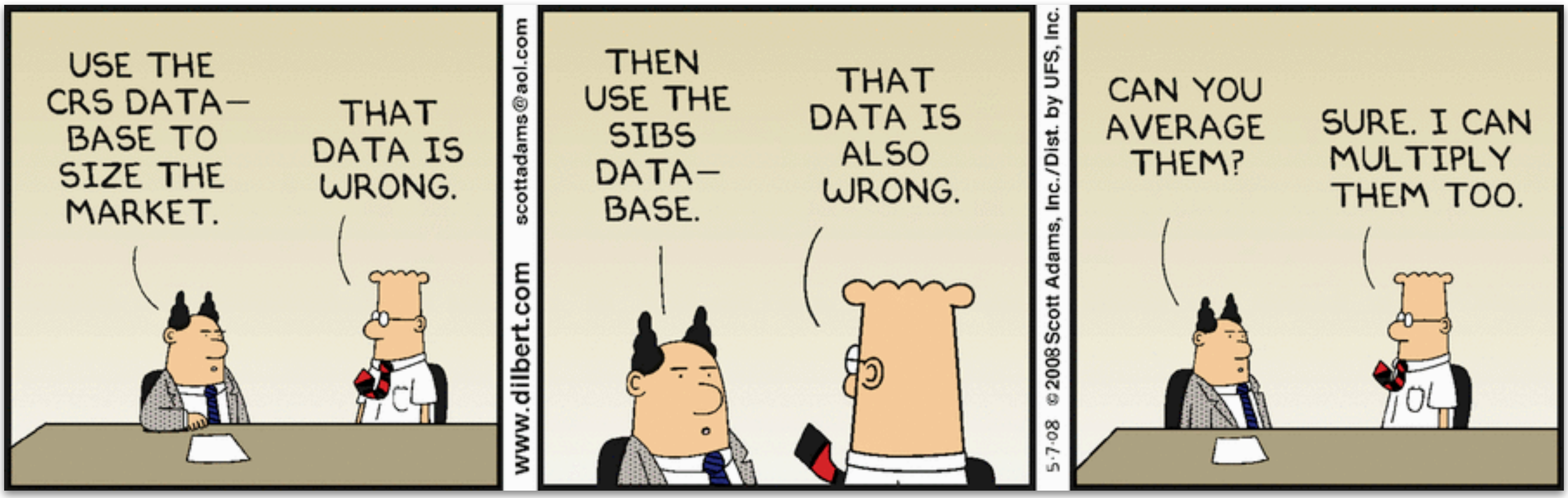
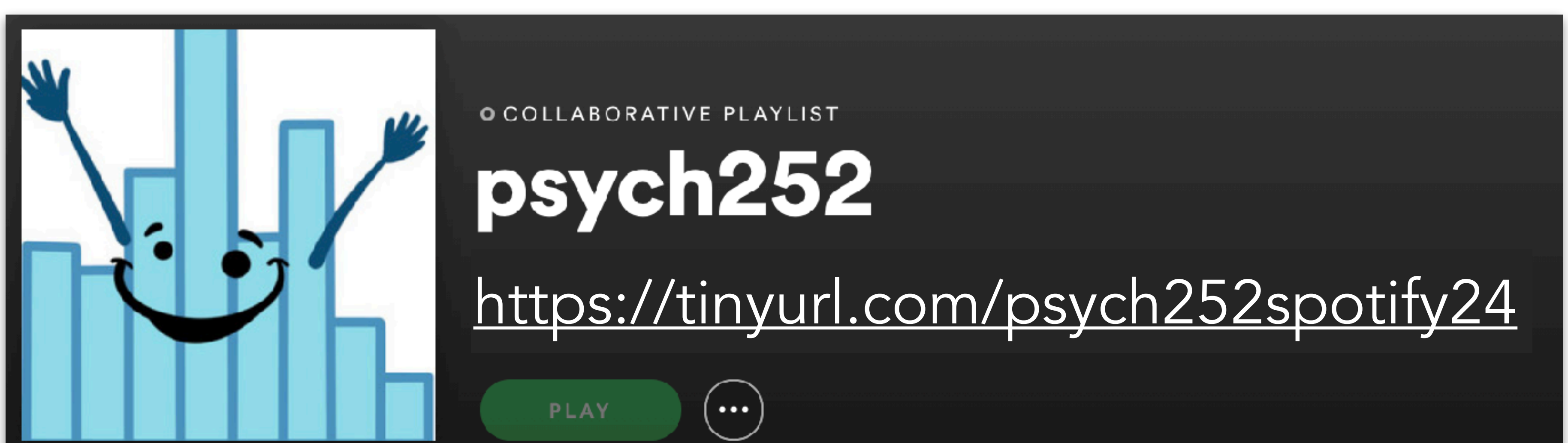
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

Simulation 2





01/26/2024

Logistics

Section on Mondays

Will from now on take place in **room 245** in **building 420.**

Where and when?

The meetings will be in person and as shown below.

Lectures: The class meets Monday, Wednesday, and Friday **10:30-11:50am** in 200-203 (Lane History Corner).

Sections: Sections are on Monday **3:30-4:20pm** in 420-245 and on Friday **12:30-1:20pm** in McMurtry Art Building 350 (attendance is optional).

updated on the website

Final presentations

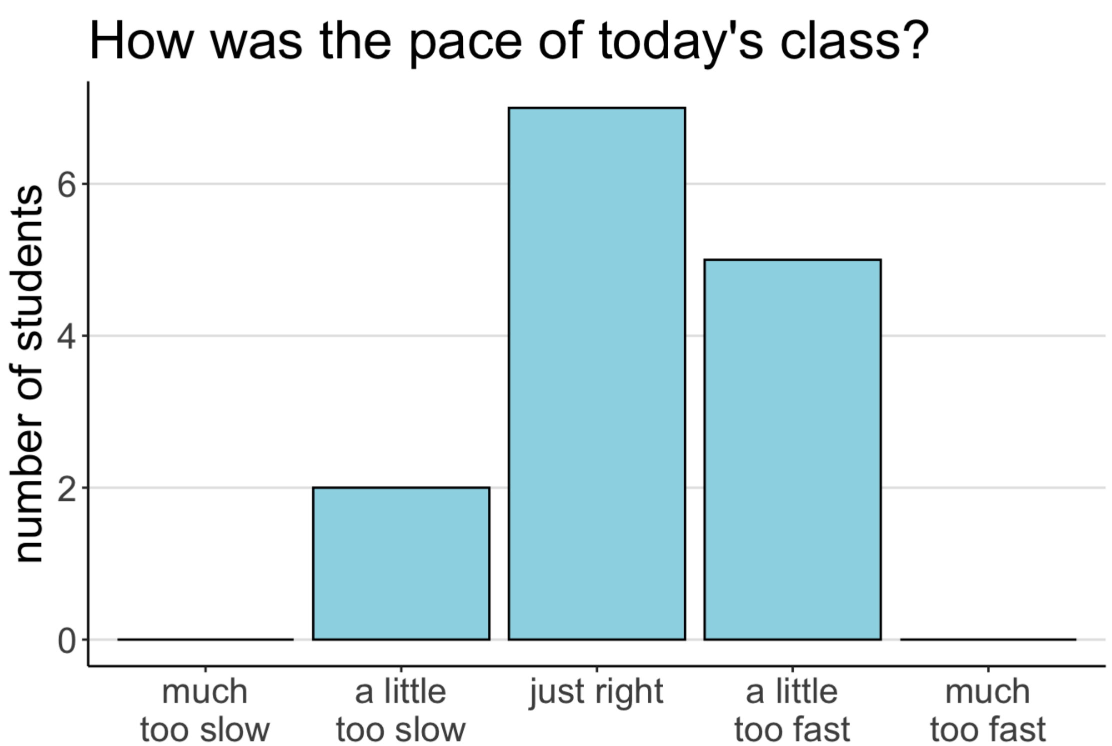
Will take place on **Monday, March 18th** from **3.30pm to 6.30pm** in Building 200, room 203 (our **usual room**).

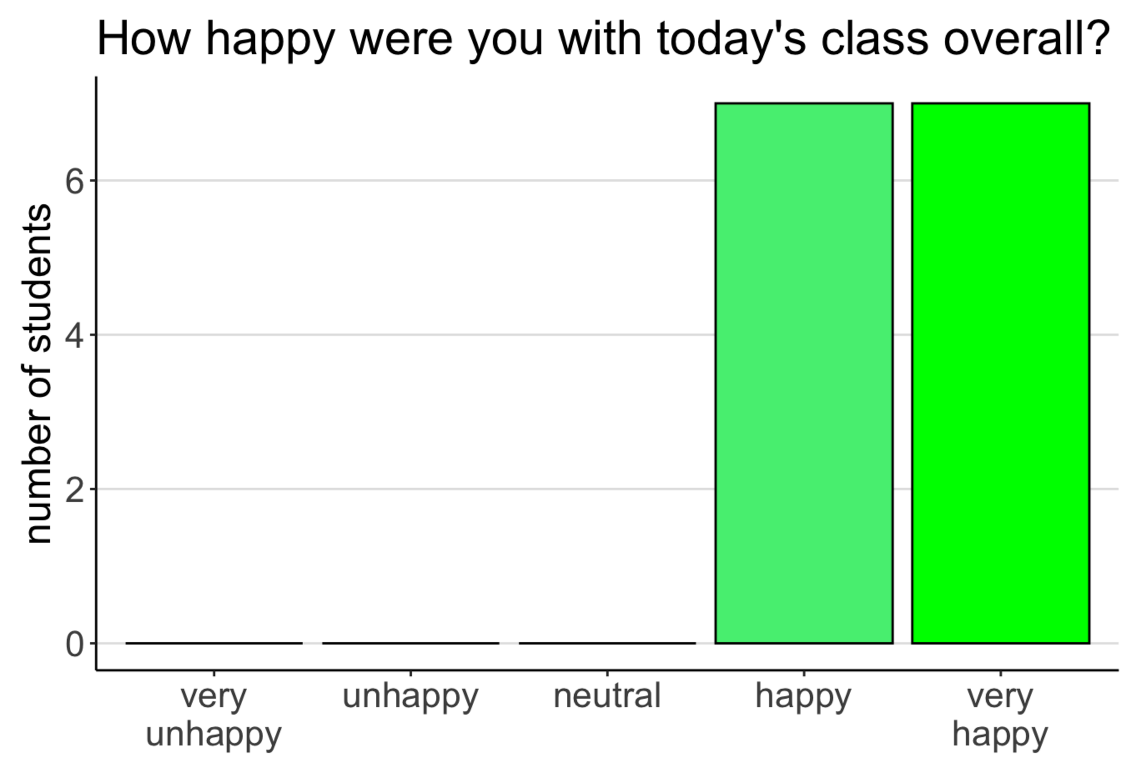
Due dates

- Thursday, January 18th: Homework 1
- Thursday, January 25th: Homework 2
- Thursday, February 1st: Homework 3
- Thursday, February 8th: Homework 4
- Thursday, February 15th: Midterm
- Thursday, February 22nd: Project proposal
- Thursday, February 29th: Homework 5
- Thursday, March 7th: Homework 6
- Thursday, March 14th: Homework 7 (optional)
- Monday, March 18th: Final project presentation
- Friday, March 22nd: Final project report

Your feedback

Your feedback





Really appreciated the recap of last class! It’s nice to get a refresher and to get to hear material twice.

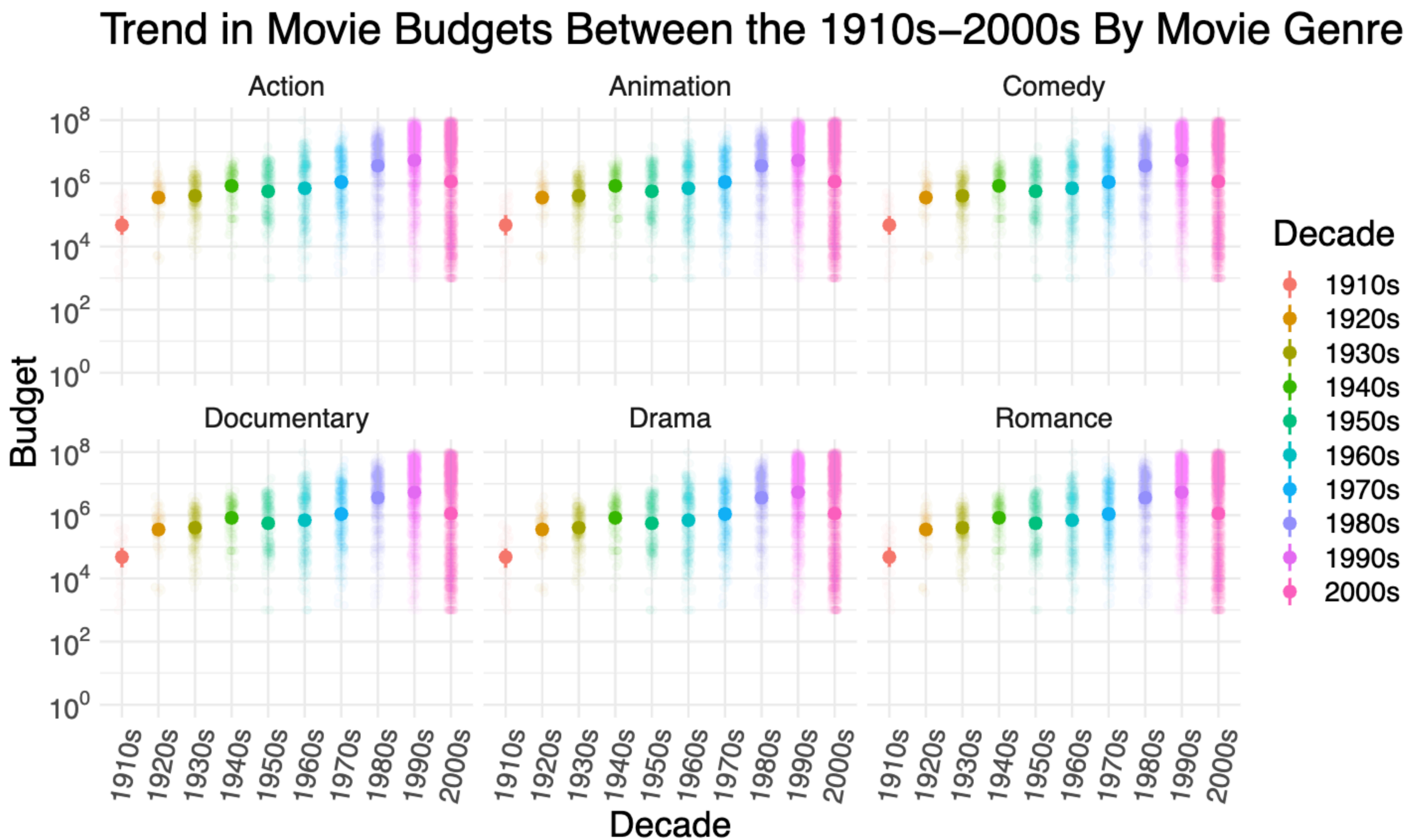
Today's class felt less rushed than the last one which I appreciated. The slide animations were helpful for understanding density!

I want the latest r-markdown file that you showed us in the class to be shared that we can see the updated code from the original html.

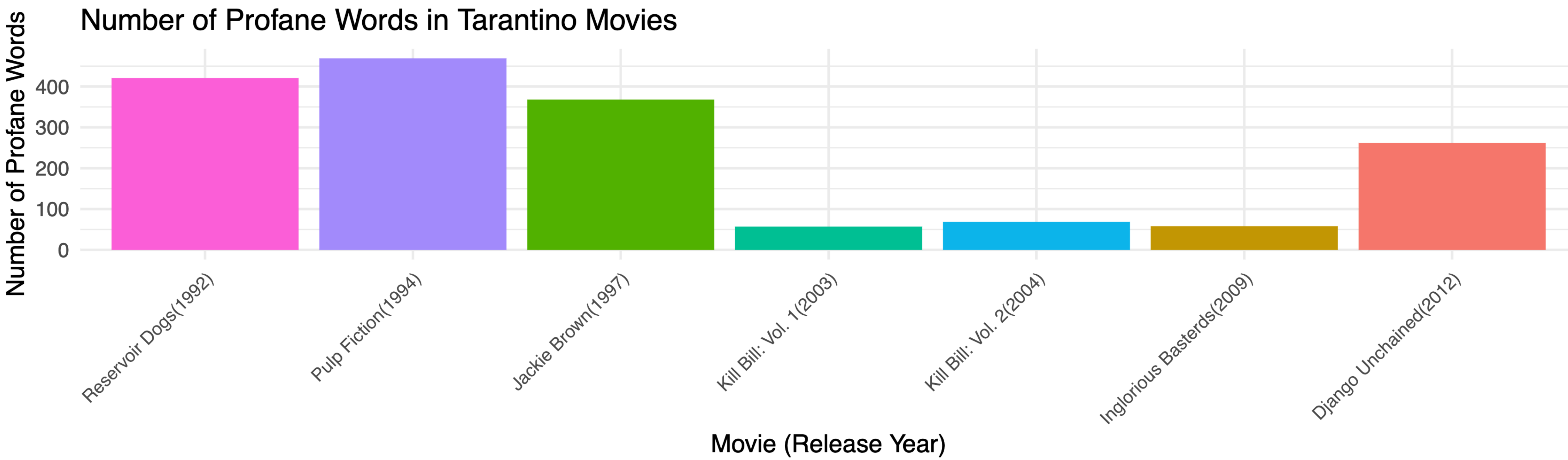
Homework

Homework 1
Show case

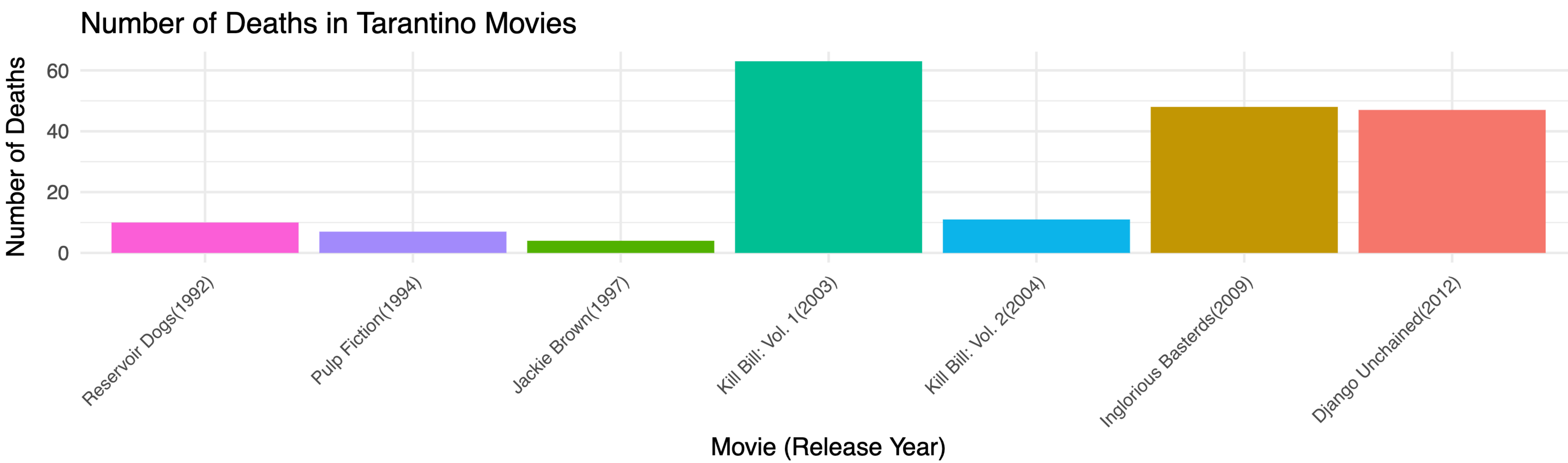
Adani Bennett Abutto



Irmak Ergin

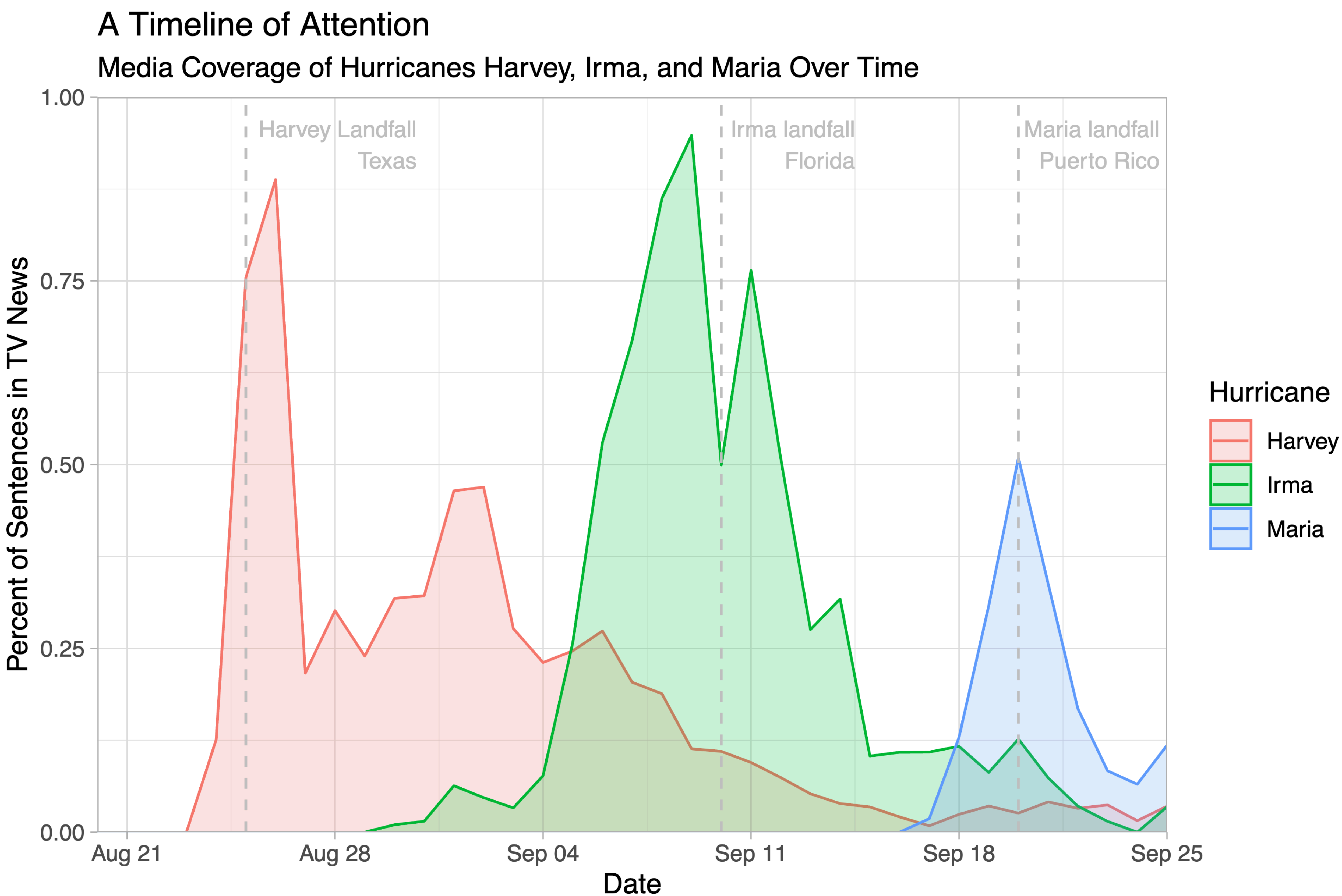


Number of Profane Words in Tarantino Movies
Number of Profane Words
400
300
200
100
0
Reservoir Dogs(1992)
Pulp Fiction(1994)
Jackie Brown(1997)
Kill Bill: Vol. 1(2003)
Kill Bill: Vol. 2(2004)
Inglorious Basterds(2009)
Django Unchained(2012)
Movie (Release Year)

