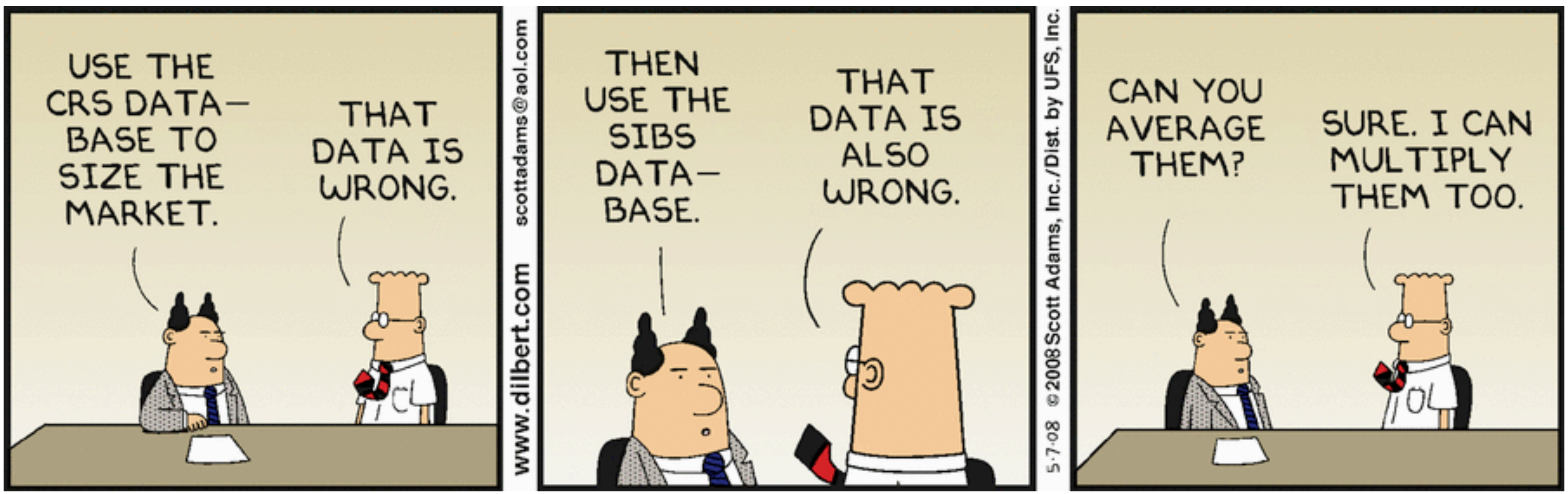
Simulation 2

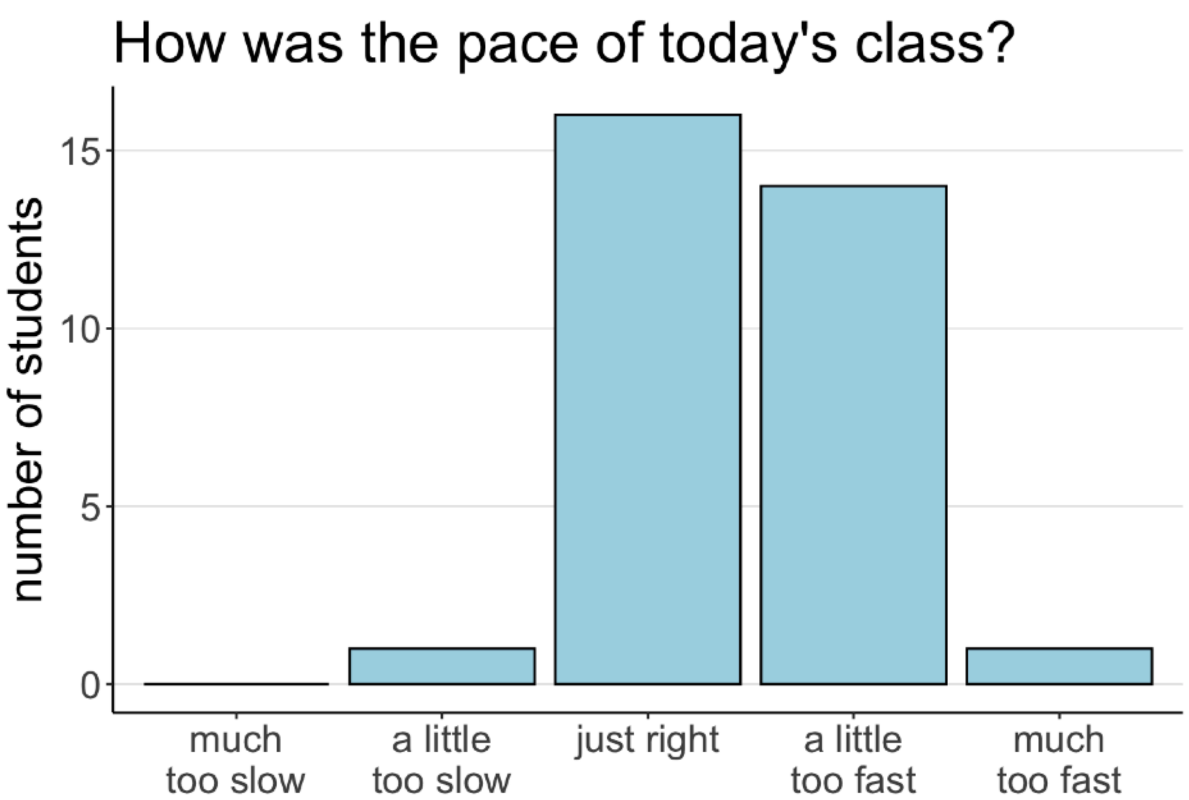


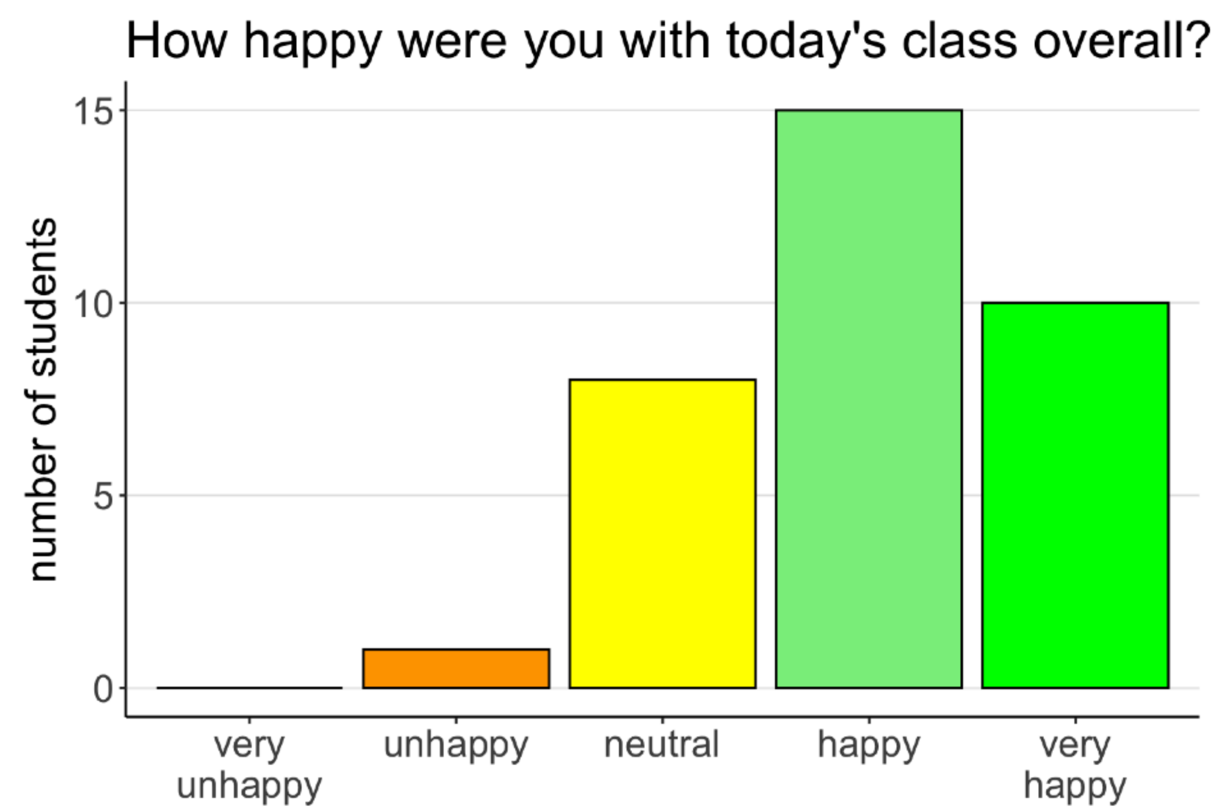
01/24/2020

Logistics

Your feedback

Your feedback





please slow me down if i go to fast!

Your feedback

Homework 2 …

Dataset description

You'll be working with a dataset collected in the *high throughput psychophysics* experimental tradition (Rajalingham et al. 2018). One variant of this approach is to collect data from hundreds, even thousands of participants, in order to generate reliable estimates of perceptual behaviors. Typically, this is accomplished using online data collection. In this dataset there were originally 345 subjects who collectively performed 42,113 trials of a visual discrimination task.

In order to generate the main plot, we'll analyze a subset of this human behavior. When this is done, we'll compare human performance with the behavior of neurons and a computational model performing the same task. The neural data is from a neuroscience/psychophysics experiment conducted in non-human primates (Majaj et al. 2015). The computational model data, in our case, was used to validate its correspondence with visual area **IT**. If you're interested in this model class, (Yamins and DiCarlo 2016) provides a nice intro. The basic idea is that we can use biologically plausible computational models to understand the relationship between the brain and behavior.

These data were collected to generate a stimulus set with an interesting pattern of perceptual properties:

1. Early visual regions (e.g. primate area **V4**) can't do the task at all
2. Late visual regions (e.g. primate aread **IT**) can do the task on some items
3. Neurotypical humans can do the task on all items (outperforming **IT** and **V4**)

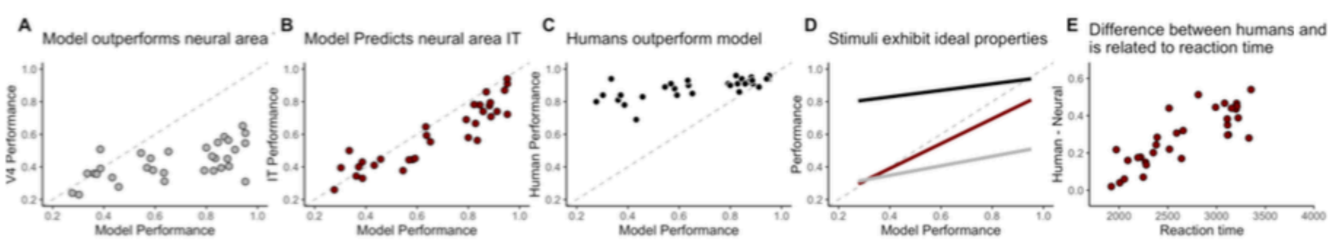


Figure 1: The beautiful plot you want to recreate.

… was challenging!

