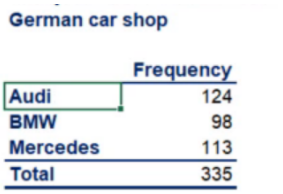
# Descriptive Statistics: Fundamentals

quick prep notes @ranton95

**Contents:**

# Representation of Categorical Variables:

## Frequency distribution tables

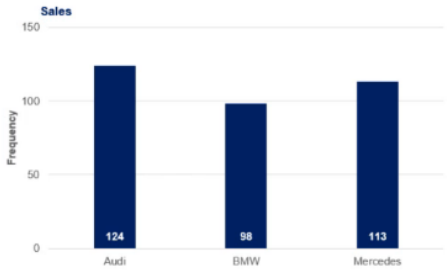


German car shop

| | Frequency |
|---|---|
| Audi | 124 |
| BMW | 98 |
| Mercedes | 113 |
| Total | 335 |

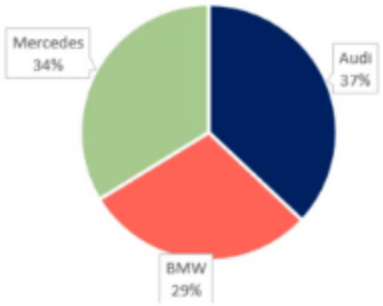
Has the categories itself and the corresponding variables. Here frequency is the number of units sold.

## Barcharts




Separates categories and visualizes them

## Pie charts




Pie charts are visualized after calculating the Relative frequencies.

**Relative frequency** is the percentage of the total frequency for each category. Naturally, all relative frequencies add up to 100%

## Pareto diagrams

A Pareto diagram is a special type of bar chart, where categories are shown in descending order of frequency.

## Graphs and tables for categorical variables

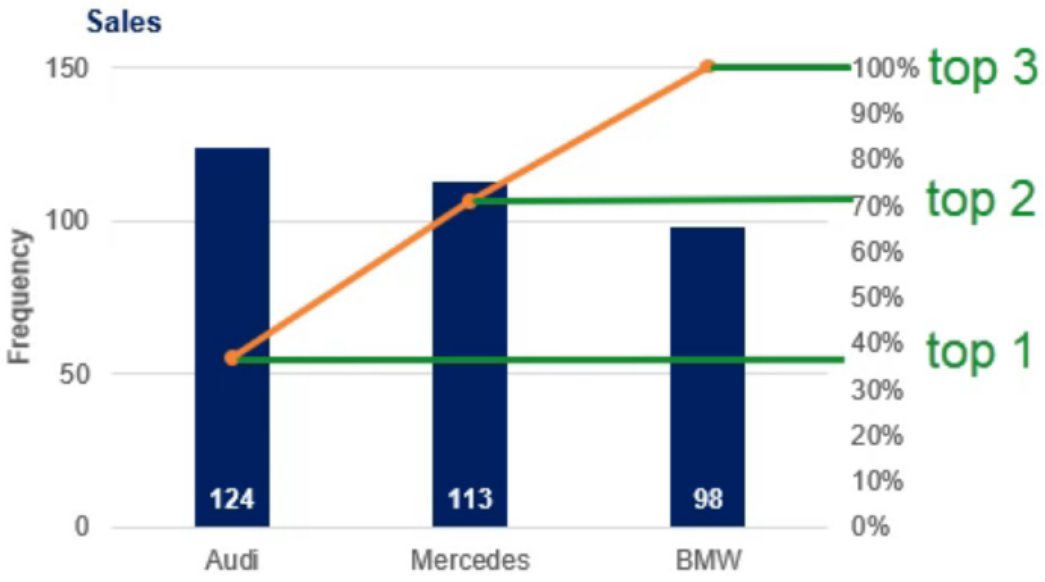
German car shop

| | Frequency | Relative frequency |
|---|---|---|
| Audi | 124 | 37% |
| BMW | 98 | 29% |
| Mercedes | 113 | 34% |
| Total | 335 | 100% |

| *Ordered* | Frequency | Relative frequency | Cumulative frequency |
|---|---|---|---|
| Audi | 124 | 37% | 37% |
| Mercedes | 113 | 34% | 71% |
| BMW | 98 | 29% | 100% |

By **frequency**, statisticians mean the number of occurrences of each item. In this case, it is the number of units sold.

