دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ یازدهم – (بخش اول)

آزمون فرض

Hypothesis Testing

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

About me

Mohammad Fozouni (Ph.D.)
Dep. of Math. & Stat.
Gonbad Kavous University

- fozouni@hotmail.com
- https://m-fozouni.ir
- http://profs.gonbad.ac.ir/fozouni/en

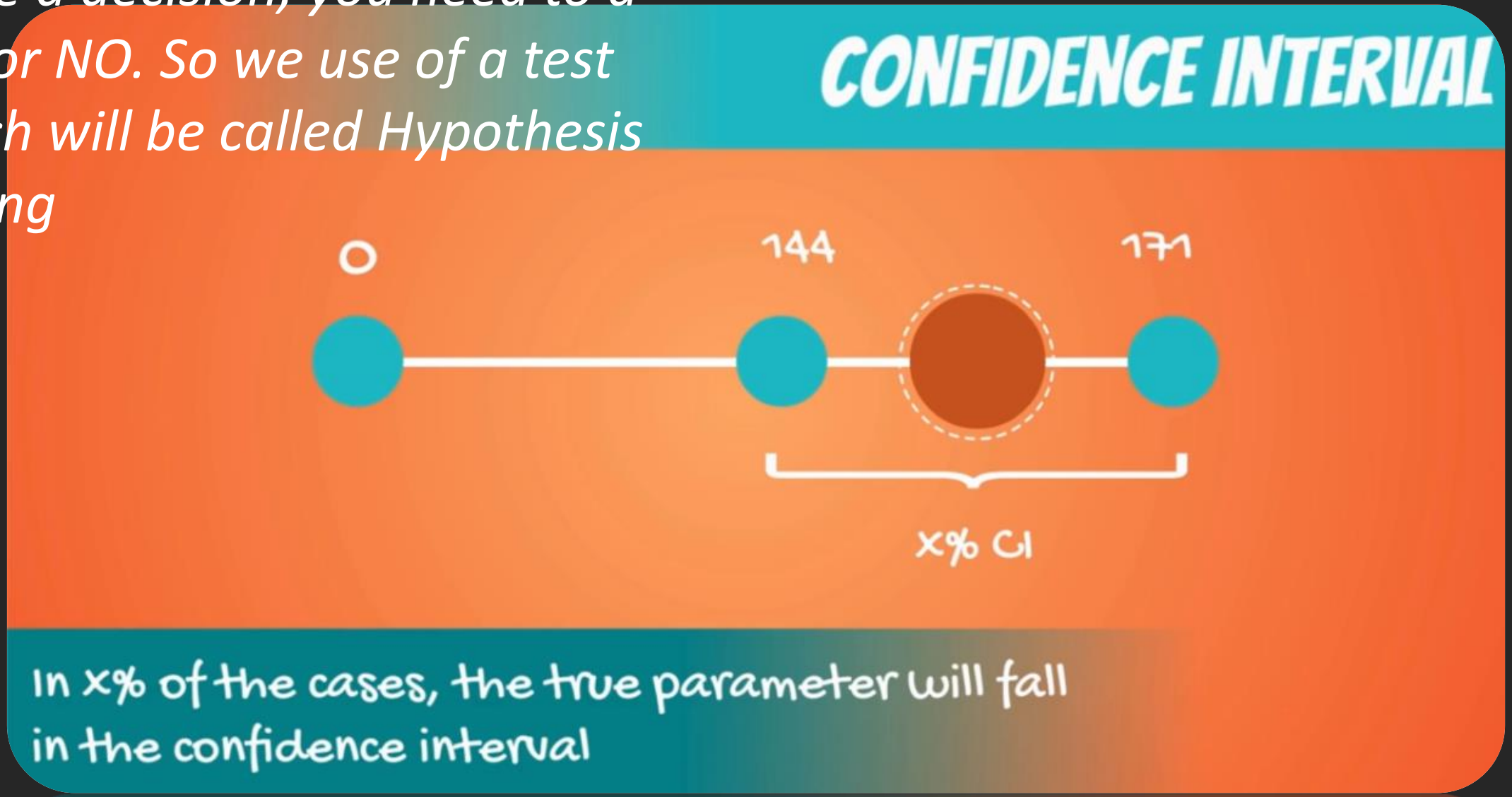
: elmedade

#data_science_fozouni

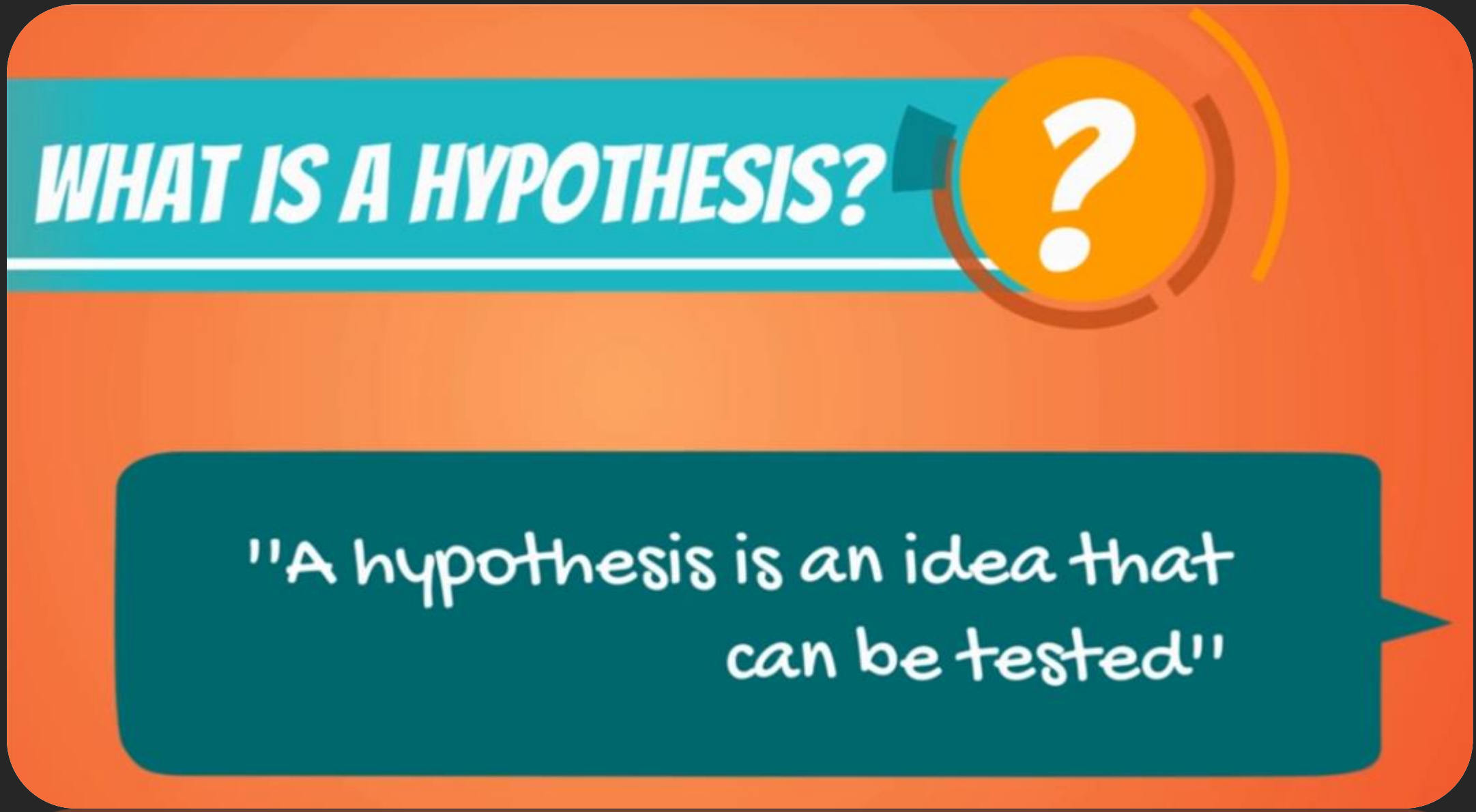
What is a Hypothesis Testing

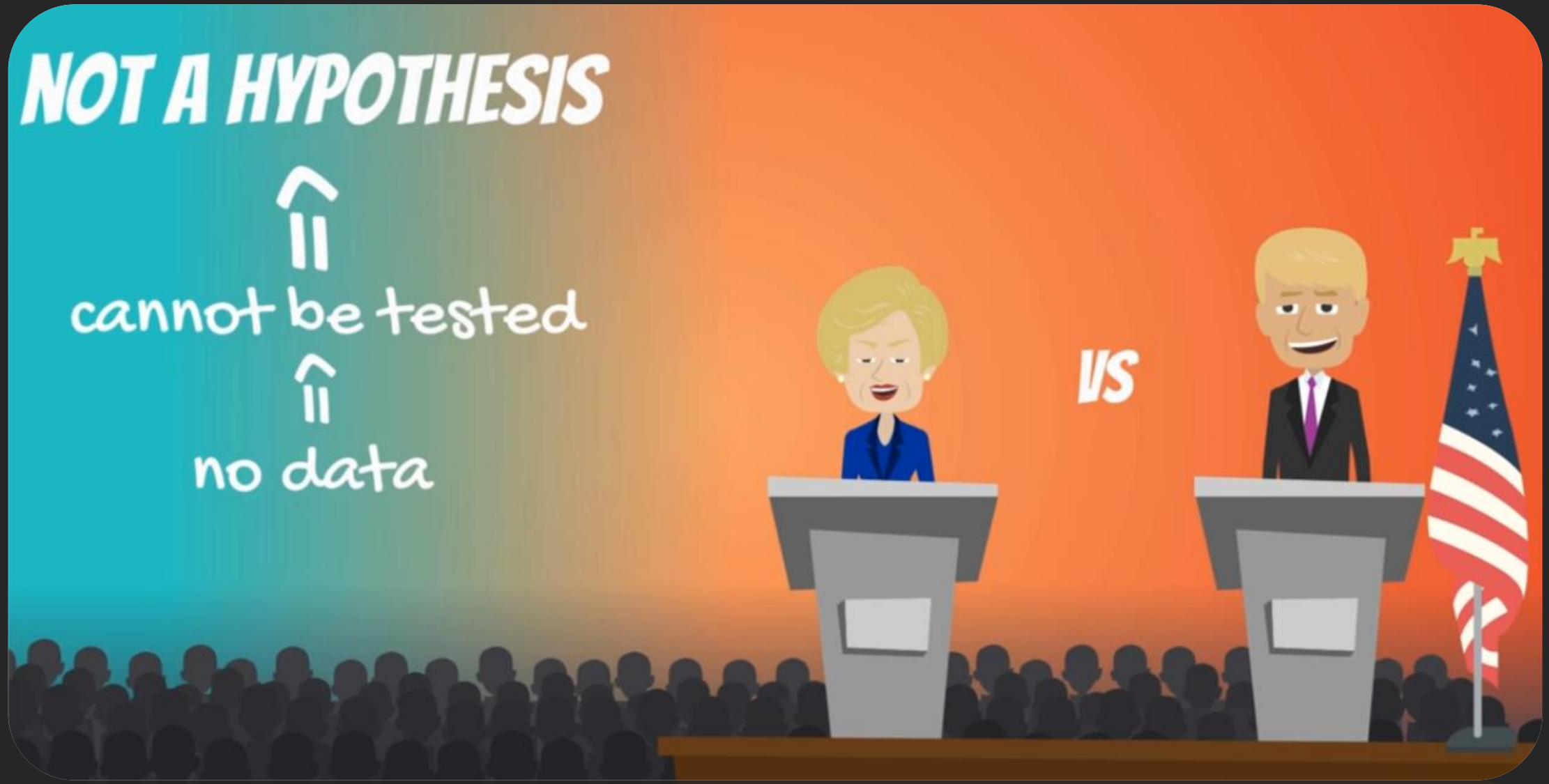