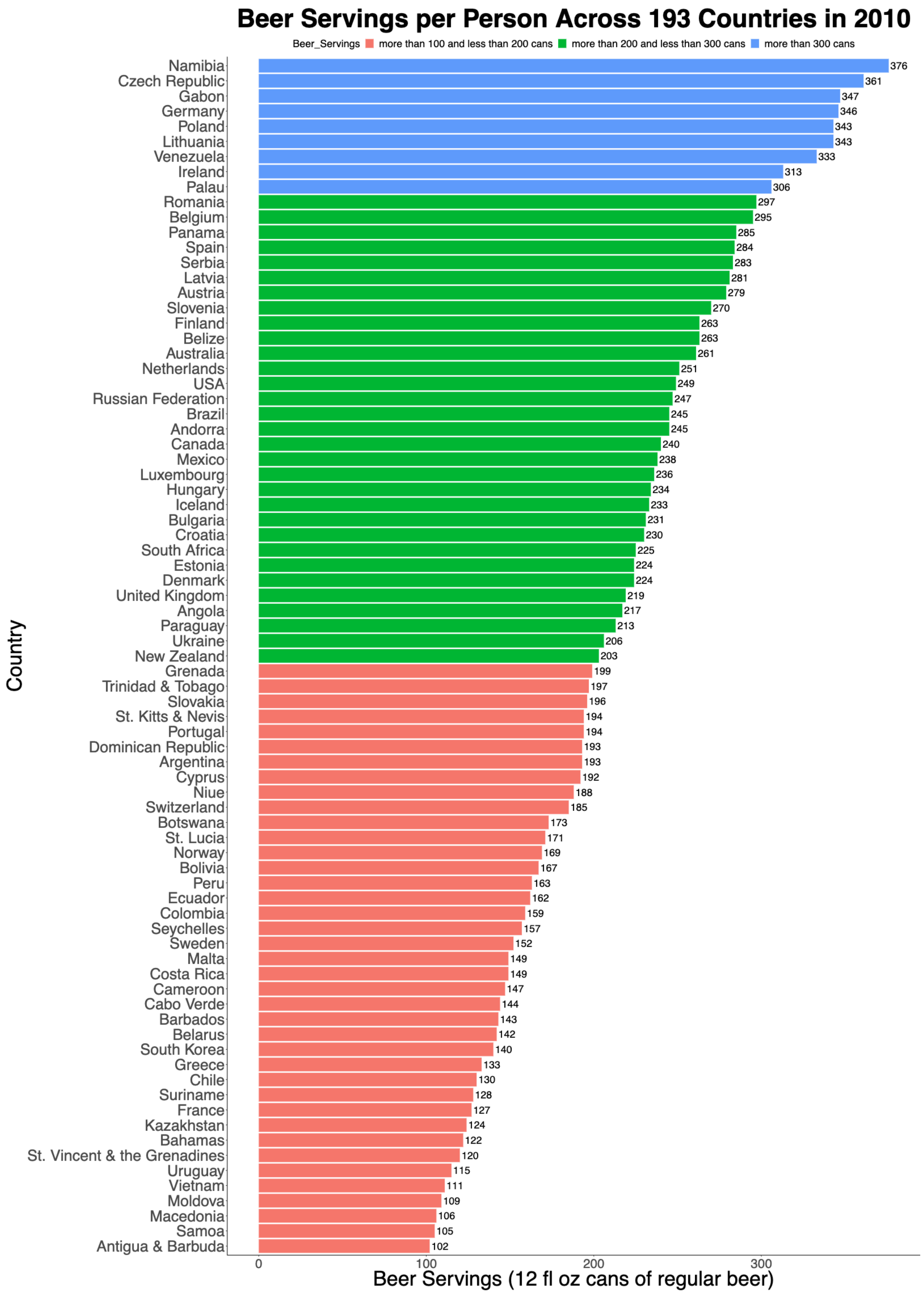


Number of Deaths in Tarantino Movies
Number of Deaths
60
40
20
0
Reservoir Dogs(1992)
Pulp Fiction(1994)
Jackie Brown(1997)
Kill Bill: Vol. 1(2003)
Kill Bill: Vol. 2(2004)
Inglorious Basterds(2009)
Django Unchained(2012)
Movie (Release Year)

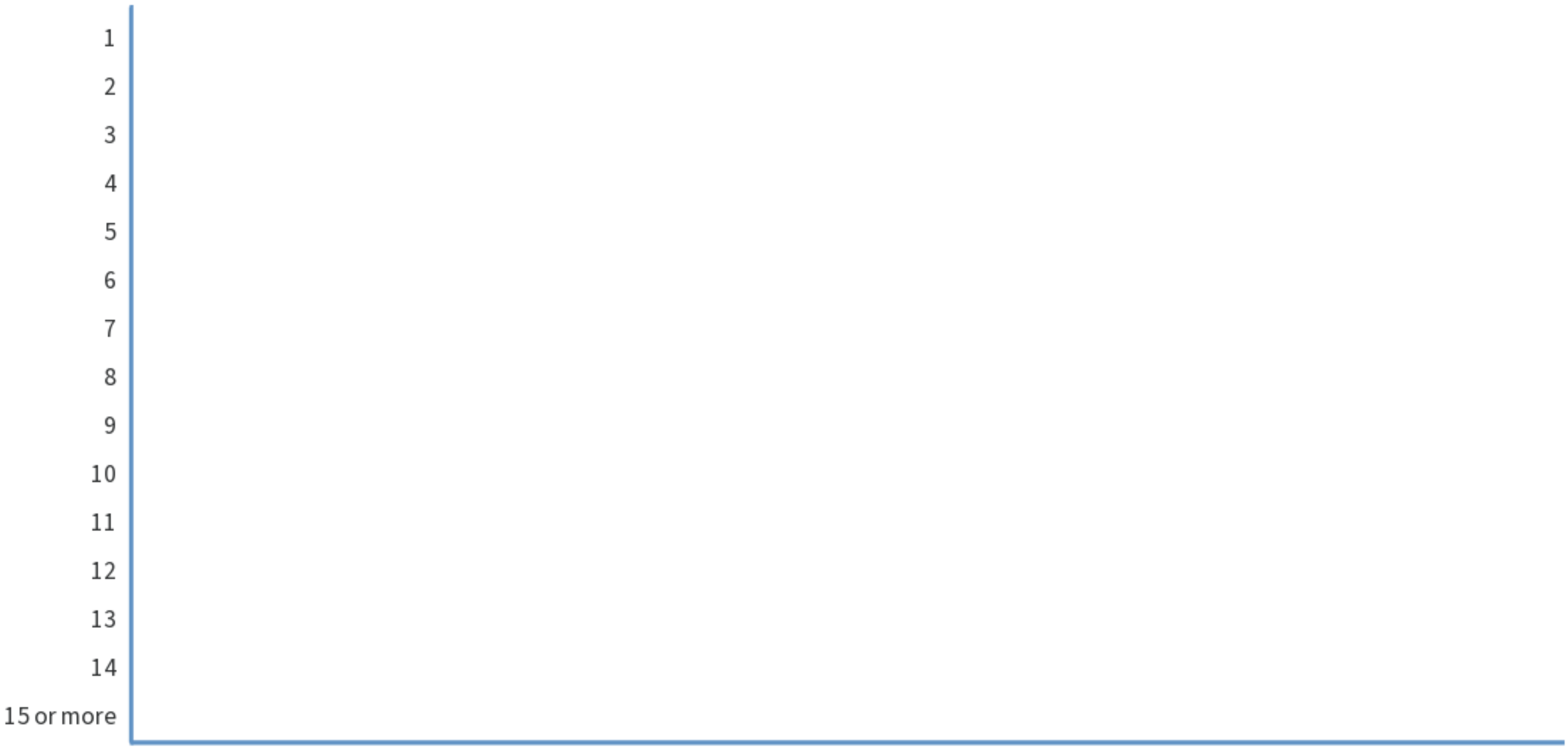
Sarah Fenrich



Eunjung Myoung



How many hours did it take you to complete Homework 2?



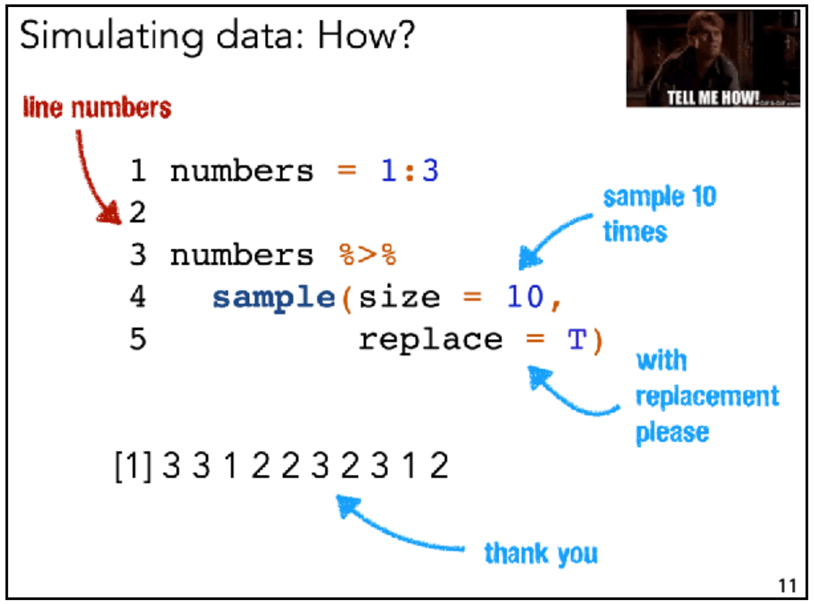
Outline

Goal: Revisit and understand key statistical concepts

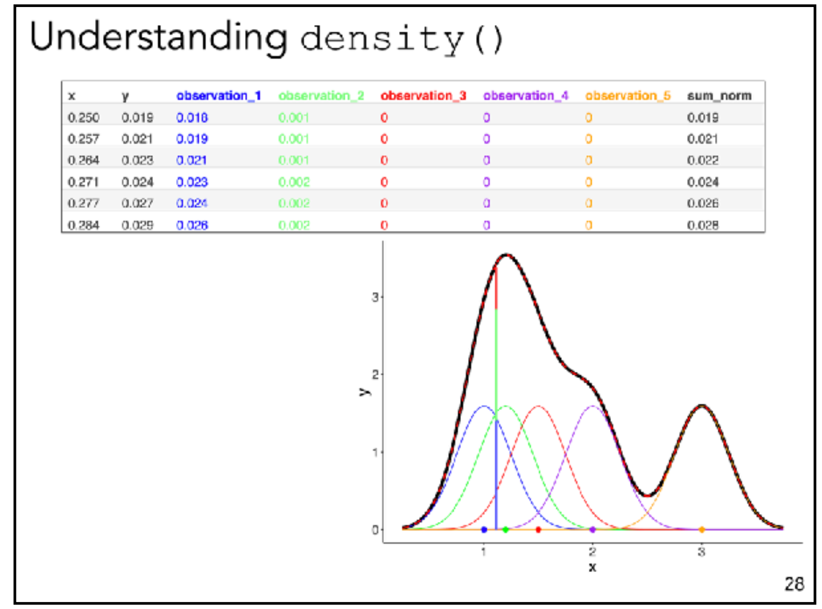
- Quick recap
- Doing Bayesian Analysis
- Inference in frequentist statistics
- Sampling distributions
- What is a p-value?
  - Permutation test

Quick recap

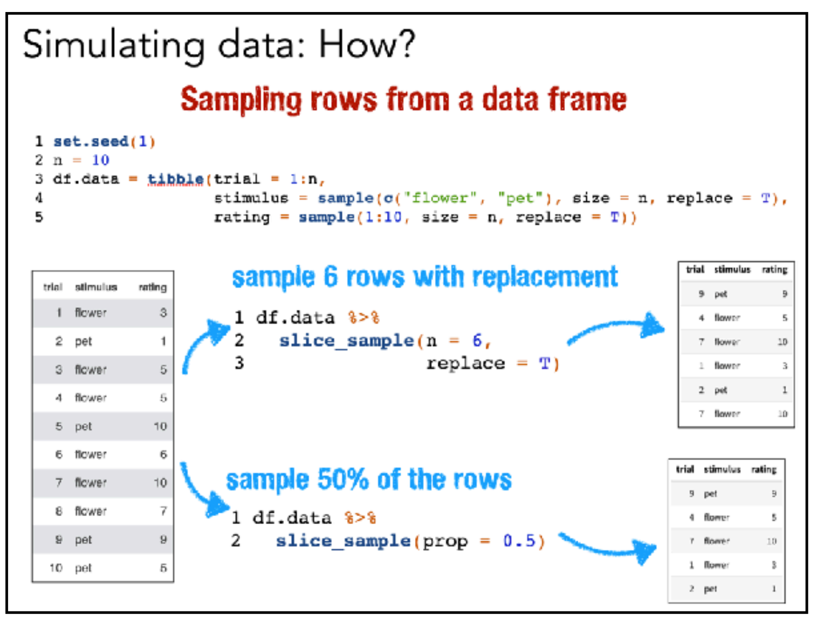
Quick recap from Simulation 1



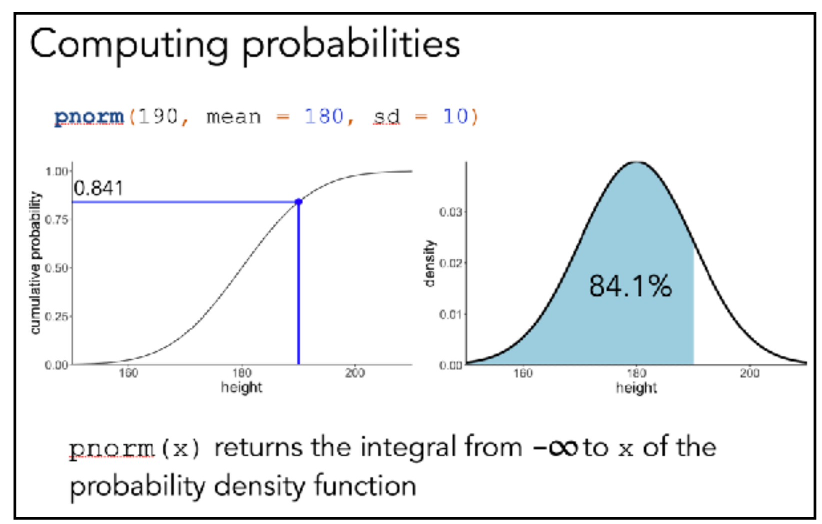
sampling values from a vector



understanding `density()`



sampling rows from a data frame



answering questions with probability distributions

or via drawing samples

`rnorm()` + wrangling

Doing Bayesian Analysis

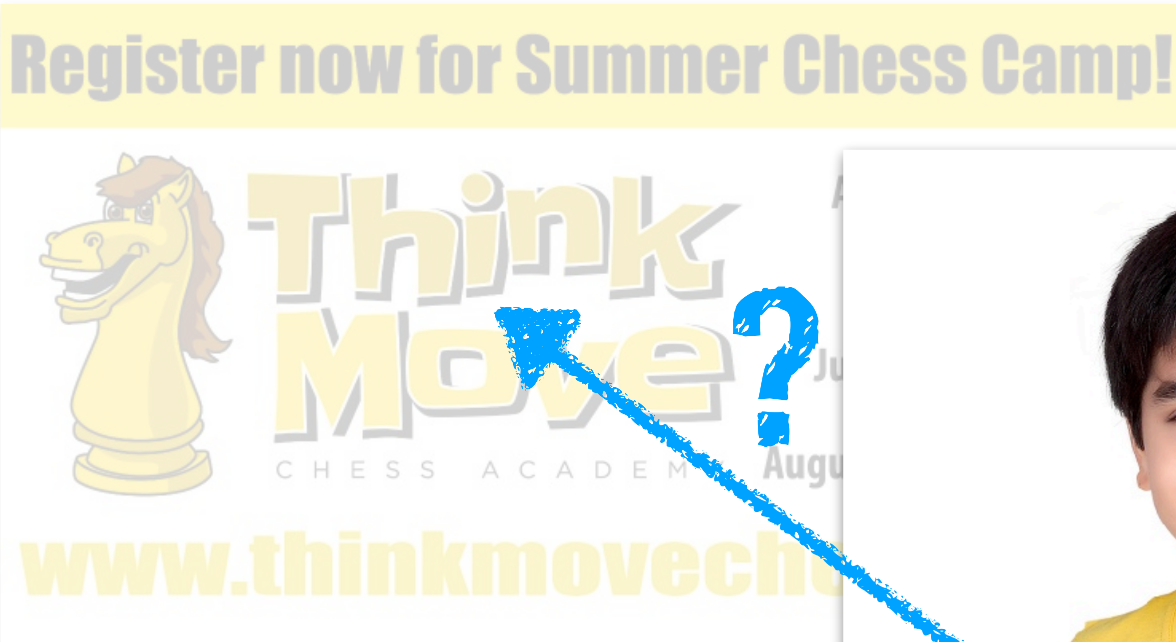
Summer camp

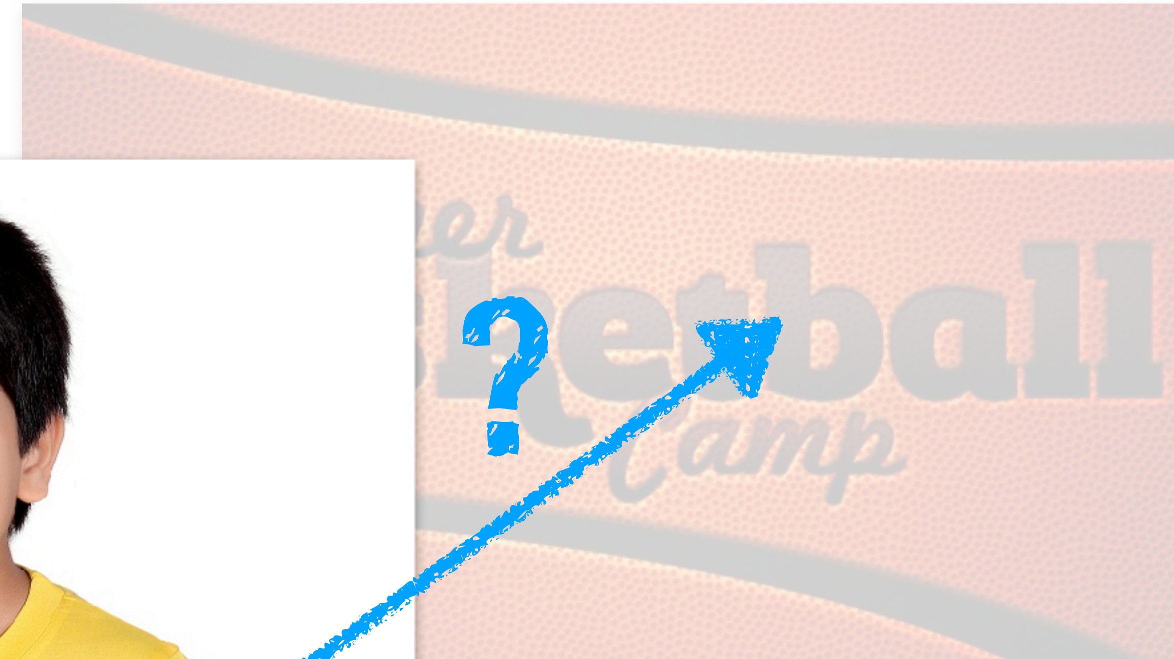
twice as many kids go to the basketball camp

$X \sim$ Normal$(\mu = 170, \sigma = 8)$

$X \sim$ Normal$(\mu = 180, \sigma = 10)$

Summer camp





twice as ma … etball camp

$X \sim$ Normal($\mu = 170$ … nal($\mu = 180$, $\sigma = 10$)

Analytic solution

Can you feel the Bayes?

