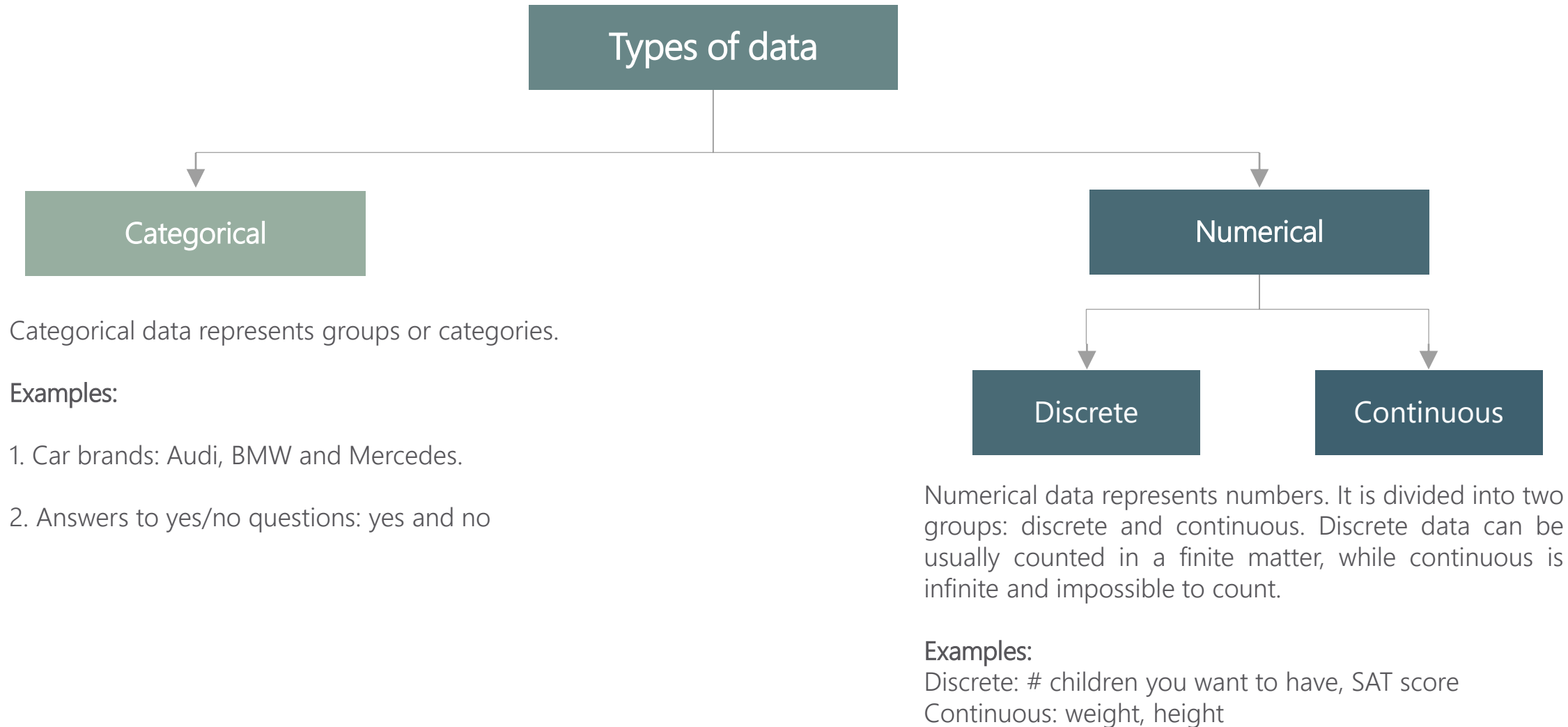
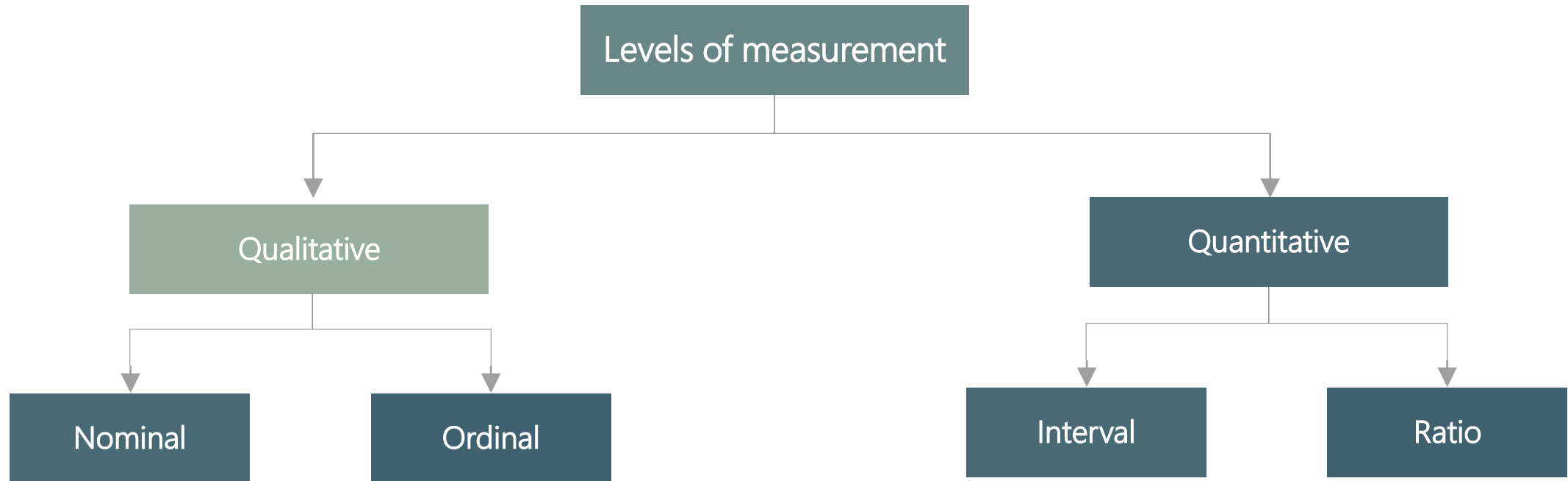


COURSE NOTES: DESCRIPTIVE STATISTICS

Types of data



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

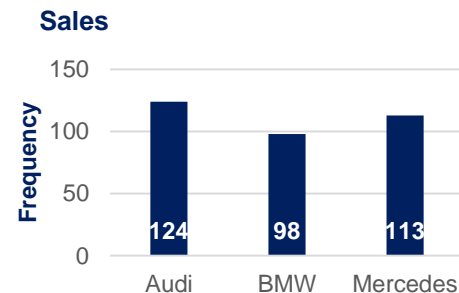
Graphs and tables that represent categorical variables

Frequency distribution tables

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

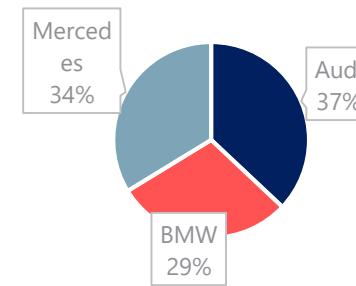
Frequency distribution tables show the category and its corresponding absolute frequency.

Bar charts



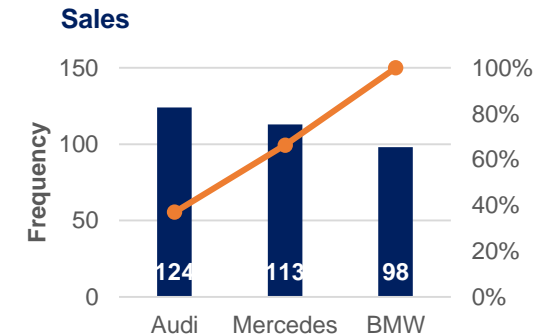
Bar charts are very common. Each bar represents a category. On the y-axis we have the absolute frequency.

Pie charts



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.

Pareto diagrams




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.