make sure to take advantage of the different resources

Piazza

Homework section

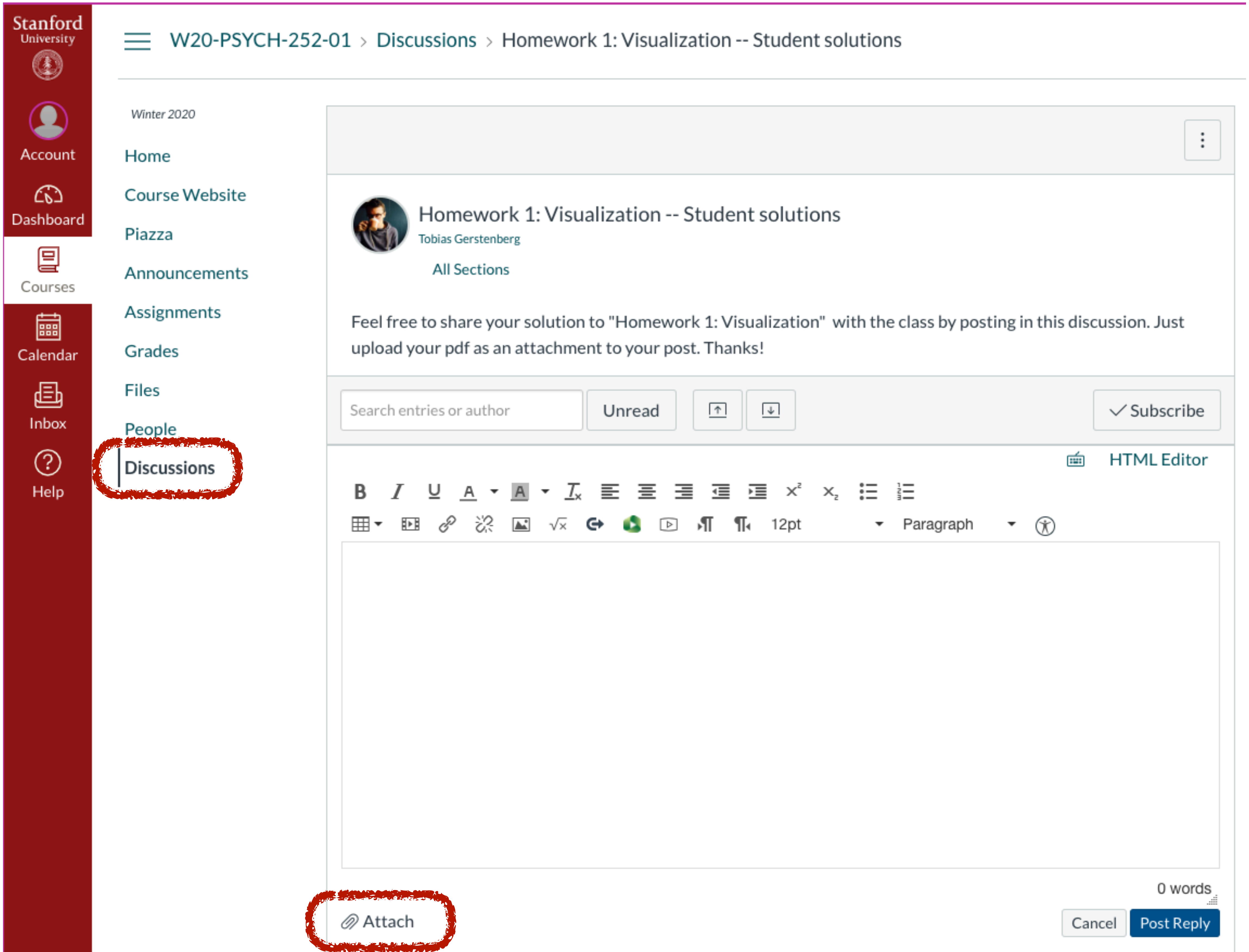
Course notes

Office hours

only students who take the class for credit should submit their homework

Homework 1

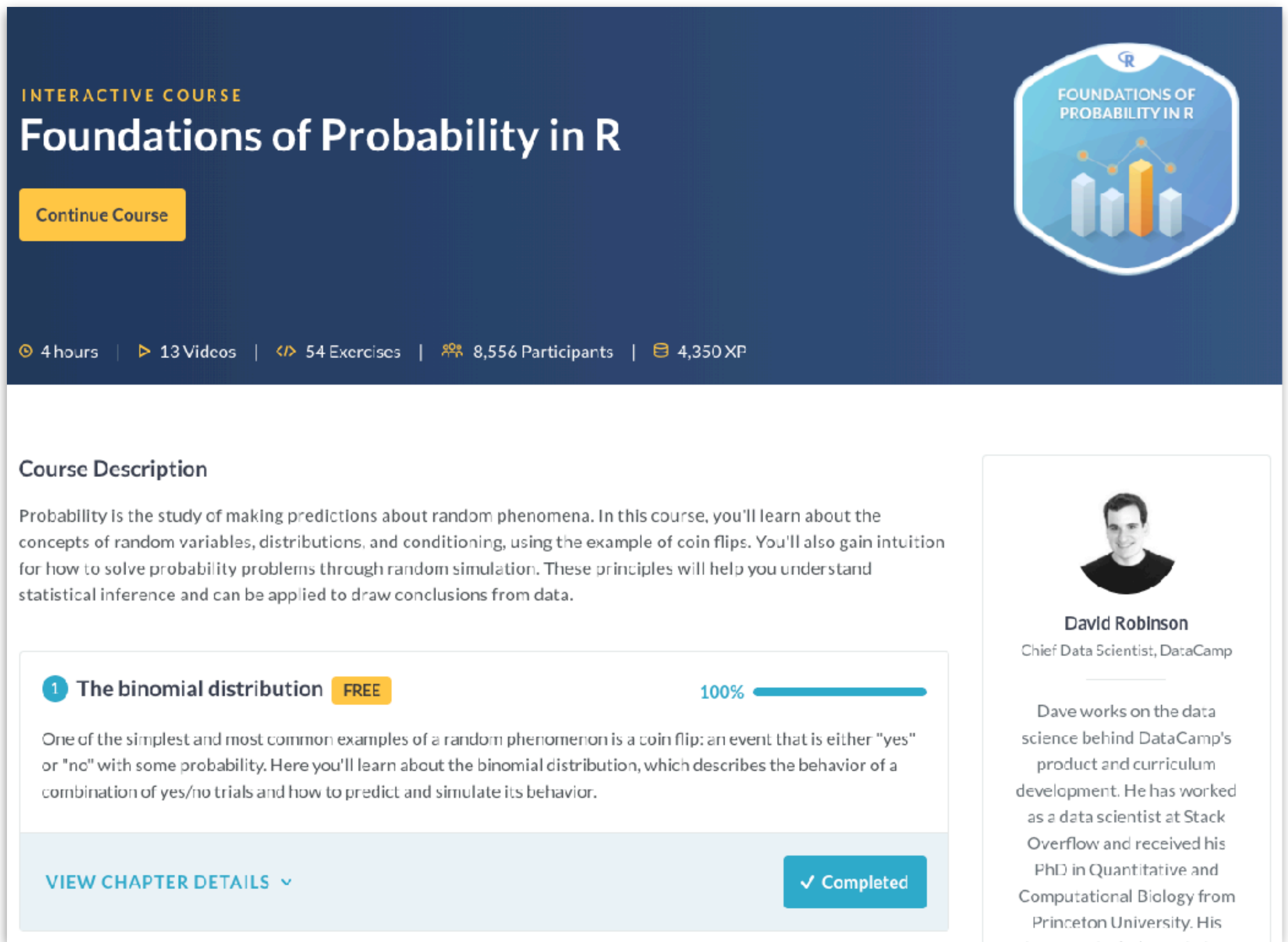
Upload your solution to HW1 (if you like)



Homework 3

Homework 3

Will be released after class today and is due this **Thursday (January 30th) at 8pm**.



Make sure to sign up for datacamp with your stanford.edu email address!

https://tinyurl.com/psych252datacamp20

Outline

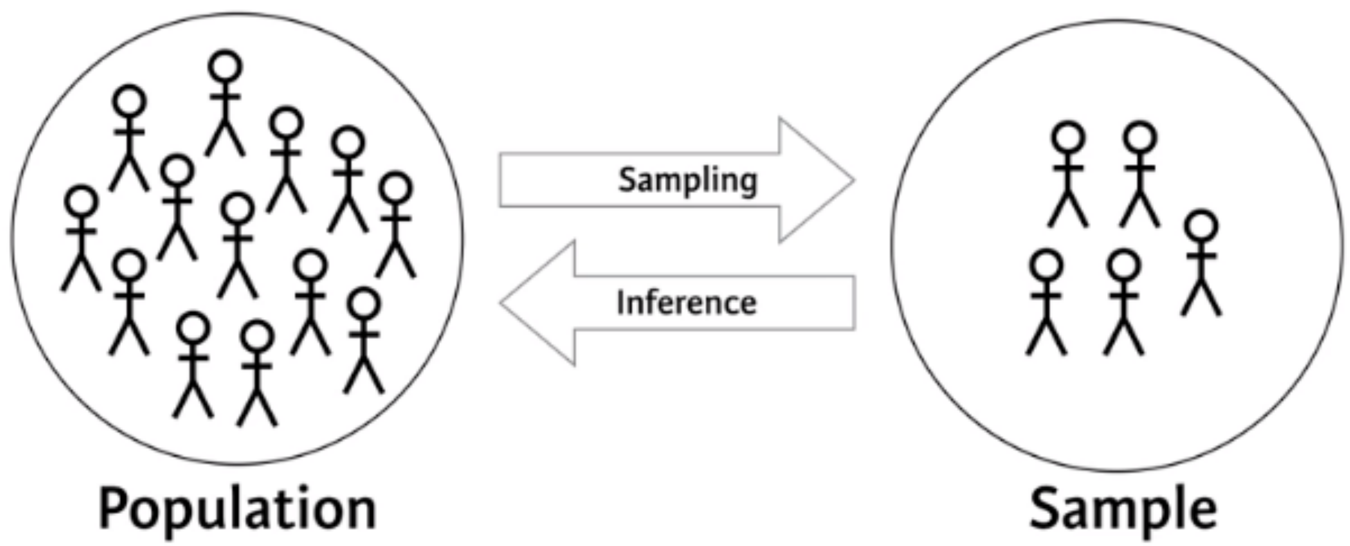
Goal: Revisit and understand key statistical concepts

- Statistical inference
- Central limit theorem
- Sampling distributions
- p-values
- Confidence intervals

Statistical inference

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?

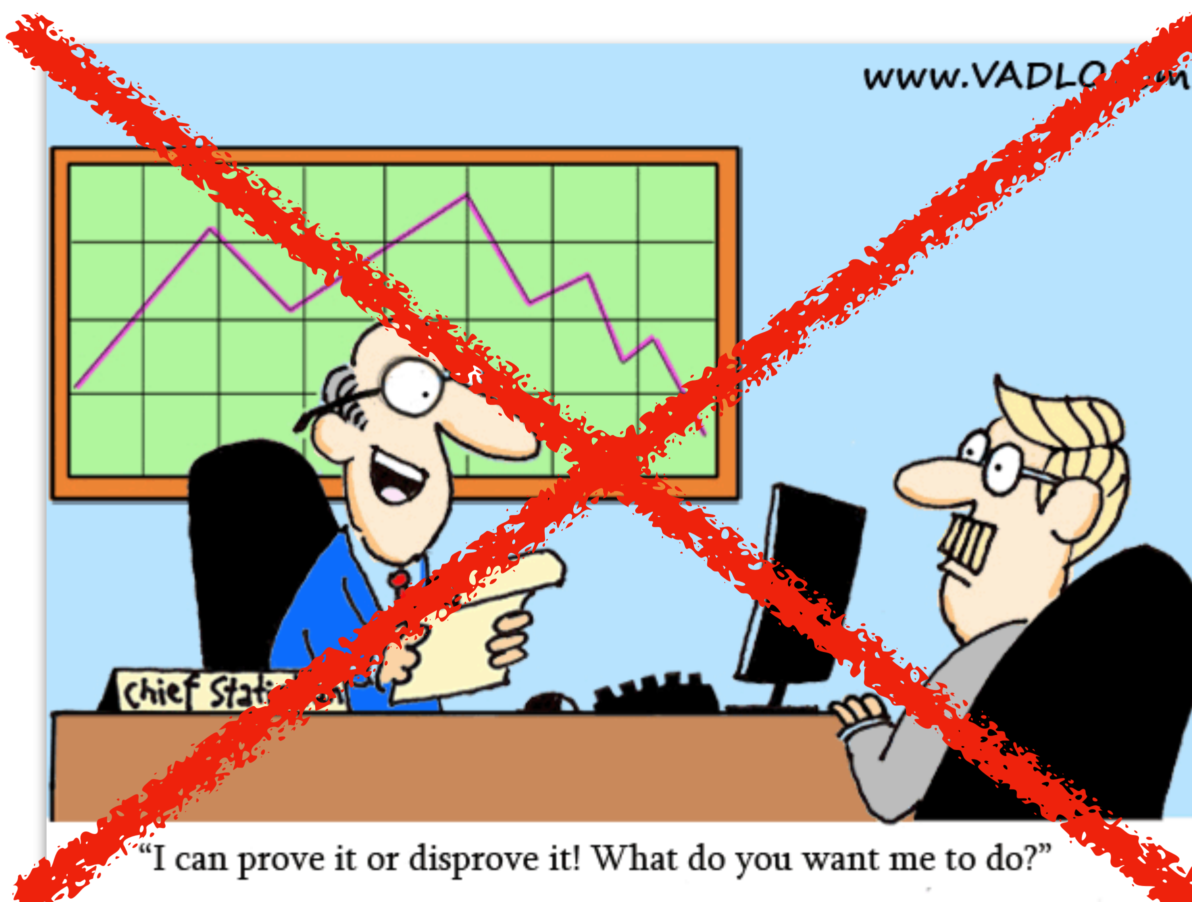
Statistical inference

Key question:

What can we infer about the population from our sample?

Answer:

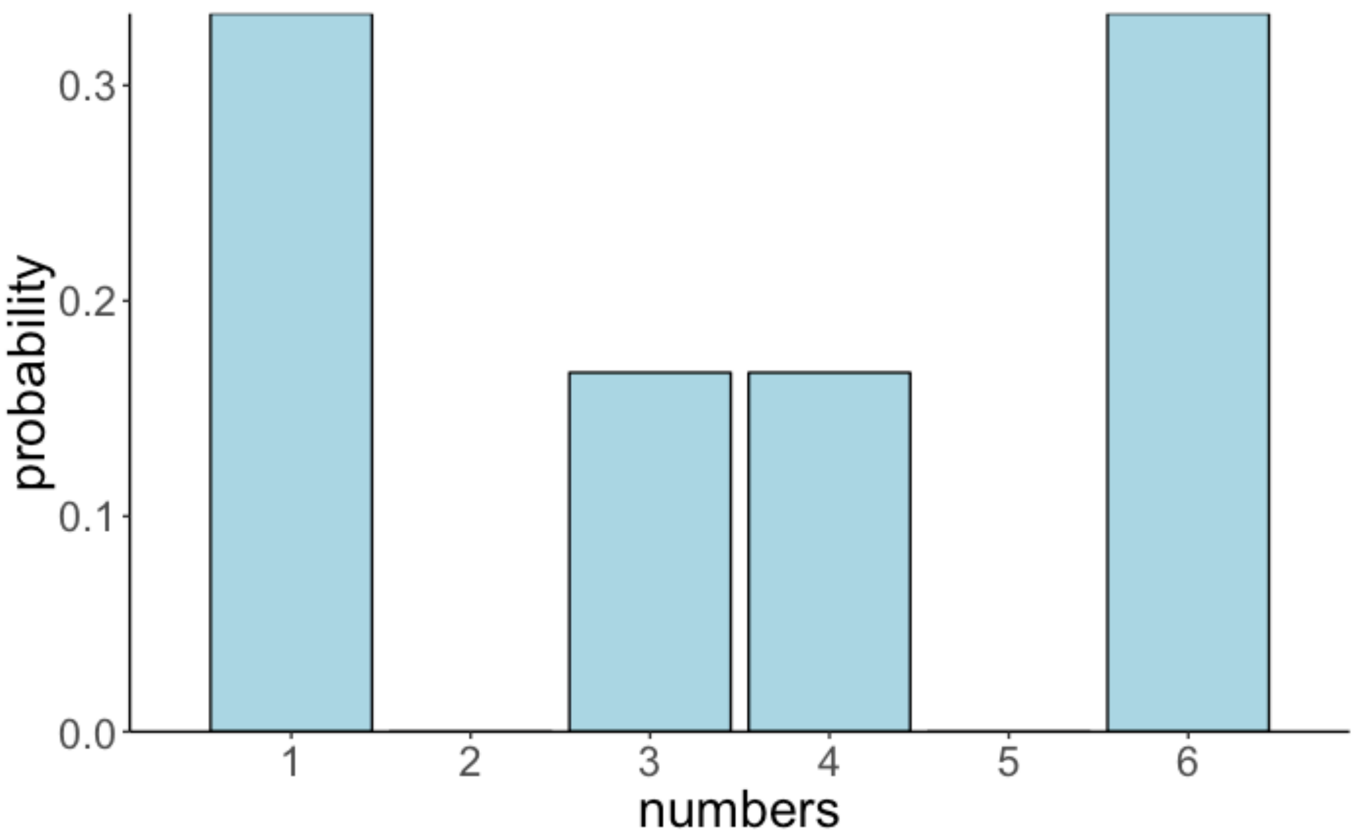
- is not trivial
- mathematical, statistical, philosophical (Bayesian vs. frequentist) machinery involved
- **important**: we can never make deterministic statements!
- we can only make probabilistic claims



