Course notes: Descriptive statistics

Types of data

```
                        Types of data


   Categorical                              Numerical


Categorical data represents groups or categories.
                                              Discrete        Continuous

Examples:
                                     Numerical data represents numbers. It is divided into two
1. Car brands: Audi, BMW and Mercedes.  groups: discrete and continuous. Discrete data can be
                                     usually counted in a finite matter, while continuous is
2. Answers to yes/no questions: yes and no  infinite and impossible to count.

                                     Examples:
                                     Discrete: # children you want to have, SAT score
                                     Continuous: weight, height
```

365 DataScience

Levels of measurement

```
                    ┌─────────────────────────┐
                    │  Levels of measurement  │
                    └─────────────────────────┘
                       │                    │
           ┌───────────┘                    └──────────────┐
           ▼                                               ▼
    ┌──────────────┐                              ┌──────────────┐
    │  Qualitative │                              │ Quantitative │
    └──────────────┘                              └──────────────┘
        │       │                                     │       │
    ┌───┘       └───┐                             ┌───┘       └───┐
    ▼               ▼                             ▼               ▼
┌─────────┐    ┌─────────┐                   ┌─────────┐    ┌─────────┐
│ Nominal │    │ Ordinal │                   │ Interval│    │  Ratio  │
└─────────┘    └─────────┘                   └─────────┘    └─────────┘
```

There are two qualitative levels: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that can be ordered.

Examples:
Nominal: four seasons (winter, spring, summer, autumn)
Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios have a true zero, while intervals don't.

Examples:
Interval: degrees Celsius and Fahrenheit
Ratio: degrees Kelvin, length

365 DataScience

Graphs and tables that represent categorical variables

Frequency distribution tables

Bar charts

Pie charts

Pareto diagrams

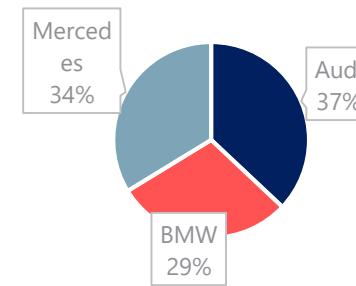|	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

Sales

Sales

Sales

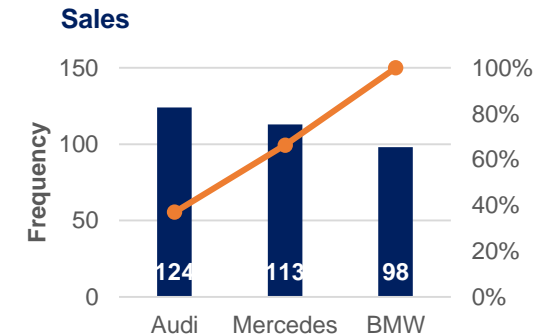Frequency distribution tables show the category and its corresponding absolute frequency.
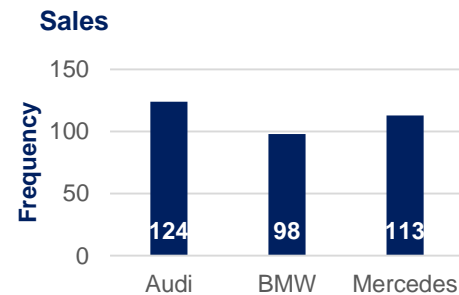
Bar charts are very common. Each bar represents a category. On the y-axis we have the absolute frequency.

Pie charts are used when we want to see the share of an item as a part of the total. Market share is almost always represented with a pie chart.

The Pareto diagram is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency.

Graphs and tables that represent categorical variables. Excel formulas

Frequency distribution tables → Bar charts → Pie charts → Pareto diagrams

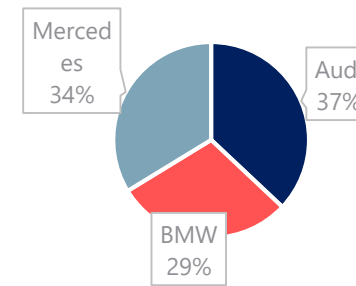|	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

In Excel, we can either hard code the frequencies or count them with a count function. This will come up later on.
Total formula: =SUM()