HYPOTHESIS TESTING
INFERENTIAL STATISTICS
DESCRIPTIVE STATISTICS

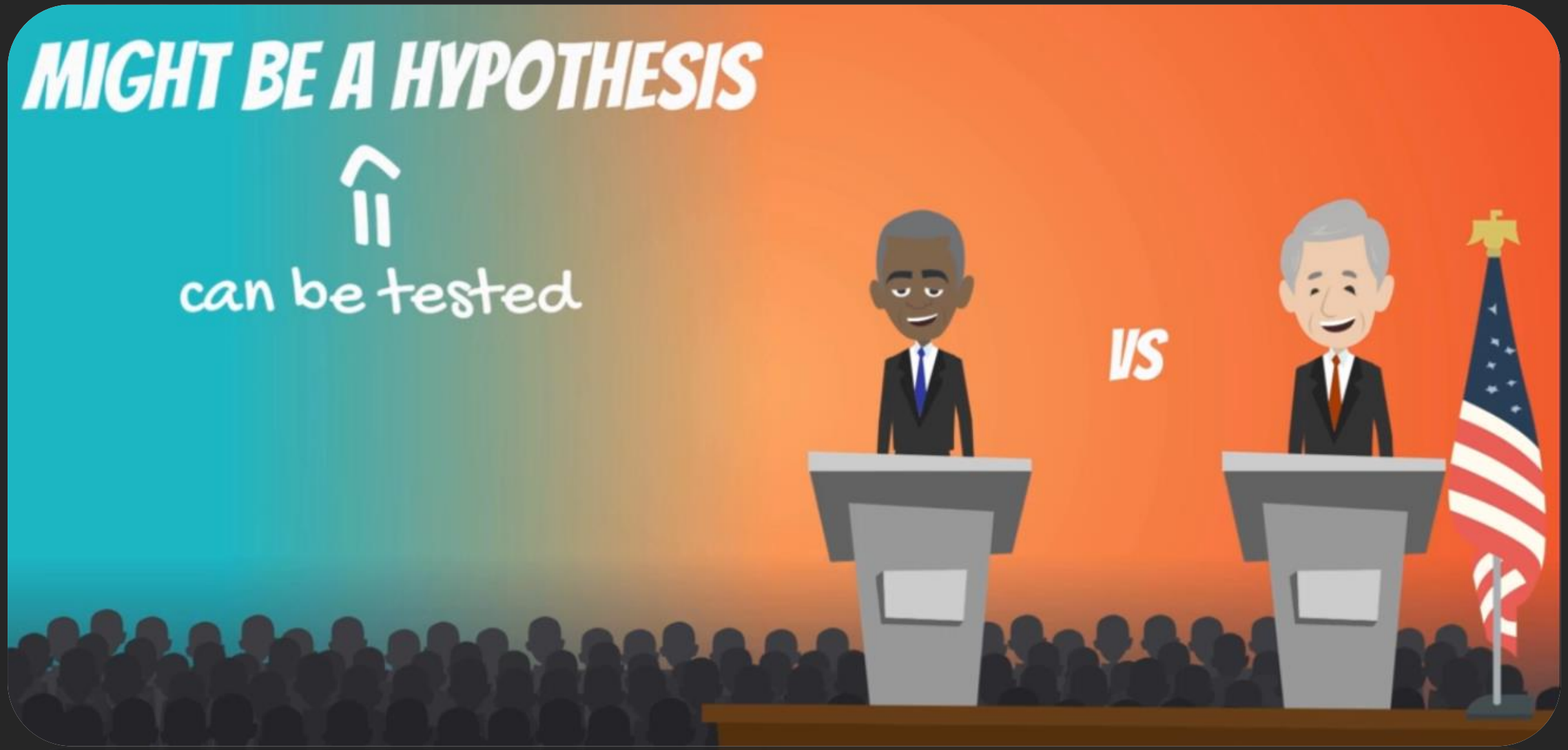
However when you are going to make a decision, you need to a YES or NO. So we use of a test which will be called Hypothesis testing

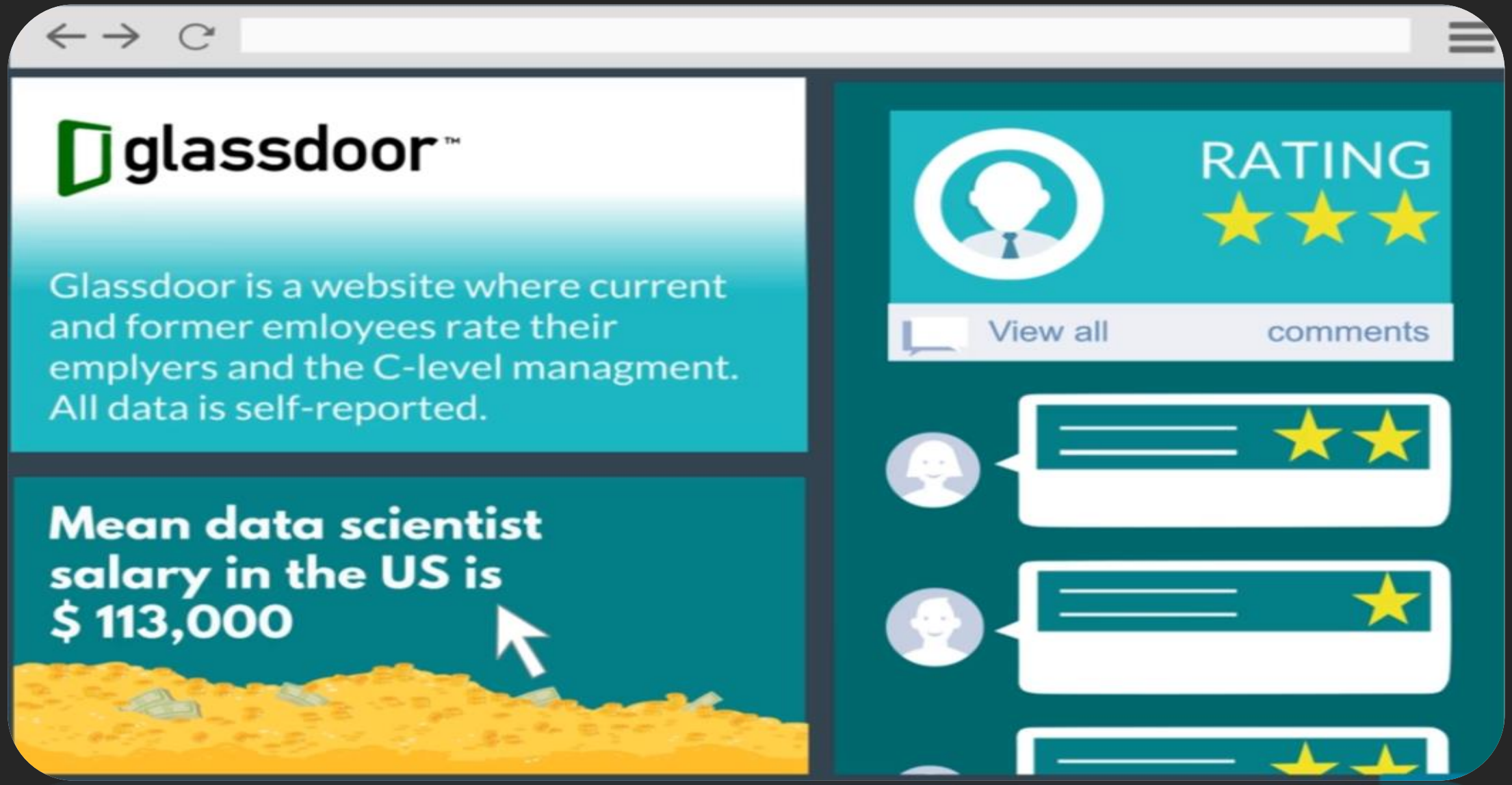



STEPS IN DATA-DRIVEN DECISION MAKING
1
2
3
4
Formulate a hypothesis
Find the right test
Execute the test
Make a decision


WHAT IS A HYPOTHESIS?
?
"A hypothesis is an idea that can be tested"

