Iliya Valchanov

Statistics

Course Notes

365 DataScience

Table of Contents

Abstract 3
1. Descriptive Statistics 4
1.1 Types of Data 4
1.2 Levels of Measurment 5
1.3 Graphs and Tables that Represent Categorical Variables 6
1.3.1 Excel Formulas 7
1.3.2 Pareto Diagrams in Excel 8
1.4 Graphs and Tables that Represent Numerical Variables 9
1.4.1 Frequency Distriubution Table and Histogram 10
1.5 Graphs and Tables for Relationships Between Variables 11
1.5.1 Cross Tables 11
1.5.2 Scatter Plots 12
1.6 Mean, Median and Mode 13
1.7 Skewness 14
1.8 Variance and Standard Deviation 15
1.9 Covariance and Correlation 16
2. Inferential Statistics 17
2.1 Distributions 17
2.1.1 The Normal Distribution 18
2.1.2 The Standard Normal Distribution 21
2.2 The Central Limit Theorem 22
2.3 Estimators and Estimates 23
2.4 Confidence Intervals and the Margin of Error 24
2.5 Student's T Distribution 25
2.6 Formulas for Confidence Intervals 26
3. Hypothesis Testing 27
3.1 The Scientific Method 27
3.2 Hypotheses 28
3.3 Null Hypotheses 29
3.4 Decisions You Can Take 30
3.5 Level of Significance and Types of Tests 31
3.6 Statistical errors (Type I Error and Type II Error) 32
3.7 P-value 33
3.8 Formulae for Hypothesis Testing 34

Abstract

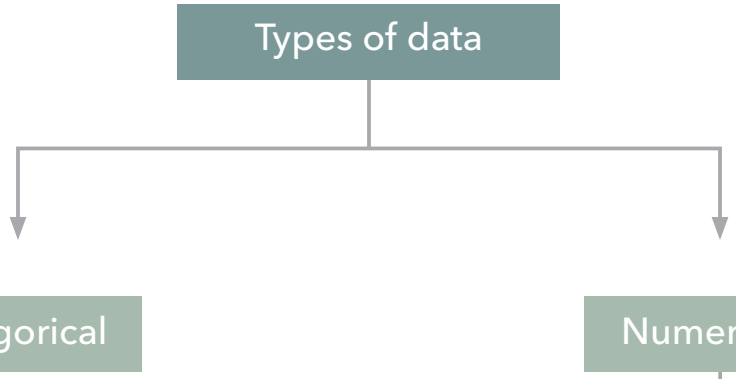
Statistics is an essential component in the ever-expanding field of data science playing an invaluable role in the making of informed business decisions. Statistical functions are applied on large sets of data to draw conclusions, make predictions, and minimize loss.

Therefore, if you want to enjoy a successful data science career, you need to have a solid grasp of statistical core concepts and basics, all covered in the statistics course notes. We start off with descriptive statistics, diving into all the associated graphs and tables for numerical and descriptive data. Then we take a look at inferential statistics, the different types of distributions, confidence intervals and respective formulas.
We finish off with the process of hypotheses testing, going into the types, examples formulas and the p-value.

Keywords: statistics, numerical data, categorical data, p-value, normal distribution, confidence interval, hypotheses testing

1. Descriptive Statistics

1.1 Types Of Data



Categorical data represents groups or categories.

Examples:

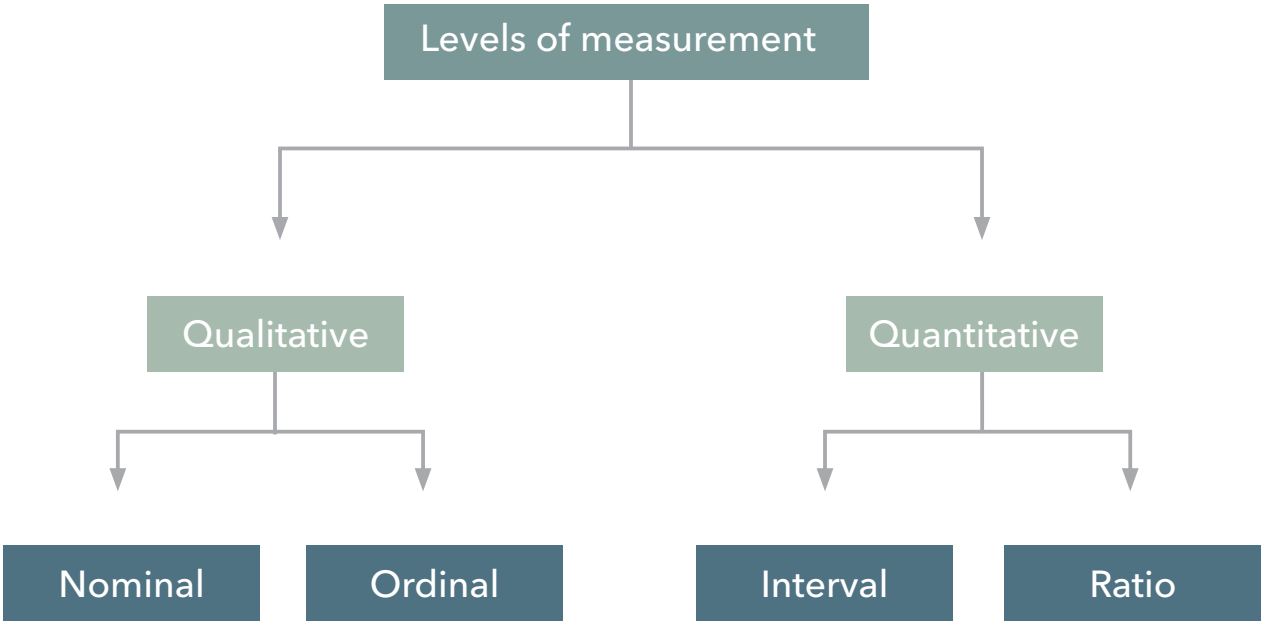
1. Car brands: Audi, BMW and Mercedes.

2. Answers to yes/no questions: yes and no

Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.

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

1.2 Levels of Measurement



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

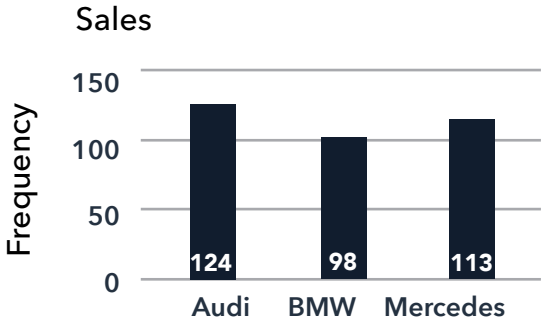
There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios have a true zero, while intervals don't.