**Cumulative frequency** (line on the graph) is the sum of relative frequencies.



The Pareto Principle (80-20 rule) derives from this. 80% of the effect come from 20% of the causes.

It shows how subtotals change with each additional category and provide us with a better understanding of our data.

# Representation of Numerical Variables:

## Frequency distribution table

As the name implies, distributing the frequencies in desired intervals. Whenever we have to plot data, it is easier to plot once we arrange them in a table. The same approach is done here. When we deal with numerical variables, it makes much more sense to group the data into intervals and then find the corresponding frequencies.

This way we have a summary of the data for which we can create a meaningful visual representation. Generally, statisticians prefer 5-20 intervals, however, it depends. The **desired intervals** can be calculated as follows:

$$\text{desired intervals: } \frac{largest number - smallest number}{number of desired intervals}$$

From these intervals, the frequency in each interval is collected into the table. A number is included in the particular interval

- If that number is **greater** than the lower bound
- is **lower** or **equal** to the upper bound

Following this, we calculate the relative frequencies in each interval.

$$\text{Relative frequencies: } \frac{Frequency}{TotalFrequency}$$

**Graphs and tables for numerical variables. Frequency distribution table**

| Dataset | Frequency |
|---|---|
| 1 | 1 |
| 9 | 1 |
| 22 | 1 |
| 24 | 1 |
| 32 | 1 |
| 41 | 1 |
| 44 | 1 |
| 48 | 1 |
| 57 | 1 |
| 66 | 1 |
| 70 | 1 |
| 73 | 1 |
| 75 | 1 |
| 76 | 1 |
| 79 | 1 |
| 82 | 1 |
| 87 | 1 |
| 89 | 1 |
| 95 | 1 |
| 100 | 1 |
| Total | 20 |

**Frequency distribution table**

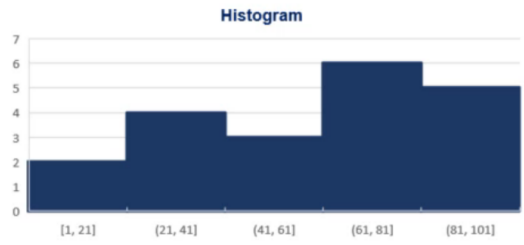
Interval width 20

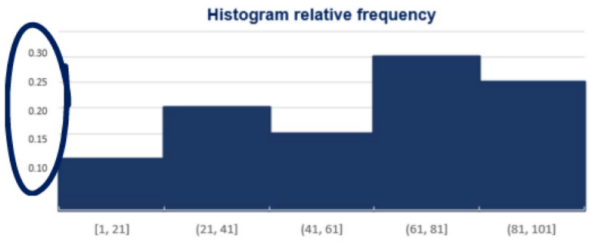
| Interval start | Interval end | Frequency | Relative frequency |
|---|---|---|---|
| 1 | 21 | 2 | 0.10 |
| 21 | 41 | 4 | 0.20 |
| 41 | 61 | 3 | 0.15 |
| 61 | 81 | 6 | 0.30 |
| 81 | 101 | 5 | 0.25 |

# The Histogram

One of the most common way to represent numerical data.

Unlike the bar chart, in Histograms, both the horizontal and vertical axes are numerical. Each bar has **width** equal to the **interval** from frequency distribution table and **height** equal to the **frequency.**