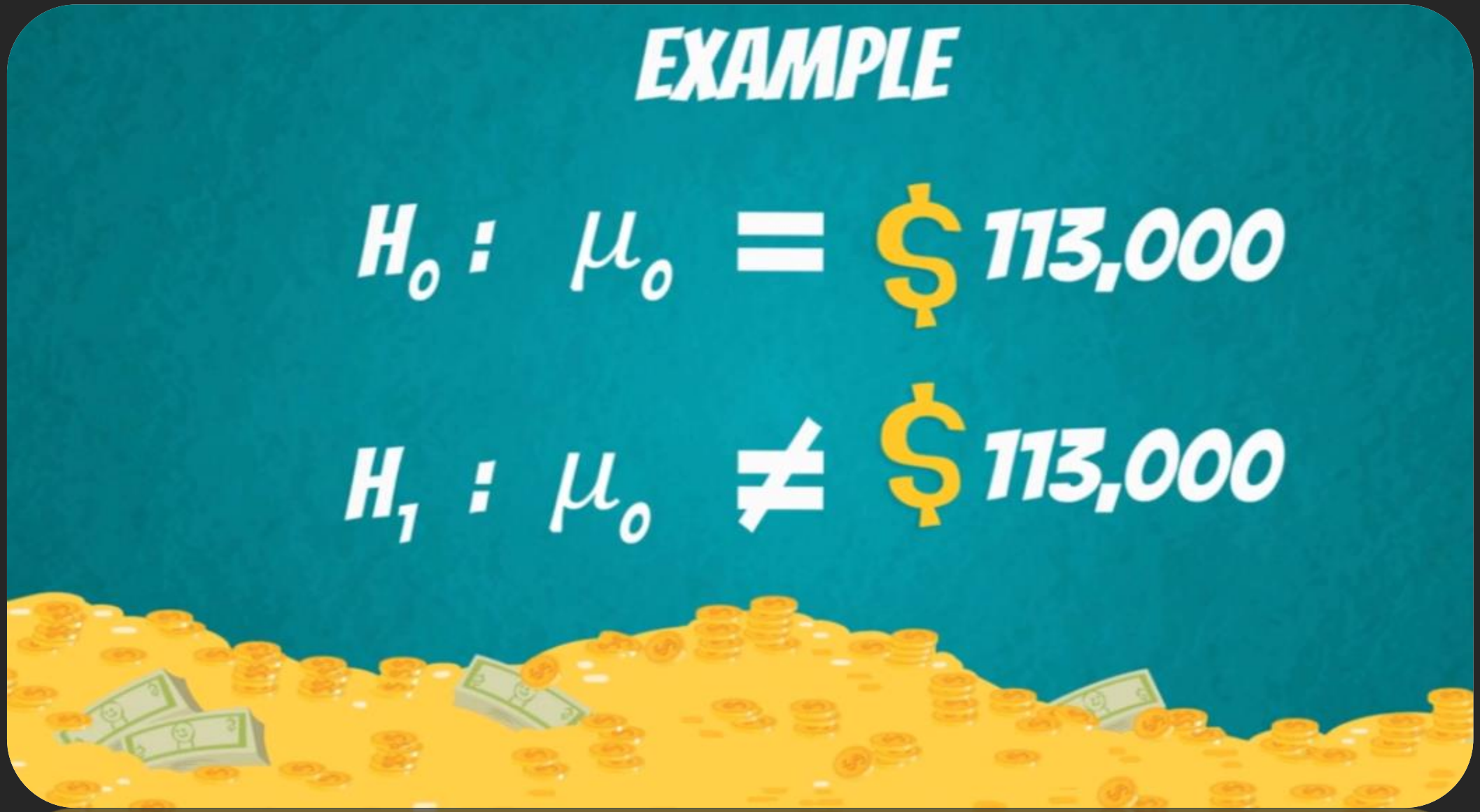

NOT A HYPOTHESIS
cannot be tested
no data
VS


MIGHT BE A HYPOTHESIS
can be tested
VS


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
RATING
View all comments

EXAMPLE

HYPOTHESES	NOTATION
Null hypothesis	H_0
Alternative hypothesis	H_1 or H_A


EXAMPLE
$H_0 : \mu_0 = \$113{,}000$
$H_1 : \mu_0 \neq \$113{,}000$

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


NULL HYPOTHESIS
CASE CLOSED
INNOCENT UNTIL PROVEN GUILTY
H_0 IS TRUE UNTIL REJECTED

EXAMPLE

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

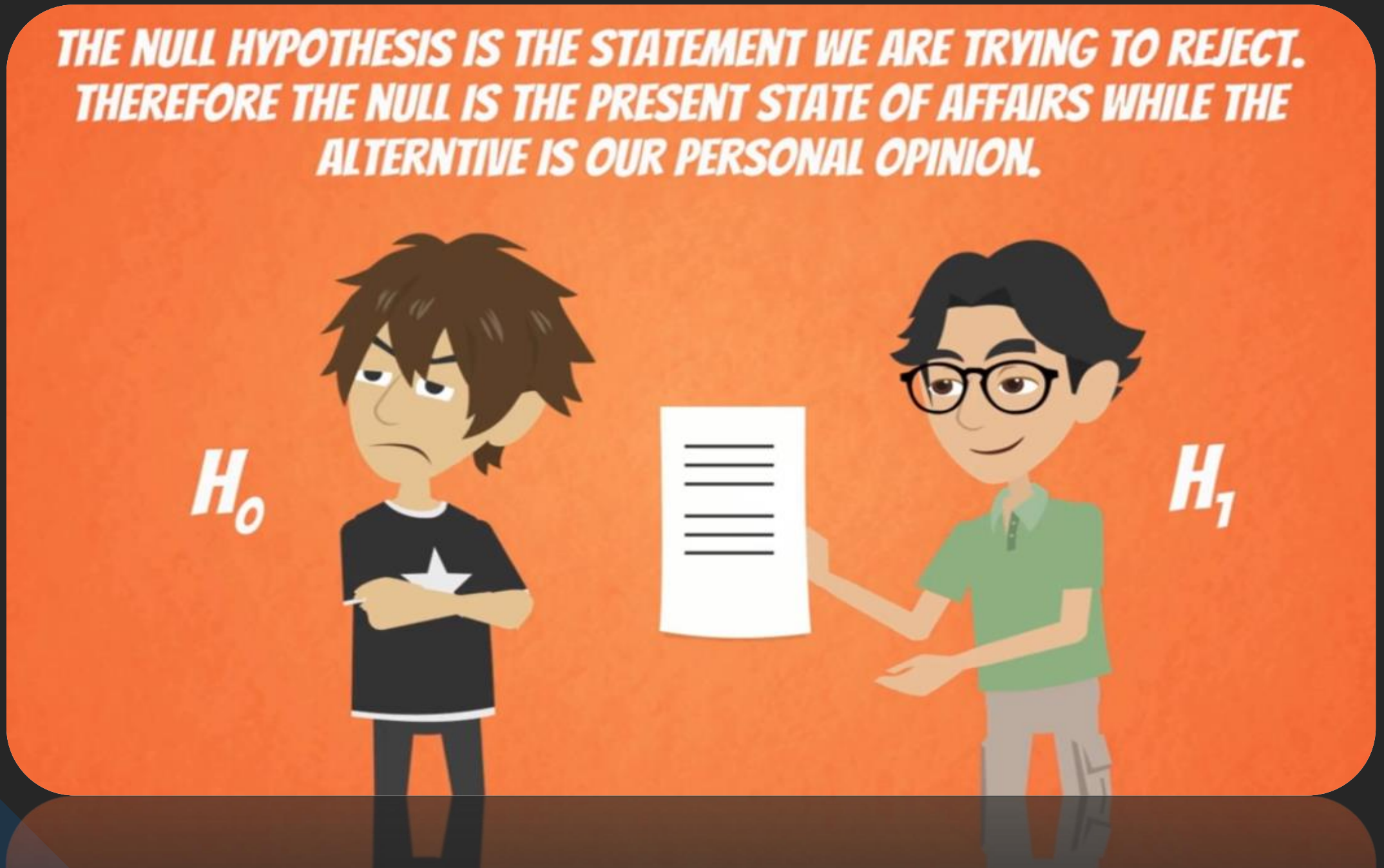
TWO-SIDED TEST
YOU CAN ALSO FORM
ONE-SIDED TESTS


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


$H_0 : \mu_0 \geq \$125,000$
$H_1 : \mu_0 < \$125,000$


$H_0 : \mu_0 \geq \$ 125,000$
$H_1 : \mu_0 < \$ 125,000$
OUTCOMES OF TESTS REFER TO POPULATION PARAMETER RATHER THAN SAMPLE STATISTIC


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


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.
H_0
H_1

Significance Level

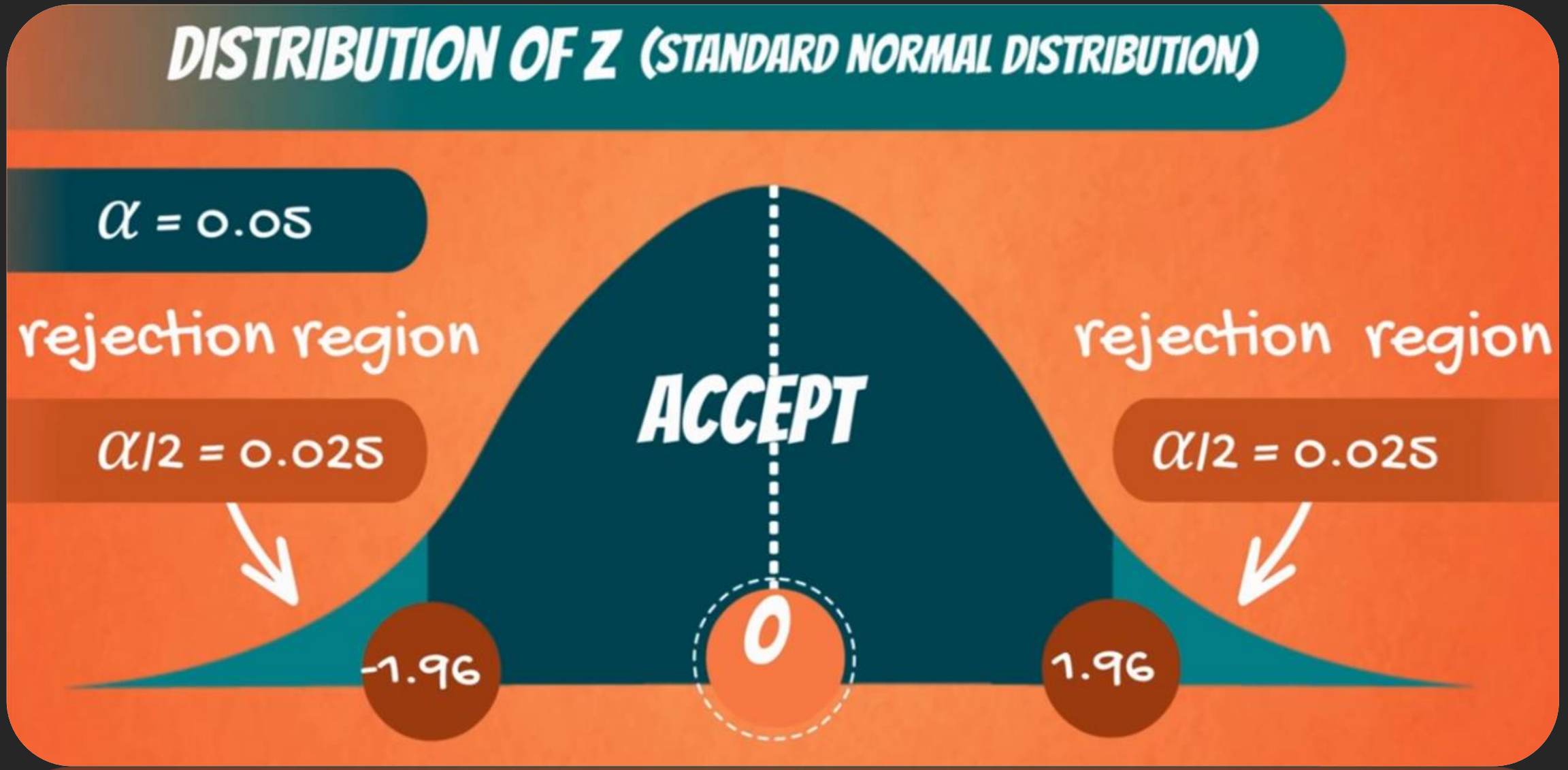

SIGNIFICANCE LEVEL
H_0
error
H_1

SIGNIFICANCE LEVEL

α ?

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

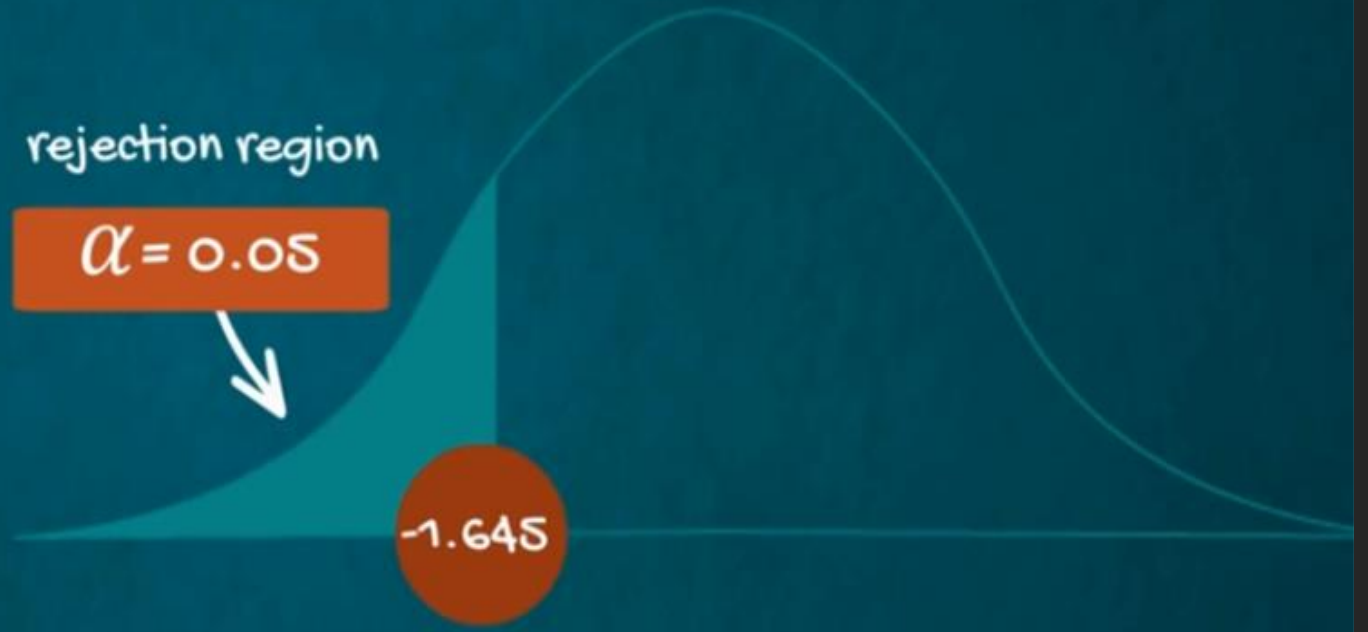
Typical values for alpha are:
0.01, 0.05, 0.1


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

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

ONE-SIDED TEST

$H_0 : \mu_0 \leq 70\%$

$H_1 : \mu_0 > 70\%$

If the test statistic is bigger than the cut-off 3-score, we would reject the null, otherwise we wouldn't.

