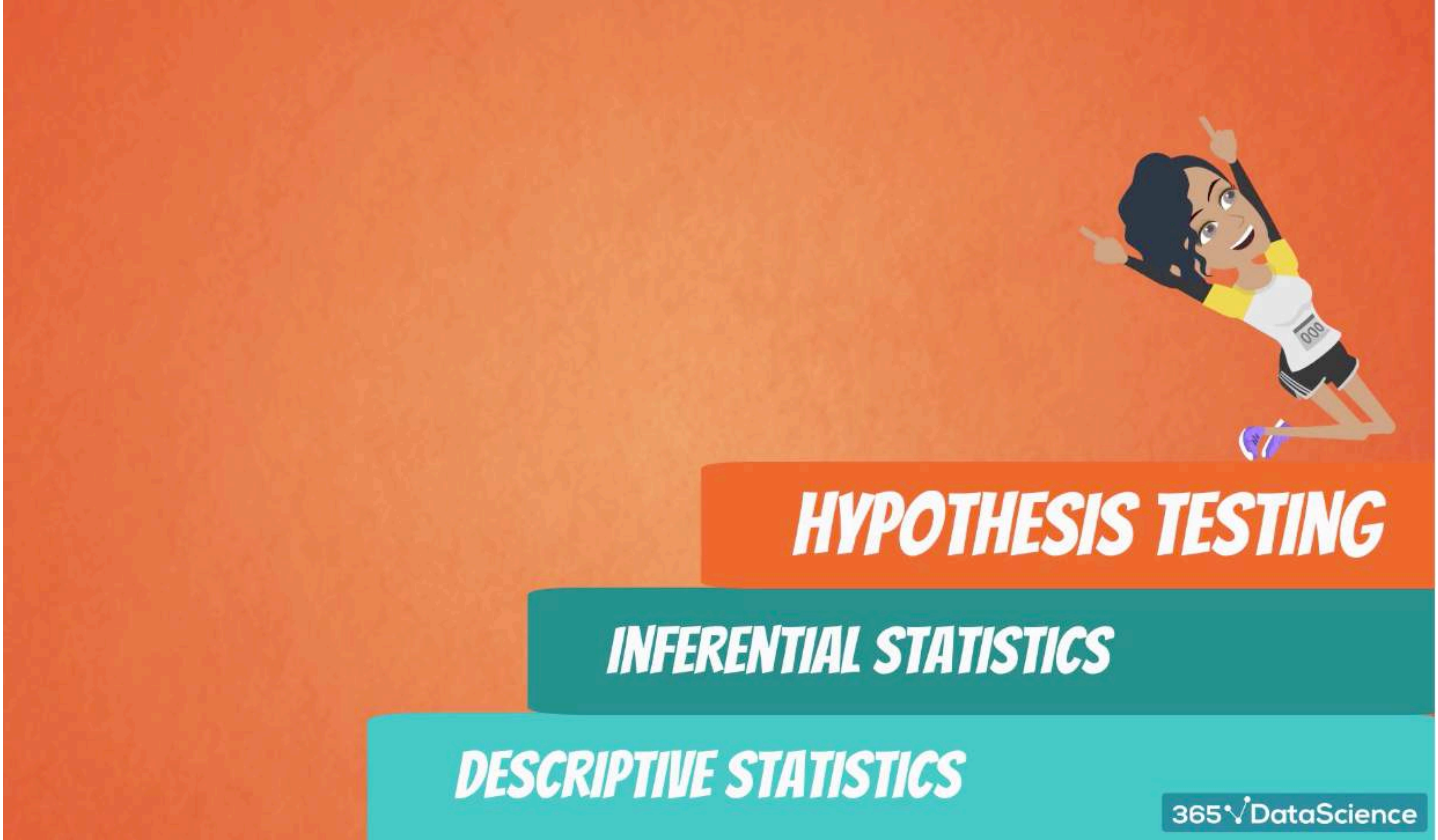
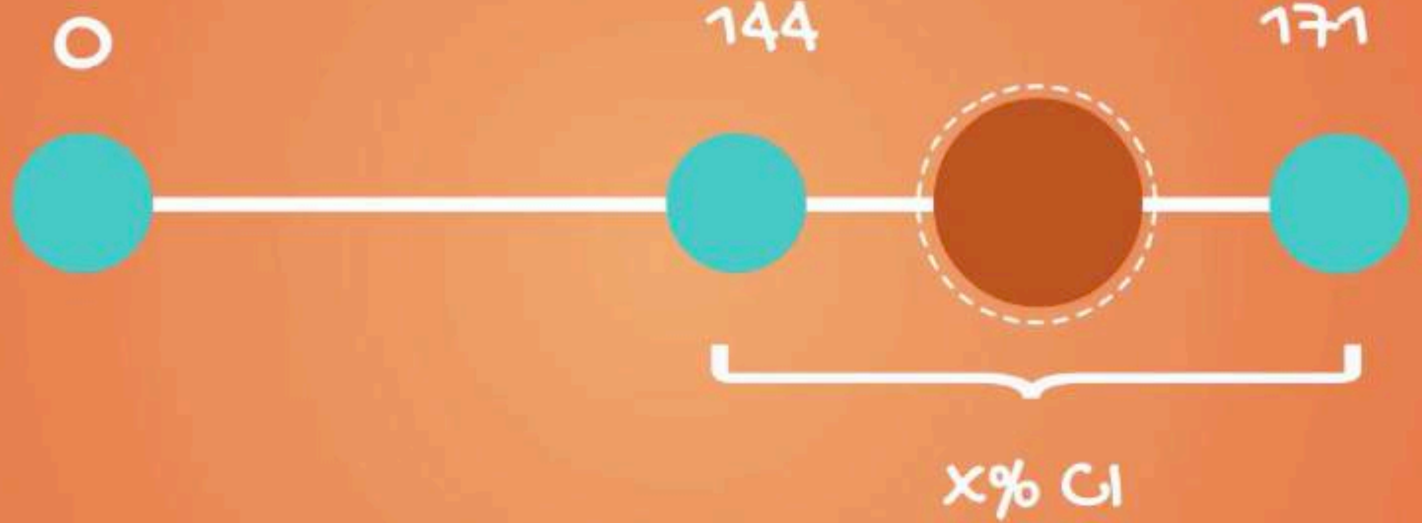
COURSE NOTES: HYPOTHESIS TESTING

365√DataScience



HYPOTHESIS TESTING
INFERENTIAL STATISTICS
DESCRIPTIVE STATISTICS
365√DataScience

CONFIDENCE INTERVAL



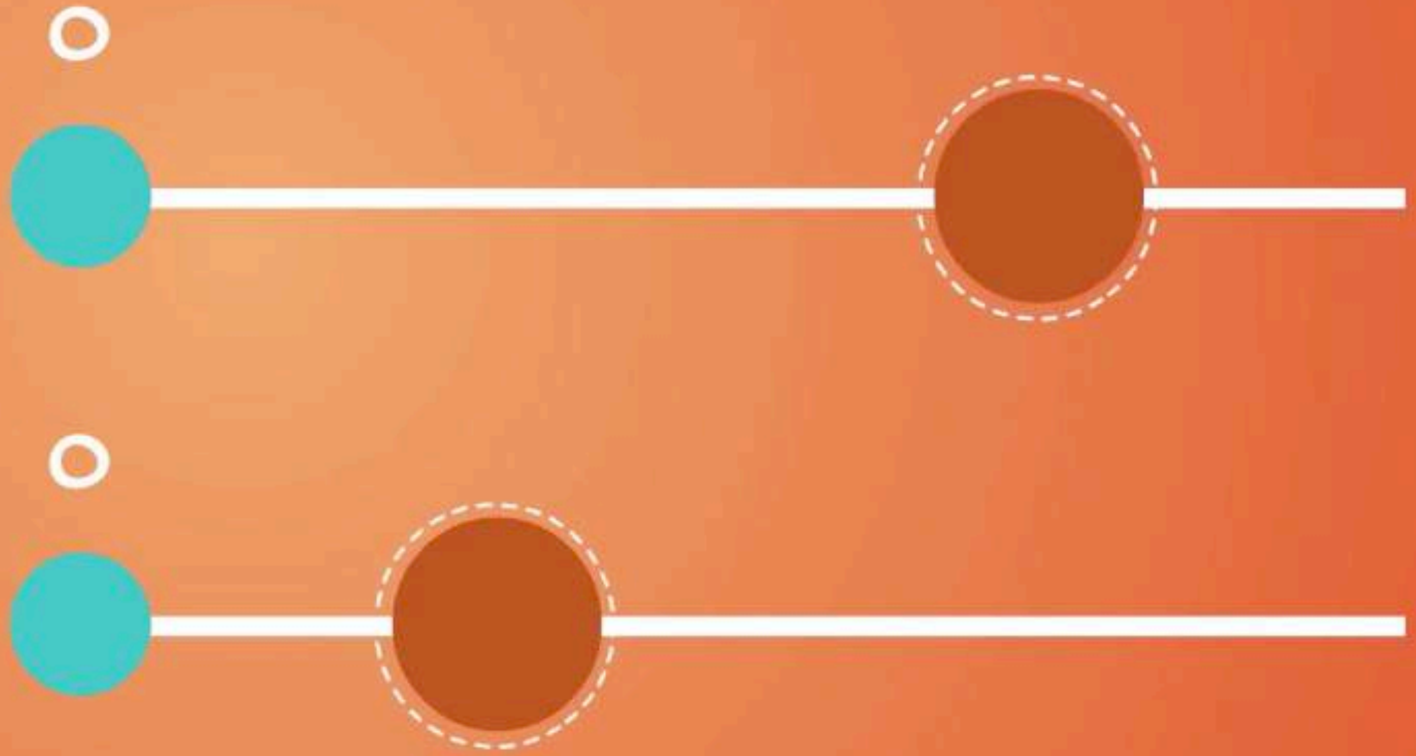
In x% of the cases, the true parameter will fall in the confidence interval

365√DataScience

HYPOTHESIS TESTING

YES

NO

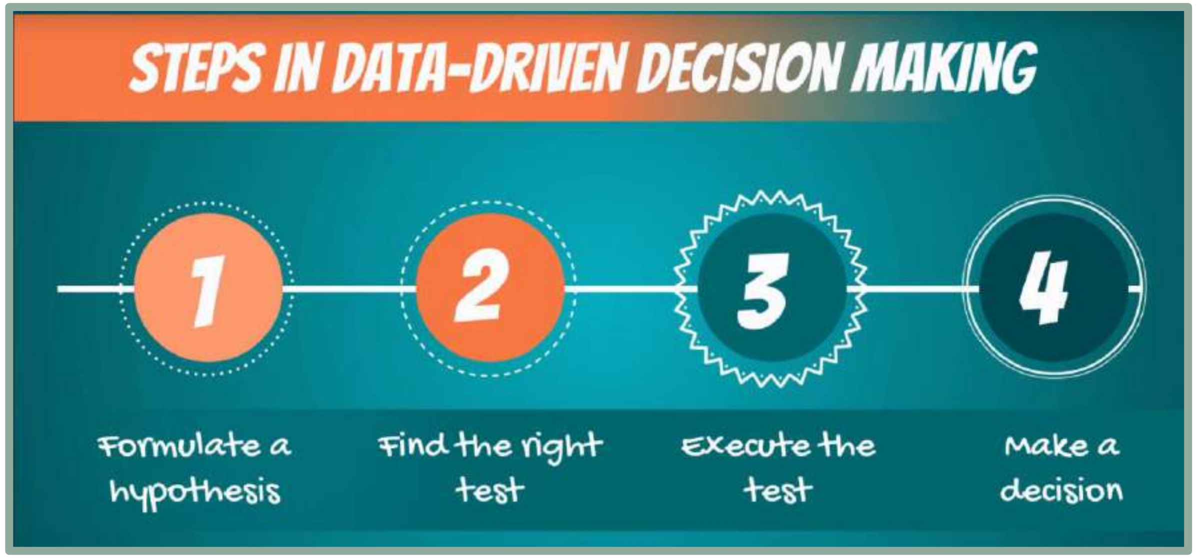


Scientific method

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)

