POLI210: Political Science Research Methods

Lecture 11.2: Null Hypothesis Significance Testing

Olivier Bergeron-Boutin

November 4th, 2021

Boring admin stuff

- Problem set 4 will be posted today; **mandatory**
- Due dates:
  - Problem set 4: December 2nd, 11:59PM
  - Quiz 2: November 30th to December 4th
  - Group project: December 6th, 11:59PM
- Group project: you should find your teams ASAP
- I post interesting articles on MyCourses

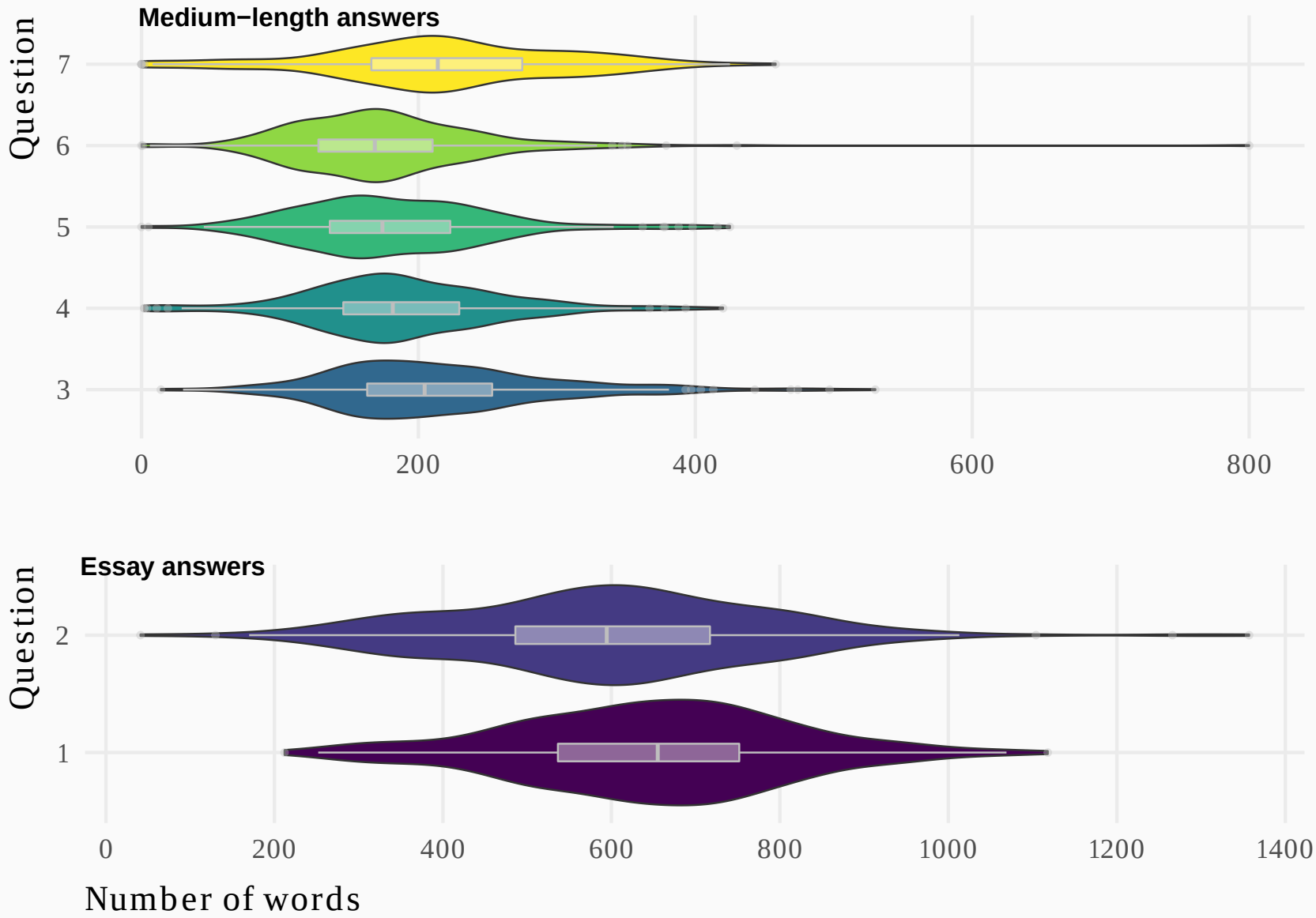
Final project

- 1500 words
- Teams of four
- Pick one article out of 5 and critique it

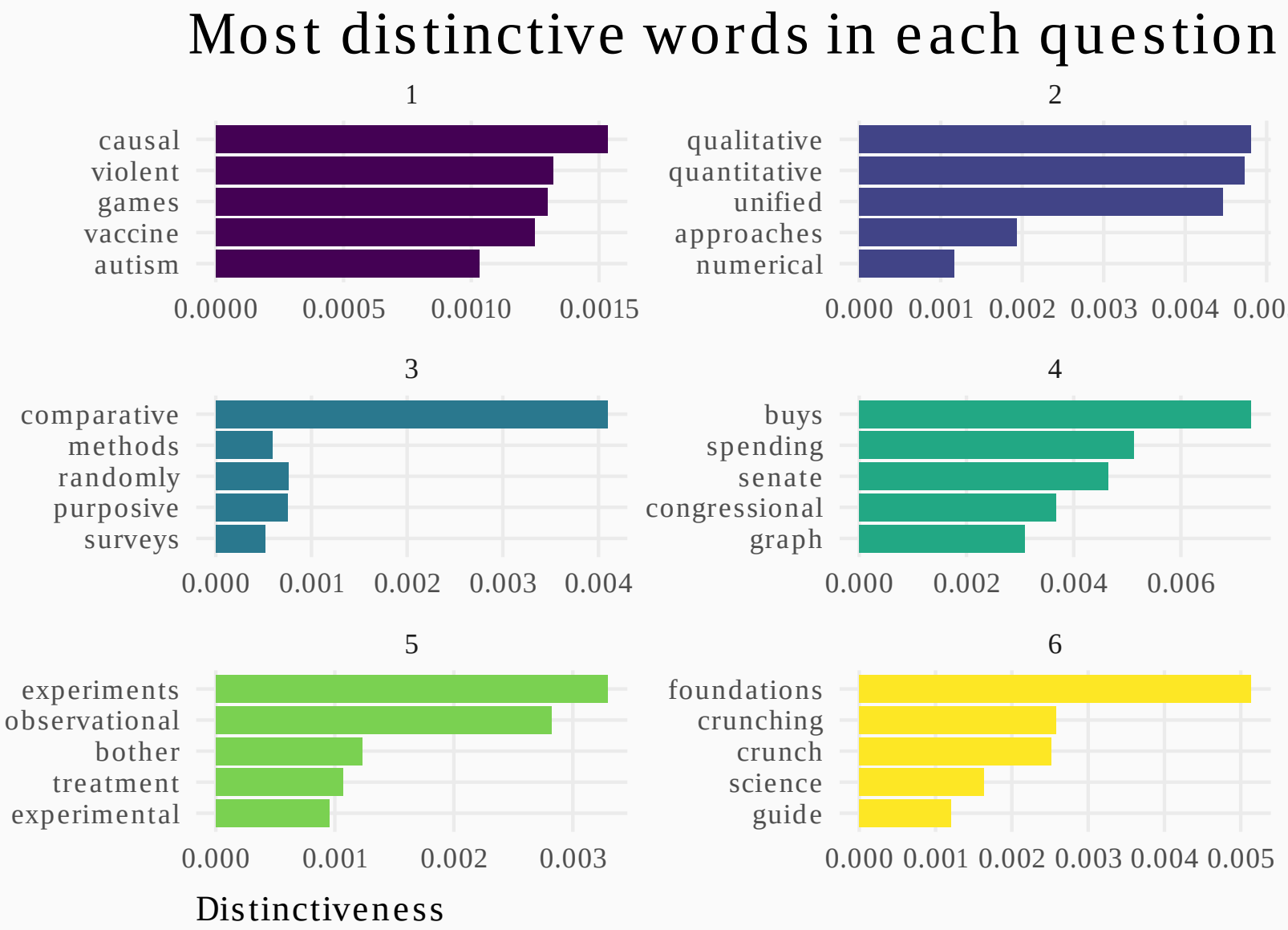
Articles to choose from

Article	Authors	Substantive content	Methodology
The Signs of Deconsolidation	Roberto Stefan Foa and Yascha Mounk	Attitudes toward democracy in advanced democracies	Quantitative analysis of survey data
The Great Divide: Literacy, Nationalism, and the Communist Collapse	Keith Darden and Anna Grzymala-Busse	Explaining cross-country variation in the success of communist parties in post-communist countries	Cross-country regression
Conceptual Models and the Cuban Missile Crisis	Graham T. Allison	Analysis of the Cuban Missile Crisis and foreign affairs decision-making	Case study using historical evidence
Sources of Authoritarian Responsiveness: A Field Experiment in China	Jidong Chen, Jennifer Pan, and Yiqing Xu	What motivates government officials in an authoritarian state to be responsive to public demands?	Field experiment in China
Democracy, Autocracy, and Revolution in Post-Soviet Eurasia	Henry E. Hale	The success or failure of transitions to democracy in post-communist Eurasia	Comparative analysis

Exploration of midterm data



Exploration of midterm data



When do you believe me?

Let's suppose that after the midterm, I tell you that the mean grade is 73

- You suspect that I'm lying, for some reason...
- But don't want to call me out unless you're quite sure
- You ask a colleague in lab about their grade...
    - Then another, and another, and another...
- When do you have enough evidence to call me out?

When do you believe me?

Let's suppose that after the midterm, I tell you that the mean grade is 73

- You suspect that I'm lying, for some reason...
- But don't want to call me out unless you're quite sure
- You ask a colleague in lab about their grade...
    - Then another, and another, and another...
- When do you have enough evidence to call me out?