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ یازدهم – (بخش دوم)

خطاها و p–مقدار

Errors and p-value

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

Errors in Hypothesis Testing


I
Reject a true null hypothesis
False positive
probability: α

II

sample size
n

variance
σ

Accept a false null hypothesis

False negative

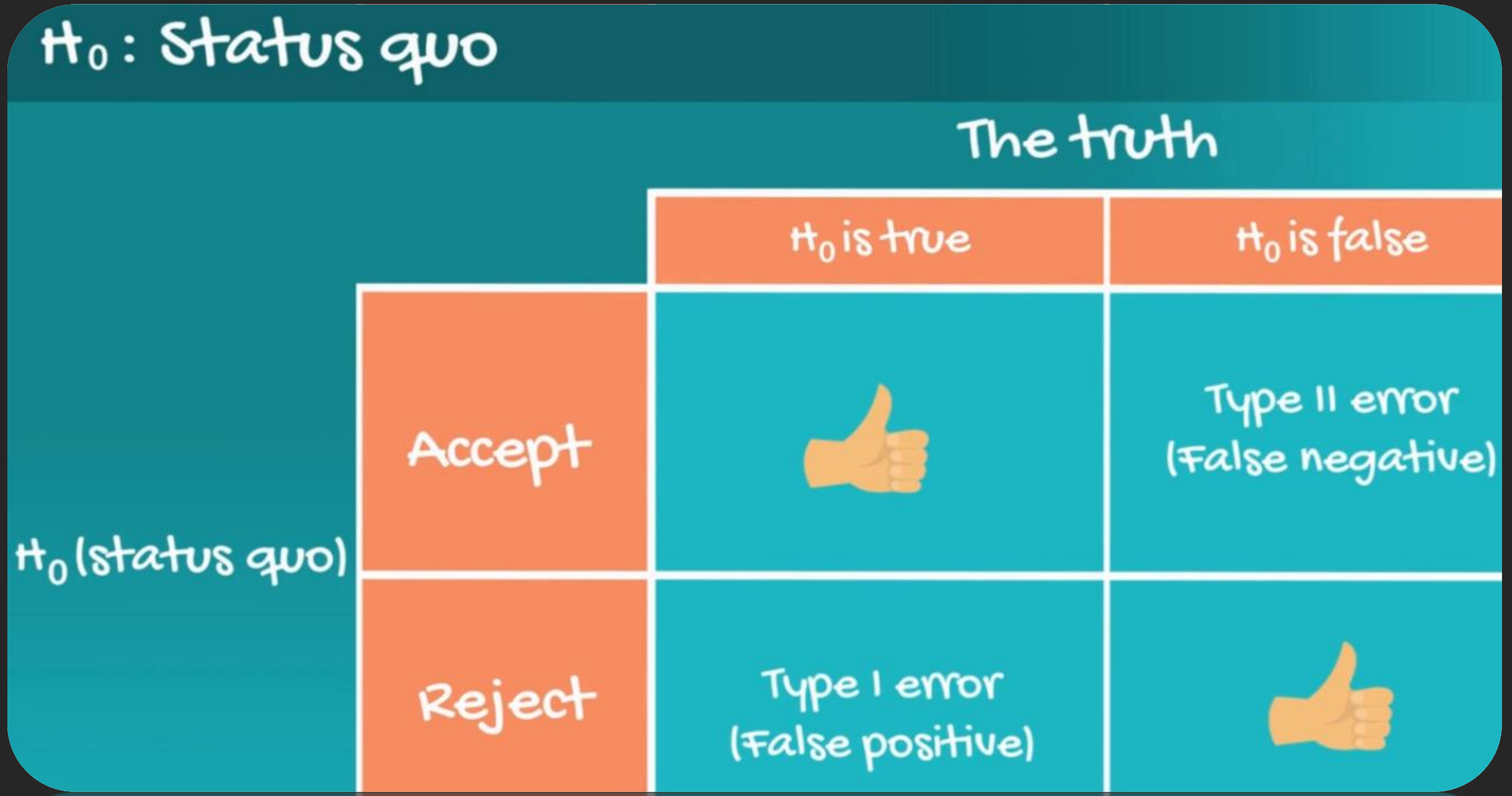
Probability: β

Goal of hypothesis testing

Rejecting a false null hypothesis

Probability:1-β

a.k.a. power of the test



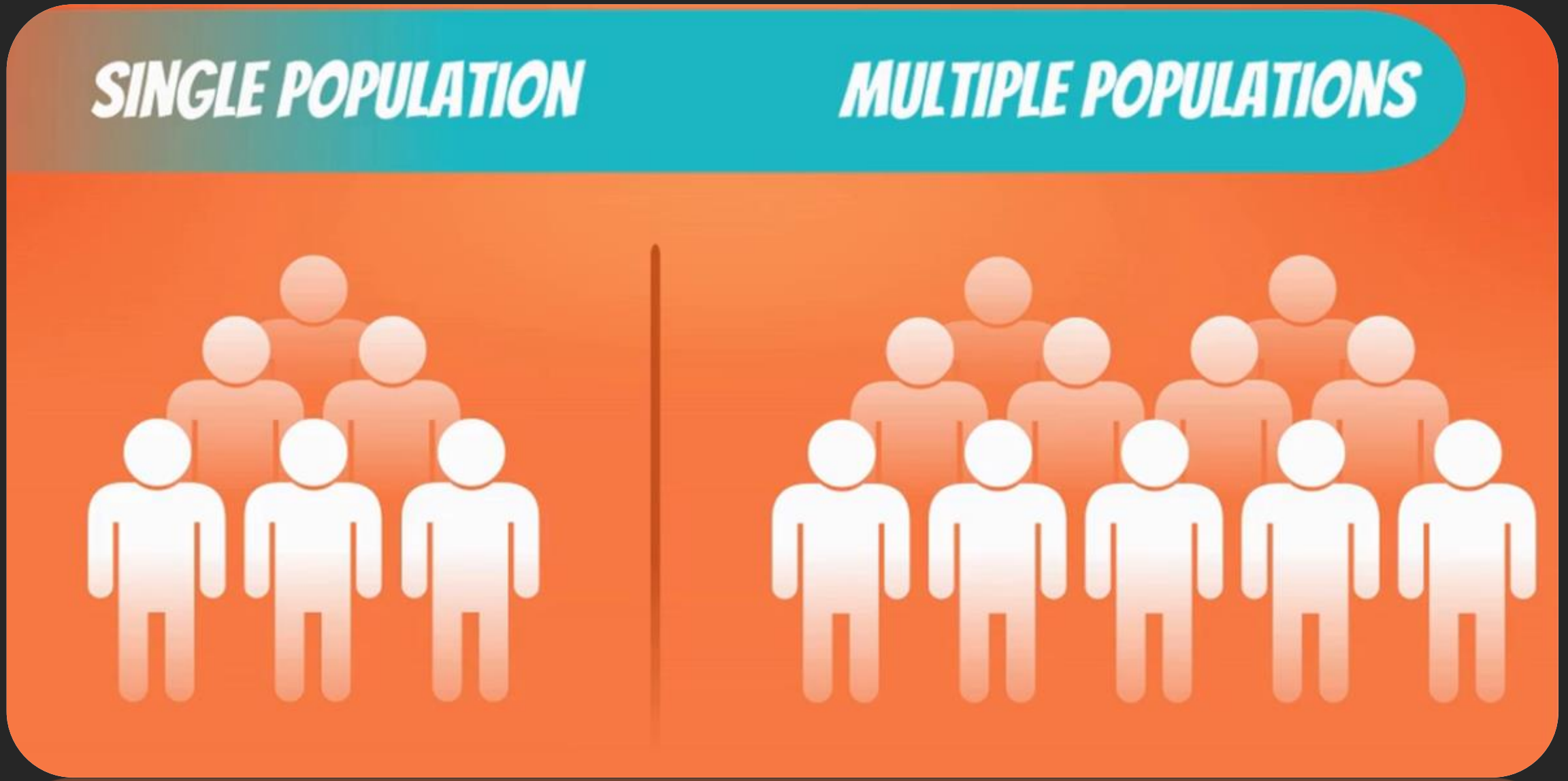
H_0 : She doesn't like you

H_0 (status quo)	The truth: She doesn't like you	The truth: She likes you
She doesn't like you (you shouldn't invite her out)	Accept (Do nothing)	
	Reject (Invite her)	

H_0 : Status quo

H_0 (status quo)	The truth: H_0 is true	The truth: H_0 is false
Accept		Type II error (False negative) missed your chance
Reject	Type I error (False positive) wrongly invited her	

Time to see some examples

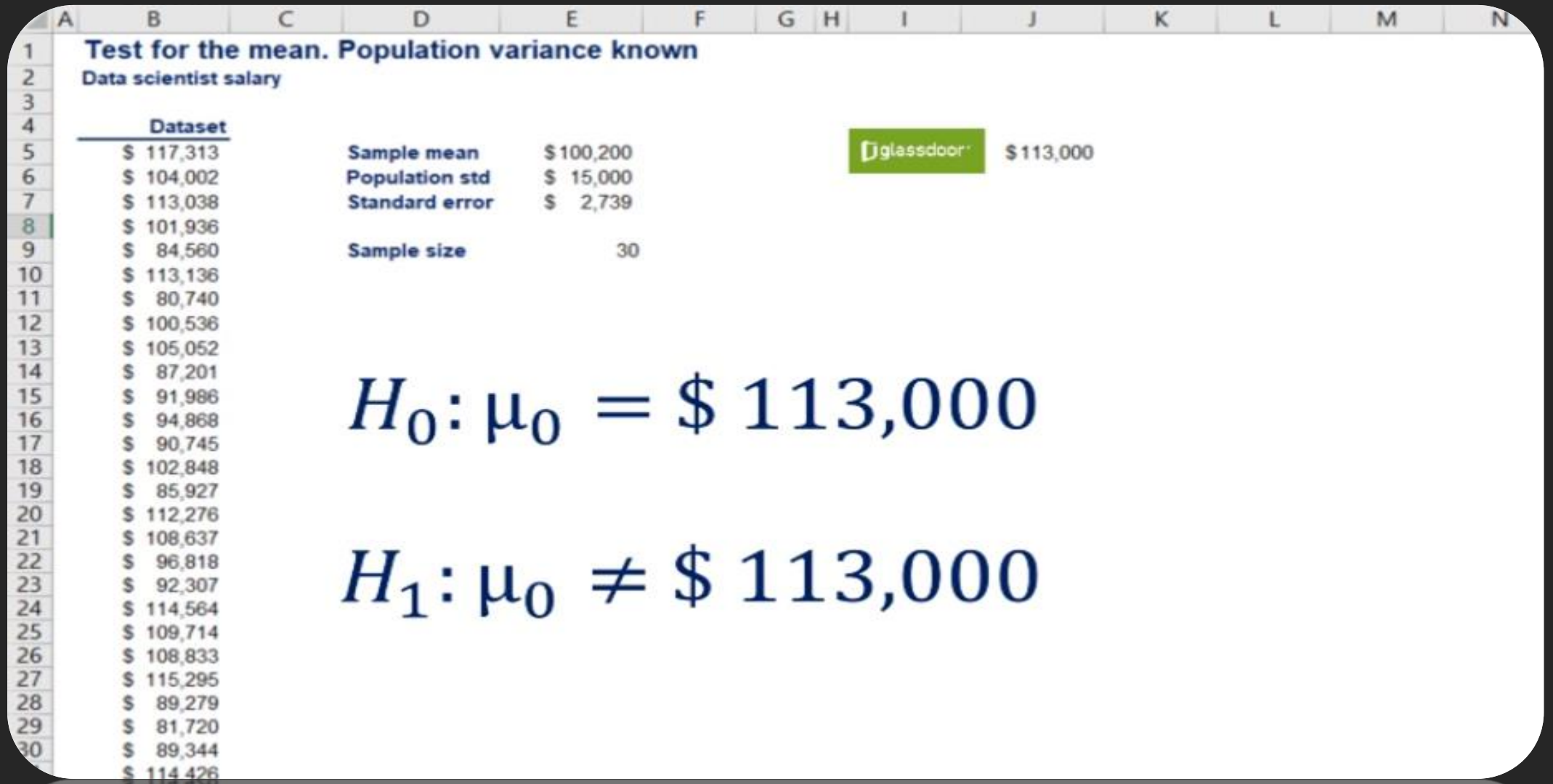

SINGLE POPULATION
MULTIPLE POPULATIONS

SINGLE POPULATION

Z
known

T
unknown

Re-visit the Data Scientist Salary problem



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

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

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

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

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

glassdoor $113,000

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

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
						0.9989	0.9989	0.9989	0.9990	0.9990