Sales (Bar chart)
Frequency — Audi: 124, BMW: 98, Mercedes: 113
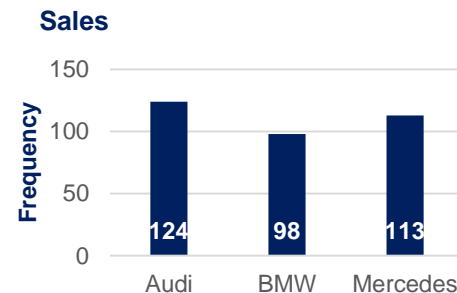
Bar charts are also called clustered column charts in Excel. Choose your data, **Insert -> Charts -> Clustered column or Bar chart.**
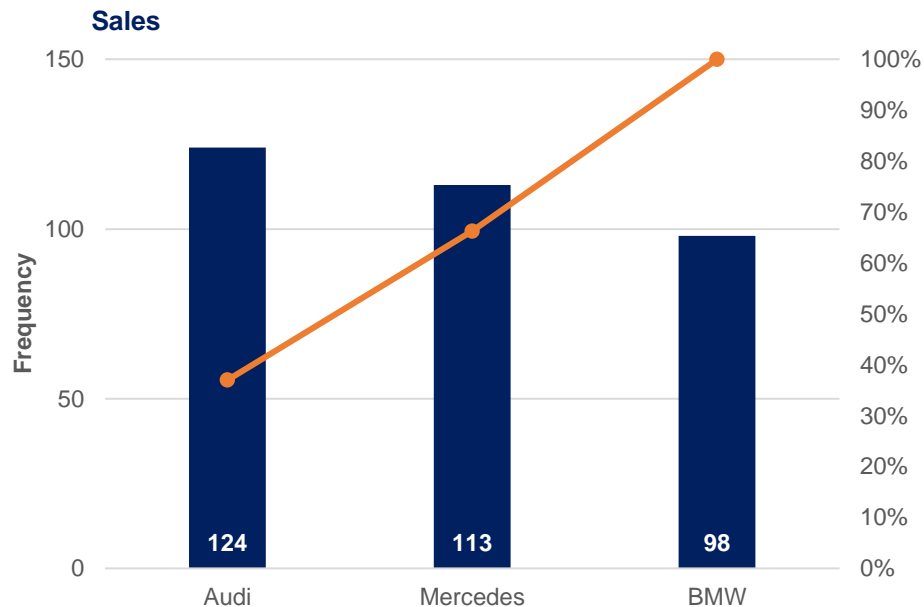
Pie chart — Mercedes 34%, Audi 37%, BMW 29%

Pie charts are created in the following way:
Choose your data, **Insert -> Charts -> Pie chart**

Sales (Pareto diagram)
Frequency — Audi: 124, Mercedes: 113, BMW: 98
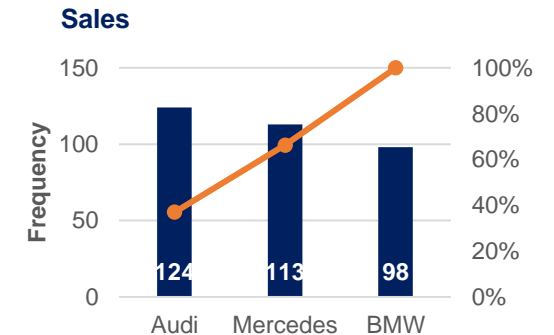
Next slide.

Pareto diagrams in Excel

Sales

Creating Pareto diagrams in Excel:

1. Order the data in your frequency distribution table in descending order.
2. Create a bar chart.
3. Add a column in your frequency distribution table that measures the cumulative frequency.
4. Select the plot area of the chart in Excel and **Right click**.
5. Choose **Select series**.
6. Click **Add**
7. Series name doesn't matter. You can put 'Line'
8. For **Series values** choose the cells that refer to the cumulative frequency.
9. Click **OK.** *You should see two side-by-side bars.*
10. Select the plot area of the chart and **Right click**.
11. Choose **Change Chart Type.**
12. Select **Combo.**
13. Choose the type of representation from the dropdown list. Your initial categories should be **'Clustered Column'.** Change the second series, that you called 'Line', to **'Line'.**
14. Done.

365 DataScience

Numerical variables. Frequency distribution table and histogram

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

Frequency distribution tables for numerical variables are different than the ones for categorical. Usually, they are divided into intervals of equal (or unequal) length. The tables show the interval, the absolute frequency and sometimes it is useful to also include the relative (and cumulative) frequencies.