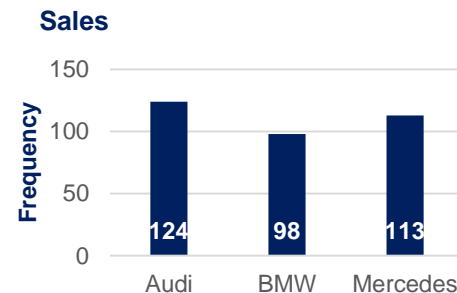
Graphs and tables that represent categorical variables. Excel formulas


Frequency distribution tables

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

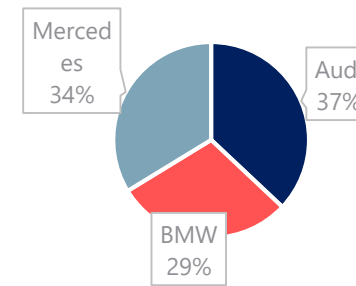
In Excel, we can either hard code the frequencies or count them with a count function. This will come up later on. Total formula: =SUM() 


Bar charts



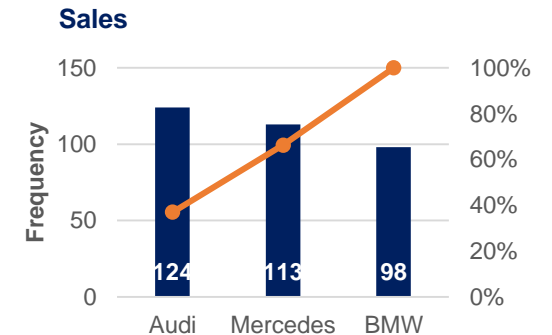
Bar charts are also called clustered column charts in Excel. Choose your data, Insert -> Charts -> Clustered column or Bar chart. 

Pie charts



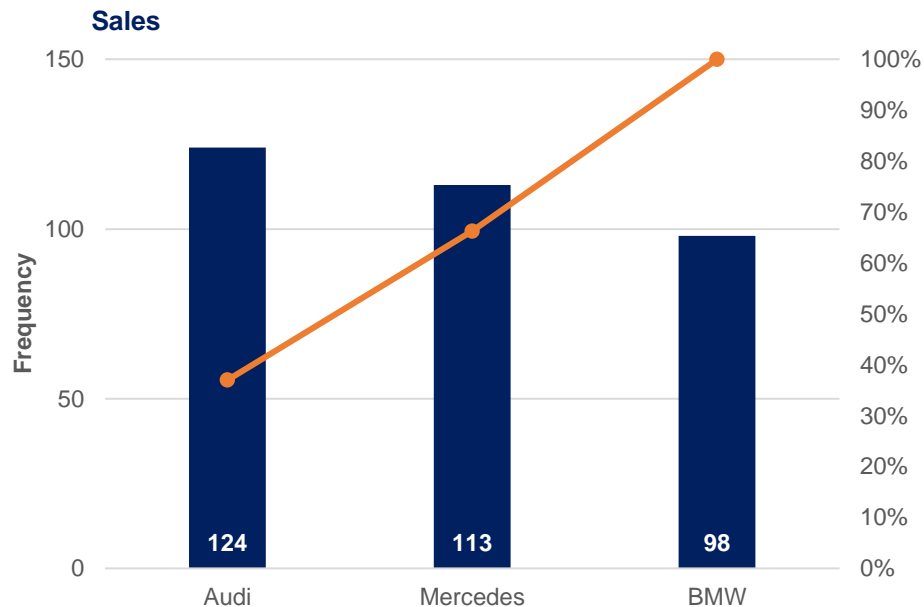
Pie charts are created in the following way: Choose your data, Insert -> Charts -> Pie chart 

Pareto diagrams



Next slide.

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.

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:

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

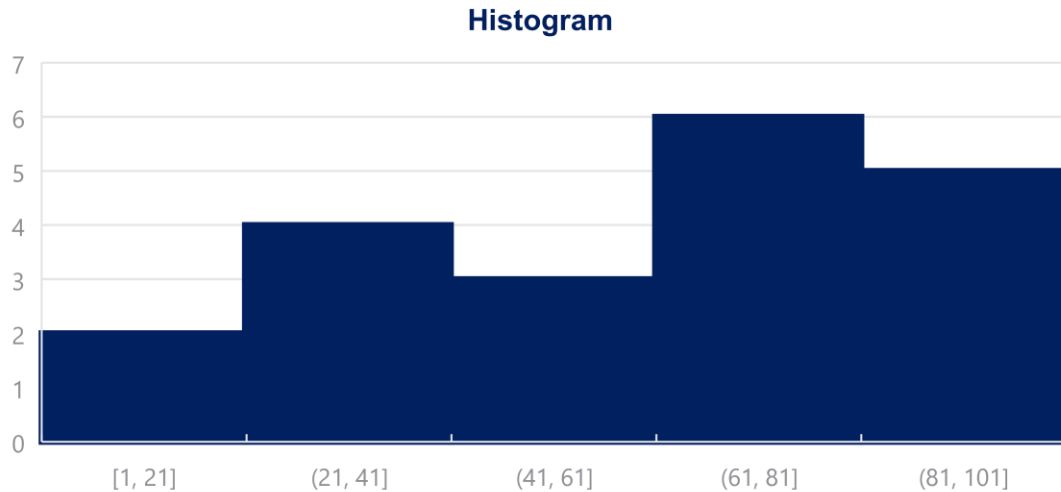


Creating the frequency distribution table in Excel:

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 consecutive 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.

Numerical variables. Frequency distribution table and histogram

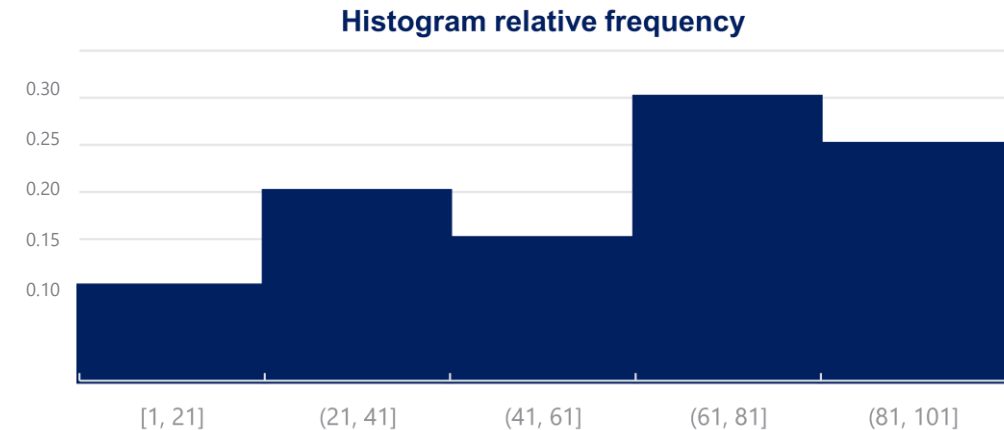


Histograms are the one of the most common ways to represent numerical data. Each bar has width equal to the width of the interval. The bars are touching as there is continuation between intervals: where one ends -> the other begins.



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.



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

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.

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

A common way to represent the data from a cross table is by using a side-by-side bar chart.

