Data Analytics Bootcamp

Week 8

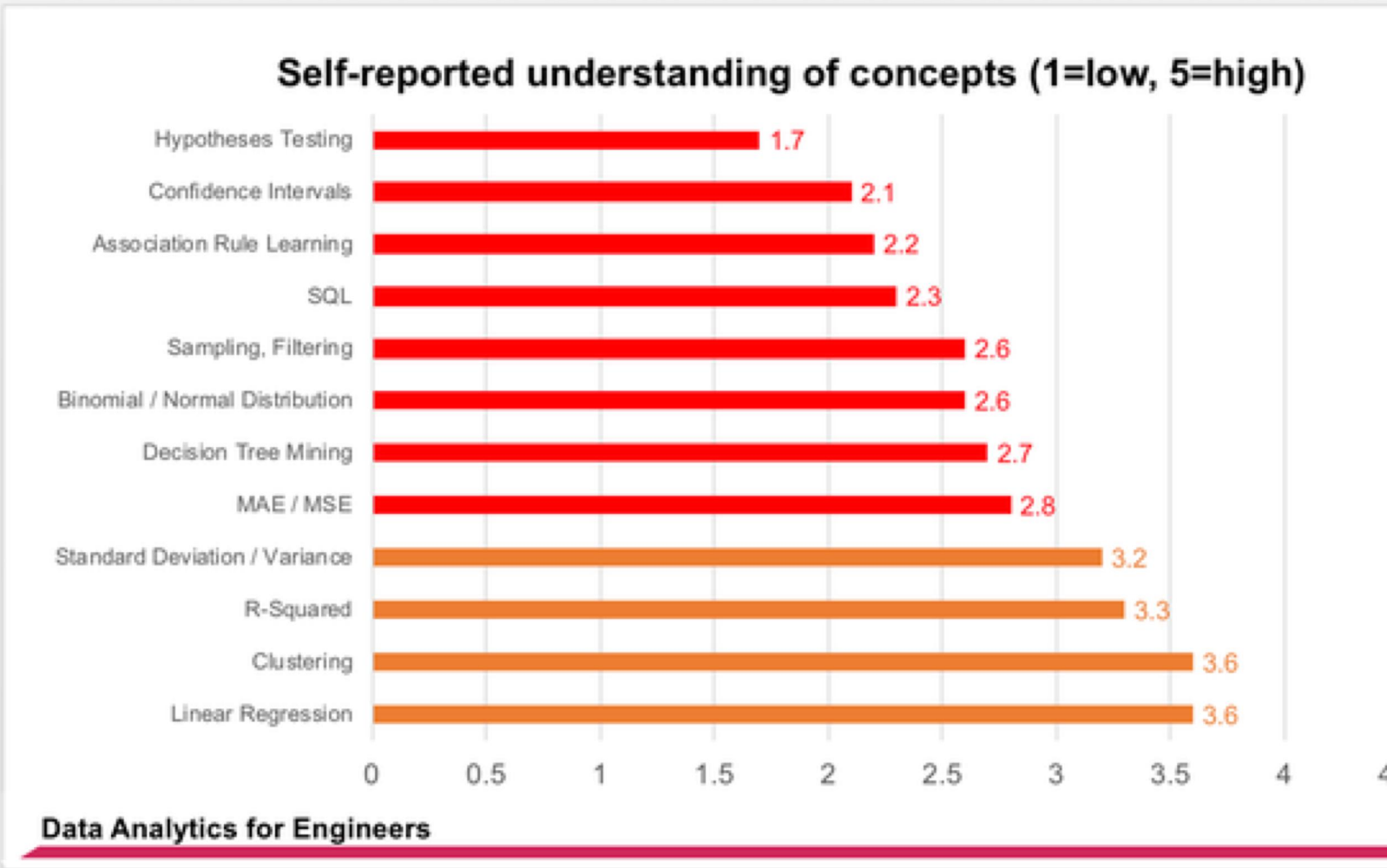
Roy Klaasse Bos

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

Why This Format?

- Already **video lectures** available
- **Practice makes perfect**
- **Limited** input on **AllAnswered**



Schedule

- **10:50 - 11:35** Take online practice exam (individually)
- **11:35 – 11:50** Break
- **11:50 – 12:30** Solutions / Discussion

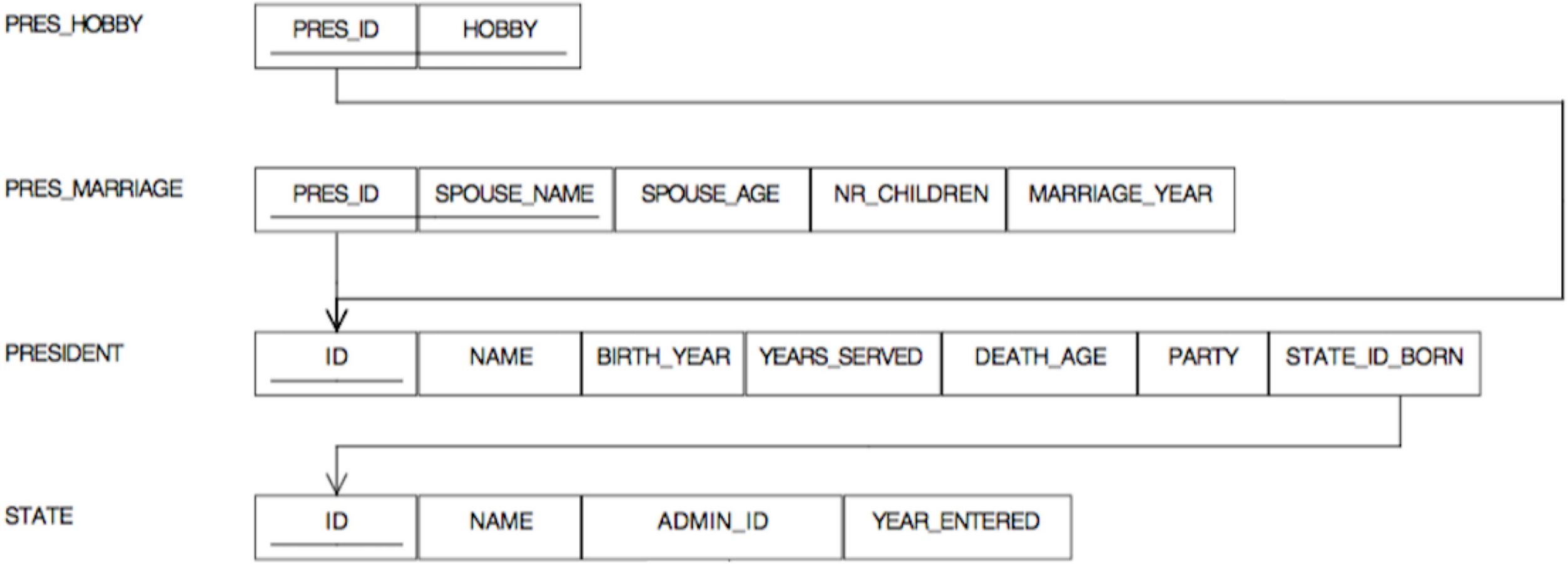
Set-up

- **Understanding** > **memorizing**
- **23 questions** (1.5-2.0 minutes/question)
- Digital **format**
- You can **skip questions** and **go back** later
- Your own answers will be automatically **sent to you**

A couple of rules...

- You are encouraged to **use slides/video lectures** etc.
- Feel free to leave the room once you're ready (in any case remain quiet)
- Try to do it on your **own**

SQL - Questions



Ready? Then, prove that you're a Data Analytics Expert!

Visit: https://goo.gl/NvK2jp and get started!