Central limit theorem

Central limit theorem

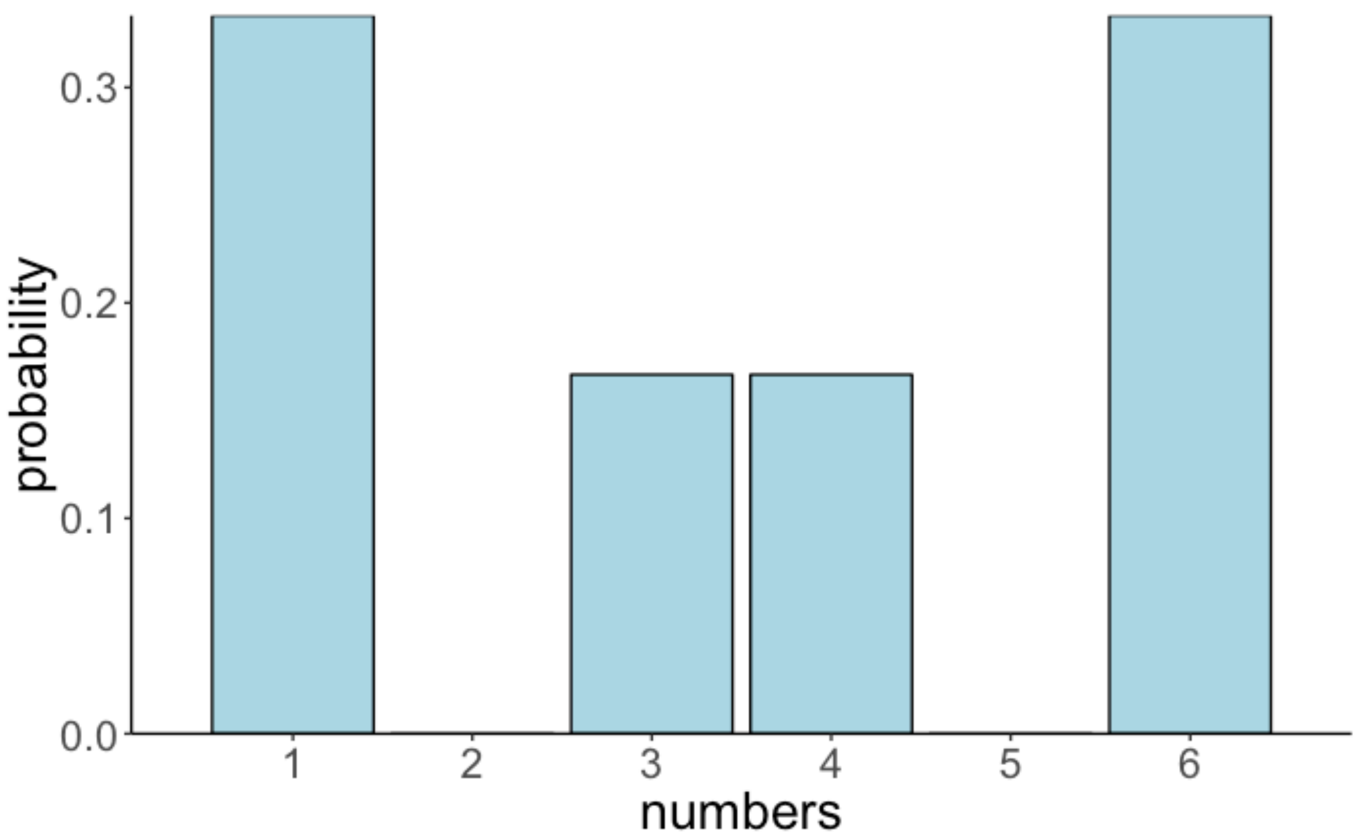
heavy metal distribution



population distribution

Central limit theorem

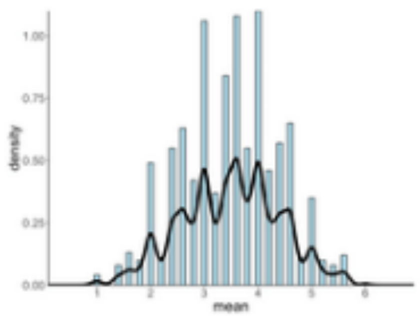
heavy metal distribution

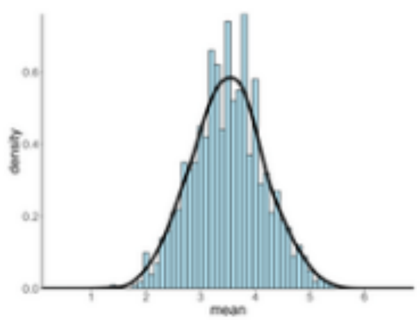


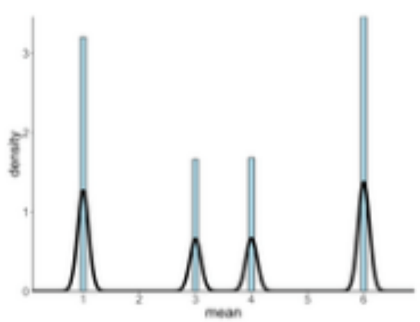
population distribution

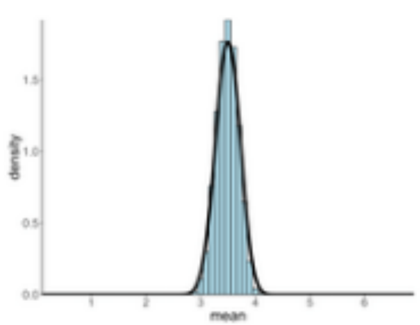
1. draw **1000 samples** of **size 1**
2. plot a histogram of the samples

What would the distribution look like? (N = 1)



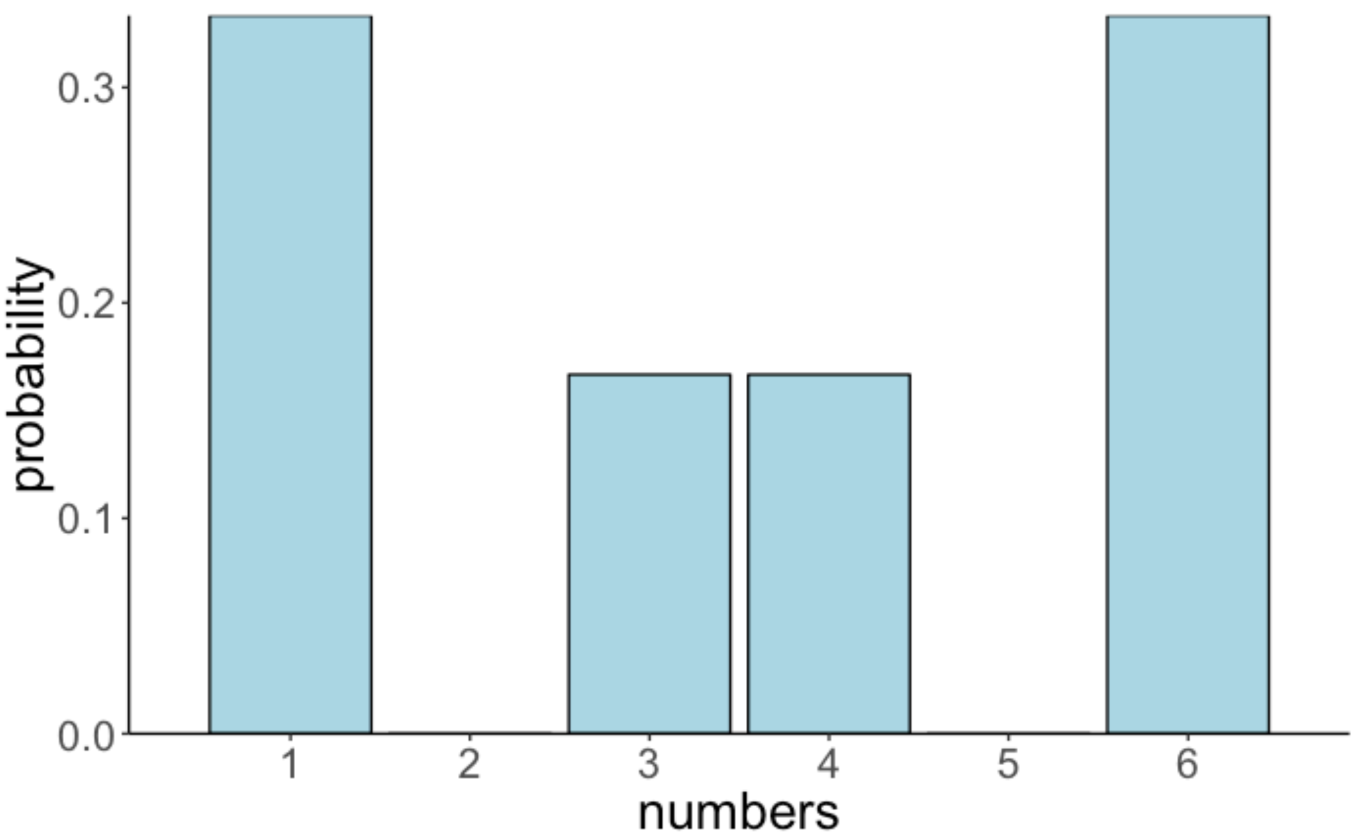






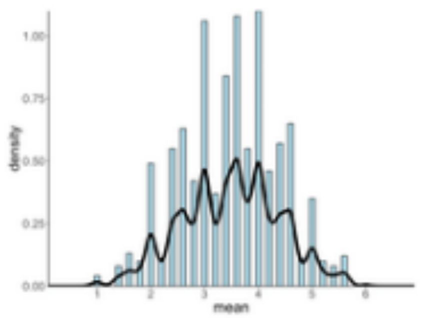
Central limit theorem

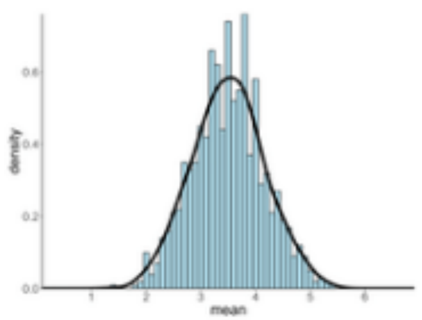
heavy metal distribution

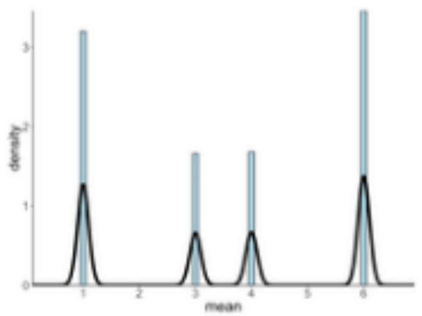


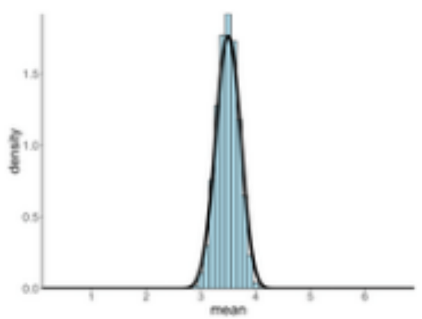
1. draw **1000 samples** of **size 100**
2. calculate the mean of each sample
3. plot a histogram of the sample means

What would the distribution look like? (N = 100)