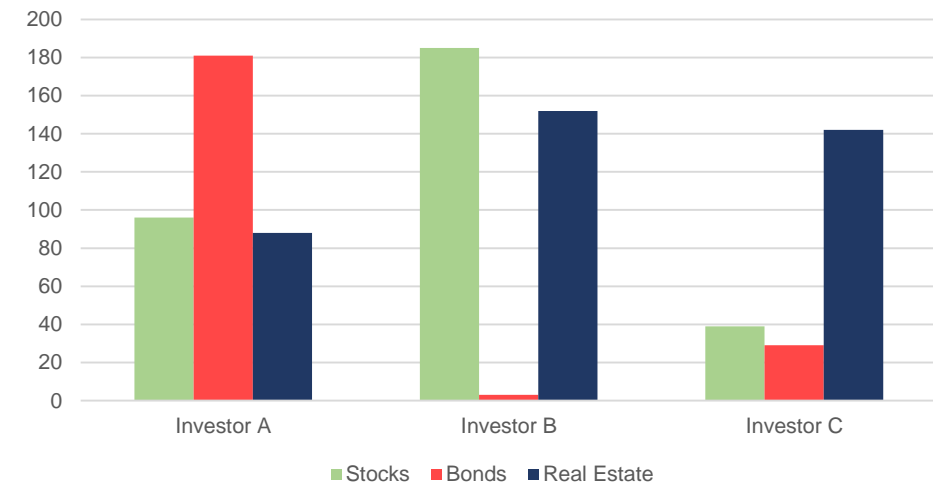


Creating a side-by-side chart in Excel:

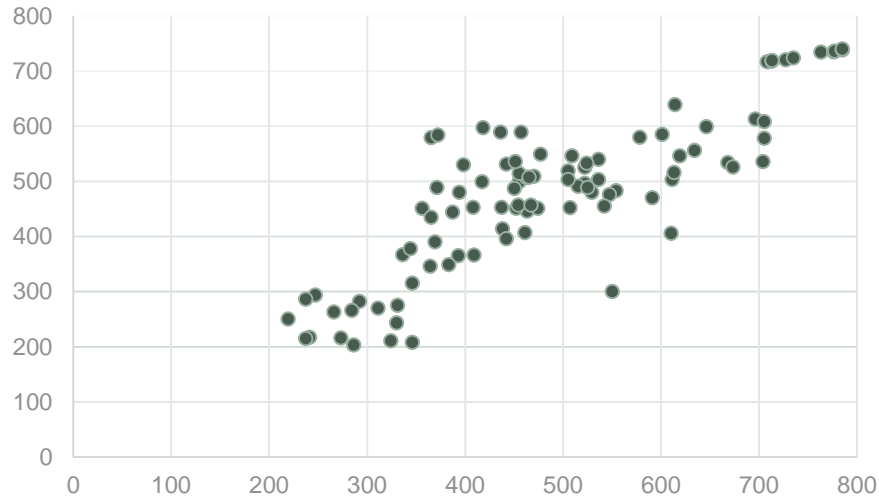
1. Choose your data
2. Insert -> Charts -> Clustered Column

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

Side-by-side bar chart



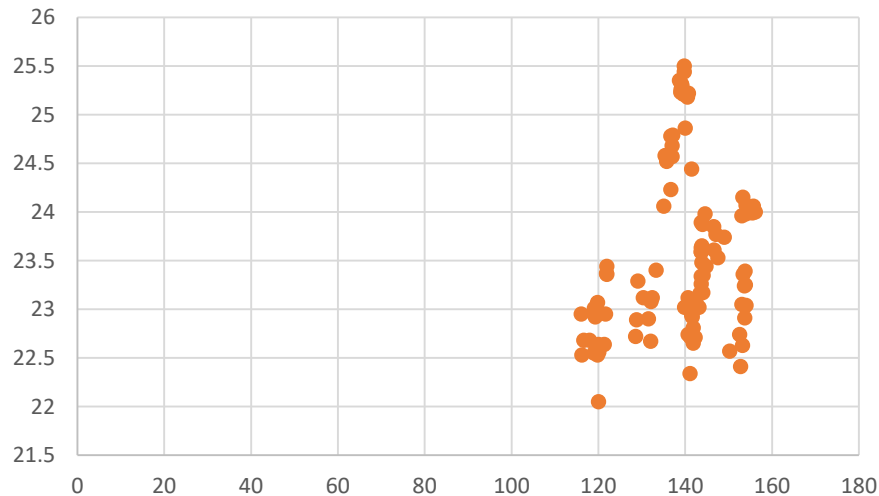
Graphs and tables for relationships between variables. Scatter plots



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:

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

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

 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.

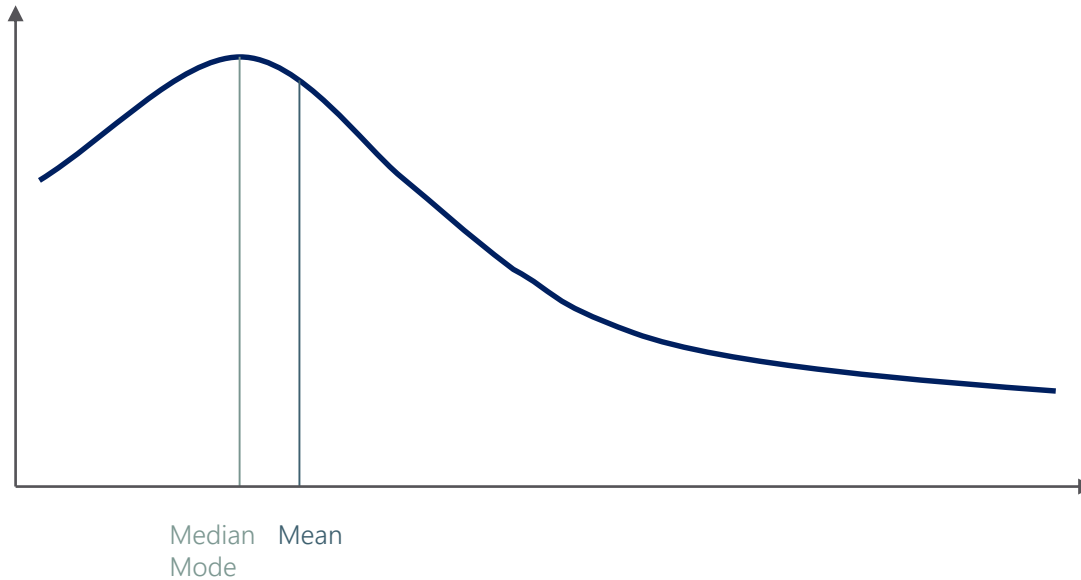
The mode is calculated simply by finding the value with the highest frequency.

 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



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.

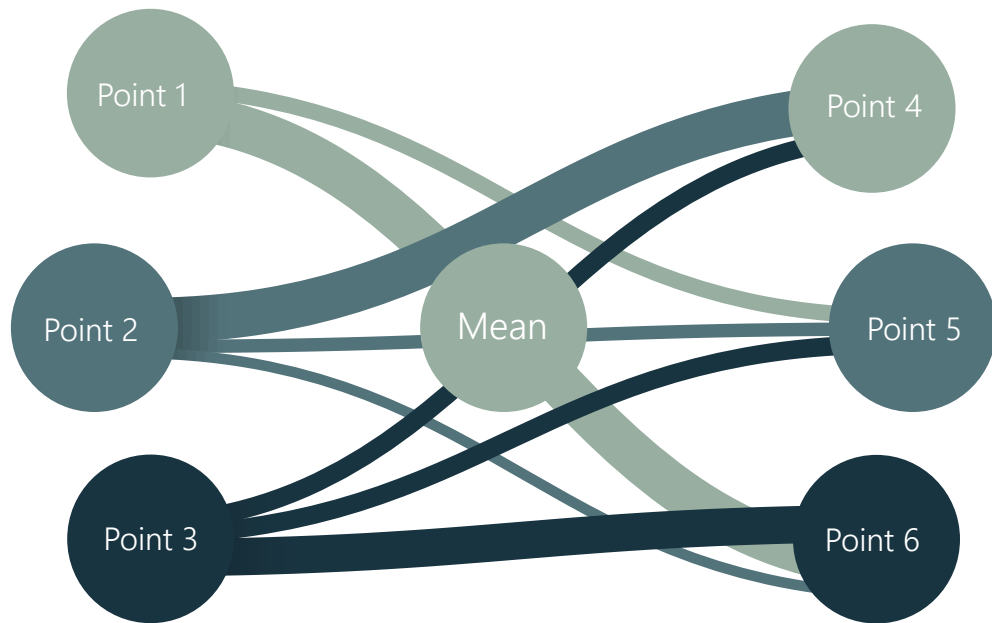
Usually, you will use software to calculate skewness.