- **10:50 - 11:35** Take online practice exam (individually)
- **11:35 – 11:50** Break
- **11:50 – 12:30** Solutions / Discussion

Solutions / Discussion

Week 8

Roy Klaasse Bos

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

```sql
-- How many children were born in marriages before 1900 and from 1950 till today,
where the spouse is in her 30s at her marriage?

-- choose right table
SELECT *
FROM pres_marriage;

-- before 1900 and from 1950 till today
SELECT *
FROM pres_marriage
WHERE marriage_year <1900 OR marriage_year >=1950

-- where the spouse is in her 30s at her marriage
SELECT *
FROM PRES_MARRIAGE
WHERE (marriage_year >=1950 OR marriage_year <1900) AND spouse_age BETWEEN 30 AND 39;

-- how many children
SELECT SUM(nr_children)
FROM PRES_MARRIAGE
WHERE (marriage_year >=1950 OR marriage_year <1900) AND spouse_age BETWEEN 30 AND 39;
```

```sql
-- Question 2
-- Determine the maximum difference between the youngest died and oldest died president that served more than 4 years.

-- choose correct table
SELECT *
FROM president;

-- that served more than 4 years
SELECT *
FROM president
WHERE years_served > 4;

-- the maximum difference between the youngest died and oldest died president
SELECT MAX(death_age)-MIN(death_age)
FROM president
WHERE years_served > 4;

-- change column name to difference
SELECT MAX(death_age) - MIN(death_age) AS difference
FROM president
WHERE years_served > 4;
```

```sql
-- Question 3
-- Select the number of marriages in which a child was born on an average of minimal once in 5 years (starting with the year the wedding took place in, up to the year the spouse reached age 45).

-- select right table
SELECT *
FROM pres_marriage;

-- average of minimal 1 child per 5 years from the marriage year to the age of 45
SELECT *
FROM pres_marriage
WHERE nr_children >= (45-spouse_age) / 5.0; -- notice the decimal (i.e. 5.0)!

-- make sure the formula always works
SELECT *
FROM pres_marriage
WHERE nr_children >= (45-spouse_age) / 5.0 AND spouse_age <= 45;

-- select the number of marriages
SELECT COUNT(*)
FROM pres_marriage
WHERE nr_children >= (45-spouse_age) / 5.0 AND spouse_age <= 45;

```

```sql
-- Question 4
-- Determine for each president married more than once his id and the greatest as well as the least number of children born in his marriages

-- select right table
SELECT *
FROM pres_marriage;

-- how many marriages?
SELECT pres_id, COUNT(pres_id)
FROM pres_marriage
GROUP BY pres_id;

-- married more than once
SELECT pres_id, COUNT(pres_id)
FROM pres_marriage
GROUP BY pres_id
HAVING COUNT(pres_id) > 1;

-- greatest number of children born
SELECT pres_id, MAX(nr_children)
FROM pres_marriage
GROUP BY pres_id
HAVING COUNT(pres_id) > 1;

-- least number of children born
SELECT pres_id, MAX(nr_children), MIN(nr_children)
FROM pres_marriage
GROUP BY pres_id
HAVING COUNT(pres_id) > 1;

-- (optional) ORDER BY pres_id
SELECT pres_id, MAX(nr_children) AS max_children, MIN(nr_children) AS min_children
FROM pres_marriage
GROUP BY pres_id
HAVING COUNT(pres_id) > 1
ORDER BY pres_id;
```

Spot the **2 mistakes**

> **Hypothesis Testing**
> Say that we want to determine whether men and women have *different* average salaries. Students are asked to write the corresponding null (H0) and alternative hypothesis (Ha).

- One of the students formulates these hypotheses as follows:
  H0: ***Xmale = Xfemale***
  Ha: ***Xmale > Xfemale***
  What two critical mistakes did the student make? What should the hypotheses be instead?

One-sided vs Two-sided

Hypothesis Testing

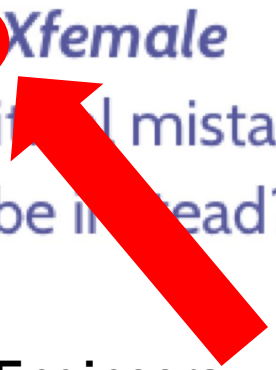
Say that we want to determine whether men and women have *different* average salaries. Students are asked to write the corresponding null (H0) and alternative hypothesis (Ha).

- One of the students formulates these hypotheses as follows:

  H0: ***Xmale = Xfemale***

  Ha: ***Xmale > Xfemale***

  What two critical mistakes did the student make? What should the hypotheses be instead?



Population vs Sample mean (Statics/Parameters)

Hypothesis Testing

Say that we want to determine whether men and women have *different* average salaries. Students are asked to write the corresponding null (H0) and alternative hypothesis (Ha).

- One of the students formulates these hypotheses as follows:

  $H0: \bar{X}_{male} = \bar{X}_{female}$

  $Ha: \bar{X}_{male} > \bar{X}_{female}$

  What two critical mistakes did the student make? What should the hypotheses be instead?

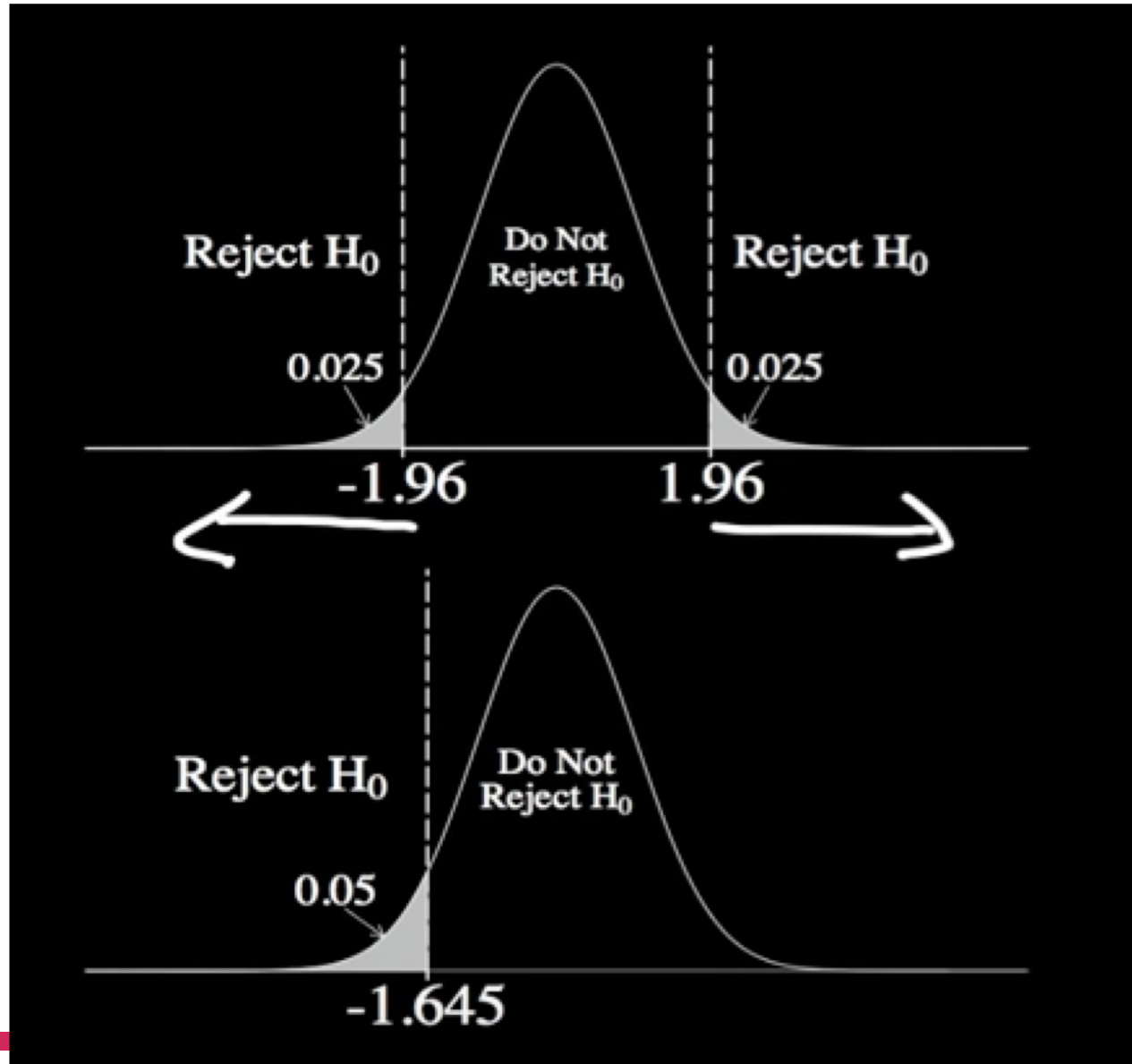
Valid conclusion?