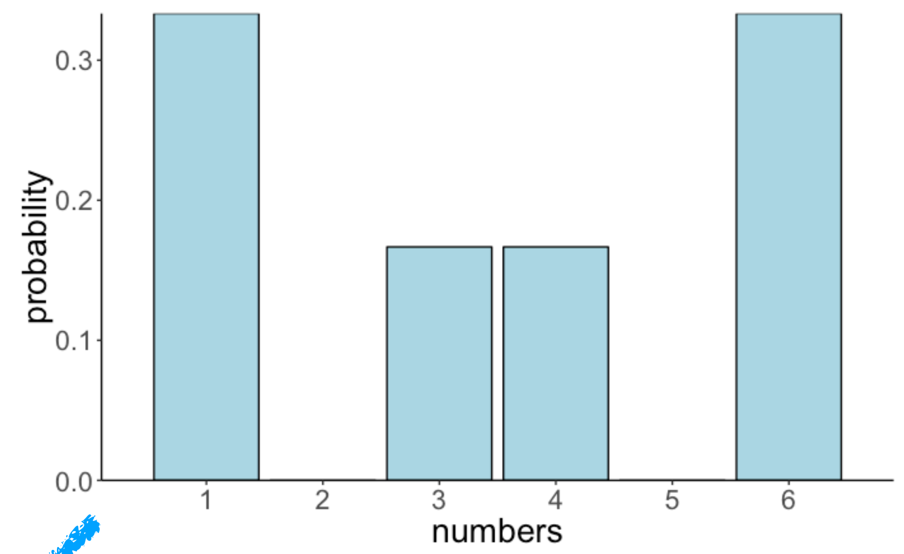






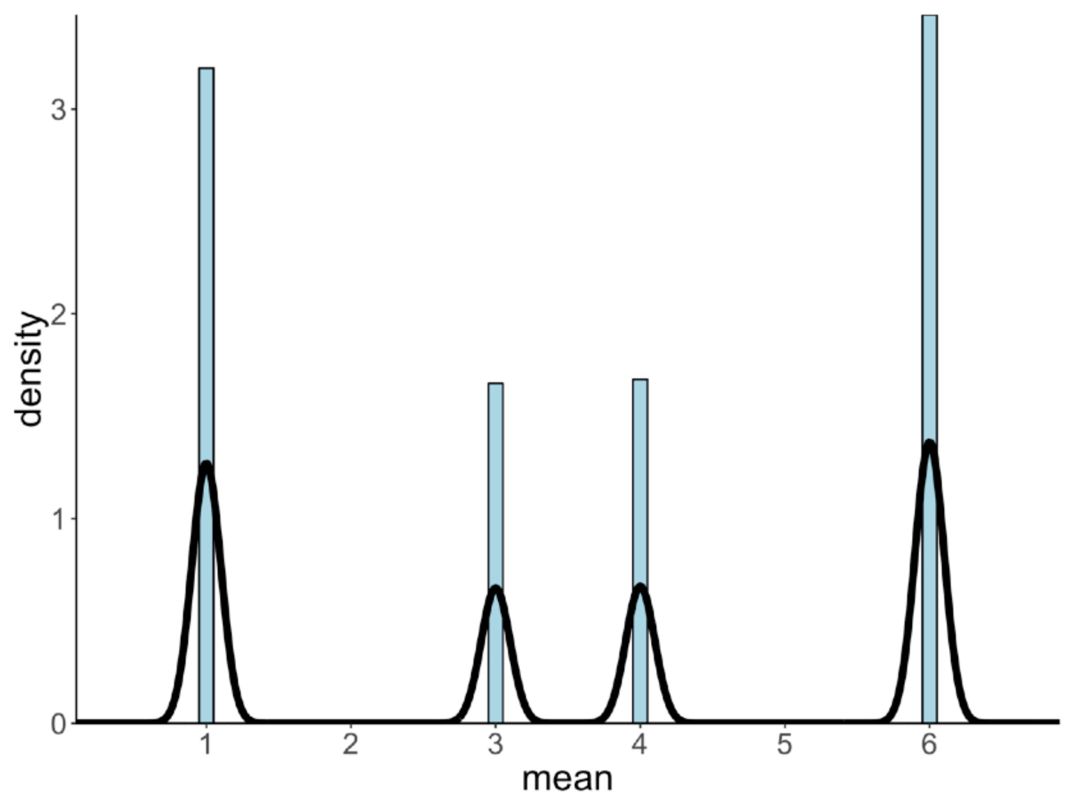
Central limit theorem

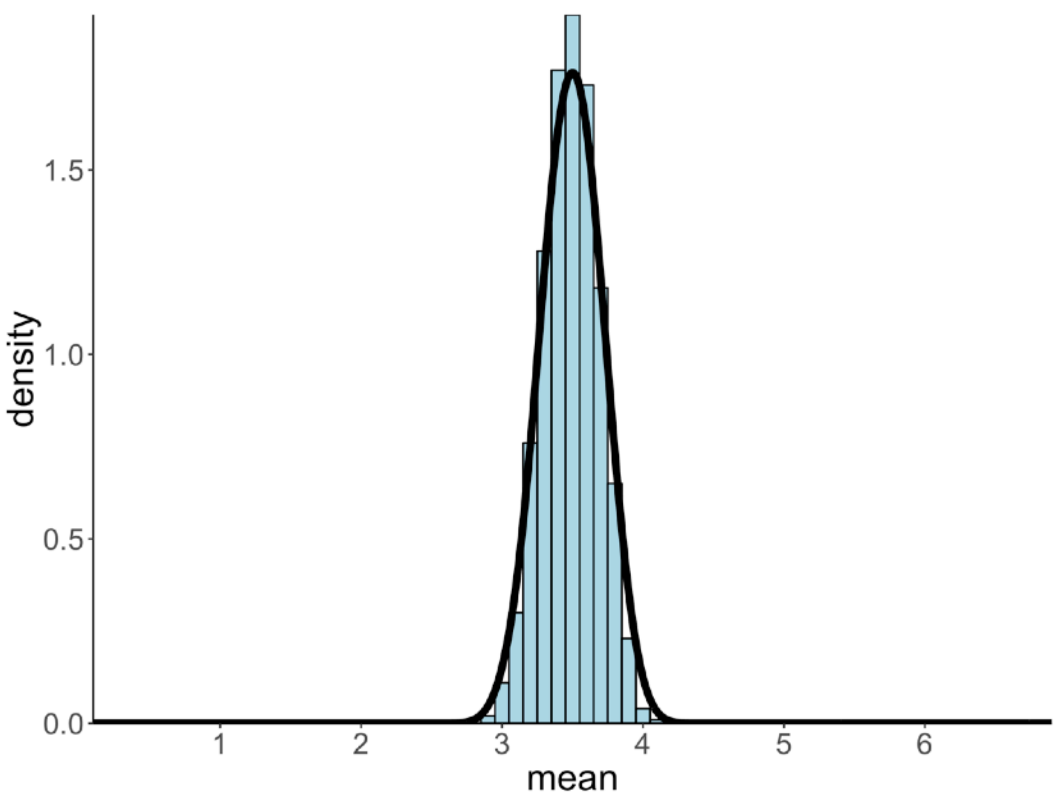
heavy metal distribution



sample size = 1

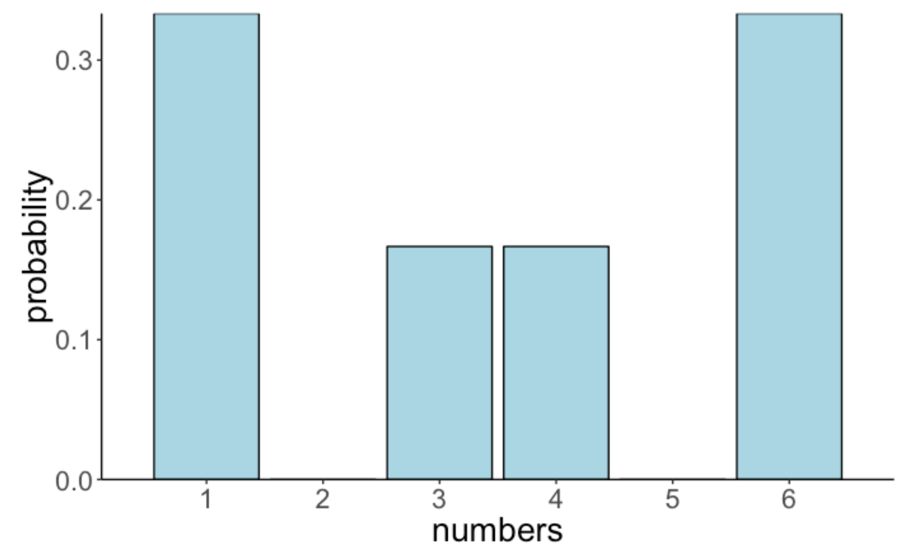
sample size = 100



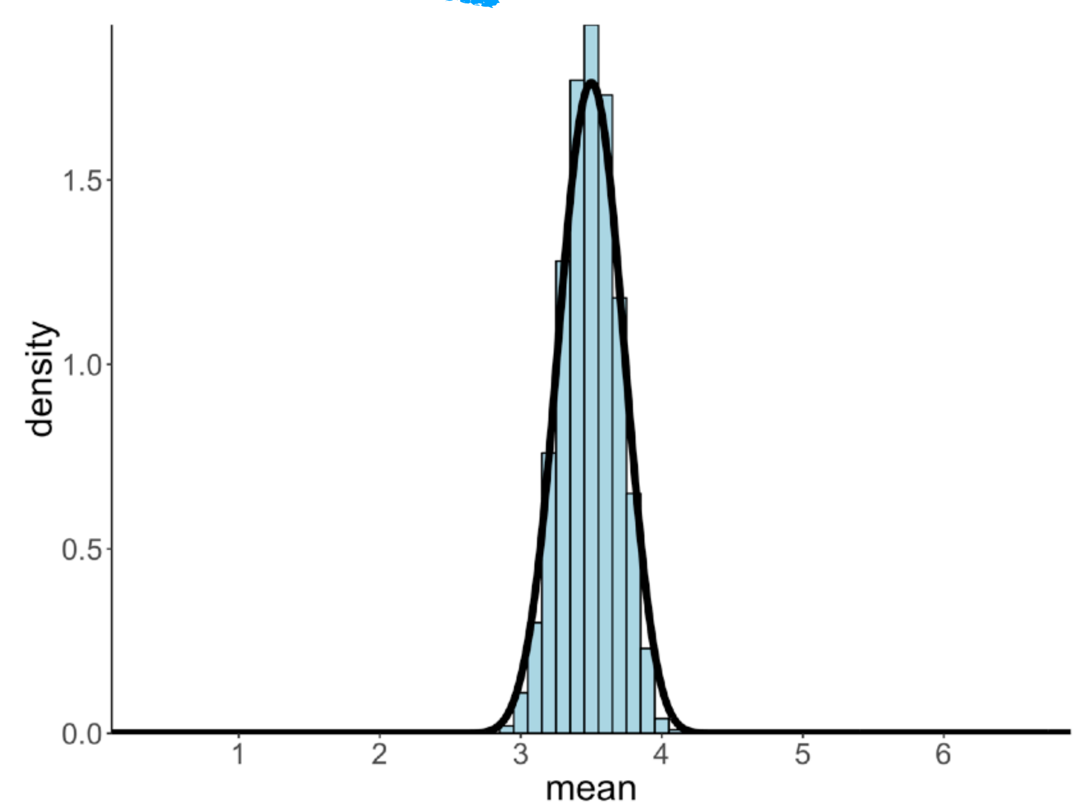


Central limit theorem

heavy metal distribution



sample size = 100



The Central Limit Theorem (CLT) states that the sample mean of a **sufficiently large number of independent and identically distributed (i.i.d.) random variables is approximately normally distributed**. The larger the sample, the better the approximation.

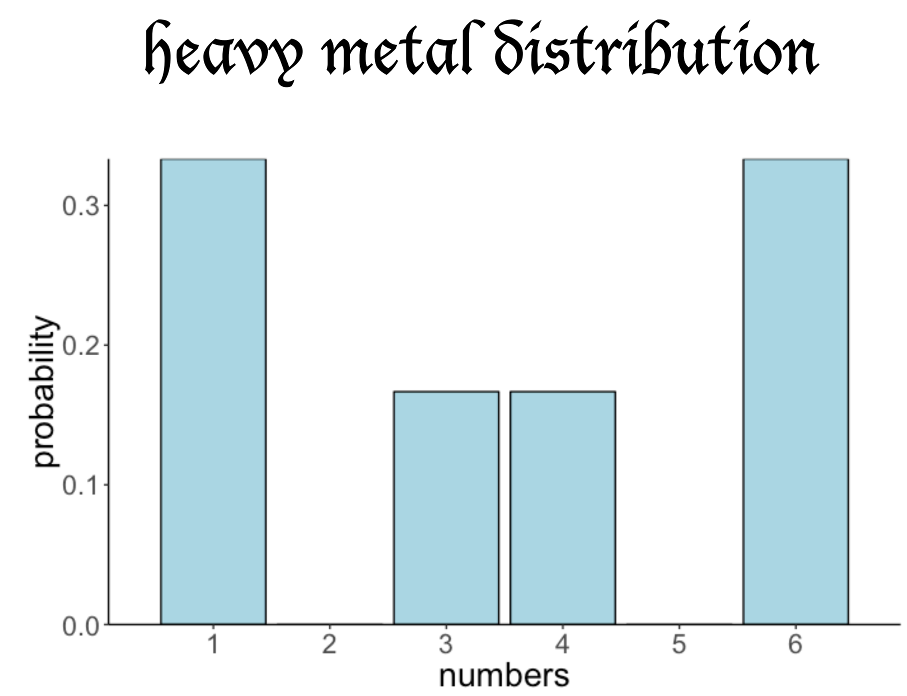
The theorem is a key ("central") concept in probability theory because it implies that **statistical methods that work for normal distributions can be applied to many problems involving other types of distributions**.