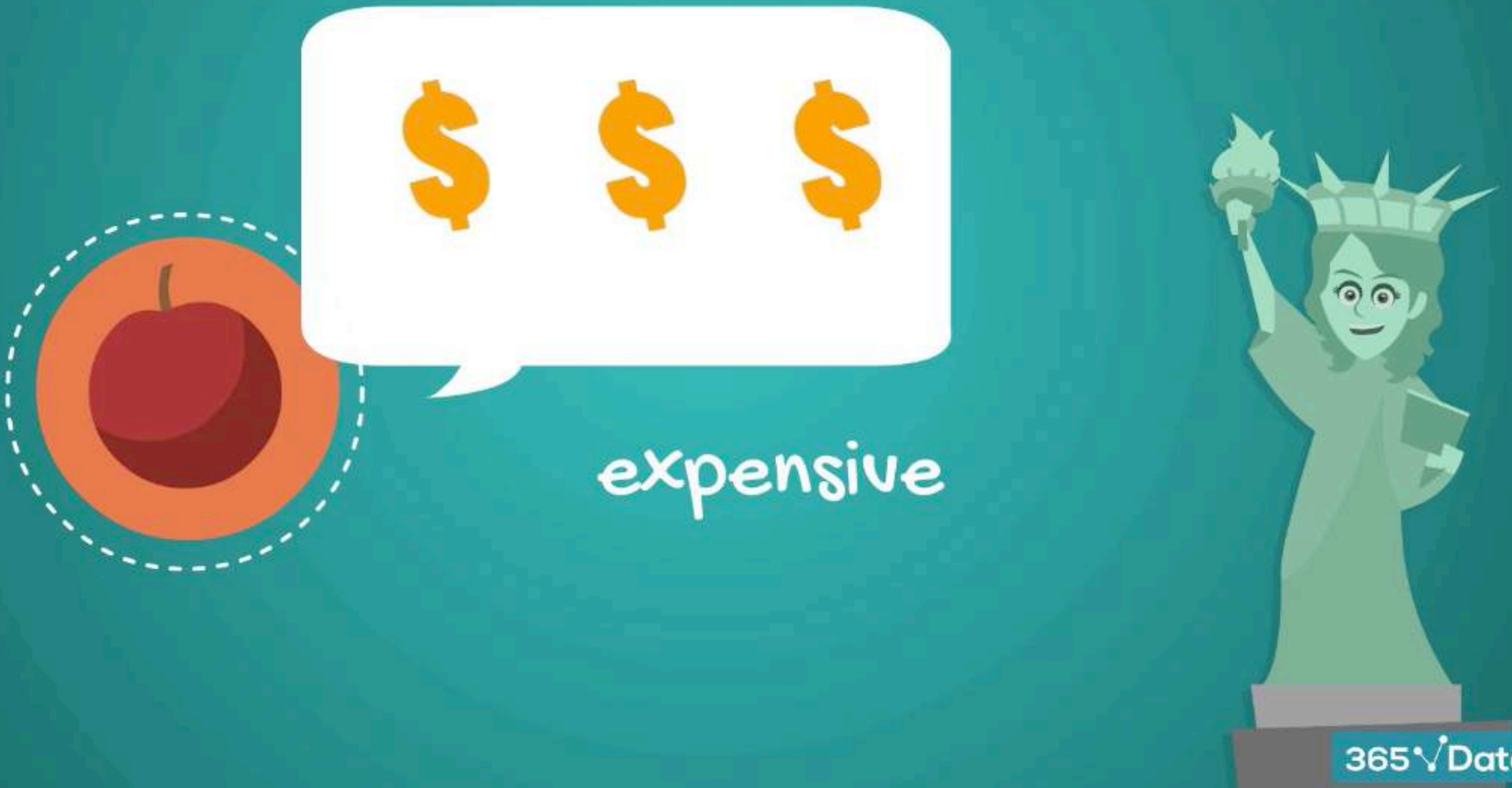
'I see why you would ask this question, as I asked the same one right after I was introduced to hypothesis testing. In statistics, the null hypothesis is the statement **we are trying to reject**. Think of it as the 'status-quo'. The alternative, therefore, is **the change** or **innovation.**

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.

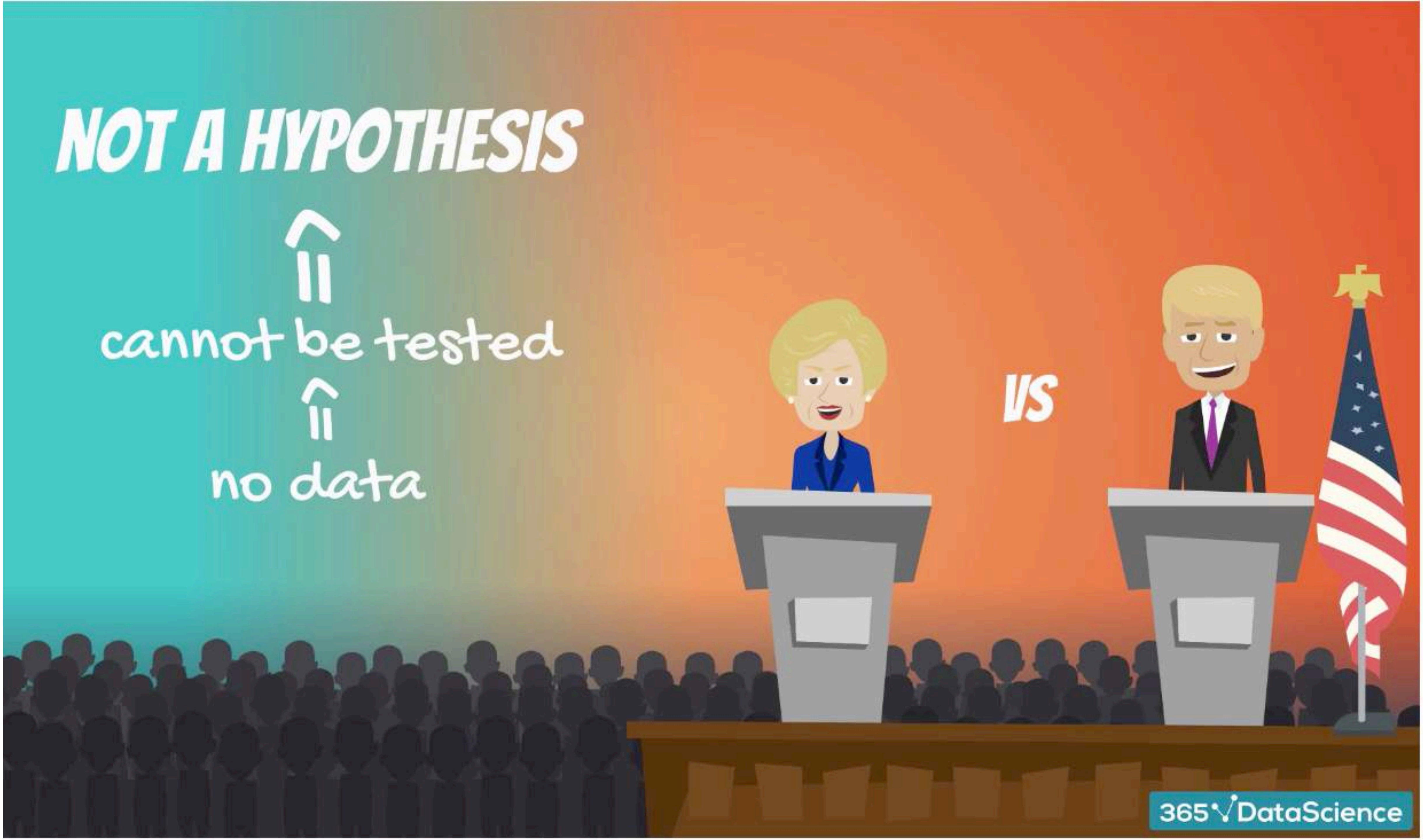
It truly is counter-intuitive in the beginning, but later on, when you start doing the exercises, you will understand the mechanics.'

STATEMENT



365 DataScience


HYPOTHESIS
$ $ $
more expensive
than $1.75


NOT A HYPOTHESIS
cannot be tested
no data
VS
365 DataScience

1.70

MIGHT BE A HYPOTHESIS

⇑

can be tested

⇑

We have data on both administrations

EXAMPLE

HYPOTHESES	NOTATION
Null hypothesis The one to be tested	H_0
Alternative hypothesis Everything else	H_1 or H_A

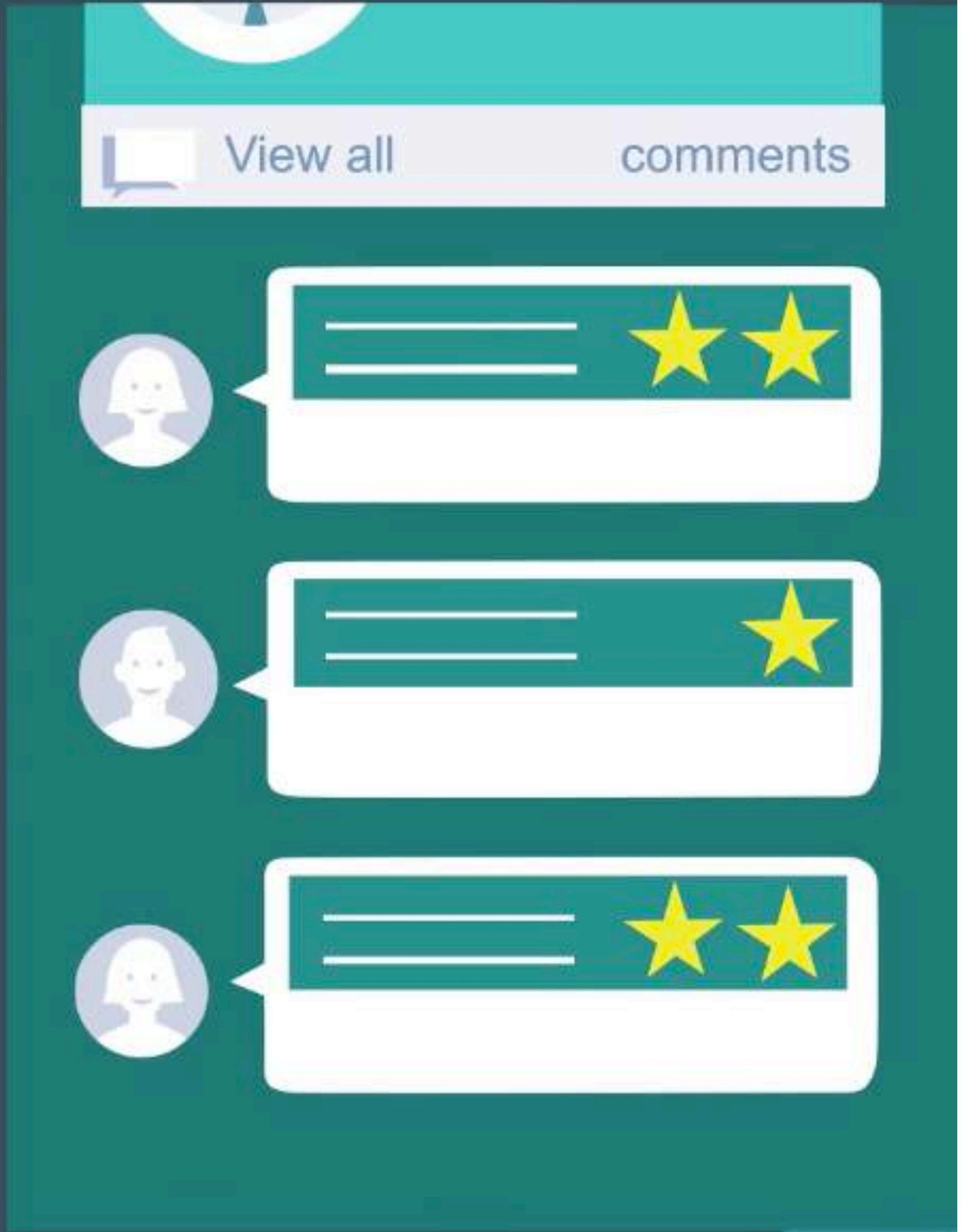
365√DataScience


1.70
NULL HYPOTHESIS
CASE CLOSED
INNOCENT UNTIL PROVEN GUILTY
H_0 IS TRUE UNTIL REJECTED
365 DataScience
In hypothesis testing we aim to reject the null hypothesis

glassdoor™

Glassdoor is a website where current and former emloyees rate their emplyers and the C-level managment. All data is self-reported.