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'

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

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

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

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)

$$Z \qquad z$$

$$4.67 > 1.96$$ => we reject the null hypothesis

At 5% significance level we have rejected the null hypothesis

p-value


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

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

$Z = -4.67$

Rule: You should reject the null hypothesis, if

$$\text{P-VALUE} < \alpha$$

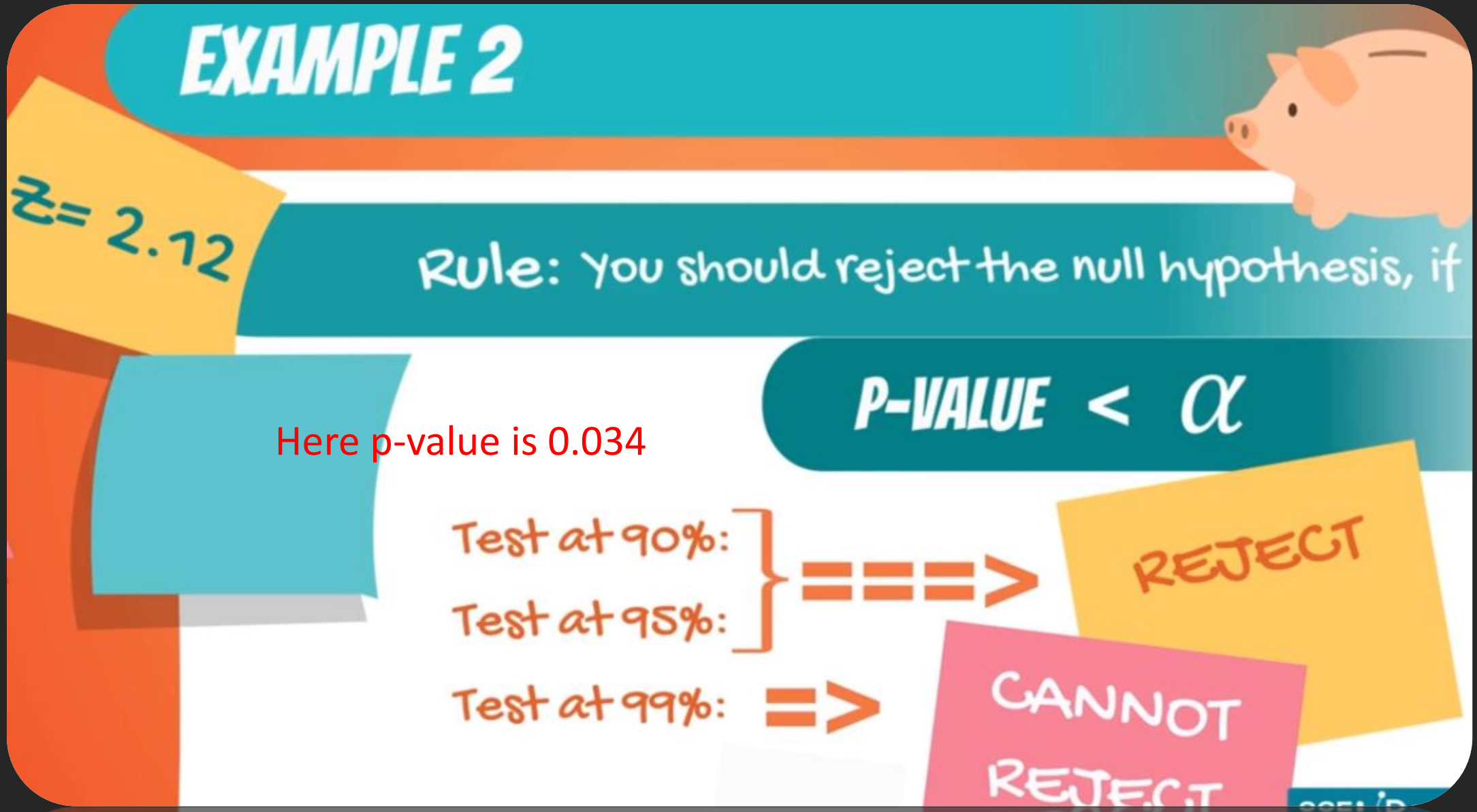
P-value = 0.0001

Test at 90%: $0.0001 < 0.1$

Test at 95%: $0.0001 < 0.05$ $\Rightarrow$

Test at 99%: $0.0001 < 0.01$

REJECT THE NULL HYPOTHESIS

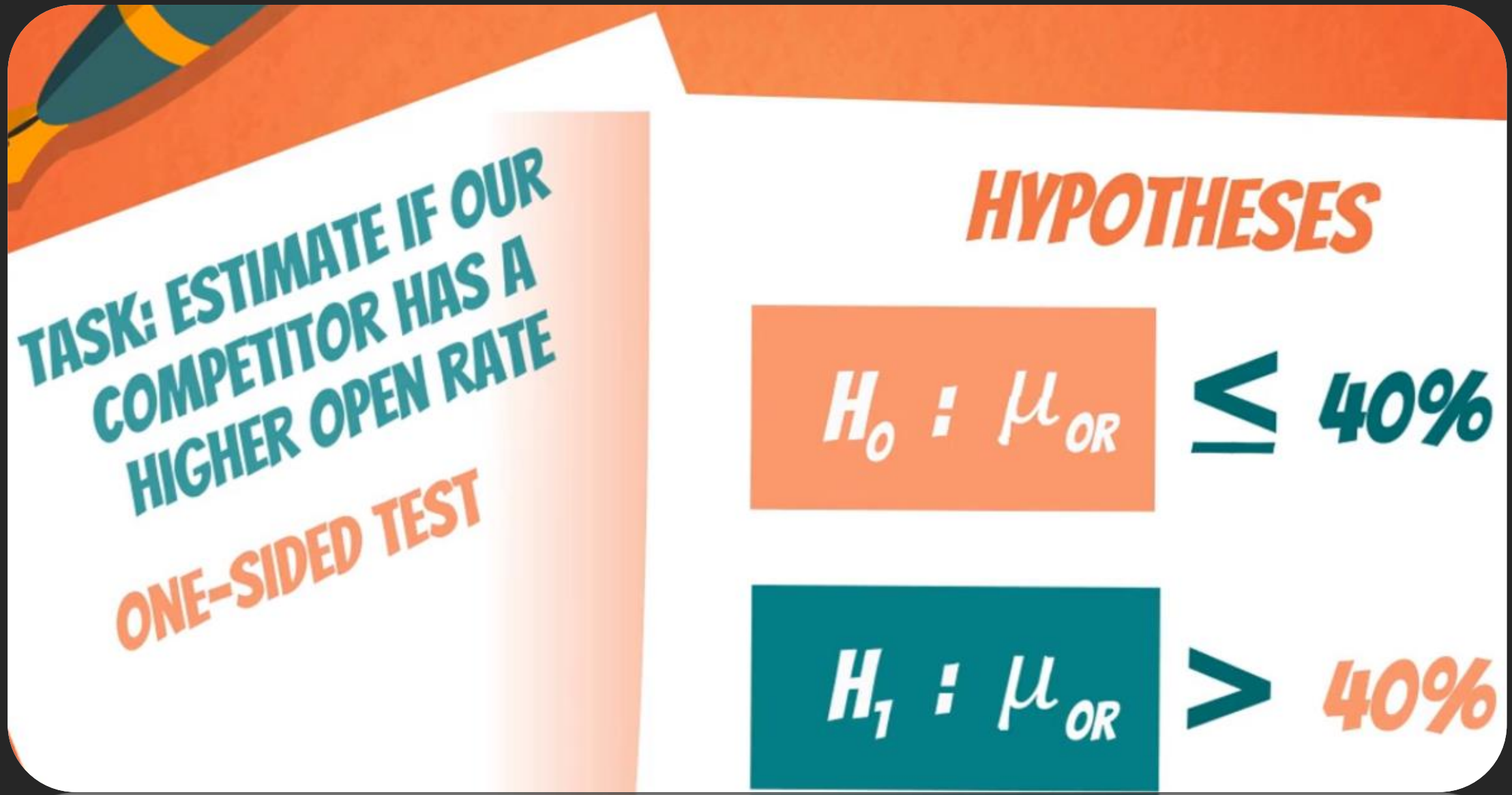

EXAMPLE 2
z= 2.12
Rule: You should reject the null hypothesis, if
P-VALUE < α
Here p-value is 0.034
Test at 90%:
Test at 95%:
===>
REJECT
Test at 99%:
=>
CANNOT
REJECT

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


TASK: ESTIMATE IF OUR COMPETITOR HAS A HIGHER OPEN RATE
ONE-SIDED TEST
HYPOTHESES
$H_0 : \mu_{OR} \leq 40\%$
$H_1 : \mu_{OR} > 40\%$

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%

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

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

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

T t

0.53 < 1.83 => we accept the null hypothesis

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

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

Our decision based on the p-value

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

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

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ یازدهم- (بخش سوم)

آزمون برای چند جامعه

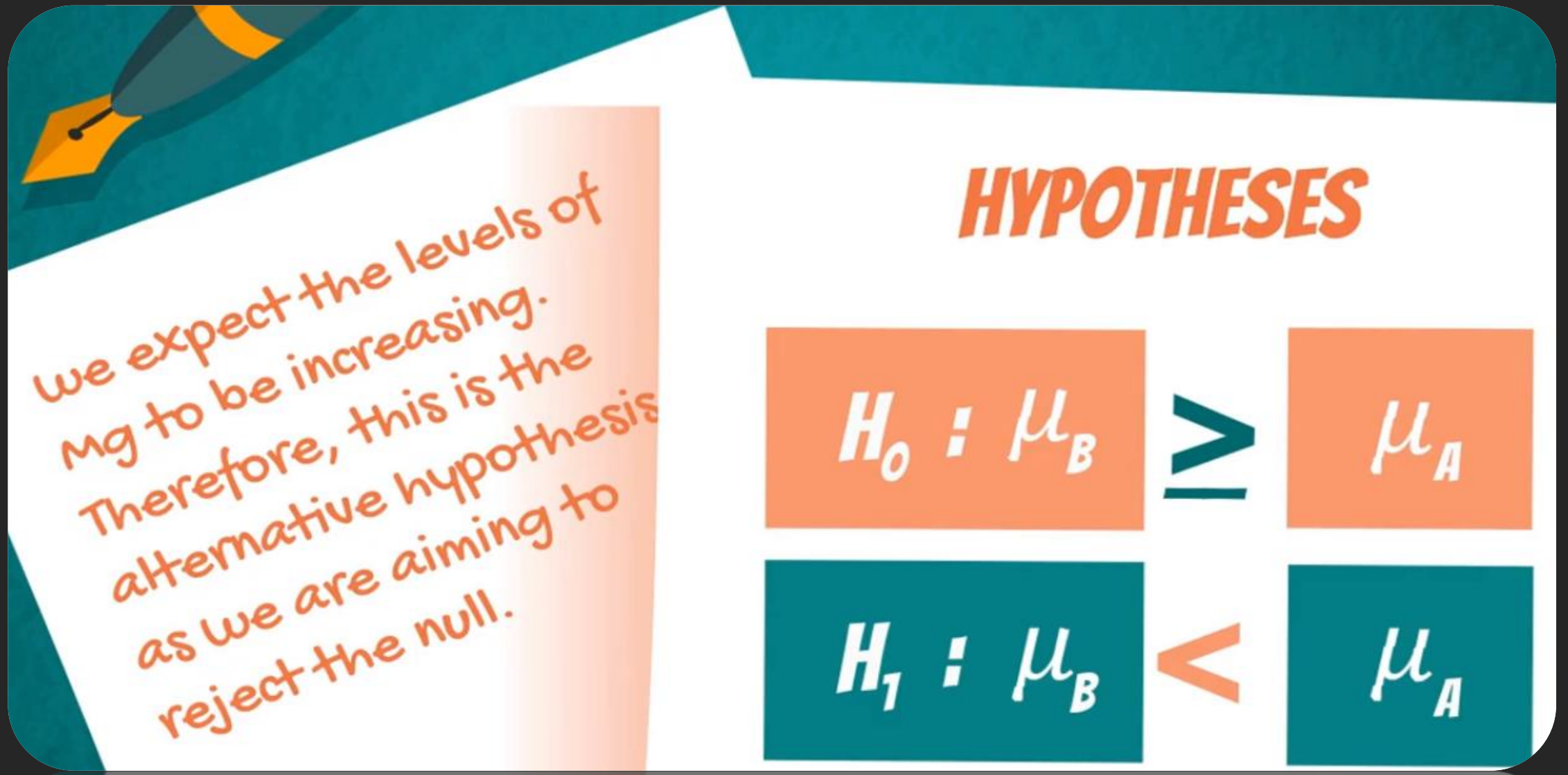
Test for Multiple Populations

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

MULTIPLE POPULATIONS

MULTIPLE POPULATIONS

DEPENDENT SAMPLES


we expect the levels of Mg to be increasing. Therefore, this is the alternative hypothesis as we are aiming to reject the null.
HYPOTHESES
$H_0 : \mu_B \geq \mu_A$
$H_1 : \mu_B < \mu_A$