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

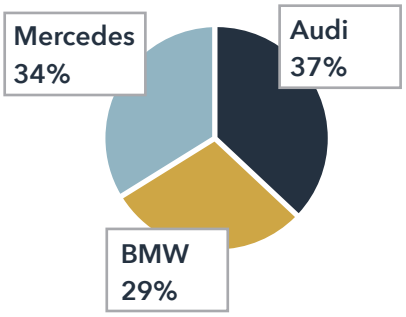
1.3 Graphs and Tables that Represent Categorical Variables

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

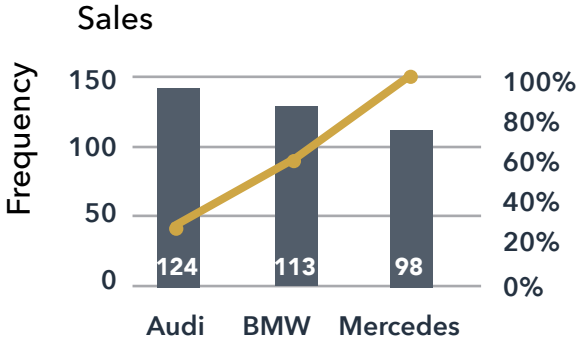
Frequency distribution tables show the category and its corresponding absolute frequency.



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.



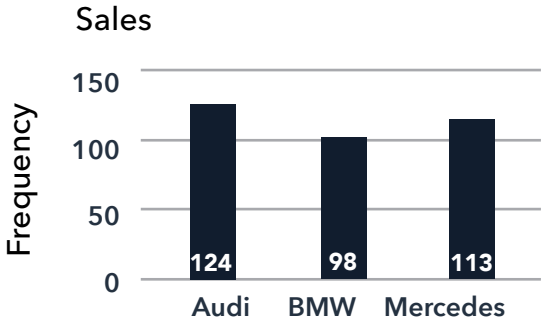
The Pareto diagram is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency.

1.3.1 Excel formulas

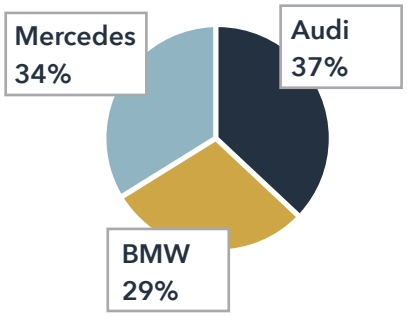
Frequency distribution tables → Bar charts → Pie charts → Pareto Diagrams

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

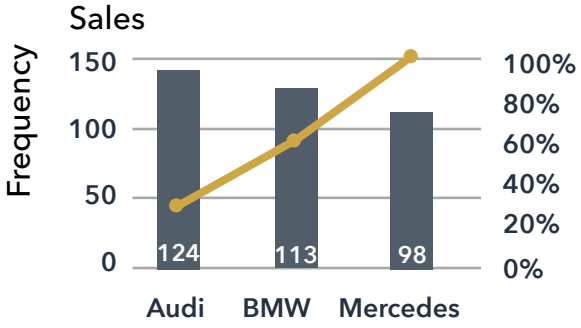
In Excel, we can either hard code the frequencies or count them with a count function. This will come up later on.
Total formula: =SUM()



Bar charts are also called clustered column charts in Excel. Choose your data, Insert -> Charts -> Clustered column or Bar chart.

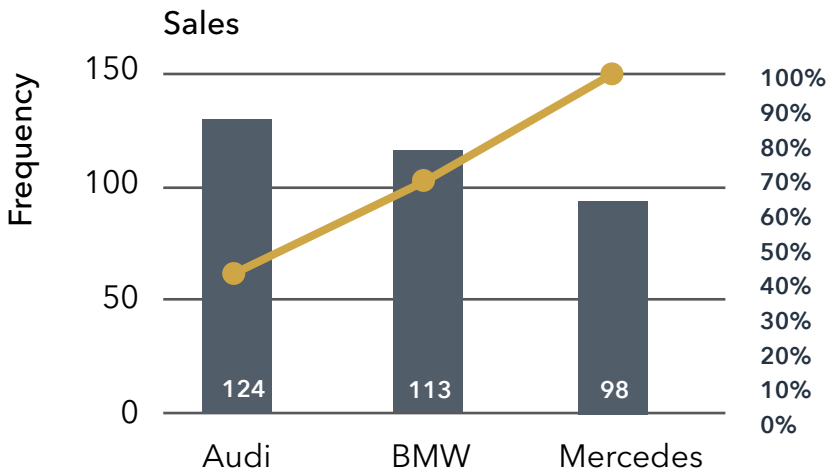


Pie charts are created in the following way:
Choose your data, Insert -> Charts -> Pie chart



Next slide

1.3.2 Pareto Diagrams in Excel



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.

1.4 Graphs and tables that represent numerical variables
1.4.1 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:

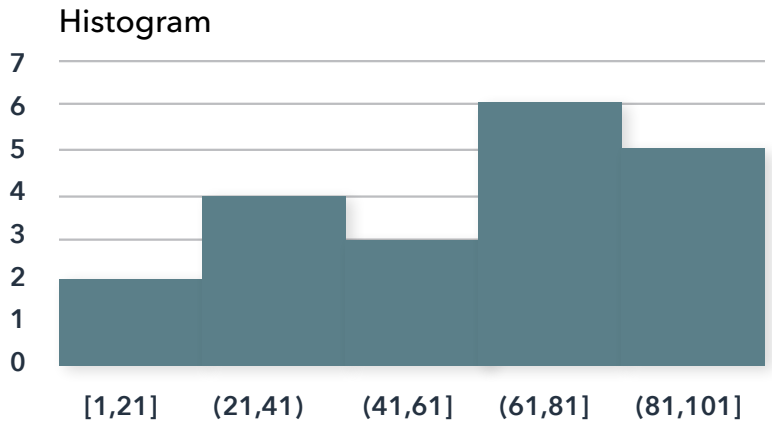
$$Interal\ width = \frac{Largest\ number - smallest\ number}{Number\ of\ desired\ intervals}$$

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:

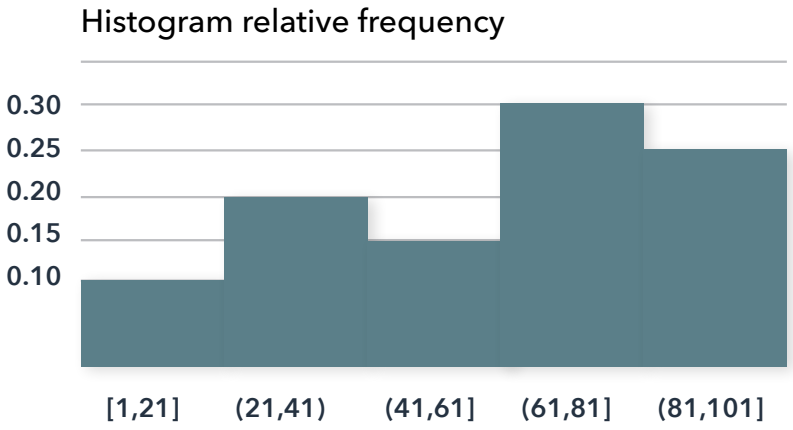
   I. The first cumulative frequency is equal to the relative frequency
   II. Each consequitive cumulative frequency = previous cumulative frequency the respective relative frequency

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



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.



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.

