# Statistics for Data Science: Descriptive Statistics

April 15th, 2023

Data Analytics

**Program Zenius Studi Independen Bersertifikat**
**Zenius Bersama Kampus Merdeka**

1. Intro to Statistics
2. Types of Data
3. Descriptive Statistics

# Intro to Statistics

# Questions that statistics can answer

1. How is the satisfaction rate of a user using this ride hailing app ?
2. What factors influencing a non-performing loan case ?

# Statistics

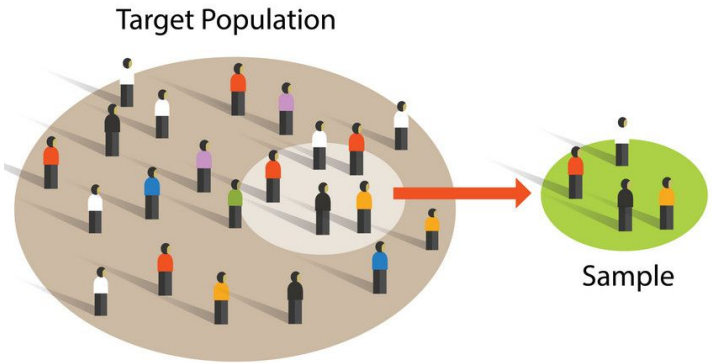
the science concerned with **developing** and **studying methods** for **collecting**, **analyzing**, **interpreting** and **presenting empirical data**.

**Source :** https://www.stat.uci.edu/what-is-statistics/

# Population vs Sample

A **population** is the **entire group** that you want to draw conclusions about.

A **sample** is the **part of population** that you will collect data from. The size of the sample is always less than the total size of the population.



# Descriptive vs Inferential Statistics

**Descriptive statistics** summarize the characteristics of a data set.

**Inferential statistics** will validate whether a population have a certain parameters based on the characteristics of a data set.

# Population vs Sample

A **sample** is the **subset of the population** and so population and sample are usually have different *characteristic*

**Population → parameter**

example : the average height of students in the class X is 161.5

**Sample → statistics**

example : the average height of 10 students in the class X is 162.3

# Types of Data

# What is Data ?

data are a **set of values** of **qualitative or quantitative variables** about **one or more persons or objects**

**Source :** https://en.wikipedia.org/wiki/Data

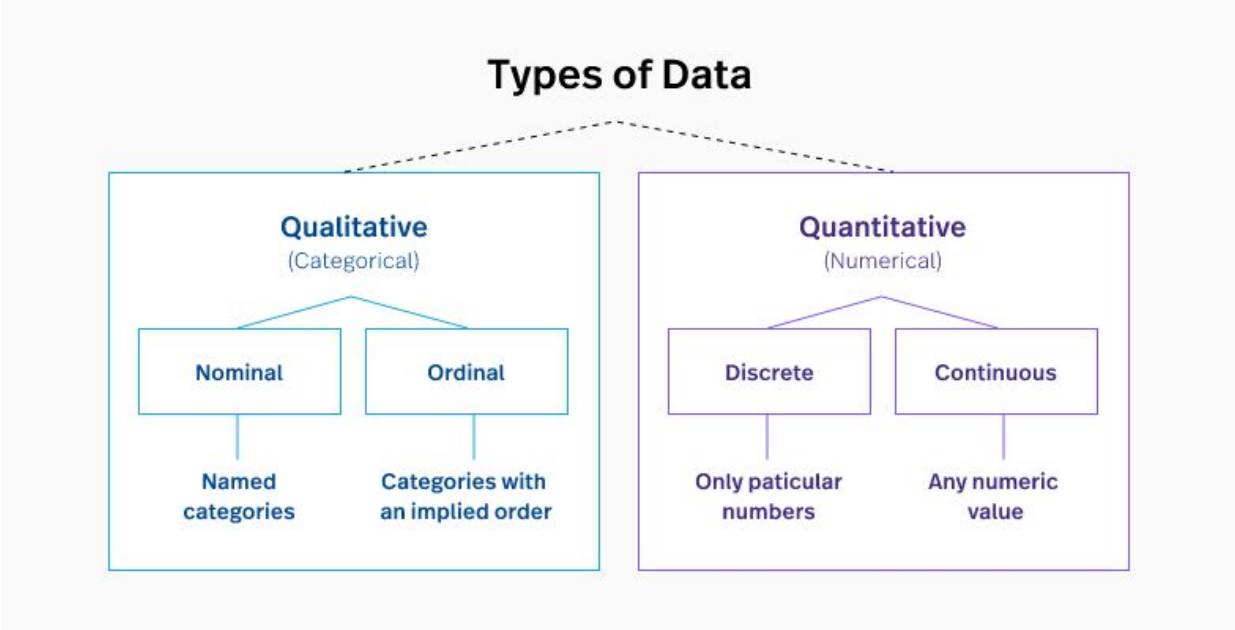
# Data Matrix

variable

observation

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup |
|---|---|---|---|---|---|---|---|---|---|---|
| 6840-RESVB | Male | 0 | Yes | Yes | 24 | Yes | Yes | DSL | Yes | No |
| 2234-XADUH | Female | 0 | Yes | Yes | 72 | Yes | Yes | Fiber optic | No | Yes |
| 4801-JZAZL | Female | 0 | Yes | Yes | 11 | No | No phone service | DSL | Yes | No |
| 8361-LTMKD | Male | 1 | Yes | No | 4 | Yes | Yes | Fiber optic | No | No |
| 3186-AJIEK | Male | 0 | No | No | 66 | Yes | No | Fiber optic | Yes | No |

# Variable Types



# Categorical Data

Categorical data is **a type of data that can be stored into groups or categories** with the aid of names or labels. This grouping is usually made according to the data characteristics and similarities of these characteristics through a method known as matching.