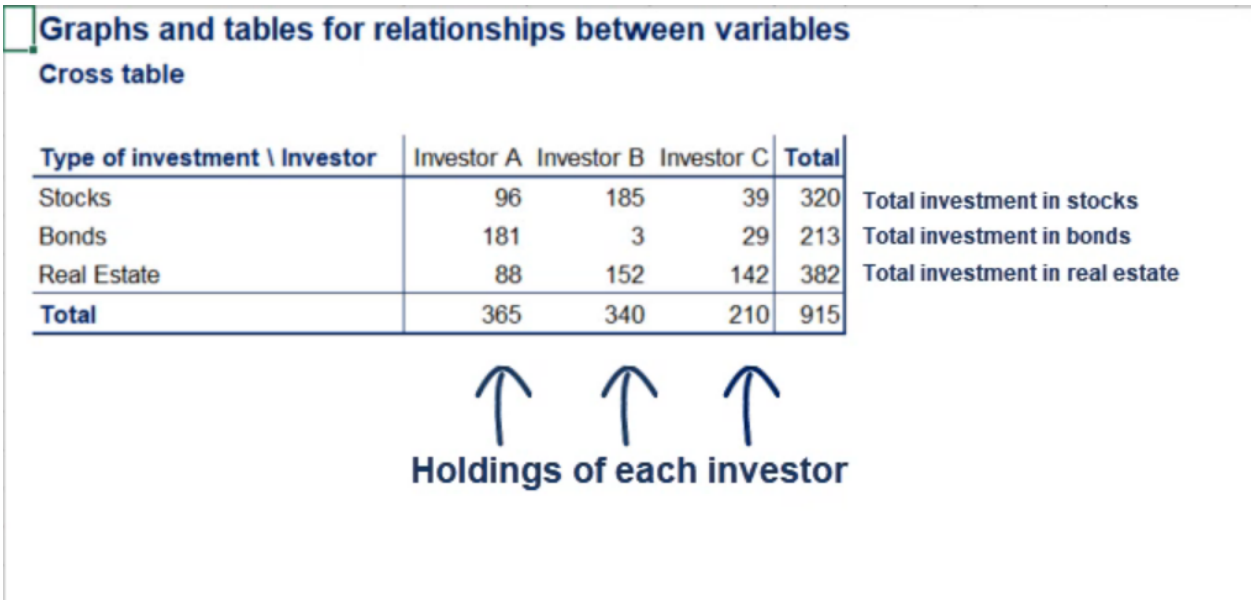




There is no gab between the bars, meaning each interval ends where the next one starts.

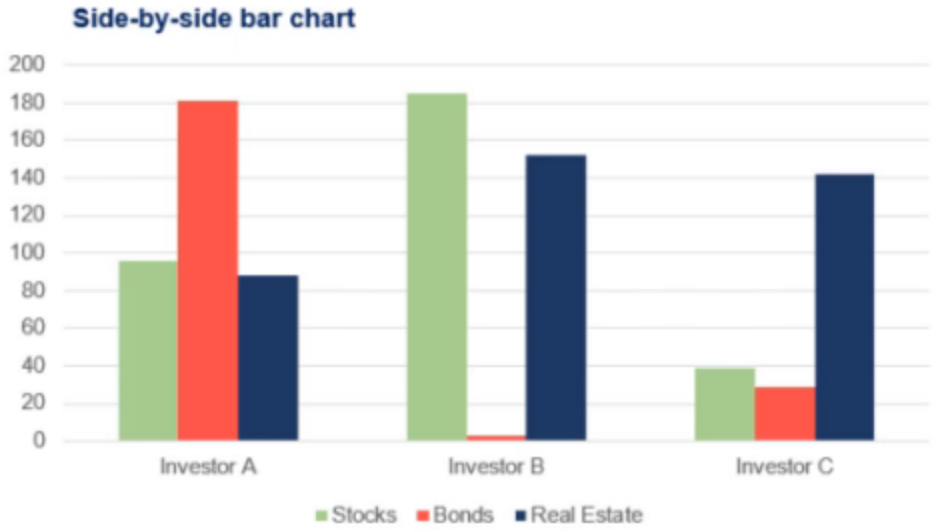
In some cases, we plot the Histograms with respect to the **relative frequency** (which is denoted as %) rather than the **absolute frequency**.