The interval width is calculated using the following formula:

$$Interal\ width = \frac{Largest\ number\ -\ smallest\ number}{Number\ of\ desired\ intervals}$$

Creating the frequency distribution table in Excel:

1. Decide on the number of intervals you would like to use.
2. Find the interval width (using a the formula above).
3. Start your 1st interval at the lowest value in your dataset.
4. Finish your 1st interval at the lowest value + the interval width. (= start_interval_cell + interval_width_cell)
5. Start your 2nd interval where the 1st stops (that's a formula as well - just make the starting cell of interval 2 = the ending of interval 1)
6. Continue in this way until you have created the desired number of intervals.
7. Count the absolute frequencies using the following COUNTIF formula:
   =COUNTIF(dataset_range,">="&interval start) -COUNTIF(dataset_range,">"&interval end).
8. In order to calculate the relative frequencies, use the following formula: = absolute_frequency_cell / number_of_observations
9. In order to calculate the cumulative frequencies:
   i. The first cumulative frequency is equal to the relative frequency
   ii. Each consequitive cumulative frequency = previous cumulative frequency + the respective relative frequency

Note that all formulas could be found in the lesson Excel files and the solutions of the exercises provided with each lesson.

365 DataScience

Numerical variables. Frequency distribution table and histogram
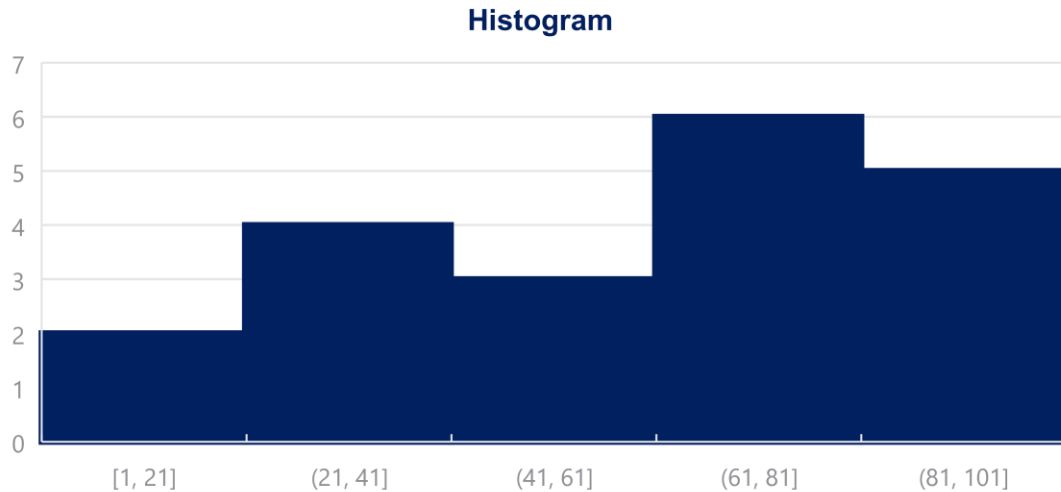
Histogram

Histograms are the one of the most common ways to represent numerical data. Each bar has width equal to the width of the interval. The bars are touching as there is continuation between intervals: where one ends -> the other begins.

Creating a histogram in Excel:

1. Choose your data
2. **Insert -> Charts -> Histogram**
3. To change the number of bins (intervals):
    1. Select the x-axis
    2. Click **Chart Tools -> Format -> Axis options**
    3. You can select the bin width (interval width), number of bins, etc.

Histogram relative frequency

365 DataScience
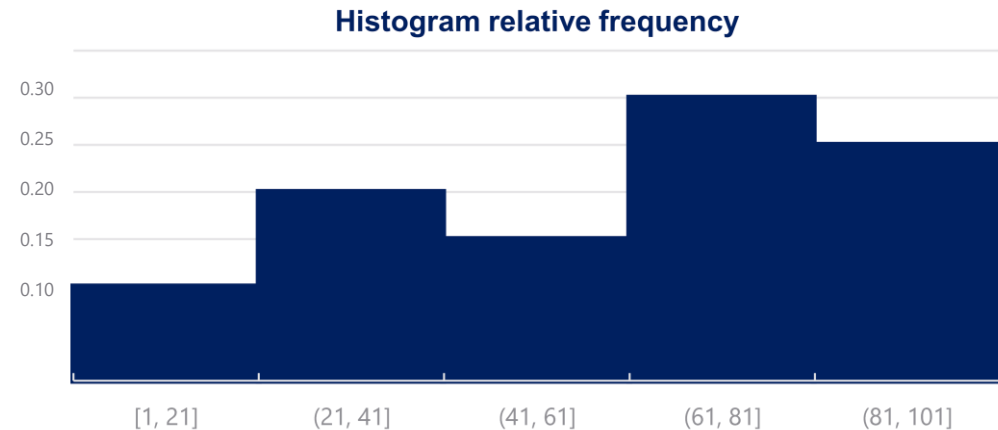
Graphs and tables for relationships between variables. Cross tables