Mean data scientist salary in the US is $ 113,000



EXAMPLE

$$H_0 : \mu_0 = \$113{,}000$$

$$H_1 : \mu_0 \neq \$113{,}000$$

EXAMPLE

$H_0: \mu_0 = \$113{,}000$

TWO-SIDED TEST
YOU CAN ALSO FORM
ONE-SIDED TESTS

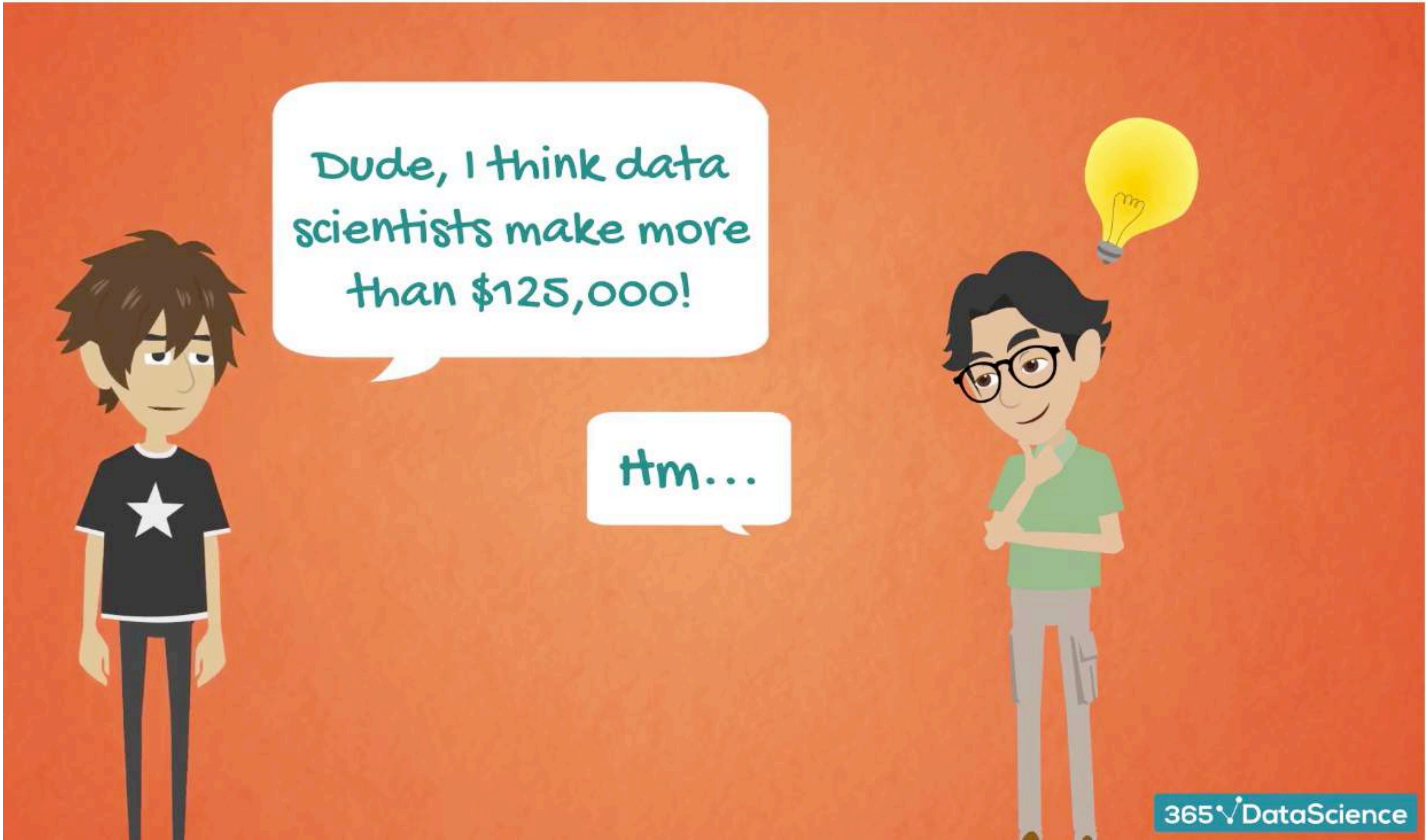
EXAMPLE

$$H_0 : \mu_0 = \$113{,}000$$

Accept if: $\bar{x}$ is close enough to the true mean

Reject if: $\bar{x}$ is too far from the true mean


1.70
NULL HYPOTHESIS
CASE CLOSED
INNOCENT UNTIL PROVEN GUILTY
H_0 IS TRUE UNTIL REJECTED
TOP SECRET
365 DataScience


Dude, I think data scientists make more than $125,000!
Hm...
365√DataScience

THE NULL HYPOTHESIS IS THE STATEMENT WE ARE TRYING TO REJECT. THEREFORE THE NULL IS THE PRESENT STATE OF AFFAIRS WHILE THE ALTERNTIVE IS OUR PERSONAL OPINION.

365√DataScience

1.70

$$H_0 : \mu_0 \geq \$125{,}000$$

$$H_1 : \mu_0 < \$125{,}000$$

OUTCOMES OF TESTS REFER TO POPULATION PARAMETER RATHER THAN SAMPLE STATISTIC

365√DataScience

The researcher is trying to REJECT the null hypothesis

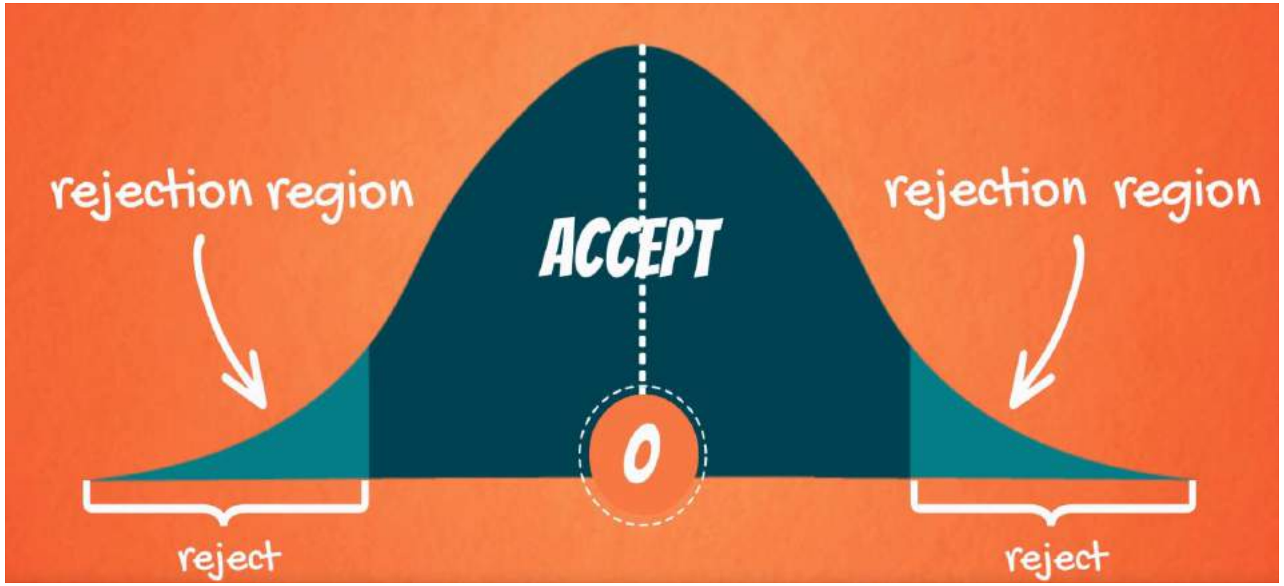
$H_0 : \mu_0 \geq \$125,000$ — STATUS QUO

$H_1 : \mu_0 < \$125,000$ — CHANGE OR INNOVATION

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)

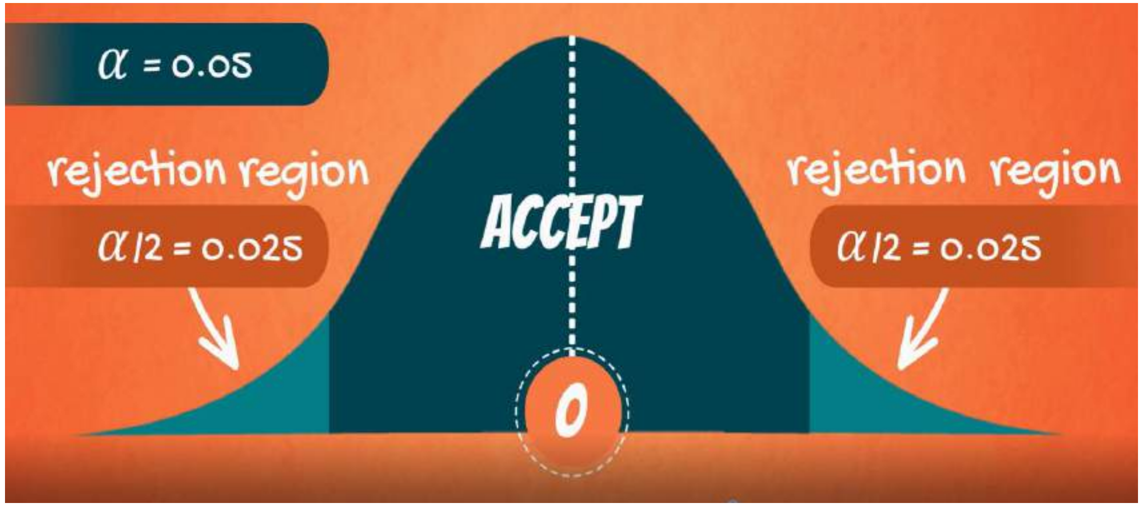
The probability of rejecting a null hypothesis that is true; the probability of making this error.

Common significance levels

0.10 | 0.05 | 0.01

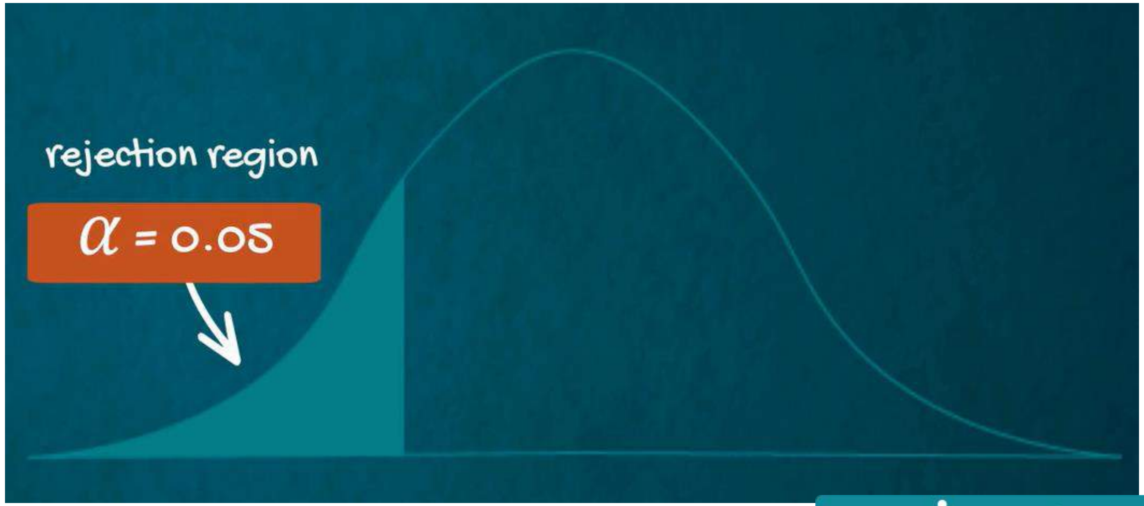
Two-sided (two-tailed) test

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

SIGNIFICANCE LEVEL

α

Probability of committing that error

The probability of rejecting the null hypothesis, if it is true.

most common

Typical values for alpha are:
0.01, 0.05, 0.1

365√DataScience

Need to test if a machine is working properly

YOU EXPECT LITTLE OR NO MISTAKES

α = 0.01

To be precise


I want to carry out an analysis on how students are performing on average.

1.60

UNIVERSITY DEAN:

Population mean grade is 70%

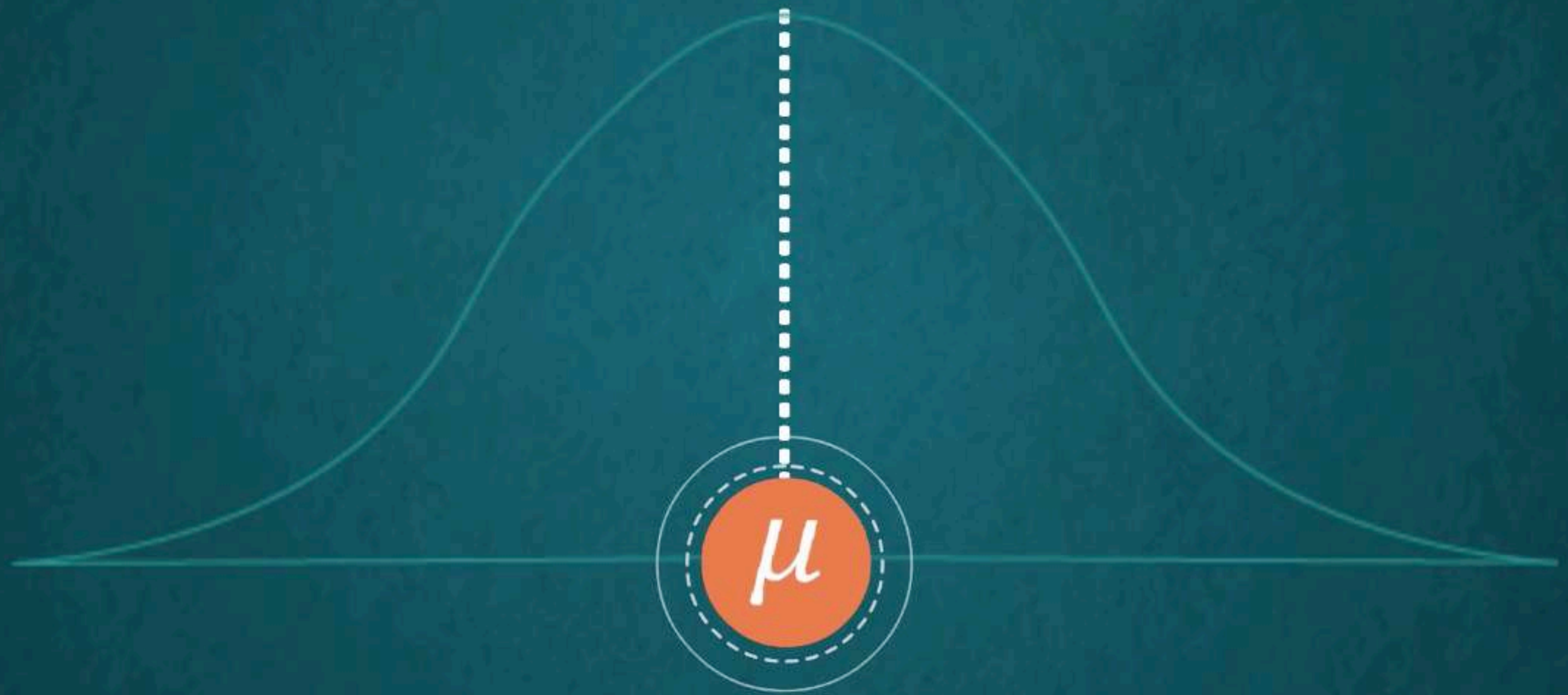
H_0 : Population mean grade is 70%

H_1 : Population mean grade is NOT 70%

$H_0 : \mu_0 = 70\%$

$H_1 : \mu_0 \neq 70\%$

DISTRIBUTION OF GRADES



DISTRIBUTION OF GRADES

sample mean

hypothesized mean

Z TEST:

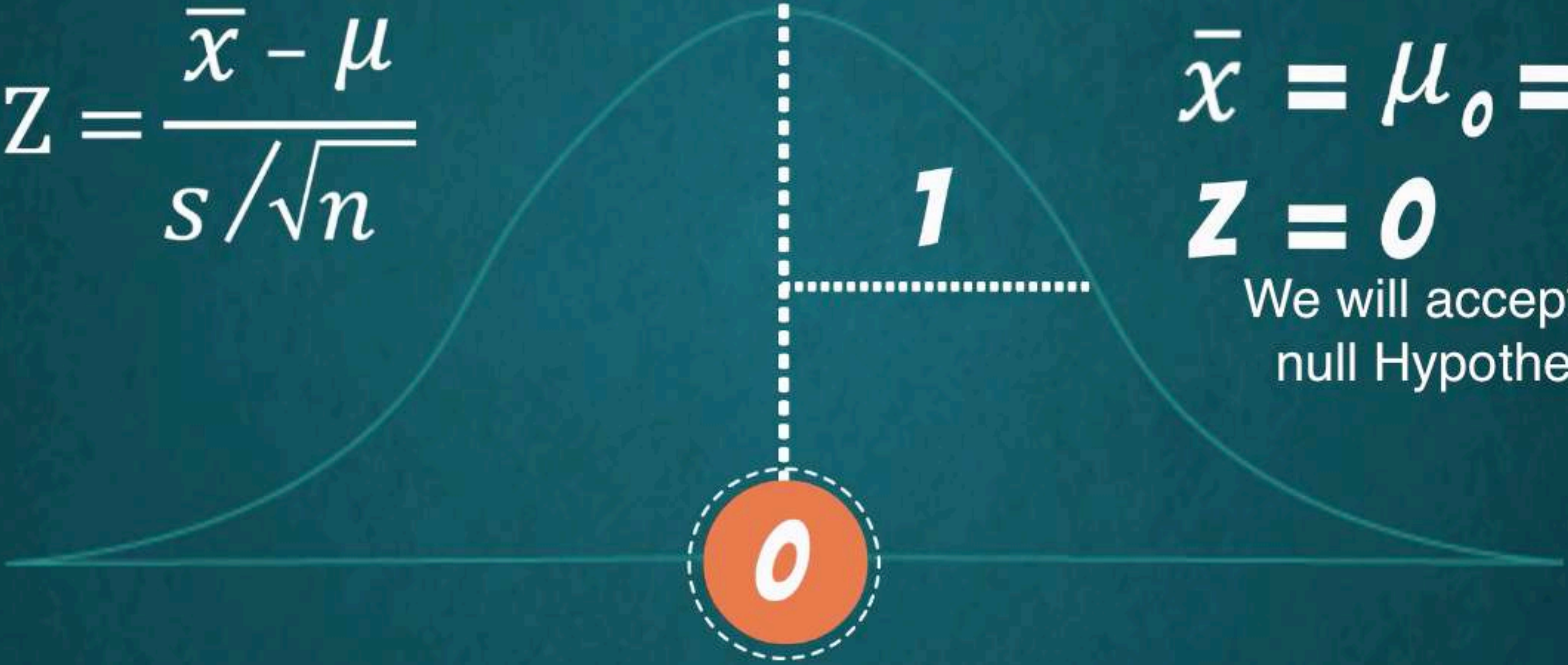
$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

standard error

1.60

DISTRIBUTION OF GRADES

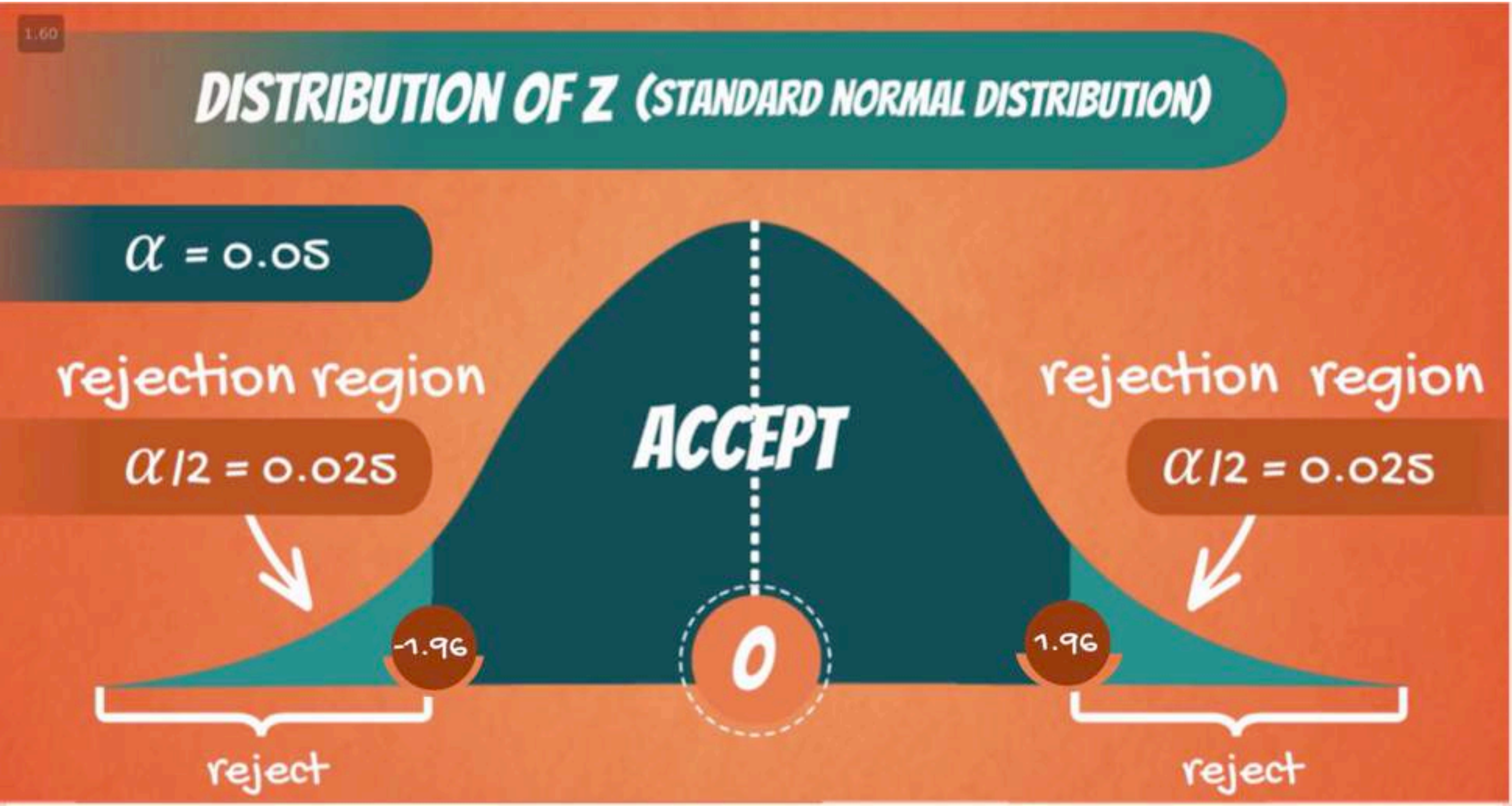
$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$



$\bar{x} = \mu_0 \Rightarrow$

$Z = 0$

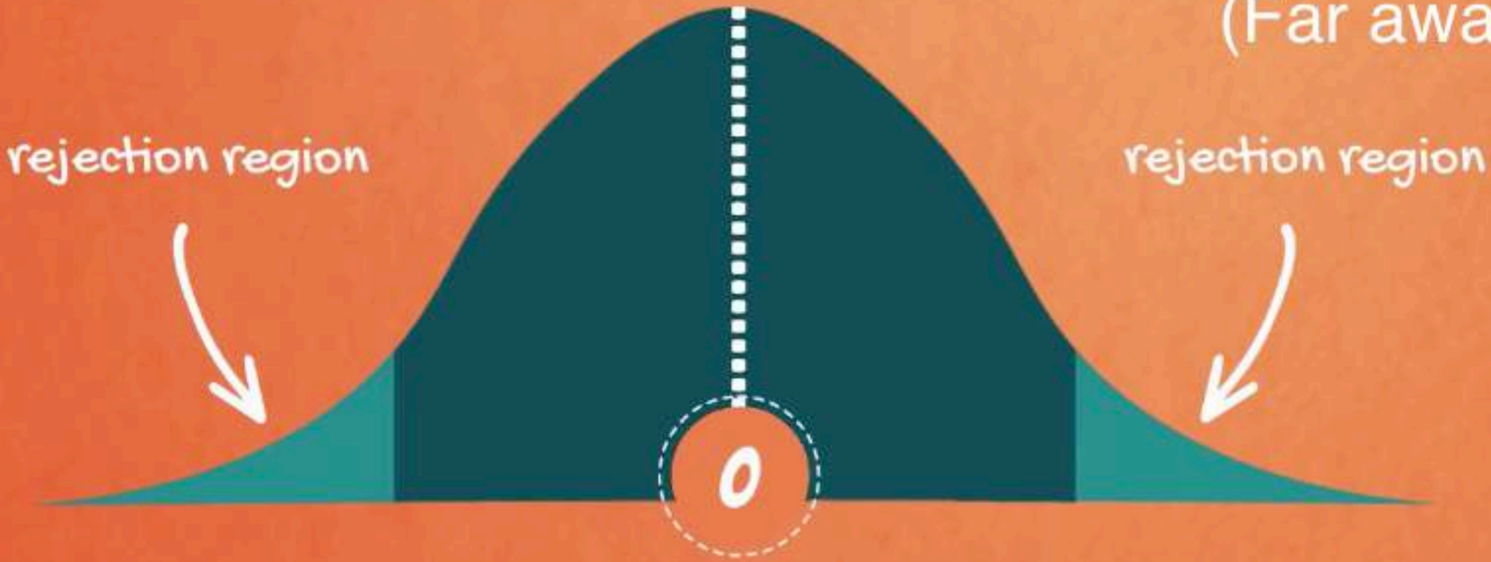
We will accept the null Hypothesis


DISTRIBUTION OF Z (STANDARD NORMAL DISTRIBUTION)
α = 0.05
rejection region
α/2 = 0.025
ACCEPT
rejection region
α/2 = 0.025
-1.96
0
1.96
reject
reject

HOW DOES HYPOTHESIS TESTING WORK?