Also known as **qualitative data.**

For example : gender is a categorical data because it can be categorized into male and female according to some unique qualities possessed by each gender.

There are 2 main types of categorical data, namely; **nominal data** and **ordinal data**.

# Categorical Data

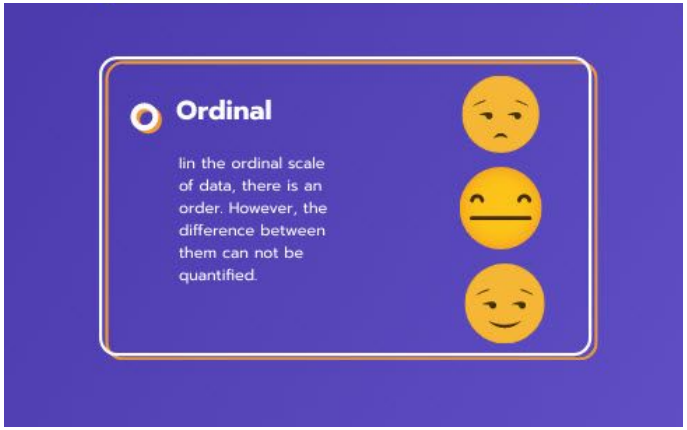
**Nominal Data**

Nominal data simply categorical data without order or rank between the category.
Example : **"pass"** or **"fail"** on a test

**Ordinal Data**

Ordinal data means a categorical data **with order** or rank.
Example : In a restaurant review, people can give ”1" for poor, "2" for below average, "3" for average, "4" for very good and "5" for excellent.

# Categorical Data

# Numerical Data

**Discrete Data**

Discrete variables are **countable in a finite amount of time**. For example, you can count the change in your pocket. You can count the money in your bank account. You could also count the amount of money in everyone's bank accounts. It might take you a long time to count that last item, but the point is—it's still countable.

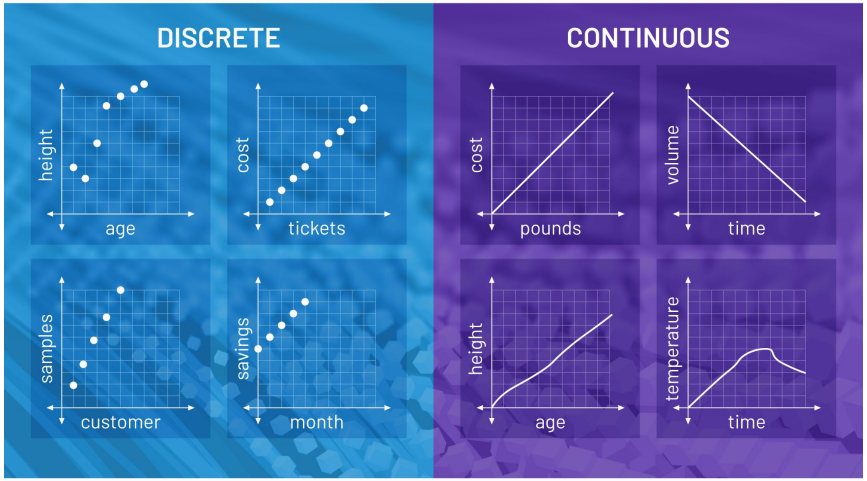
**Continuous Data**

Continuous data is a type of numerical data that refers to the unspecified number of possible measurements between two realistic points. Non-countable but measured.

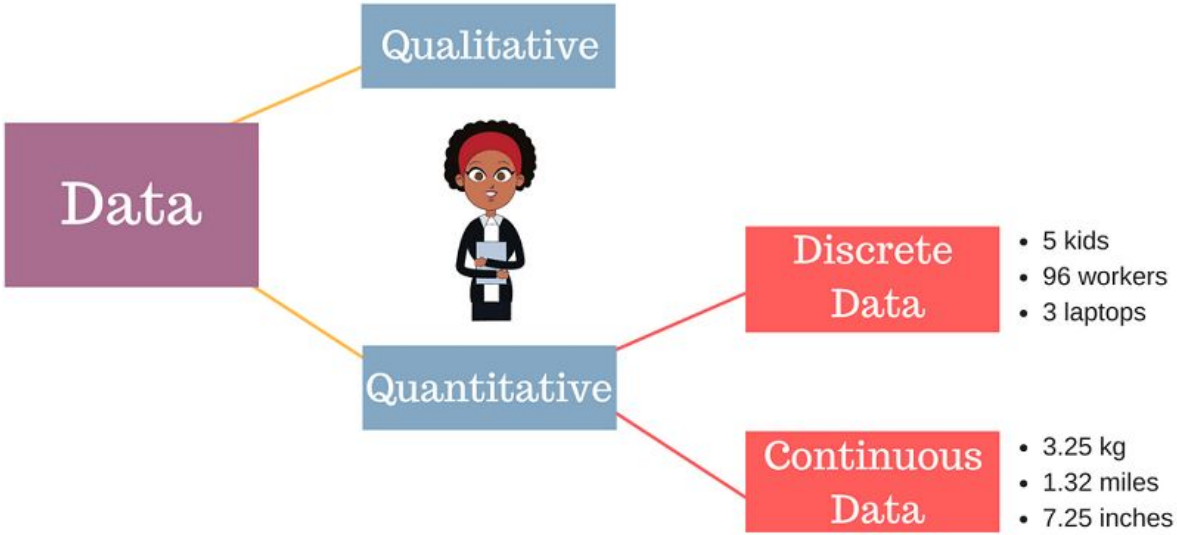
These numbers are not always clean and tidy like those in discrete data, as they're usually collected from precise measurements. Over time, measuring a particular subject allows us to create a defined range, where we can reasonably expect to collect more data.