1.5 Graphs and Tables for Relationships Between Variables.
1.5.1 Cross Tables

Type of investment / Investor	Investor A	Investor B	Investor C	Total
Stoks	96	185	39	320
Bonds	181	388	29	213
real Estate	88	152	142	382
Total	365	340	210	915

Type of investment / Investor	Investor A	Investor B	Investor C	Total
Stoks	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

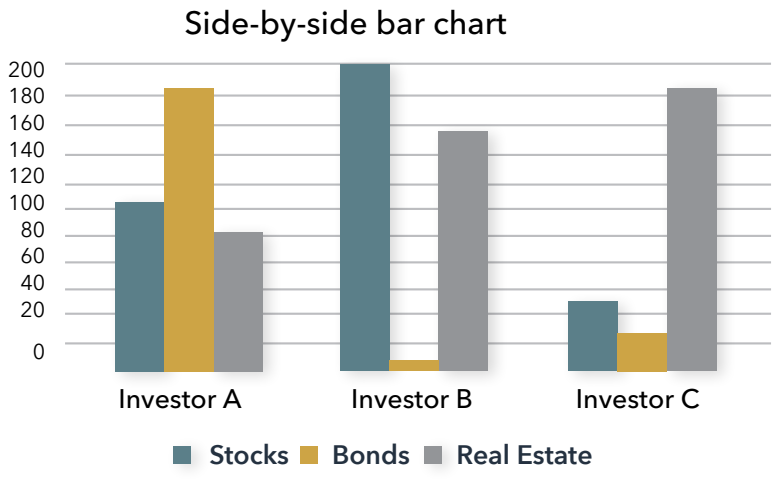
Cross tables (or contingency tables) are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals. Sometimes, these tables are constructed with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

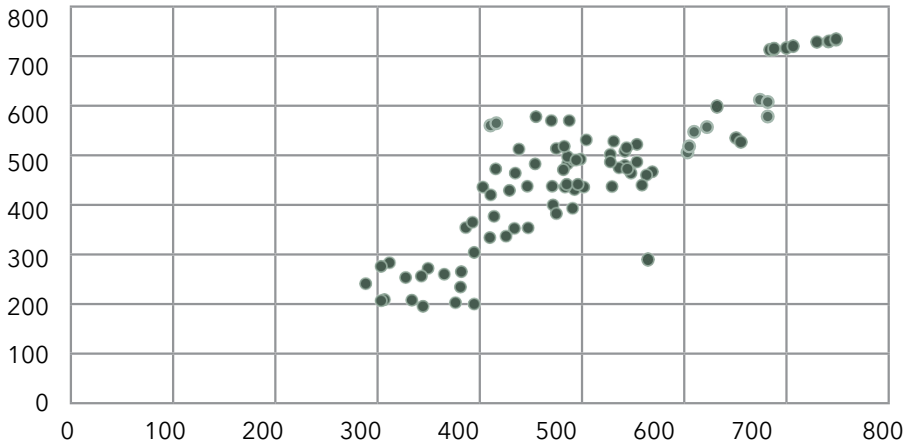
Creating a side-by-side chart in Excel:

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.



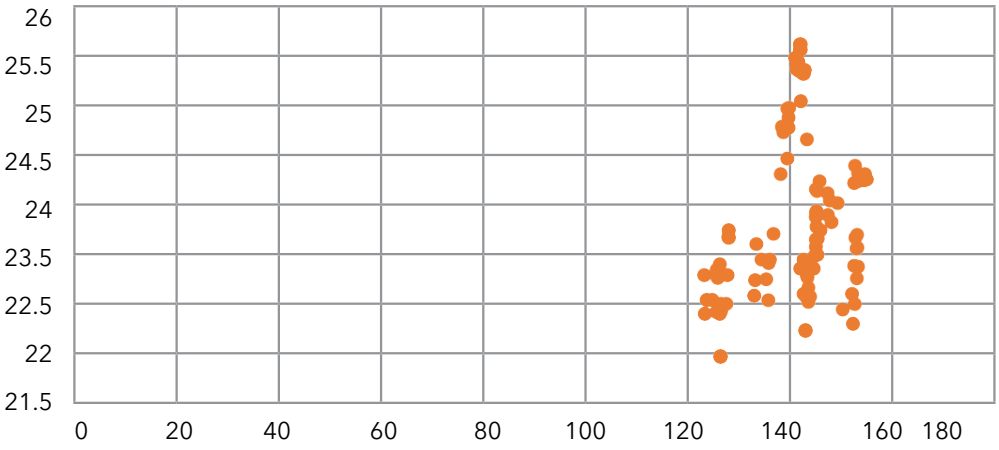
1.5.2 Scatter Plots



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.

1.6 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:

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

$$\frac{x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n}{n}$$

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.

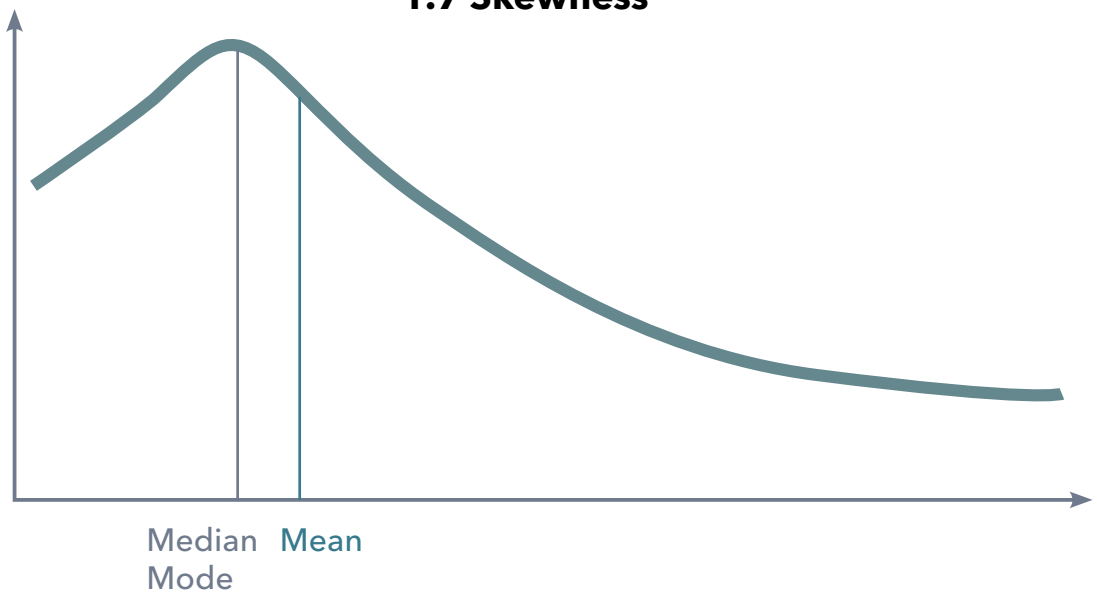
The mode is calculated simply by finding the value with the highest frequency.

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.

1.7 Skewness



Calculating skewness in Excel:

=SKEW()

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).