- Calculate a statistic (e.g. $\bar{x}$)
- Scale it (e.g. $Z = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$)
- Check if Z is in the rejection region

(Far away from Zero)

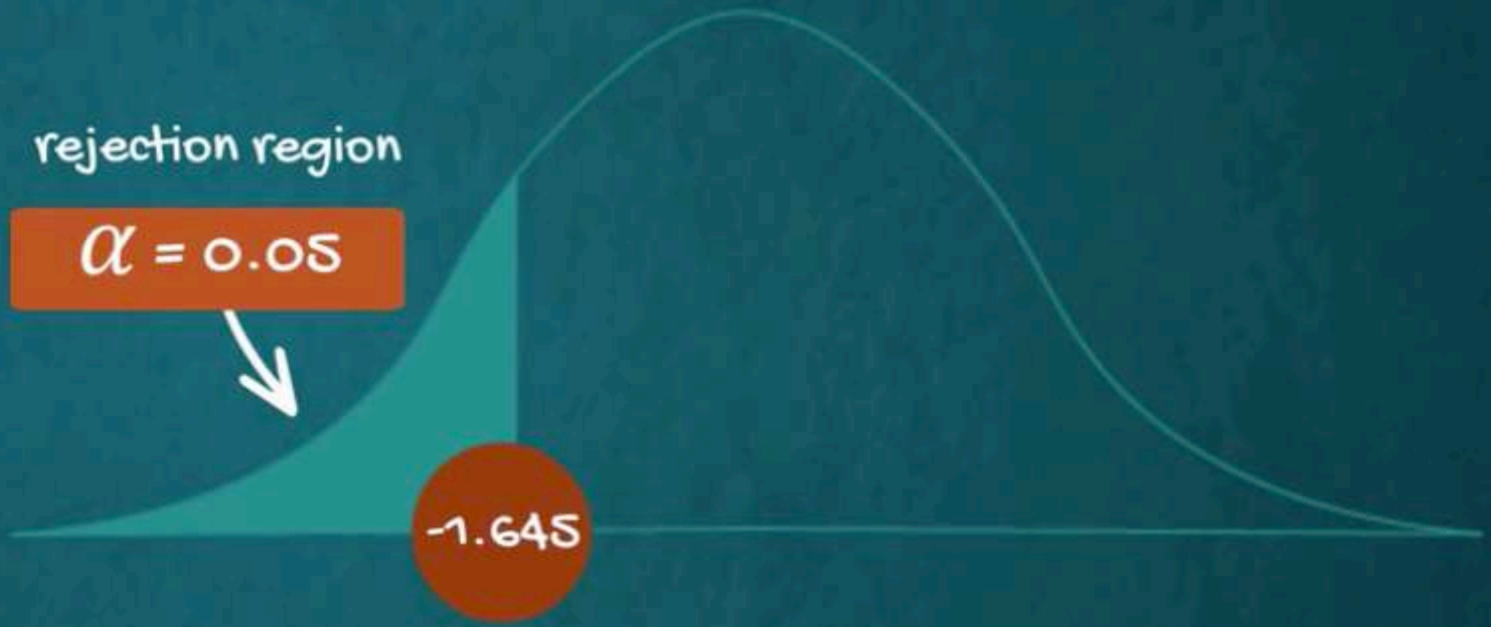


ONE-SIDED TEST

ONE-SIDED TEST

$H_0 : \mu_0 \geq \$125{,}000$

$H_1 : \mu_0 < \$125{,}000$



If $z < -1.645$, we would reject the null hypothesis

1.60

UNIVERSITY DEAN:

H_0 Population mean grade LOWER than 70%

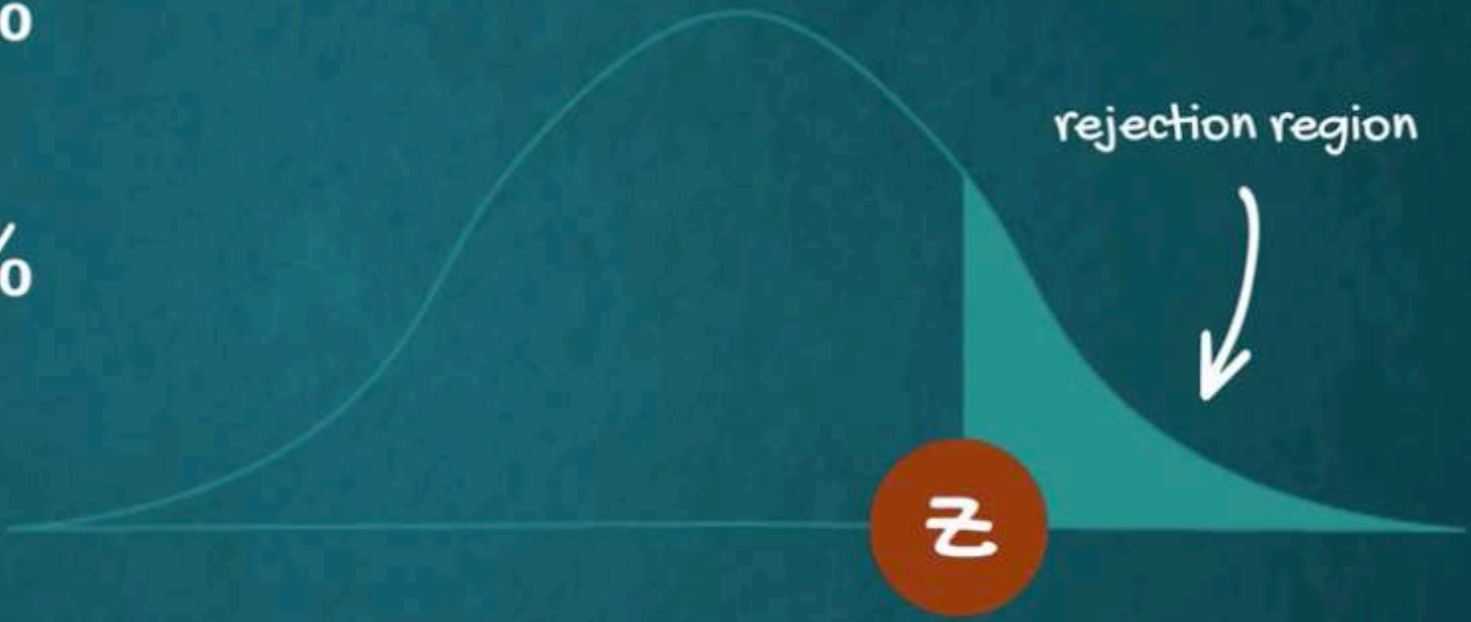
$H_0 : \mu_0 \leq 70\%$

$H_1 : \mu_0 > 70\%$

ONE-SIDED TEST

$H_0 : \mu_0 < 70\ \%$

$H_1 : \mu_0 \geq 70\ \%$



If the test statistic is bigger than the cut-off z-score, we would reject the null, otherwise we wouldn't.

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.

365√DataScience

I

Researcher responsibility of making this Error as he chose alpha.

Reject a true null hypothesis

False positive

probability: α
(Level of significance)

1.50

II

sample size

n

variance

σ

Accept a false null hypothesis

False negative

Probability: β

Goal of hypothesis testing

Rejecting a false null hypothesis

Probability: $1-\beta$

a.k.a. power of the test

1.50

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)	

365√DataScience

H_0 : Status quo

H_0 : She doesn't like you

The truth

		She doesn't like you	She likes you
		H_0 is true	H_0 is false
H_0 (status quo) She doesn't like you (you shouldn't invite her out)	Accept (Do nothing)		Type II error (False negative)  missed your chance
	Reject (Invite her)	Type I error (False positive)  wrongly invited her	

1.50

SINGLE POPULATION

Z

known

T

unknown

SINGLE POPULATION

Z

known

Hypothesis testing.
Single Population. Z known

Testing is done by standardizing the variable at hand and comparing it to the z

We standardize a variable by subtracting the mean and dividing by the standard deviation

$$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

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

$$Z \sim N(\bar{x}-\mu_0, 1)$$

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

$$Z \sim N(\bar{x}-\mu_0, 1)$$

The closest the value to Zero, implies a higher chance to accept the null hypothesis

Z -> standardized variable associated with the test called the Z-score
z -> one from the table and will be referred to as 'the critical value'

Standardization lets us compare the means

Now we will compare Z-score with $z_{\alpha/2}$

Decision rule:
Reject if: absolute value of Z-score > positive critical value (z)

Test for the mean. Population variance known

Data scientist salary

Dataset
$ 117,313
$ 104,002
$ 113,038
$ 101,936
$ 84,560
$ 113,136
$ 80,740
$ 100,536
$ 105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$ 102,848
$ 85,927
$ 112,276
$ 108,637
$ 96,818
$ 92,307
$ 114,564
$ 109,714
$ 108,833
$ 115,295
$ 89,279
$ 81,720

Sample mean	$ 100,200
Population std	$ 15,000
Standard error	$ 2,739
Sample size	30

glassdoor $ 113,000

$$H_0: \mu_0 = \$\,113{,}000$$

$$H_1: \mu_0 \neq \$\,113{,}000$$

Data on Glassdoor is usually self-reported

*Glassdoor is a popular salary and career opportunities information website

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{100200 - 113000}{2739} = -4.67$$

Now we will compare |-4.67| with $z_{\alpha/2}$

365√DataScience

Some z-tables don't include negative values (like this one)

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

As the standard normal distribution is symmetrical around 0, the two statemets are equivalent:

$-4.67 <$ a negative z $\iff$ $4.67 >$ a positive z

Decision rule:
Reject if: absolute value of Z-score > positive critical value (z)

Now we will compare $|-4.67|$ with $z_{\alpha/2}$

365√DataScience

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Decision rule:
Reject if: absolute value of Z-score > positive critical value (z)

5% significance

$\alpha = 0.05$

$z_{0.025} = 1.96$

Test for the mean. Population variance known

Data scientist salary

Dataset
$ 117,313
$ 104,002
$ 113,038
$ 101,936
$ 84,560
$ 113,136
$ 80,740
$ 100,536
$ 105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$ 102,848
$ 85,927
$ 112,276
$ 108,637
$ 96,818
$ 92,307
$ 114,564
$ 109,714
$ 108,833
$ 115,295
$ 89,279
$ 81,720

Sample mean	$100,200
Population std	$ 15,000
Standard error	$ 2,739
Sample size	30

glassdoor $113,000

Decision rule:
Reject if: absolute value of Z-score > positive critical value (z)

5% significance

$\alpha = 0.05$

$z_{0.025} = 1.96$

Z z

4.67 > 1.96 => we reject the null hypothesis

At 5% significance level we have rejected the null hypothesis

At 5% significance level there is no statistical evidence that the mean salary is $113,000

Test for the mean. Population variance known

Data scientist salary

Dataset
$ 117,313
$ 104,002
$ 113,038
$ 101,936
$ 84,560
$ 113,136
$ 80,740
$ 100,536
$ 105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$ 102,848
$ 85,927
$ 112,276
$ 108,637
$ 96,818
$ 92,307
$ 114,564
$ 109,714
$ 108,833
$ 115,295
$ 89,279
$ 81,720
$ 89,344
$ 114,426

Sample mean	$100,200
Population std	$ 15,000
Standard error	$ 2,739
Sample size	30

glassdoor $113,000

1% significance

$\alpha = 0.01$

$z_{0.005} = 2.58$

Z z

4.67 > 2.58 => we reject the null hypothesis

Decision rule:
Reject if: absolute value of Z-score > positive critical value (z)

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.

WE DON'T KNOW

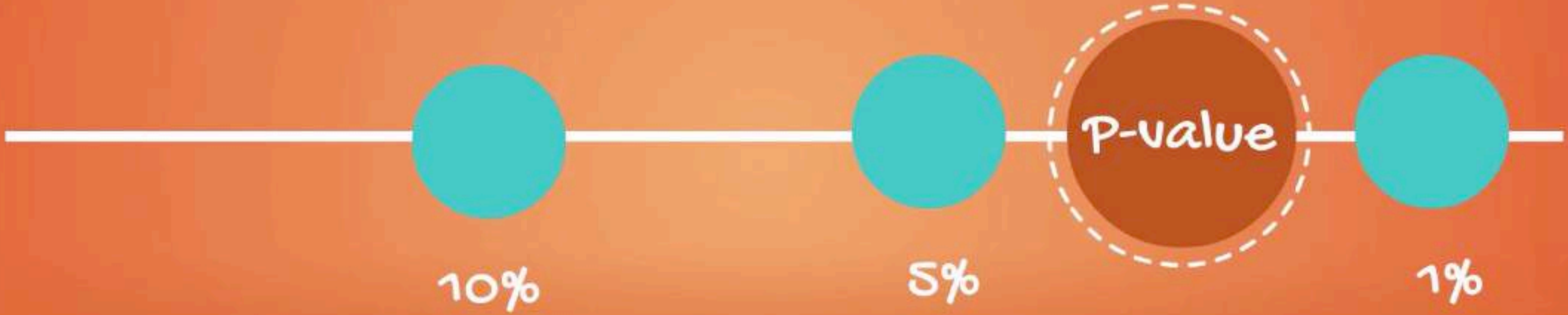
? A level of significance after which we can no longer do it

WE KNOW

- ✔ How to test hypotheses
- ✔ How to reject them...
- ✔ ... at various levels of significance

365√DataScience

P-VALUE



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

EXAMPLE

Standard error = 2739

Population std = 15000

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

n = 30

z= - 4.67

We rejected the null at 0.05 and 0.01

HOW MUCH?

365√DataScience

1.50

Standard normal distribution

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

p-value = 1 - (number from table) = 0.001

1.50

EXAMPLE

$z = -4.67$

Rule: You should reject the null hypothesis, if

P-VALUE < α

P-value = 0.0001

Test at 90%: $0.0001 < 0.1$

Test at 95%: $0.0001 < 0.05$ => REJECT THE NULL HYPOTHESIS

Test at 99%: $0.0001 < 0.01$

365√DataScience

EXAMPLE 2

$z = 2.12$

Rule: You should reject the null hypothesis, if

P-VALUE < α

Test at 90%:
Test at 95%: ===> REJECT

Test at 99%: => CANNOT REJECT

Standard normal distribution

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

1.50

EXAMPLE 2

How to find the p-value manually

One-sided p-value: 1 - the number from the table => 1-0.983= = 0.017

Two-sided p-value: (1 - the number from the table) x2 => (1-0.983)x2= = 0.034

365√DataScience

WHERE AND HOW ARE P-VALUES USED?