# Numerical Data

**Discrete vs Continuous**



# Numerical Data



# Data Matrix

**Try to examine which is their respective variable type of each variable**

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup |
|---|---|---|---|---|---|---|---|---|---|---|
| 6840-RESVB | Male | 0 | Yes | Yes | 24 | Yes | Yes | DSL | Yes | No |
| 2234-XADUH | Female | 0 | Yes | Yes | 72 | Yes | Yes | Fiber optic | No | Yes |
| 4801-JZAZL | Female | 0 | Yes | Yes | 11 | No | No phone service | DSL | Yes | No |
| 8361-LTMKD | Male | 1 | Yes | No | 4 | Yes | Yes | Fiber optic | No | No |
| 3186-AJIEK | Male | 0 | No | No | 66 | Yes | No | Fiber optic | Yes | No |

# Descriptive Statistics

# Descriptive Statistics

## Measure of Central Tendency

## Measure of Spread

# Measure of Central Tendency

Position Statistics measure the data central tendency. Central tendency refers to where the data is centered. You may have calculated an average of some kind.

Despite the common use of average, there are different statistics by which we can describe the average of a data set:

- **Mean**
- **Median**
- **Mode**

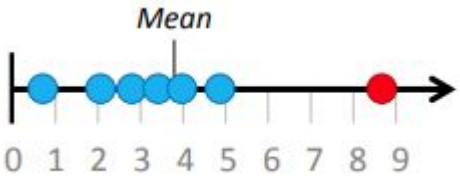
# Mean

The sum of all the values divided by the size of the data set.

The mean of a sample usually denoted by ‘x-bar’.

The mean of a population usually denoted by ‘μ’.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$



# Median

The middle value where exactly half of the data values are above it and half are below it.

A useful statistic due to its robustness.

Tp calculate median :

1. Order the values first from low to high
2. If number of sample is odd, take the middle value
   With an even number of values, take the mean of the two middle values.

| |
|---|
| 23 |
| 33 |
| 34 |
| 36 |
| **38** |
| 40 |
| 41 |
| 41 |
| 44 |
| |

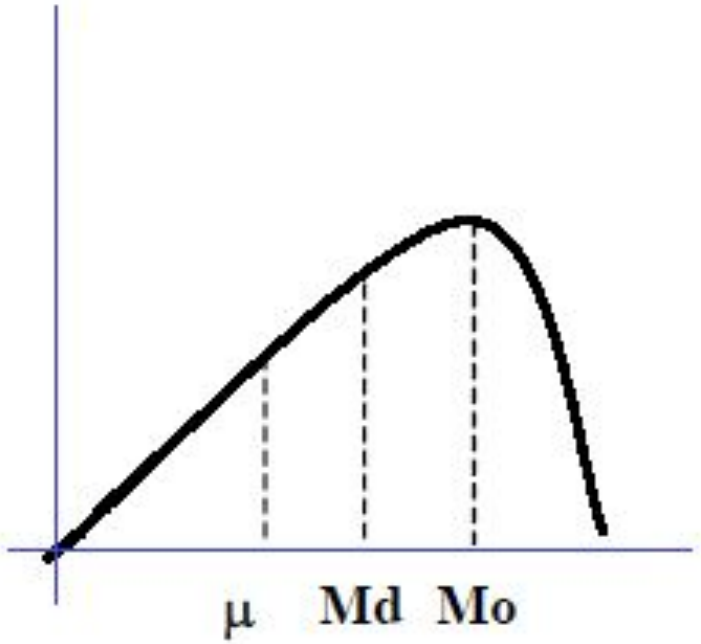
| |
|---|
| 12 |
| 30 |
| 31 |
| 37 |
| **38** |
| **40** |
| 41 |
| 41 |
| 44 |
| 45 |