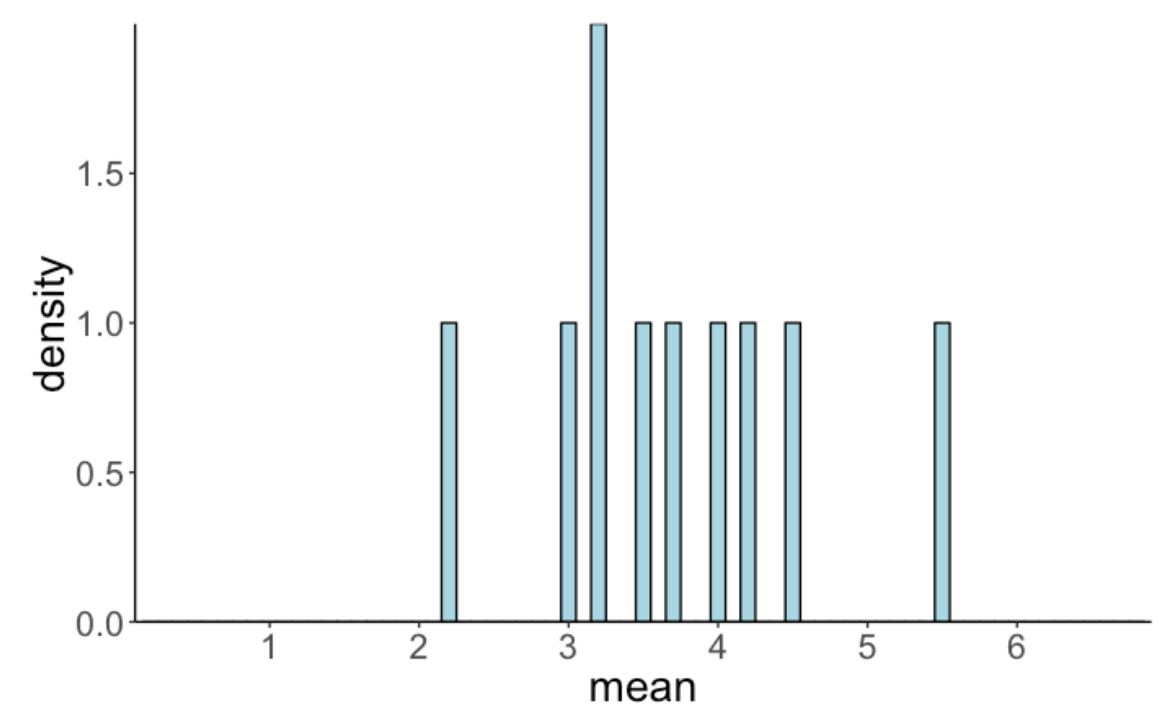
Central limit theorem

sample size = 4

number of samples = 10

sample	draw_1	draw_2	draw_3	draw_4
1	1	6	6	4

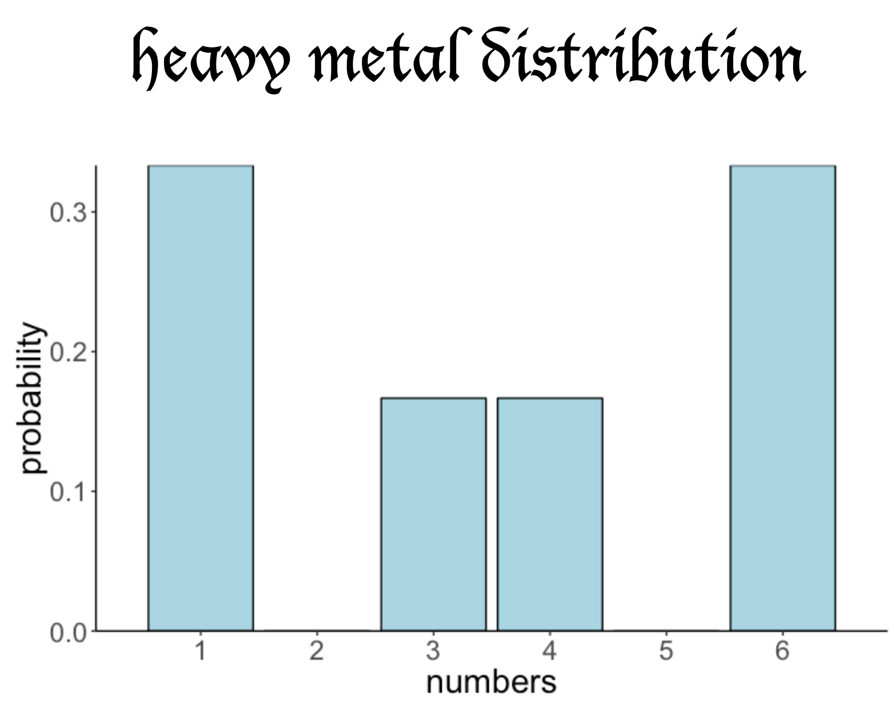




Central limit theorem

sample size = 100

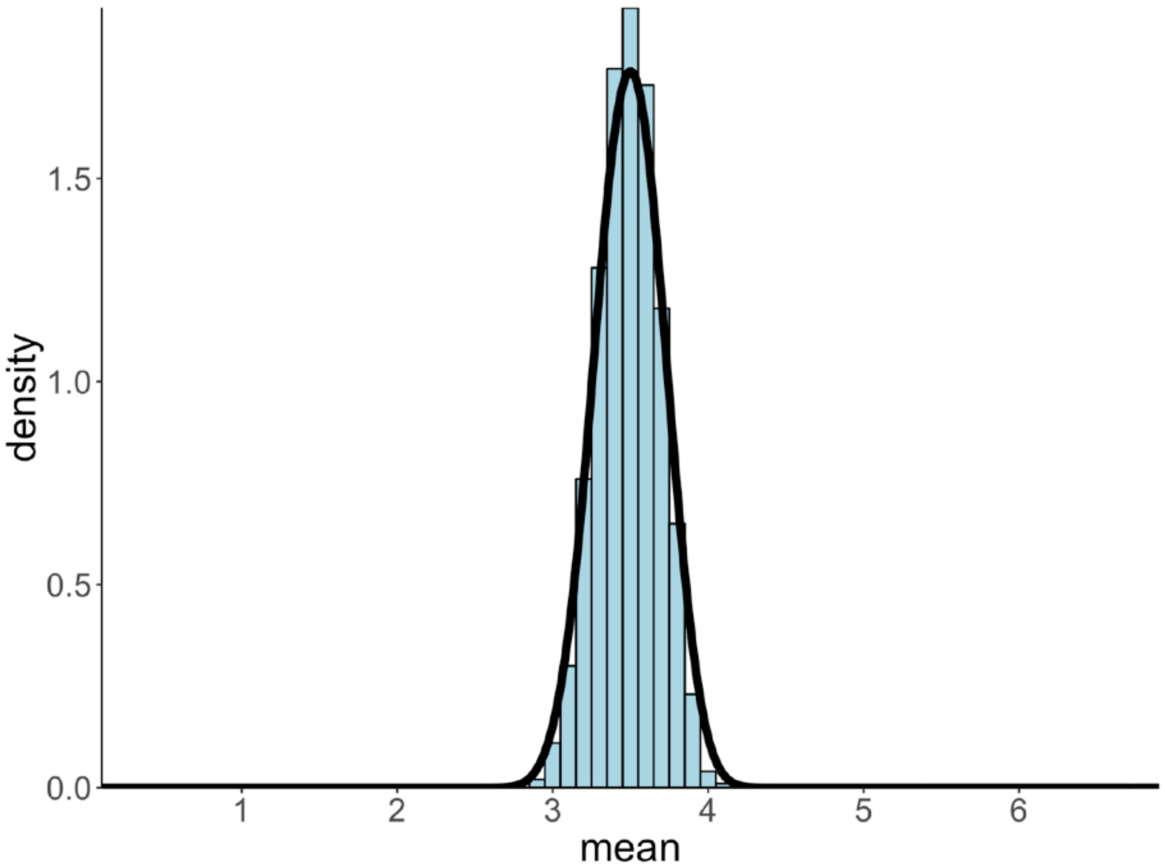
number of samples = 1000



...

sample	draw_1	draw_2	draw_3	draw_4	sample_mean
1	1	6	6	4	4.25
2	1	4	4	6	3.75
3	6	1	1	1	2.25
4	3	6	3	6	4.50
5	3	4	6	3	4.00
6	4	1	6	1	3.00
7	1	6	1	6	3.50
8	4	6	6	6	5.50
9	6	1	3	3	3.25
10	3	1	3	6	3.25

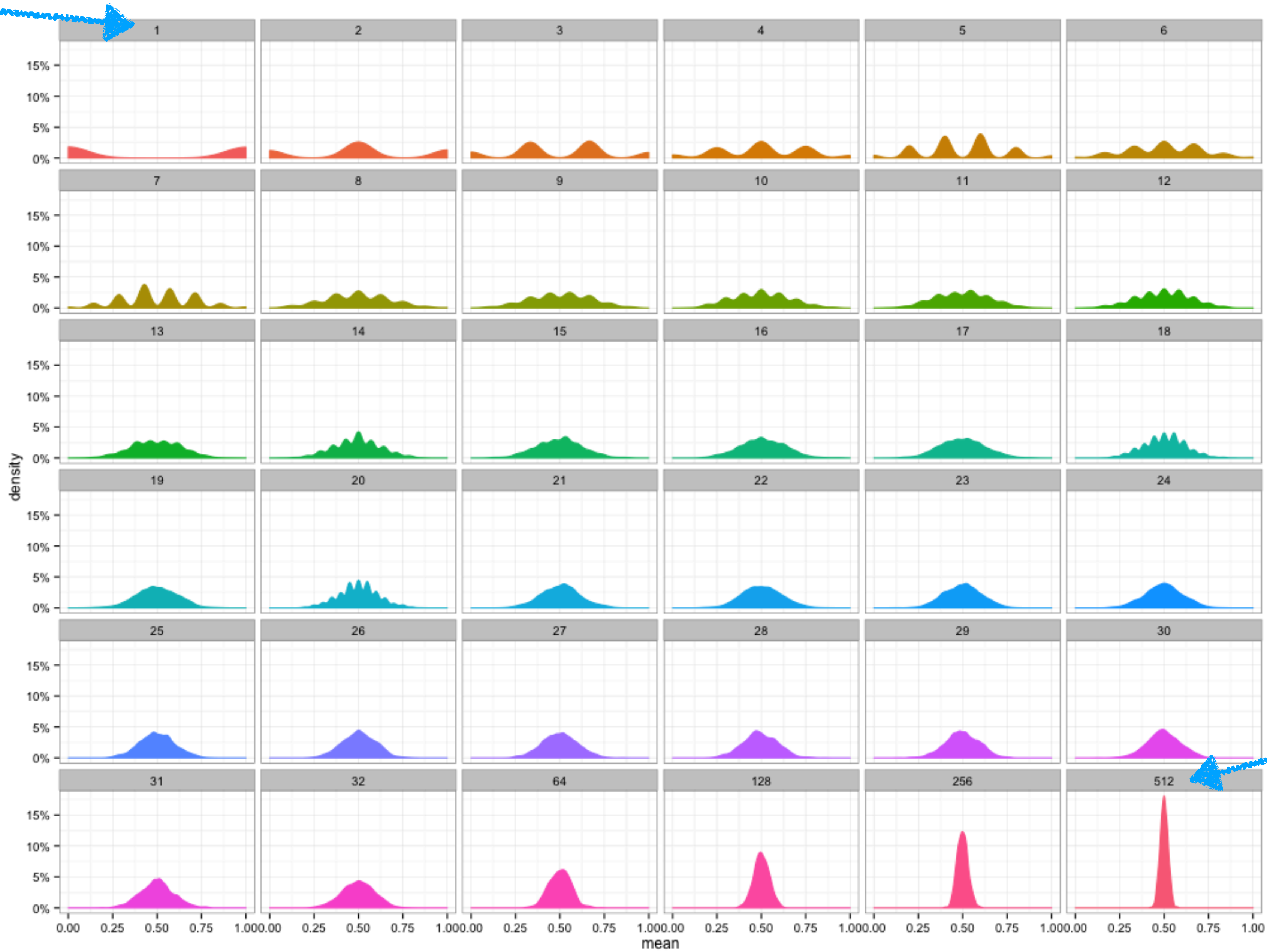
⋮



Central limit theorem

Binomial distribution: generate 0s and 1s and calculate their mean

sample size



more normal
& narrower

The larger the sample, the better the approximation.

Central limit theorem

Central limit theorem

seeing theory demo

Central limit theorem

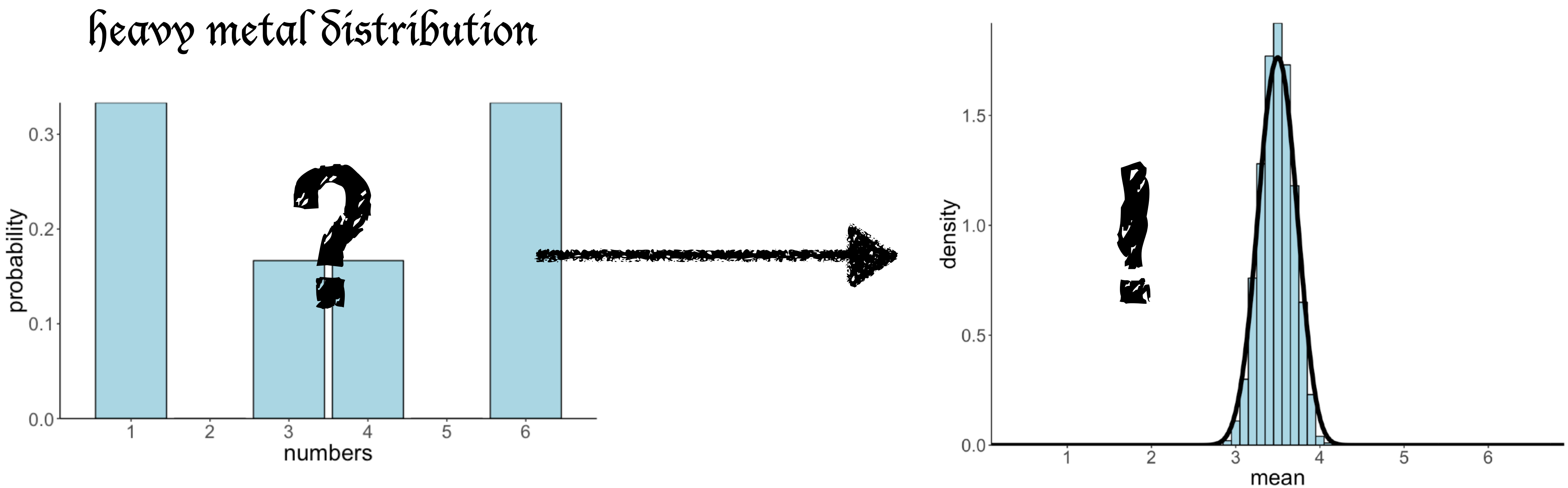
Why do we care?

- normal distributions **are everywhere**
  - outcomes that are affected by many factors that combine additively will tend to be normally distributed

Central limit theorem

Why do we care?

- even if we don't know the true underlying probability distribution (and we cannot accurately predict any individual observation), we can still make inferences about the mean of the distribution (even when that distribution is far from normal)



Central limit theorem

Where it works

- sufficiently large number of i.i.d. factors that combine additively

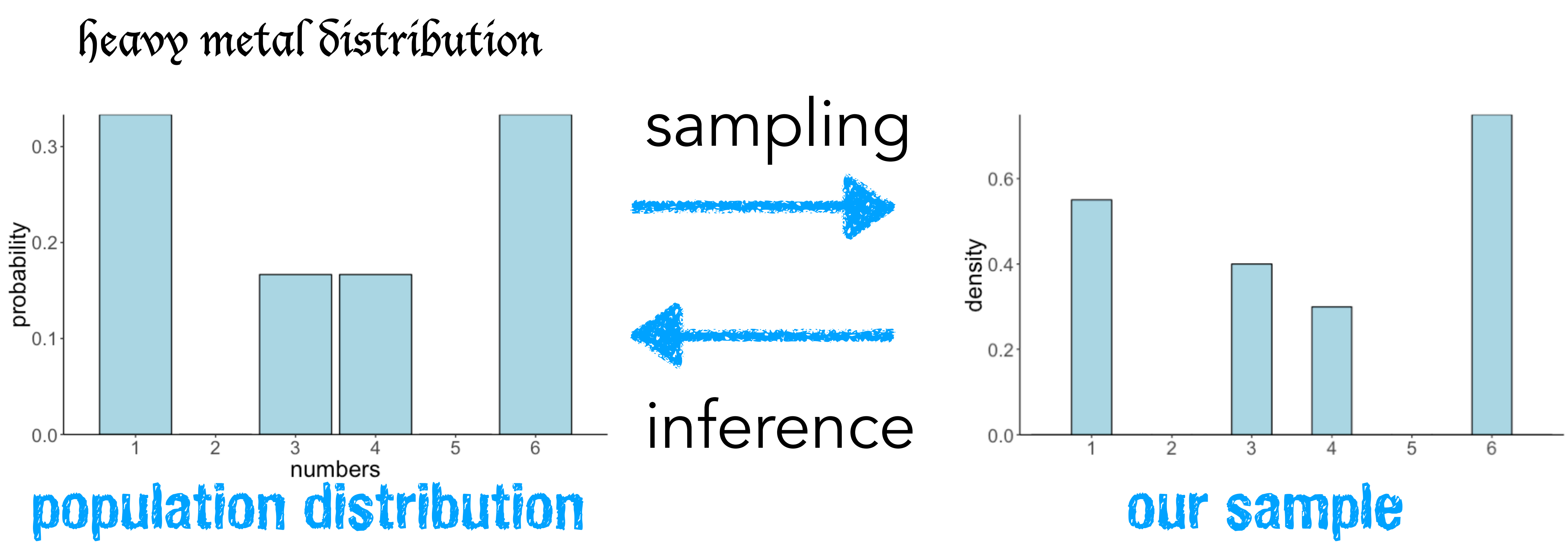
Where it breaks down

1. when one factor affects the outcome (much) more strongly than others
2. when processes involve strong dependence
   - e.g. rich get richer dynamics (distribution of wealth)
3. when factors combine multiplicatively
   - many diseases (e.g. how cancer cells divide and grow)
   - (such phenomena often follow a log-normal distribution)

