Type II error
	Rejected	Type I Error	Correct

t-Test

Analyzes whether there is a significant difference between the mean values of two groups.

- **Types**
  - *One sample t-Test* - compare the mean of a sample with a known reference mean.
  - *Independent samples t-Test* - compare the means of two independent samples.
  - *Paired samples t-Test* - compare the means of two dependent groups - different measurements of the same subject at two different times
- **Assumptions**
  - Variables are metric
  - Normal distribution
  - Variances in groups must be approximately equal (independent samples t-test)

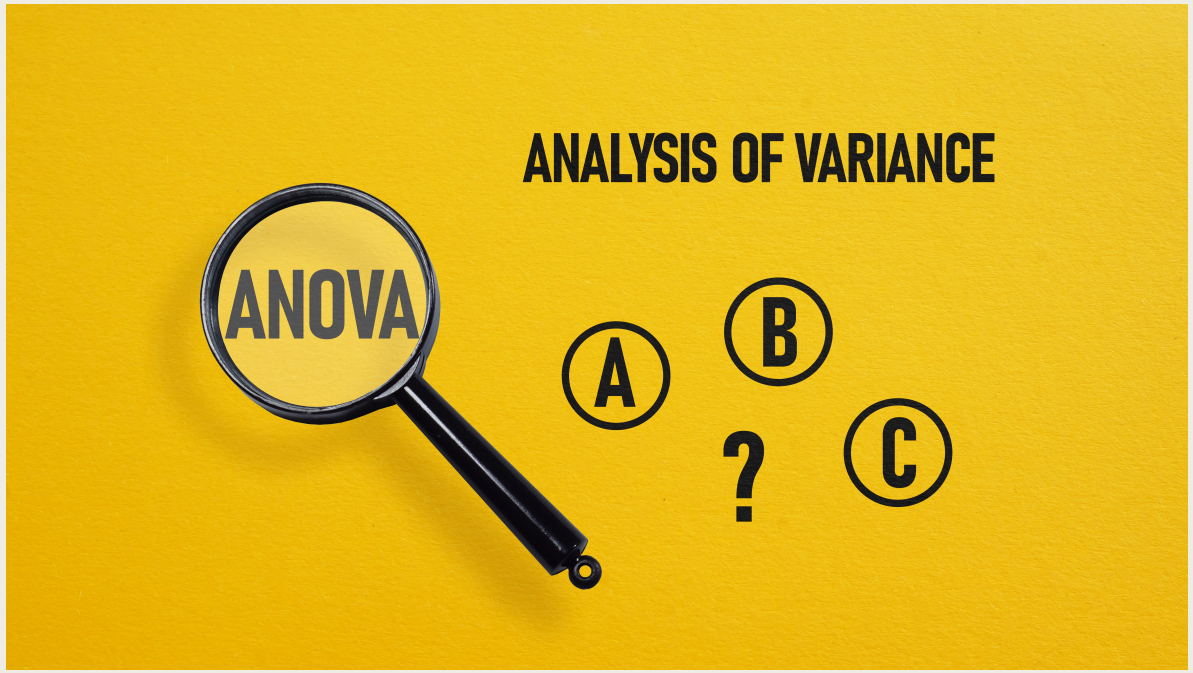


Jupyter Notebook
Hypothesis Testing - t-Test

ANOVA

Determine if there are statistically significant differences between results from 3 or more unrelated samples/groups.

- **Types**
  - Single factorial without measurement repetition
  - Single factorial with measurement repetition
  - Two-factorial without measurement repetition
  - Two-factorial with measurement repetition
- **Assumptions**
  - Normality
  - Homogeneity of Variances
  - Measurements should be independent
  - Dependent variable should have a metric scale level