Meeting students

Table 1: Grades of students you meet in lab

Student #	Grade
1	63

Do you call me a liar?

Meeting students

Table 2: Grades of students you meet in lab

Student #	Grade
1	63
2	67

Do you call me a liar?

Meeting students

Table 3: Grades of students you meet in lab

Student #	Grade
1	63
2	67
3	71

Do you call me a liar?

Meeting students

Table 4: Grades of students you meet in lab

Student #	Grade
1	63
2	67
3	71
4	56

Do you call me a liar?

Meeting students

Table 5: Grades of students you meet in lab

Student #	Grade
1	63
2	67
3	71
4	56
5	77

Do you call me a liar?

Meeting students

Table 6: Grades of students you meet in lab

Student #	Grade
1	63
2	67
3	71
4	56
5	77
6	47

Do you call me a liar?

The null hypothesis

The setup:

- We set a **null hypothesis**, also referred to as H_0
  - The null hypothesis is our reference point – it is arbitrary!
  - It's a sort of statistical "strawman"
- We then set an **alternative hypothesis**, or H_1
  - If the null is not true, then the alternative hypothesis must be true
- We start from the premise that the null hypothesis is true
  - The key question: How surprised are you to see the data that you have, if the null hypothesis is true?
  - Evidence is inconsistent with the null $\rightsquigarrow$ reject the null
  - Evidence is not inconsistent with the null $\rightsquigarrow$ fail to reject the null
- This is the framework of **hypothesis testing**
  - Start from the null
  - Think about what the data should look like, if the null were true
  - Analyze the data; reject/fail to reject the null

The null hypothesis in our example

What was the null hypothesis in the example above?

The null hypothesis in our example

What was the null hypothesis in the example above?

- H_0: $\mu_{exam} = 73$

What was the alternative hypothesis?

The null hypothesis in our example

What was the null hypothesis in the example above?

- H_0: $\mu_{exam} = 73$

What was the alternative hypothesis?

- H_1: $\mu_{exam} \neq 73$ (non-directional hypothesis)
- H_1: $\mu_{exam} > 73$ (directional hypothesis)
- H_1: $\mu_{exam} < 73$ (directional hypothesis)

The null hypothesis in our example

What was the null hypothesis in the example above?

- H_0: $\mu_{exam} = 73$

What was the alternative hypothesis?

- H_1: $\mu_{exam} \neq 73$ (non-directional hypothesis)
- H_1: $\mu_{exam} > 73$ (directional hypothesis)
- H_1: $\mu_{exam} < 73$ (directional hypothesis)

Hypothesis testing in our example

Assume that the null is true – i.e. the true mean is 73

- What do you expect to see when talking to your peers?

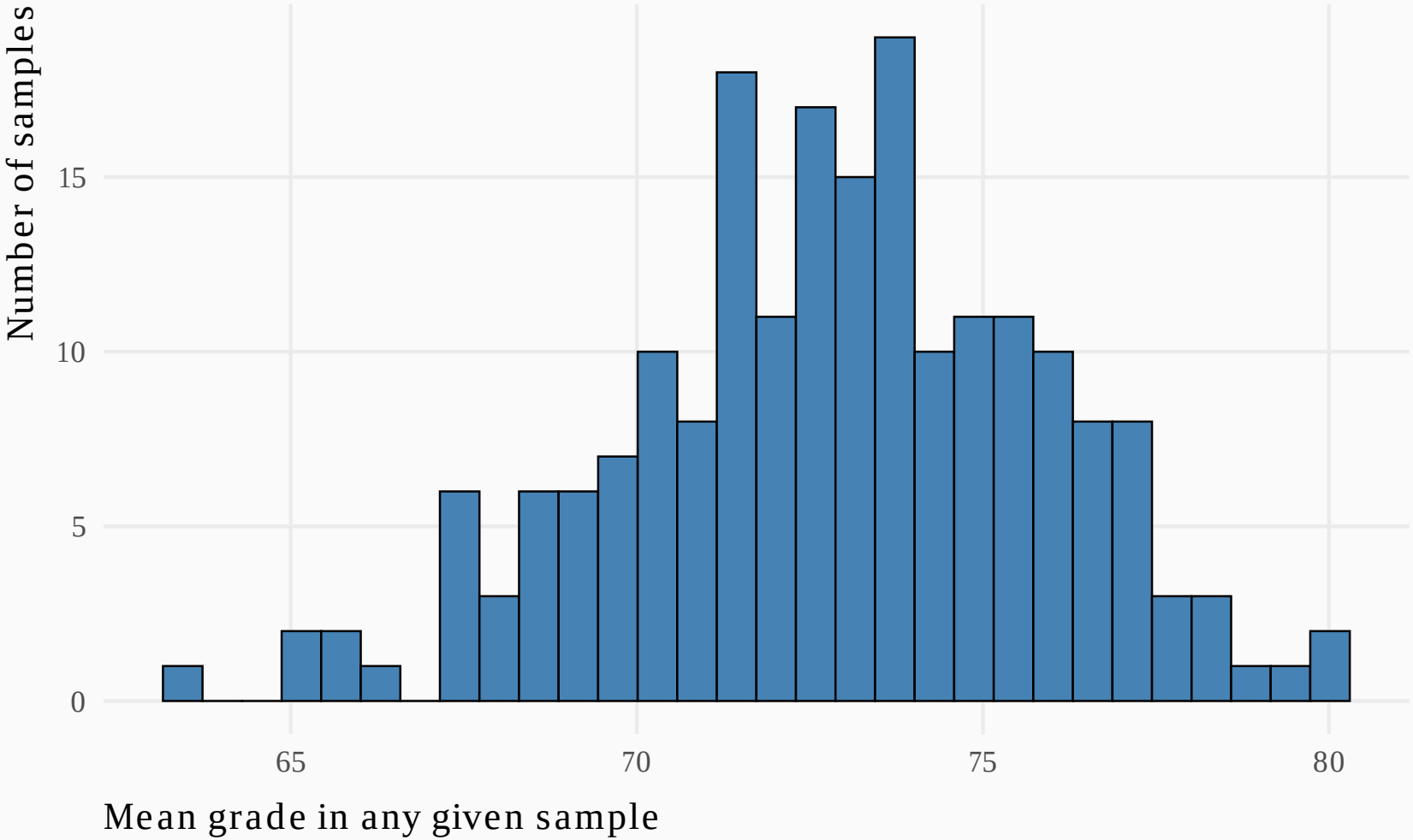
Hypothesis testing in our example

Assume that the null is true – i.e. the true mean is 73

- What do you expect to see when talking to your peers?
  - You expect to see have a sample mean of roughly 73!
  - It might be 71, it might be 75
    - Central limit theorem: the sampling distribution is normal and centered on the true population parameter
  - But you would be surprised to talk to 20 random students and learn that their mean grade is 59
  - The data would be inconsistent with the null hypothesis
  - At some point, the data is so inconsistent with the null hypothesis that we are comfortable rejecting it
    - How much we need to see before rejecting the null depends on the confidence level that we set

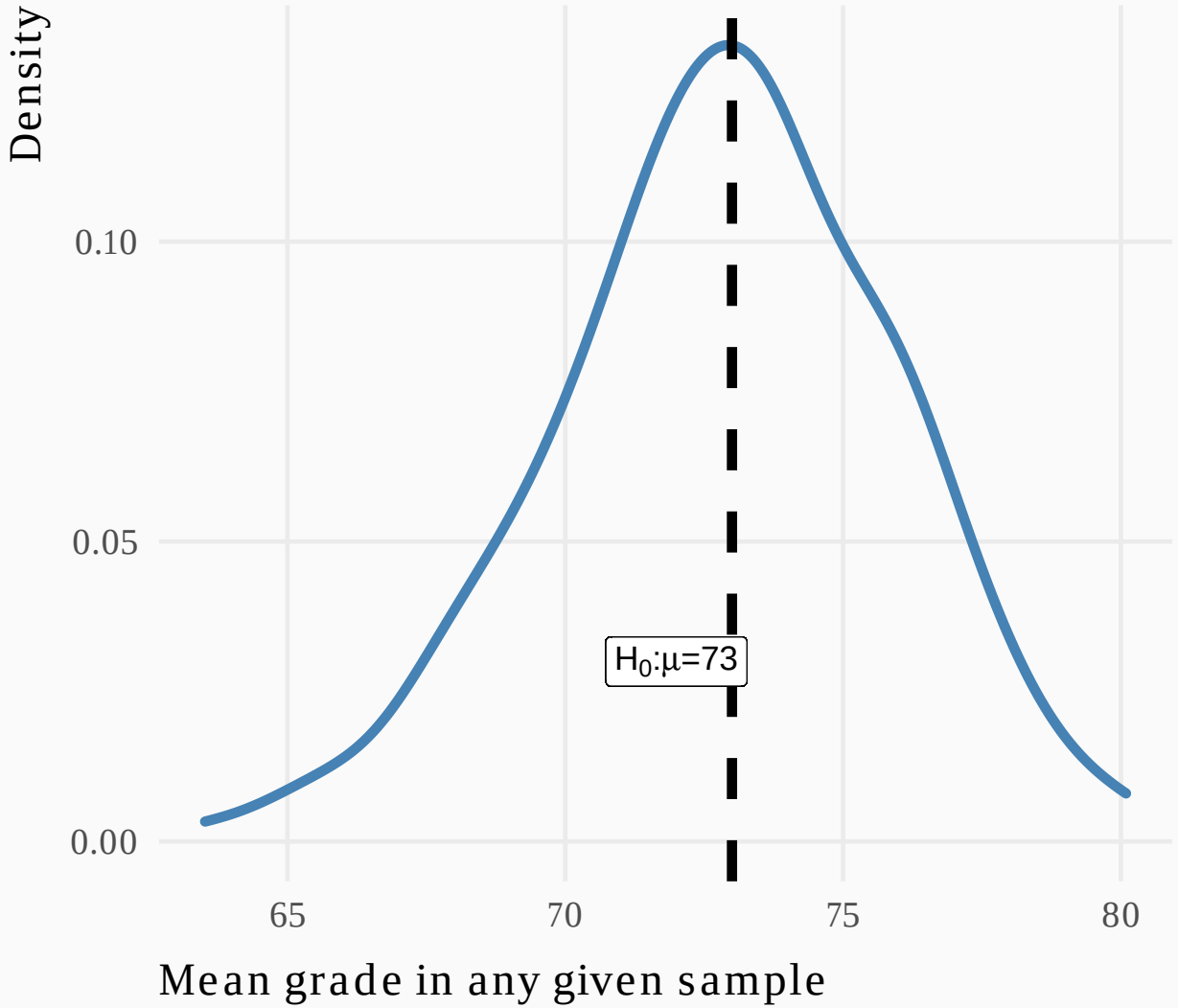
Sampling distributions

If the midterm average really is 73, the sampling distribution should look like this:



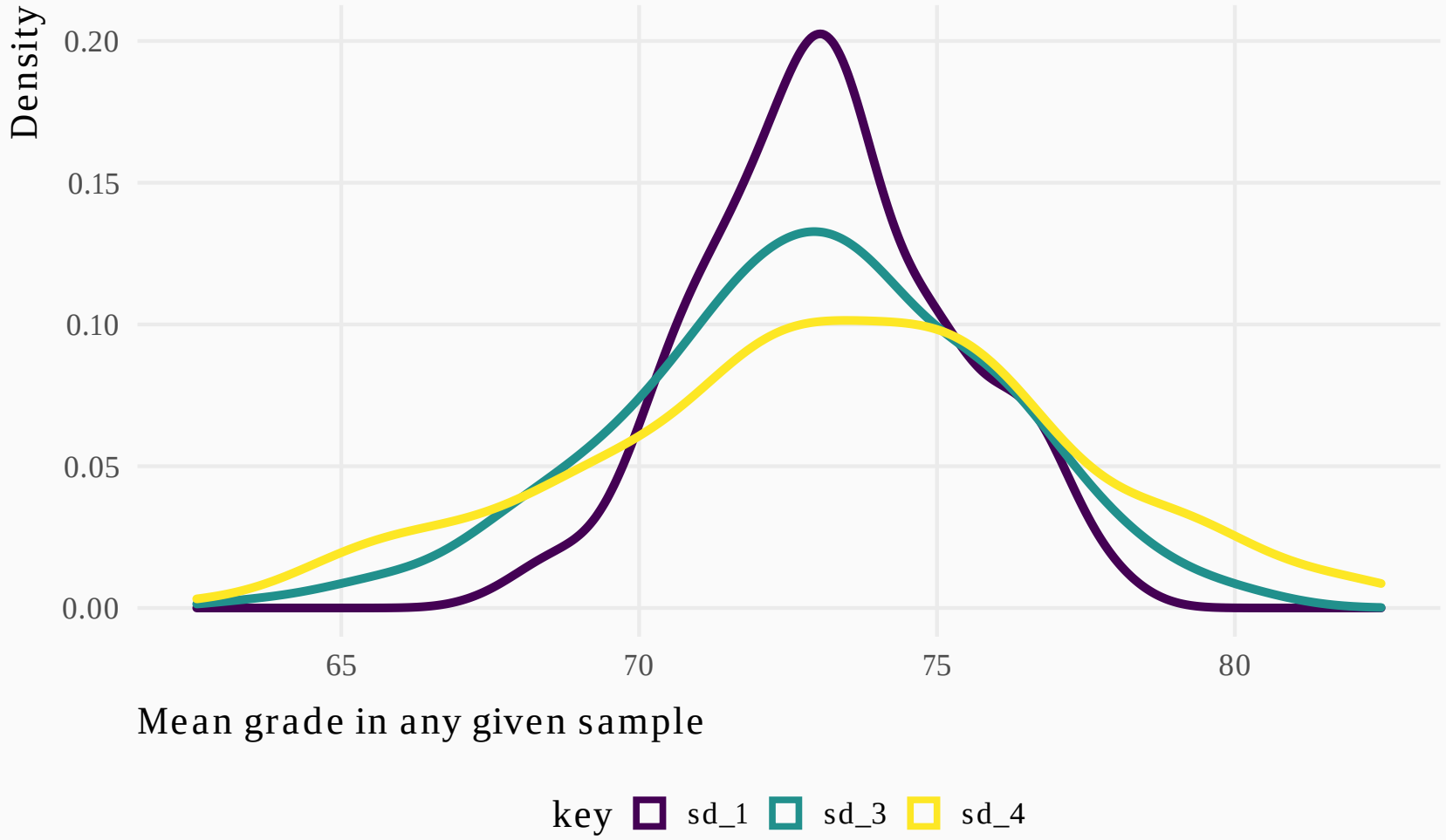
Sampling distributions

I can also show this using a density plot:



Sampling distributions with different SD

My sampling distribution may have a different standard deviation:



The sampling distribution and hypothesis

Whatever the particular SD of sampling distribution...

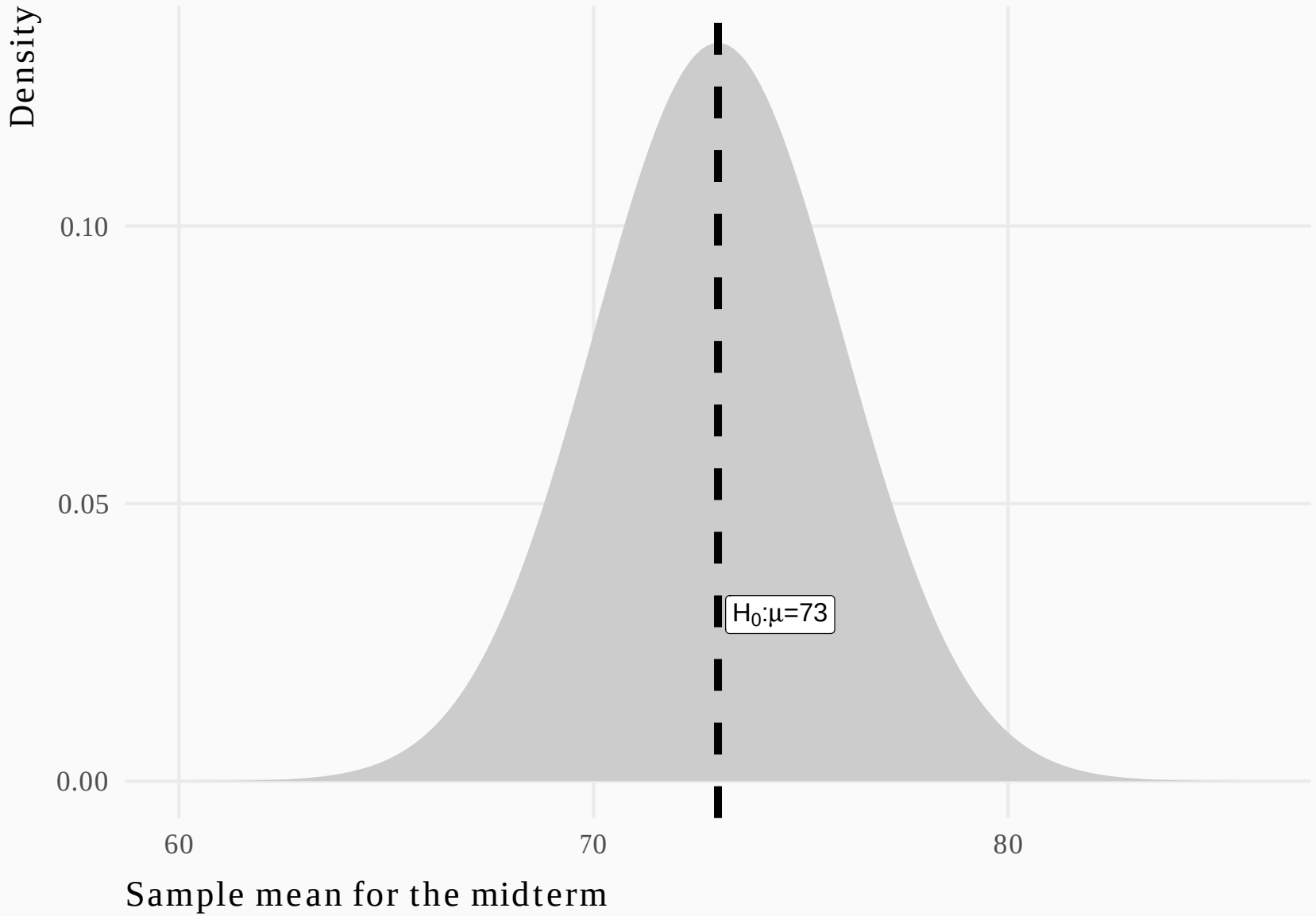
- It should approximate a normal distribution and be centered on the true parameter
- The key feature of a normal distribution:
  - About 68.4% of the data is within 1SD of the mean
  - About 95% of the data is within 2SD of the mean
  - About 99.7% of the data is within 3SD of the mean
- Therefore, if the null is true, I am...
  - Not surprised to observe a sample statistic that's 1SD away from the null
  - Surprised to observe a sample statistic that's 2 SDs away from the null
  - Very surprised to observe a sample statistic that is 3 SDs away from the null

What does my sampling distribution looks like?