Sampling distributions

Statistical inference

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



Statistical inference

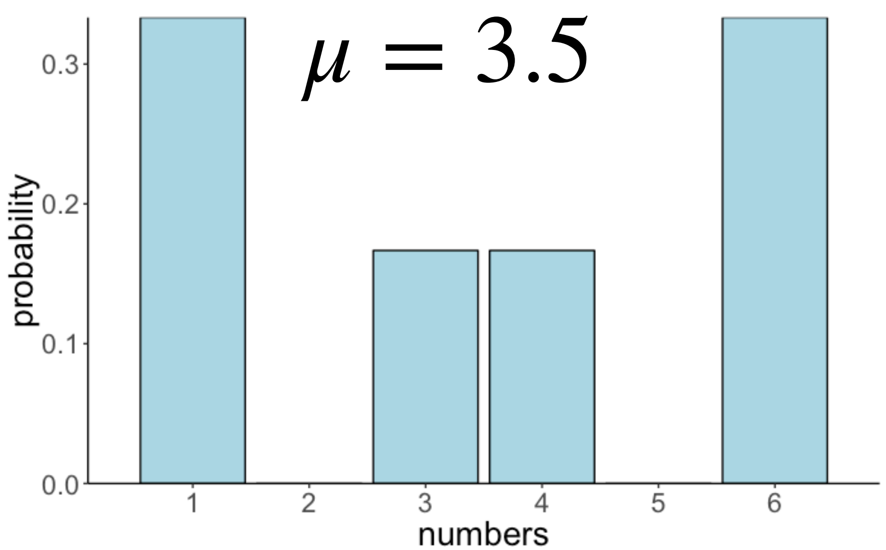
what's the
population mean?

sample mean = 3.725

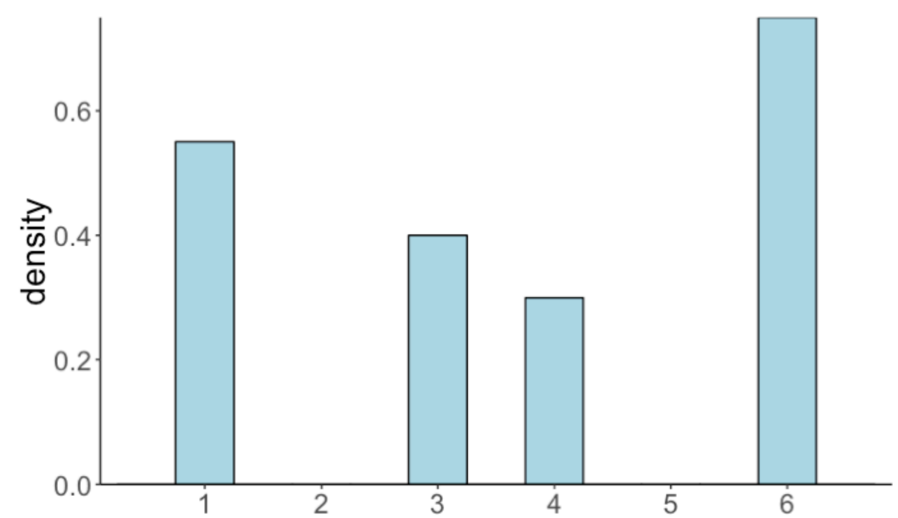
standard deviation = 2.05

n = 40

heavy metal distribution

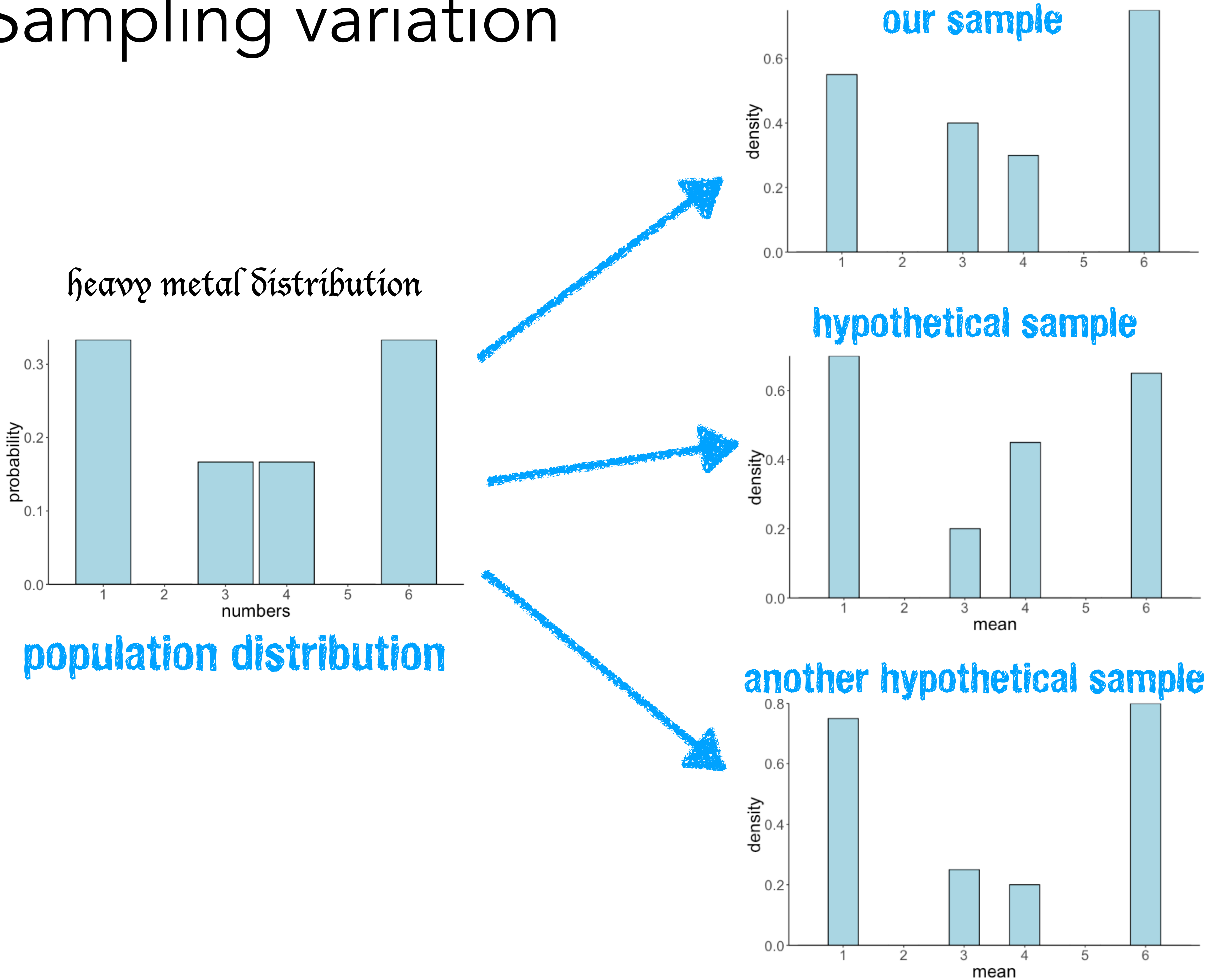


true unknown distribution



our sample

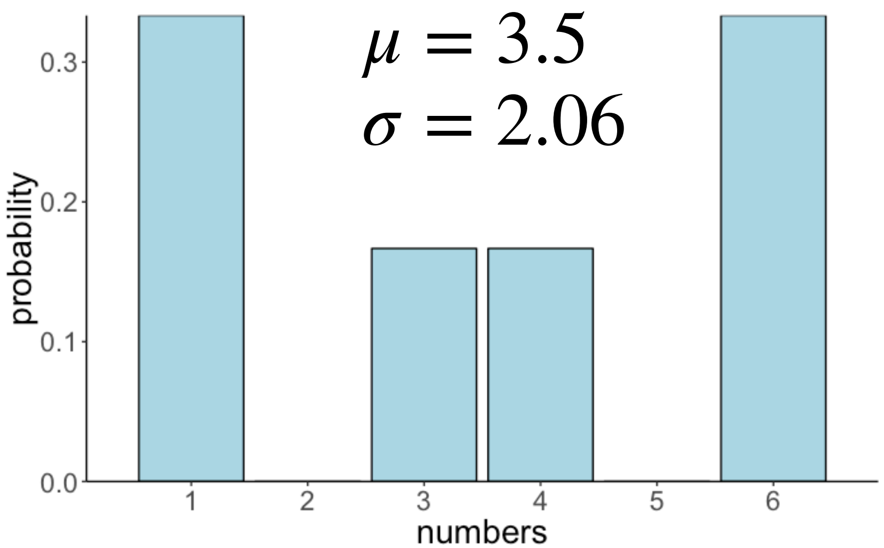
Sampling variation



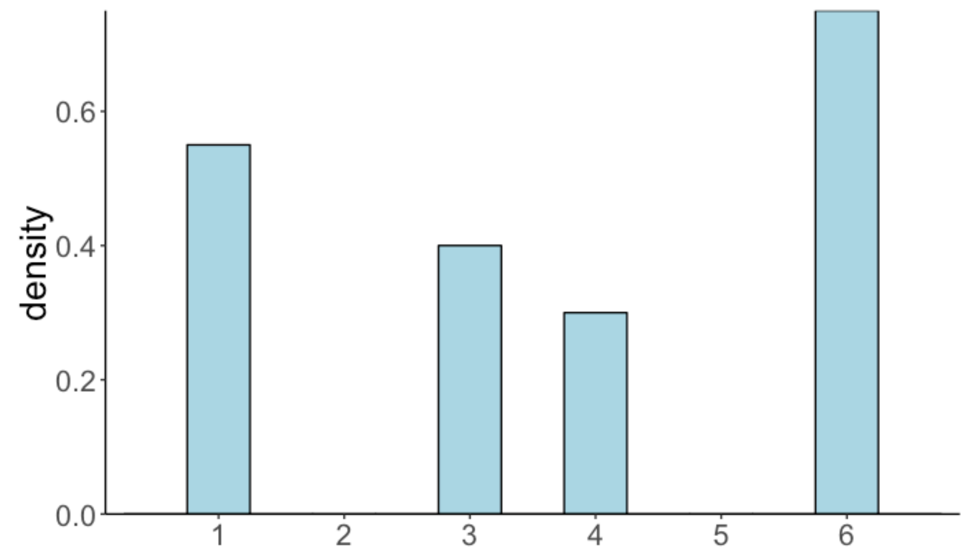
Sampling distribution

population distribution

heavy metal distribution

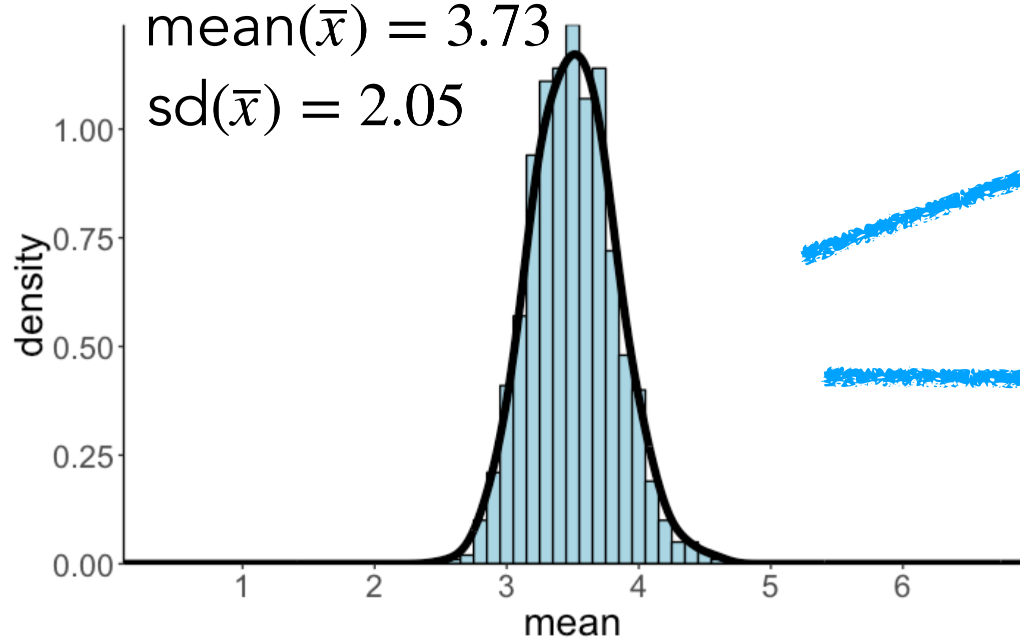


our sample



$\text{mean}(x) = 3.73$
$\text{sd}(x) = 2.05$
$\text{n} = 40$

sampling distribution



p-values

confidence intervals

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. Make a decision (either reject or don't reject H_0) based on pre-specified significance level α

The magic 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" (cf. Central Limit Theorem CLT).

instead: simulation-based approach

Central limit theorem

sample size = 100

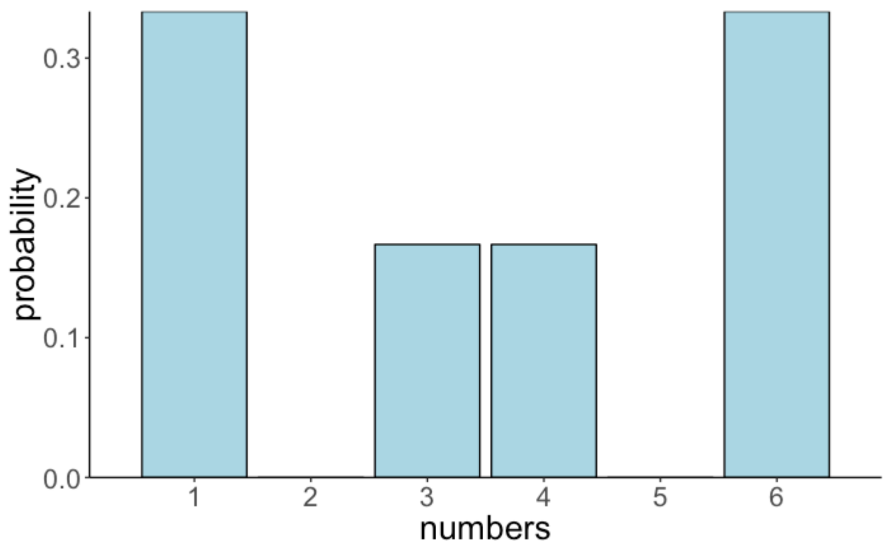
number of samples = 1000

...