 Calculating skewness in Excel:

=SKEW()

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

Variance and standard deviation



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:

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

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

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

 In Excel, correlation is calculated by:

`=CORREL()`

COURSE NOTES: INFERENCEAL STATISTICS

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.

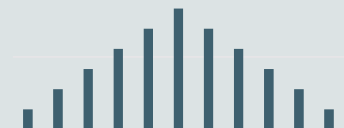
Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them.

Examples:

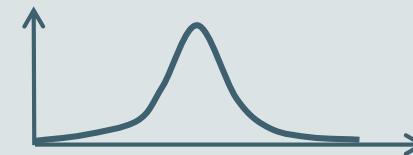
Uniform distribution



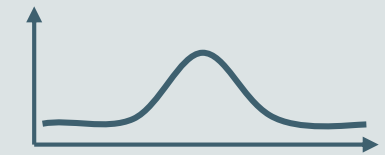
Binomial distribution



Normal distribution



Student's T distribution



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

Examples:

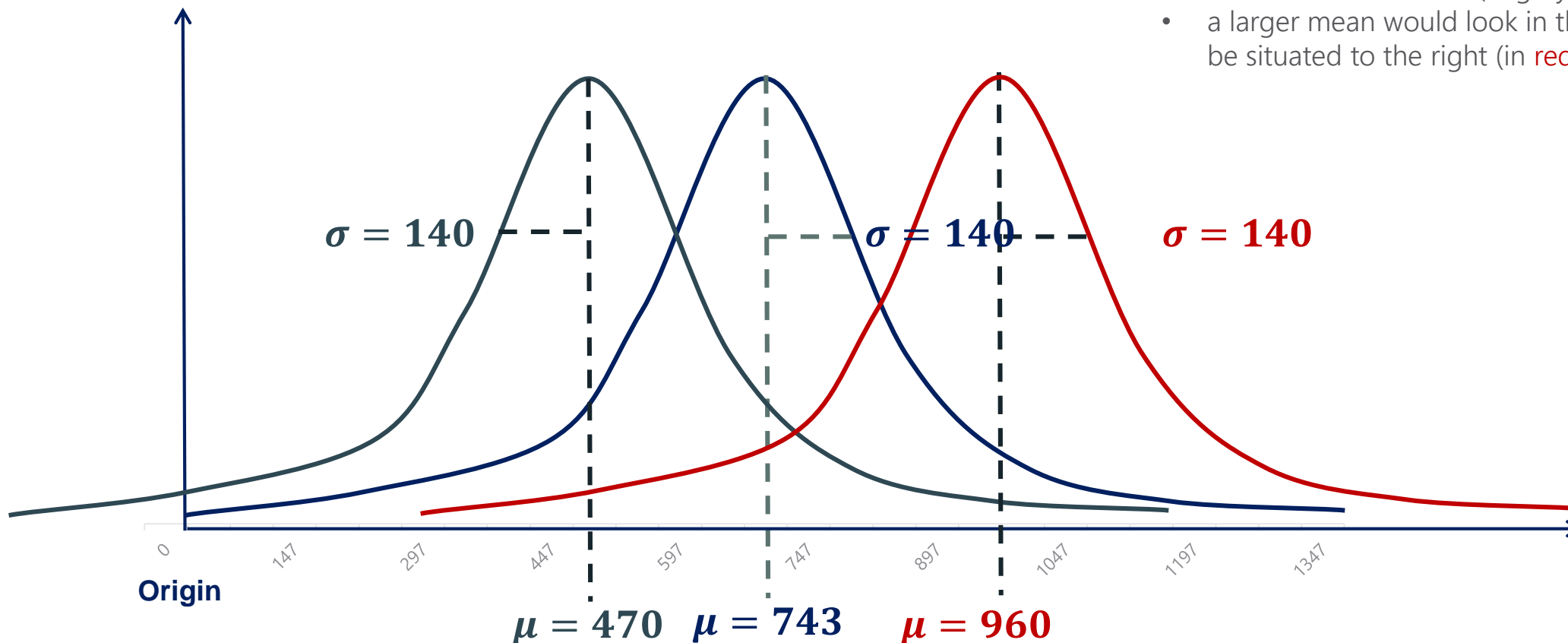
- Biology. Most biological measures are normally distributed, such as: height; length of arms, legs, nails; blood pressure; thickness of tree barks, etc.
- IQ tests
- Stock market information


$$N \sim (\mu, \sigma^2)$$

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

The Normal Distribution

Controlling for the standard deviation

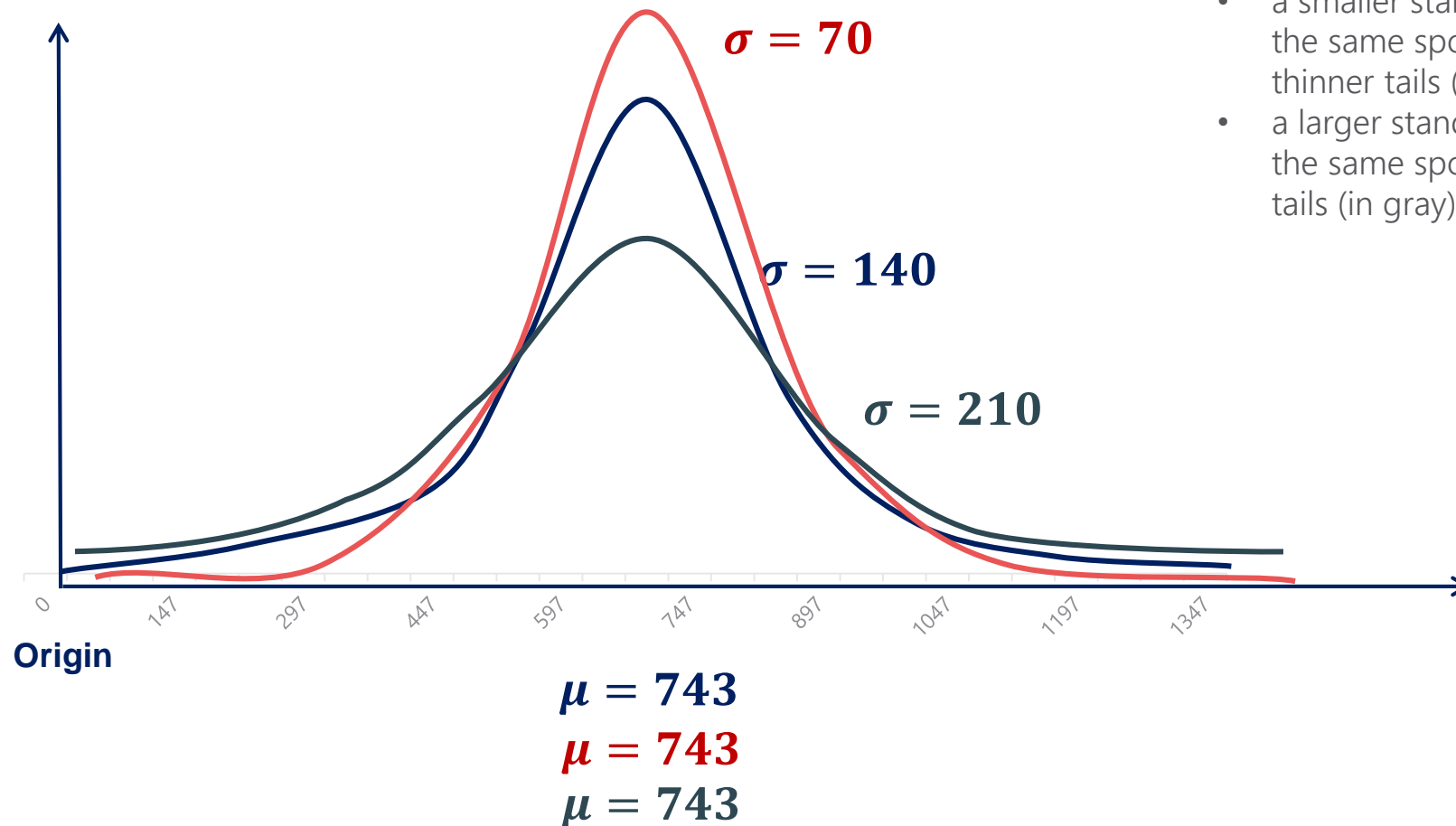


Keeping the standard deviation constant, the graph of a normal distribution with:

- 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)

The Normal Distribution

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)

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.

$$N \sim (0, 1)$$

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.