H = {basketball, chess}

D = 175 cm

$$\overset{\text{posterior}}{p(H \mid D)} = \frac{\overset{\text{likelihood}}{p(D \mid H)} \cdot \overset{\text{prior}}{p(H)}}{\underset{\text{probability of the data?!}}{p(D)}}$$

H = Hypothesis

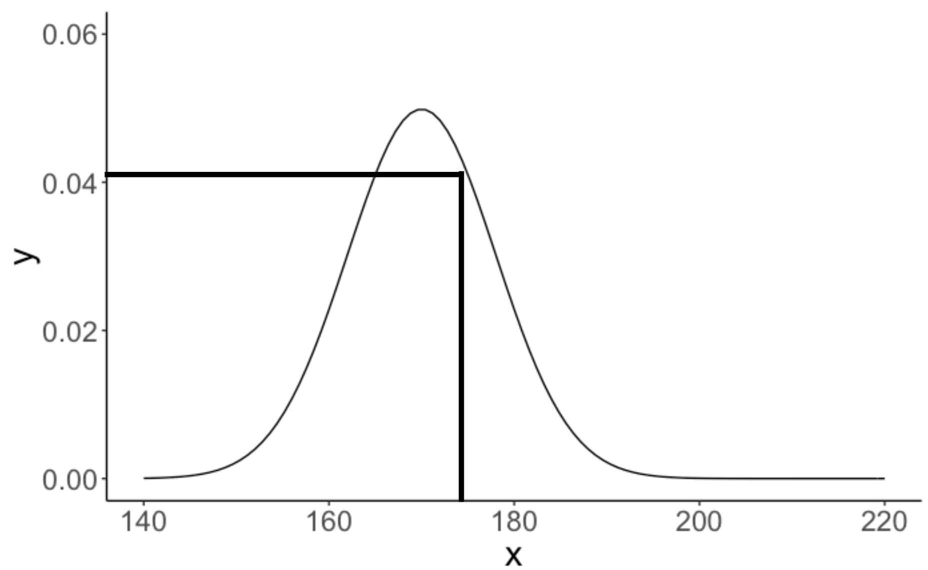
D = Data

Summer camp

prior

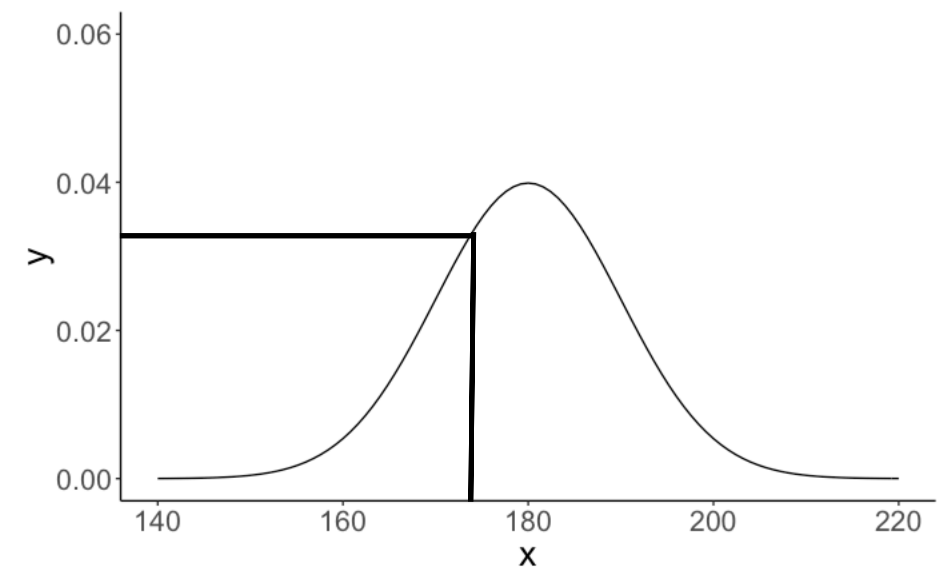
$$p(\text{chess}) = \frac{1}{3} \qquad p(\text{basketball}) = \frac{2}{3}$$

likelihood



dnorm(175, mean = 170, sd = 8)

= 0.041



dnorm(175, mean = 180, sd = 10)

= 0.035

posterior

likelihood prior

$$p(\text{sport} = \text{basketball} \,|\, \text{height} = 175) = \frac{p(175 \,|\, \text{basketball}) \cdot p(\text{basketball})}{p(175)}$$

data

$$p(\text{basketball} \,|\, 175) = \frac{p(175 \,|\, \text{basketball}) \cdot p(\text{basketball})}{p(175 \,|\, \text{basketball}) \cdot p(\text{basketball}) + p(175 \,|\, \text{chess}) \cdot p(\text{chess})}$$

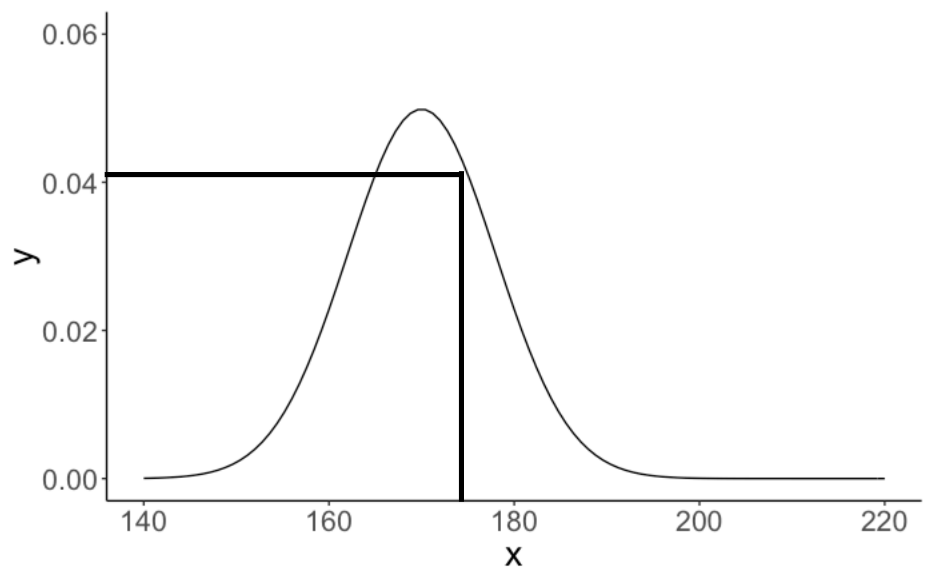
Summer camp

prior

$$p(\text{chess}) = \frac{1}{3}$$

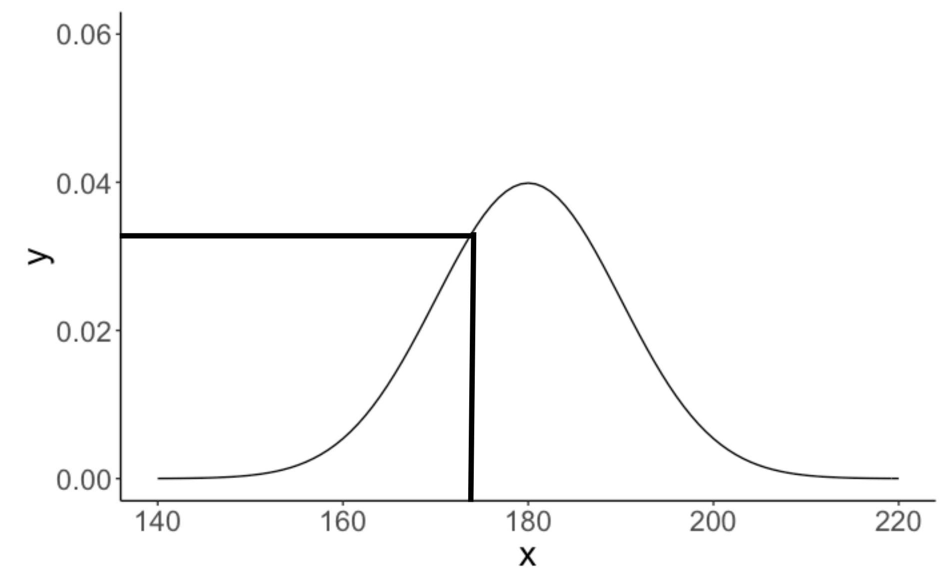
$$p(\text{basketball}) = \frac{2}{3}$$

likelihood




```
dnorm(175, mean = 170, sd = 8)
```

= 0.041




```
dnorm(175, mean = 180, sd = 10)
```

= 0.035

posterior

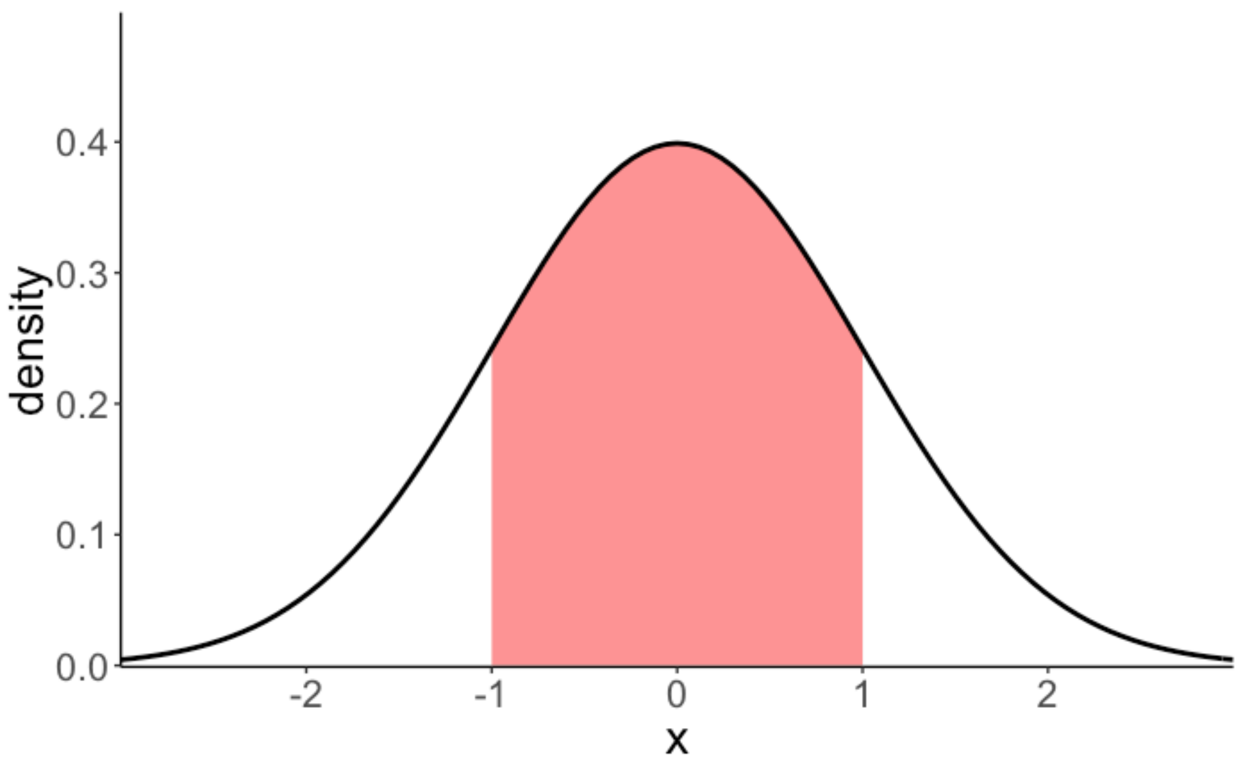
$$p(\text{basketball} \mid 175) = \frac{p(175 \mid \text{basketball}) \cdot p(\text{basketball})}{p(175 \mid \text{basketball}) \cdot p(\text{basketball}) + p(175 \mid \text{chess}) \cdot p(\text{chess})}$$

$$p(\text{basketball} \mid 175) = \frac{0.035 \cdot 2/3}{0.035 \cdot 2/3 + 0.041 \cdot 1/3} \approx 0.63$$

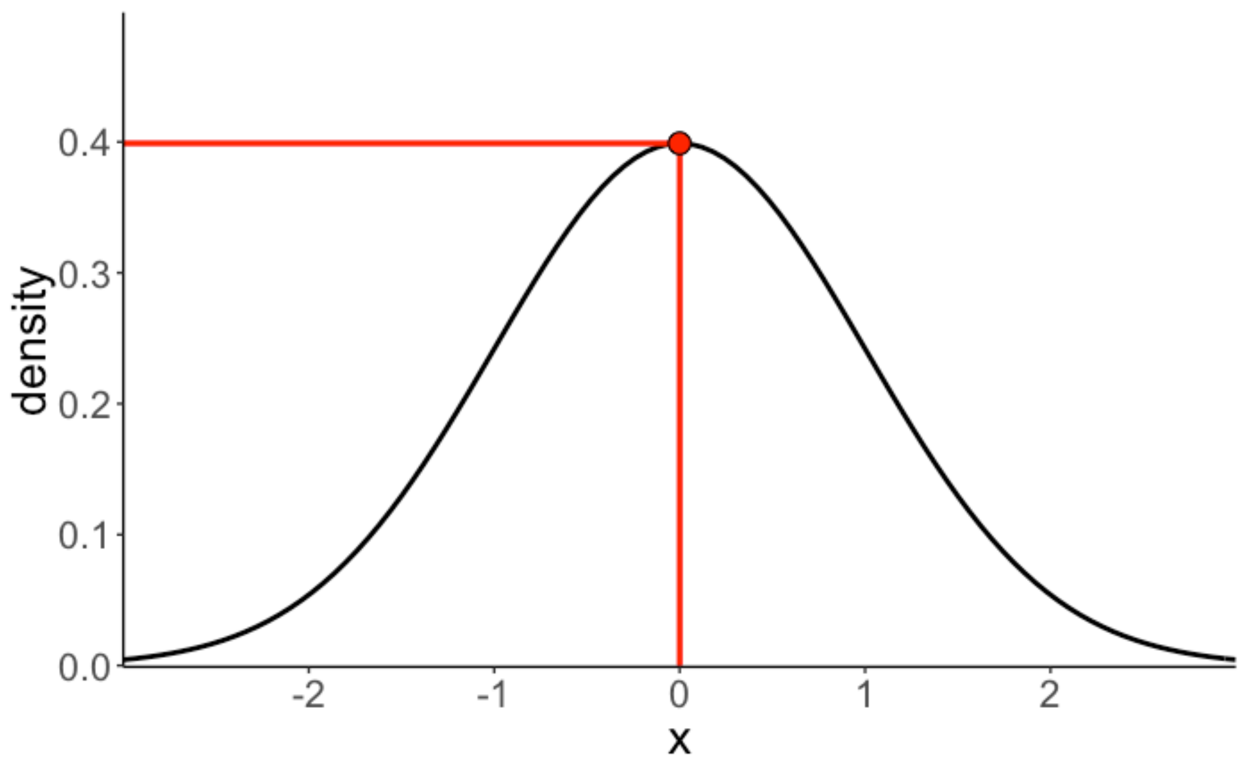
send the kid to the basketball gym!