Median = 38 + 40 / 2 = 39

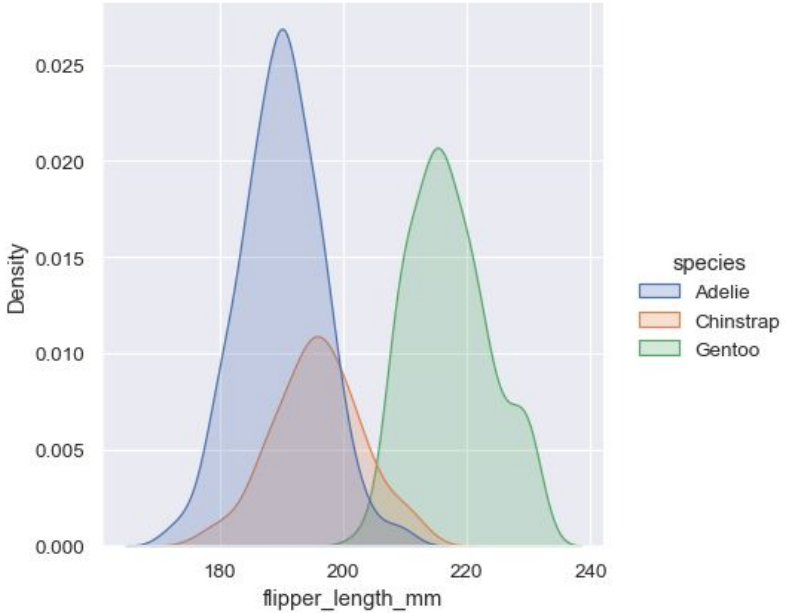
# Mode

The value that occurs the most often in a data set.

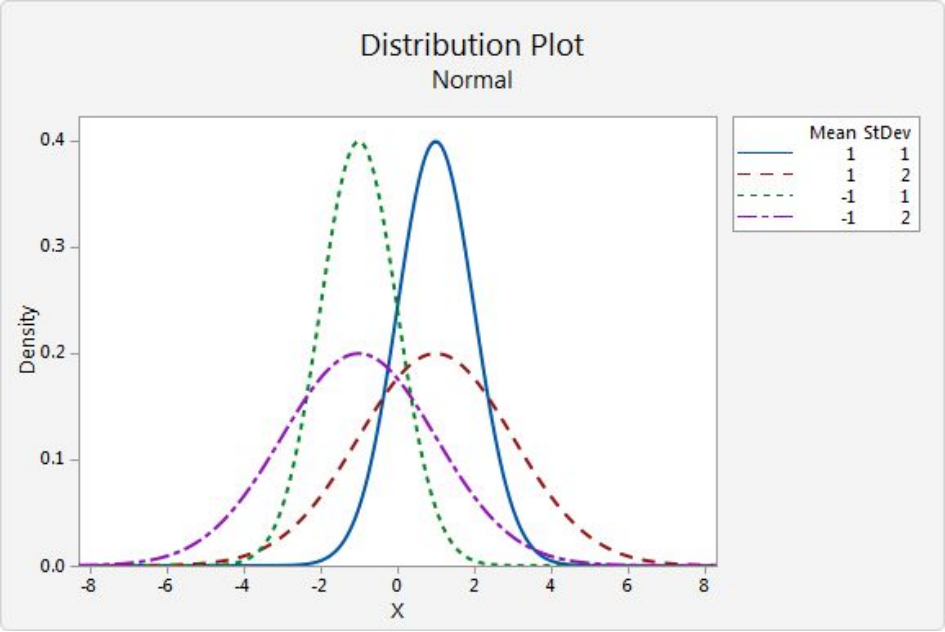
It is rarely used as a central tendency measure