Test the mean. Dependent Samples

Magnesium levels example

$$H_0: D_0 \geq 0$$

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

1. Small sample

2. We assume normal distribution of the population

=> t-statistic

3. Variance unknown

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

Test the mean. Dependent Samples

Magnesium levels example

$$H_0: D_0 \geq 0$$

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

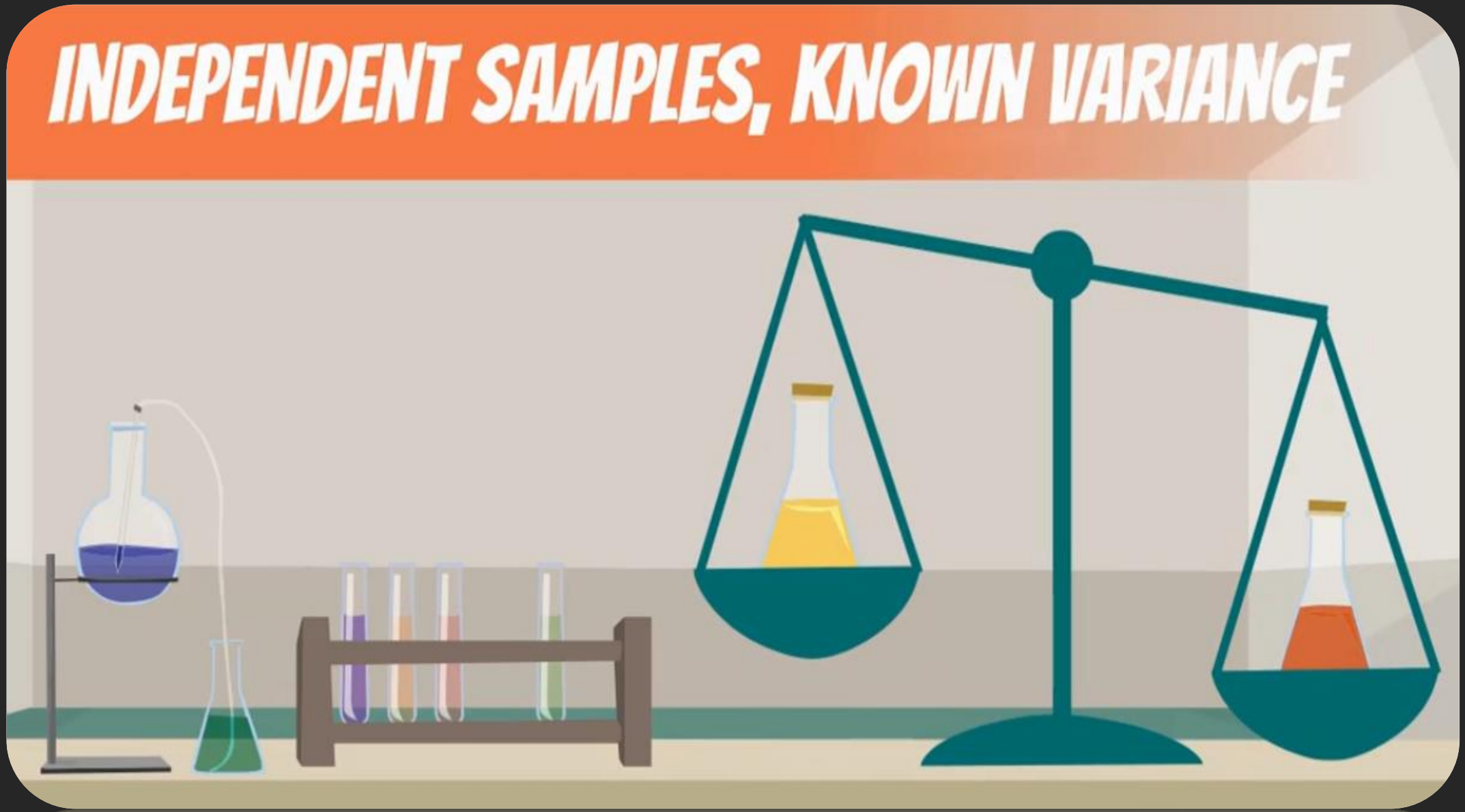
5% significance $0.024 < 0.05$ => reject the null hypothesis

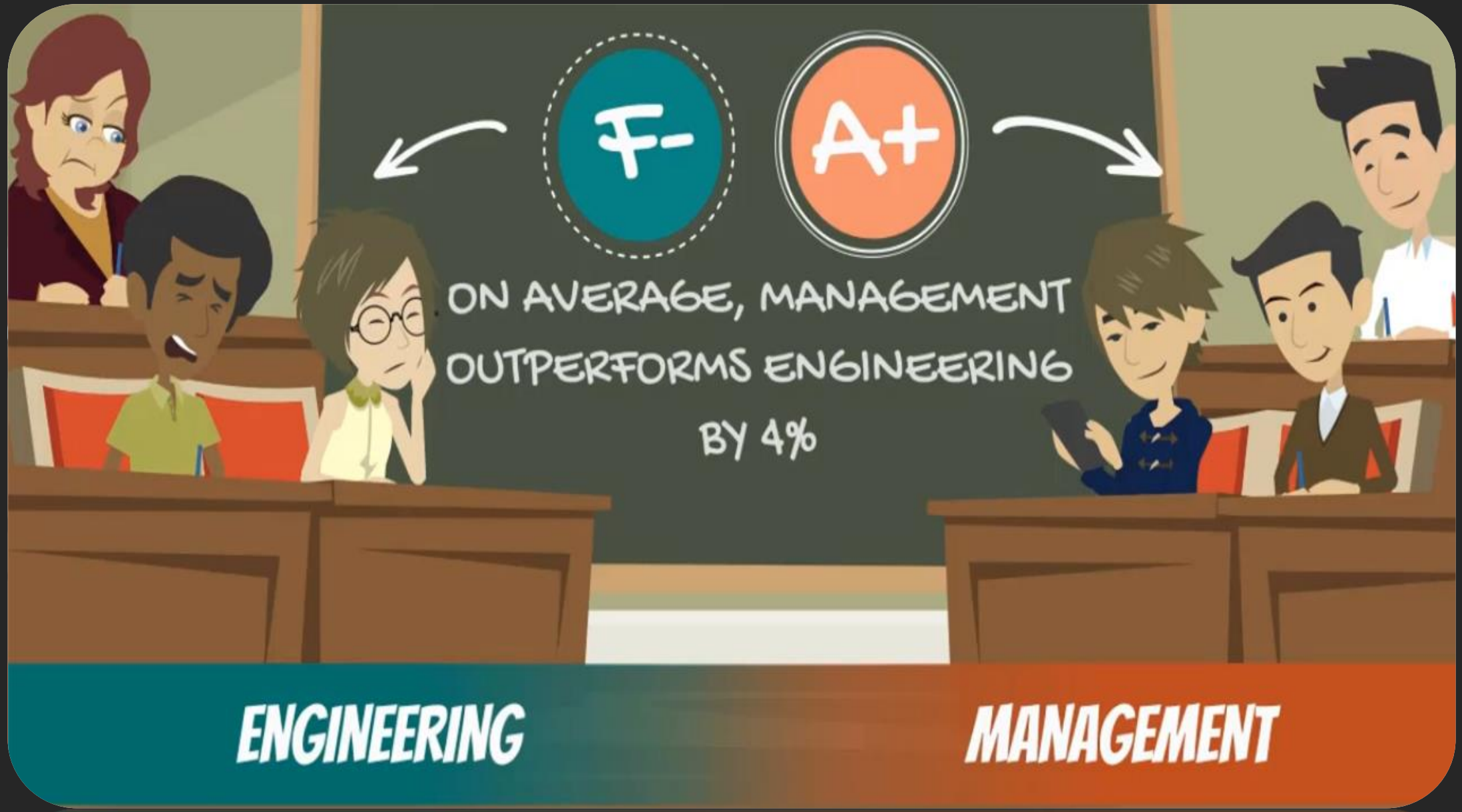
1% significance $0.024 > 0.01$ => accept the null hypothesis

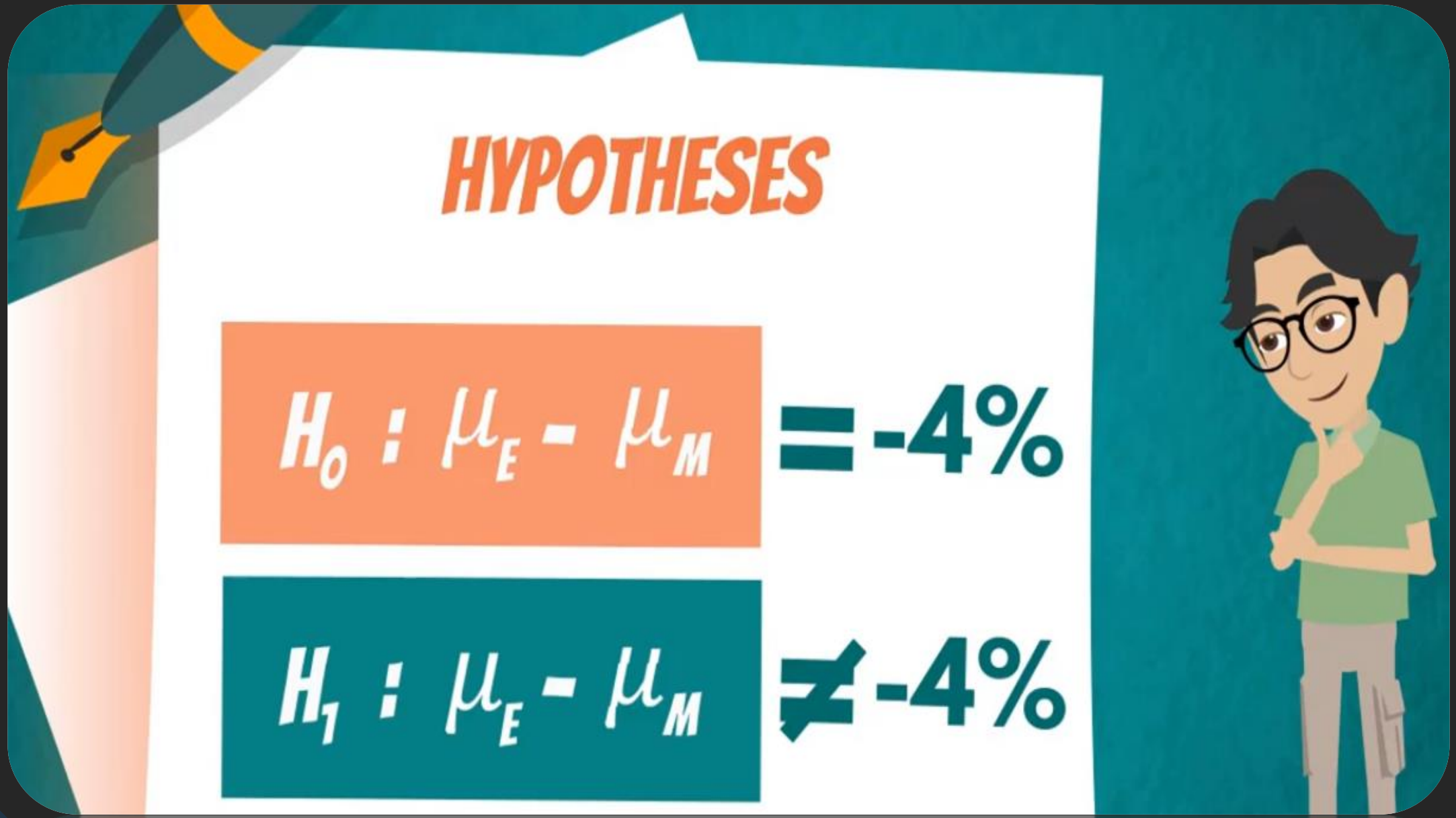
Decision rule:
Accept if: $p > \alpha$
Reject if: $p < \alpha$