Probability vs. likelihood

Probability

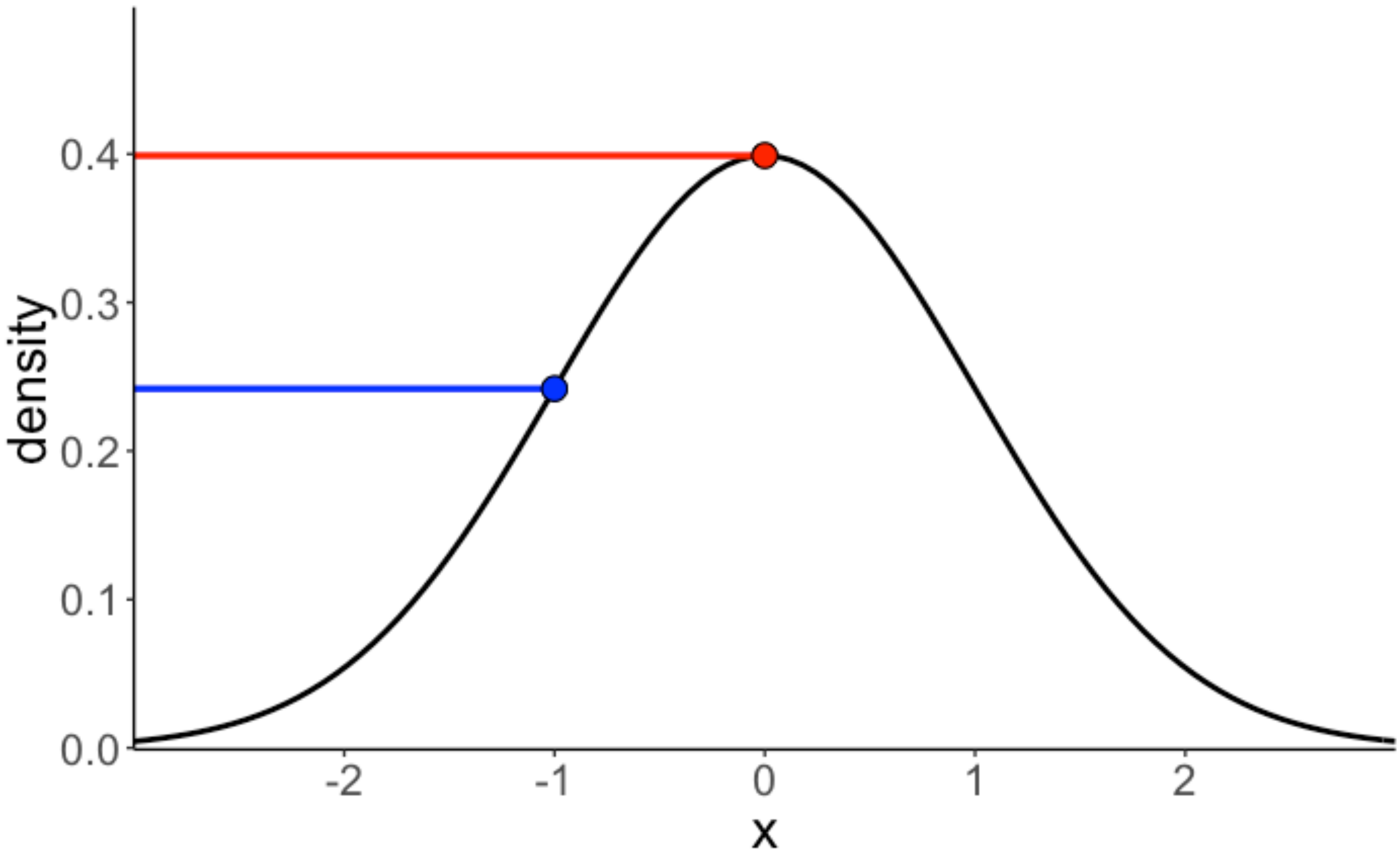


Likelihood



Probability vs. likelihood

`dnorm(0)/dnorm(-1) = 1.6487`

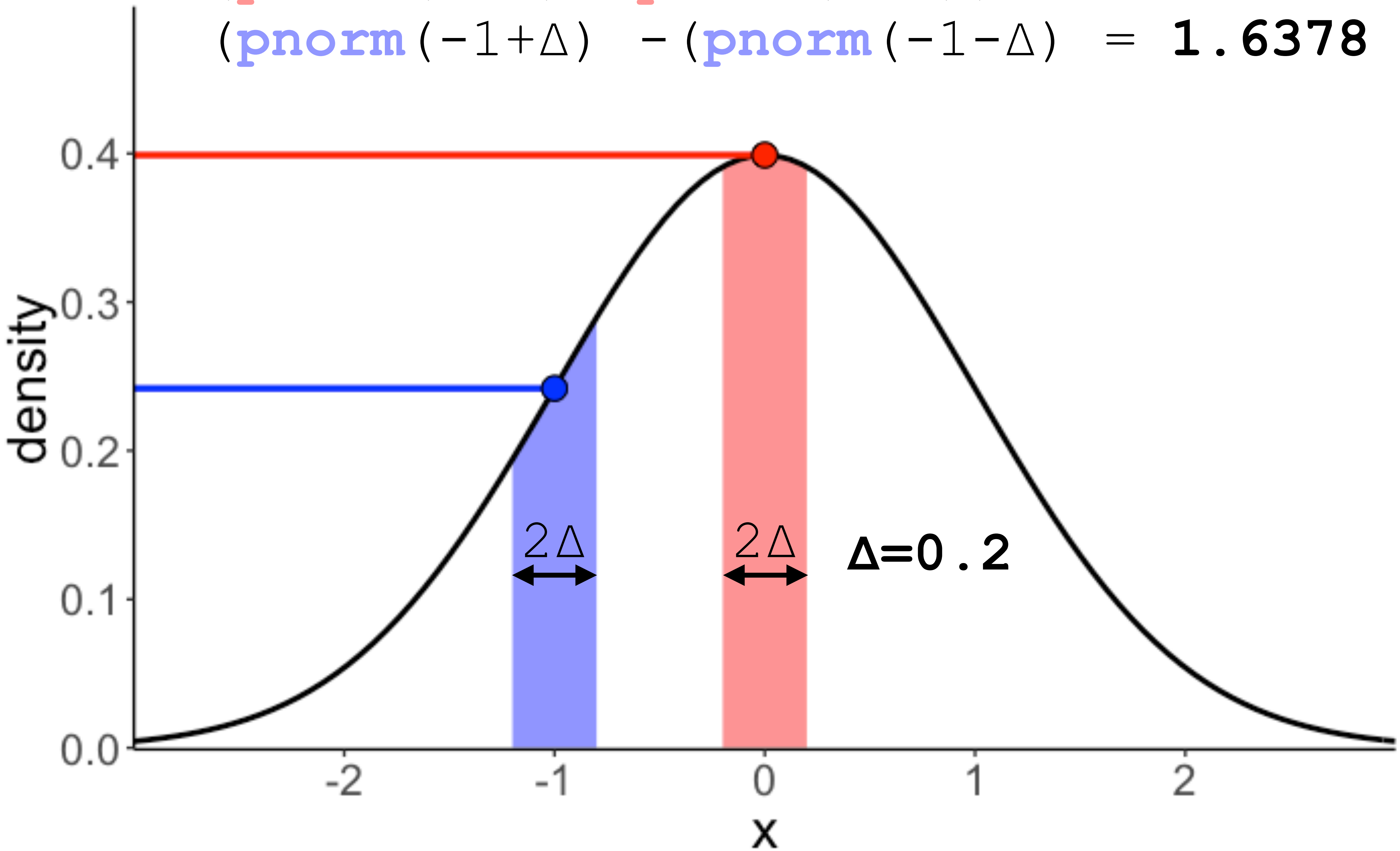


relative probability of one value vs. another

Probability vs. likelihood

```
dnorm(0)/dnorm(-1) = 1.6487
```

```
(pnorm(0+Δ)- pnorm(0-Δ))/
(pnorm(-1+Δ) -(pnorm(-1-Δ) = 1.6378
```

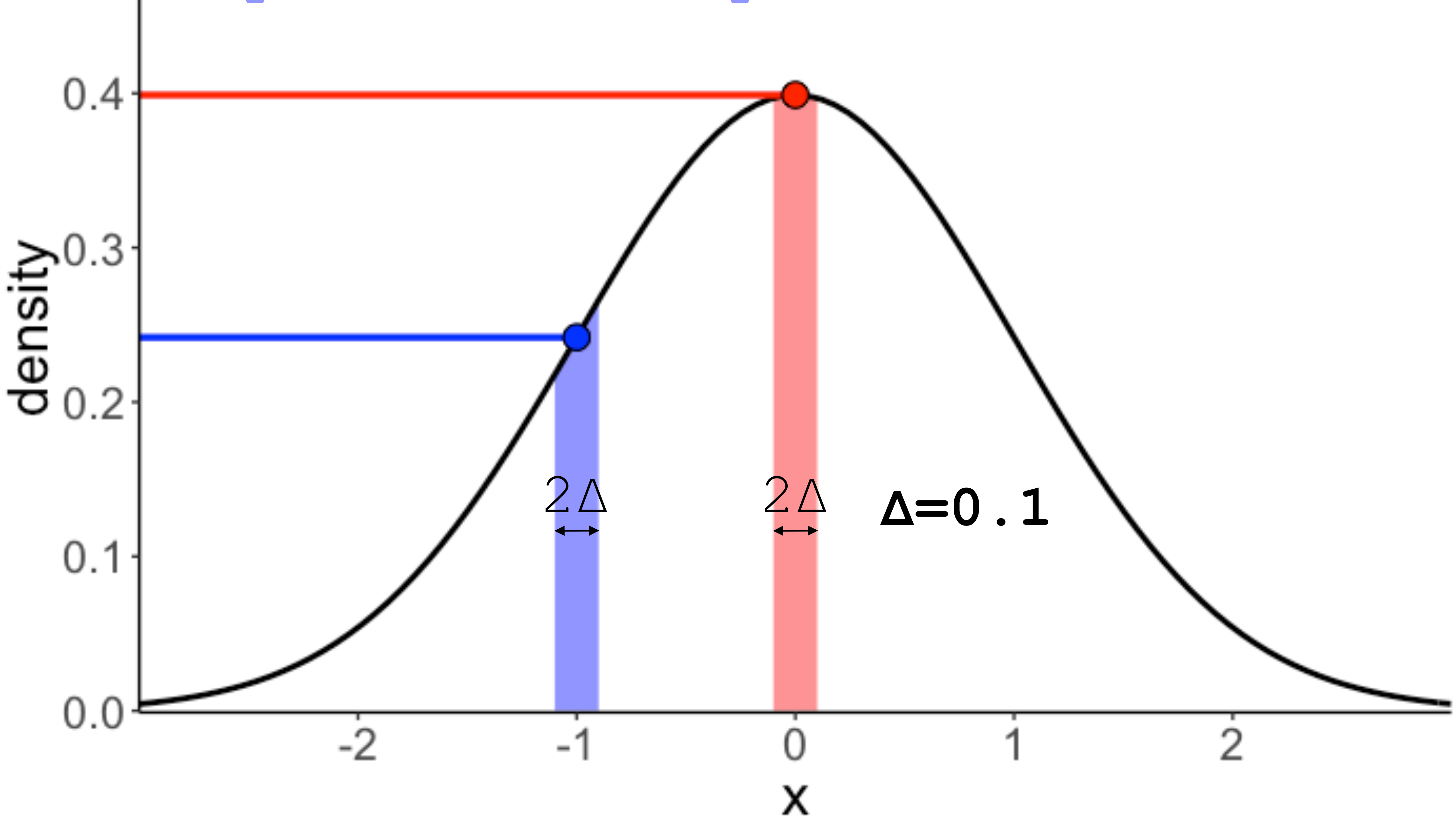


relative probability of one value vs. another

Probability vs. likelihood

`dnorm(0)/dnorm(-1) = 1.6487`

`(pnorm(0+Δ)- pnorm(0-Δ))/`
`(pnorm(-1+Δ) -(pnorm(-1-Δ) =` **`1.6459`**

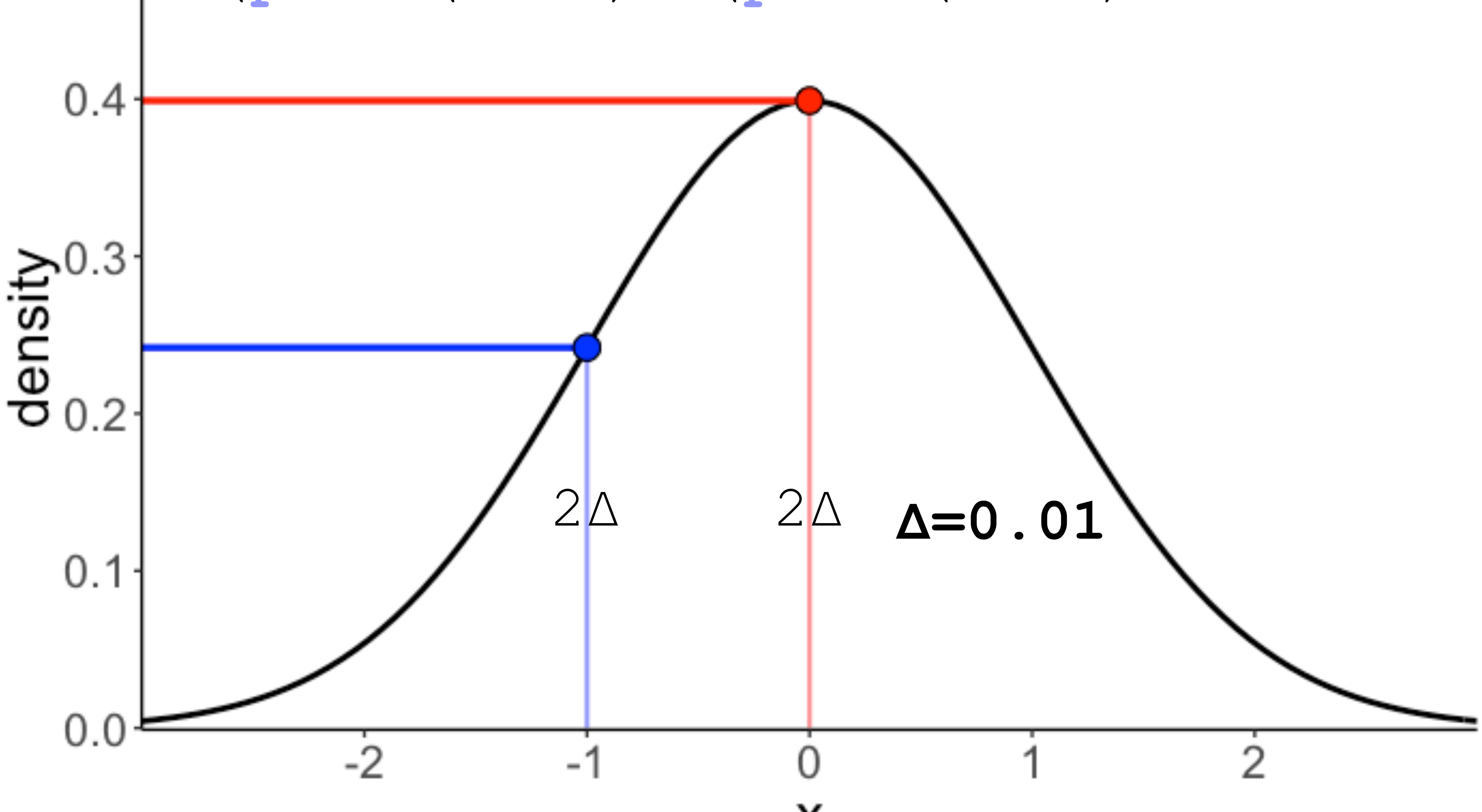


relative probability of one value vs. another

Probability vs. likelihood

`dnorm(0)/dnorm(-1) = 1.6487`

`(pnorm(0+Δ)- pnorm(0-Δ))/`
`(pnorm(-1+Δ) -(pnorm(-1-Δ) =` **`1.6486`**



relative probability of one value vs. another

Sampling solution

Summer camp: Via sampling

```r
df.camp = tibble(
  kid = 1:1000,
  sport = sample(c("chess", "basketball"),
                 size = 1000,
                 replace = T,
                 prob = c(1/3, 2/3))) %>%
  rowwise() %>%
  mutate(height = ifelse(test = sport == "chess",
                         yes = rnorm(., mean = 170, sd = 8),
                         no = rnorm(., mean = 180, sd = 10))) %>%
  ungroup()
```

kid	sport	height
1	basketball	164.84
2	basketball	163.22
3	basketball	191.18
4	chess	160.16
5	basketball	182.99
6	chess	163.54
7	chess	168.56
8	basketball	192.99
9	basketball	171.91
10	basketball	177.12

```r
df.camp %>%
  filter(height == 175) %>%
  count(sport)
```

doesn't work!

Summer camp: Via sampling

```
df.camp = tibble(
  kid = 1:100000,
  sport = sample(c("chess", "basketball"),
                 size = 100000,
                 replace = T,
                 prob = c(1/3, 2/3))) %>%
  rowwise() %>%
  mutate(height = ifelse(test = sport == "chess",
                         yes = rnorm(., mean = 170, sd = 8),
                         no = rnorm(., mean = 180, sd = 10))) %>%
  ungroup()
```

kid	sport	height
1	basketball	164.84
2	basketball	163.22
3	basketball	191.18
4	chess	160.16
5	basketball	182.99
6	chess	163.54
7	chess	168.56
8	basketball	192.99
9	basketball	171.91
10	basketball	177.12

```
df.camp %>%
  filter(between(height,
                 left = 174,
                 right = 176)) %>%
  count(sport)
```

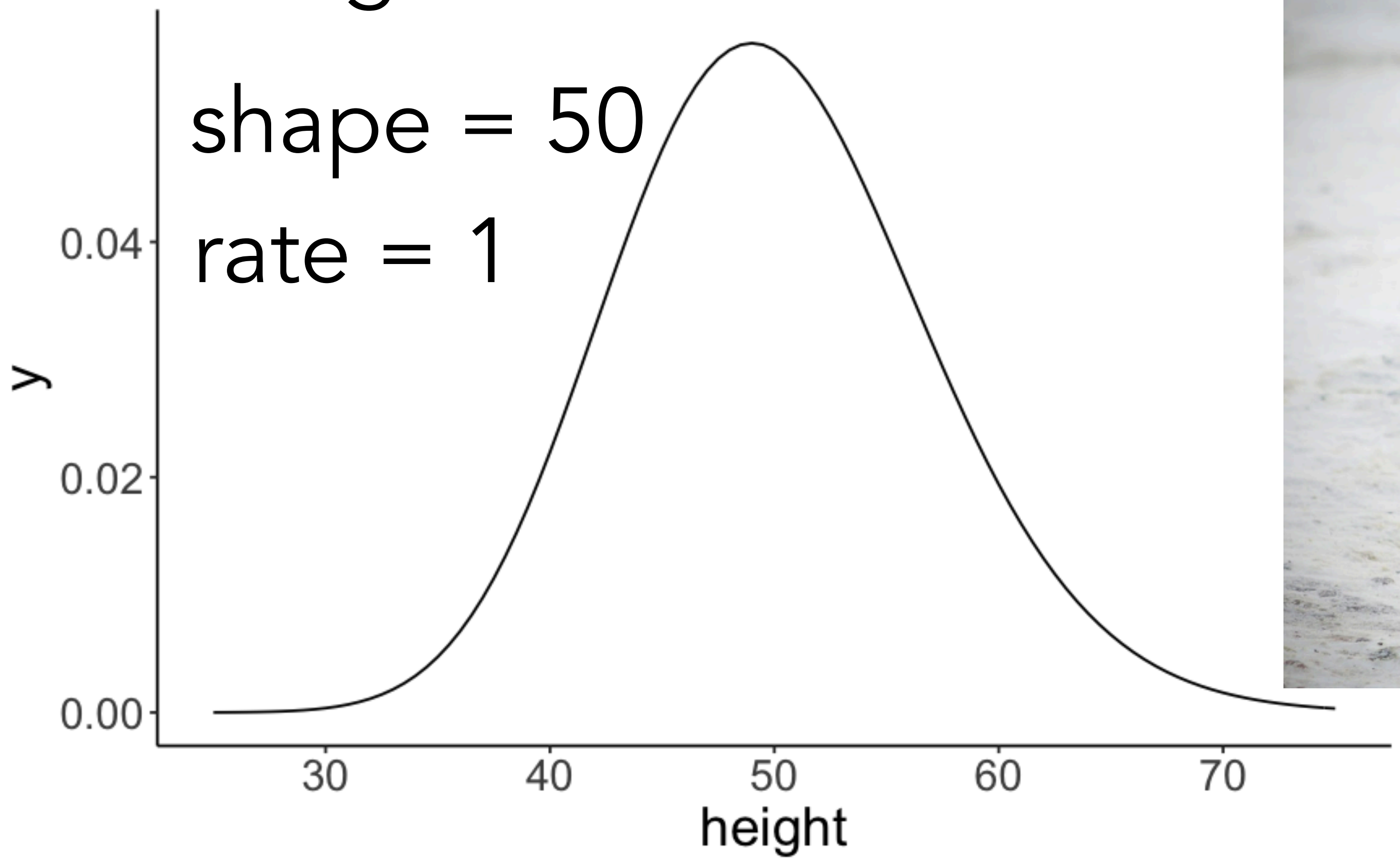
this works!

sport	n
basketball	469
chess	273

$$\frac{\text{basketball}}{\text{basketball} + \text{chess}} \approx 0.63$$

Answering questions about Penguins

gamma distribution



10:00

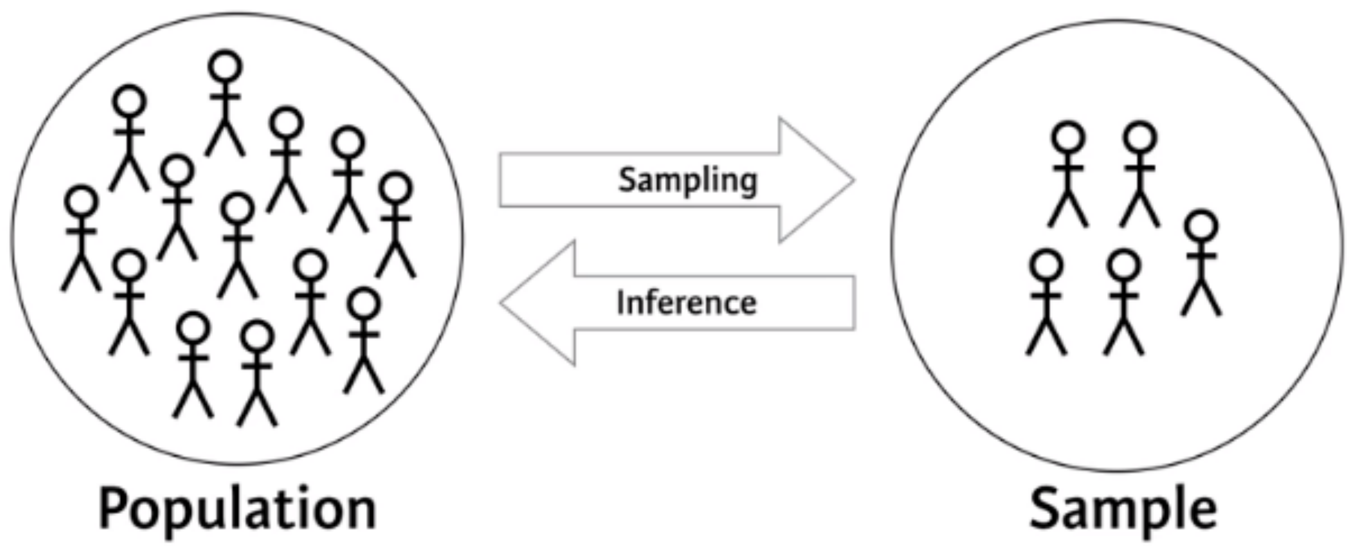
Solve via drawing random samples + data wrangling

1. A 60cm tall Penguin claims that no more than 10% are taller than her. Is she correct?
2. Are there more penguins between 50 and 55cm or between 55 and 65cm?
3. What size is a Penguin who is taller than 75% of the rest?

Inference in frequentist statistics

Statistical inference

The process of making claims about a population based on information from a sample.



Life would be easy if we were able to observe the whole population -- we could simply do descriptive analyses!

Key question:

What can we infer about the population from our sample?