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

- No matter the distribution
- The distribution of $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \dots, \bar{x}_k$ would tend to $N\left(\mu, \frac{\sigma^2}{n}\right)$
- The more samples, the closer to Normal ($k \rightarrow \infty$)
- The bigger the samples, the closer to Normal ($n \rightarrow \infty$)

Why is it useful?

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.](#)

Where can we see it?

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.

Estimators and Estimates

Estimators

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

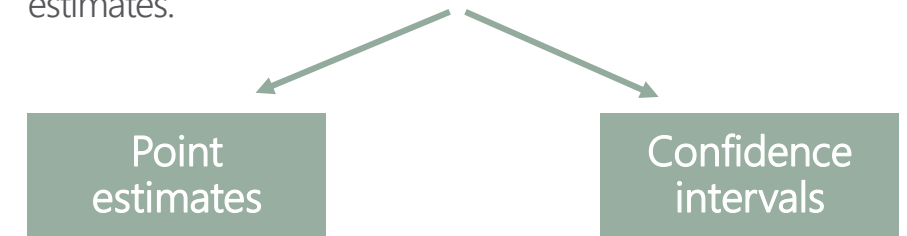
The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.

- Efficiency

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

Estimates

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

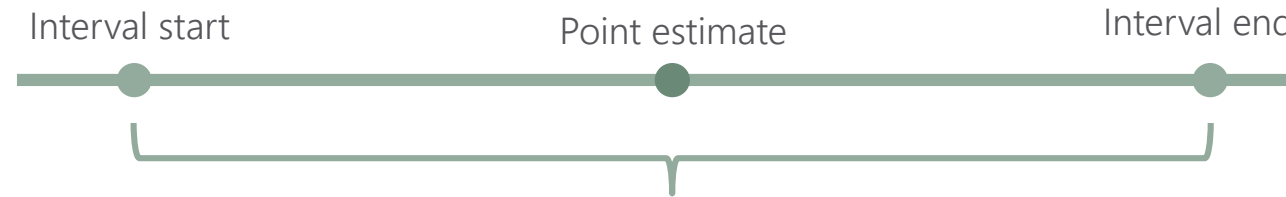
An interval.

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.

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.

$(1-\alpha)$ is the level of confidence. We are $(1-\alpha)*100\%$ confident that the population parameter will fall in the specified interval. Common alphas are: 0.01, 0.05, 0.1.

General formula:

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

$$\text{ME} = \text{reliability factor} * \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

$$z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

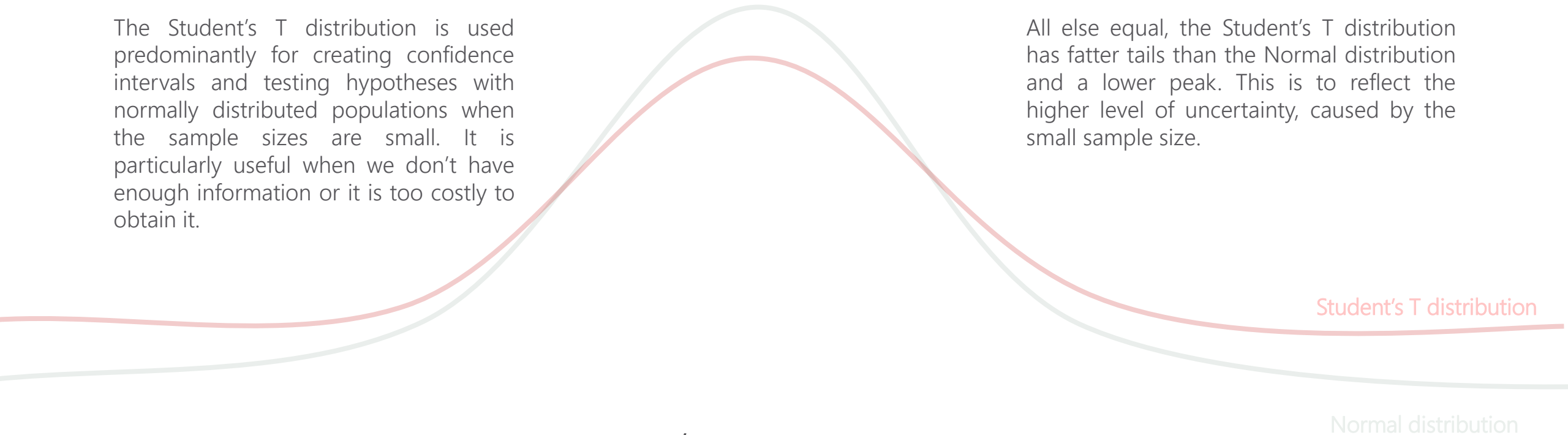
$$t_{v,\alpha/2} * \frac{s}{\sqrt{n}}$$

Term	Effect on width of CI
$(1-\alpha) \uparrow$	\uparrow
$\sigma \uparrow$	\uparrow
$n \uparrow$	\downarrow

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 $t_{\nu, \alpha}$, where ν are the degrees of freedom.

We can obtain the student's T distribution for a variable with a Normally distributed population using the formula: $t_{\nu, \alpha} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

Formulas for Confidence Intervals

# populations	Population variance	Samples	Statistic	Variance	Formula
One	known	-	z	σ^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	$s_{difference}^2$	$\bar{d} \pm t_{n-1, \alpha/2} \frac{s_d}{\sqrt{n}}$
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	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$(\bar{x} - \bar{y}) \pm t_{n_x + n_y - 2, \alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
Two	unknown, assumed different	independent	t	s_x^2, s_y^2	$(\bar{x} - \bar{y}) \pm t_{v, \alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

A photograph of a modern conference room with large windows and a long table, overlaid with a blue tint. The room features a long, dark wooden conference table surrounded by several black office chairs. Large windows on the left and right sides offer a view of a cityscape. The ceiling has a grid pattern with recessed lights. The entire image is covered with a semi-transparent blue overlay.

COALESCE() - Preamble

COALESCE() - Preamble

Here we will study something a bit more sophisticated.

IF NULL() and COALESCE() are among the advanced SQL functions in the toolkit of SQL professionals. They are used when null values are dispersed in your data table and you would like to substitute the null values with another value.

So, let's adjust the "Departments" duplicate in a way that suits the purposes of the next video, in which we will work with IF NULL() and COALESCE().

First, let's look at our table and see what we have there.

COALESCE() - Preamble



SQL

```
SELECT * FROM departments_dup;
```

Result Grid		Filter Rows:	Export:	Wrap Cell Content:
dept_no	dept_name			
d001	Marketing			
d002	Finance			
d003	Human Resources			
d004	Production			
d005	Development			
d006	Quality Management			
d007	Sales			
d008	Research			
d009	Customer Service			

Nine departments, with their department numbers and names provided. Ok!

COALESCE() - Preamble

Currently, as shown in the DDL statement of this table, the “Department name” field is with a NOT NULL constraint, which naturally means we must insert a value in each of its rows.

DDL for employees.departments_dup

```
1 CREATE TABLE `departments_dup` (  
2   `dept_no` char(4) NOT NULL,  
3   `dept_name` varchar(40) NOT NULL  
4 ) ENGINE=InnoDB DEFAULT CHARSET=utf8
```

COALESCE() - Preamble

Now, with the ALTER TABLE statement and the CHANGE COLUMN command, we will modify this constraint and allow null values to be registered in the “department name” column.