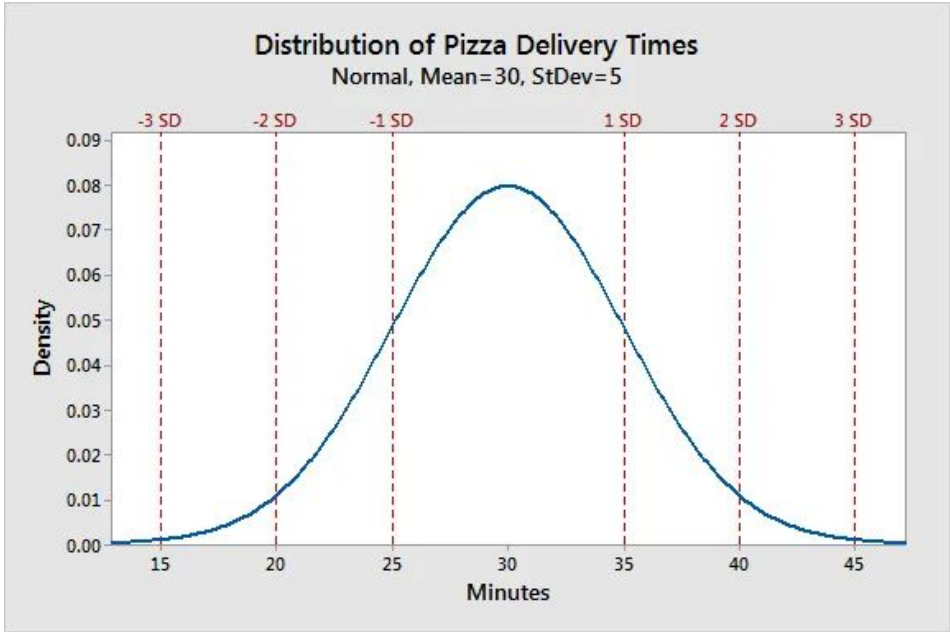
# Distribution Plot



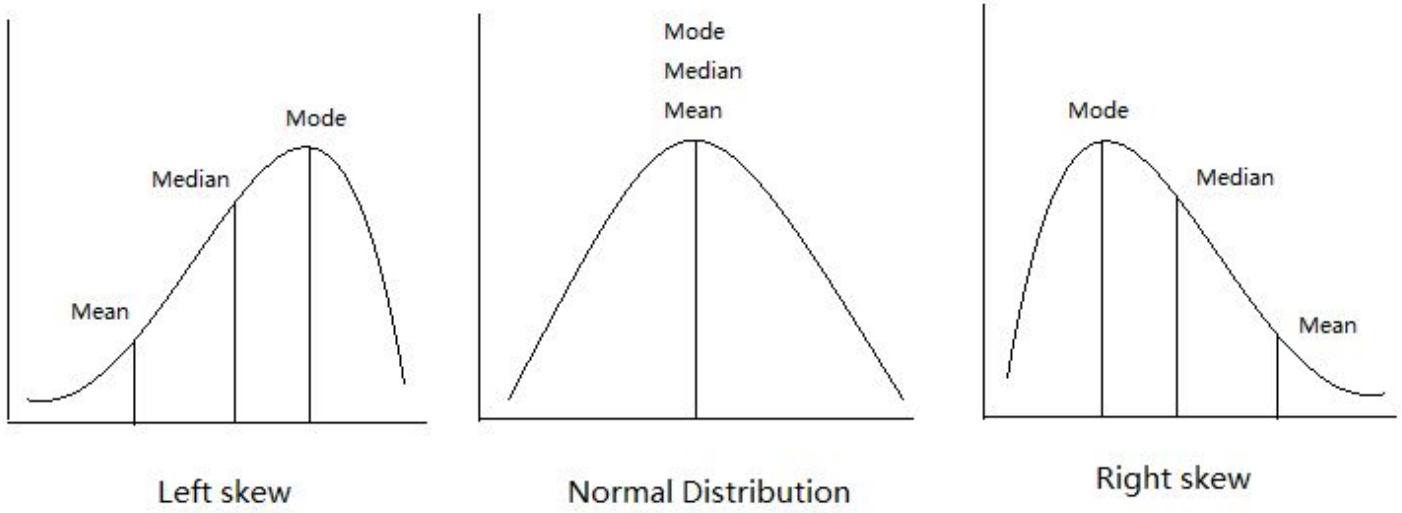
# Distribution Plot



# Distribution Plot



# Mean, Median, Mode



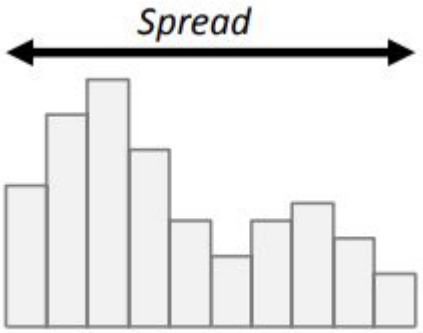
# Measure of Spread

The Spread refers to how the data deviates from the position measure.

It gives an indication of the amount of variation in the process.

- An important indicator of quality.
- Used to control process variability and improve quality.

Metrics used : **Range, Interquartile Range and Standard Deviation**



# Range

The difference between the highest and the lowest values.

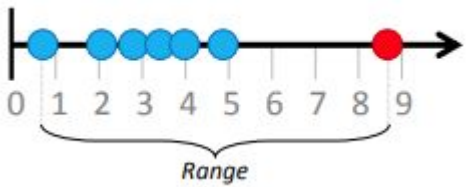
The simplest measure of variability.

It is good enough in many practical cases.

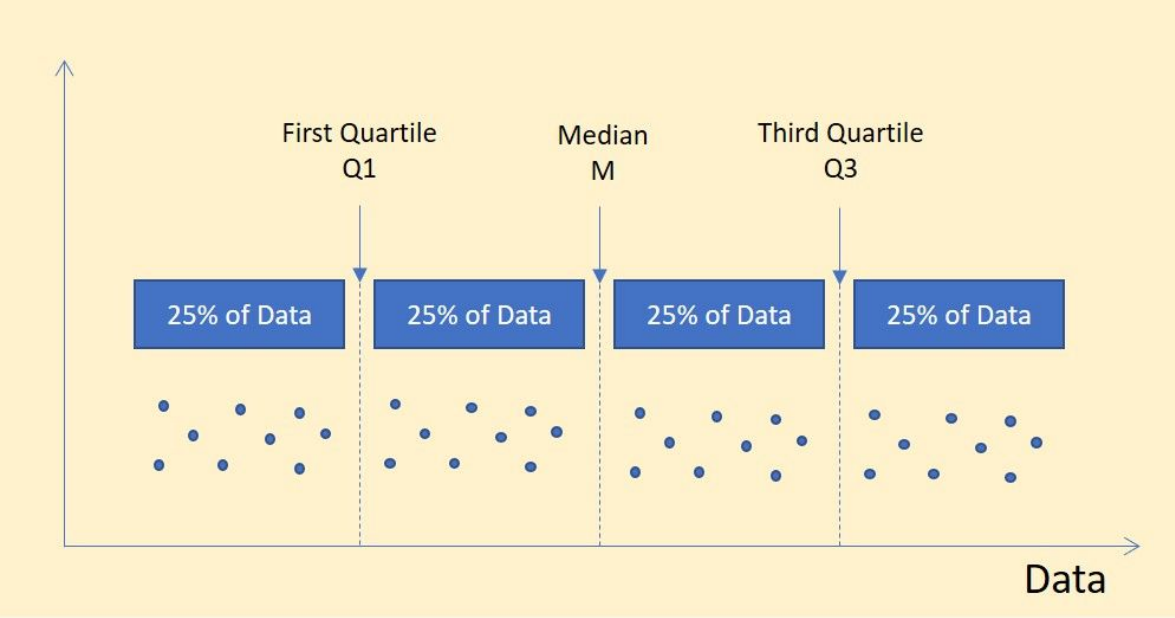
It does not make full use of the available data.