> **Hypothesis Testing**
> Say that we want to determine whether men and women have *different* average salaries. Students are asked to write the corresponding null and alternative hypothesis.

After conducting the right hypotheses tests, the student finds a p-value of **0.19** and thus concludes that there is strong evidence that men earn as much as women. Is this statement correct? If so/not, why?

What's the effect on the rejection region if we go from a two-sided to a one-sided hypothesis test and the significance level remains the same (e.g. 0.05). What does that imply?

What's the effect on the rejection region if we go from a two-sided to a one-sided hypothesis test and the significance level remains the same (e.g. 0.05). What does that imply?



Read the description below. What could cause such a **low p-value** in this case?

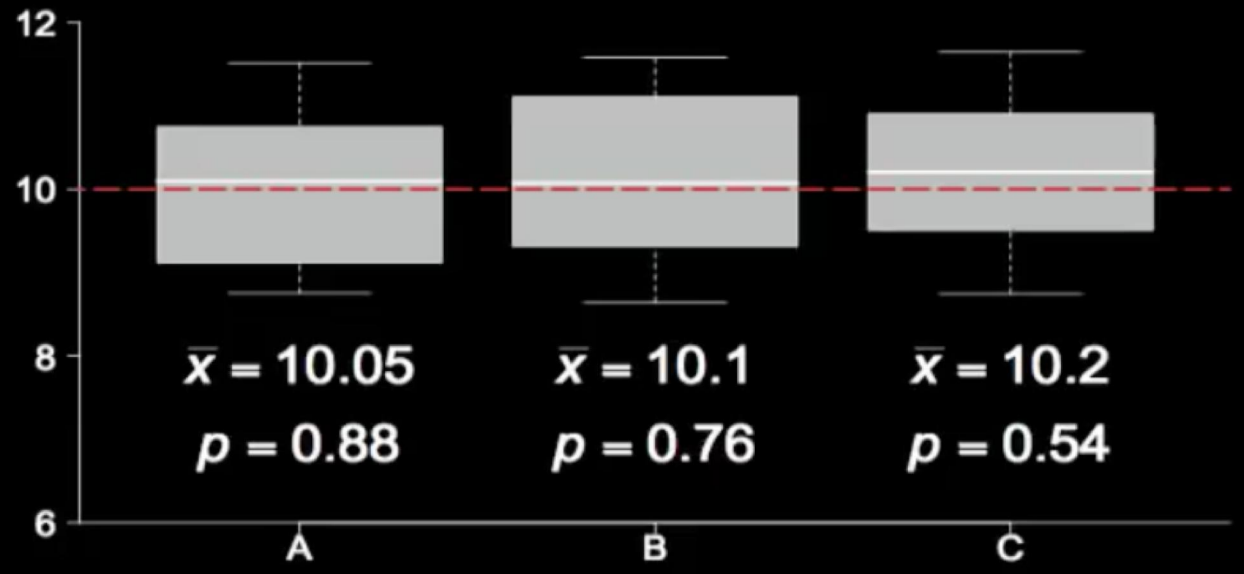
Suppose a call centre claims their average wait time is 30 seconds. We decide to test:

$$H_0: \mu = 30$$
$$H_a: \mu > 30$$

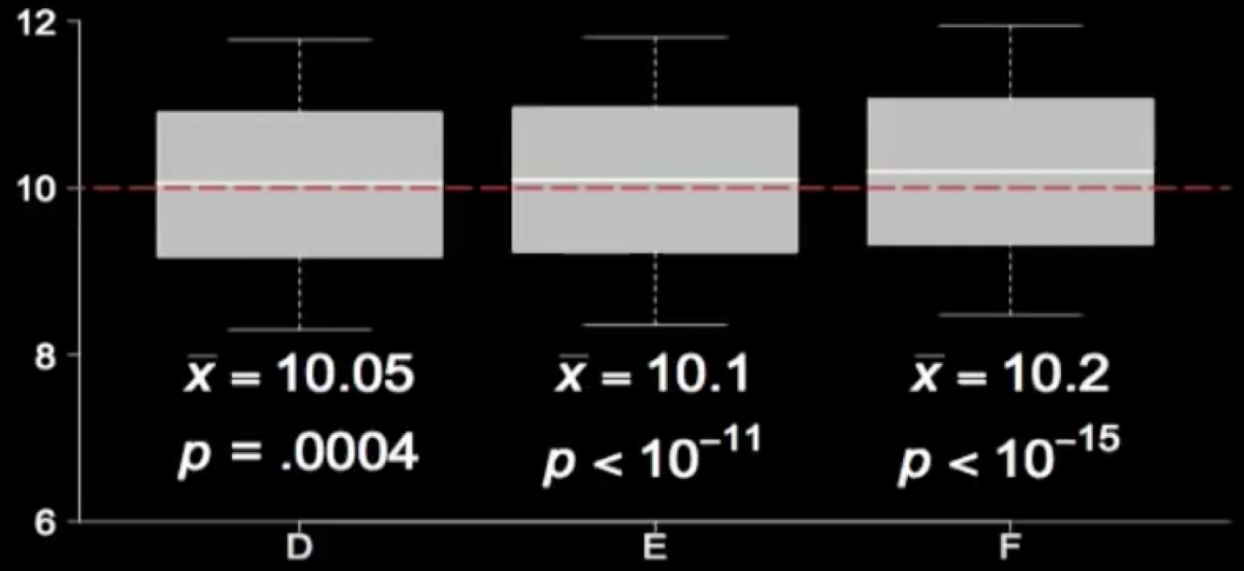
We find $\bar{X} = 30.6$, and a p-value of 0.002.

Testing H_0: $\mu = 10$ for 6 samples.

$n = 10$



$n = 5000$



TU/e

In hypotheses testing we distinguish the **Z-test** and **T-test**. Explain when to use which one. Which one do you think is more common?

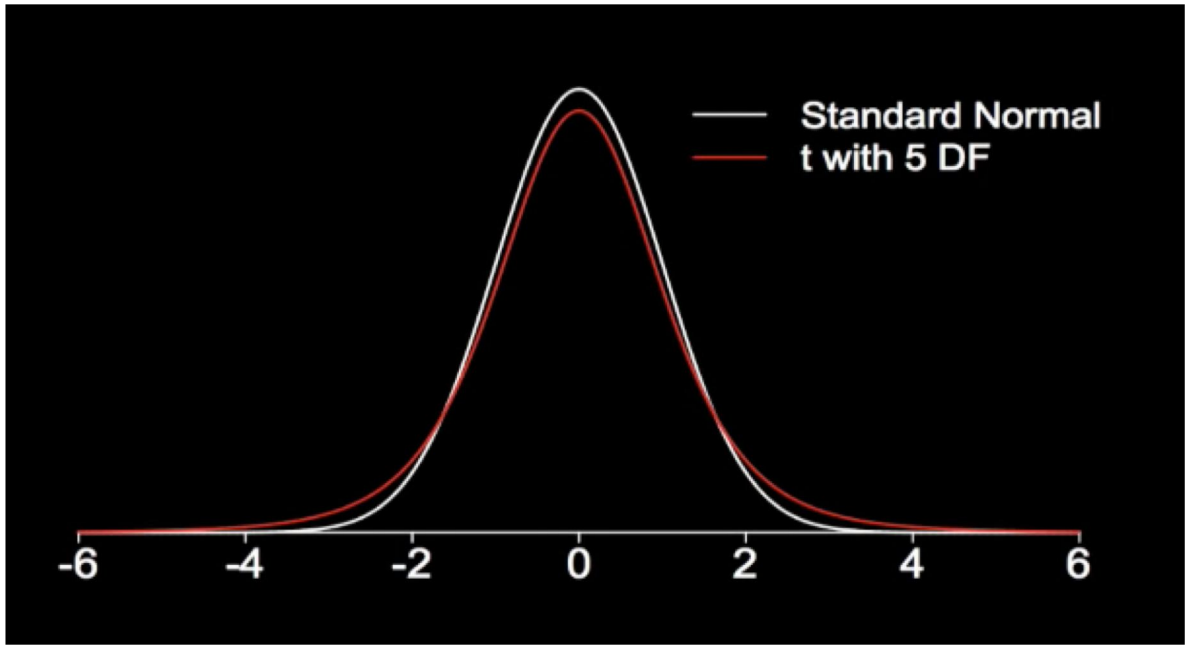
In hypotheses testing we distinguish the **Z-test** and **T-test**. Explain when to use which one. Which one do you think is more common?