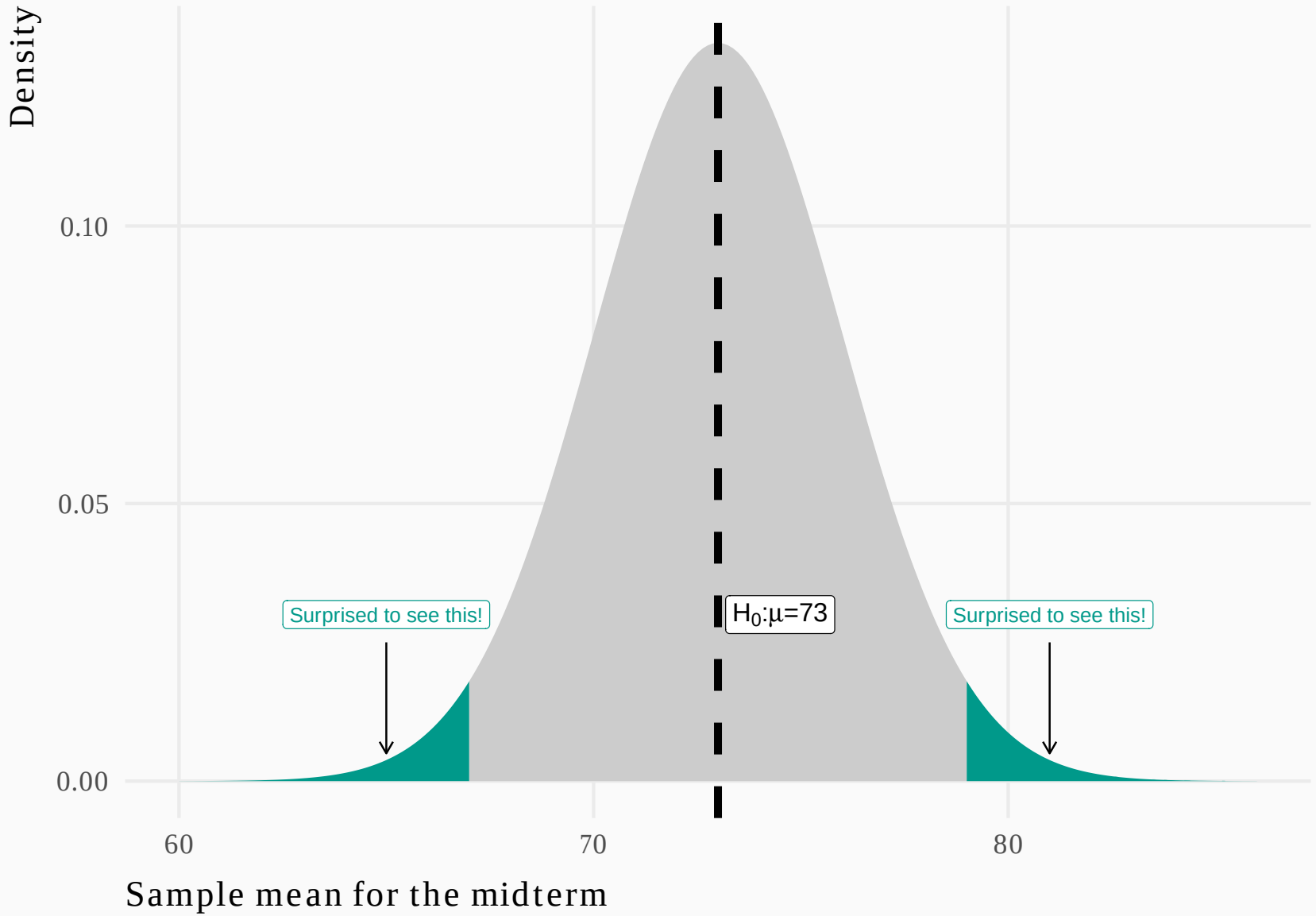
Remember that, in practice, we only draw a single sample

- We do not observe the sampling distribution
- But, the sampling distribution has 2 properties:
  - The mean
    - We will assume that the mean is equal to whatever the null hypothesis indicates
  - Standard deviation, for which we have a good guess:
    - $\hat{SE} = \frac{\hat{\sigma}}{\sqrt{n}}$
- With this in mind, we have a good idea of what the sampling distribution should look like if the null were true
  - And therefore we know how unlikely it is to have drawn the sample that we drew