It can be misleading when the data is skewed or in the presence of outliers.

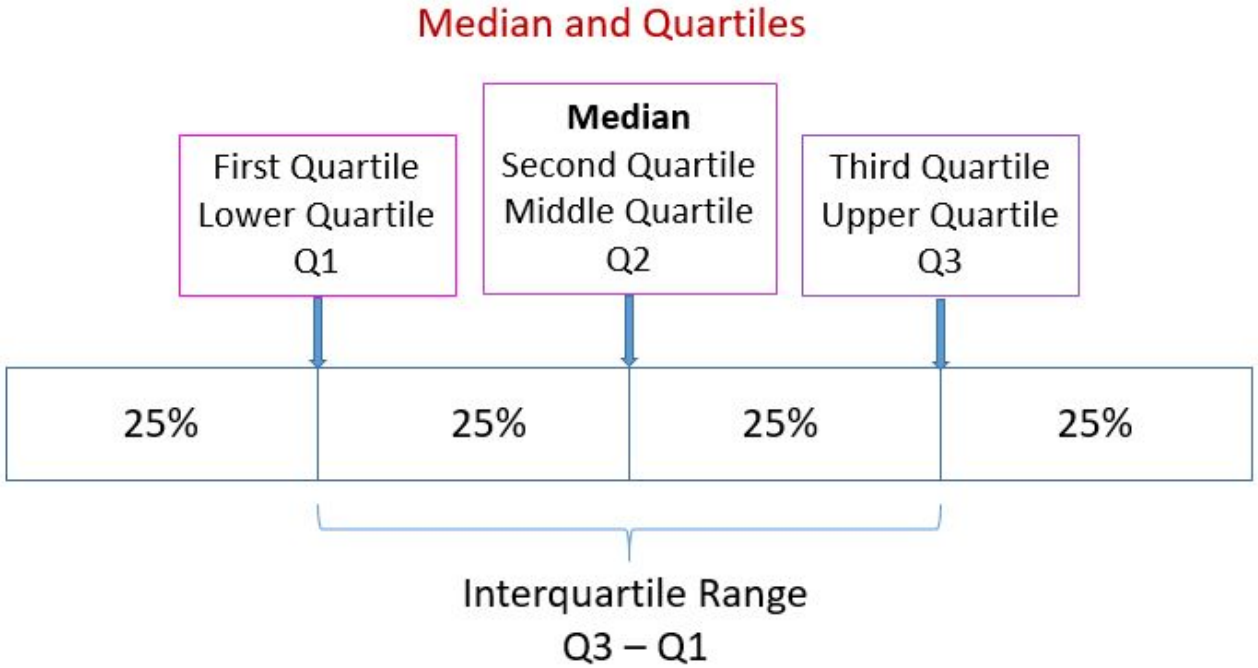
- Just one outlier will increase the range dramatically.



# Quartile



# Interquartile Range

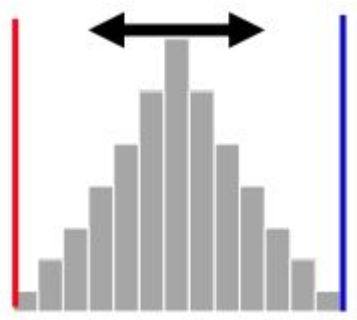


# Standard Deviation

Standard Deviation Formula

cuemath
THE MATH EXPERT

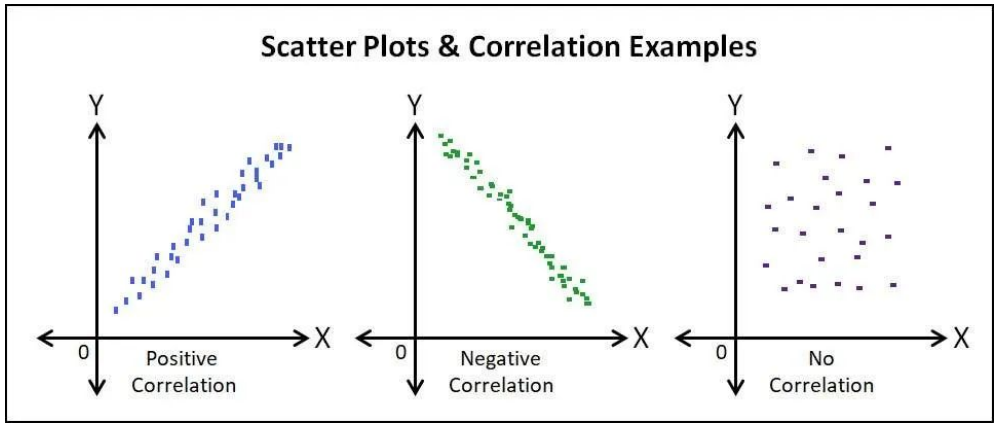
| Population | Sample |
| --- | --- |
| $\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$ | $s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$ |
| X - The Value in the data distribution<br>$\mu$ - The population Mean<br>N - Total Number of Observations | X - The Value in the data distribution<br>$\bar{x}$ - The Sample Mean<br>n - Total Number of Observations |