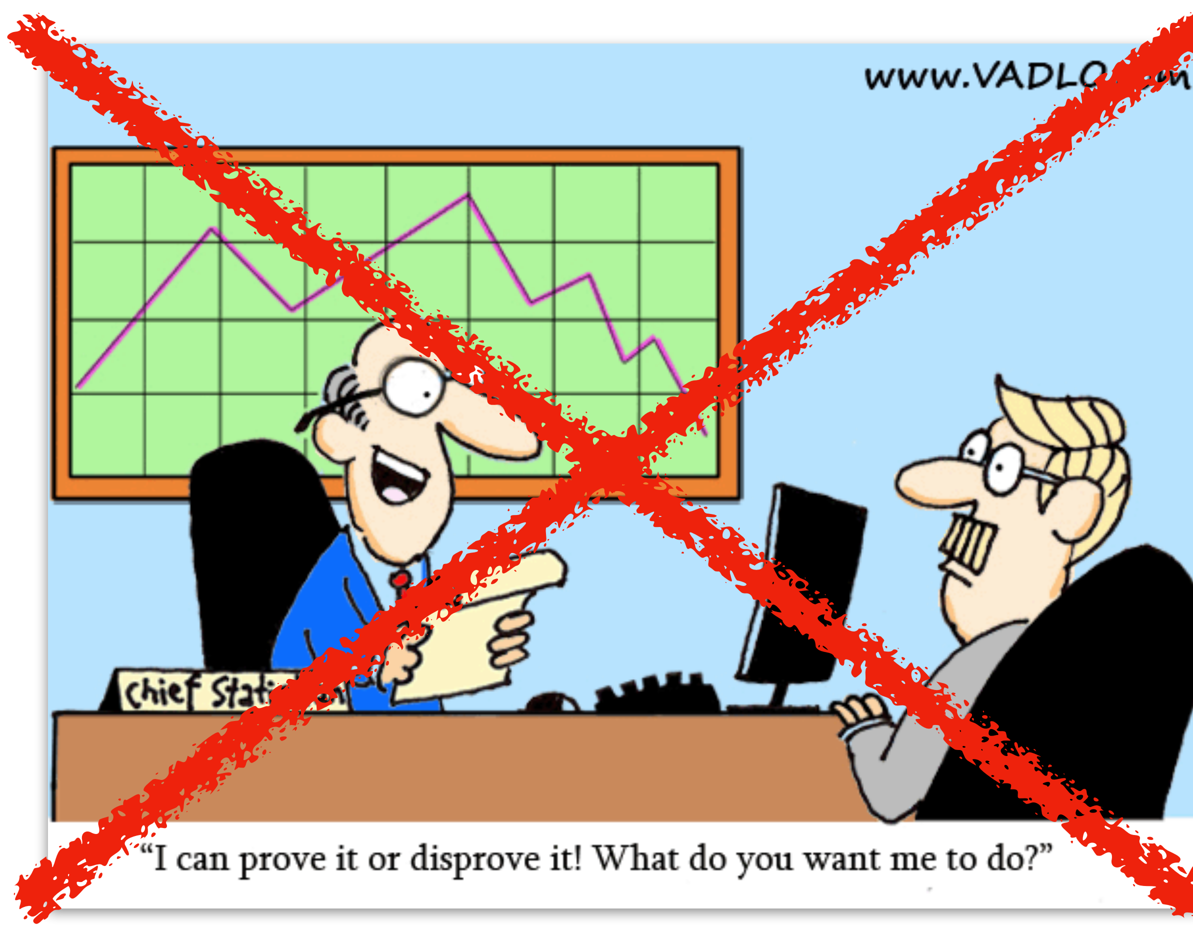
Statistical inference

Key question:

What can we infer about the population from our sample?

Answer:

- is not trivial
- mathematical, statistical, philosophical (Bayesian vs. frequentist) machinery involved
- **important**: we can never make deterministic statements!



Underlying principle of statistical testing

1. Define population, state hypotheses
2. Draw one (ideally large) random sample
3. Compute measure of interest (e.g. mean, correlation coefficient, difference between condition means), and then the test statistic
4. Apply statistical distribution theory to get the **sampling distribution of a test statistic**
5. Evaluate the observed test statistic on the sampling distribution; make a decision (either reject or don't reject H_0) based on pre-specified significance level α

The magical component

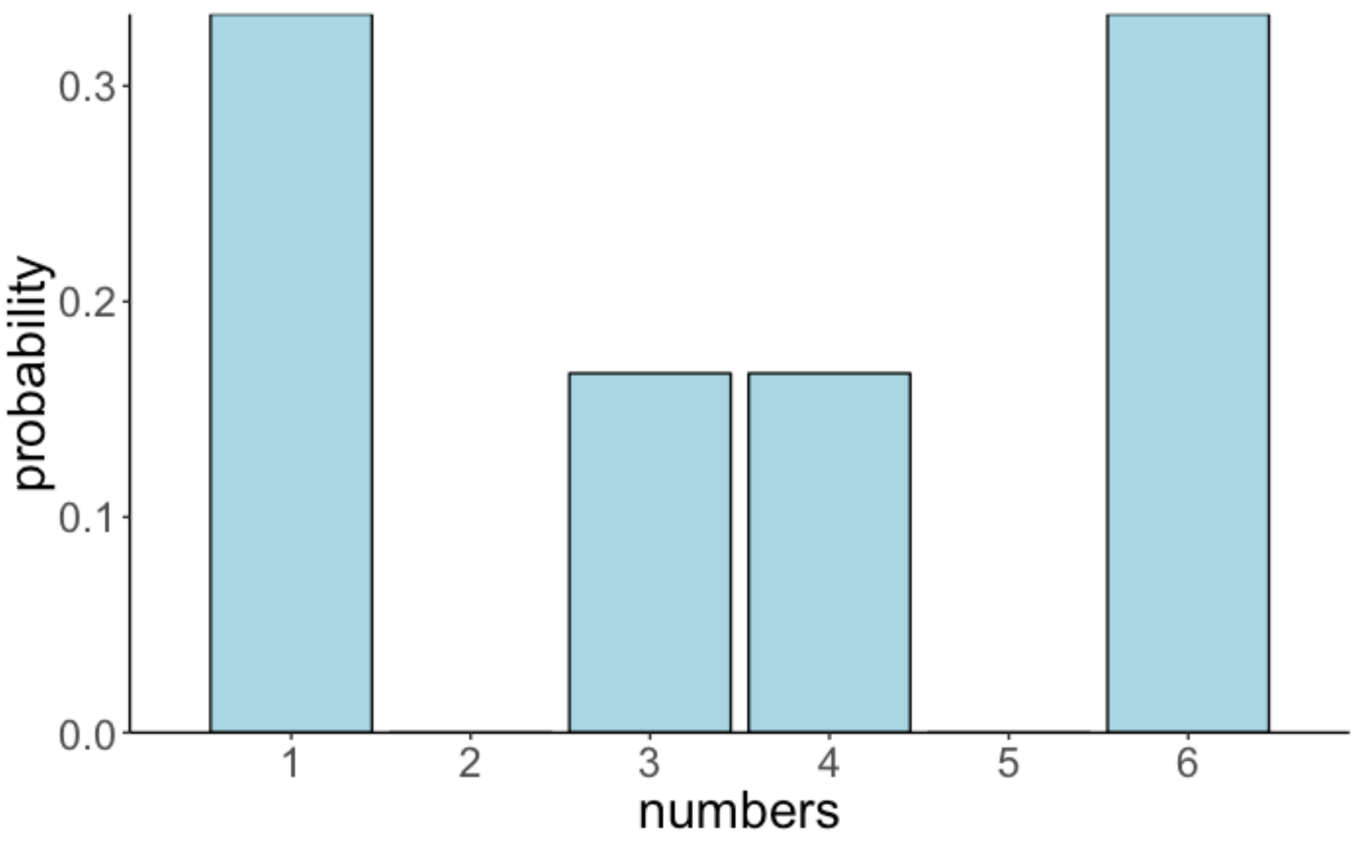
"4. Apply statistical distribution theory to get the **sampling distribution of a test statistic**"

This dates back to pre-computer era where statisticians derived mathematically the distribution of statistical measures for an infinite amount of samples! That's are tricky thing to do and these approximations are typically tied to assumptions such as normality, homoscedasticity, independent observations, and: the sample needs to be "large".

instead: simulation-based approach

Sampling distributions

heavy metal distribution

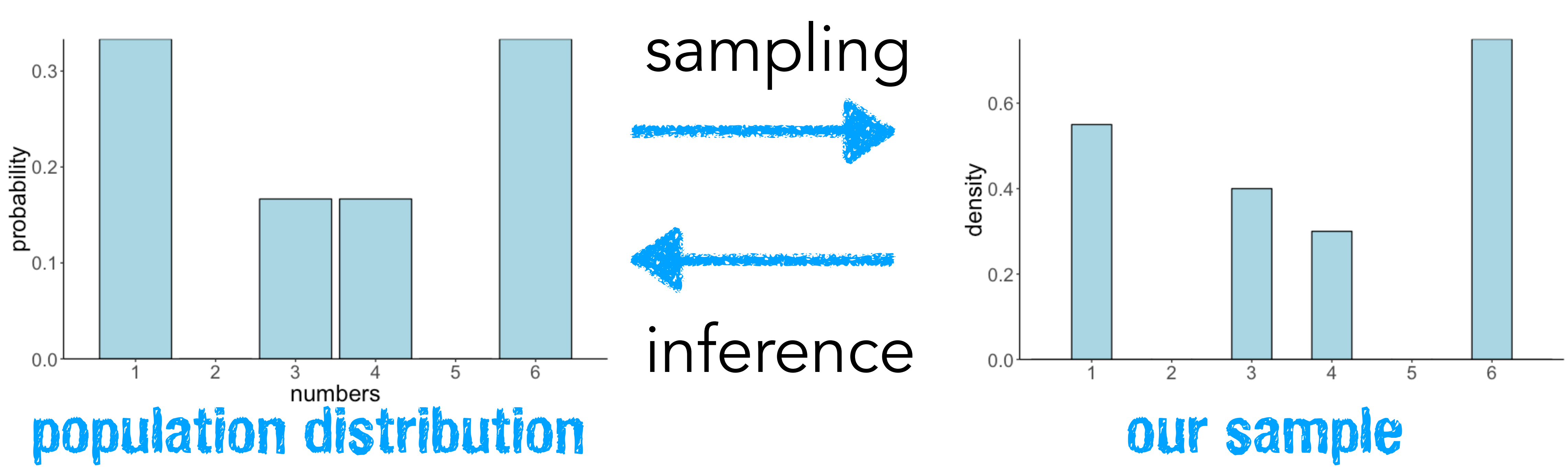


population distribution

Statistical inference

The process of making claims about a population based on information from a sample.

heavy metal distribution



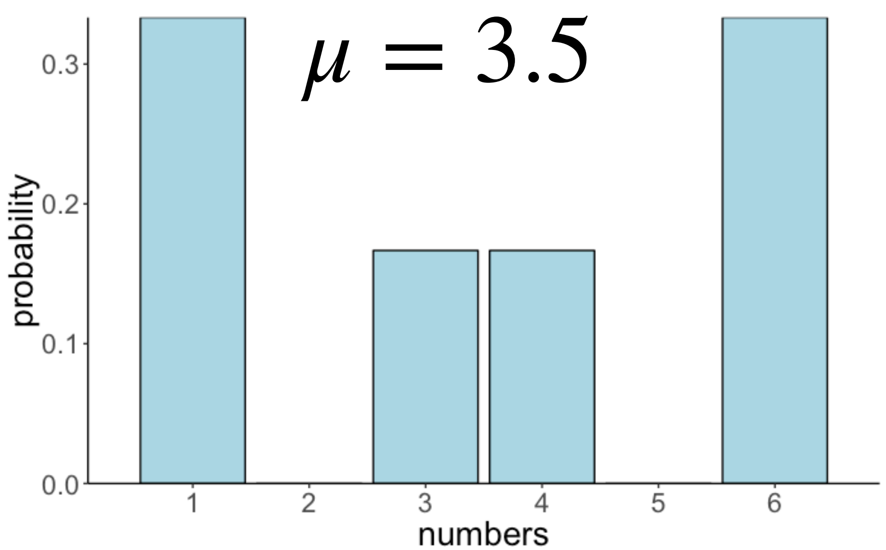
Statistical inference

what's the
population mean?

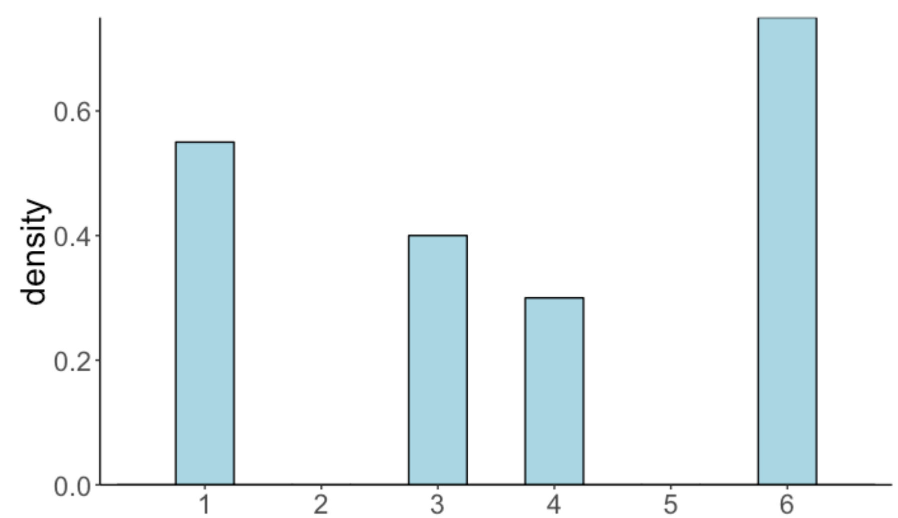
```
sample mean = 3.725
standard deviation = 2.05
n = 40
```