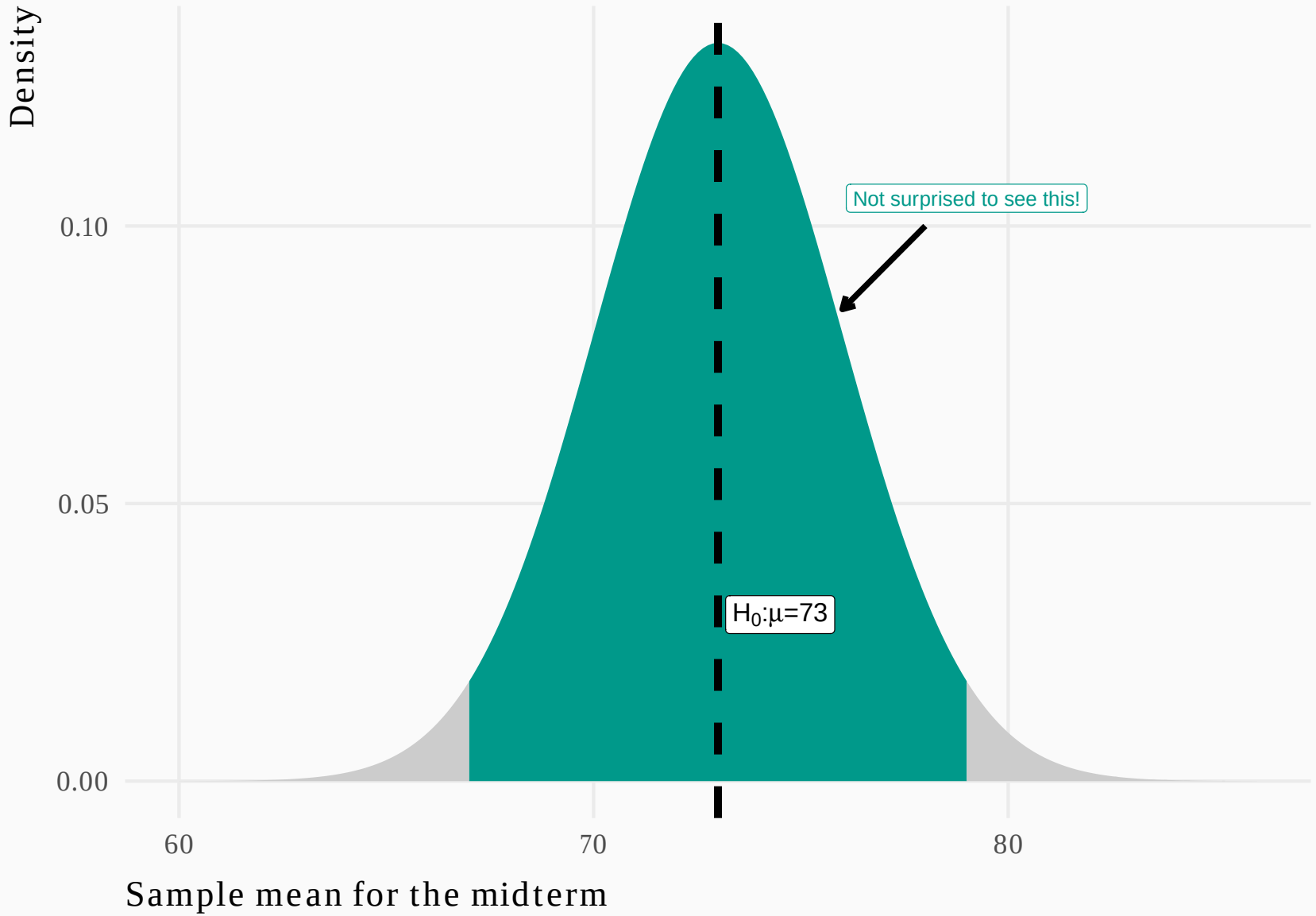
A hypothetical sampling distribution



A hypothetical sampling distribution



A hypothetical sampling distribution

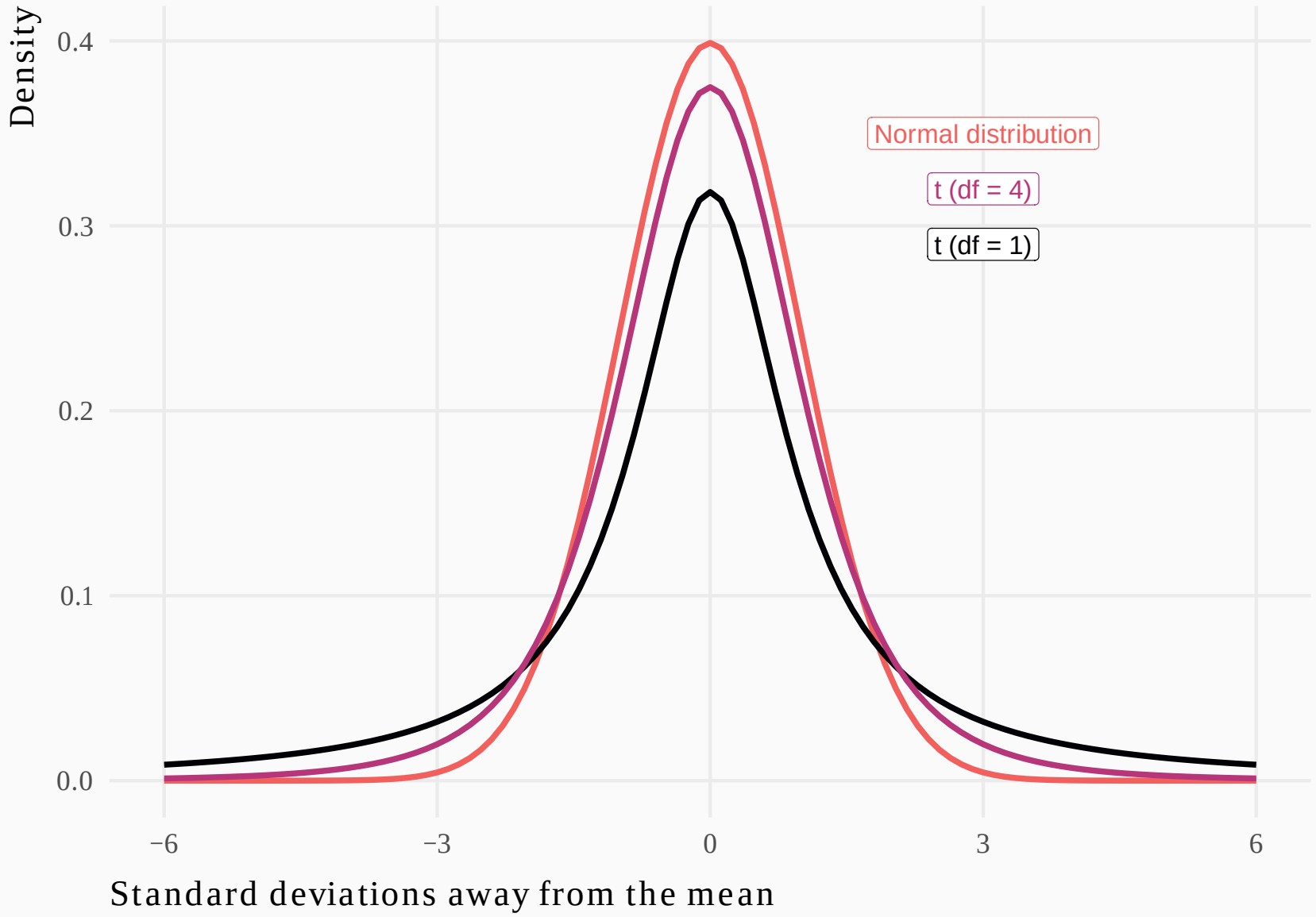


But what about small samples?

Any problem with the previous figure?

- If we draw a single outlying value, should we be surprised?
- Not enough for us to reject the null hypothesis! It's just a single value
- So what is the problem with the normal distribution?
  - It doesn't take into account sample size
- So instead, we'll use the **t-distribution**
  - It has an additional parameter: **degrees of freedom**
  - For our purposes, "degrees of freedom" refers to sample size
  - With a very high number of degrees of freedom, the t-distribution is just like the normal
  - With lower "df", the t-distribution has "fatter tails" $\rightsquigarrow$ higher probability of extreme values

The t-distribution



p-values

I now "know" what the sampling distribution would look like under the null:

- I know where it peaks (at the null hypothesis, e.g. H_0: $\mu = 73$)
- I "know" its standard deviation by estimating the standard error
  - $SE = \frac{\hat{\sigma}}{\sqrt{n}}$
- I know the "degrees of freedom" parameter (the sample size)

The next step: how likely is the data I observe, if the null is true?

- If $\mu_{\text{true}} = 73$, I'm not surprised to draw a sample with mean - 73
- At some point, the data I observe is so unlikely to have been produced by sampling from a population with $\mu_{\text{true}} = 73$ that I must reject the null

p-values

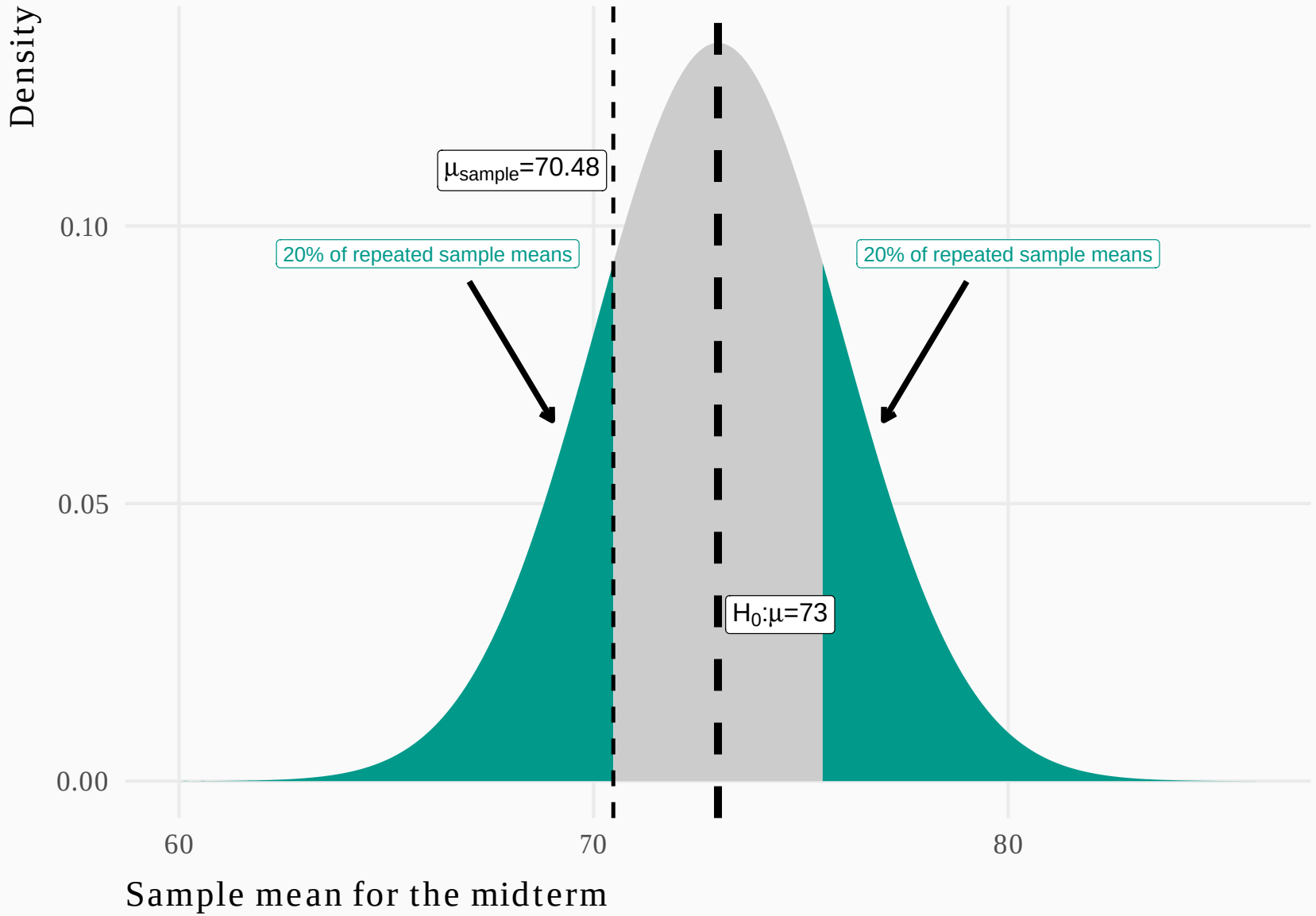
- Even if I draw a sample that's far from H_0, it's possible I drew a weird sample by chance
- e.g. it is possible to draw 20 random students with $\mu_{\text{grade}} = 62$ even if the true mean is 73
- But there is a point where it's so unlikely that I'm comfortable rejecting the null
    - This is our prespecified **significance level** (often $\alpha = 0.05$)
- When looking at our data, we can compute a **p-value**
    - The p-value is a number between 0 and 1
    - It represents the expected probability of observing the sample data, if the null hypothesis were true
    - p-value close to 1: given the null, we're not surprised to see this $\rightsquigarrow$ fail to reject the null
    - p-value close to 0: given the null, we're surprised to see this $\rightsquigarrow$ reject the null
    - $p < \alpha$: reject the null; $p > alpha$: fail to reject the null