SQL

```
ALTER TABLE departments_dup
```

```
CHANGE COLUMN dept_name dept_name VARCHAR(40) NULL;
```

COALESCE() - Preamble

Right after that, we will insert into the department number column of this table a couple of data values – D-10 and D-11, the numbers of the next two potential departments in the “Departments Duplicate” table.



SQL

```
INSERT INTO departments_dup(dept_no) VALUES ('d010'), ('d011');
```

COALESCE() - Preamble

By running this SELECT query over here, you can see whether this operation was carried out successfully.



SQL

```
SELECT
    *
FROM
    departments_dup
ORDER BY dept_no ASC;
```


COALESCE() - Preamble

We have the two new department numbers listed below, and in the “Department name” column we can see two null values. The latter happened because we allowed for null values to exist in this field, “Department name”. Thus, Workbench will indicate that a value in a cell is missing by attaching a “null” label to it. Great!

Result Grid		Filter Rows:	Export:	Wrap Cell Content:
	dept_no	dept_name		
	d001	Marketing		
	d002	Finance		
	d003	Human Resources		
	d004	Production		
	d005	Development		
	d006	Quality Management		
	d007	Sales		
	d008	Research		
	d009	Customer Service		
	d010	NULL		
	d011	NULL		

COALESCE() - Preamble

The next adjustment we'll have to make is adding a third column called "Department manager". It will indicate the manager of the respective department. For now, we will leave it empty, and will add the NULL constraint. Finally, we will place it next to the "Department name" column by typing "AFTER "Department name".



SQL

```
ALTER TABLE employees.departments_dup  
ADD COLUMN dept_manager VARCHAR(255) NULL AFTER dept_name;
```

COALESCE() - Preamble

Let's check the state of the "Departments duplicate" table now.



SQL

```
SELECT
    *
FROM
    departments_dup
ORDER BY dept_no ASC;
```

COALESCE() - Preamble

Exactly as we wanted, right? The third column is completely empty and we have null values in the last two records. These are the “department name” and “manager” fields.

Result Grid	Filter Rows:	Export:	Wrap Cell Content:
dept_no	dept_name	dept_manager	
d001	Marketing	NULL	
d002	Finance	NULL	
d003	Human Resources	NULL	
d004	Production	NULL	
d005	Development	NULL	
d006	Quality Management	NULL	
d007	Sales	NULL	
d008	Research	NULL	
d009	Customer Service	NULL	
d010	NULL	NULL	
d011	NULL	NULL	

COALESCE() - Preamble

To save the “Departments duplicate” table in its current state, execute a COMMIT statement.



SQL

```
COMMIT;
```

Here we'll end the setup for the video about IF NULL() and COALESCE().

Good Luck!

COURSE NOTES: HYPOTHESIS TESTING

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).



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.

Examples of 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.

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

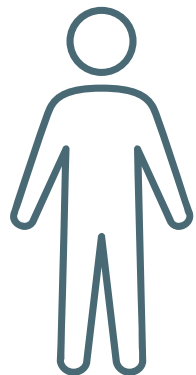
Instructor's answer (with some adjustments)

'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 or equal to** \$125,000.

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



Student's question

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.

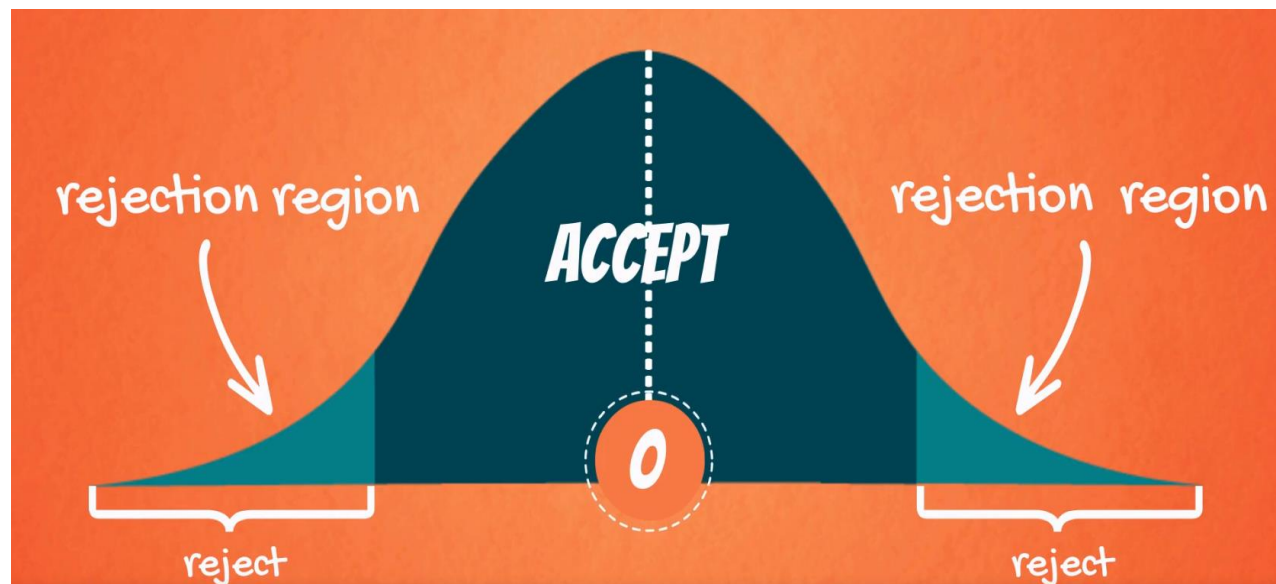
Please explain further.

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

At x% significance, we accept the null hypothesis

At x% significance, A is not significantly different from B

At x% significance, there is not enough statistical evidence that...

At x% significance, we cannot reject the null hypothesis

Reject

At x% significance, we reject the null hypothesis

At x% significance, A is significantly different from B

At x% significance, there is enough statistical evidence...