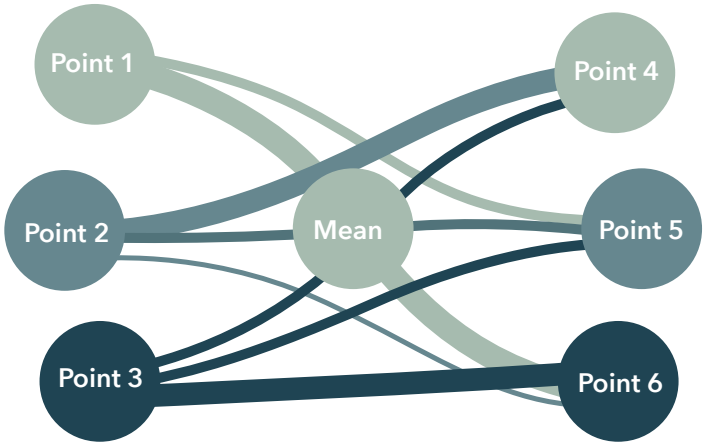
Left (negative) skewness means that the outliers are to the left.

Usually, you will use software to calculate skewness.

Formula to calculate skewness:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sqrt[3]{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

1.8 Variance and Standard Deviation



Calculating variance in Excel:

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

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:

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

1.9 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 -∞ to +∞ . 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:

$$r = \frac{s_{xy}}{s_x s_y}$$

Population correlation formula:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

In Excel, correlation is calculated by:
=CORREL()

2. Inferential Statistics

2.1 Distributions

Definition

In statistics, when we talk about distributions we usually mean probability distributions.

Definition (informal): A distribution is a function that shows the possible values for a variable and how often they occur.

Definition (Wikipedia): In probability theory and statistics, a probability distribution is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

Examples: Normal distribution, Student's T distribution, Poisson distribution, Uniform distribution, Binomial distribution

Graphical representation

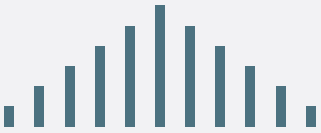
It is a common mistake to believe that the distribution is the graph. In fact the distribution is the 'rule' that determines how values are positioned in relation to each other.

Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them.

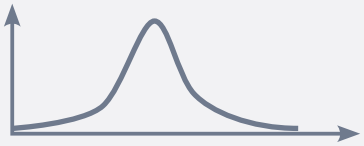
Examples:

Uniform distribution

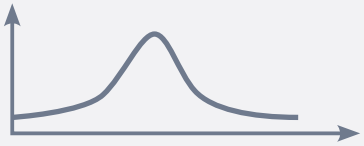
Binomial distribution



Normal distribution



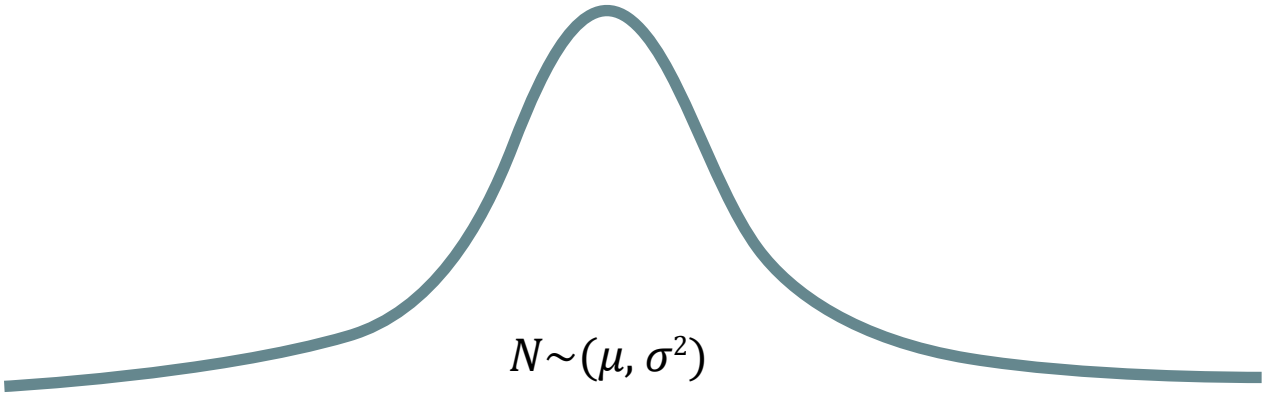
Student's T distribution



2.1.1 The Normal Distribution

The Normal distribution is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons:

- It approximates a wide variety of random variables
- Distributions of sample means with large enough samples sizes could be approximated to normal
- All computable statistics are elegant
- Heavily used in regression analysis
- Good track record

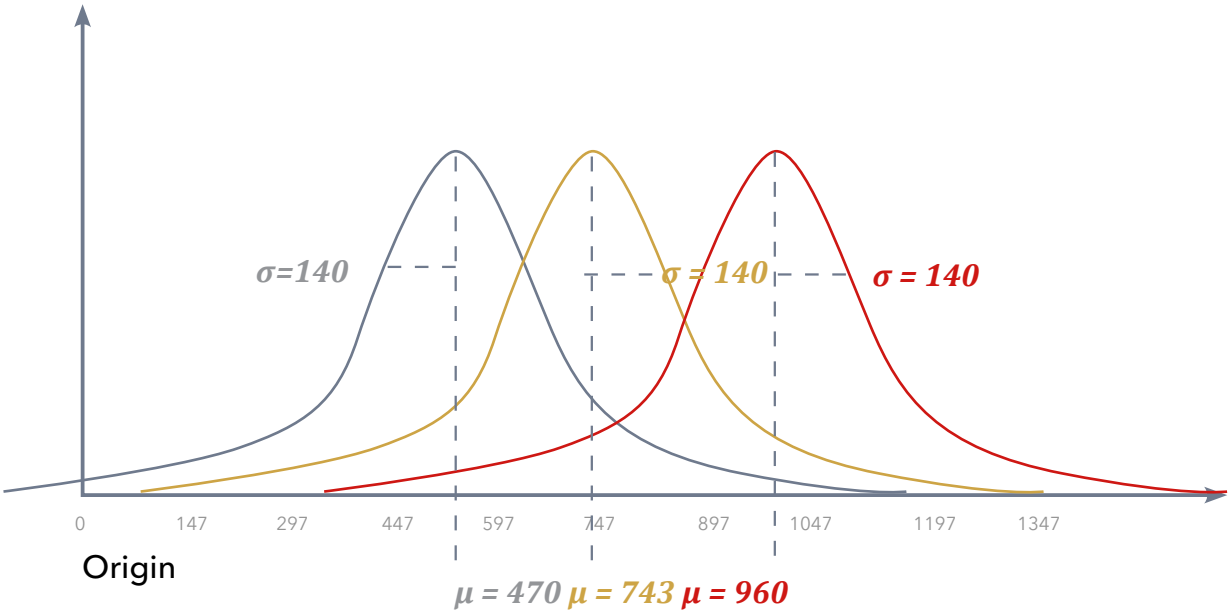


N stands for normal;
$\sim$ stands for a distribution;
μ is the mean;
$\sigma 2$ is the variance.