Interpreting p-values

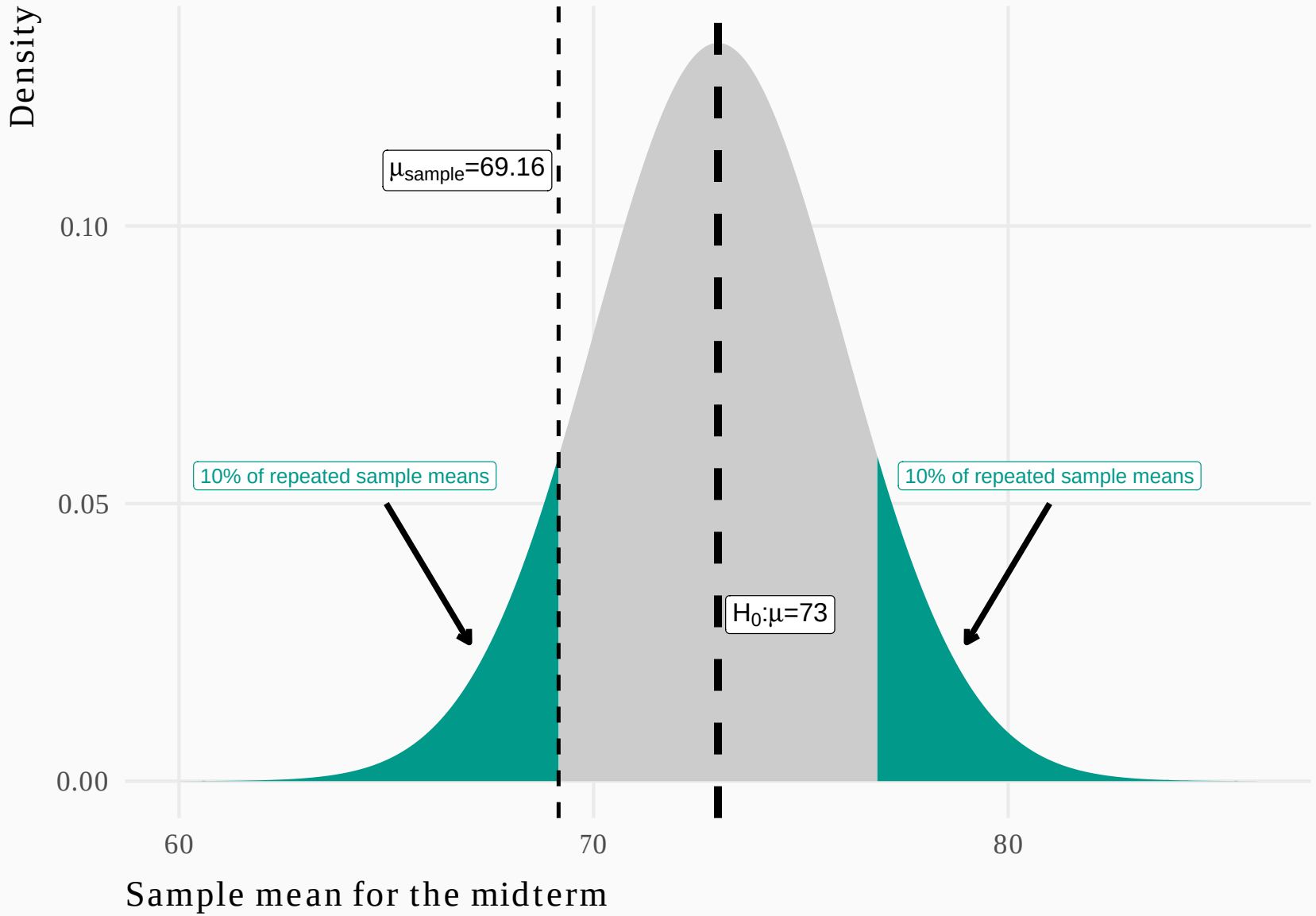
Let's say I randomly sample students and compute a mean grade of 67

- H_0: $\mu_{\text{true}} = 73$
- Let's say I get a p-value of 0.13; what I can say:
  - If I were to repeatedly sample from our population (students who took the midterm)...
  - I would expect to get a result as "extreme" as this (extreme = far away from the null hypothesis)...
  - In about 13% of repeated samples...
  - If the null hypothesis is true
- In other words: it's somewhat unlikely, but very much possible
- With $\alpha = 0.05$, we fail to reject the null that the mean is 73
  - Can we conclude that the mean is 73?
    - NO! We do NOT "accept" the null; we "fail to reject"
  - There is no **statistically significant** difference between our sample mean and the null hypothesis

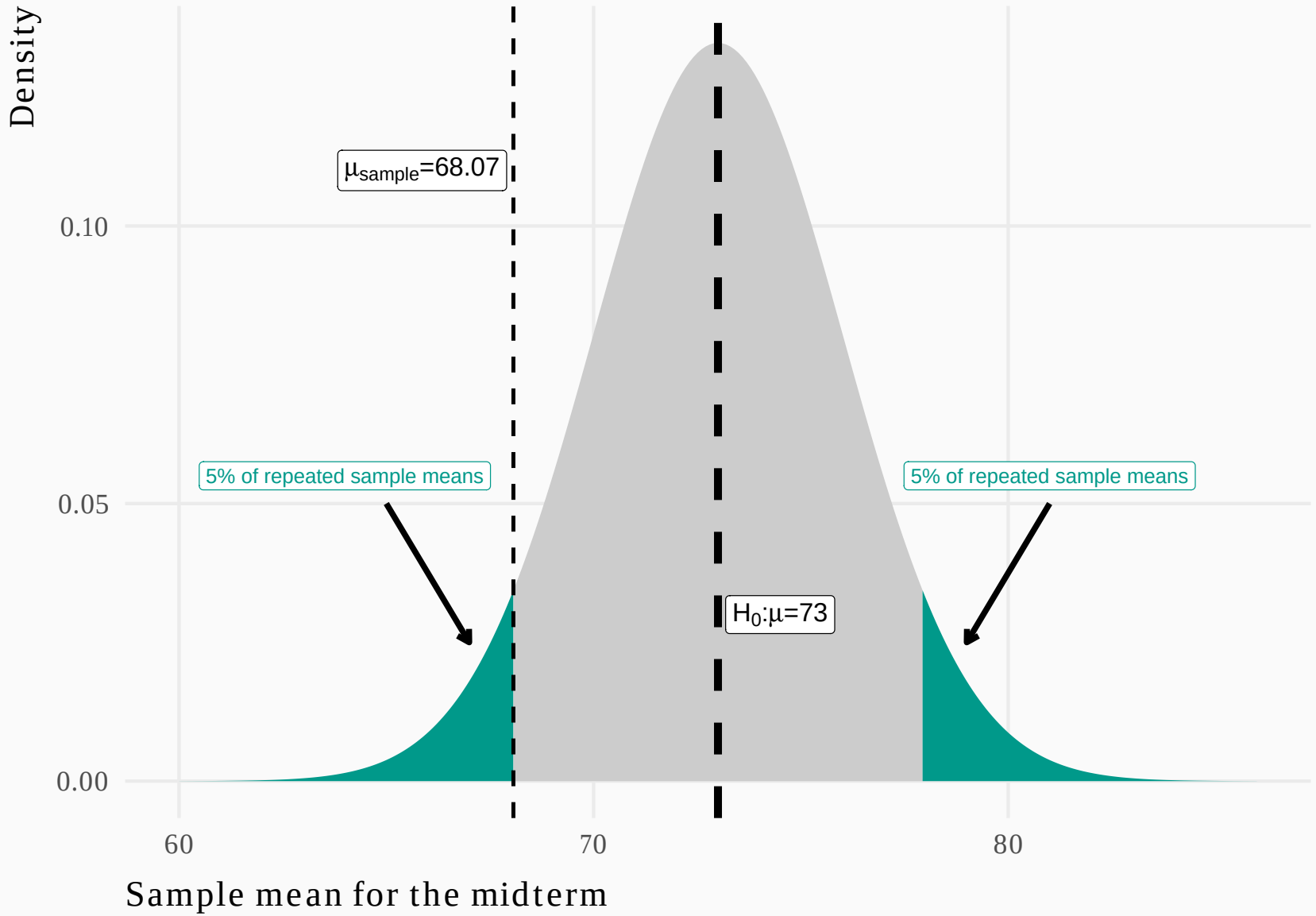
p = 0.4 (with non-directional hypothesis)



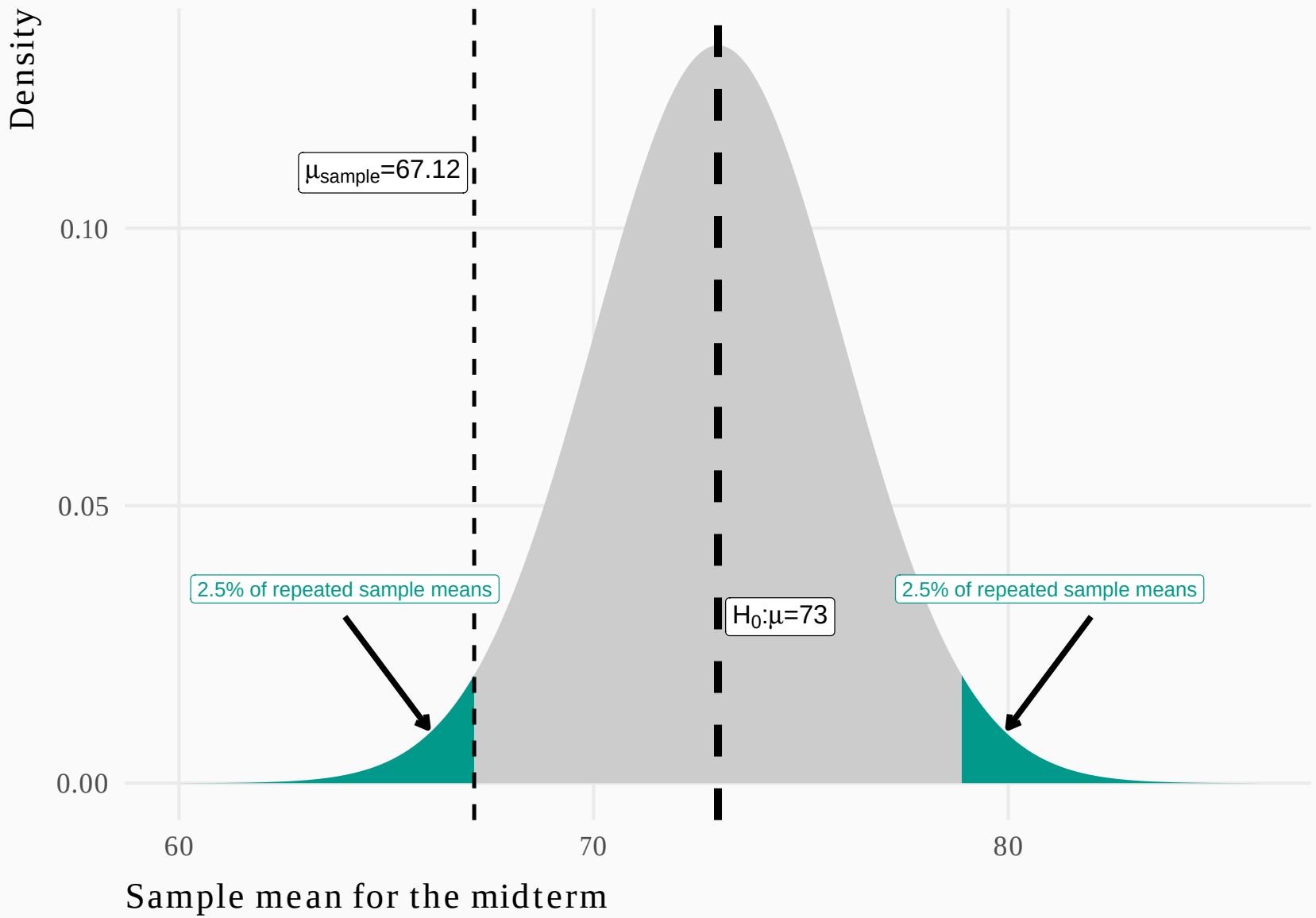
p = 0.2 (with non-directional hypothesis)



p = 0.1 (with non-directional hypothesis)



p = 0.05 (with non-directional hypothesis)