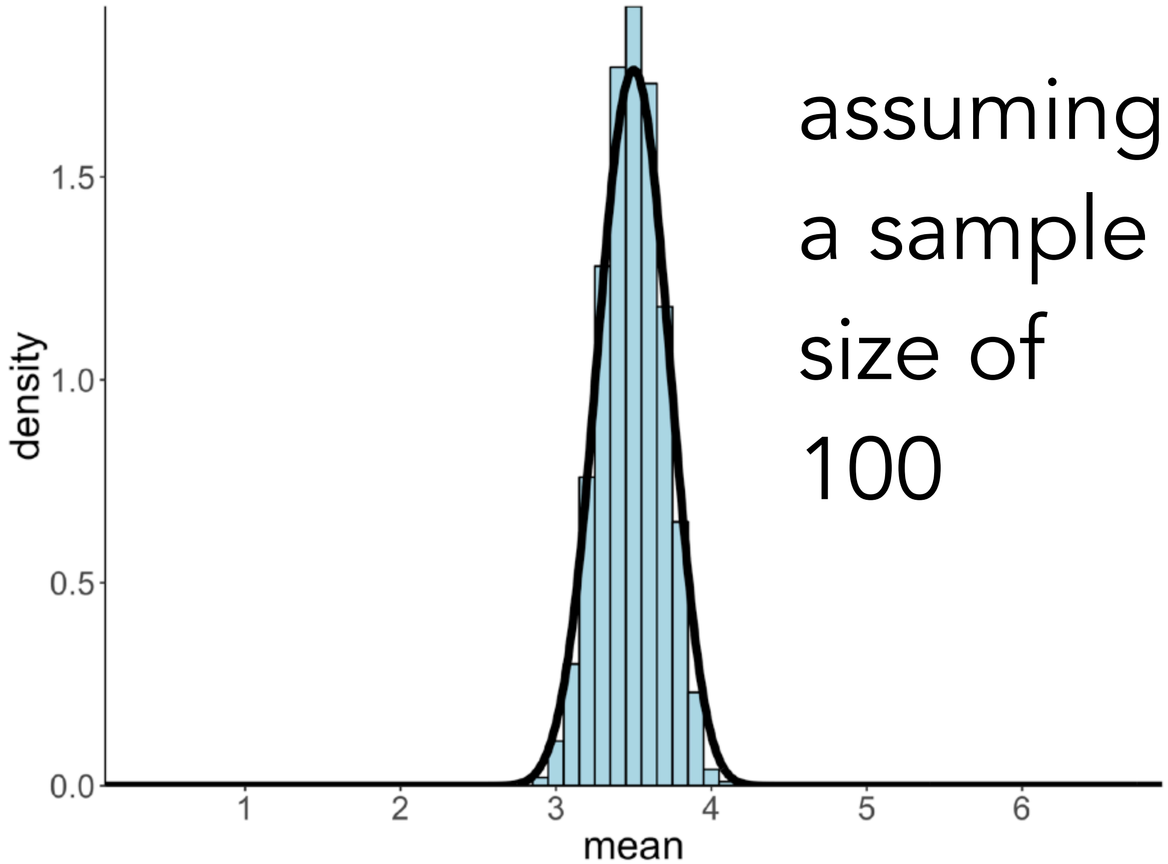
sample	draw_1	draw_2	draw_3	draw_4	sample_mean
1	1	6	6	4	4.25
2	1	4	4	6	3.75
3	6	1	1	1	2.25
4	3	6	3	6	4.50
5	3	4	6	3	4.00
6	4	1	6	1	3.00
7	1	6	1	6	3.50
8	4	6	6	6	5.50
9	6	1	3	3	3.25
10	3	1	3	6	3.25

⋮

heavy metal distribution



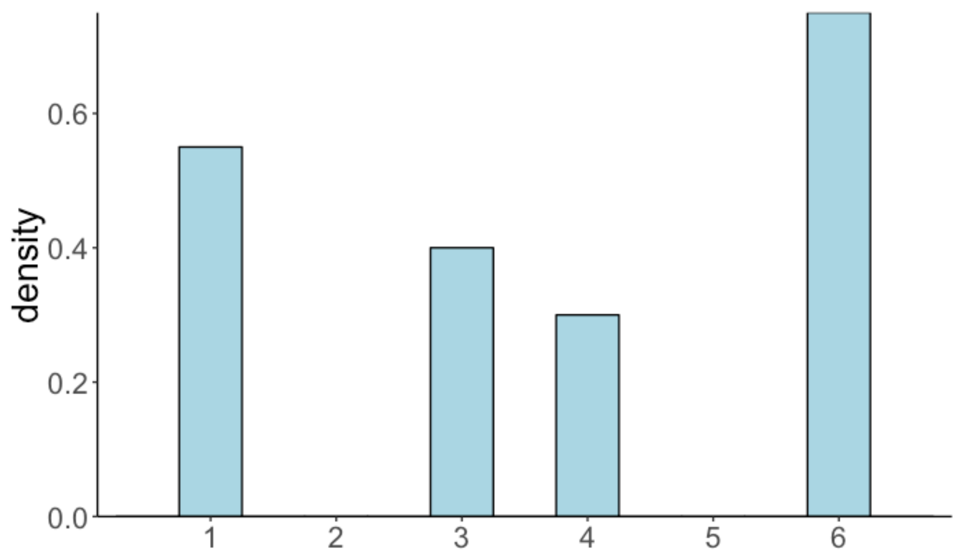
population distribution



assuming a sample size of 100

sampling distribution

our sample

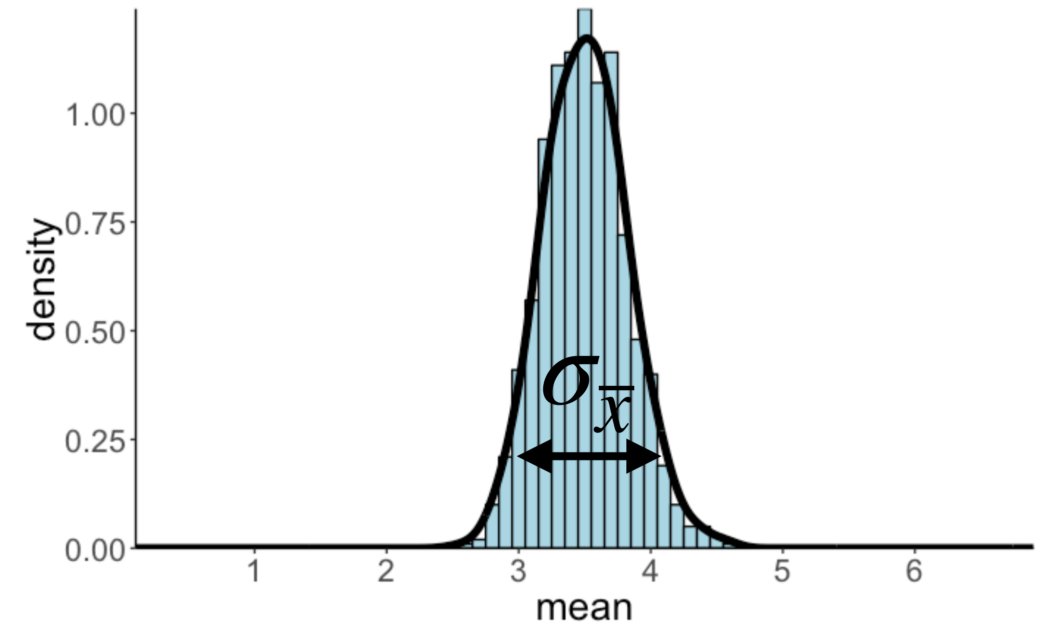


standard deviation

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

gives a sense for how well the mean summarizes the data

sampling distribution



standard error

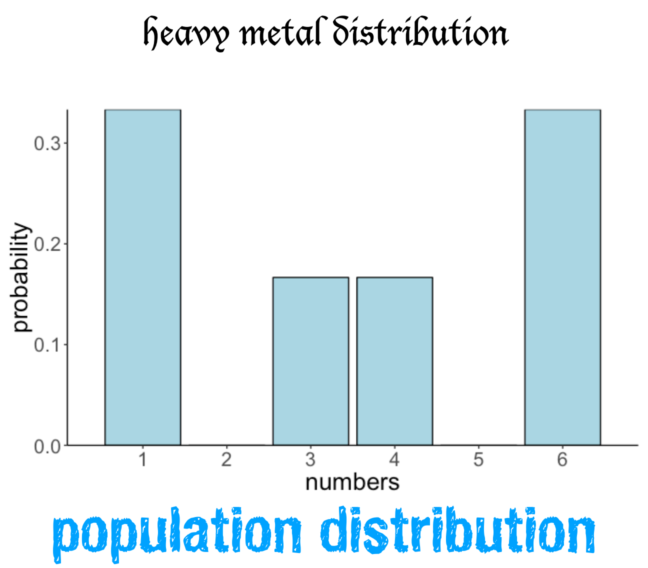
$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$$

the standard deviation of the sampling distribution

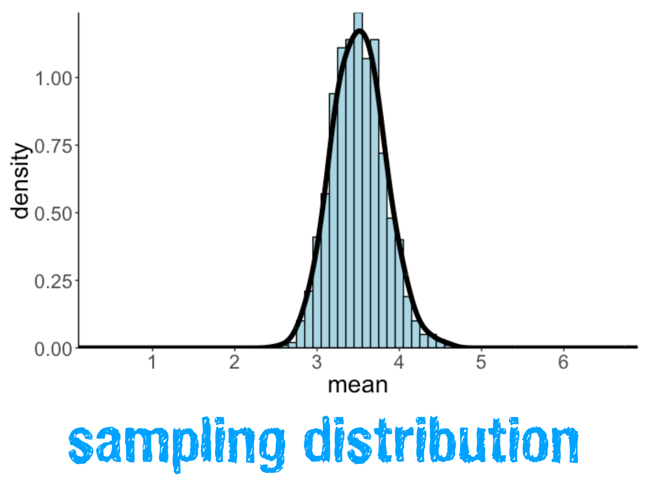
how much variation do we expect between the means of different samples

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

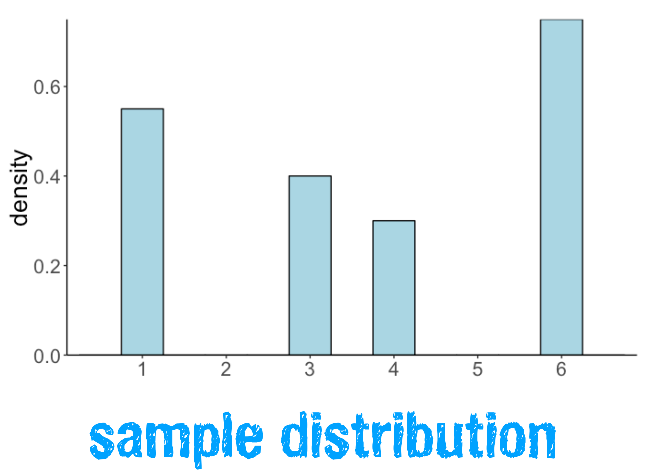
3 distributions in statistical inference



- unknown
- our target for inference



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



- our observed sample
- we compute statistics of interest (mean, variance, correlation, ...)

Take a look at the course notes

```
### The sampling distribution

And let's now create the sampling distribution (making the unrealistic assumption that we know the
population distribution).

```{r clt6}
# make example reproducible
set.seed(1)

# parameters
sample_size = 40 # size of each sample
sample_n = 1000 # number of samples

# define a function that draws samples from a discrete distribution
fun.draw_sample = function(sample_size, distribution){
  x = sample(distribution$numbers,
             size = sample_size,
             replace = T,
             prob = distribution$probability)
  return(x)
}

# generate many samples
samples = replicate(n = sample_n,
                    fun.draw_sample(sample_size, df.population))

# set up a data frame with samples
df.sampling_distribution = matrix(samples, ncol = sample_n) %>%
  as_tibble(.name_repair = ~ str_c(1:sample_n)) %>%
  pivot_longer(cols = everything(),
               names_to = "sample",
               values_to = "number") %>%
  mutate(sample = as.numeric(sample)) %>%
  group_by(sample) %>%
  mutate(draw = 1:n()) %>%
  select(sample, draw, number) %>%
  ungroup()

# turn the data frame into long format and calculate the means of each sample
df.sampling_distribution_means = df.sampling_distribution %>%
  group_by(sample) %>%
  summarize(mean = mean(number)) %>%
  ungroup()
```
```

p-values

Statistical inference

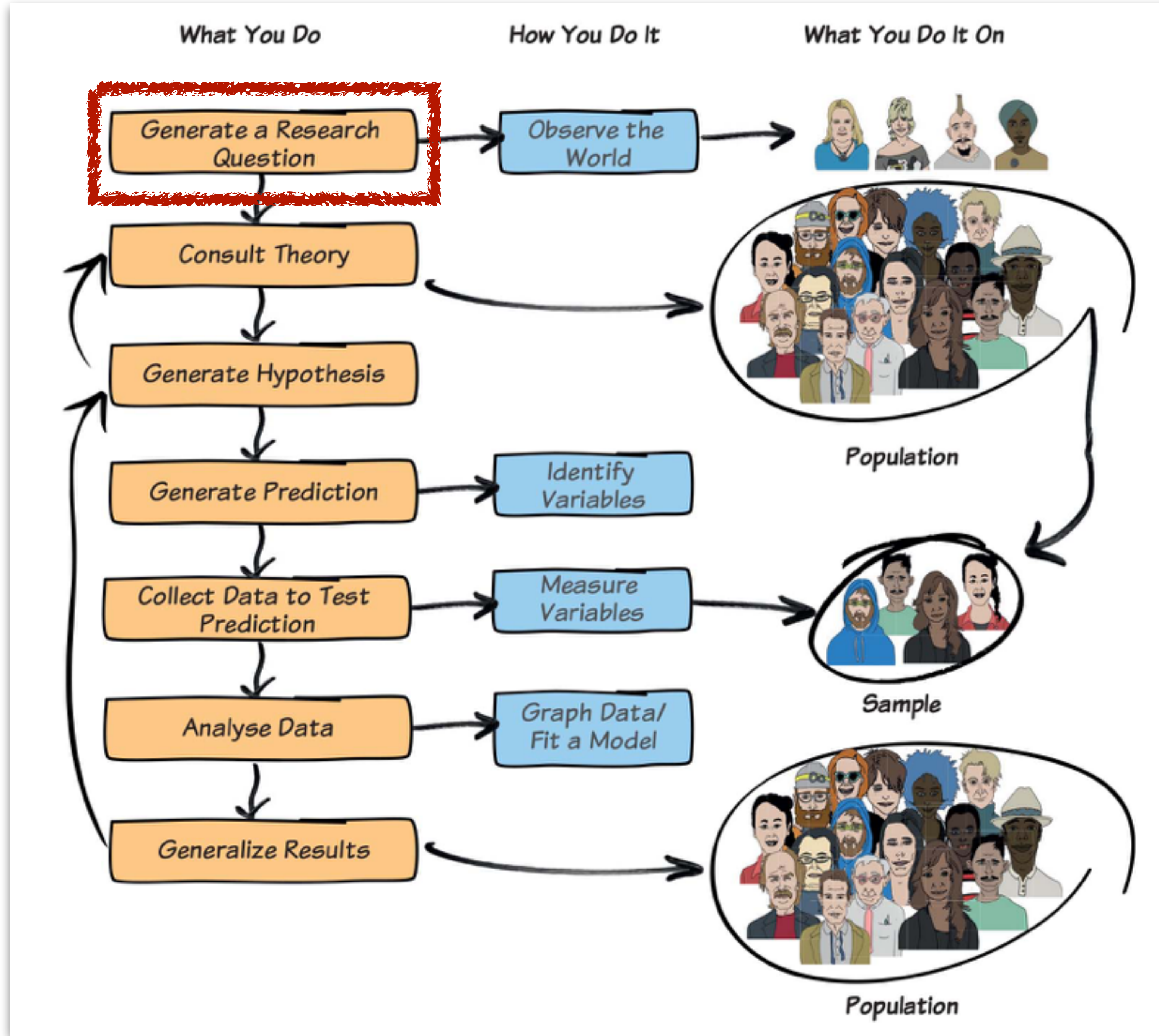
null hypothesis

$$H_0 : \mu_1 = \mu_2$$

alternative hypothesis

$$H_1 : \mu_1 < \mu_2$$

Greek letters
population parameters!