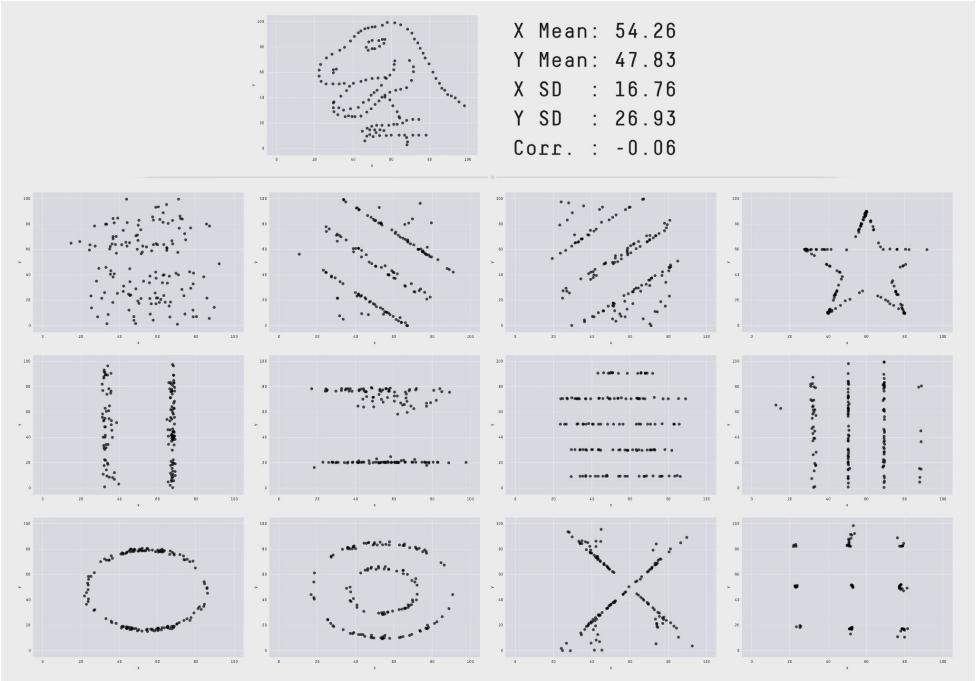
# Correlation

Correlation refers to the degree of association or relationship between two variables.

One of the most popular correlation used is **Pearson Correlation** which to measure **linear relationship** between two variables where the values range from -1 to 1



# Trap of numeric descriptive statistics



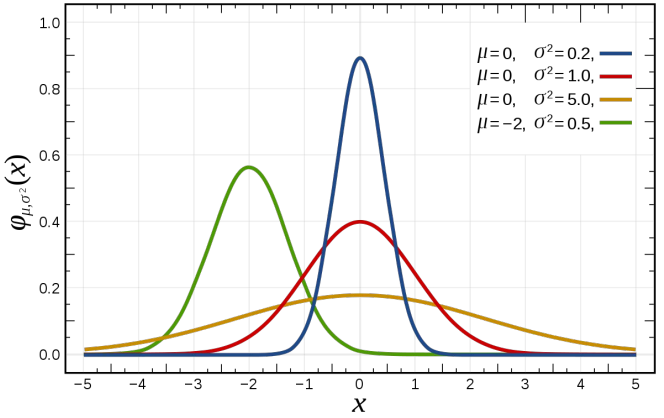
# Distribution functions

# Normal Distribution

The normal distribution is a probability distribution that is symmetric around the mean and characterized by its mean and standard deviation.

Many natural phenomena and statistical models follow this distribution due to the Central Limit Theorem.

Normal distribution configured by its **mean** and **variance.**

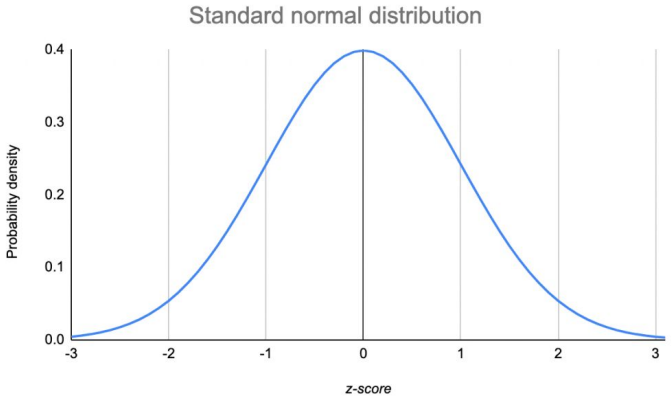



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

**Simple explanation :** https://www.youtube.com/watch?v=xgQhefFOXrM

# Standard Normal Distribution

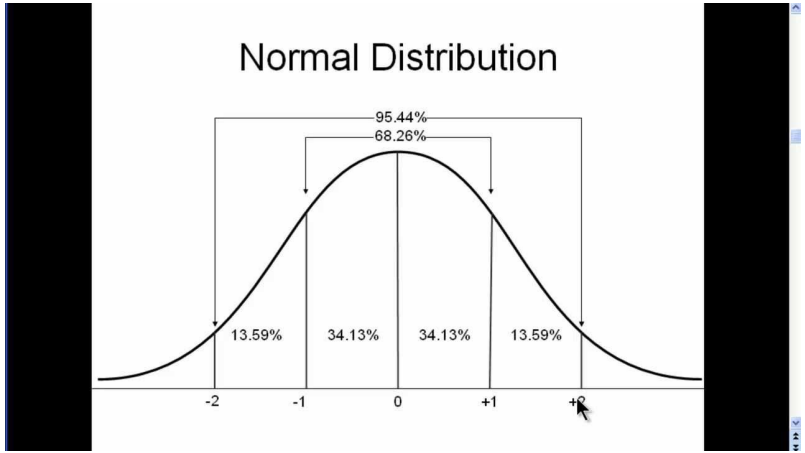
Standard Normal Distribution is a Normal Distribution where the mean = 0 and variance = 1