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	0.10	0.20	0.04	0.35
Bonds	0.20	0.00	0.03	0.23
Real Estate	0.10	0.17	0.16	0.42
Total	0.40	0.37	0.23	1.00

Cross tables (or contingency tables) are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals. Sometimes, these tables are constructed with the *relative frequencies* as shown in the table below.

A common way to represent the data from a cross table is by using a side-by-side bar chart.
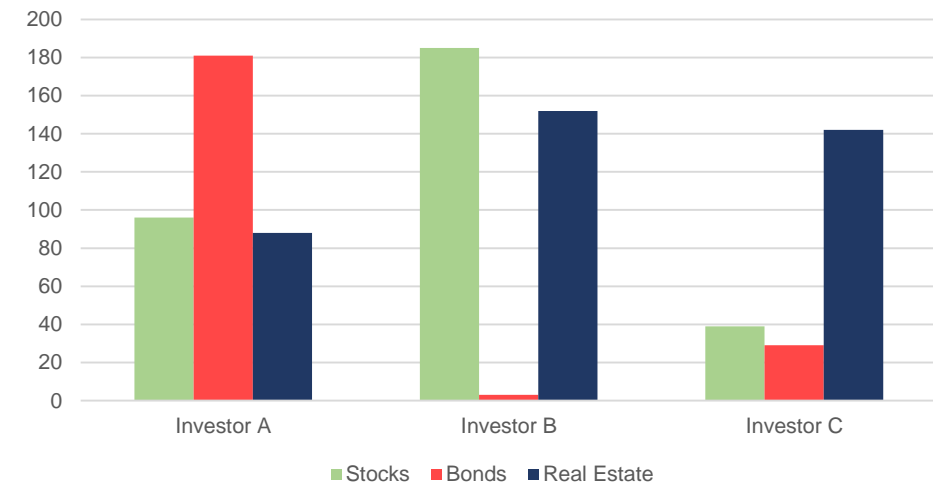
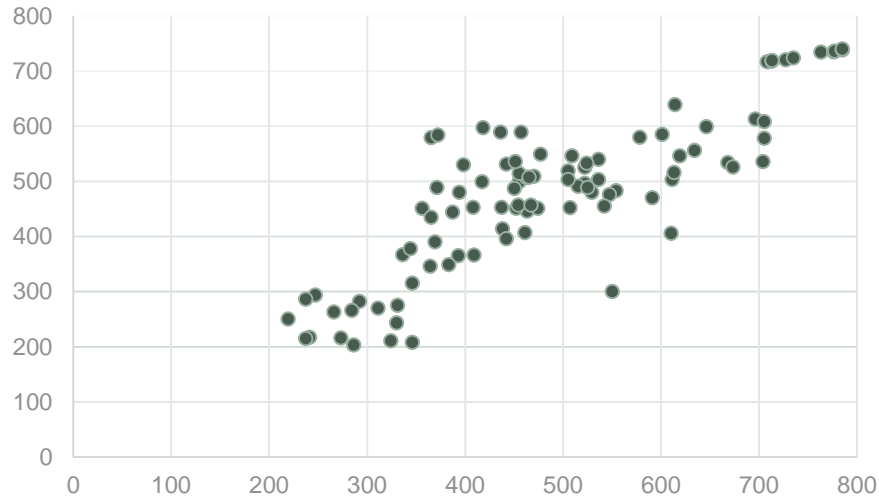X⊞ Creating a side-by-side chart in Excel:

1. Choose your data
2. Insert -> Charts -> Clustered Column

Selecting more than one series (groups of data) will automatically prompt Excel to create a side-by-side bar (column) chart.


Side-by-side bar chart

365 DataScience

Graphs and tables for relationships between variables. Scatter plots

When we want to represent two numerical variables on the same graph, we usually use a scatter plot. Scatter plots are useful especially later on, when we talk about regression analysis, as they help us detect patterns (linearity, homoscedasticity).