At x% significance, we cannot say that *restate the null*

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

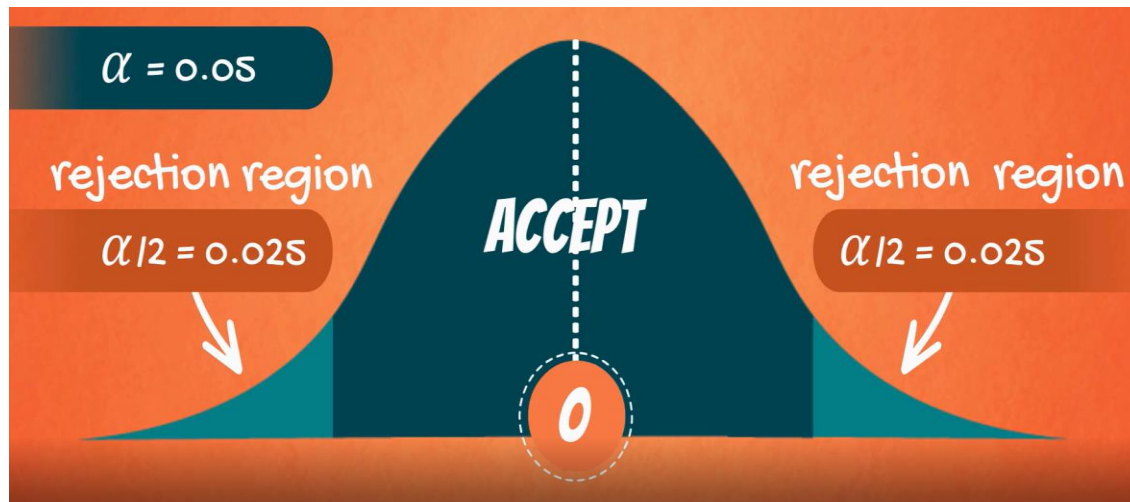
0.10

0.05

0.01

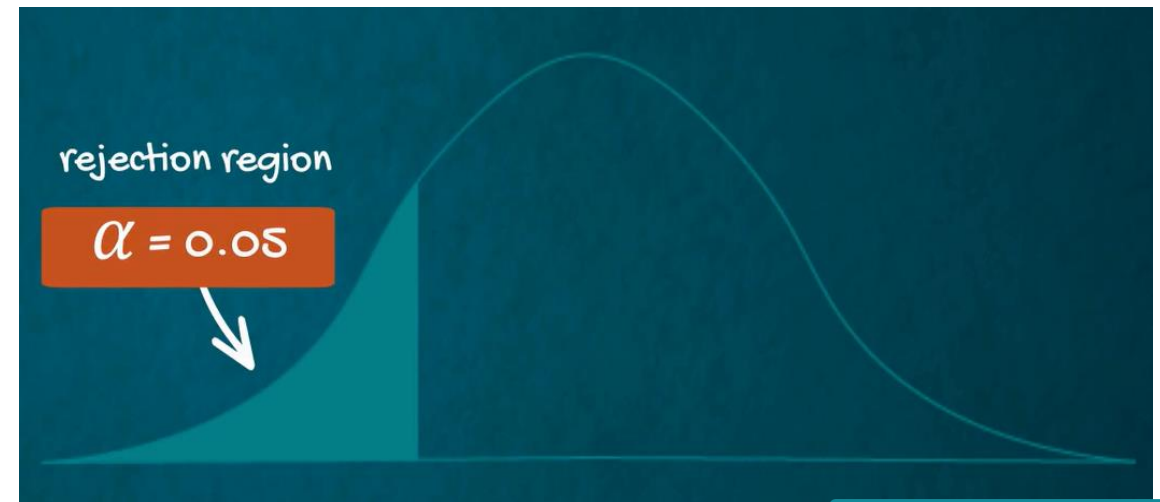
Two-sided (two-tailed) test

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



One-sided (one-tailed) test



Used when the null doesn't contain equality or inequality sign ($<$, $>$, \leq , \geq)





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:

H_0 : Status quo		The truth	
		H_0 is true	H_0 is false
H_0 (status quo)	Accept		Type II error (False negative)
	Reject	Type I error (False positive)	

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

H_0 : She doesn't like you		The truth	
		She doesn't like you	She likes you
H_0 (status quo) She doesn't like you (you shouldn't invite her out)	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) and is called 'power of the test'.

[If you want to find out more about statistical errors, just follow this link for an article written by your instructor.](#)

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

Notable p-values

0.000

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

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](#).

Formulae for Hypothesis Testing

# populations	Population variance	Samples	Statistic	Variance	Formula for test statistic	Decision rule
One	known	-	z	σ^2	$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	<p>There are several ways to phrase the decision rule and they all have the same meaning.</p> <p>Reject the null if:</p> <ol style="list-style-type: none"> 1) $\text{test statistic} > \text{critical value}$ 2) The absolute value of the test statistic is bigger than the absolute critical value 3) $\text{p-value} < \text{some significance level}$ <i>most often 0.05</i> <p>Usually, you will be using the p-value to make a decision.</p>
One	unknown	-	t	s^2	$T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$	
Two	-	dependent	t	$s_{\text{difference}}^2$	$T = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}$	
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	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$T = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$	

Statistics for Data Science and Business Analysis



**Online p-value
calculators**

What is a p-value calculator?

A p-value calculator is a software that has an input your test statistic, (and the degrees of freedom, if applicable) that returns the p-value for the test.

There are various online solutions, however, the one I would normally use is by socstatistics @



How to use the socstatistics online p-value calculator?

Step 1: Go to

<http://www.socscistatistics.com/pvalues/>

There are several choices available depending on the test you need.

Step 2: Choose the test applicable to your problem and click on the link.

In this course we cover Z-score, t-score and the F-ratio score.

The screenshot shows the 'Social Science Statistics' website. At the top, there is a navigation bar with links: Home, Statistical Calculators, Test Yourself Quizzes, Which Statistics Test?, Descriptive Statistics, P Value Calculators, Donate, About, and Contact. Below the navigation bar, there are three buttons: 'AdChoices', 'P Value', and 'SPSS Statistics'. The 'P Value' button is highlighted. Underneath, the section is titled 'Quick P-Value Calculators' with a subtitle: 'This is a set of very simple calculators that generate p-values from various test scores (i.e., t test, chi-square, etc)'. A list of five options is provided: 'P-value from Z score.', 'P-value from t score.', 'P-value from chi-square score.', 'P-value from F-ratio score.', and 'P-value from Pearson (r) score.'. To the right of the list, three arrows point from text labels to the corresponding options: 'Z-score, Normal distribution' points to 'P-value from Z score.', 't-score, Student's T distribution' points to 'P-value from t score.', and 'F-ratio score, F distribution' points to 'P-value from F-ratio score.'. The website also features a formula for the t-test at the top:
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Quick P-Value Calculators

This is a set of very simple calculators that generate p-values from various test scores (i.e., t test, chi-square, etc).

- P-value from Z score.
- P-value from t score.
- P-value from chi-square score.
- P-value from F-ratio score.
- P-value from Pearson (r) score.

Z-score, Normal distribution

t-score, Student's T distribution

F-ratio score, F distribution

365 DataScience

P-value from Z-score

Step 1: Type in the Z-score you got from your test.

Step 2 (optional): Choose the significance level, if you want to get the result for your test.

Step 3: Choose if this is a one-tailed or two-tailed test.

Step 4: Click calculate.

The screenshot shows the 'P Value from Z Score Calculator' interface. It includes a navigation bar with links like 'Home', 'Statistical Calculators', and 'P Value Calculators'. Below the navigation bar are buttons for 'P Value Calculator', 'Z Score', and 'T Test Calculator'. The main section is titled 'P Value from Z Score Calculator' and contains instructions: 'This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!'. It also provides a link to a Z test calculator. The form has four input fields with arrows pointing to them from the right, labeled 'Step 1' through 'Step 4':
1. 'Z score:' with an empty text box.
2. 'Significance Level:' with radio buttons for 0.01, 0.05 (selected), and 0.10.
3. 'One-tailed or two-tailed hypothesis?:' with radio buttons for One-tailed (selected) and Two-tailed.
4. A 'Calculate' button.
Below the form, it says 'Enter your z score value, and then press the button.'

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

Home Statistical Calculators Test Yourself Quizzes Which Statistics Test? Descriptive Statistics P Value Calculators Donate About Contact

AdChoices P Value Calculator Z Score T Test Calculator

P Value from Z Score Calculator

This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a Z score from raw data, [you can find a Z test calculator here.](#)

Z score: ← Step 1

Significance Level:

☐ 0.01
☒ 0.05 ← Step 2
☐ 0.10

One-tailed or two-tailed hypothesis?:

☒ One-tailed ← Step 3
☐ Two-tailed

Enter your z score value, and then press the button.

Calculate ← Step 4

P-value from Z-score

Step 1: Type in the Z-score you got from your test.

Step 2 (optional): Choose the significance level, if you want to get the decision for your test.

Step 3: Choose if this is a one-tailed or two-tailed test.

Step 4: Click calculate.

The screenshot shows the 'P Value from Z Score Calculator' interface. It includes a navigation bar with links like 'Home', 'Statistical Calculators', and 'P Value Calculators'. Below the navigation bar are buttons for 'P Value Calculator', 'Z Score', and 'T Test Calculator'. The main section is titled 'P Value from Z Score Calculator' and contains instructions: 'This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!'. It also provides a link to a Z test calculator. The form has four input fields with arrows pointing to them from the right, labeled 'Step 1' through 'Step 4':
1. 'Z score:' with an empty text box.
2. 'Significance Level:' with radio buttons for 0.01, 0.05 (selected), and 0.10.
3. 'One-tailed or two-tailed hypothesis?:' with radio buttons for One-tailed (selected) and Two-tailed.
4. A 'Calculate' button.
Below the form, it says 'Enter your z score value, and then press the button.'