Examples:

- Biology. Most biological measures are normally distributed, such as: height; length of arms, legs, nails; blood pressure; thickness of tree barks, etc.
- IQ tests
- Stock market information

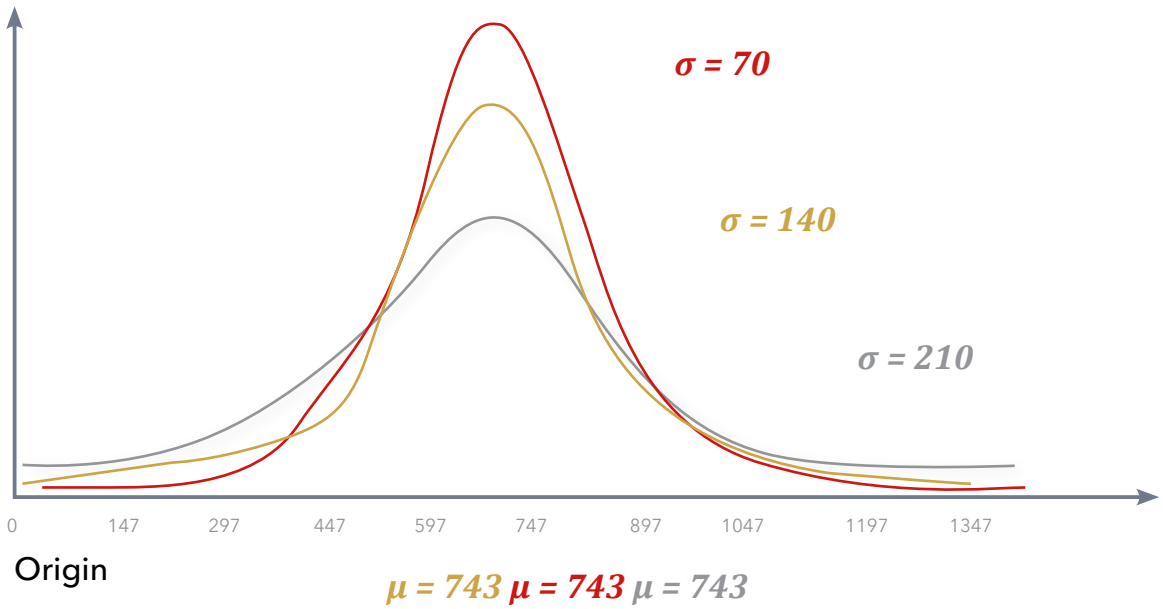
Controlling for the standard deviation



Keeping the standard deviation constant, the graph of a normal distribution with:

- a smaller mean would look in the same way, but be situated to the left (in gray)
- a larger mean would look in the same way, but be situated to the right (in red)

Controlling for the mean



Keeping the mean constant, a normal distribution with:

- a smaller standard deviation would be situated in the same spot, but have a higher peak and thinner tails (in red)
- a larger standard deviation would be situated in the same spot, but have a lower peak and fatter tails (in gray)

2.1.2 The Standard Normal Distribution

The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.

Every Normal distribution can be 'standardized' using the standardization formula:

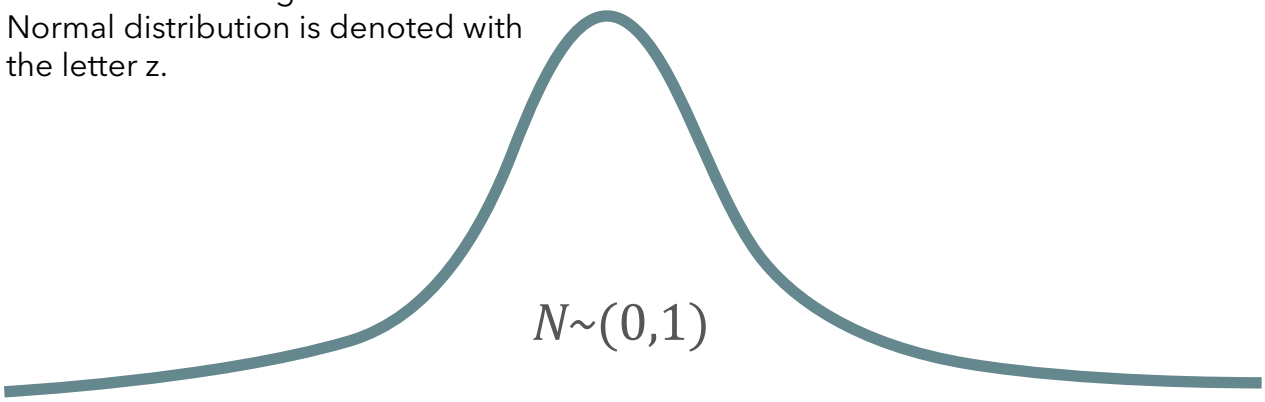
$$z = \frac{x - \mu}{\sigma}$$

A variable following the Standard Normal distribution is denoted with the letter z.

Why standardize?

Standardization allows us to:

- compare different normally distributed datasets
- detect normality
- detect outliers
- create confidence intervals
- test hypotheses
- perform regression analysis



Rationale of the formula for standardization:
We want to transform a random variable from $N\sim \mu, \sigma^2$ *to* $N\sim(0,1)$.
Subtracting the mean from all observations would cause a transformation from $N\sim \mu,\sigma^2$ to $N\sim 0, \sigma^2$, moving the graph to the origin.
Subsequently, dividing all observations by the standard deviation would cause a transformation from $N\sim 0, \sigma^2$ to $N\sim 0,1$, standardizing the peak and the tails of the graph.

2.2 The Central Limit Theorem

The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. sum of rolled numbers when rolling dice).

The theorem	Why is it useful?	Where can we see it?
• No matter the distribution • The distribution of $\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, \ldots, \overline{x_n}$ would tend to $N \sim \left(\mu, \frac{\sigma^2}{n}\right)$ • The more samples, the closer to Normal $(k \to \infty)$ • The bigger the samples, the closer to Normal $(n \to \infty)$	The CLT allows us to assume normality for many different variables. That is very useful for confidence intervals, hypothesis testing, and regression analysis. In fact, the Normal distribution is so predominantly observed around us due to the fact that following the CLT, many variables converge to Normal. Click here for a CLT simulator.	Since many concepts and events are a sum or an average of different effects, CLT applies and we observe normality all the time. For example, in regression analysis, the dependent variable is explained through the sum of error terms.

2.3 Estimators and Estimates

Estimators ⟷ Estimates

Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information.

Examples of estimators and the corresponding parameters:

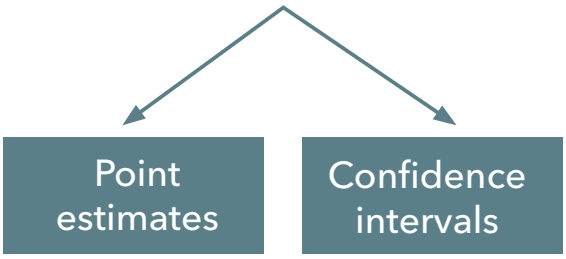
Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	r	ρ

Estimators have two important properties:

- Bias
  The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.