Jupyter Notebook
Hypothesis Testing - ANOVA-Test

Non-parametric tests

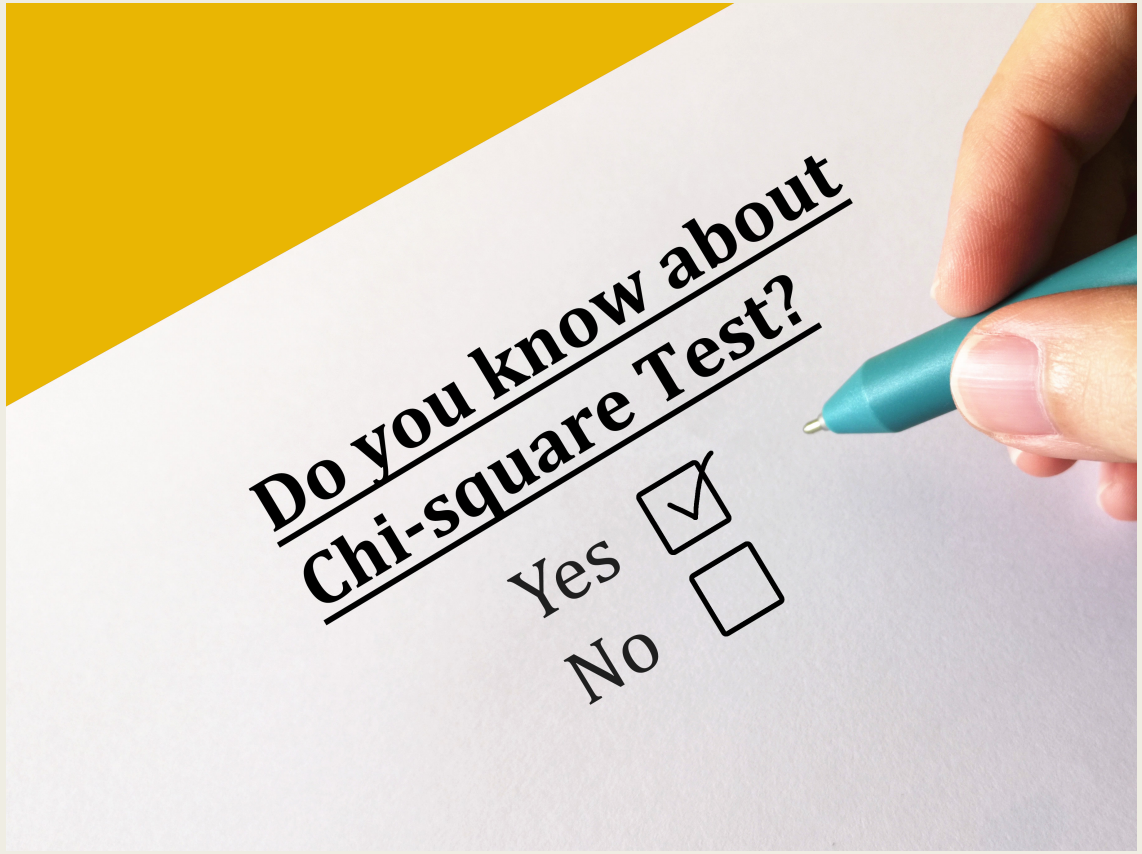
May be used with data that are **not** normally distributed

Non-parametric test	Corresponding parametric test
Mann-Whiteney U Test	t-test for independent samples
Wilcoxon Test	t-test for dependent samples
Kruskal-Wallis Test	ANOVA for more than two independent samples
Friedman Test	ANOVA with repeated measures

Chi^2 Test

A hypothesis test that is used when you want to determine if there is a relationship between **two categorical variables**.

- **Categorical Variables**
  - Variables with characteristics and descriptors that can't be easily measured, but can be observed subjectively.
  - Examples:
    - Gender
    - Favorite newspaper
    - Education level
- **Assumptions**
  - The expected frequencies per cell are greater than 5
  - Uses only the categories and not the rankings