Most statistical software calculates p-values for each test

Researcher decides significance post-factum

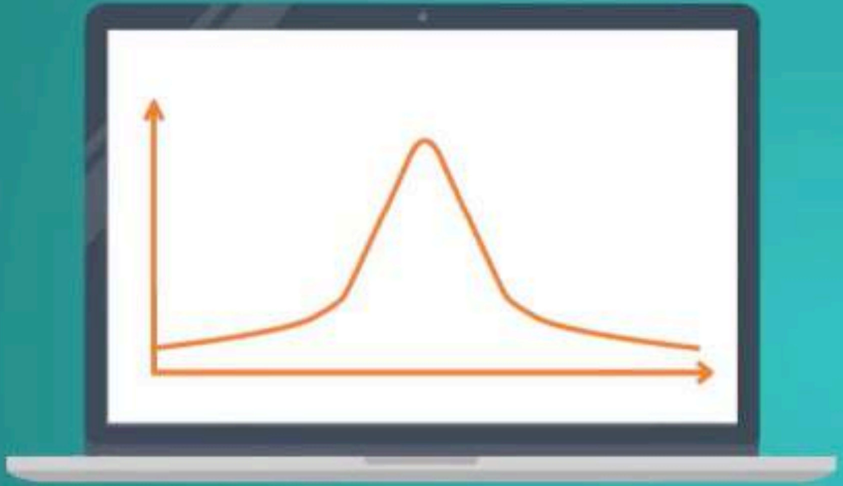
P-values are usually found with 3 digits after the dot x.xxx

The closer to 0.000 the p-value, the better

The closer the p-value to Zero, the more significant is the result that we obtained.

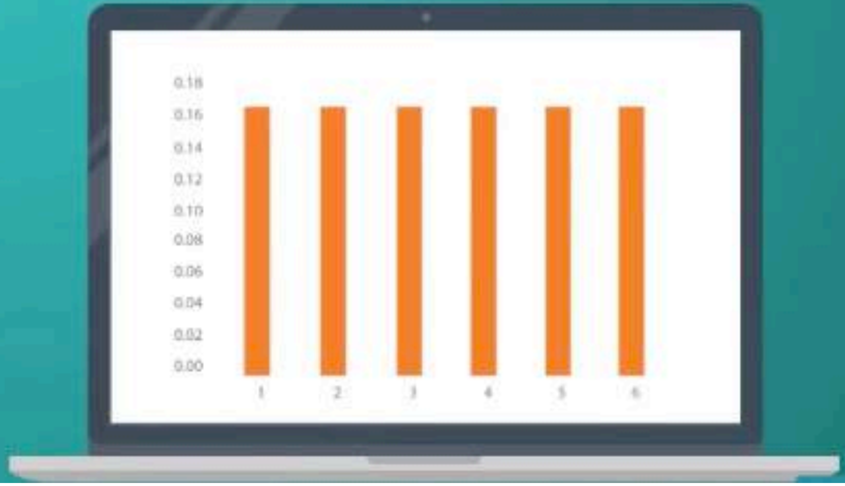
P-VALUE IS A UNIVERSAL CONCEPT THAT WORKS WITH EVERY DISTRIBUTION

normal



student's T

binomial



uniform

IF THE P-VALUE IS LOWER THAN THE LEVEL OF SIGNIFICANCE

YOU REJECT THE NULL HYPOTHESIS

SINGLE POPULATION

TASK: ESTIMATE IF OUR COMPETITOR HAS A HIGHER OPEN RATE

MARKETING ANALYST

YOUR COMPANY: 40%

Definition: measure of how many people on the email list actually opened an email

365√DataScience

1.60

One side test

In hypothesis testing we aim to reject the null hypothesis

When we want to test that the rating is higher than 40%, the null hypothesis actually says the opposite statement.

HYPOTHESES

$$H_0 : \mu_{OR} \leq 40\%$$

$$H_1 : \mu_{OR} > 40\%$$

365√DataScience

Test for the mean. Population variance unknown

Email spying example

Hypothesis testing.
Population variance unknown. Small sample (t)

$H_0: \mu_{OR} \leq 40\%$

Open rate
26%
23%
42%
49%
23%
59%
29%
29%
57%
40%

Sample mean	37.70%
Sample standard dev	13.74%
Standard error	4.34%
Null hypothesis value	40%

$$T = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{37.70\% - 40\%}{4.34\%} = -0.53$$

Like confidence intervals with varaince unknown and a small sample, the correct statistic to use is the t-statistic

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
inf.	1.282	1.645	1.960	2.326	2.576

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{37.70\% - 40\%}{4.34\%} = -0.53$$

degrees of freedom = n - 1 = 9

0.05 one-sided significance

t = 1.833

Test for the mean. Population variance unknown

Email spying example

Open rate
26%
23%
42%
49%
23%
59%
29%
29%
57%
40%

Sample mean	37.70%
Sample standard dev	13.74%
Standard error	4.34%
Null hypothesis value	40%
T-score	-0.53

Tests

	One sided
t-stat 95%	1.83

$$H_0: \mu_{OR} \leq 40\%$$

T t

$$0.53 < 1.83$$

=> we accept the null hypothesis

At this level of significance, statistically we cannot say that the email open rate of the competitor is higher than 40%.

Decision rule:

Accept if: The absolute value of the T-score < critical value t

Reject if: The absolute value of the T-score > critical value t

dy

Display Settings

365√DataScience

Test for the mean. Population variance unknown
Email spying example

H_0: $\mu_{OR} \leq 40\%$

Open rate
26%
23%
42%
49%
23%
59%
29%
29%
57%
40%

Sample mean	37.70%
Sample standard dev	13.74%
Standard error	4.34%
Null hypothesis value	40%
T-score	-0.53

Tests

	One sided
t-stat 95%	1.83
t-stat 99%	2.82
p-value	0.304

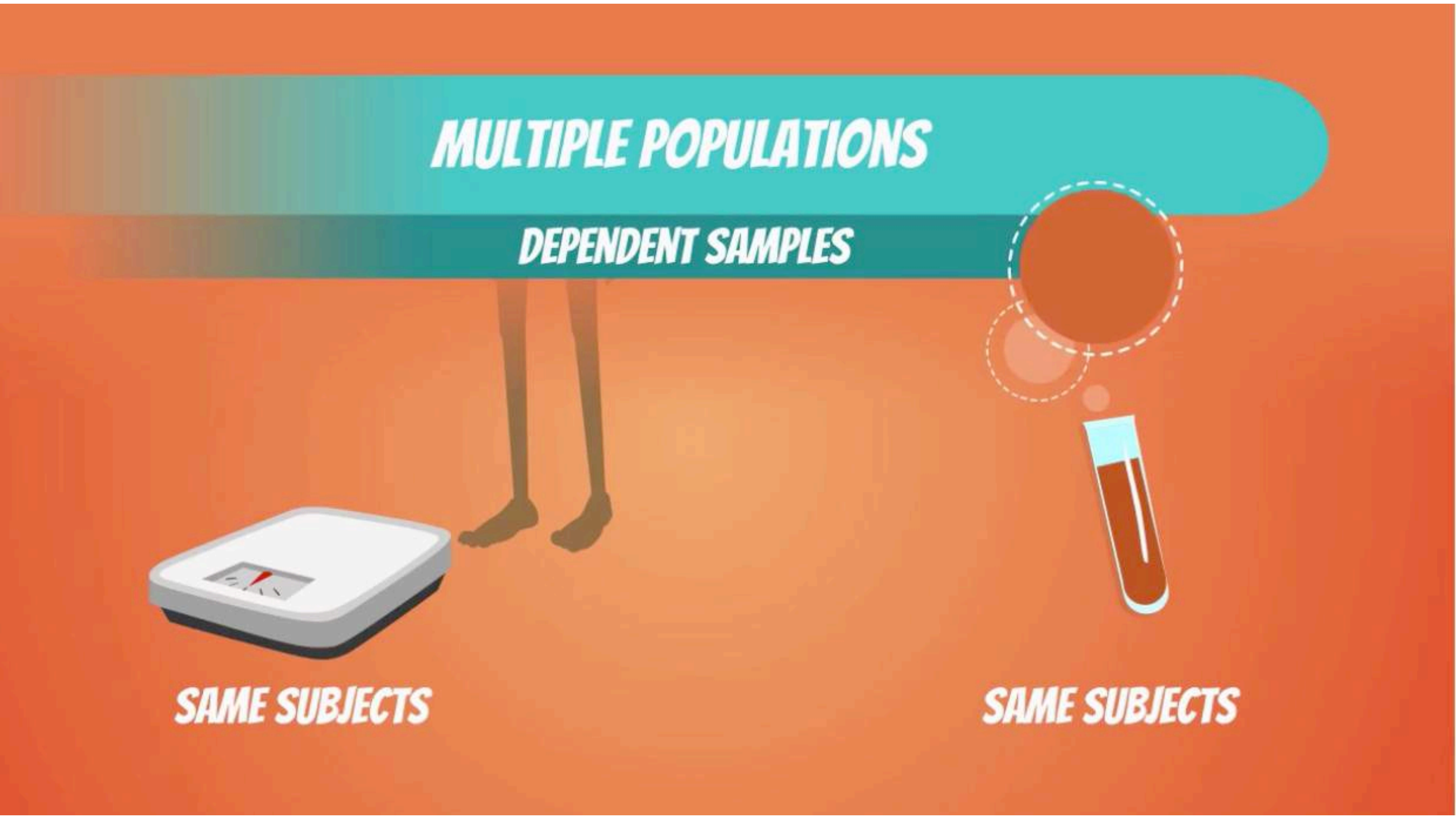
p-value = 0.304 > 0.01 => we accept the null hypothesis

p-value = 0.304 > 0.05 => we accept the null hypothesis

Decision rule:
Accept if: p-value > α
Reject if: p-value < α

If we cannot reject a test at 0.05, we cannot reject it at smaller levels either

MULTIPLE POPULATIONS


MULTIPLE POPULATIONS
DEPENDENT SAMPLES
SAME SUBJECTS
SAME SUBJECTS

DEPENDENT SAMPLES

1.60

we expect the levels of mg to be increasing. Therefore, this is the alternative hypothesis as we are aiming to reject the null.

HYPOTHESES

Population Before		Population After
$H_0 : \mu_B$	$\geq$	μ_A
$H_1 : \mu_B$	$<$	μ_A

365√DataScience

HYPOTHESES

$H_0 : \mu_B \geq \mu_A$

$H_1 : \mu_B < \mu_A$

$H_0 : \mu_B - \mu_A \geq 0$

$H_0 : D_0 \geq 0$

$H_1 : D_0 < 0$

Test the mean. Dependent Samples

Magnesium levels example

Before	After	Difference (B - A)
2	1.7	0.3
1.4	1.7	-0.3
1.3	1.8	-0.5
1.1	1.3	-0.2
1.8	1.7	0.1
1.6	1.5	0.1
1.5	1.6	-0.1
0.7	1.7	-1
0.9	1.7	-0.8
1.5	2.4	-0.9

Sample mean	-0.33
Standard deviation	0.45
Standard error	0.14

Hypothesis testing.
Population variance unknown.
Dependent samples. Small sample (t)

$$H_0: D_0 \geq 0$$

1. Small sample
2. We assume normal distribution of the population
3. Variance unknown

=> t-statistic

$$T = \frac{\bar{d} - \mu_0}{St.error} = \frac{-0.33 - 0}{0.14} = -2.29$$

*Note that with rounding to two decimal places, the result is -2.36. In Excel, however, numbers are calculated with a higher precision. Therefore, -2.29 is the exact result that we get.

Dataset | Difference | Mean | Standard error | T-score | p-value

365√DataScience

HYPOTHESES

$H_0 : \mu_B \geq \mu_A$

$H_1 : \mu_B < \mu_A$

$H_0 : \mu_B - \mu_A \geq 0$

$H_0 : D_0 \geq 0$

$H_1 : D_0 < 0$

Test the mean. Dependent Samples
Magnesium levels example

Before	After	Difference (B - A)
2	1.7	0.3
1.4	1.7	-0.3
1.3	1.8	-0.5
1.1	1.3	-0.2
1.8	1.7	0.1
1.6	1.5	0.1
1.5	1.6	-0.1
0.7	1.7	-1
0.9	1.7	-0.8
1.5	2.4	-0.9

Sample mean	-0.33
Standard deviation	0.45
Standard error	0.14
T-score	-2.29
p-value	0.024

$$H_0: D_0 \geq 0$$

Magnesium pill:

1. Researcher should be very cautious
2. Medicine entails more precise tests
3. Increasing sample size always leads to a better study