Scatter plots usually represent lots and lots of data. Typically, we are not interested in single observations, but rather in the structure of the dataset.

Creating a scatter plot in Excel:

1. Choose the two datasets you want to plot.
2. Insert -> Charts -> Scatter
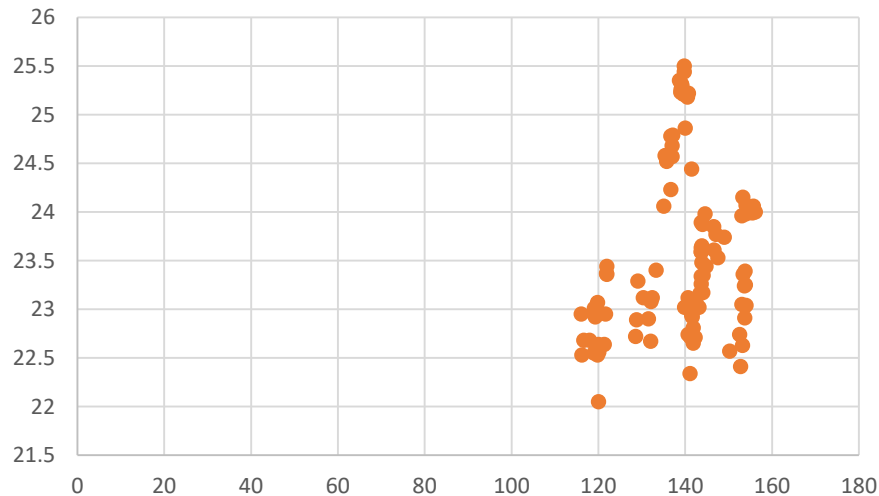
A scatter plot that looks in the following way (down) represents data that **doesn't have a pattern.** Completely vertical 'forms' show no association.

Conversely, the plot above shows a linear pattern, meaning that the observations move together.

Mean, median, mode

Mean

The mean is the most widely spread measure of central tendency. It is the simple average of the dataset.

Note: easily affected by outliers

The formula to calculate the mean is:

$$\frac{\sum_{i=1}^{N} x_i}{N} \quad \text{or}$$

$$\frac{x_1 + x_2 + x_3 + \cdots + x_{N-1} + x_N}{N}$$

In Excel, the mean is calculated by:

=AVERAGE()

Median

The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.

In an ordered dataset, the median is the number at position $\frac{n+1}{2}$.

If this position is not a whole number, it, the median is the simple average of the two numbers at positions closest to the calculated value.

In Excel, the median is calculated by:

=MEDIAN()

Mode

The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes.
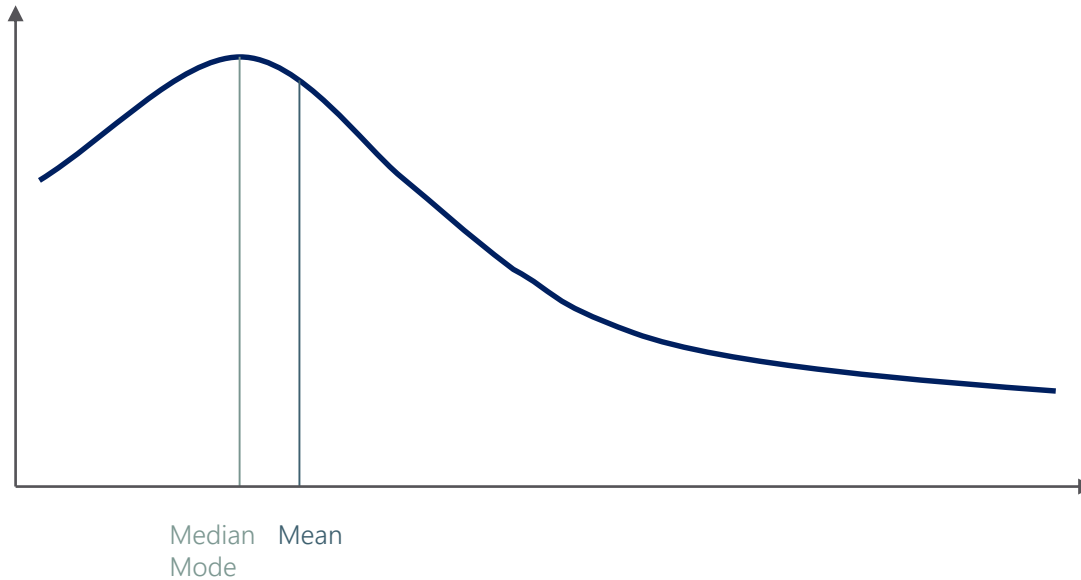
The mode is calculated simply by finding the value with the highest frequency.