If σ is known:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

If σ is unknown:

$$t = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$



What's the effect on the **rejection region** if the **sample size increases** for a **Z-test** and **T-test**.

What's the effect on the **rejection region** if the **sample size increases** for a **Z-test** and **T-test**.

Standard normal distribution

Example: $\Phi(0.31) = P(Z \leq 0.31) = 0.6217$.

	0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224

What's the effect on the **rejection region** if the **sample size increases** for a **Z-test** and **T-test**.

10.2 Student t-distribution $(t_{\nu;\alpha})$

Example: $P(T_3 \geq 1.638) = 0.1$, thus $t_{3;0.1} = 1.638$.

ν \ α	0.3	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001
1	0.727	1.376	1.963	3.078	6.314	12.71	15.90	31.82	63.66	127.3	318.3
2	0.617	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.10	22.33
3	0.584	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215
4	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173
5	0.559	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893

Type I and II errors

Consider a criminal trial. We test the hypotheses:

H0: The defendant did not commit the crime.

Ha: The defendant committed the crime.

Explain in your own words **what** the corresponding **Type I and II errors** are in this context.

Type I and II errors

Consider a criminal trial. We test the hypotheses:

H0: The defendant did not commit the crime.

Ha: The defendant committed the crime.

Explain in your own words **what** the corresponding **Type I and II errors** are in this context.

A Type I error is rejecting H_0 when, in reality, it is true.

A Type II error is failing to reject H_0 when, in reality, it is false.

Type I and II errors

Consider a criminal trial. We test the hypotheses:

H0: The defendant did not commit the crime.

Ha: The defendant committed the crime.

Explain in your own words **what** the corresponding **Type I and II errors** are in this context.

		Underlying reality	
		H_0 is false	H_0 is true
Conclusion from test	Reject H_0	Correct decision	Type I error
	Do not reject H_0	Type II error	Correct decision

Type I and II errors

Consider a criminal trial. We test the hypotheses:

H0: The defendant did not commit the crime.

Ha: The defendant committed the crime.

Explain in your own words **what** the corresponding **Type I and II errors** are in this context.

Type I error: Convicting a person who, in reality, did not commit the crime.

Type II error: Acquitting a person who, in reality, committed the crime.

Given the context which **type** of error (**I** or **II**) do you believe is **worse**?
Why?

Given the context which **type** of error (**I** or **II**) do you believe is **worse**? Why?

Type I error: Convicting a person who, in reality, did not commit the crime.

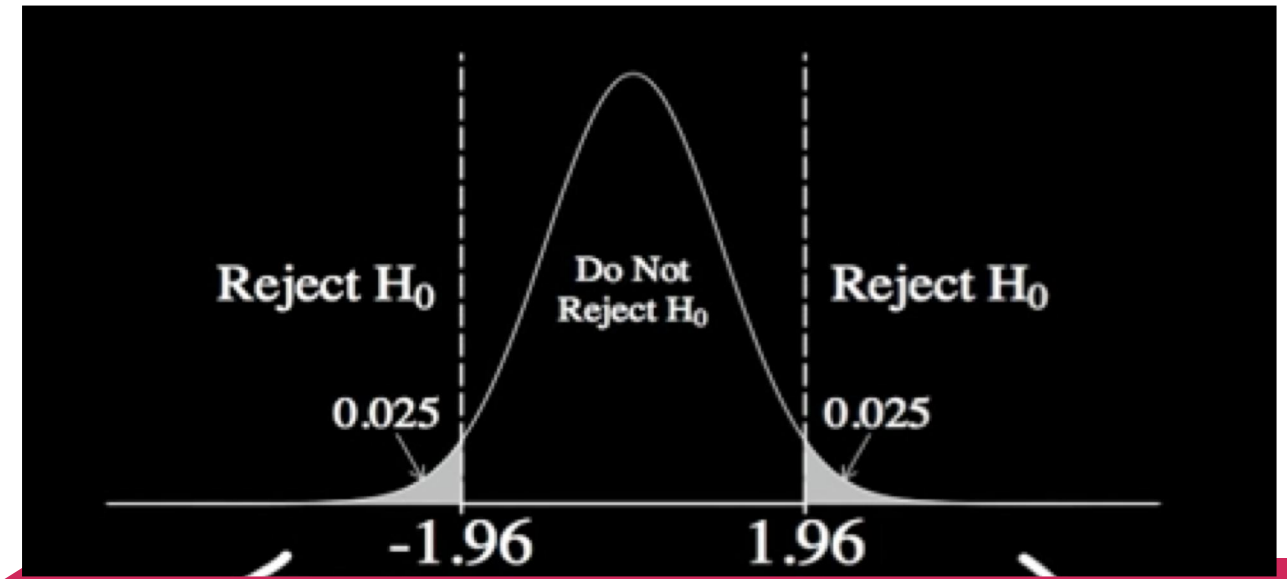
Type II error: Acquitting a person who, in reality, committed the crime.

Assume that the court uses hypothesis testing to form its final verdict. What can we say about the **significance level** of this test if the court wants to minimize the number of **Type I errors**?

Assume that the court uses hypothesis testing to form its final verdict. What can we say about the **significance level** of this test if the court wants to minimize the number of **Type I errors**?

Type I error: Convicting a person who, in reality, did not commit the crime.

A Type I error is rejecting H_0 when, in reality, it is true.

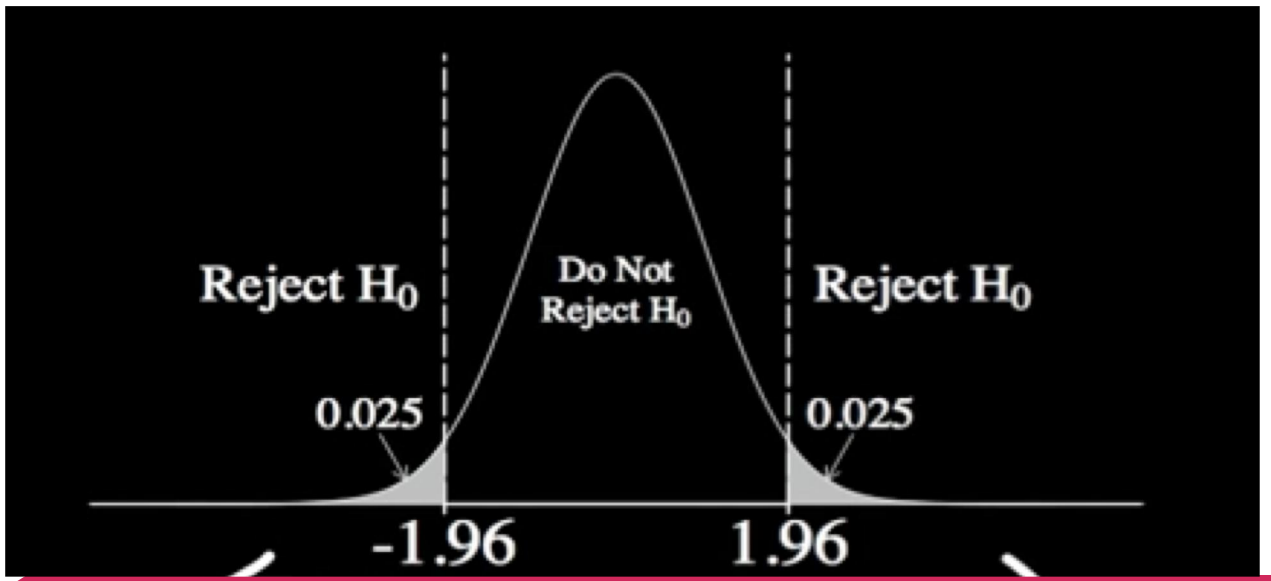


Explain the consequences of choosing a significance level of **0.10** instead of **0.05** in this context. What will be the effect on the number of Type I and II errors?