One-sample t-test in R

```r
# the hypothetical grades I gave you earlier
grades <- c(63, 67, 71, 56, 77, 47)
t.test(grades, mu = 73)
```

```
##
##  One Sample t-test
##
## data:  grades
## t = -2.1615, df = 5, p-value = 0.08303
## alternative hypothesis: true mean is not equal to 73
## 95 percent confidence interval:
##  52.20211 74.79789
## sample estimates:
## mean of x
##      63.5
```

Interpret the confidence interval and the p-value

- Should you call me a liar?

When should you have called me a liar?

Table 7: Grades of students you meet in lab

Student #	Grade	p_value
1	63	NA
2	67	0.156
3	71	0.122
4	56	0.072
5	77	0.156
6	47	0.083
7	55	0.034

You can call me a liar when you get the 7th data point!

- (Assuming $\alpha = 0.05$)

Type I and Type II errors

When p is very small, we're very surprised by the data we're seeing

- But weird samples happen!
- It's not impossible that the null is true given the data; it's just very unlikely

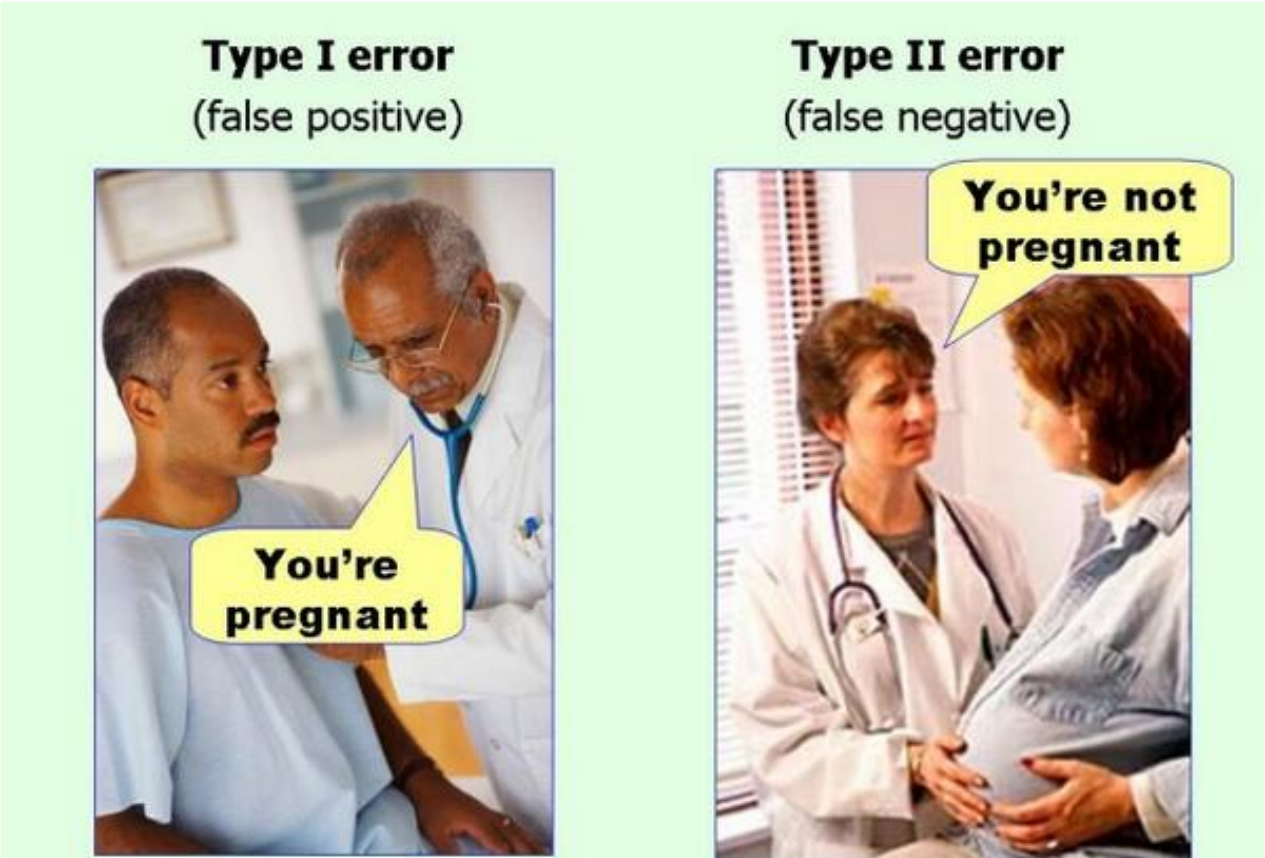
Take the example above

- If 100 of you talk to peers and ask about their midterm grade
- Each person sets $\alpha = 0.05$
- 5 people will accuse me of lying even if the true mean is 73
    - i.e. they will draw data that is inconsistent with what I said, even if what I said is true
- This is **Type I error**: I reject the null when the null is actually true
    - Also known as a **false positive**
    - By setting a lower α, I reduce the chances of Type I errors

Type I and Type II errors

Type II error is the opposite

- You fail to reject the null when the null is actually not true
- Also known as a **false negative**
- By setting a lower α, you increase the chances of Type II error



Differences-in-means

I just presented an example of one hypothesis test where we examined the mean of a variable against some null hypothesis

- Hypothesis tests can be conducted for many different hypotheses
- Another important application: differences-in-means
- Remember what we did in assignment 2 (causality)?

```
druckman <- read_csv("lectures/lecture_11.1/druckman_2003.csv")
druckman %>%
  group_by(tv) %>%
  summarise(who_won = mean(won2,na.rm = T) %>% round(3))
```

```
## # A tibble: 2 x 2
##      tv who_won
##   <dbl>   <dbl>
## 1     0   0.38
## 2     1   0.262
```

The setup

Again, we have a null hypothesis; what is it?

The setup

Again, we have a null hypothesis; what is it?

- H_0: $\mu_1 = \mu_2$

The setup

Again, we have a null hypothesis; what is it?

- H_0: $\mu_1 = \mu_2$

And we have an alternative hypothesis: H_1: $\mu_1 \neq \mu_2$

If the null is true, what do we expect to see?

- If we draw many repeated samples...
- And compute the difference-in-means for each...
- The sampling distribution should be centered on 0

And again, depending on how surprising the data is given the null, we decide to reject the null or fail to reject it

The difference-in-means in R

```
t.test(druckman$won2[druckman$tv==0], druckman$won2[druckman$tv==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  druckman$won2[druckman$tv == 0] and druckman$won2[druckman$tv == 1]
## t = 3.4387, df = 166.76, p-value = 0.0007382
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05022681 0.18565360
## sample estimates:
## mean of x mean of y
## 0.3798450 0.2619048
```

The difference-in-means is different from 0 in a **statistically significant** manner

The difference-in-means in R

```r
# equivalent to the above
t.test(druckman$won2 ~ druckman$tv) # mu = 0 is the default
```

```
##
##  Welch Two Sample t-test
##
## data:  druckman$won2 by druckman$tv
## t = 3.4387, df = 166.76, p-value = 0.0007382
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
## 95 percent confidence interval:
##  0.05022681 0.18565360
## sample estimates:
## mean in group 0 mean in group 1
##       0.3798450       0.2619048
```

How to interpret the difference-in-means

Are differences-in-means causal quantities?

- Well, it depends!
- If they're means from experimental conditions $\rightsquigarrow$ causal interpretation
- If not, it's probably hard to interpret them causally
- But they're still interesting!

The dangers of hypothesis testing

Null hypothesis statistical testing is, by far, the dominant approach

- But it is easy to misinterpret what our statistical tests are saying
- Much discussion recently in the scientific community!