## Cross table and scatter plot

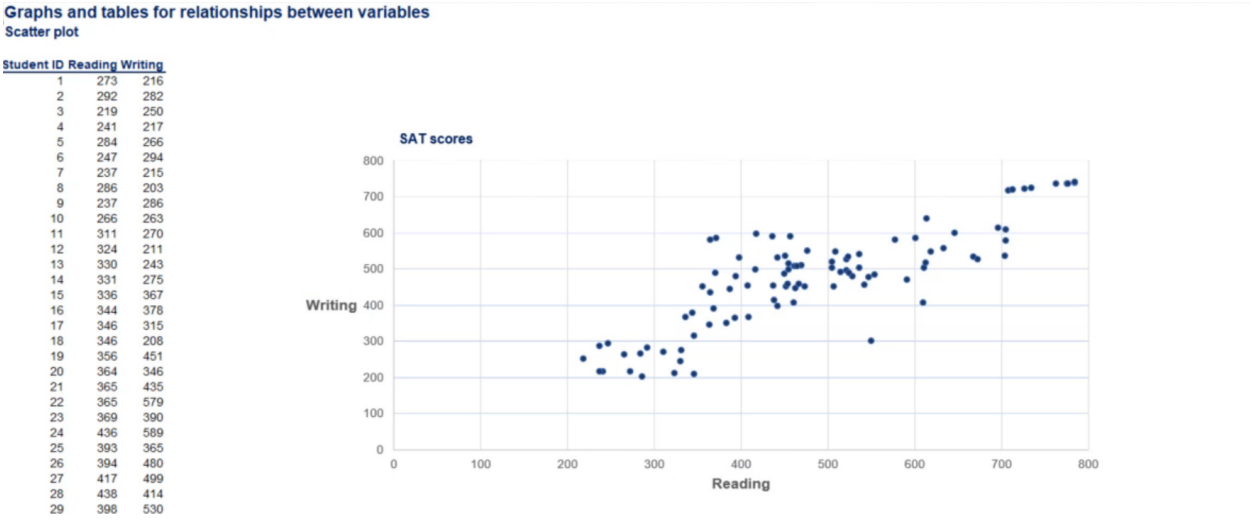
**Cross table:** The most common way to represent Categorical Variable. Also known as *contingency tables*.




The most common way to represent cross table **side-by-side chart**, which is a variation of the bar chart.




All graphs are very easy to create and read, once you have identified the type of data you are dealing with.



