COURSE NOTES: HYPOTHESIS TESTING

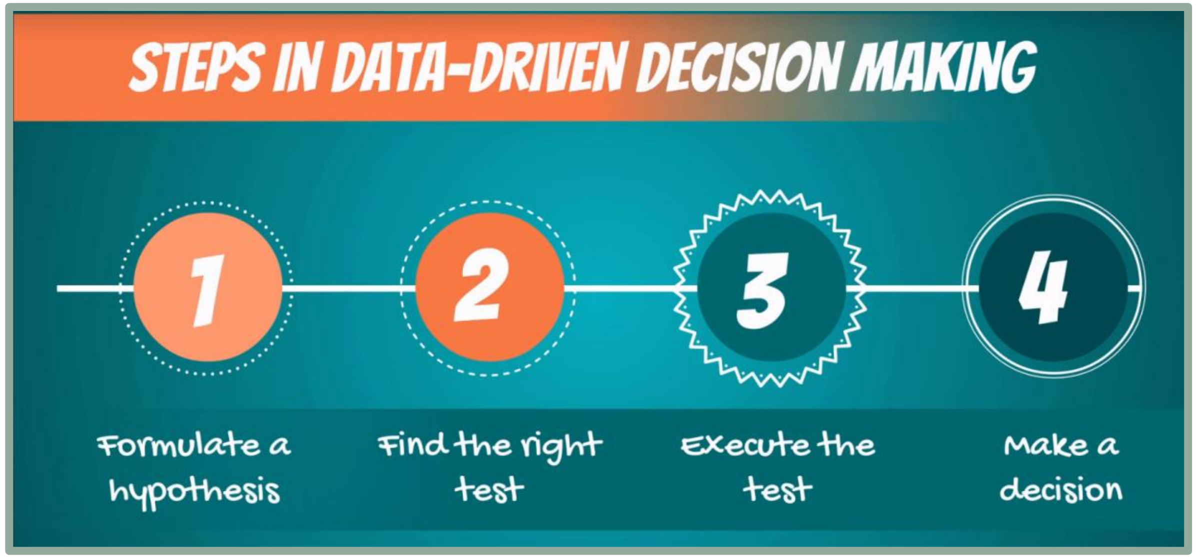
365√DataScience

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.

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.**

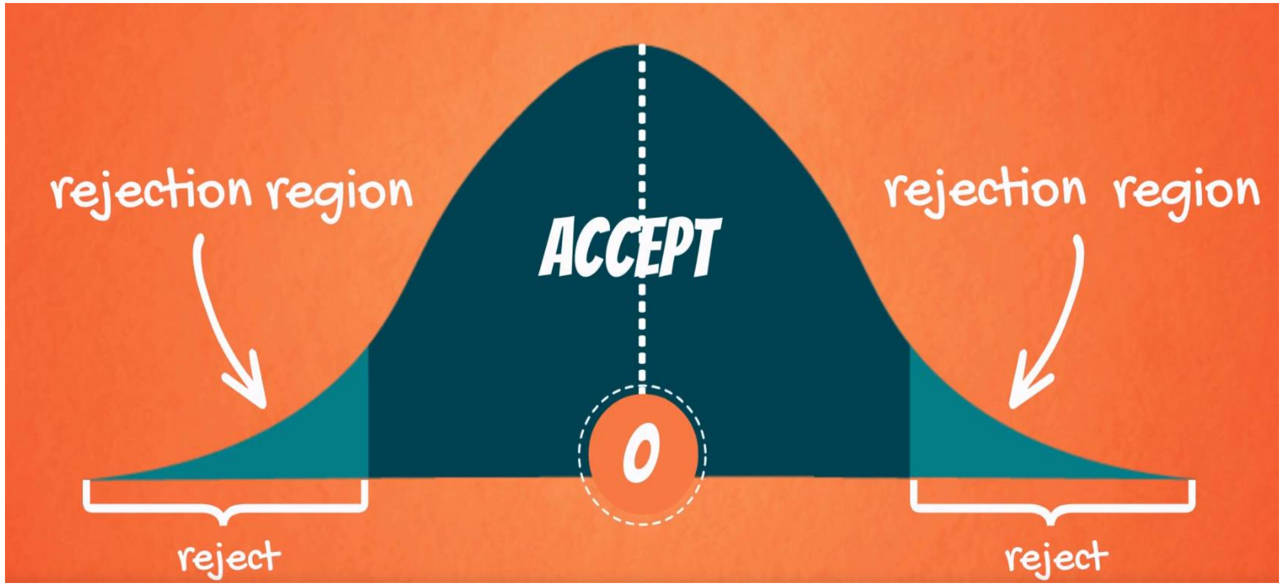
Example 2: Our friend Paul told us that the mean salary is >$125,000 (status-quo, null). Our opinion is that he may be wrong, so we are testing that. Therefore, the alternative is: the mean data scientist salary **is lower or equal to** $125,000.