Political Analysis

Political Analysis
@polanalysis

Political Analysis will no longer report p values in regression tables or elsewhere. There are many reasons for this change—most notably that a p value alone does not give evidence in support of a given model or the associated hypotheses. See Editorial in Issue 26.1 for more info

10:15 AM · Jan 22, 2018 · Twitter Web Client

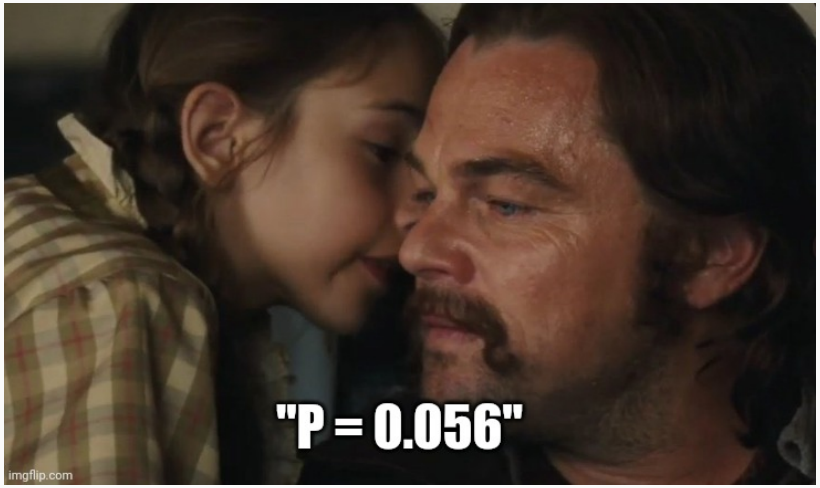
326 Retweets **159** Quote Tweets **469** Likes

“p-hacking”

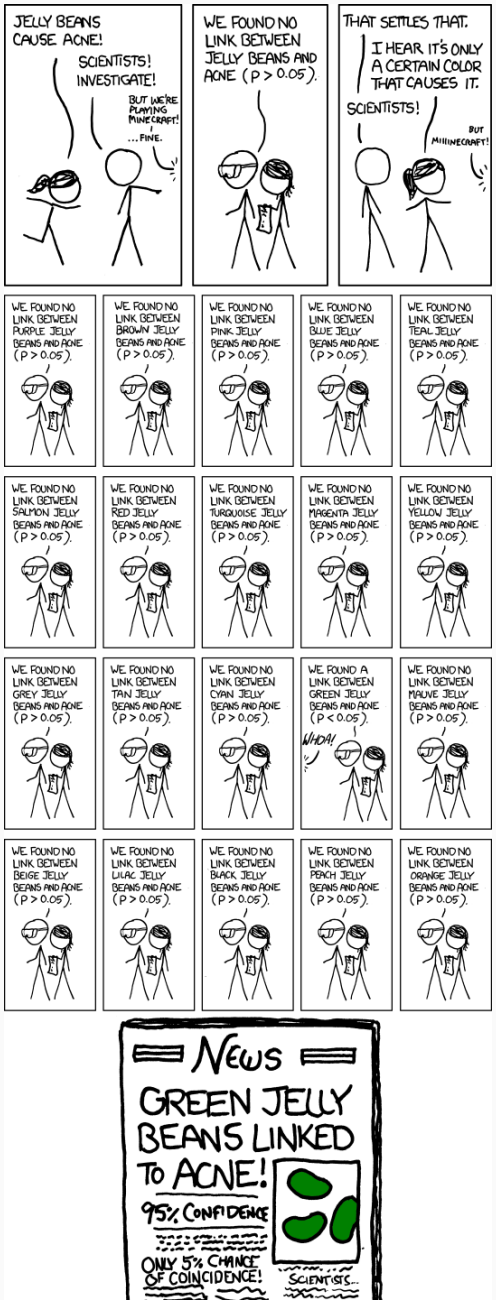
We’ve seen that there a (completely arbitrary) threshold below which results are considered “statistically significant”

- Publication is much easier if you achieve statisticaly significance
- Incentive: play around with data until you achieve $p < 0.05$
    - Play around: add/remove control variables, remove observations, use alternative measures...
- Called: p-hacking, researcher degrees of freedom, garden of forking paths...
- This is a **widespread** problem that we are just starting to grapple with
- But when you think about it...
    - Are you really more certain of your result if $p = 0.049$ compared to $p = 0.051$?

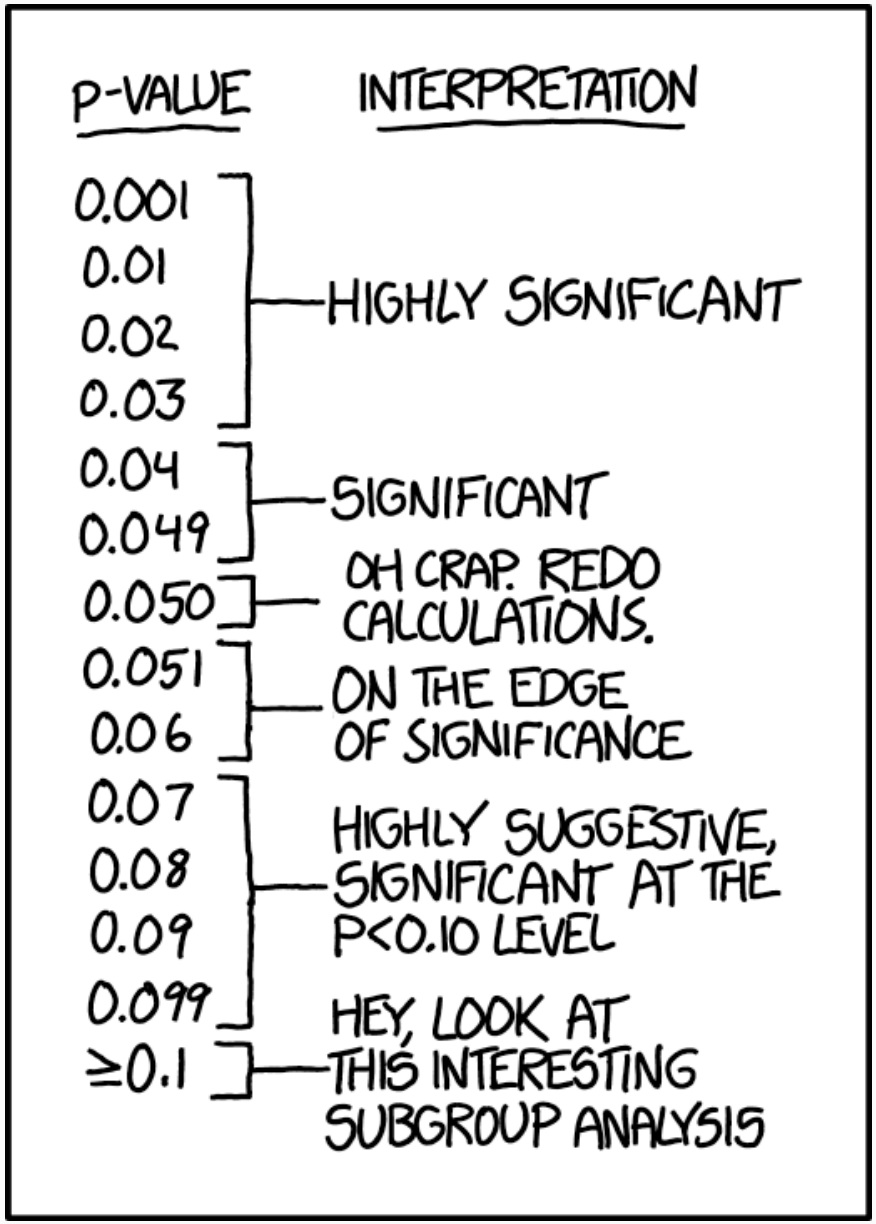
Wrong side of the arbitrary threshold



Jelly beans and acne



Significance



Significance as a dummy

Your results are either statistically significant or they're not

- If p = 0.06 and you set $\alpha = 0.05$, your finding is not statistically significant
  - What should we do? More research!
  - Hopefully with a larger sample that will be able to detect an effect

You will often see papers that ignore this because they want significant results (remember publication bias?)

- "approaches significance", "marginally significant"...
- For a longer (and hilarious) list, see here
- "As well as being statistically flawed (results are either significant or not and can't be qualified), the wording is linguistically interesting, often describing an aspect of the result that just doesn't exist. For example, "a trend towards significance" expresses non-significance as some sort of motion towards significance, which it isn't: there is no 'trend', in any direction, and nowhere for the trend to be 'towards'."

Statistical vs substantive significance

Above all, **we may not care about a statistically significant finding**

Statistical significance $\neq$ substantive significance

It is possible to have a statistically significant difference that is substantively not meaningful

- e.g. a large survey (60,000) shows that mean happiness for Facebook users is 7.64 on 1-10 scale and 7.68 for non-Facebook users
- Given the sample size, we may find a statistically significant different difference, e.g. $p < 0.01$
- But do we actually care?
    - **Substantive** significance: does it pass the "so what" test?

Why should I care?

Real-world decisions and our understanding of the world depend on our interpretation of ours results

- And our interpretation depends on whether we have statistical significance

« On vient de fournir à la planète un espoir ! s'exclame au bout du fil le Dr Jean-Claude Tardif, chercheur principal de l'étude COLCORONA et directeur du centre de recherche de l'Institut de cardiologie de Montréal (ICM). On a finalement un premier traitement qui peut aider les patients atteints de la COVID-19 avant leur admission à l'hôpital pour prévenir les hospitalisations, prévenir les intubations et prévenir les décès. »

Chez 4159 patients qui présentaient un facteur de risque de complications et dont le diagnostic de COVID-19 avait été validé par un test PCR, la colchicine a entraîné une baisse des hospitalisations de 25 %, une baisse du besoin de ventilation de 50 % et une diminution des décès de 44 % par rapport au groupe témoin. « C'est une percée majeure », déclare le Dr Tardif.

Why should I care?

Table 2. Rates and Odds Ratios for Major Clinical Outcomes.

Clinical Outcome	Colchicine	Placebo	Odds Ratio (95% CI)	P Value
ITT population	N=2235	N=2253		
Primary composite endpoint - no. (%)	104 (4.7%)	131 (5.8%)	0.79 (0.61-1.03)	0.08
Components of primary endpoint:				
Death - no. (%)	5 (0.2%)	9 (0.4%)	0.56 (0.19-1.67)	
Hospitalization for COVID-19 no. (%)	101 (4.5%)	128 (5.7%)	0.79 (0.60-1.03)	
Secondary endpoint:				
Mechanical ventilation - no. (%)	11 (0.5%)	21 (0.9%)	0.53 (0.25-1.09)	

References i