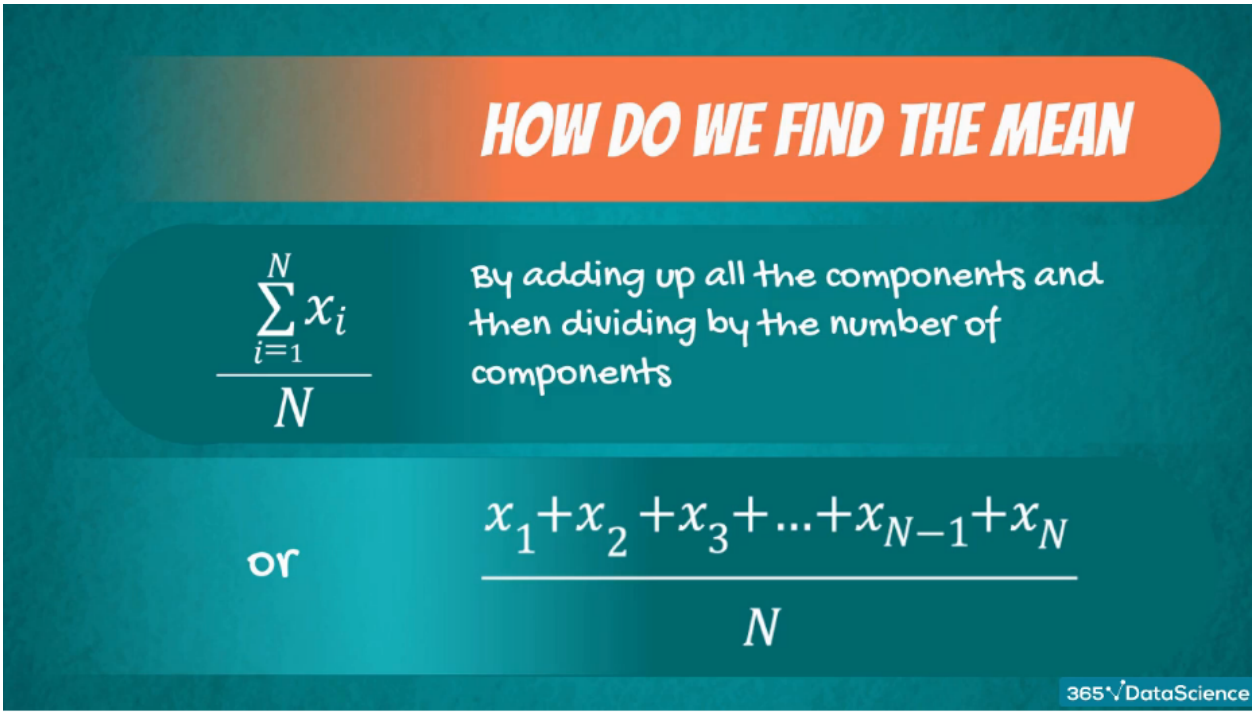

**Scatter plots** are used when representing two numerical variables. Each point gives us information about a particular students performance. When interpreting a scatterplot, a statistician is not expected to look at a single data point. The main idea here is to extract how the data is distributed.

Scatterplots helps us determine the **Outliers**. These are data points that go against the logic of the whole dataset.

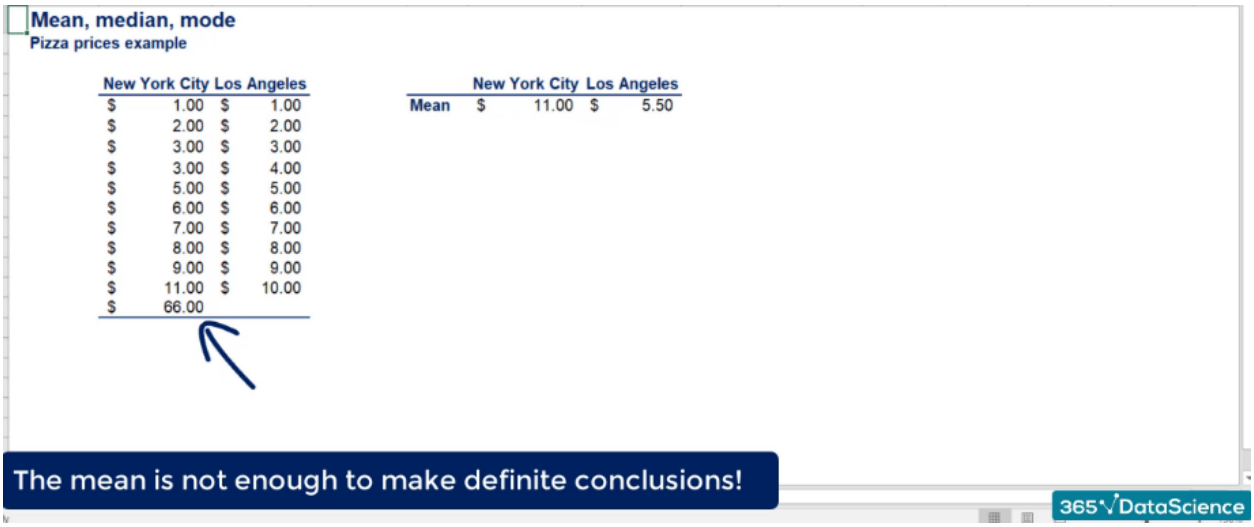
# Measurements of Central Tendency

## Mean

- Population Mean (**μ**)
- Sample Mean (**x̄**)




The Mean is the most common measure of central tendency. However it has a huge downside as it is **easily affected** when an outlier exists in the dataset.