- Efficiency
  The most efficient estimator is the one with the smallest variance.

An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.

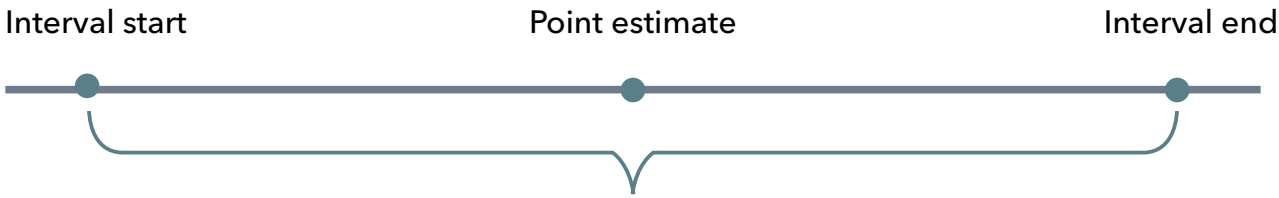


A single value.
Examples:
- 1
- 5
- 122.67
- 0.32

An interval.
Examples:
- (1 , 5)
- (12 , 33)
- (221.78 , 745.66)
- (- 0.71 , 0.11)

Confidence intervals are much more precise than point estimates. That is why they are preferred when making inferences.

2.4 Confidence Intervals and the Margin of Error



Definition: A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall.
We build the confidence interval around the point estimate.
(1-α) is the level of confidence. We are (1-α)*100% confident that the population parameter will fall in the specified interval. Common alphas are: 0.01, 0.05, 0.1.

General formula:

$[\bar{x} - ME, \bar{x} + ME\ ]$, where ME is the margin of error.

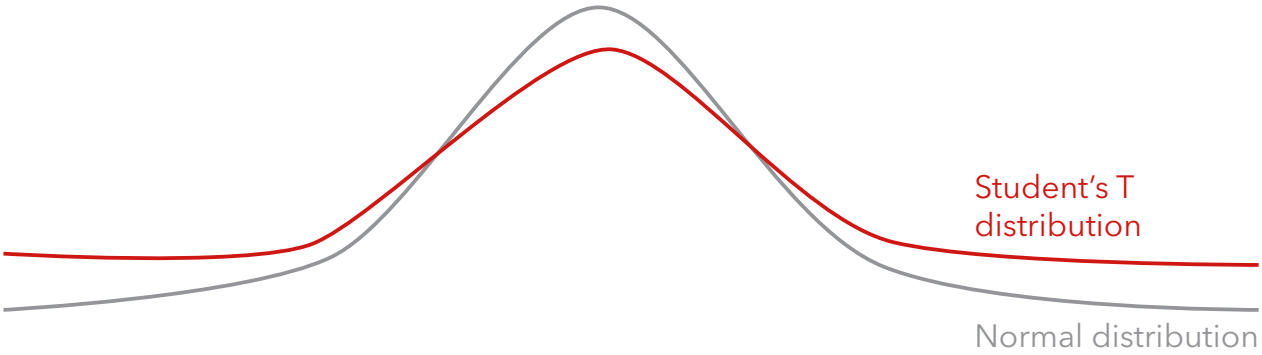
Term	Effect on width of CI
$(1\text{-}\alpha)\uparrow$	↑
$\boldsymbol{\sigma}\uparrow$	↑
$n\uparrow$	↓

$$ME = \text{reliability factor} * = \frac{standard\ deviation}{\sqrt{sample\ size}} \rightarrow z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad t_{v,\alpha/2} * \frac{s}{\sqrt{n}}$$

2.5 Student's T Distribution

The Student's T distribution is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are small. It is particularly useful when we don't have enough information or it is too costly to obtain it.

All else equal, the Student's T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size.



A random variable following the t-distribution is denoted $t_{\upsilon,\alpha}$, where υ are the degrees of freedom.

We can obtain the student's T distribution for a variable with a Normally distributed population using the formula:

$$t_{\upsilon,\alpha} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

2.6 Formulas for Confidence Intervals

# populations	Population variance	Samples	Statistic	Variance	Formula for test statistic
One	known	-	z	$\boldsymbol{\sigma^2}$	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
One	unknown	-	t	s^2	$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$
Two	-	dependent	t	$s^2 difference$	$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$
Two	known	independent	z	σ_x^2, σ_y^2	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$(\bar{x} - \bar{y}) \pm t_{v,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$
Two	unknown, assumed different	independent	t	s_x^2, s_y^2	$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$

3. Hypothesis Testing

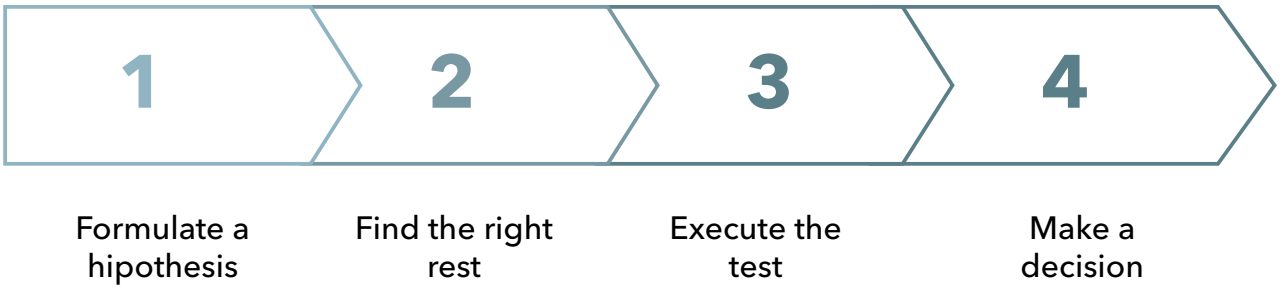
3.1 Scientific method

The 'scientific method' is a procedure that has characterized natural science since the 17th century. It consists in systematic observation, measurement, experiment, and the formulation, testing and modification of hypotheses.

Since then we've evolved to the point where most people and especially professionals realize that pure observation can be deceiving. Therefore, business decisions are increasingly driven by data. That's also the purpose of data science.

While we don't 'name' the scientific method in the videos, that's the underlying idea. There are several steps you would follow to reach a data-driven decision (pictured).

Steps in data-driven decision making



3.2 Hypotheses

A hypothesis is
"an idea that can be tested"

It is a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

Null hypothesis (H_0)

The null hypothesis is the hypothesis to be tested.

It is the status-quo. Everything which was believed until now that we are contesting with our test.

The concept of the null is similar to: innocent until proven guilty. We assume innocence until we have enough evidence to prove that a suspect is guilty.

Alternative hypothesis (H_1 or H_A)

The alternative hypothesis is the change or innovation that is contesting the status-quo.

Usually the alternative is our own opinion. The idea is the following:

If the null is the status-quo (i.e., what is generally believed), then the act of performing a test, shows we have doubts about the truthfulness of the null. More often than not the researcher's opinion is contained in the alternative hypothesis.