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

Home Statistical Calculators Test Yourself Quizzes Which Statistics Test? Descriptive Statistics P Value Calculators Donate About Contact

AdChoices P Value Calculator Z Score T Test Calculator

P Value from Z Score Calculator

This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a Z score from raw data, [you can find a Z test calculator here.](#)

Z score: ← Step 1

Significance Level:

☐ 0.01
☒ 0.05 ← Step 2
☐ 0.10

One-tailed or two-tailed hypothesis?:

☒ One-tailed ← Step 3
☐ Two-tailed

Enter your z score value, and then press the button.

Calculate ← Step 4

P-value from Z-score. Example result (Part 1)

After clicking 'Calculate', you would instantly get two results.

Result 1: The p-value of the test.

Result 2: The decision, based on the information you entered above.

Note: When using this online p-value calculator, a **red** color of the text means that the result is **not significant**, given the significance level you have chosen.

Seth's Blog
sethgodin.typepad.com

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

Home | Statistical Calculators | Test Yourself Quizzes | Which Statistics Test? | Descriptive Statistics | P Value Calculators | Donate | About | Contact

AdChoices | P Value | Z Score | T Test

P Value from Z Score Calculator

This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a Z score from raw data, [you can find a Z test calculator here.](#)

Z score:

Significance Level:

☐ 0.01
☒ 0.05
☐ 0.10

One-tailed or two-tailed hypothesis?:

☒ One-tailed
☐ Two-tailed

The P-Value is 0.109349.

The result is not significant at p < 0.05.

Calculate

Result 1

Result 2

P-value from Z-score. Example result (Part 2)

After clicking 'Calculate', you would instantly get two results.

Result 1: The p-value of the test.

Result 2: The decision, based on the information you entered above.

Note: When using this online p-value calculator, a **blue** color of the text means that the result is **significant**, given the significance level you have chosen.

The screenshot shows the 'Social Science Statistics' website's 'P Value from Z Score Calculator'. The page has a navigation bar with links: Home, Statistical Calculators, Test Yourself Quizzes, Which Statistics Test?, Descriptive Statistics, P Value Calculators, Donate, About, and Contact. Below the navigation bar are three buttons: 'AdChoices', 'P Value' (selected), 'Z Score', and 'T Test'. The main heading is 'P Value from Z Score Calculator'. The instructions state: 'This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!'. A link is provided: 'If you need to derive a Z score from raw data, you can find a Z test calculator here.' The 'Z score' input field contains '3.54'. The 'Significance Level' section has three radio buttons: '0.01', '0.05' (selected), and '0.10'. The 'One-tailed or two-tailed hypothesis?:' section has two radio buttons: 'One-tailed' (selected) and 'Two-tailed'. The results are displayed in blue text: 'The P-Value is 0.0002.' and 'The result is significant at p < 0.05.'. A 'Calculate' button is at the bottom. Two arrows point from the text 'Result 1' and 'Result 2' to the two lines of blue result text.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

Home | Statistical Calculators | Test Yourself Quizzes | Which Statistics Test? | Descriptive Statistics | P Value Calculators | Donate | About | Contact

AdChoices | P Value | Z Score | T Test

P Value from Z Score Calculator

This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a Z score from raw data, you can find a Z test calculator here.

Z score:

Significance Level:

☐ 0.01
☒ 0.05
☐ 0.10

One-tailed or two-tailed hypothesis?:

☒ One-tailed
☐ Two-tailed

The P-Value is 0.0002.

The result is significant at p < 0.05.

Calculate

Result 1

Result 2

P-value from t-score

Step 1: Type in the t-score you got from your test.

Step 2: Type in the degrees of freedom associated with your test.

Step 3 (optional): Choose the significance level, if you want to get the decision for your test.

Step 4: Choose if this is a one-tailed or two-tailed test.

Step 5: Click calculate.

The screenshot shows the 'P Value from T Score Calculator' interface. It includes a navigation bar with links like 'Home', 'Statistical Calculators', and 'P Value Calculators'. The main content area has tabs for 'P Value', 'T Test', and 'SPSS Statistics'. Below the tabs, there is a title 'P Value from T Score Calculator' and a detailed instruction paragraph. The form contains input fields for 'T Score' and 'DF', a 'Significance Level' section with radio buttons for .01, .05, and .10, and a 'One-tailed or two-tailed hypothesis?' section with radio buttons for 'One-tailed' and 'Two-tailed'. A 'Calculate' button is at the bottom. Five numbered arrows point to specific elements: Step 1 points to the 'T Score' input field, Step 2 points to the 'DF' input field, Step 3 points to the '.05' significance level radio button, Step 4 points to the 'One-tailed' hypothesis radio button, and Step 5 points to the 'Calculate' button.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

Home | Statistical Calculators | Test Yourself Quizzes | Which Statistics Test? | Descriptive Statistics | P Value Calculators | Donate | About | Contact

AdChoices | P Value | T Test | SPSS Statistics

P Value from T Score Calculator

This should be self-explanatory, but just in case it's not: your T Score goes in the T Score box, you stick your degrees of freedom in the DF box ($N - 1$ for single sample and dependent pairs, $(N_1 - 1) + (N_2 - 1)$ for independent samples), select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a T Score from raw data, then you can find t test calculators here.

T Score:

DF:

Significance Level:

☐ .01

☒ .05

☐ .10

One-tailed or two-tailed hypothesis?:

☒ One-tailed

☐ Two-tailed

Enter your values for T Score and degrees of freedom, and then press the button.

Calculate

Step 1

Step 2

Step 3

Step 4

Step 5

P-value from t-score. Example result (Part 1)

After clicking 'Calculate', you would instantly get two results.

Result 1: The p-value of the test.

Result 2: The decision, based on the information you entered above.

Note: When using this online p-value calculator, a **red** color of the text means that the result is **not significant**, given the significance level you have chosen.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

Home | Statistical Calculators | Test Yourself Quizzes | Which Statistics Test? | Descriptive Statistics | P Value Calculators | Donate | About | Contact

AdChoices | P Value | T Test | SPSS Statistics

P Value from T Score Calculator

This should be self-explanatory, but just in case it's not: your T Score goes in the T Score box, you stick your degrees of freedom in the DF box ($N - 1$ for single sample and dependent pairs, $(N_1 - 1) + (N_2 - 1)$ for independent samples), select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a T Score from raw data, [then you can find t test calculators here.](#)

T Score:

DF:

Significance Level:

☐ .01

☒ .05

☐ .10

One-tailed or two-tailed hypothesis?:

☒ One-tailed

☐ Two-tailed

The P-Value is .093417.

The result is not significant at $p < .05$.

Calculate

Result 1

Result 2

365 DataScience

P-value from t-score. Example result (Part 2)

After clicking 'Calculate', you would instantly get two results.

Result 1: The p-value of the test.

Result 2: The decision, based on the information you entered above.

Note: When using this online p-value calculator, a **blue** color of the text means that the result is **significant**, given the significance level you have chosen.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Social Science Statistics

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Home | Statistical Calculators | Test Yourself Quizzes | Which Statistics Test? | Descriptive Statistics | P Value Calculators | Donate | About | Contact

AdChoices | P Value | T Test | SPSS Statistics

P Value from T Score Calculator

This should be self-explanatory, but just in case it's not: your T Score goes in the T Score box, you stick your degrees of freedom in the DF box ($N - 1$ for single sample and dependent pairs, $(N_1 - 1) + (N_2 - 1)$ for independent samples), select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a T Score from raw data, [then you can find t test calculators here.](#)

T Score:

DF:

Significance Level:

☐ .01

☒ .05

☐ .10

One-tailed or two-tailed hypothesis?:

☒ One-tailed

☐ Two-tailed

The P-Value is .001482.

The result is significant at $p < .05$.

Note: If you wish to calculate the effect size, [this calculator](#) will do the job.

Calculate

Result 1

Result 2