Explain the consequences of choosing a significance level of **0.10** instead of **0.05** in this context. What will be the effect on the number of Type I and II errors?

A Type I error is rejecting H_0 when, in reality, it is true.

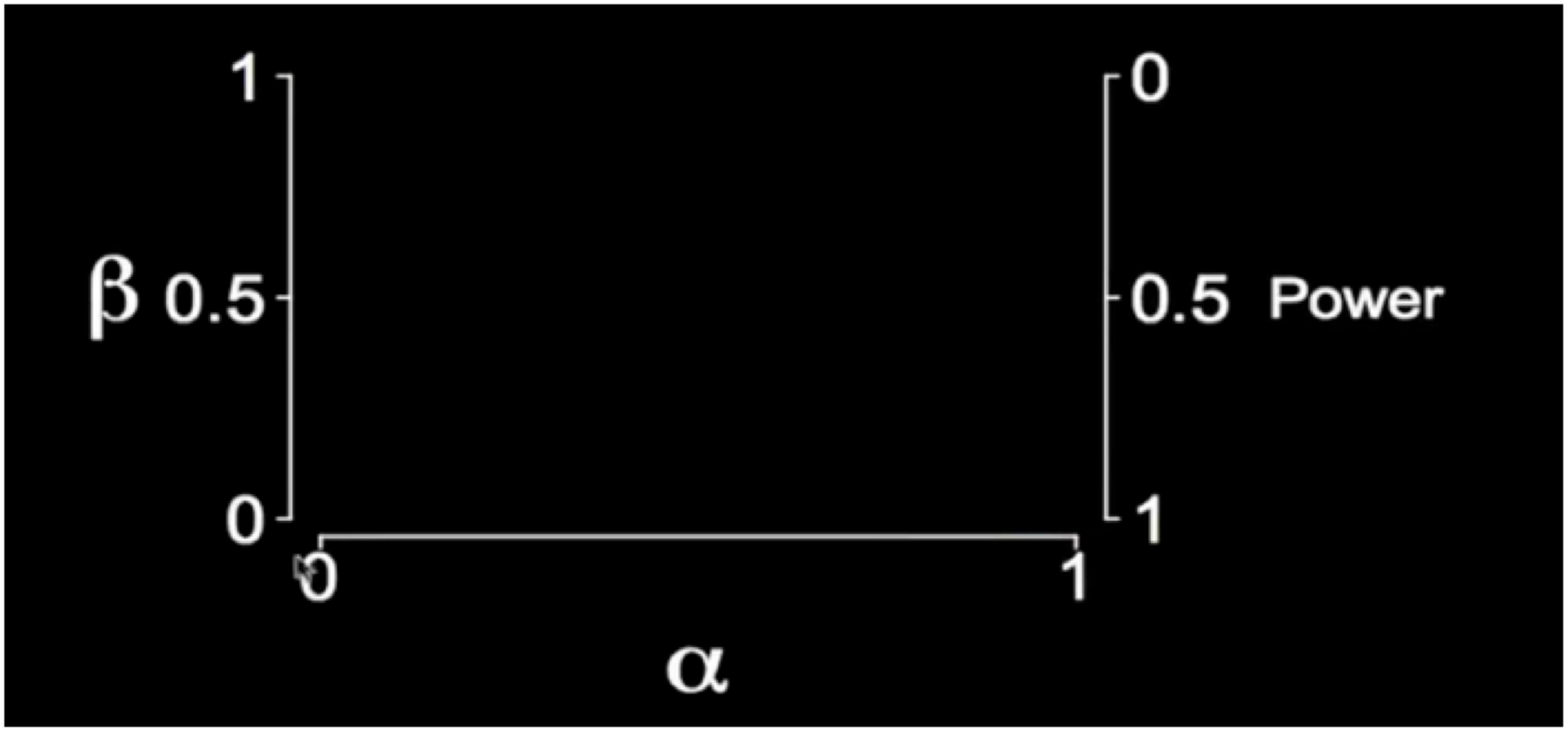
A Type II error is failing to reject H_0 when, in reality, it is false.



Here you see an empty chart with 3 axes:

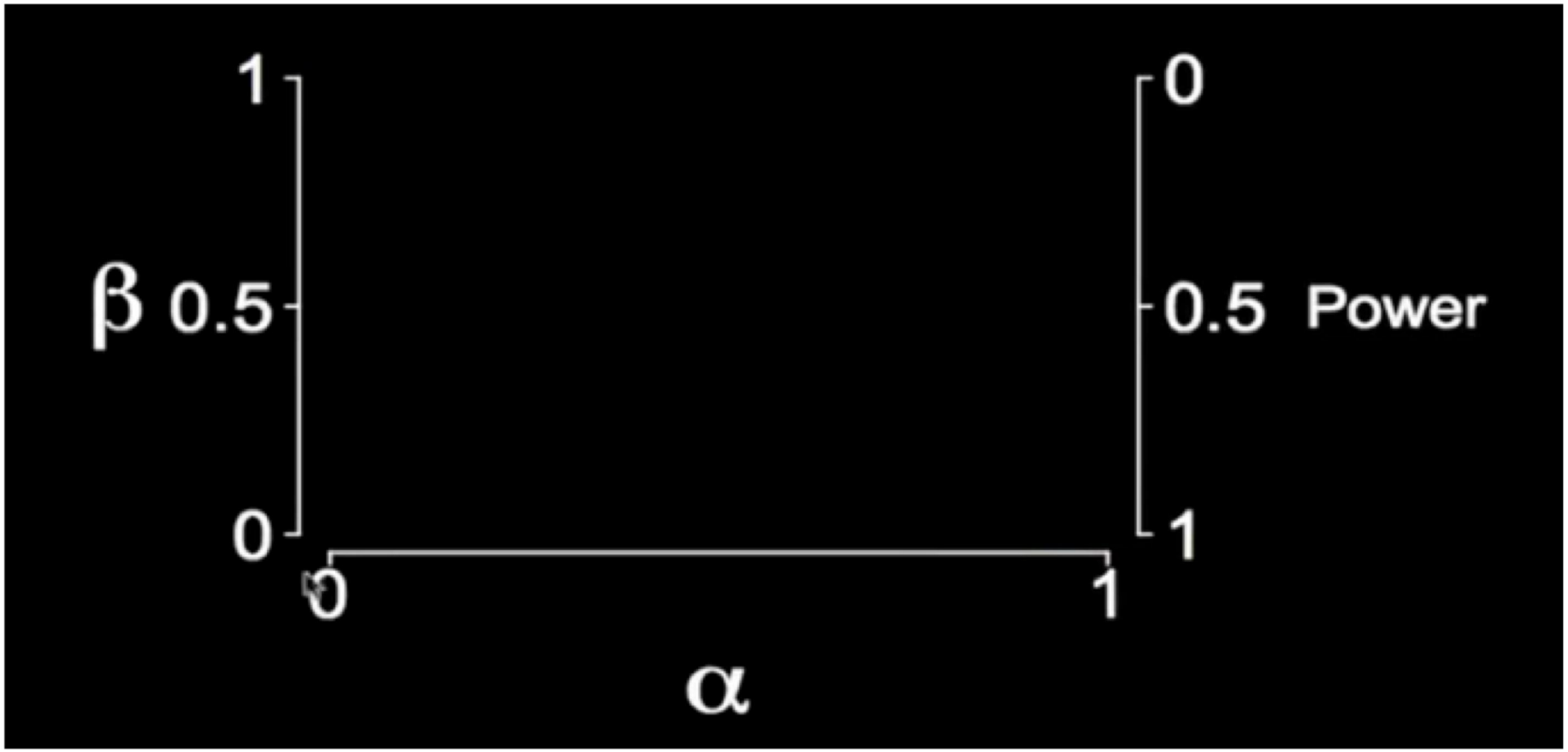
- Type I error (horizontal - alpha)
- Type II error (left hand side - beta)
- power (right hand side)

If you were to draw a line in this chart how would it look like?