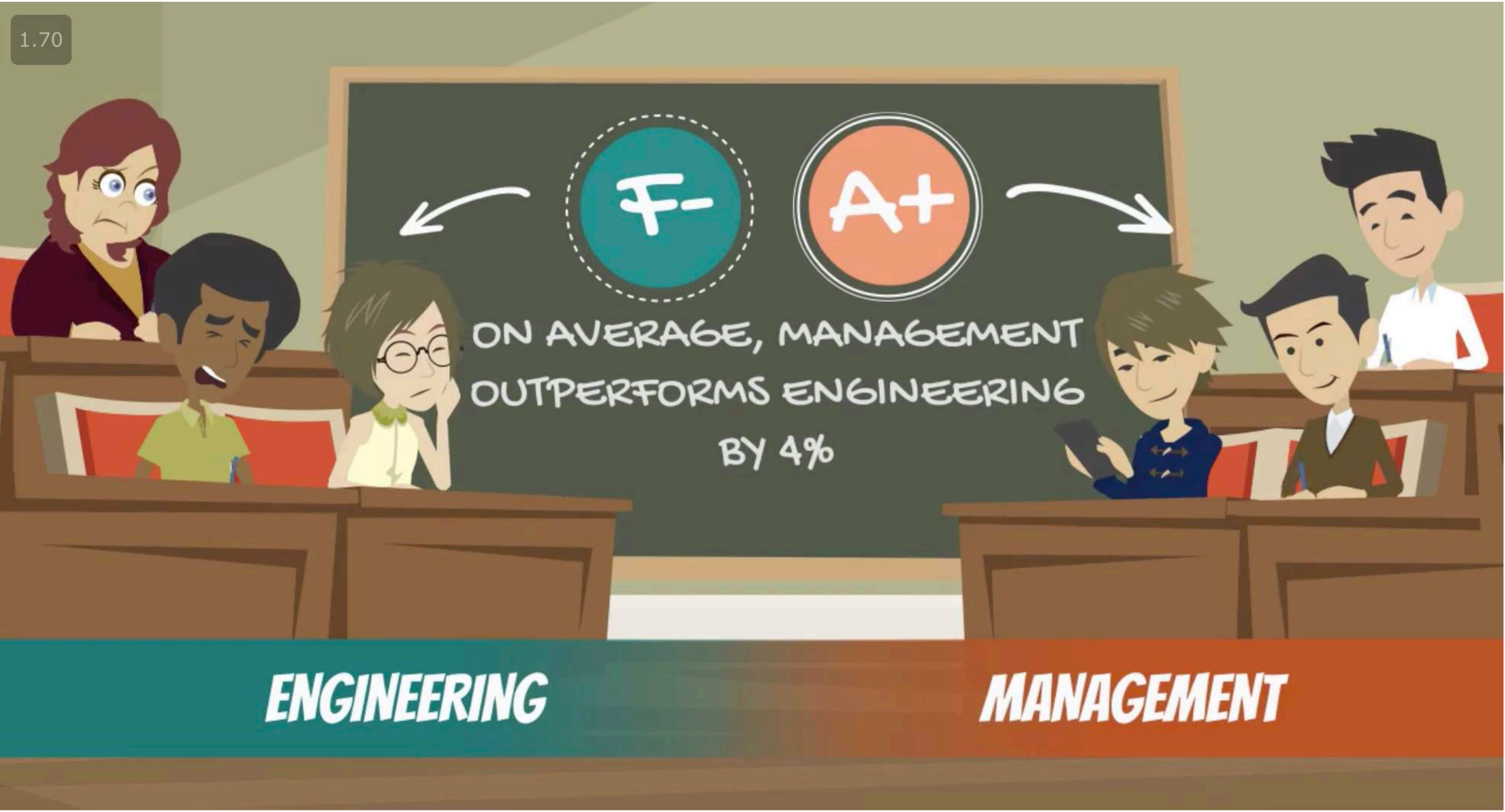
5% significance 0.024 < 0.05 => reject the null hypothesis

1% significance 0.024 > 0.01 => accept the null hypothesis

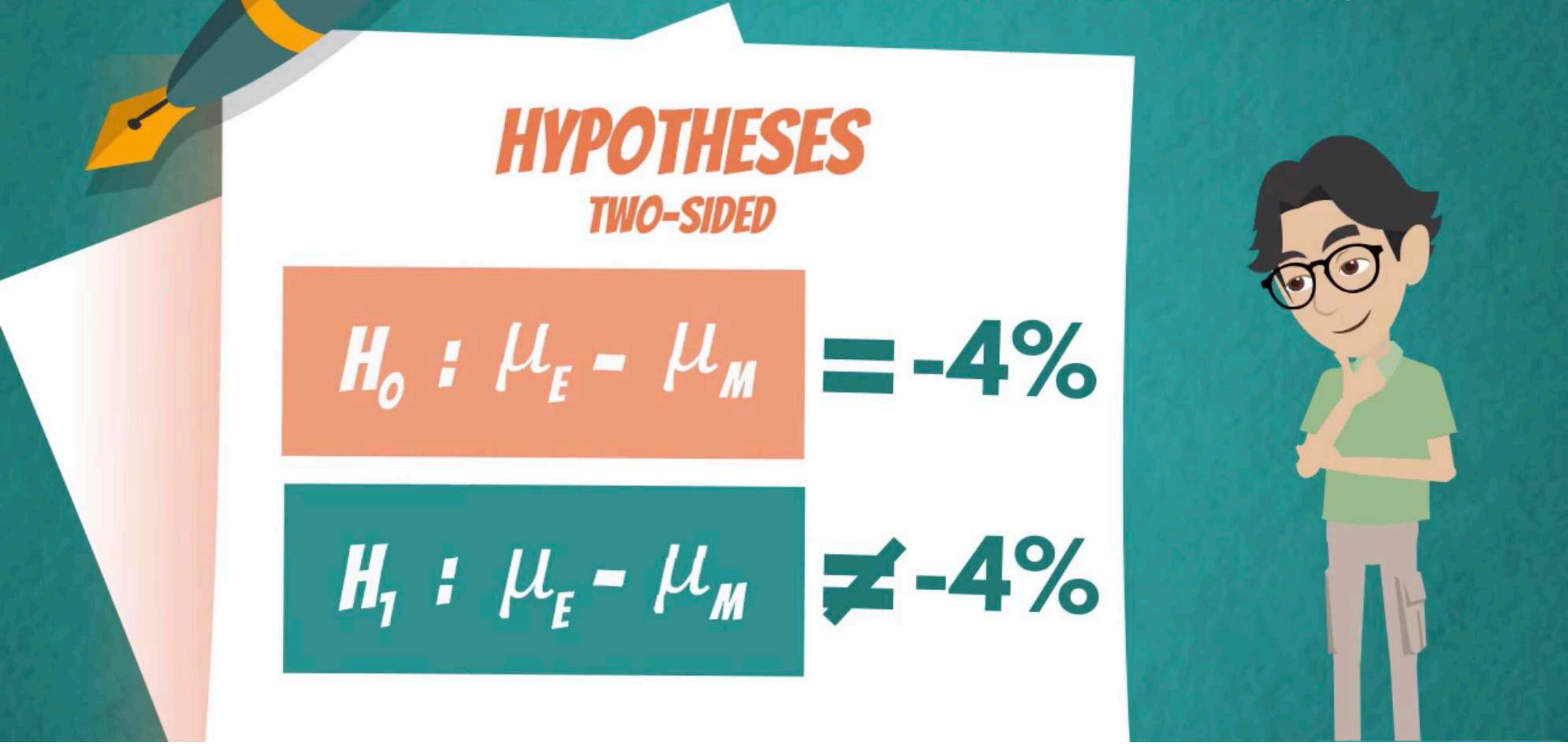
Decision rule:
Accept if: $p > \alpha$
Reject if: $p < \alpha$

The lowest significance level at which we can reject the null hypothesis is 0.024.
This is exactly the p-value

INDEPENDENT SAMPLES, KNOWN VARIANCE


1.70
F-
A+
ON AVERAGE, MANAGEMENT OUTPERFORMS ENGINEERING BY 4%
ENGINEERING
MANAGEMENT

Two sided Hypothesis: Research question is not to find the difference but to check if it is exactly 4%.



Test for two means. Independent samples, variance known

University example

	Engineering	Management	Difference
Size	100	70	?
Mean	58%	65%	-7.00%
Population std	10%	6%	1.23%

Hypothesized diff.	-4%
Z-score	-2.44
p-value	0.015

$$\sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}$$

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}}$$

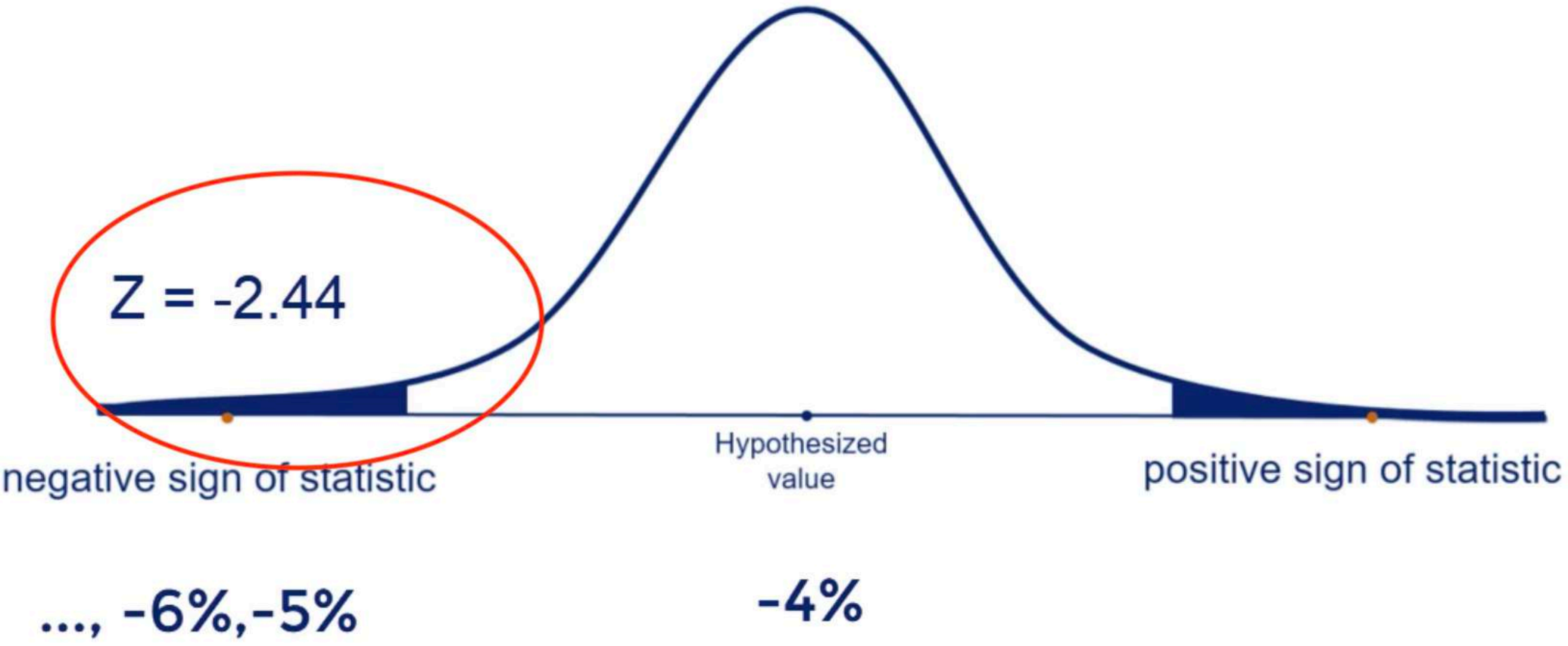
0.015 < 0.05 => we reject the null

There is enough statistical evidence that the mean difference is NOT 4%

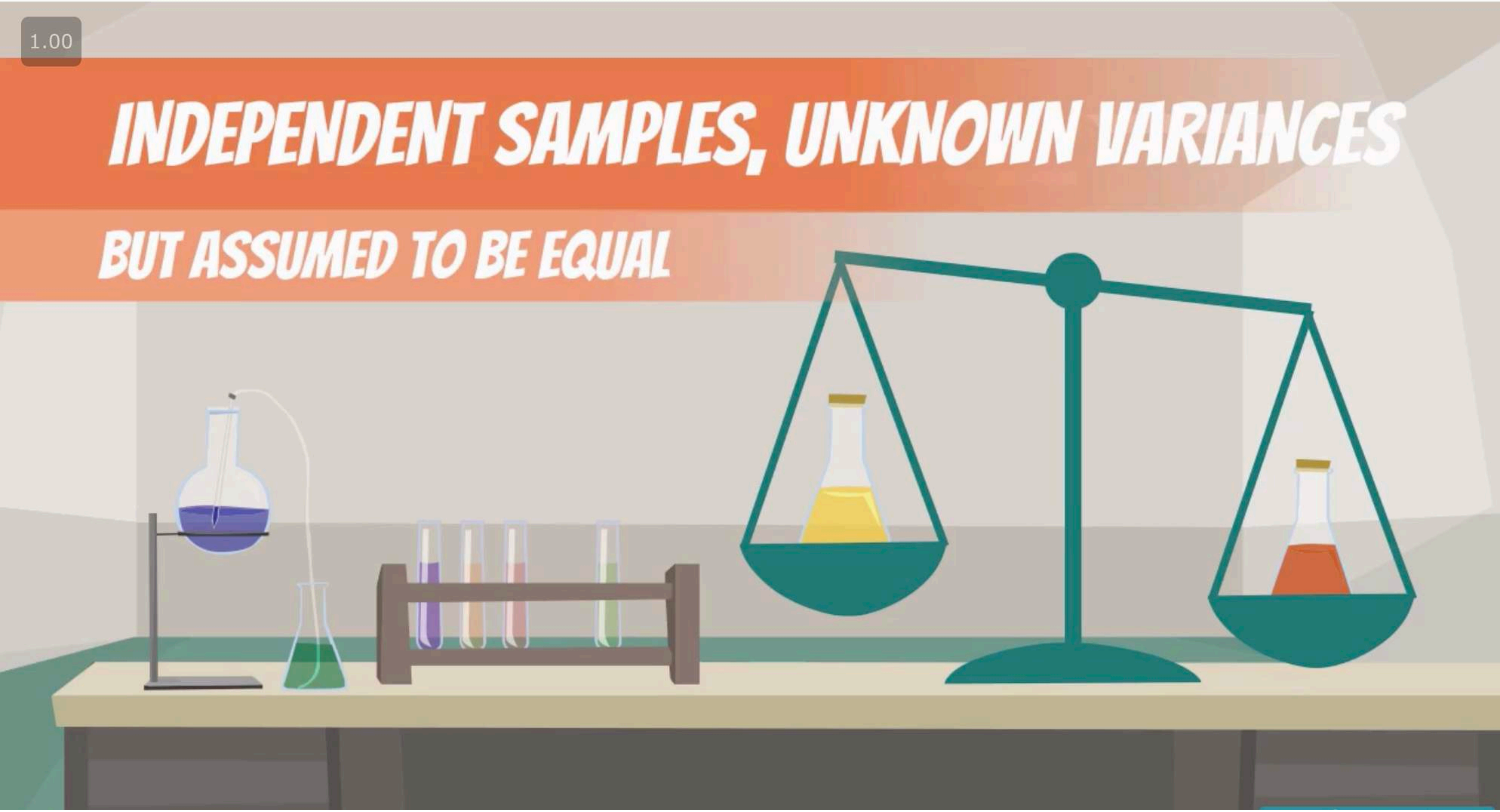
z-statistic	t-statistic	Up to the researcher z-statistic
Big samples	Small samples	Big samples
Known variances	Unknown variances	Unknown variances

Is the difference higher or lower than 4%?