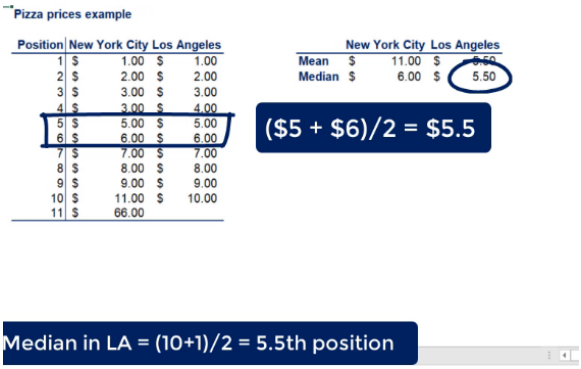



This is where we use Median.

## Median

The median is the middle number in an ordered dataset.

- In order to calculate median, we arrange the dataset first in the ascending order.
- The median is the number at position (n+1)/2 in the ordered list where n is the number of observations




## Mode

The Mode is the value that occurs the most often. It can be used for both numerical and categorical data.

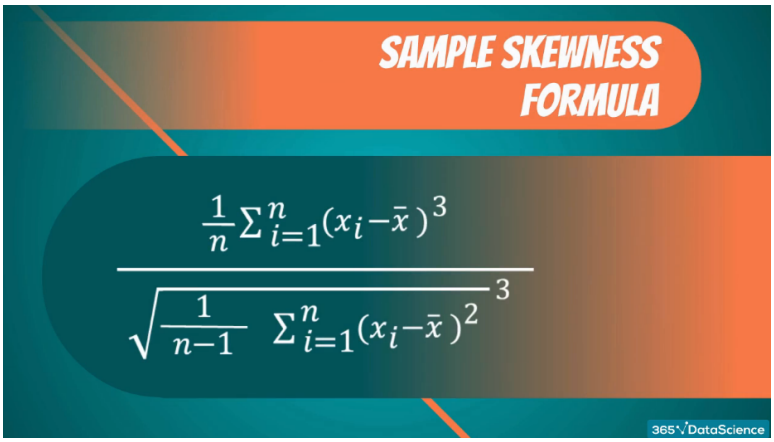
| Position | New York City | Los Angeles |
|---|---|---|
| 1 | $ 1.00 | $ 1.00 |
| 2 | $ 2.00 | $ 2.00 |
| 3 | $ 3.00 | $ 3.00 |
| 4 | $ 3.00 | $ 4.00 |
| 5 | $ 5.00 | $ 5.00 |
| 6 | $ 6.00 | $ 6.00 |
| 7 | $ 7.00 | $ 7.00 |
| 8 | $ 8.00 | $ 8.00 |
| 9 | $ 9.00 | $ 9.00 |
| 10 | $ 11.00 | $ 10.00 |
| 11 | $ 66.00 | |

| | New York City | Los Angeles |
|---|---|---|
| Mean | $ 11.00 | $ 5.50 |
| Median | $ 6.00 | $ 5.50 |
| Mode | $ 3.00 | - |

When data point appears only once, there occurs the point where we say there is **no** mode.