heavy metal distribution

$\mu = 3.5$

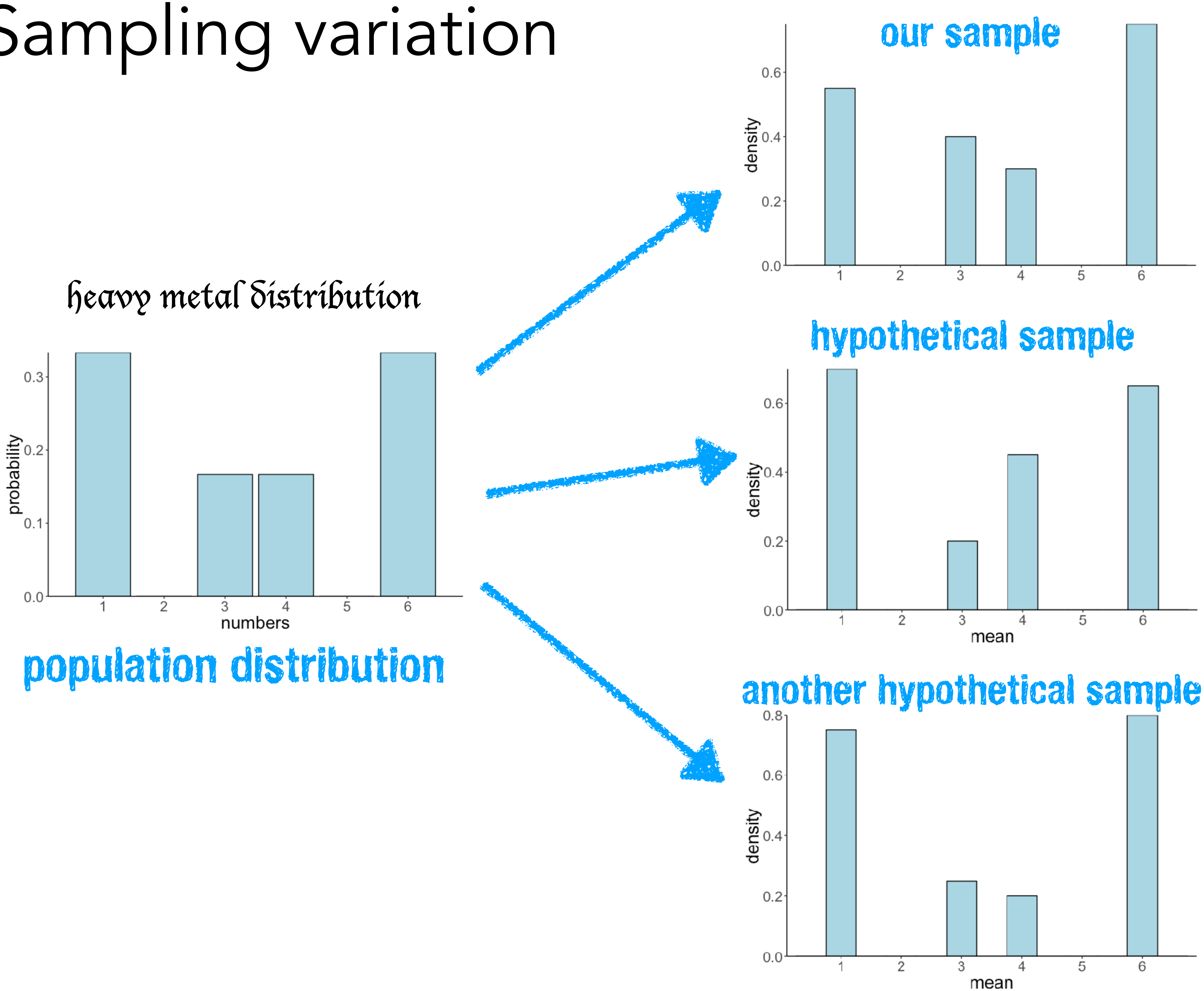


true unknown distribution



our sample

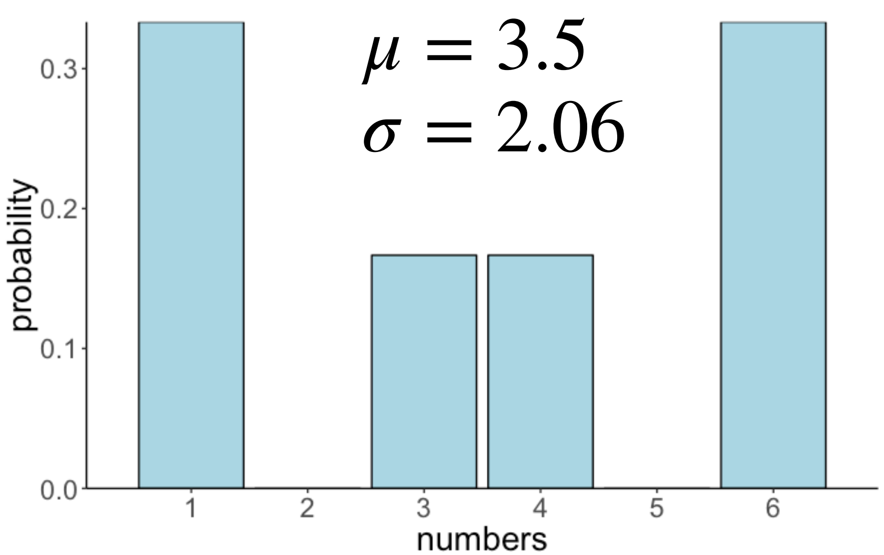
Sampling variation



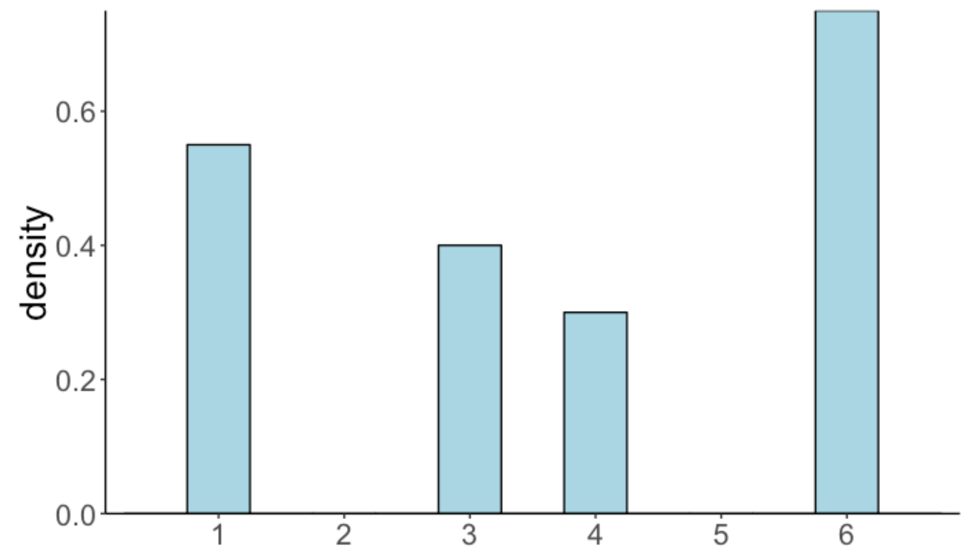
Sampling distribution

population distribution

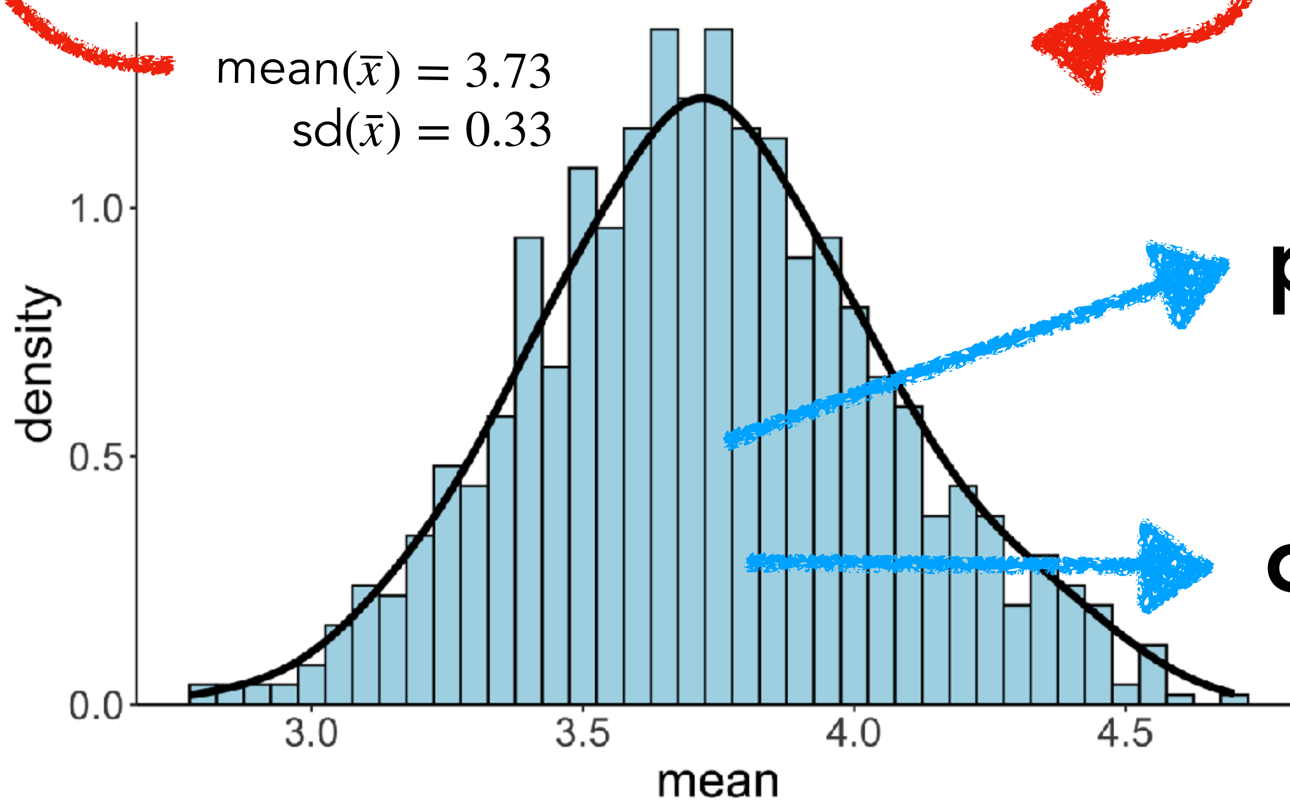
heavy metal distribution



our sample



$\text{mean}(x) = 3.72$
$\text{sd}(x) = 2.05$
$\text{n} = 40$

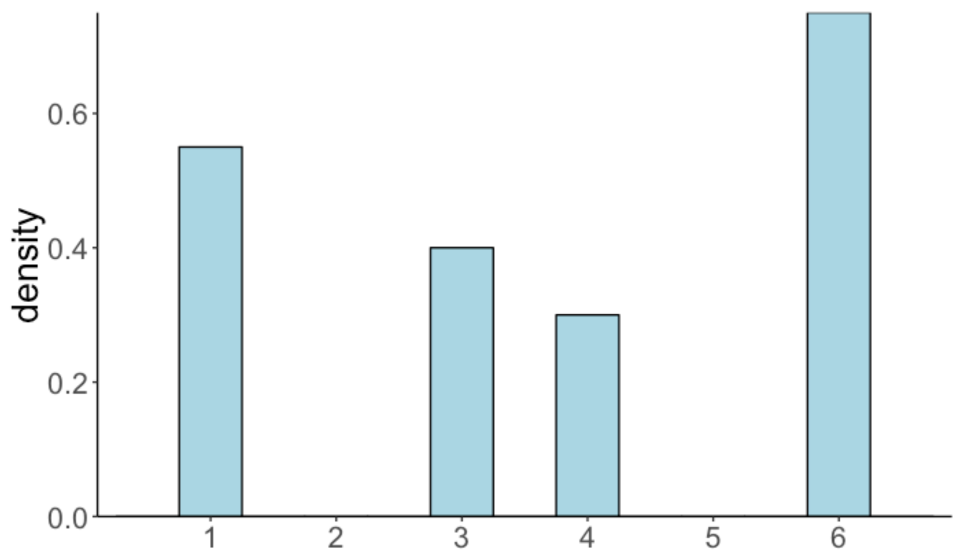


sampling distribution of the mean

p-values

confidence intervals

our sample

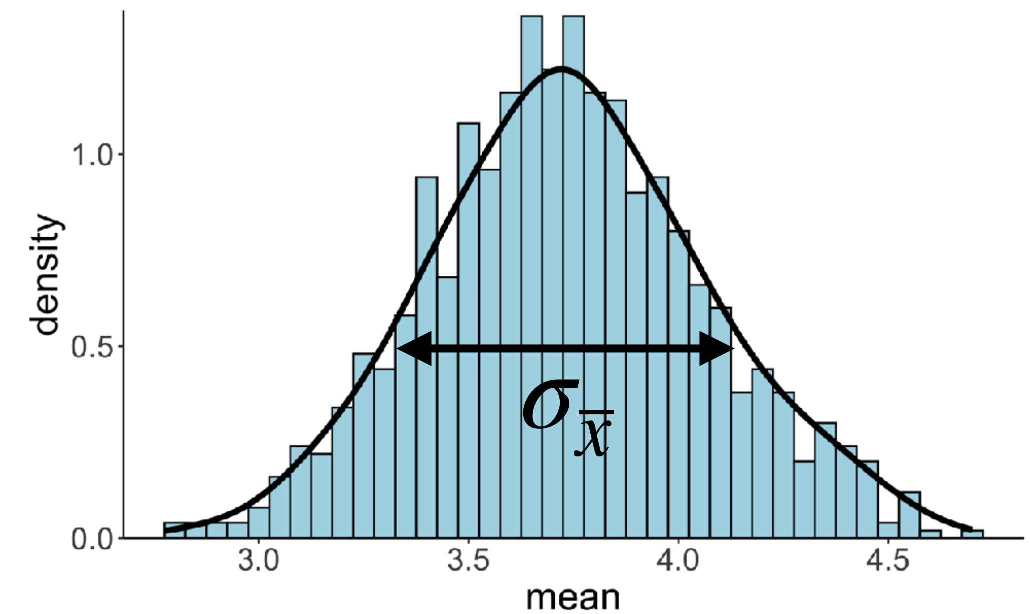


standard deviation

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{N - 1}}$$

gives a sense for how well the mean summarizes the data

sampling distribution



standard error

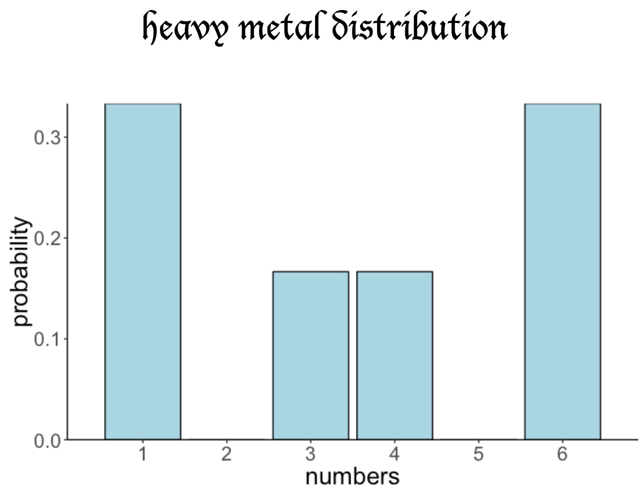
$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$$

the standard deviation of the sampling distribution

how much variation would we expect between the means of different samples

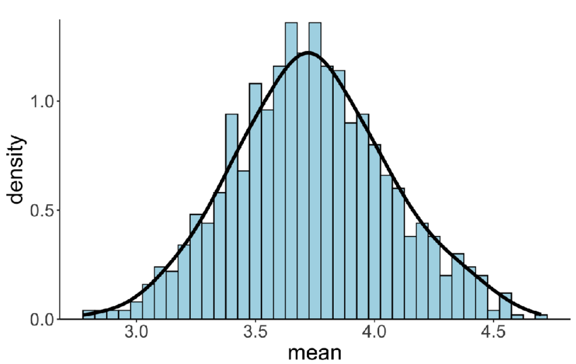
how likely is it that our sample mean is representative of the population mean?

3 distributions in statistical inference



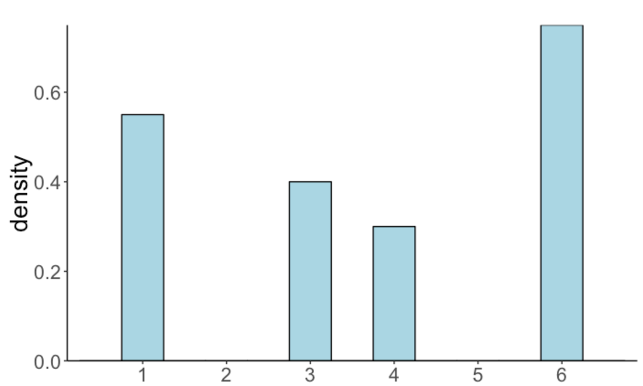
population distribution

- unknown
- our target for inference
- e.g. we might be interested in the mean of the population distribution



sampling distribution

- bridge between sample and population
- derived mathematically / computationally
- asymptotic distribution theory or resampling approaches
- shows how test statistic varies between samples

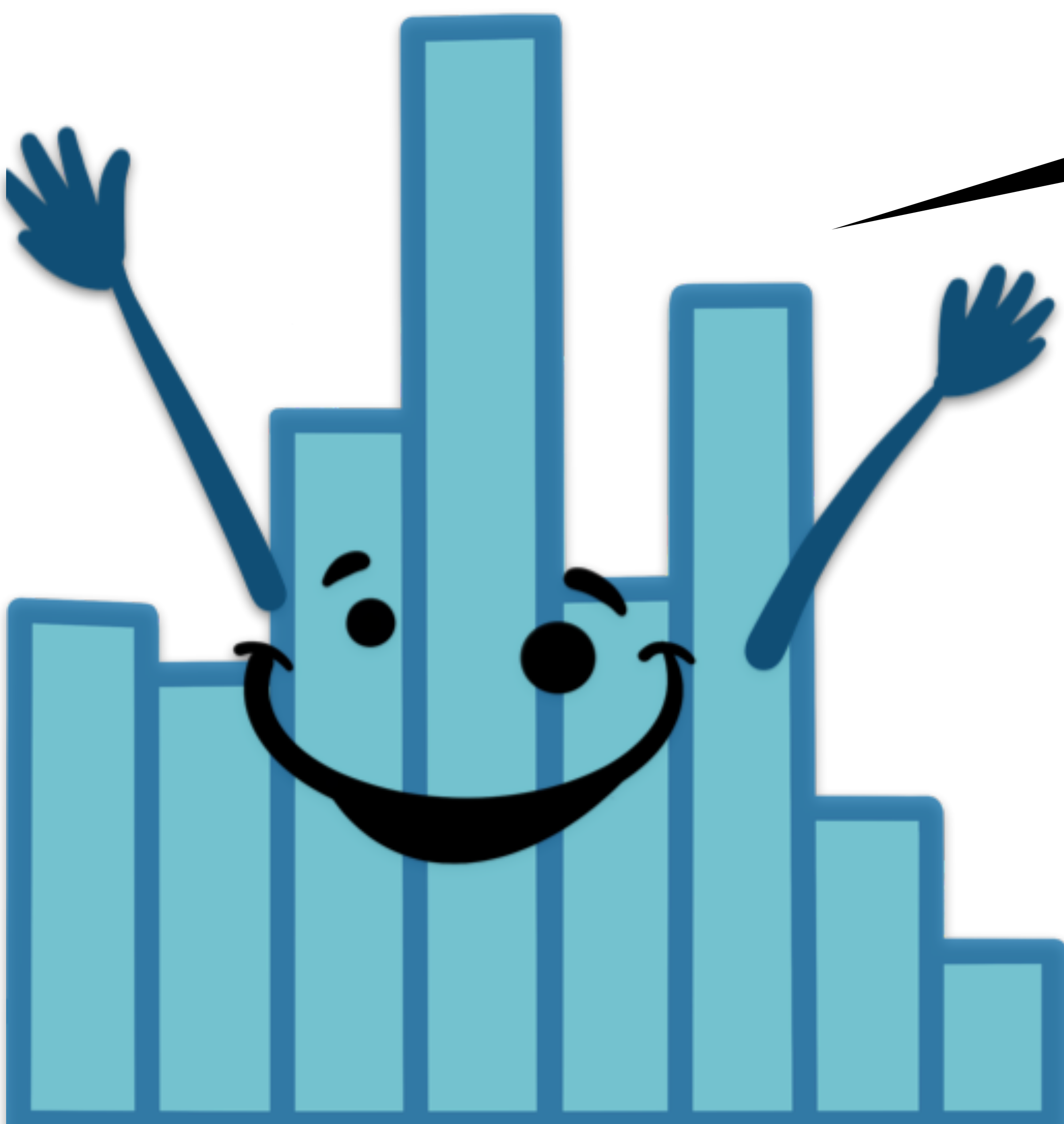


sample distribution

- our observed sample
- we compute statistics of interest (mean, variance, correlation, ...)
- make an inference about the population via the sampling distribution

02:00

stretch break!



What is a p-value?

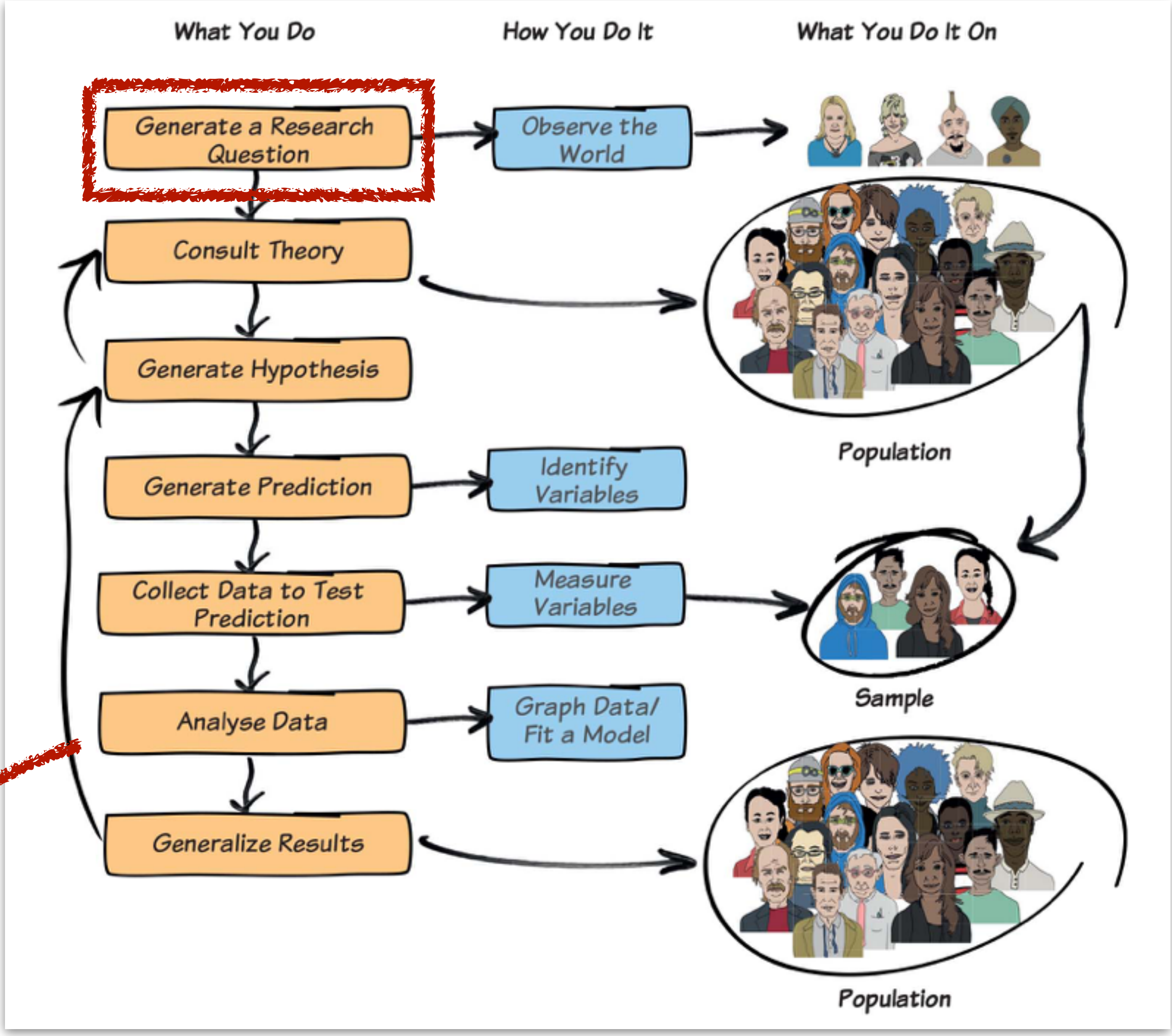
Statistical inference

null hypothesis

$$H_0 : \mu_1 = \mu_2$$

alternative hypothesis

$$H_1 : \mu_1 < \mu_2$$



a p-value, yay!

Which of the following statements about the p-value do you believe to be true?

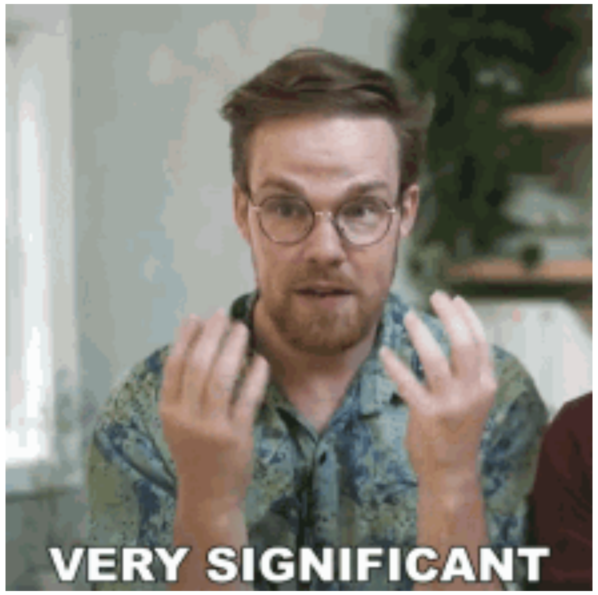
The p-value is the probability that the null hypothesis is true.

The p-value is the probability that the alternative hypothesis is true.

The p-value is the probability of obtaining the observed or more extreme results if the alternative hypothesis is true.

The p-value is the probability of obtaining the observed results or results which are more extreme if the null hypothesis is true.

What is a p-value? Your answers



- Used in hypothesis significance testing. It's quantifies the strength of the evidence against the null hypothesis provided by the data.
- The probability that your data occurred by chance and that there is a statistical difference
- An indication of statistical power related to how likely the null hypothesis is true
- p-value is the likelihood of one data point fall into the sampling distribution
- Probability of getting results that are significant above random chance
- level of significance, usually set to $p < .05$, $p <.01$, or $p <.001$
- Statistical measurement used to validate a hypothesis.
- probability of a given observation to occur
- A value representing confidence in a statistical test's results. E.g., "$p < 0.05$" means that there is less than a 5% chance that the results were due to random chance rather than a statistically significant pattern

What is a p-value?