The sign of the test statistic shows if the mean is lower or higher than the hypothesized value



365√DataScience


INDEPENDENT SAMPLES, UNKNOWN VARIANCES
BUT ASSUMED TO BE EQUAL

STATEMENT

Are appples in NY as expensive as in LA

HYPOTHESES

$$H_0 : \mu_{NY} - \mu_{LA} = 0$$

$$H_1 : \mu_{NY} - \mu_{LA} \neq 0$$

Testing of two means. Independent samples, variances unknown but assumed to be equal

Apples example

NY apples	LA apples
$ 3.80	$ 3.02
$ 3.76	$ 3.22
$ 3.87	$ 3.24
$ 3.99	$ 3.02
$ 4.02	$ 3.06
$ 4.25	$ 3.15
$ 4.13	$ 3.81
$ 3.98	$ 3.44
$ 3.99	
$ 3.62	

	NY	LA
Mean	$3.94	$3.25
Std. deviation	$0.18	$0.27
Sample size	10	8

Pooled variance	0.05
Pooled std	0.11
T-score	6.53
p-value	0.000

Did this hypothesis make much sense?
We could easily say that the prices are different, no testing needed.

$$s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2} = \frac{(10-1)0.0324+(8-1)0.0729}{10+8-2} = 0.05$$

$$\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} = \sqrt{\frac{0.05}{10} + \frac{0.05}{8}} = 0.11$$

$$\mathrm{T} = \frac{\bar{d}-\mu_0}{St.error} = \frac{0.69-0}{0.11} = 6.53$$

The p-value is somewhere around 0.000001

Researchers are always looking for those .000

It means that the test is extremally significant and the probability of making a type I error is virtually 0.

We reject the null hypothesis at all common and many uncommon levels of significance

t-statistic

Small samples

Unknown variances

Degrees of freedom = combined sample size - number of variables = 10 + 8 - 2 = 16

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
inf.	1.282	1.645	1.960	2.326	2.576

Generally, for Z and T, a value higher than 4 is extremely significant

Rule of thumb: reject the null hypothesis when the T-score is bigger than 2

GENDER PAY GAP

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

1.70

We are going to test if there is a significant difference in the salaries, based on gender

A1

Practical example. Hypothesis testing.

Spark Fortress Inc. HR data

Surname	Name	Age	Gender	Country	Ethnicity	Start_date	Department	Position	Salary
Warfield	Sarah	39	Female	United States	Asian	3/30/2015	IT/IS	Sr. Network Engineer	$114,816.00
South	Joe	52	Male	United States	White	11/10/2014	IT/IS	Sr. Network Engineer	$110,240.00
Boutwell	Bonalyn	30	Female	United States	Asian	2/16/2015	Admin Offices	Sr. Accountant	$72,696.00
Foster-Baker	Amy	38	Female	United States	White	1/5/2009	Admin Offices	Sr. Accountant	$72,696.00
Sweetwater	Alex	51	Male	United States	White	8/15/2011	Software Engineering	Software Engineering Mana	$56,160.00
Del Bosque	Keyla	38	Female	United States	Black or African American	1/9/2012	Software Engineering	Software Engineer	$118,809.60
Carabbio	Judith	30	Female	United States	White	11/11/2013	Software Engineering	Software Engineer	$116,480.00
Martin	Sandra	30	Female	United States	Asian	11/11/2013	Software Engineering	Software Engineer	$115,460.80
Saada	Adell	31	Female	United States	White	11/5/2012	Software Engineering	Software Engineer	$102,440.00
Szabo	Andrew	34	Male	United States	White	7/7/2014	Software Engineering	Software Engineer	$99,840.00
Andreola	Colby	38	Female	United States	White	11/10/2014	Software Engineering	Software Engineer	$99,008.00
LeBlanc	Brandon R	33	Male	United States	White	1/5/2016	Admin Offices	Shared Services Manager	$114,400.00
Smith	John	33	Male	United States	Black or African American	5/18/2014	Sales	Sales Manager	$116,480.00
Daneault	Lynn	27	Female	United States	White	5/5/2014	Sales	Sales Manager	$112,320.00
Moumanil	Maliki	43	Male	United States	Black or African American	5/13/2013	Production	Production Technician II	$60,320.00
Sahoo	Adil	31	Male	United States	White	8/30/2010	Production	Production Technician II	$60,320.00
Blount	Dianna	27	Female	United States	White	4/4/2011	Production	Production Technician II	$56,160.00
Faller	Megan	39	Female	United States	Black or African American	7/7/2014	Production	Production Technician II	$56,160.00
Monkfish	Erasumus	25	Male	United States	White	11/7/2011	Production	Production Technician II	$56,160.00
Nowlan	Kristie	32	Female	United States	White	11/10/2014	Production	Production Technician II	$54,891.20
Lunquist	Lisa	35	Female	United States	Black or African American	8/19/2013	Production	Production Technician II	$54,288.00
Burkett	Benjamin	40	Male	United States	White	4/4/2011	Production	Production Technician II	$54,080.00
							Production	Production Technician II	$54,080.00
							Production	Production Technician II	$54,080.00
							Production	Production Technician II	$52,000.00
							Production	Production Technician II	$52,000.00

sample

male female

independent samples

$$H_0: \mu_m - \mu_f = 0$$

$$H_1: \mu_m - \mu_f \neq 0$$

We will use a hypothesis test for mean salary

The test we should use is the t-test for independent samples, var unknown but assumed equal

If the values in one sample reveal no information about the other, they are independent

Ready

365√DataScience

We are going to test if there is a significant difference in the salaries, based on gender

Practical example. Hypothesis testing.

Is there a gender bias in Spark Fortress

Overall	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	98	$ 65,736.91	$² 932,705,380.02	$² 1,025,188,119.29	1.34	0.182
Male	76	$ 72,300.53	$² 1,144,799,128.75			

$$H_0: \mu_m - \mu_f = 0$$

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

$$T = \frac{(\bar{x}-\bar{y})-(\mu_m-\mu_f)}{\sqrt{\frac{s_p^2}{n_m}+\frac{s_p^2}{n_f}}}$$

Degrees of freedom= 172

0.182 >> all common levels of significance

We cannot reject the null hypothesis =>
there isn't enough statistical evidence that there is a gender wage gap in this firm

Once we have surpassed 50 degrees of freedom, the Student's T ~ Normal distribution

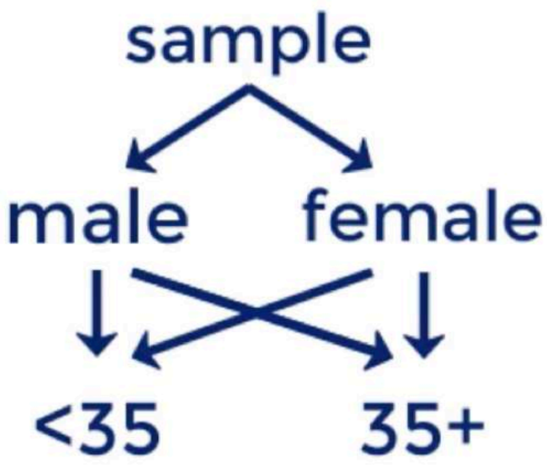
What if there is a hidden wage gap?

Practical example. Hypothesis testing.

Spark Fortress Inc. HR data

Surname	Name	Age	Gender	Country	Ethnicity	Start_date	Department	Position	Salary
Bold	Caroline	63	Female	United States	White	7/2/2012	Executive Office	President & CEO	$166,400.00
Foss	Jason	37	Male	United States	Black or African American	4/15/2011	IT/IS	IT Director	$135,200.00
Zamora	Jennifer	38	Female	United States	White	4/10/2010	IT/IS	CIO	$135,200.00
Dougall	Eric	47	Male	United States	White	1/5/2014	IT/IS	IT Manager - Support	$133,120.00
Monroe	Peter	31	Male	France	Hispanic	2/15/2012	IT/IS	IT Manager - Infra	$131,040.00
Roup	Simon	44	Male	United States	White	1/20/2013	IT/IS	IT Manager - DB	$128,960.00
Houlihan	Debra	51	Female	United States	White	5/5/2014	Sales	Director of Sales	$124,800.00
Bramante	Elisa	34	Female	United States	Black or African American	1/5/2009	Production	Director of Operations	$124,800.00
Del Bosque	Keyla	38	Female	United States	Black or African American	1/9/2012	Software Engineering	Software Engineer	$118,809.60
Onque	Jasmine	27	Female	United States	White	9/30/2013	Sales	Area Sales Manager	$118,560.00
Carabbio	Judith	30	Female	United States	White	11/11/2013	Software Engineering	Software Engineer	$116,480.00
Smith	John	33	Male	United States	Black or African American	5/18/2014	Sales	Sales Manager	$116,480.00
Digitale	Alfred	29	Male	United States					
Jeremy	Peter	43	Male	United States					
Martins	Joseph	47	Male	Israel					
Villanueva	Noah	28	Male	United States					
Martin	Sandra	30	Female	United States					
Friedman	Gerry	48	Male	United States	Two or more races	3/7/2011	Sales	Area Sales Manager	$115,440.00
Gonzales	Ricardo	63	Male	United States	White	5/12/2014	Sales	Area Sales Manager	$115,440.00
Warfield	Sarah	39	Female	United States	Asian	3/30/2015	IT/IS	Sr. Network Engineer	$114,816.00
LeBlanc	Brandon R	33	Male	United States	White	1/5/2016	Admin Offices	Shared Services Manager	$114,400.00
Butler	Webster L	34	Male	United States	White	1/28/2016	Production	Production Manager	$114,400.00
Liebig	Ketsia	36	Female	United States	White	9/30/2013	Production	Production Manager	$114,400.00
Sullivan	Kissy	39	Female	United States	Black or African American	1/8/2009	Production	Production Manager	$114,400.00
Buck	Edward	42	Male	United States	White	9/29/2014	Sales	Area Sales Manager	$114,400.00
Carter	Michelle	54	Female	United States	White	8/18/2014	Sales	Area Sales Manager	$114,400.00
									$114,400.00
									$114,400.00
									$114,400.00
									$114,400.00
									$114,400.00
									$114,400.00

When in doubt, always segment!



This may explain the egalitarian culture of the company...
and / or ... Caroline Bold's high salary biased the data

Practical example. Hypothesis testing.

Is there a gender bias in Spark Fortress

$$H_0: \mu_m - \mu_f = 0$$

$$H_1: \mu_m - \mu_f \neq 0$$

Overall	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	98	$ 65,736.91	$² 932,705,380.02	$² 1,025,188,119.29	1.34	0.182
Male	76	$ 72,300.53	$² 1,144,799,128.75			

Below35	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	46	$ 66,775.23	$² 1,063,144,850.86	$² 1,048,675,919.03	0.43	0.668
Male	37	$ 69,859.89	$² 1,030,589,754.25			

there is virtually no wage gap on a gender basis

Over 35	n	Mean	Sample variance	Pooled variance	T-score	p-value
Female	52	$ 63,448.22	$² 1,129,668,678.18	$² 1,210,601,529.12	2.00	0.048
Male	40	$ 78,049.60	$² 1,316,436,795.73			

at 5% significance, we reject the null hypothesis

a wage gap does exist for older employees

This was a two-sided test, so we are not sure who gets more money…

The T-score (2.00) > 0 => the difference in pay is positive (males earn more)

Overall, there is no wage gap in Spark Fortress (driven by younger employees)

Limitations (we omitted important factors):

1. Position
2. Ethnicity

… etc.

Formulae for Hypothesis Testing

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

365√DataScience