3.3 Null Hypotheses

A hypothesis is
"an idea that can be tested"

After a discussion in the Q&A, we have decided to include further clarifications regarding the null and alternative hypotheses.

As per the above logic, in the video tutorial about the salary of the data scientist, the null hypothesis should have been: Data Scientists do not make an average of $113,000.
In the second example the null Hypothesis should have been: The average salary should be less than or equal to $125,000.
Please explain further.

Student's question

Now note that the statement in the question is **NOT** true.

Instructor's answer (with some adjustments)

'I see why you would ask this question, as I asked the same one right after I was introduced to hypothesis testing. In statistics, the null hypothesis is the statement **we are trying to reject.** Think of it as the 'status-quo'. The alternative, therefore, is **the change or innovation.**

Example 1: So, for the data scientist salary example, the null would be: **the mean data scientist salary is $113,000.** Then we will try to reject the null with a statistical test. So, usually, your personal opinion (e.g. data scientists don't earn exactly that much) is the **alternative hypothesis.**

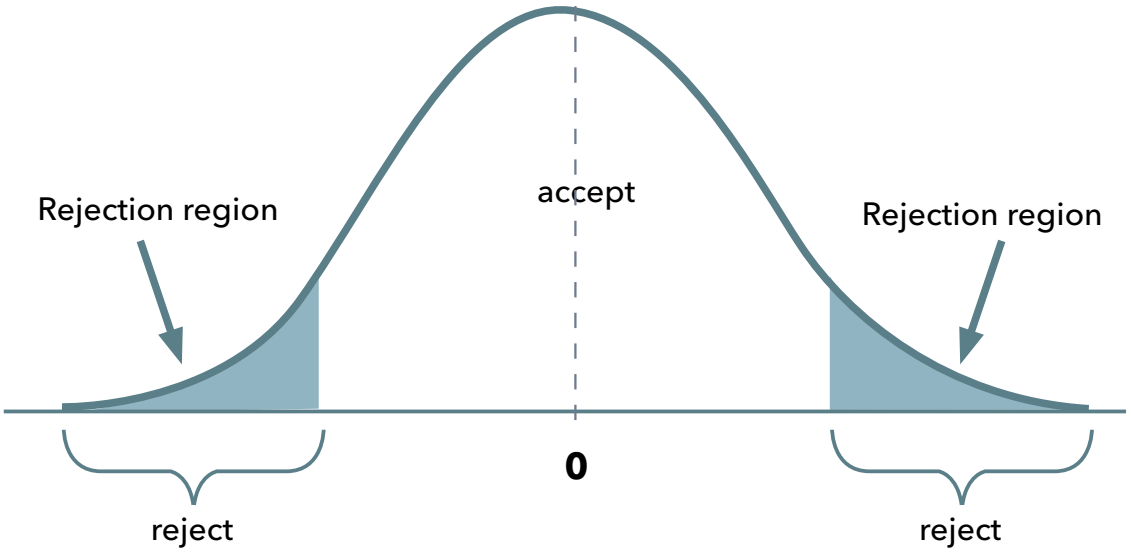
Example 2: Our friend Paul told us that the mean salary is >=$125,000 (status- quo, null). Our opinion is that he may be wrong, so we are testing that. Therefore, the alternative is: the mean data scientist salary **is lower than $125,000.**

It truly is counter-intuitive in the beginning, but later on, when you start doing the exercises, you will understand the mechanics.'

3.4 Decisions You Can Take

When testing, there are two decisions that can be made: to **accept** the null hypothesis or to **reject** the null hypothesis.

To accept the null means that there isn't enough data to support the change or the innovation brought by the alternative. To reject the null means that there is enough statistical evidence that the status-quo is not representative of the truth.



Given a two-tailed test:

Graphically, the tails of the distribution show when we reject the null hypothesis ('rejection region').

Everything which remains in the middle is the 'acceptance region'.

The rationale is: if the observed statistic is too far away from 0 (depending on the significance level), we reject the null. Otherwise, we accept it

Different ways of reporting the result:

Accept	Reject
At x% significance, we accept the null hypothesis	At x% significance, we reject the null hypothesis
At x% significance, A is not significantly different from B	At x% significance, A is significantly different from B
At x% significance, there is not enough statistical evidence that… At x% significance, we cannot reject the null hypothesis	At x% significance, there is enough statistical evidence… At x% significance, we cannot say that *restate the null*

3.5 Level of Significance and Types of Tests

Level of significance (α)

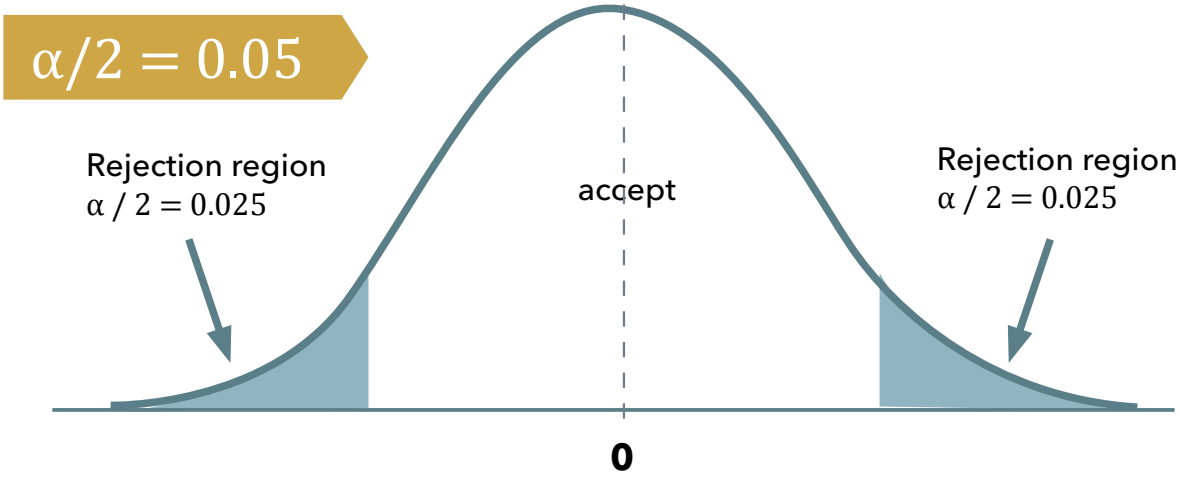
The probability of rejecting a null hypothesis that is true; the probability of making this error.