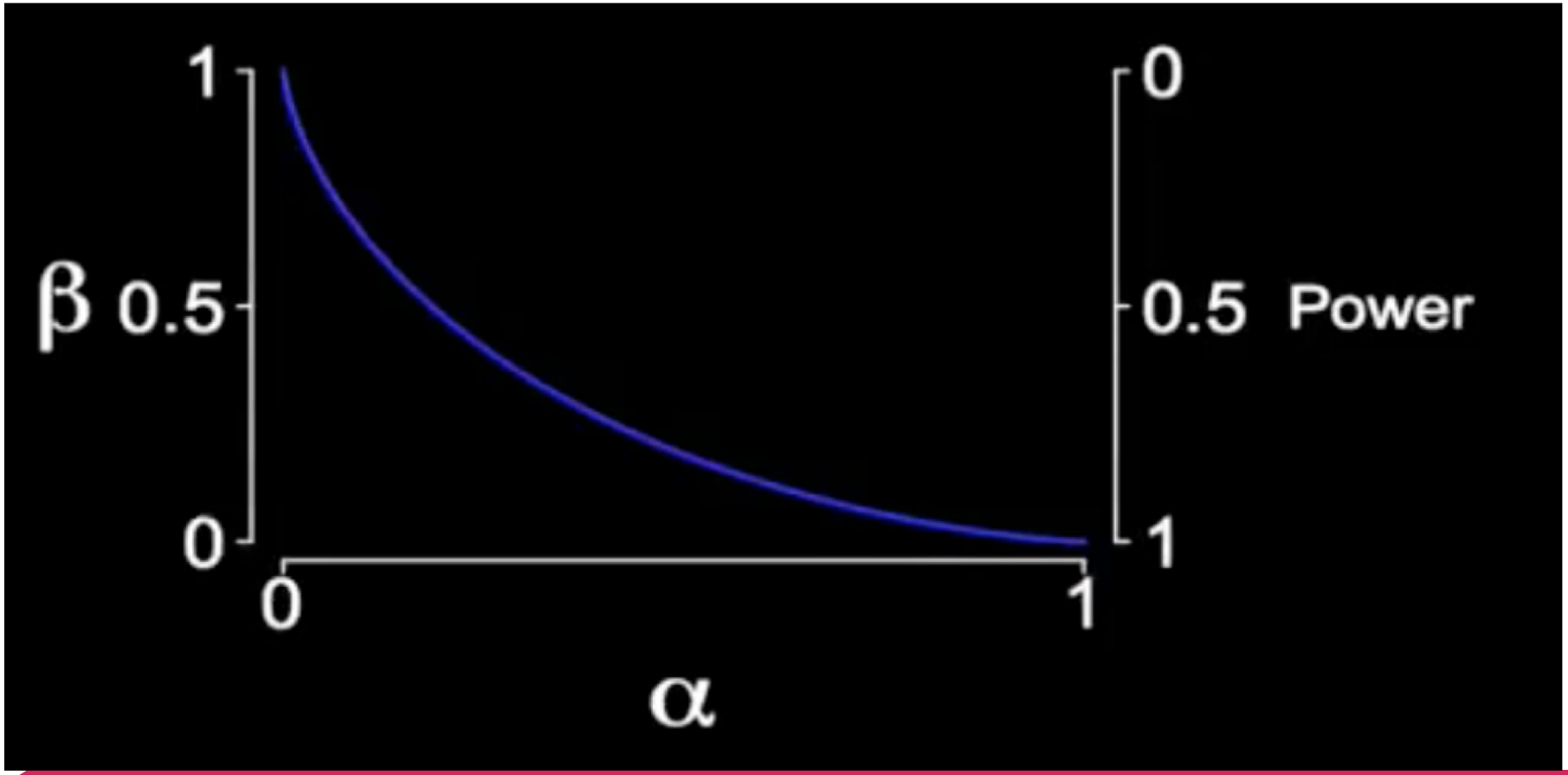
A Type I error is rejecting H_0 when, in reality, it is true.

A Type II error is failing to reject H_0 when, in reality, it is false.



A Type I error is rejecting H_0 when, in reality, it is true.

A Type II error is failing to reject H_0 when, in reality, it is false.



“ Standard Deviation / Confidence Intervals

Suppose two professsors (A and B) look into the grades for GA1.
Professor **A** looks at a sample A of **10 DAE-students** and professor **B** at a sample B of **100 DAE-students**.

A student comments on the sample sizes of both professors: "the **standard deviation** for A must be **larger** than for B because sample B contains 10 times as many data points". Do you agree? Why?

Standard Deviation / Confidence Intervals

Suppose two professsors (A and B) look into the grades for GA1.
Professor **A** looks at a sample A of **10 DAE-students** and professor **B** at a sample B of **100 DAE-students**.

A student comments on the sample sizes of both professors: "the **standard deviation** for A must be **larger** than for B because sample B contains 10 times as many data points". Do you agree? Why?

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set,
N = number of values in the data set.

Standard Deviation / Confidence Intervals

Suppose two professsors (A and B) look into the grades for GA1.
Professor **A** looks at a sample A of **10 DAE-students** and professor **B** at a sample B of **100 DAE-students**.

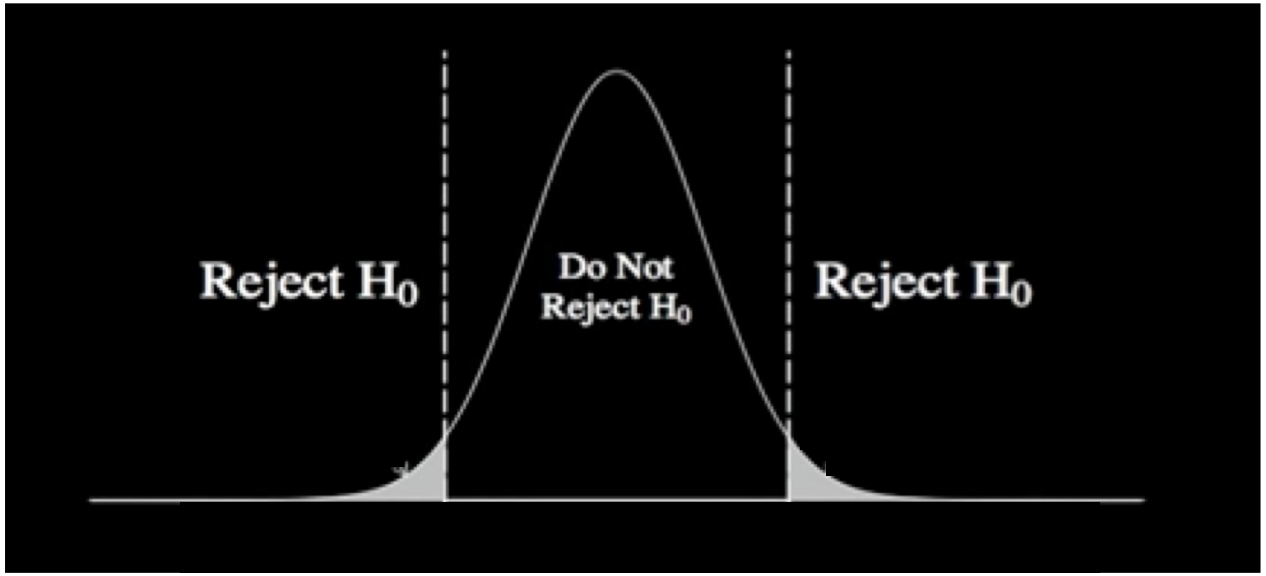
The professors find a **90% confidence** interval of **6.9-8.3** for the GA1-grades.
What can we conclude about the p-value for **alpha=0.10** given the following two hypotheses: **H0: μ=7.1** and **Ha: μ≠7.1**?