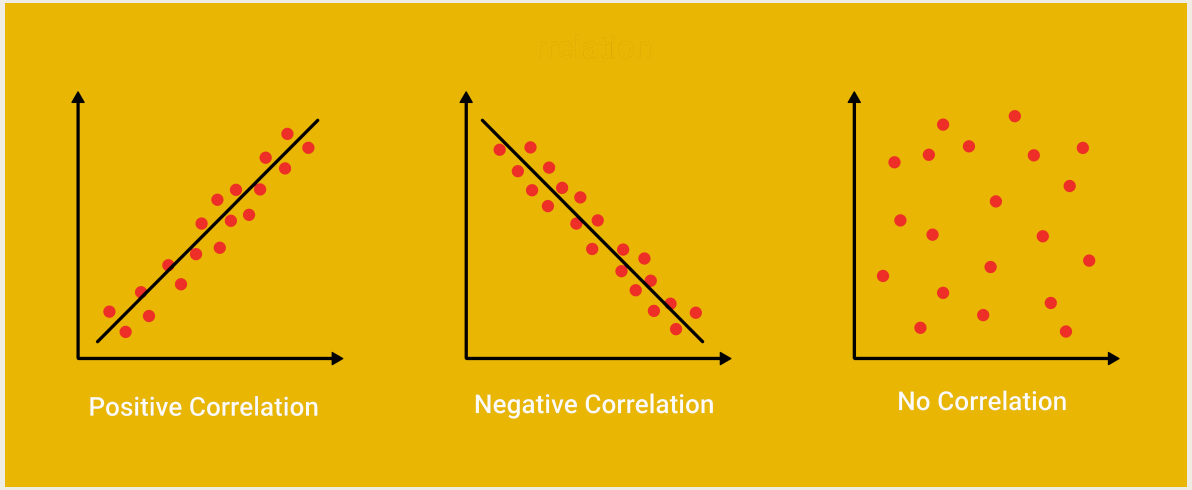


Jupyter Notebook
Hypothesis Testing - Chi² Test

Correlation Analysis

A statistical method used to measure the relationship between two variables.

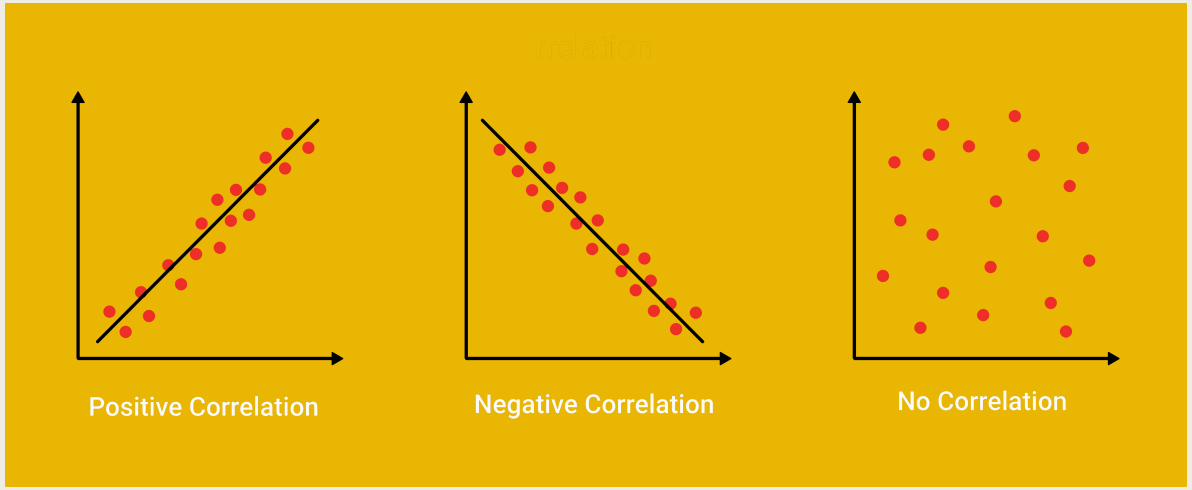
- **Types**
  - Pearson Correlation (r)
    - Measures the linear relationship between two variables
    - Requires normally distributed data
  - Spearman Correlation
    - Non-normally distributed data
    - Correlation based on rankings
  - Kendalls Tau
  - Point Biserial Correlation



Correlation Analysis

- **Output**
  - How strong the correlation is
  - Which direction the correlation goes
  - Correlation coefficient: -1 to 1
    - Strength: distance from 0
    - Direction: plus or minus

Coefficient (r)	Strength of the correlation
0.0 ± 0.1	No correlation
0.1 ± 0.3	Low correlation
0.3 ± 0.5	Medium correlation
0.5 ± 0.7	High correlation
0.7 ± 1	Very high correlation



Jupyter Notebook
Hypothesis Testing - Correlation