Common significance levels

0.10 0.05 0.01

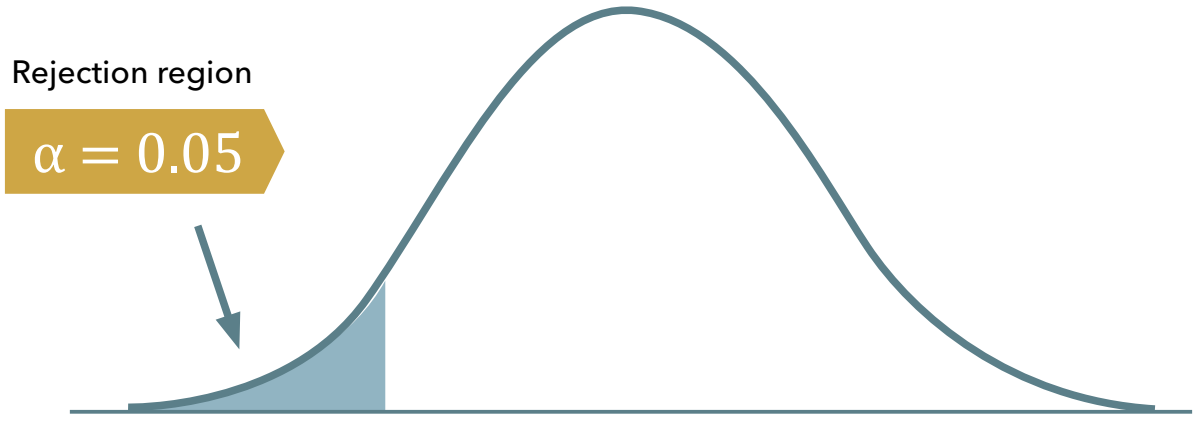
Two-sided (two-tailed) test

Used when the null contains an equality (=) or an inequality sign (≠)



One-sided (one-tailed) test

Used when the null doesn't contain equality or inequality sign (<,>,≤,≥)



3.6 Statistical Errors (Type I Error and Type II Error)

In general, there are two types of errors we can make while testing: Type I error (False positive) and Type II Error (False negative).

Statisticians summarize the errors in the following table:

Ho: Status quo			
		The truth	
		Ho is true	Ho is false
Ho (Status quo)	Accept		Type II error (False negative)
	Reject	Type I error (False positive)	

Here's the table with the example from the lesson:

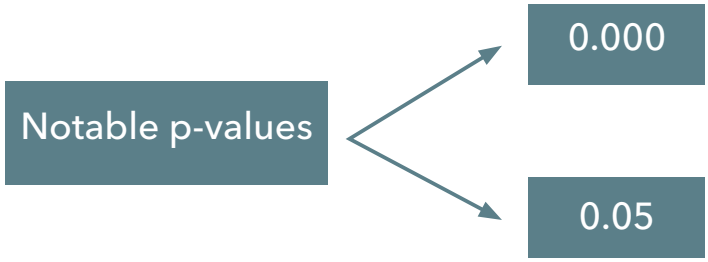
Ho: She doesn't like you			
		The truth	
		She doesn't like you	She likes you
Ho (Status quo) She doesn't like you (you should not invite her)	Accept (do nothing)		Type II error (False negative)
	Reject (invite her)	Type I error (False positive)	

The probability of committing Type I error (False positive) is equal to the significance level (α).
The probability of committing Type II error (False negative) is equal to the beta (β).
If you want to find out more about statistical errors, just follow this link for an article written by your instructor.

3.7 P-value

P-value

The p-value is the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic



When we are testing a hypothesis, we always strive for those 'three zeros after the dot'. This indicates that we reject the null at all significance levels.
0.05 is often the 'cut-off line'. If our p-value is higher than 0.05 we would normally accept the null hypothesis (equivalent to testing at 5% significance level). If the p-value is lower than 0.05 we would reject the null.

Where and how are p-values used?

- Most statistical software calculates p-values for each test
- The researcher can decide the significance level post-factum
- p-values are usually found with 3 digits after the dot (x.xxx)
- The closer to 0.000 the p-value, the better

Should you need to calculate a p-value 'manually', we suggest using an online p-value calculator, e.g. this one.

3.8 Formulae for Hypothesis Testing

# populations	Population variance	Samples	Statistic	Variance	Formula for test statistic
One	known	-	z	σ^2	$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
One	unknown	-	t	s^2	
Two		dependent	t	$s^2_{difference}$	$T = \frac{\bar{d} - \mu_0}{\frac{s_d}{\sqrt{n}}}$
Two	known	independent	z	σ_x^2, σ_y^2	$Z = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$T = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$

Copyright 2022 365 Data Science Ltd. Reproduction is forbidden unless authorized. All rights reserved.

Decision rule
There are several ways to phrase the decision rule and they all have the same meaning.

Reject the null if:
1) |test statistic| > |critical value|
2) The absolute value of the test statistic is bigger than the absolute critical value
3) p-value < some significance level
most often 0.05

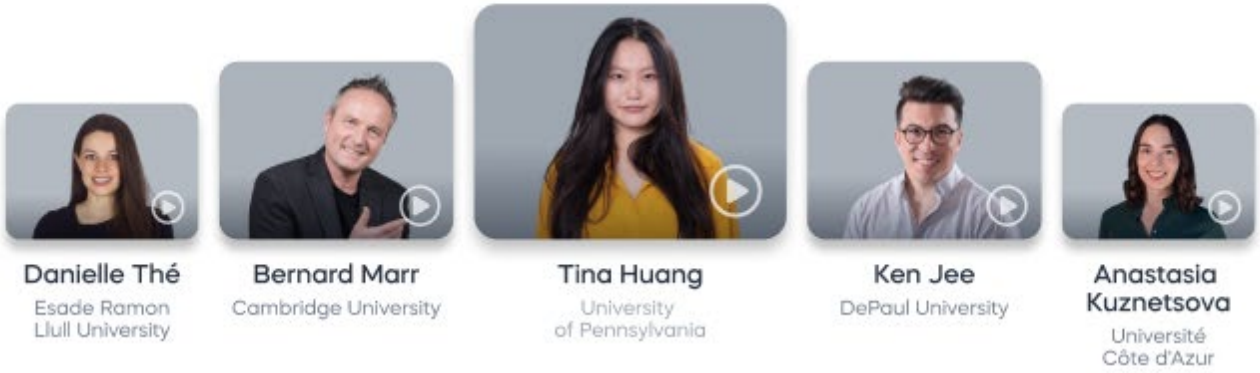
Usually, you will be using the p-value
to make a decision.

Copyright 2022 365 Data Science Ltd. Reproduction is forbidden unless authorized. All rights reserved.

Learn **DATA SCIENCE** anytime, anywhere, at your own pace.

If you found this resource useful, check out our **e-learning program**. We have everything you need to succeed in data science.

Learn the most sought-after data science skills from **the best experts in the field**! Earn a **verifiable certificate** of achievement trusted by employers worldwide and future proof your career.



Danielle Thé
Esade Ramon Llull University

Bernard Marr
Cambridge University

Tina Huang
University of Pennsylvania

Ken Jee
DePaul University

Anastasia Kuznetsova
Université Côte d'Azur

Comprehensive training, exams, certificates.

- ✓ 162 hours of video
- ✓ 599+ Exercises
- ✓ Downloadables
- ✓ Exams & Certification
- ✓ Personalized support
- ✓ Resume Builder & Feedback
- ✓ Portfolio advice
- ✓ New content
- ✓ Career tracks

Join a global community of 1.8 M successful students with an annual subscription at 60% OFF with coupon code **365RESOURCES**.

~~$432~~ **$172.80**/year

Start at 60% Off

VAT may be applied

365√DataScience

Iliya Valchanov

Email: team@365datascience.com

365 DataScience