**Simple explanation :** https://www.youtube.com/watch?v=xgQhefFOXrM

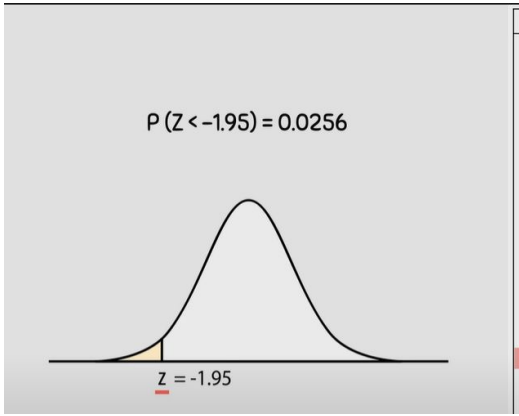
# Z-Score

Z-score simply tell us how many standard deviation away a point from its sample




**Simple explanation :** https://www.youtube.com/watch?v=xgQhefFOXrM

# Probability Calculation & Z-score



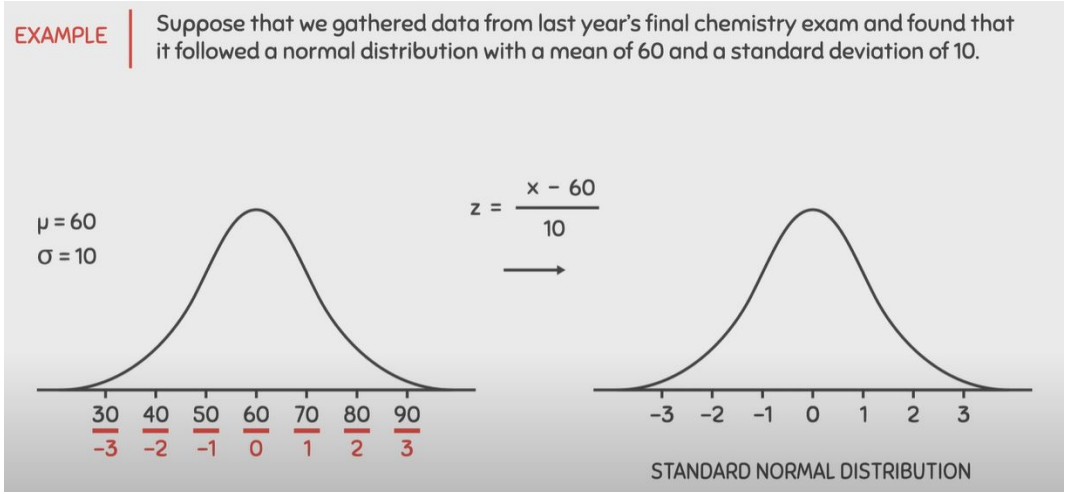

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 |
|---|---|---|---|---|---|---|---|---|
| -3.4 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 |
| -3.3 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 |
| -3.2 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 |
| -3.1 | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 |
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 |
| -2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 |
| -2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 |
| -2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 |
| -2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 |
| -2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 |
| -2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 |
| -2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 |
| -2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 |
| -2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 |
| -2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 |
| -1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0224 |
| -1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 |
| -1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 |

Statistician in the past use this table to map z-score value to the integral of the distribution

* Integral of the distribution function from a to b = probability of a observation occurred between a to b

**Simple explanation :** https://www.youtube.com/watch?v=2tuBREK_mgE

# Standardization




Common procedure :

1. We want to know the probability of **X** happened
2. The data have normal distribution
3. We standardize it
4. Then lookup the probability value based on the z-score of **X**

**Simple explanation :** https://www.youtube.com/watch?v=2tuBREK_mgE

# Expected Values

In statistics, expectation is a mathematical concept that represents the mean value of a random variable.

For discrete random variable the expected value is **calculated by multiplying each of the possible outcomes** by **the likelihood each outcome will occur** and then **summing all of those values**.

For continuous random variable, we take the integral instead of the sum

$$E[X] = \sum_i x_i f(x_i)$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

# Expected Values

General Formula:

| Outcome | $X_1$ | $X_2$ | $X_3$ | $X_4$ | ... | $X_n$ |
|---|---|---|---|---|---|---|
| Probability | $P_1$ | $P_2$ | $P_3$ | $P_4$ | ... | $P_n$ |

Expected Value = $P_1X_1 + P_2X_2 + P_3X_3 + P_4X_4 + ... + P_nX_n$

Example: A single fair six-sided die is rolled.

| Outcome | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Probability | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

Expected Value = $1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

# Thank you

## Any Questions ?