Magnesium pill:

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


INDEPENDENT SAMPLES, KNOWN VARIANCE


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


HYPOTHESES
$H_0 : \mu_E - \mu_M = -4\%$
$H_1 : \mu_E - \mu_M \neq -4\%$

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%		

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

Standard Error

z-statistic	t-statistic
Big samples	Small samples
Known variances	Unknown variances

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		

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}} = \frac{(-7\%) - (-4\%)}{1.23\%} = -2.44$$

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		

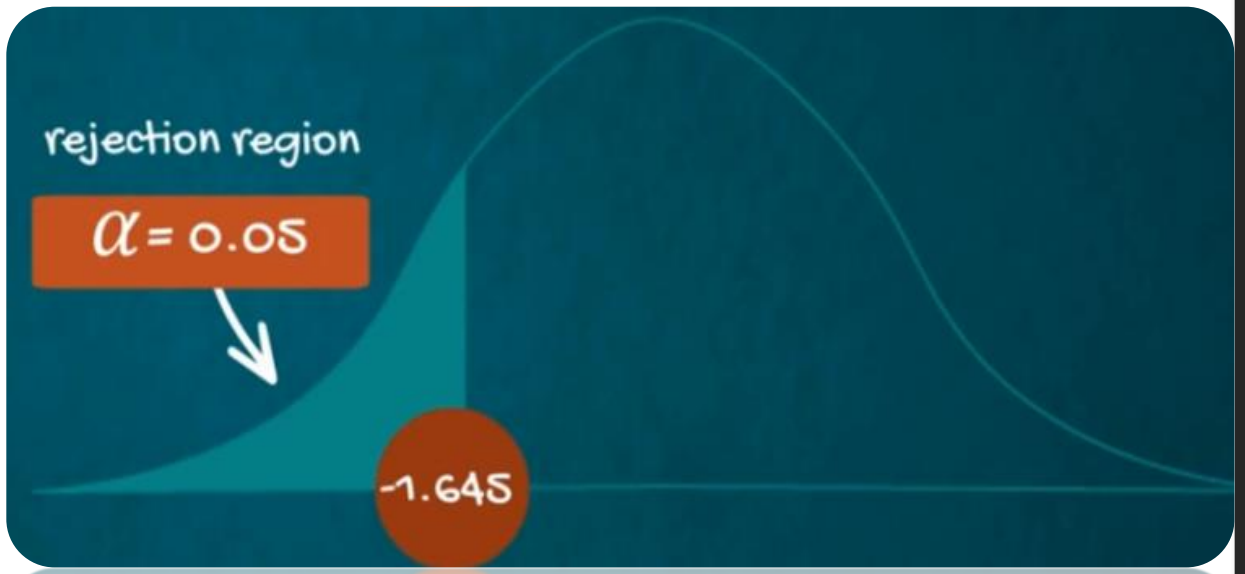
0.015 < 0.05 => we reject the null

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

Test for two means. Independent samples, variance known

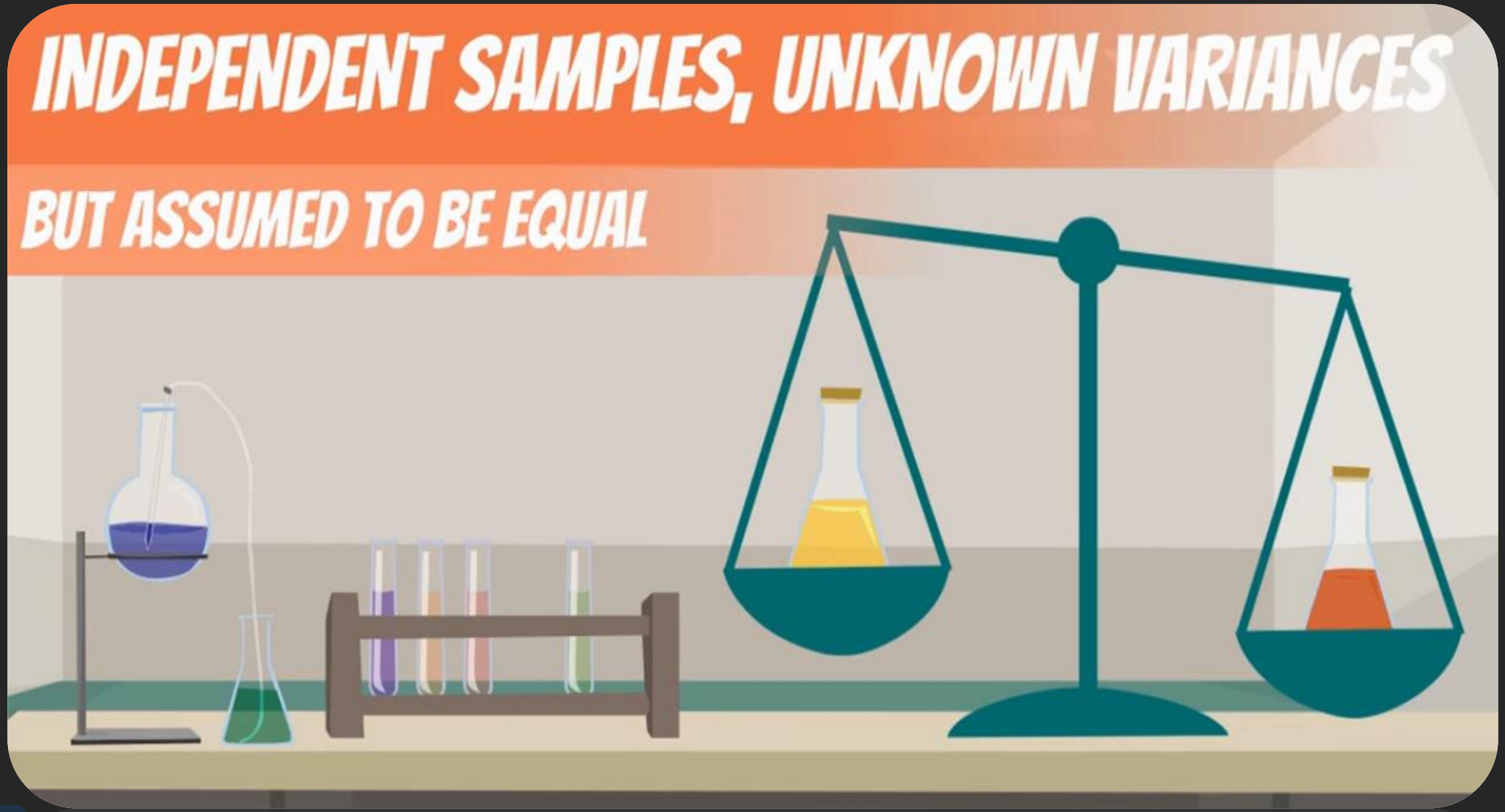
University example

	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		

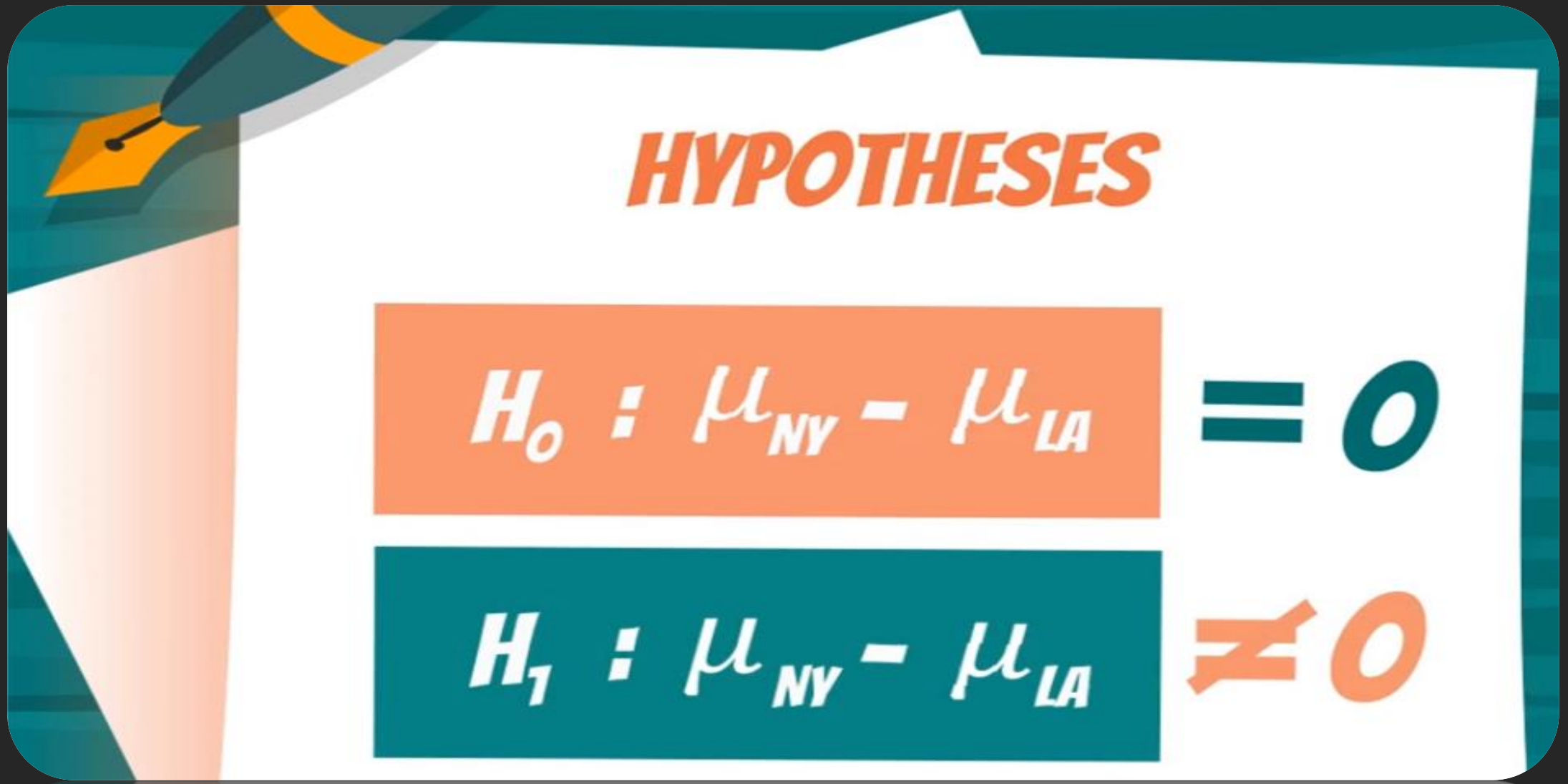


Is the difference higher or lower than 4%?