Standard Deviation / Confidence Intervals

Suppose two professsors (A and B) look into the grades for GA1.
Professor **A** looks at a sample A of **10 DAE-students** and professor **B** at a sample B of **100 DAE-students**.

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The professors find a **90% confidence** interval of **6.9-8.3** for the GA1-grades. What can we conclude about the p-value for **alpha=0.10** given the following two hypotheses: **H0: μ=7.1** and **Ha: μ≠7.1**?



Binomial Distribution

Variance: ***n*p*(1-p)***

Mean: ***n*p***

Where p is the probability of success and n the number of individuals.

Derive the formula for the **standard deviation** (sd) for a binomial distribution and determine for which value of p the sd is largest.

Binomial Distribution

Variance: ***n*p*(1-p)***

Mean: ***n*p***

Where p is the probability of success and n the number of individuals.

Explain in your own words why this value for p makes sense.

Association Rule Learning

Suppose we want to find out common relationships between disease A and B. Two metrics you should consider are support and confidence:

$$Support\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Total\ number\ of\ disease}$$

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

Association Rule Learning

Suppose we want to find out common relationships between disease A and B. Two metrics you should consider are support and confidence:

$$Support\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Total\ number\ of\ disease}$$

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

What is the effect of choosing a (too) **low support cut-off level**?

What is the effect of choosing a (too) **high confidence cut-off level**?

$$Support\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Total\ number\ of\ disease}$$

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

Disease A	1	1	1	1	0	0	0	0
Disease B	0	1	1	1	1	1	1	1

$$Support\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Total\ number\ of\ disease}$$

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

Disease A	1	1	1	1	0	0	0	0
Disease B	0	1	1	1	1	1	1	1

Rule	Support	Confidence
A => B		

$$Support\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Total\ number\ of\ disease}$$

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

Disease A	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**
Disease B	0	1	1	1	1	1	1	1

Rule	Support	Confidence
A => B	3/8	

$$Support\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Total\ number\ of\ disease}$$

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

Disease A	1	1	1	1	0	0	0	0
Disease B	0	1	1	1	1	1	1	1

Rule	Support	Confidence
A => B	3/8	3/4

A student argues that a **high accuracy** automatically means you have a **high precision**. Do you agree? Illustrate with an example.

Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

A student argues that a **high accuracy** automatically means you have a **high precision**. Do you agree? Illustrate with an example.

Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

		Prediction	
		Yes	No
Real Value	Yes	TP	FN
	No	FP	TN

		Prediction	
		Yes	No
Real Value	Yes	1	4
	No	11	299

A student argues that a **high accuracy** automatically means you have a **high precision**. Do you agree? Illustrate with an example.

Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

		Prediction	
		Yes	No
Real Value	Yes	TP	FN
	No	FP	TN

		Prediction	
		Yes	No
Real Value	Yes	1	4
	No	11	299

Accuracy: (1+299)/(1+299+4+11) ≈ **0.95**

TU/e

A student argues that a **high accuracy** automatically means you have a **high precision**. Do you agree? Illustrate with an example.

Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

		Prediction	
		Yes	No
Real Value	Yes	TP	FN
	No	FP	TN

		Prediction	
		Yes	No
Real Value	Yes	1	4
	No	11	299

Accuracy: (1+299)/(1+299+4+11) **≈ 0.95**

Precision: 1/(1+11) **≈ 0.08**

“ MAE / RMSE

Two metrics used to measure accuracy for continuous variables are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE).

Continue press ENTER

Can you explain the **difference** between the two metrics **in your own words**?

MAE / RMSE

Two metrics used to measure accuracy for continuous variables are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE).

Continue press ENTER

Can you explain the **difference** between the two metrics **in your own words?**

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

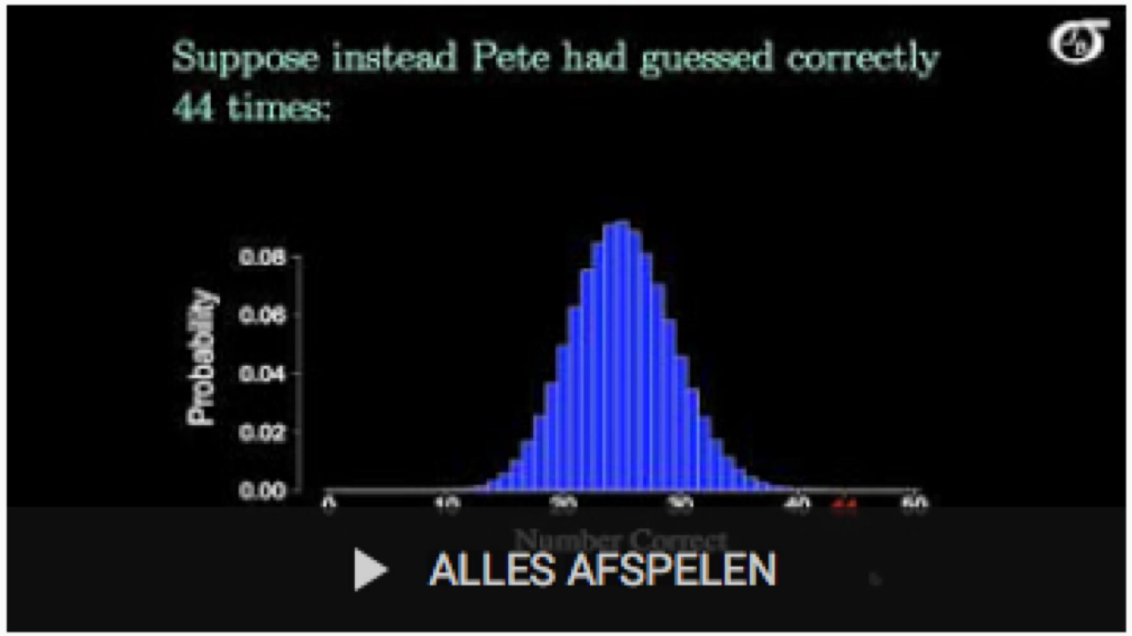
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Real value	Predicted Value	Absolute Error	Squared Error
1	4	3	9
2	1	1	1
3	19	16	256
4	4	0	0
		MAE: 20/4 = 5	**RMSE: (266/4)^.5 ≈ 8.2**

Looking for Excellent Explanations?



jbstatistics GEABONNEERD 63K

An introduction to hypothesis testing. I discuss the concept of hypothesis testing, rejection regions, p-values, Type I and Type II errors, power calculations, statistical significance, Z tests for one mean, t tests for one mean, the assumptions of the tests.