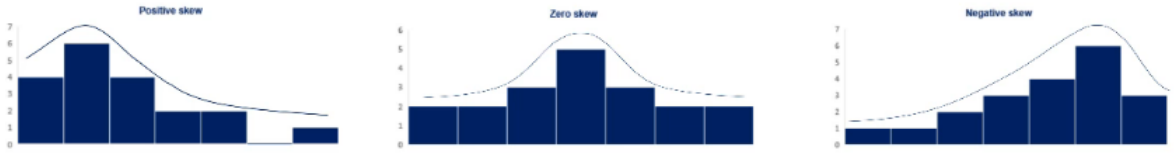
# Measurements of Asymmetry

## Skewness



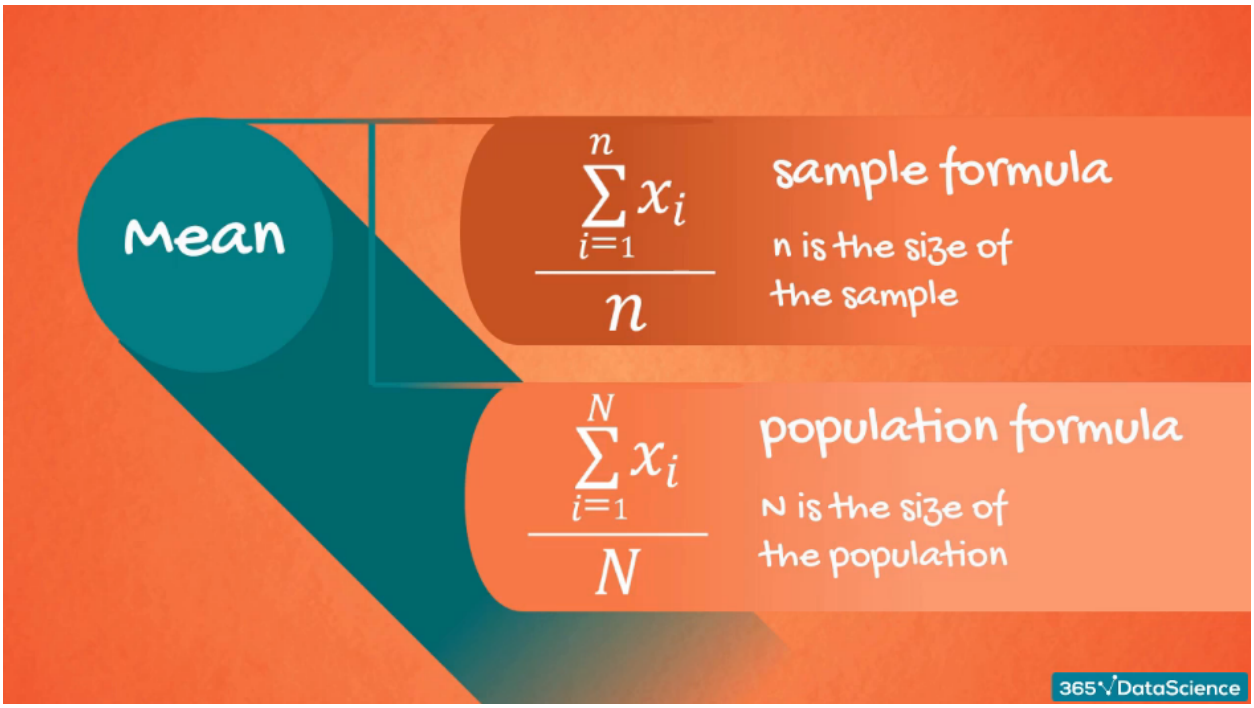

Skewness indicates whether the data is concentrated on one side

- mean > median: Positive skew/Right skew (outliers to the right)
- mean = median: mode: No skew asymmetrical
- mean < median: Negative skew/Left skew (outliers to the left)




# Measurements of Variability

When we take sample data for measurements the **formulas** applied for **population** and **sample** are **different**. For example:




In **population data** we are 100% sure of the the measures we are calculating. However for **sample data** it is always an approximation of the population parameter. Hence the formulas are adjusted for both cases.

## Variance

Variance measures the dispersion(spread) of a set of data points around their mean value.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

Notice for the numerator, the difference indicates the dispersion from the mean and squaring to produce non-negative values.

# Standard Deviation

Standard deviation is the most common measure of variability for a single dataset.

It is simply the **square root** of **variance**:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$\sigma$ = population standard deviation
$N$ = the size of the population
$x_i$ = each value from the population
$\mu$ = the population mean

Population standard deviation formula

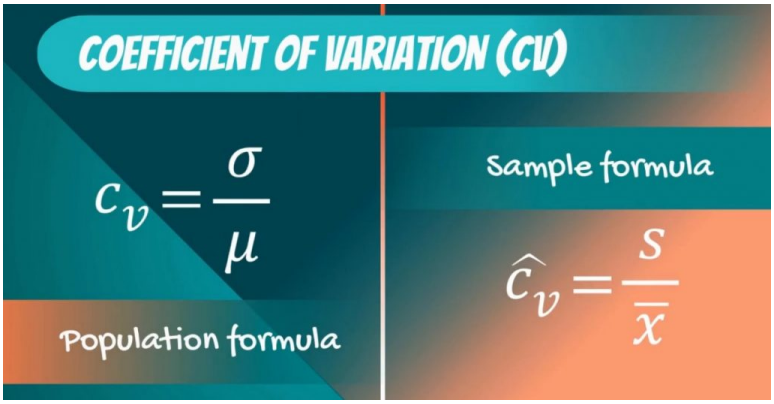
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