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


INDEPENDENT SAMPLES, UNKNOWN VARIANCES
BUT ASSUMED TO BE EQUAL

iPhone Price



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

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

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-statistic

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

Small samples

Unknown variances

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

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

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

Burn Calories and Solve the Following Problem

فرض کنید اطلاعات میزان کلیک افراد در ۲۴ دوشنبه و ۲۱ شنبه را طی چند سال روی پاپ-آپ‌های یک پلتفرم آموزشی در اختیار دارید. همچنین فرض کنید که نمونه‌ها را بطور مجزا استخراج کرده‌اید. آیا به اندازه‌ی کافی اطلاعات آماری دارید که نتیجه بگیرید تعداد کلیک‌ها در دوشنبه‌ها نسبت به شنبه‌ها بیشتر است؟

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

Shopping example

Background	You have data on the amount of times people click on a pop-up add on 24 Mondays and 21 Saturdays on an e-learning platform for several years. The samples are drawn independently.
Task	Statistically speaking, is there strong evidence that the number of clicks the add records on Mondays is higher than the number of clicks on Saturdays?

Solution:

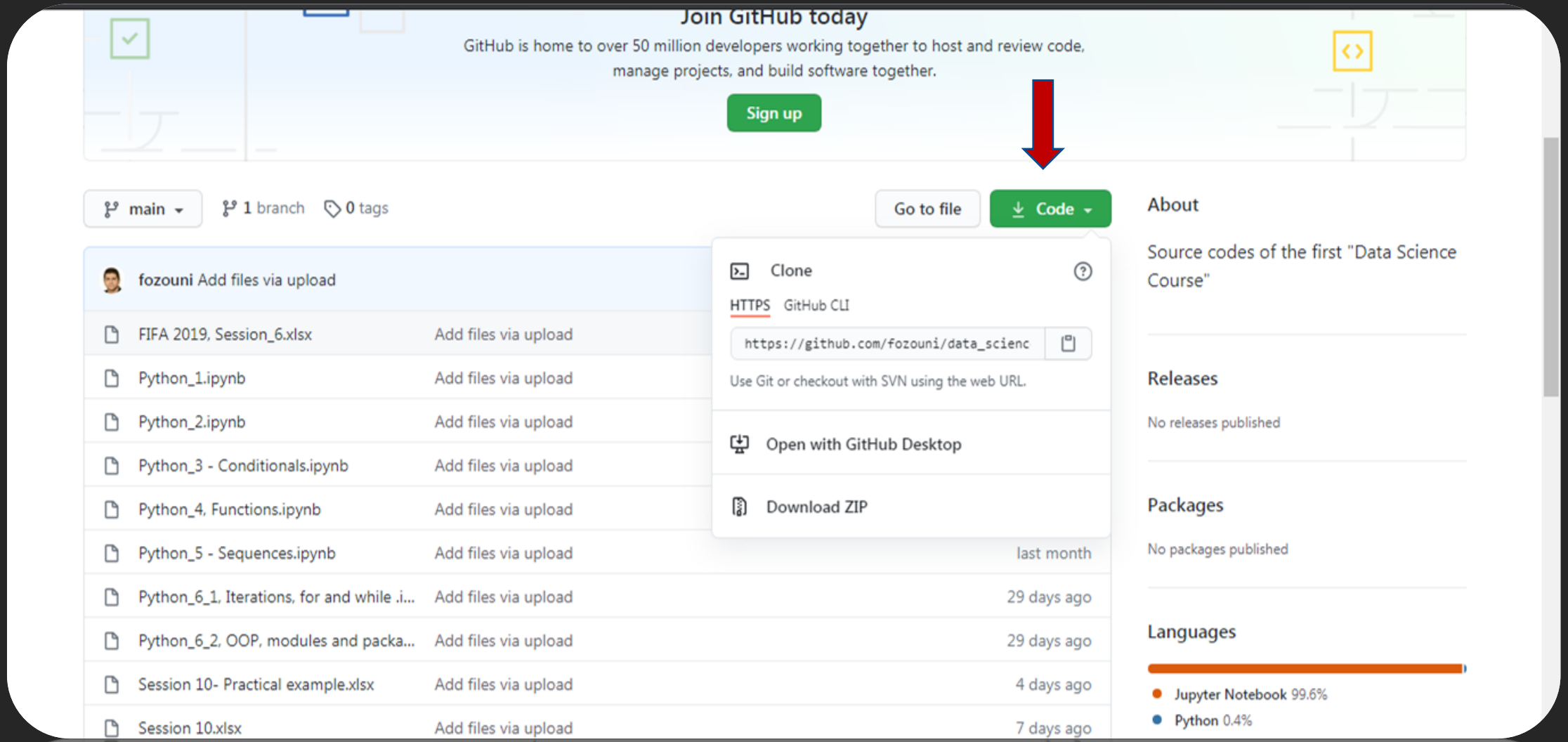
	Monday	Saturday
Mean	1,078.00	908.20
Std. deviation	633.00	469.80
Sample size	24	21
Sample Variances	400,689.00	220,712.04
Pooled Variance	316,978.79	
T-score	1.01	
p-value	0.16	

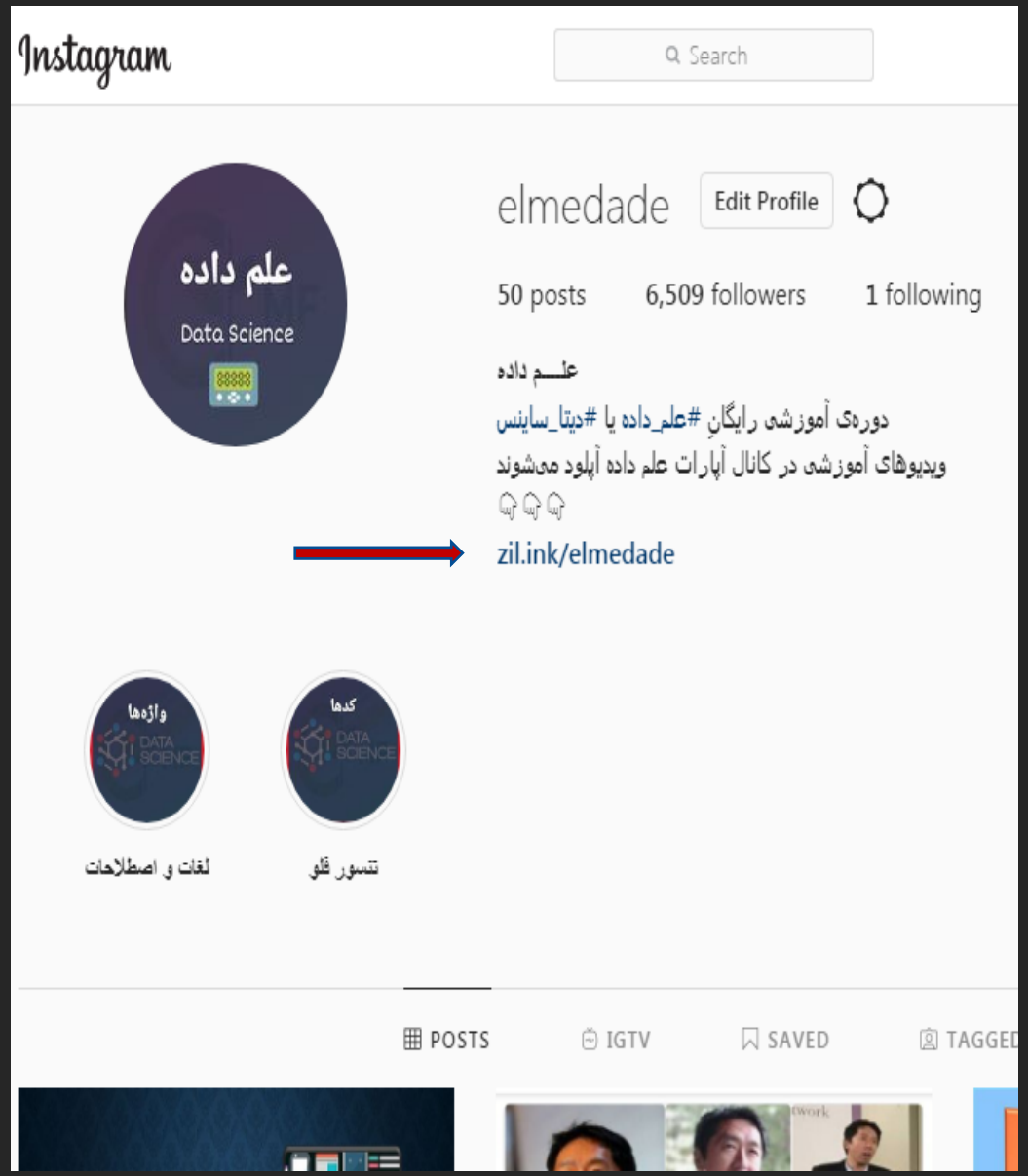
Steps to complete the task

1. Find the Sample Variances
2. State the Null Hypothesis — Ho: $\mu m - \mu s \leq 0$
3. Calculate the pooled variance
4. Find the t-score
5. Find the p-value
6. Interpret the result

The p-value shows the reuslt is not significant.
Therefore, there is no reason to assume more people click on Mondays than on Saturdays.

Two Points Regarding the Educational Materials


Join GitHub today
GitHub is home to over 50 million developers working together to host and review code, manage projects, and build software together.
Sign up
main
1 branch
0 tags
Go to file
Code
fozouni Add files via upload
FIFA 2019, Session_6.xlsx Add files via upload
Python_1.ipynb Add files via upload
Python_2.ipynb Add files via upload
Python_3 - Conditionals.ipynb Add files via upload
Python_4, Functions.ipynb Add files via upload
Python_5 - Sequences.ipynb Add files via upload last month
Python_6_1, Iterations, for and while .i... Add files via upload 29 days ago
Python_6_2, OOP, modules and packa... Add files via upload 29 days ago
Session 10- Practical example.xlsx Add files via upload 4 days ago
Session 10.xlsx Add files via upload 7 days ago
Clone
HTTPS GitHub CLI
https://github.com/fozouni/data_scienc
Use Git or checkout with SVN using the web URL.
Open with GitHub Desktop
Download ZIP
About
Source codes of the first "Data Science Course"
Releases
No releases published
Packages
No packages published
Languages
Jupyter Notebook 99.6%
Python 0.4%


Instagram
Search
elmedade
Edit Profile
50 posts
6,509 followers
1 following
علم داده
Data Science
علـــم داده
دوره‌ی آموزشی رایگانِ #علم_داده یا #دیتا_ساینس
ویدیوهای آموزشی در کانال آپارات علم داده آپلود می‌شوند
zil.ink/elmedade
واژه‌ها
کدها
لغات و اصطلاحات
تنسور فلو
POSTS
IGTV
SAVED
TAGGED


آموزش رایگان «علم داده»
در این صفحه موارد مختلف آموزشی در خصوص علم داده را مشاهده خواهید نمود. ویدیوها توسط «گروه علم داده» به مدیریت دکتر محمد فزونی، عضو هیات علمی دانشگاه گنبدکاووس تهیه و منتشر می‌گردند.
کانال آپارات
کانال آپارات «علم داده»
جابجا کردن
گیت‌هاب علم داده
کدها و دیتاست‌ها
جابجا کردن
پست آموزش رایگان علم داده
پست اصلی در سایت
جابجا کردن

Thanks for watching AMIGOS, see you and take care

Also thanks to Data Science Course 2020 by Udemy and Data Science 365 team. Almost all the slides have been duplicated from this wonderful course.