Hypothesis Testing

16 video's • 236.948 weergaven • Laatst geüpdatet op 25 jun. 2014

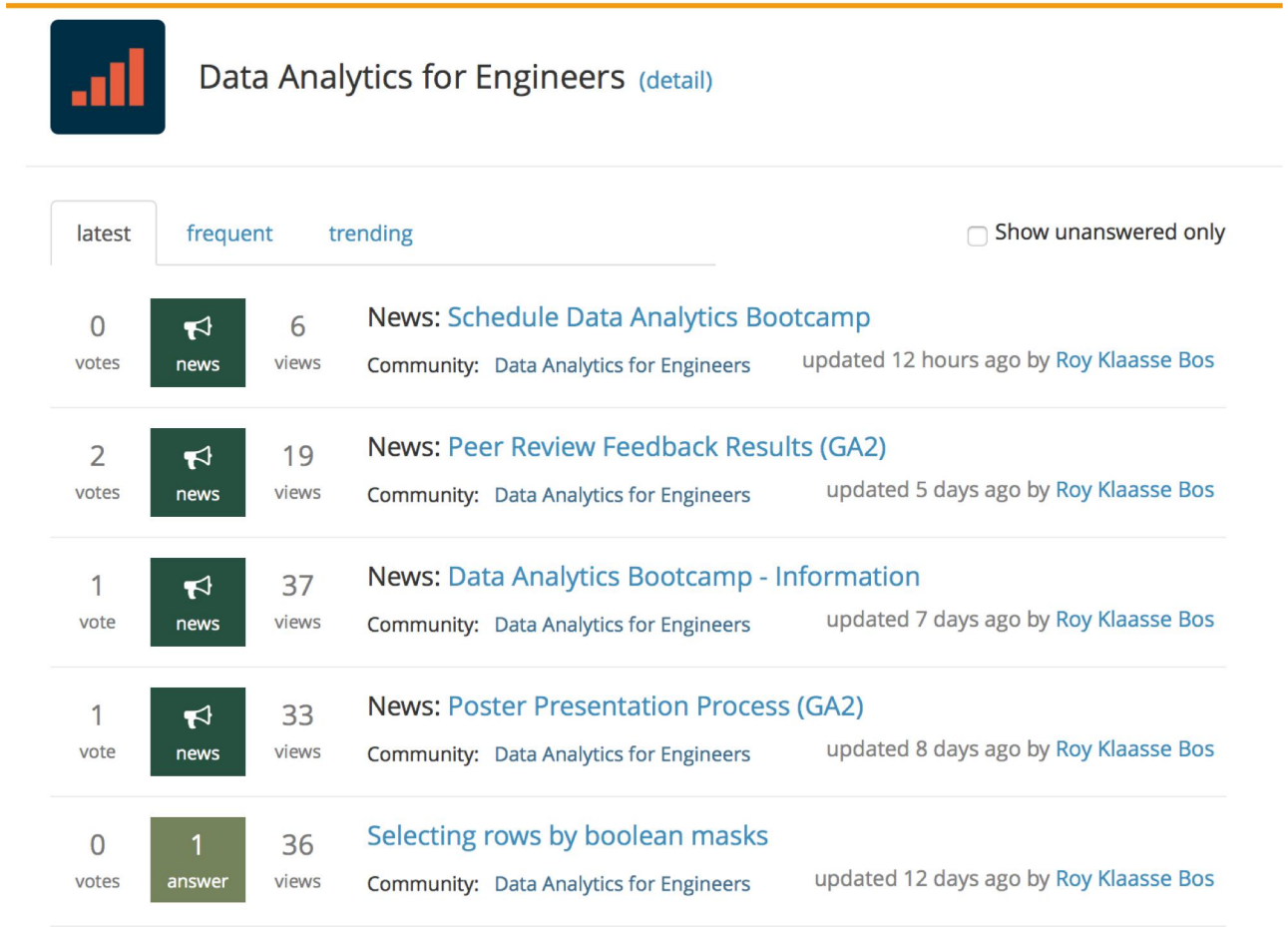
Solutions

24 → That was it! Would you like to receive the (correct) **solutions** by email?

They will be send to you next week.

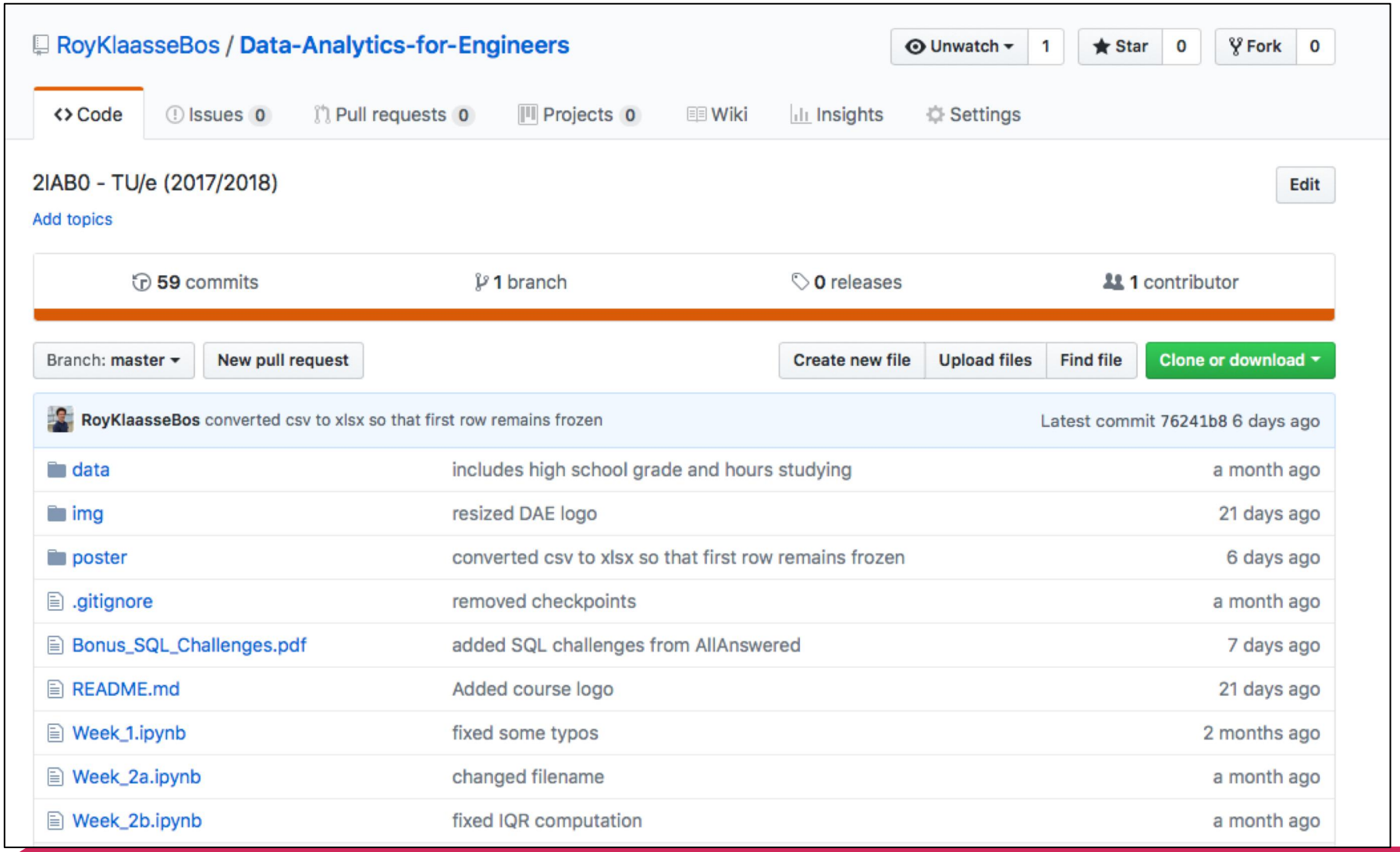
Y Yes

N No



https://www.allanswered.com/community/s/data-analytics-for-engineers/

Bonus SQL Challenges



RoyKlaasseBos / Data-Analytics-for-Engineers

Unwatch 1 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

2IAB0 - TU/e (2017/2018)

Edit

Add topics

59 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

RoyKlaasseBos converted csv to xlsx so that first row remains frozen — Latest commit 76241b8 6 days ago

data	includes high school grade and hours studying	a month ago
img	resized DAE logo	21 days ago
poster	converted csv to xlsx so that first row remains frozen	6 days ago
.gitignore	removed checkpoints	a month ago
Bonus_SQL_Challenges.pdf	added SQL challenges from AllAnswered	7 days ago
README.md	Added course logo	21 days ago
Week_1.ipynb	fixed some typos	2 months ago
Week_2a.ipynb	changed filename	a month ago
Week_2b.ipynb	fixed IQR computation	a month ago

Thanks and …