The **p-value** is the probability of finding the observed (or more extreme) results when the null hypothesis (H_0) is true.

$$p(\text{test statistic} \geq \text{observed value} \,|\, H_0 = \text{true})$$

what we're actually interested in!

$$\cancel{p(H_1 = \text{true} \,|\, \text{test statistic} \geq \text{observed value})}$$

... we'll have to wait for Reverend Bayes

$$p(H \,|\, D) = \frac{p(D \,|\, H) \cdot p(H)}{p(D)}$$

$H =$ Hypothesis

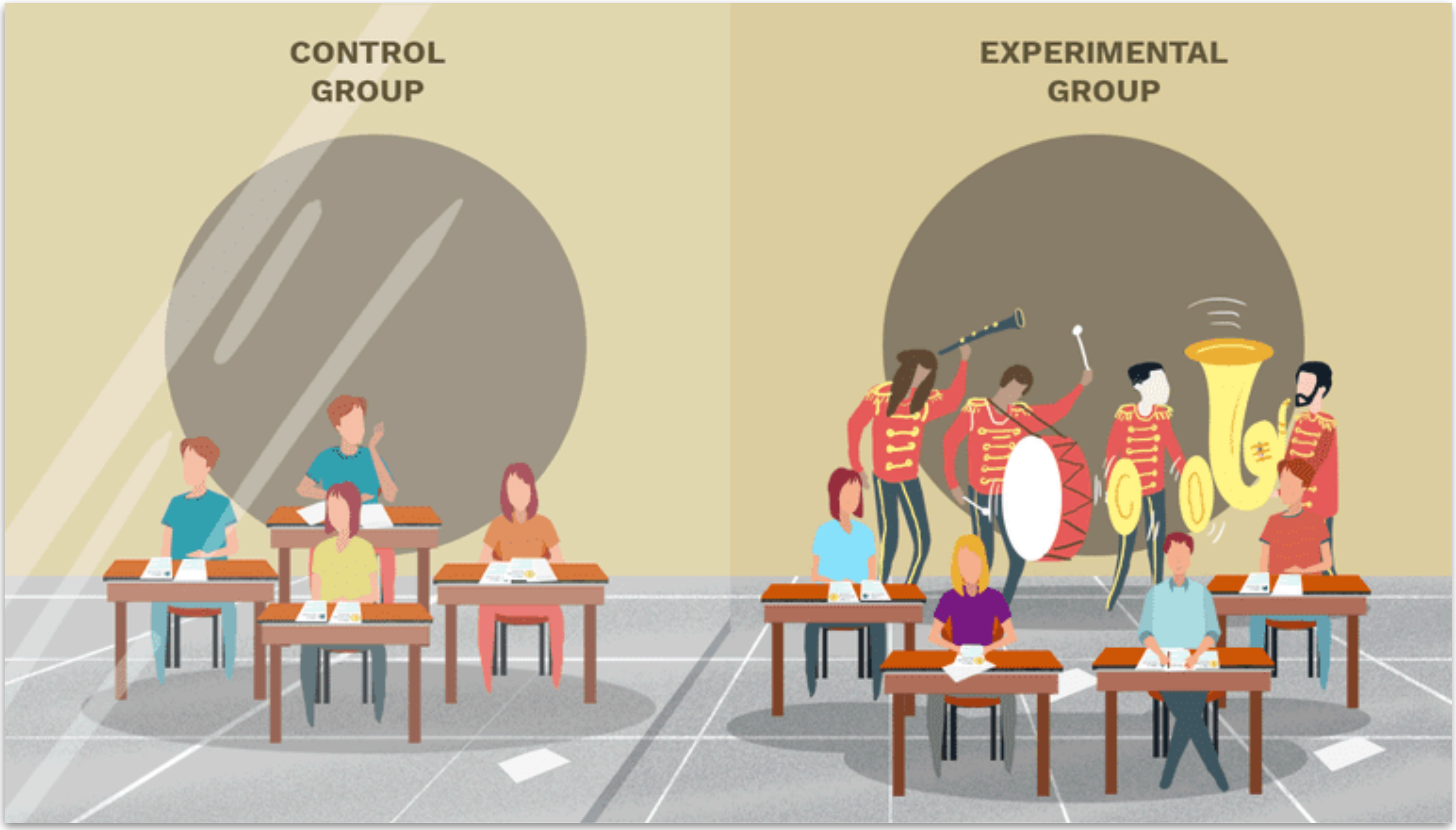
$D =$ Data

Logic of inference

- calculate a **test statistic** based on the sample
  - for example, the difference between the means of two conditions
- build a **sampling distribution** of this statistic *assuming that the null hypothesis is true*
  - use math or resampling methods
- **calculate the probability** of the observed statistic on the sampling distribution
- reject the null hypothesis if the probability of the observed statistic is less than the pre-specified α level

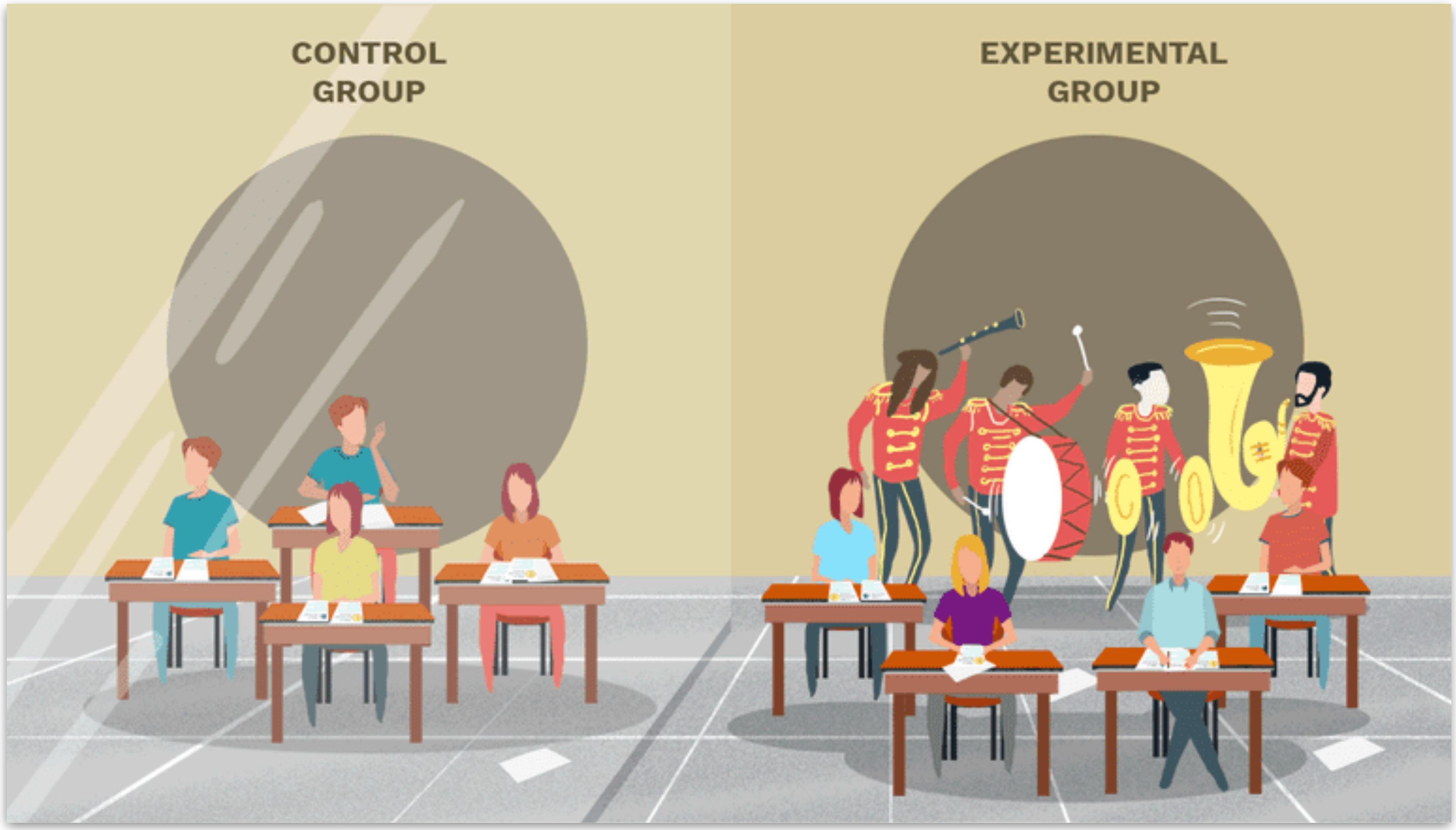
Permutation test

Permutation test



Research question:
Will student test scores be affected by distracting sounds (e.g. the Stanford marching band)?

Permutation test



$$H_0 : \mu_{\text{control}} = \mu_{\text{experimental}}$$

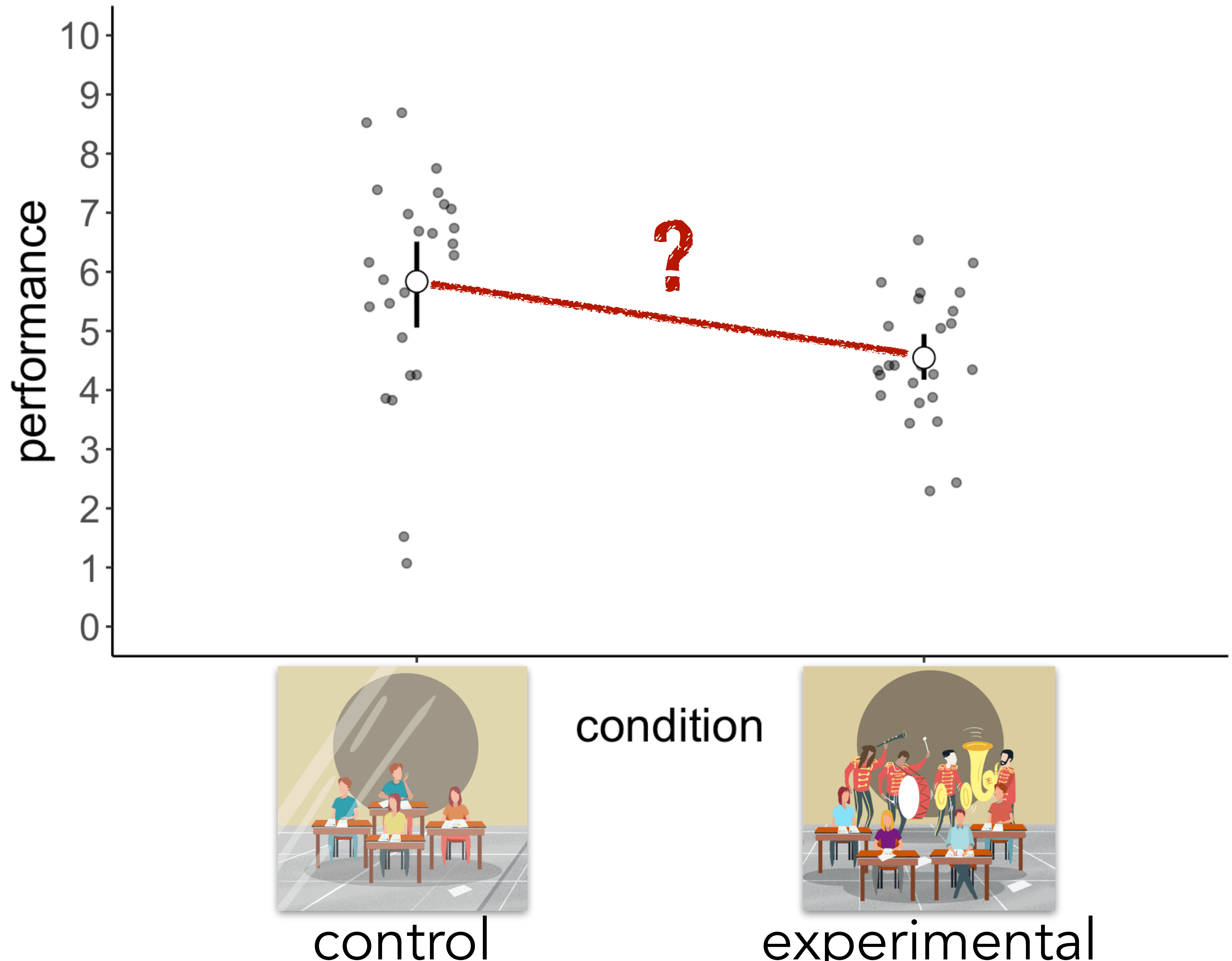
There is no difference between the control group and the experimental group

$$H_1 : \mu_{\text{control}} > \mu_{\text{experimental}}$$

Performance in the control group is better than in the experimental group

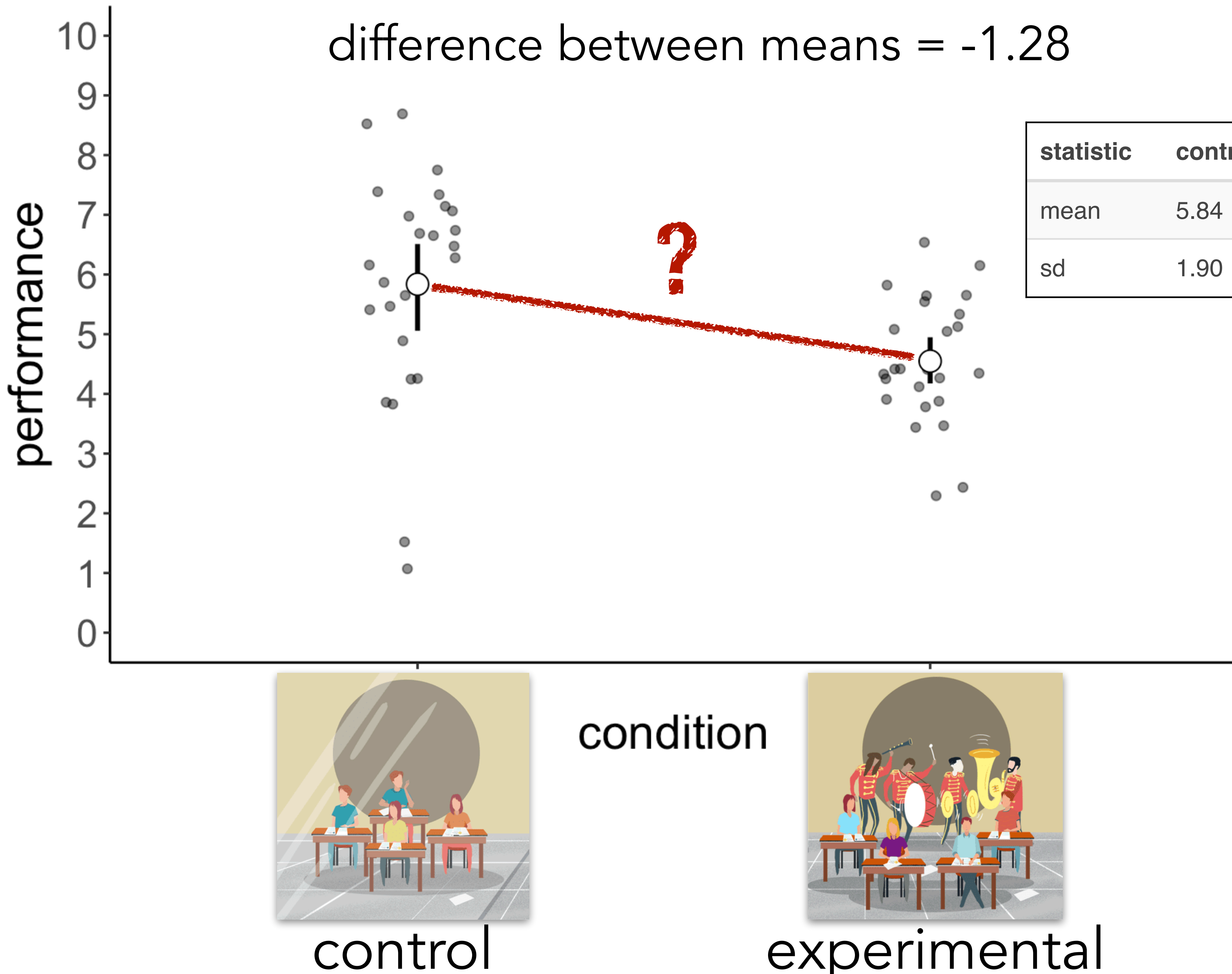
Permutation test

Is the difference in performance statistically significant?



Permutation test

Is the difference in performance statistically significant?



statistic	control	experimental
mean	5.84	4.55
sd	1.90	1.06

Permutation test

- **logic**:
    - assuming our experimental manipulation made no difference, what would be the probability of observing the data that we did?
    - if, assuming that the null hypothesis is true, the probability of observing the data (or data that is more extreme) is less than 5%, we reject the null hypothesis

we need a sampling distribution of our test statistic (difference between means)

Permutation test

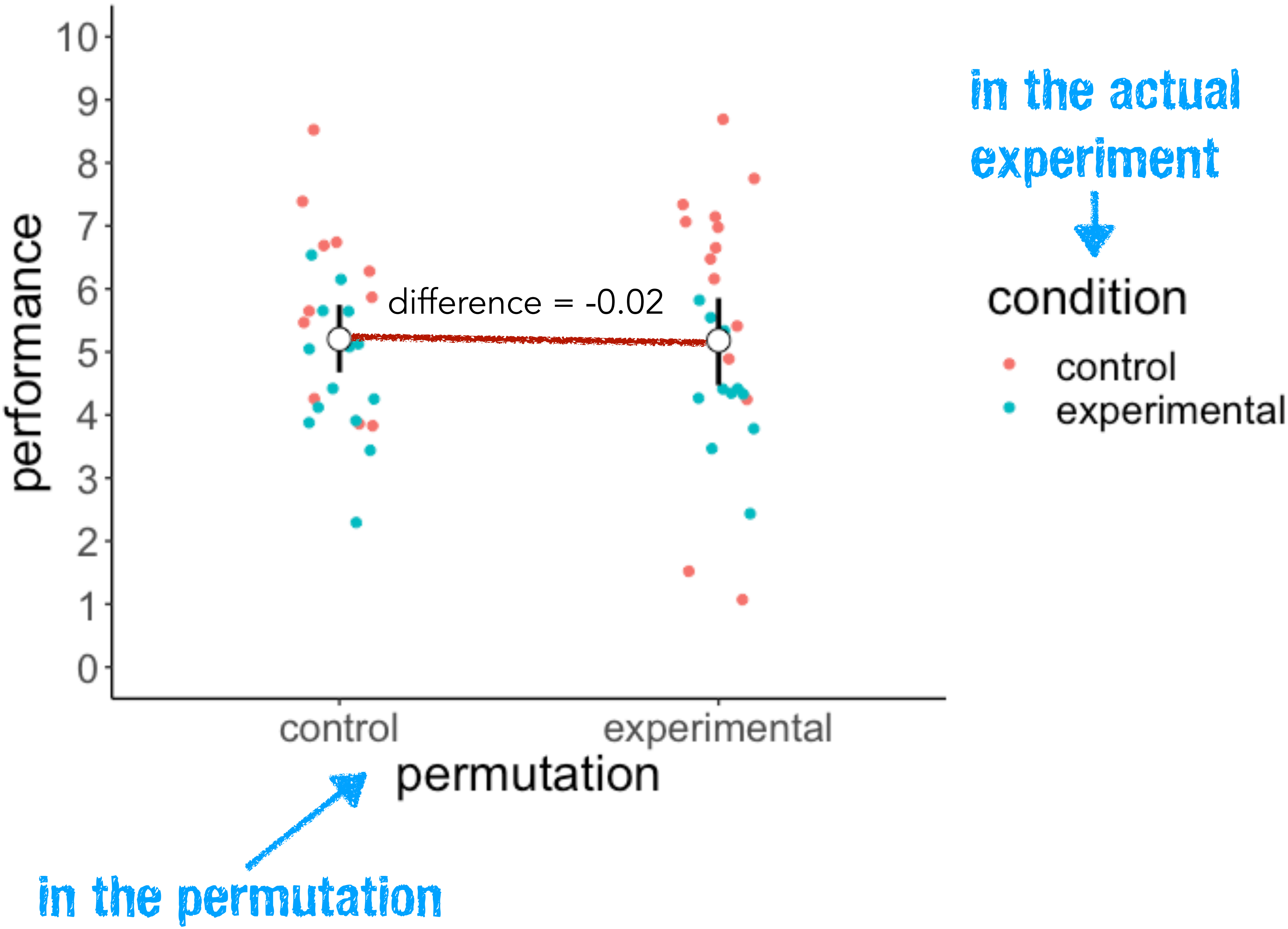
observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

random permutation

participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	experimental	3.78
30	experimental	5.13

Permutation test



Permutation test

observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

1

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	control	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

2

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	experimental	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	control	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

3

participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	experimental	3.78
30	experimental	5.13