$s$ = sample standard deviation
$N$ = the number of observations
$x_i$ = the observed values of a sample item
$\overline{x}$ = the mean value of the observations

Sample standard deviation formula

# Coefficient of Variation (CV)

**Coefficient of variation** also known as **Relative standard deviation** is equal to the standard deviation divided by the mean:



| NY Dollars | | Pesos |
|---|---|---|
| $ 1.00 | MXN | 18.81 |
| $ 2.00 | MXN | 37.62 |
| $ 3.00 | MXN | 56.43 |
| $ 3.00 | MXN | 56.43 |
| $ 5.00 | MXN | 94.05 |
| $ 6.00 | MXN | 112.86 |
| $ 7.00 | MXN | 131.67 |
| $ 8.00 | MXN | 150.48 |
| $ 9.00 | MXN | 169.29 |
| $ 11.00 | MXN | 206.91 |

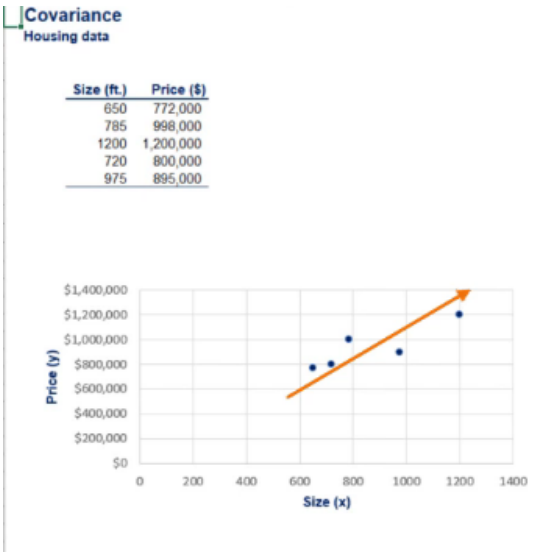
| | Dollars | Pesos |
|---|---|---|
| Mean | $ 5.50 | MXN 103.46 |
| Sample variance | $^2$ 10.72 | MXN$^2$ 3793.69 |
| Sample standard deviation | $ 3.27 | MXN 61.59 |
| Sample coefficient of variation | 0.60 | 0.60 |

- does not have a unit of measurement
- universal across datasets
- perfect for comparisons

# Measurements of Relation Between Variables

## Covariance

Helpful for between understanding the clear relationship between two variables. Example, Housing prices. Size vs. Price scatter plot.




We say that two variables are correlated and the main statistics to measure this correlation is called **covariance.**

Covariance can be **positive** or **negative** or even equal to **zero**.

**Sample formula**

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

**Population formula**

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$

Covariance gives a sense of direction

> 0, the two variables move together
< 0, the two variables move in opposite directions
= 0, the two variables are independent

## Linear Correlation Coefficient

Correlation adjusts covariance, so that the relationship between the two variables becomes easy and intuitive to interpret.

$$\frac{Cov(x,y)}{Stdev(x) * Stdev(y)}$$

$$\frac{S_{xy}}{S_x S_y} \qquad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

sample population

**-1 ≤ correlation coefficient ≤ 1**

Basically we manipulate the covariance to get a better result. However keep in mind there is a strong relationship between the two variables.

Correlation coeff. = 1 means the entire variability of one variable is explained by the other

Correlation coeff. = 0 : Absolutely independent (prices of coffee in Brazil, to prices of Houses in London)

Correlation coeff. = -1 perfect negative correlation. Think about your crypto trading of altcoin pairs and bitcoin.

## Causality

Important to understand the direction of causal relationships.

**Correlation does not imply causation.**