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

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

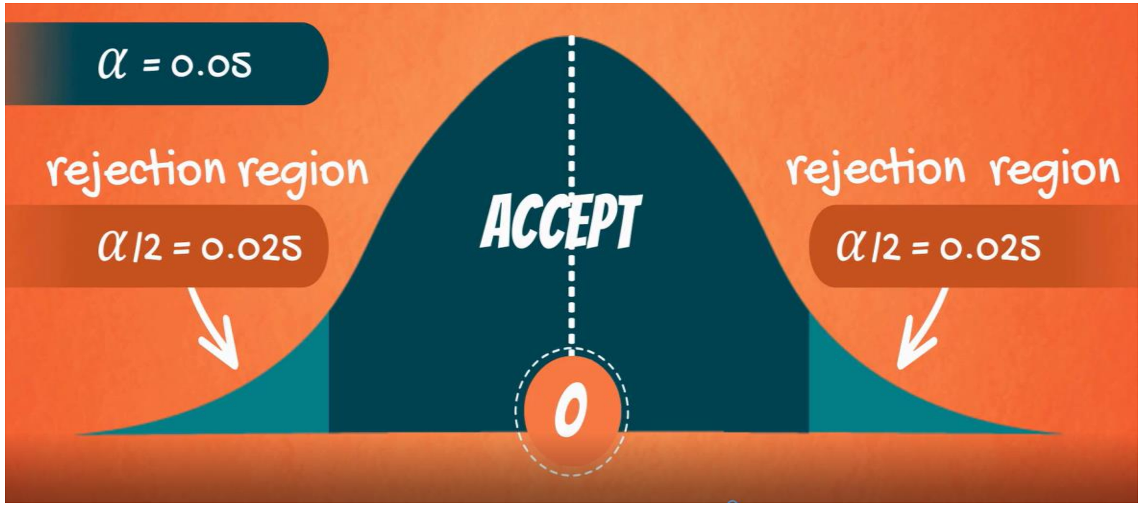
0.10

0.05

0.01

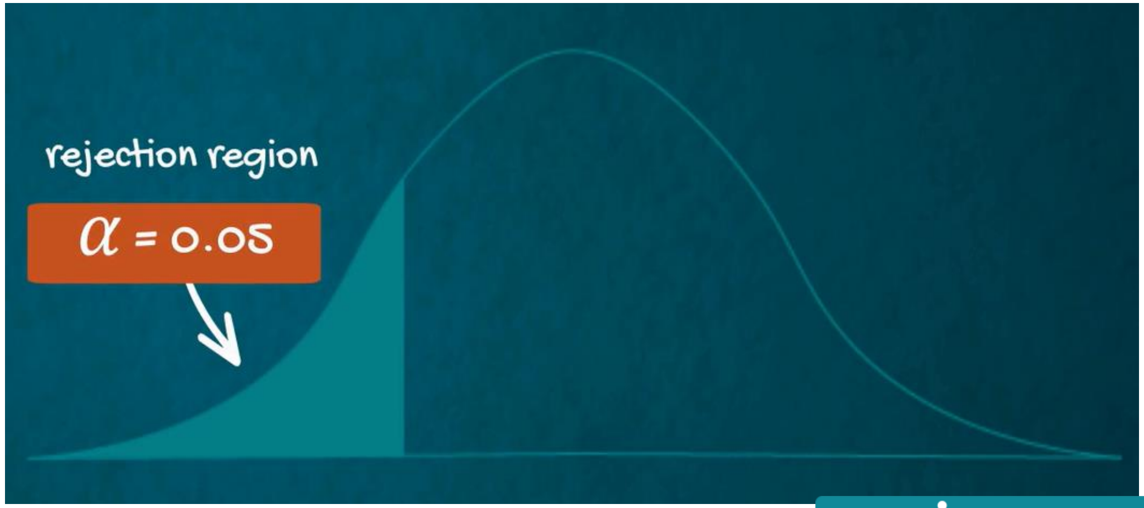
Two-sided (two-tailed) test

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



One-sided (one-tailed) test

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



365√DataScience

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	$Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{\mathrm{n}}}$	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.
One	unknown	-	t	s^2	$T = \dfrac{\bar{x} - \mu_0}{s/\sqrt{\mathrm{n}}}$	
Two	-	dependent	t	$s^2_{difference}$	$T = \dfrac{\bar{d} - \mu_0}{s_d/\sqrt{\mathrm{n}}}$	
Two	Known	independent	z	$\sigma_x^2\,,\ \sigma_y^2$	$Z = \dfrac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}}$	
Two	unknown, assumed equal	independent	t	$s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$T = \dfrac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\dfrac{s_p^2}{n_x} + \dfrac{s_p^2}{n_y}}}$	