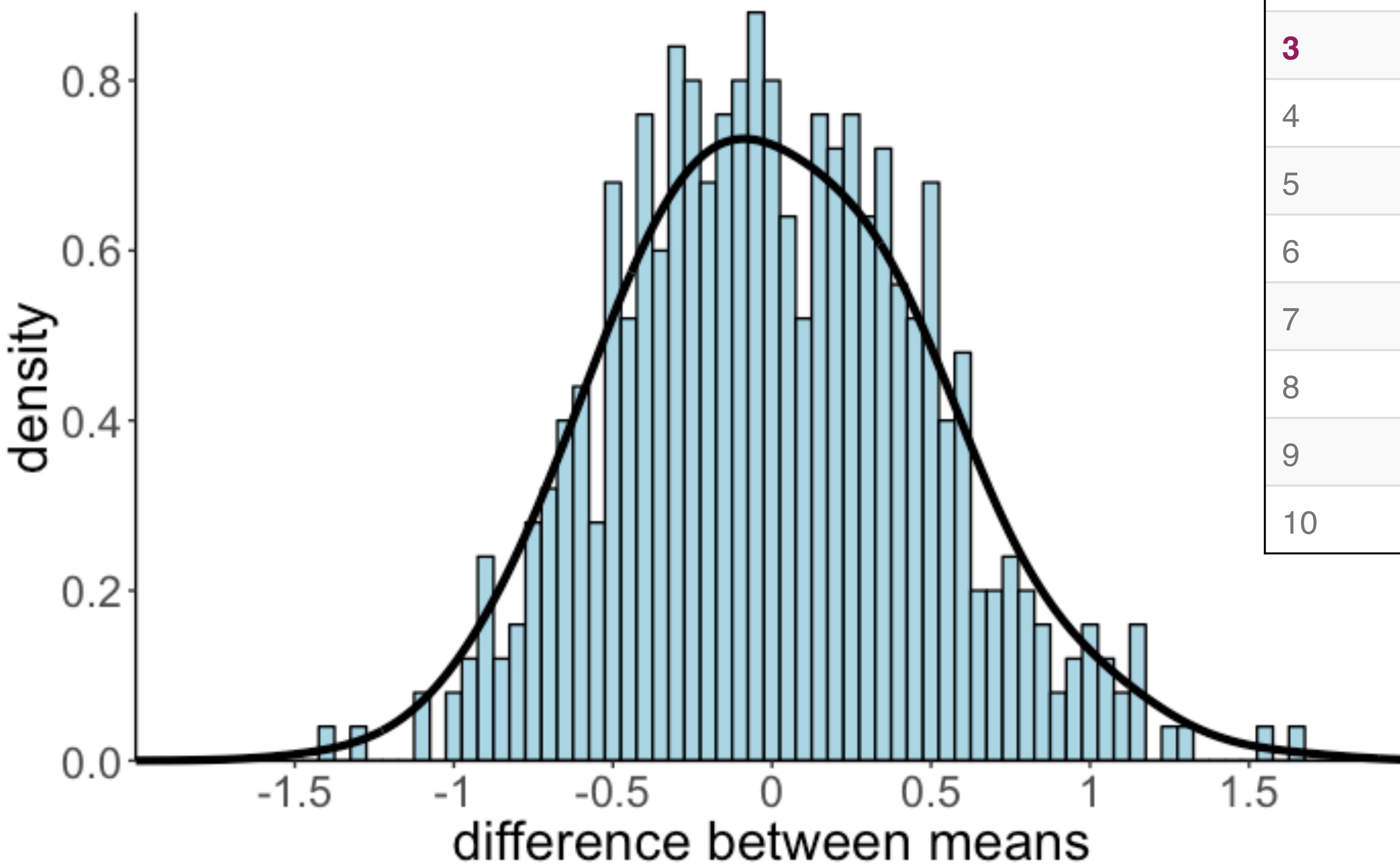
⋮

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

Permutation test

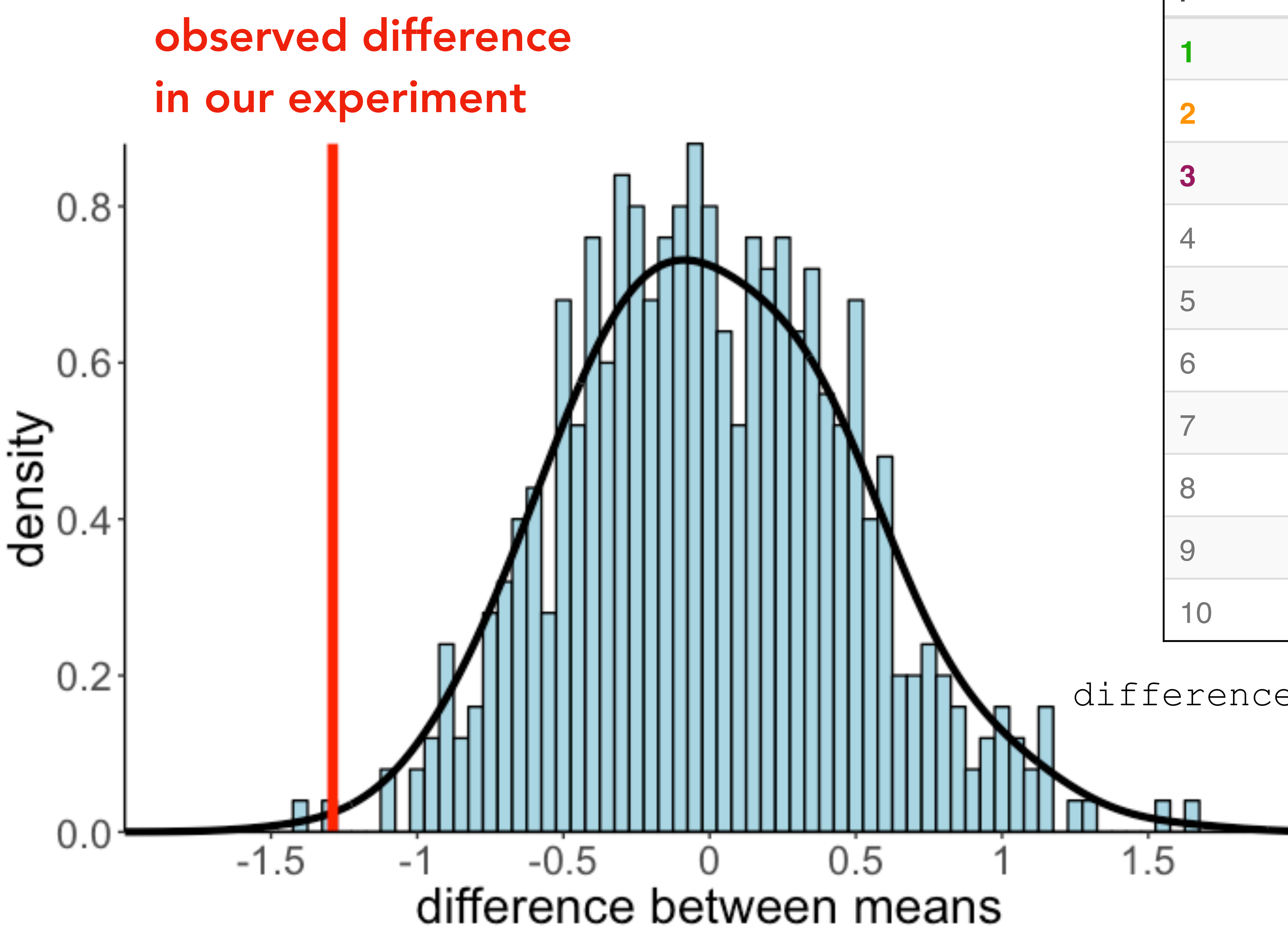
df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31



Sampling distribution of differences
(expected differences if the null hypothesis was true)

Permutation test



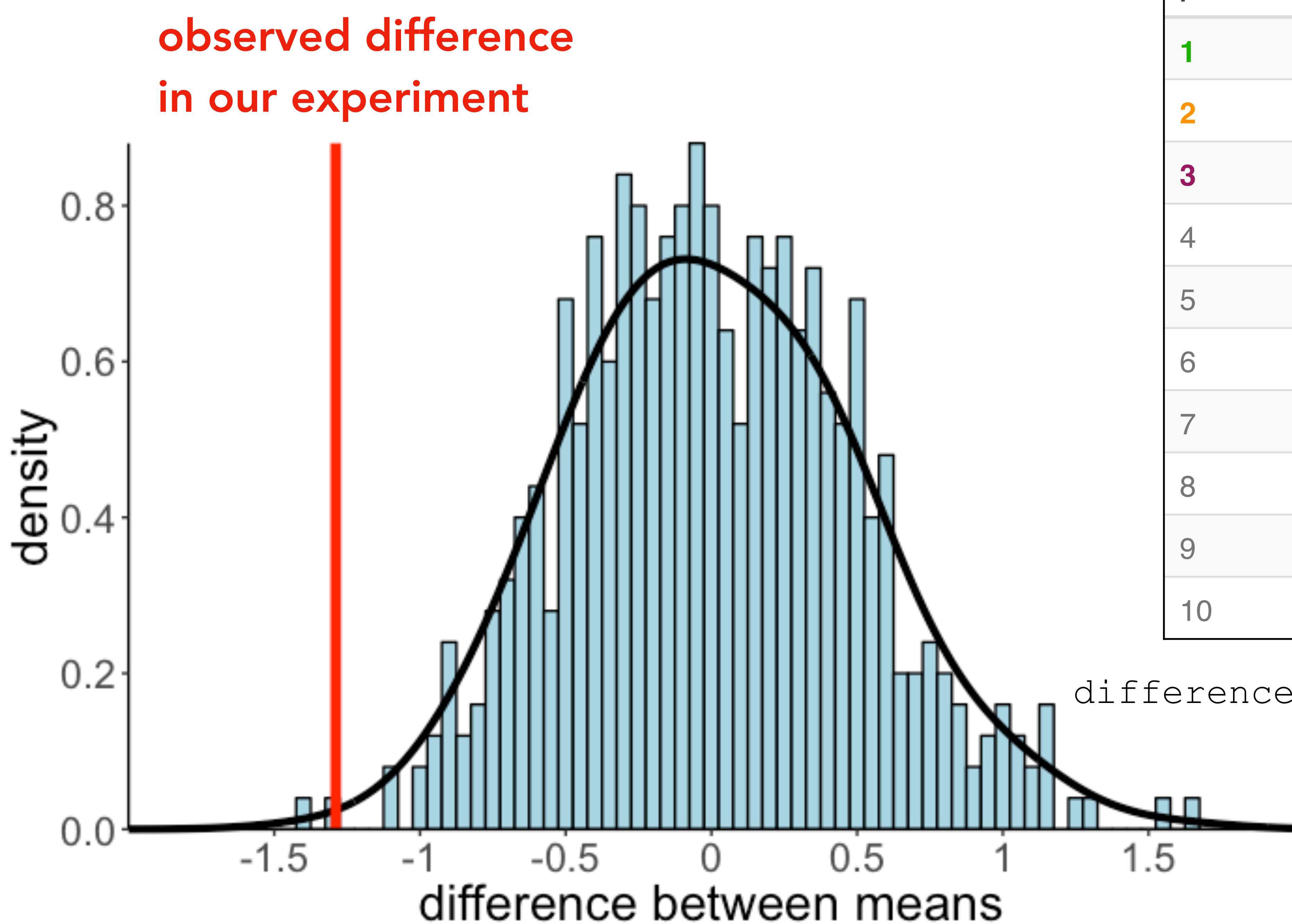
df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

`difference_actual = -1.28`

Sampling distribution of differences
(expected differences if the null hypothesis is true)

Permutation test



df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

difference_actual = -1.28

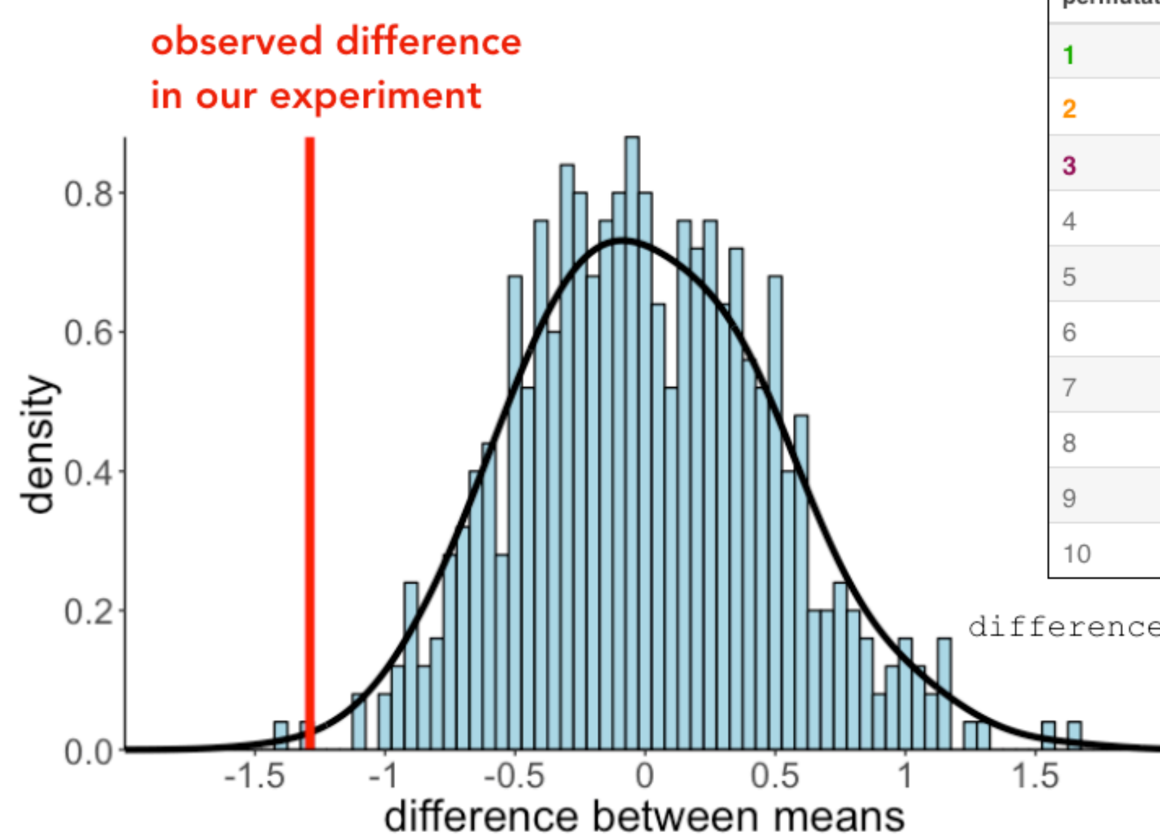
```
#calculate p-value of our observed result
df.permutations %>%
  summarize(p_value = sum(mean_difference <= difference_actual)/n())
```

p-value = .002

What is a p-value?

The **p-value** is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) is true.

Permutation test



df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

```
difference_actual = -1.28
```

```
#calculate p-value of our observed result
df.permutations %>%
  summarize(p_value = sum(mean_difference <= difference_actual)/n())
```

p-value = .002

57

Permutation test

```r
n_permutations = 500

# permutation function
func_permutations = function(df){
  df %>%
    mutate(condition = sample(condition)) %>%
    group_by(condition) %>%
    summarize(mean = mean(performance)) %>%
    pull(mean) %>%
    diff()
}
```

set the number of permutations

shuffle the condition labels

calculate difference between group means

observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

random permutation

participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	experimental	3.78
30	experimental	5.13

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	control	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	experimental	4.27
28	control	2.20
29	control	3.78
30	experimental	5.13

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	experimental	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	control	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.20
29	experimental	3.78
30	experimental	5.13

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

Permutation test

```r
n_permutations = 500

# permutation function
func_permutations = function(df){
  df %>%
    mutate(condition = sample(condition)) %>%
    group_by(condition) %>%
    summarize(mean = mean(performance)) %>%
    pull(mean) %>%
    diff()
}

# data frame with permutation results
df.permutations = tibble(
  permutation = 1:n_permutations,
  mean_difference = replicate(n = n_permutations, func_permutations(df.data))
)
```

df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

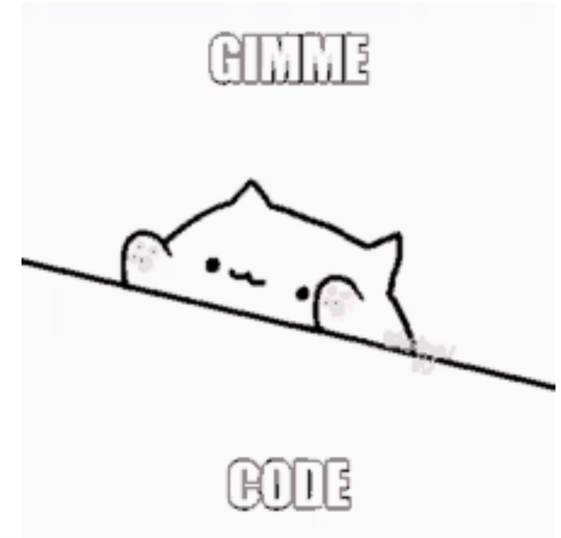
run the `func_permutations()` function many times

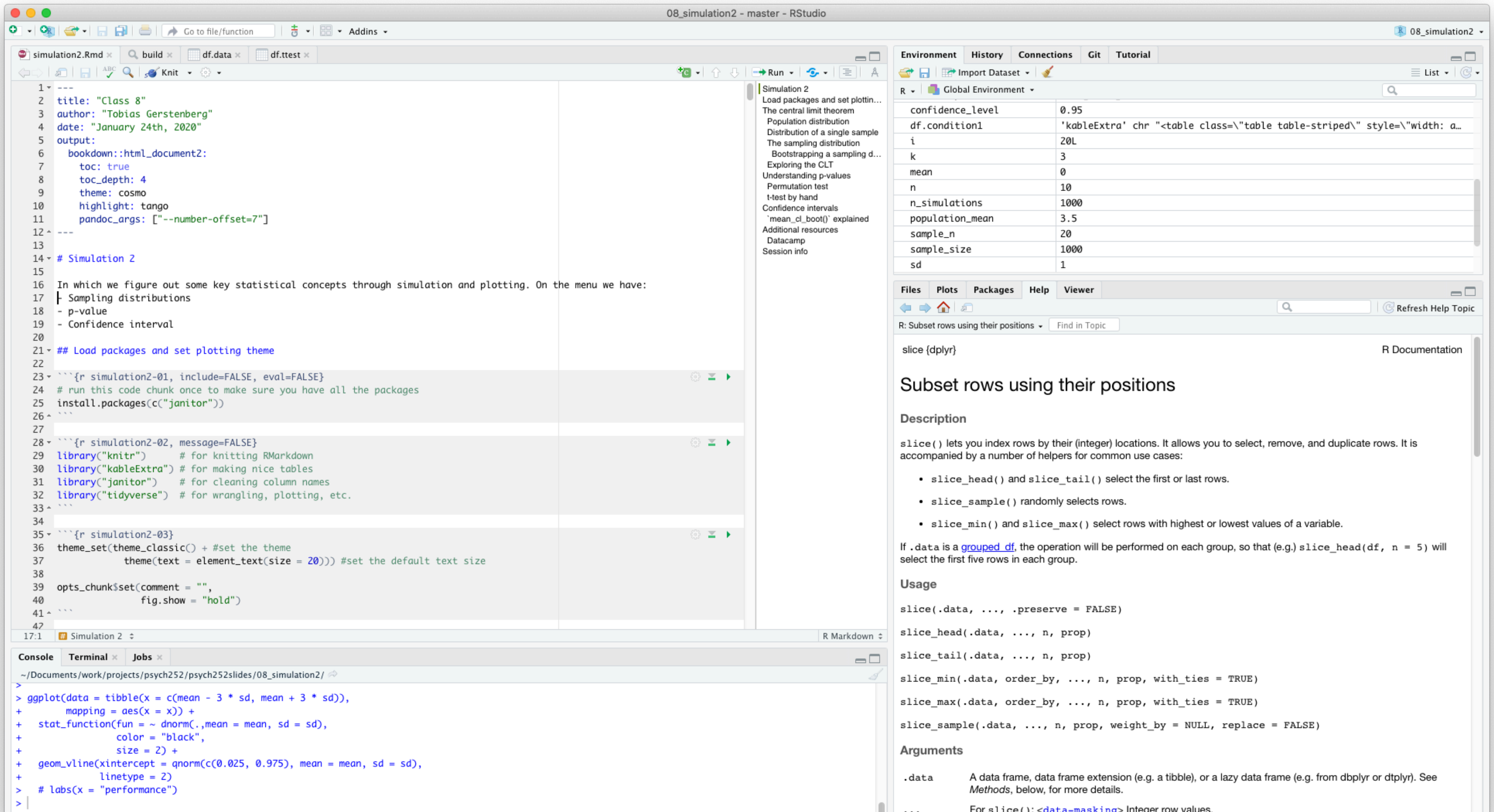
(instead of using a for loop)

Summary Revisit and understand key statistical concepts

- **Inference in frequentist statistics**
  - goal is to make inference from sample to population
  - we do so via a complicated procedure that involves sampling distributions
- **Sampling distributions**
  - the link between sample and population in frequentist statistics
  - theoretical (or simulated) distribution of a test statistic
- **What is a p-value?**
  - the probability of the observed test result (or a more extreme result) assuming that the H_0 is true
- **Confidence interval (of the mean)**
  - "If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the true mean."

How to better understand!





Datacamp

INTERACTIVE COURSE

Foundations of Inference

Continue Course

4 hours | 17 Videos | 58 Exercises | 12,551 Participants | 4,350 XP

Course Description

One of the foundational aspects of statistical analysis is inference, or the process of drawing conclusions about a larger population from a sample of data. Although counter intuitive, the standard practice is to attempt to disprove a research claim that is not of interest. For example, to show that one medical treatment is better than another, we can assume that the two treatments lead to equal survival rates only to then be disproved by the data. Additionally, we introduce the idea of a p-value, or the degree of disagreement between the data and the hypothesis. We also dive into confidence intervals, which measure the magnitude of the effect of interest (e.g. how much better one treatment is than another).

1 Introduction to ideas of inference FREE 100%

In this chapter, you will investigate how repeated samples taken from a population can vary. It is the variability in

This course is part of these tracks:

Intro to Statistics with R

Jo Hardin

Professor at Pomona College

https://www.datacamp.com/courses/foundations-of-inference

Feedback

How was the pace of today's class?

much too slow

a little too slow

just right

a little too fast

much too fast

How happy were you with today's class overall?

What did you like about today's class? What could be improved next time?

Thank you!