In Excel, the mode is calculated by:

=MODE.SNGL() -> returns one mode

=MODE.MULT() -> returns an array with the modes. It is used when we have more than 1 mode.

Skewness

Median Mean
Mode

Skewness is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side.

Right (positive) skewness looks like the one in the graph. It means that the **outliers** are to the right (long tail to the right).

Left (negative) skewness means that the outliers are to the left.
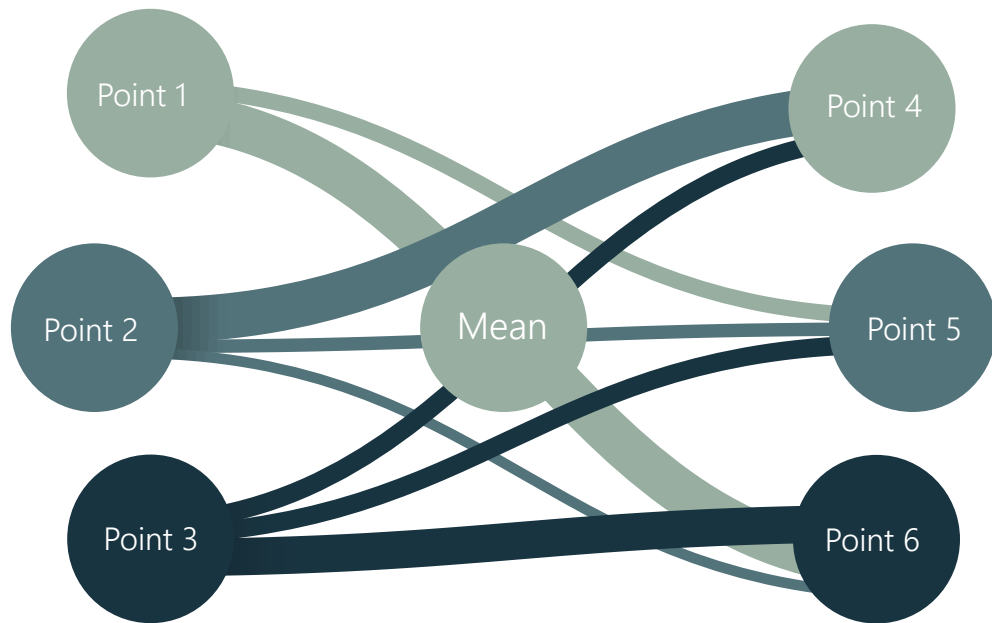
Usually, you will use software to calculate skewness.

Calculating skewness in Excel:

=SKEW()

Formula to calculate skewness: $\dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^3}$

Variance and standard deviation

Variance and standard deviation measure the dispersion of a set of data points around its mean value.

There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas. More on the mathematics behind it.

Sample variance formula:

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

Population variance formula:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample standard deviation formula:

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

Population standard deviation formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

📊 Calculating variance in Excel:

Sample variance: =VAR.S()
Population variance: =VAR.P()
Sample standard deviation: = STDEV.S()
Population standard deviation: =STDEV.P()

365√DataScience

Covariance and correlation

Covariance

Covariance is a measure of the joint variability of two variables.

➢ A positive covariance means that the two variables move together.
➢ A covariance of 0 means that the two variables are independent.
➢ A negative covariance means that the two variables move in opposite directions.

Covariance can take on values from $-\infty$ to $+\infty$. This is a problem as it is very hard to put such numbers into perspective.

Sample covariance formula:

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

Population covariance formula:

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

In Excel, the covariance is calculated by:

Sample covariance: =COVARIANCE.S()
Population covariance: =COVARIANCE.P()

Correlation

Correlation is a measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result.

➢ A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other.
➢ A correlation of 0 means that the variables are independent.
➢ A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.

Sample correlation formula: $\mathbf{r} = \dfrac{S_{xy}}{S_x S_y}$

Population correlation formula: $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

In Excel, correlation is calculated by:

=CORREL()