What is a p-value?

- The p-value is the probability of the data given the null hypothesis.
- Given the null hypothesis is true, the probability to observe the data we have collected
- a number between 0 and 1 that indicates evidence in relation to a hypothesis.
- Probability of observing the results obtained or more significant if the null hypothesis is true
- the probability that, given that the null hypothesis is true, you would observe results that are favorable to your H1 hypothesis
- It measures how probable a particular event is to occur not due to random chance
- The likelihood that the hypothesis and null hypothesis are not different.
- a p-value is the probability that the results in your sample reflect a real difference in the population.

⋮ we could go on and on

What is a p-value?

so let's go on and on

- The p value is the probability of observing the data in your sample assuming that the **bill hypothesis** was true.

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

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

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

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

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

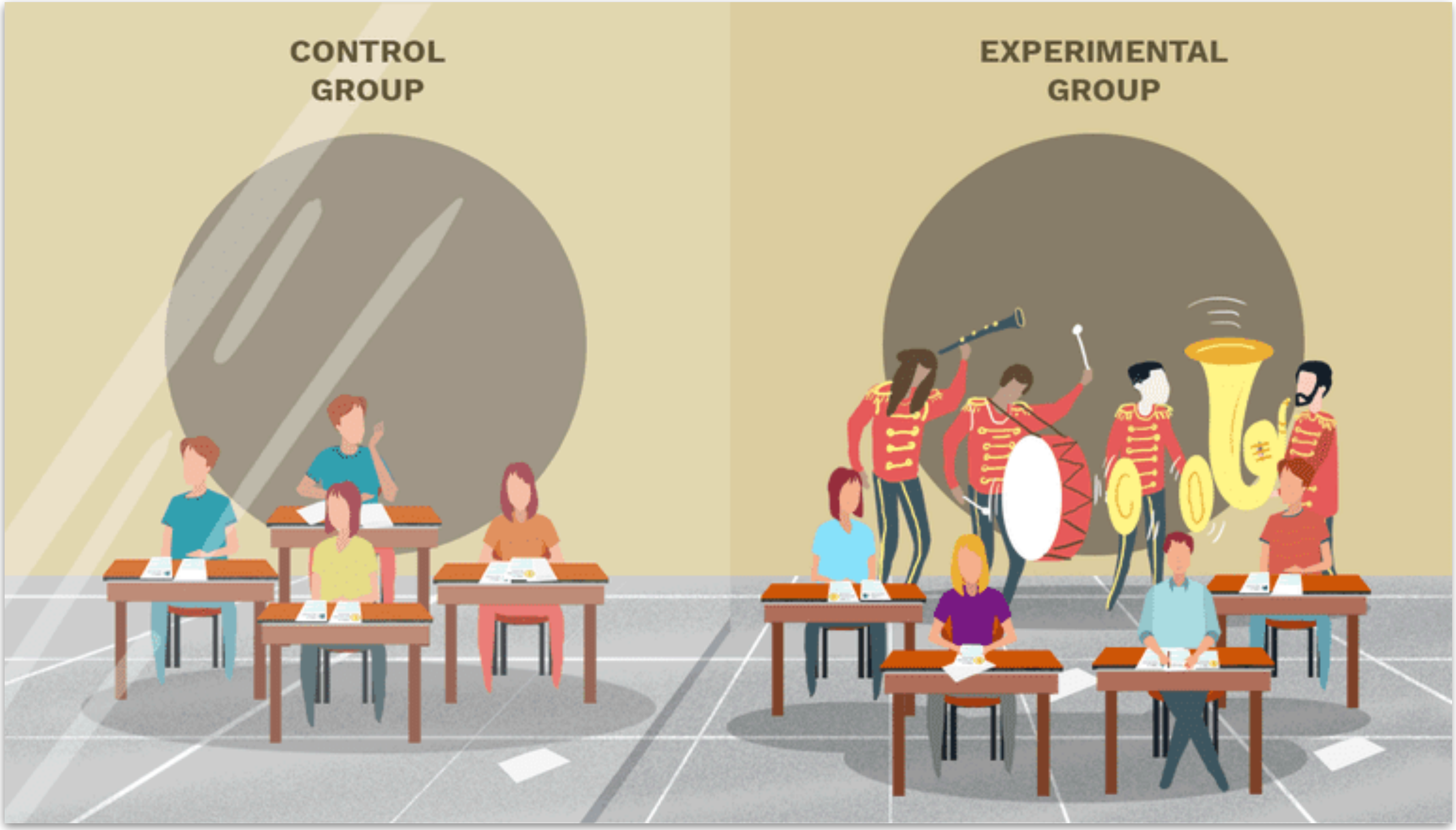
$H = \text{Hypothesis}$

$D = \text{Data}$

Logic of inference

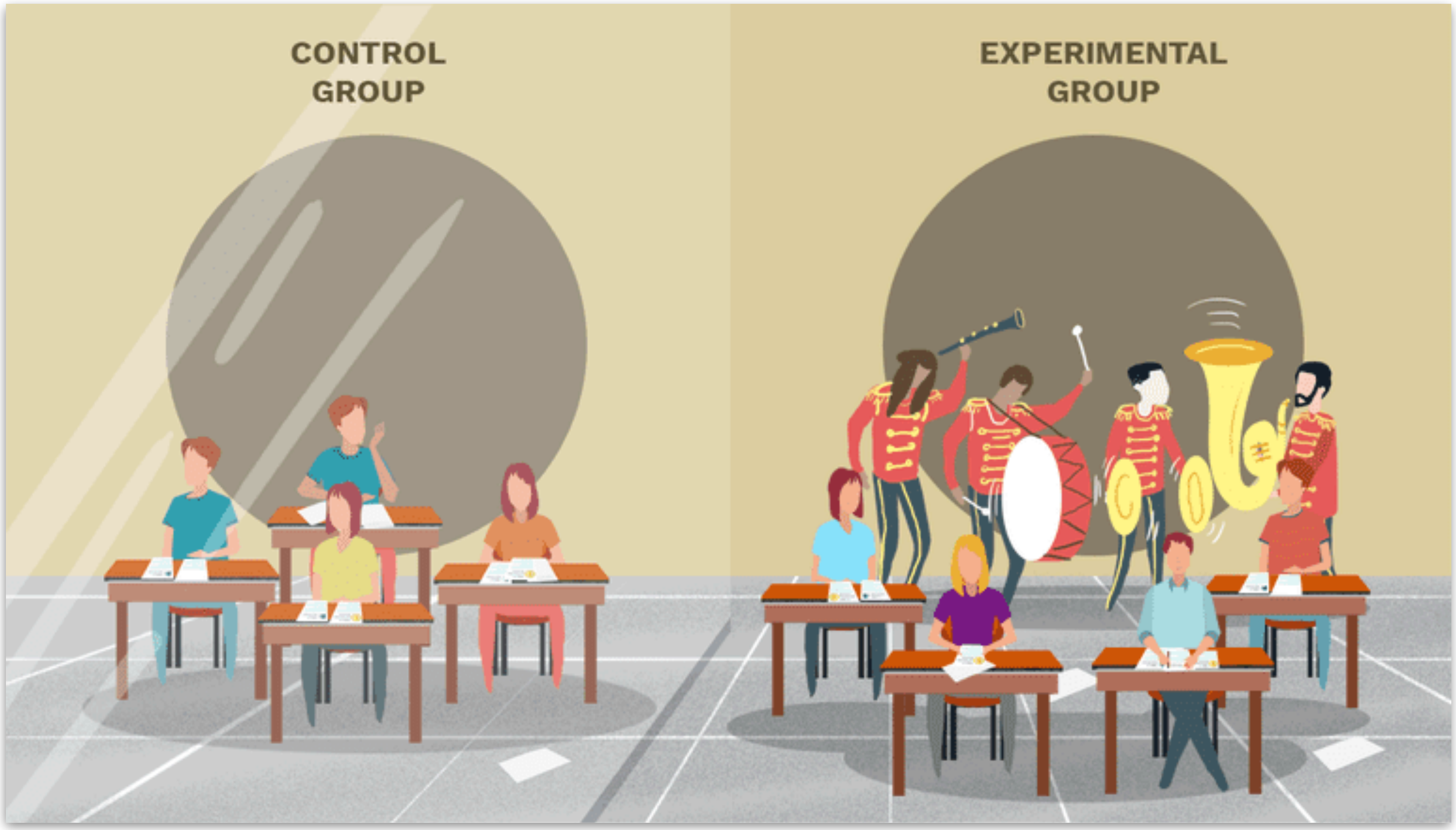
- calculate a **test statistic** based on the sample
  - for example, the difference between the means of two conditions
- build a **sampling distribution** of this statistic assuming that the null hypothesis is true
  - use math or resampling methods
- **calculate the probability** of the observed statistic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



Research question:
Will student test scores be affected by distracting sounds (aka the Stanford band)?

Permutation test



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

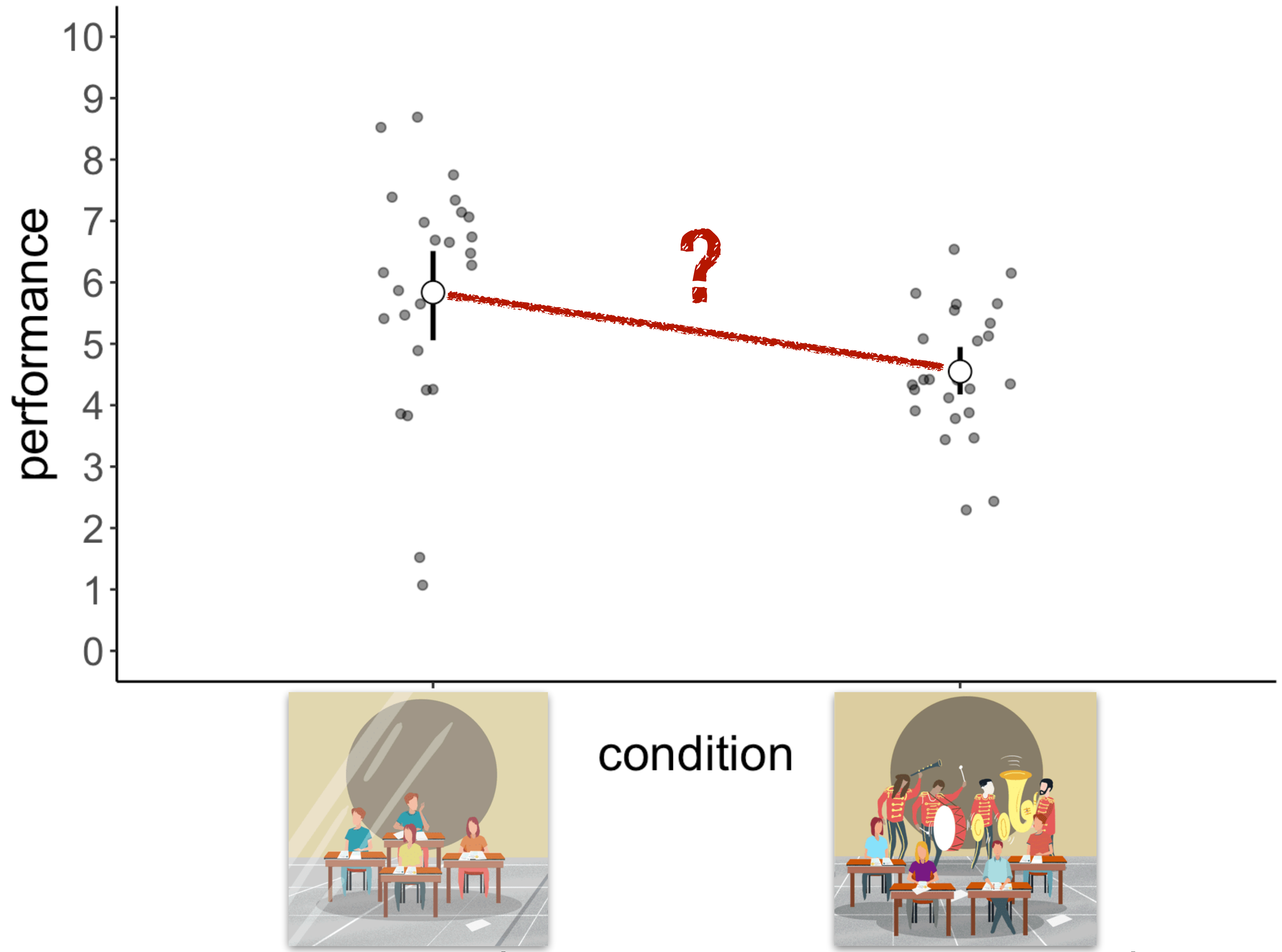
There is no difference between the control group and the experimental group

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

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

Permutation test

Is the difference in performance statistically significant?

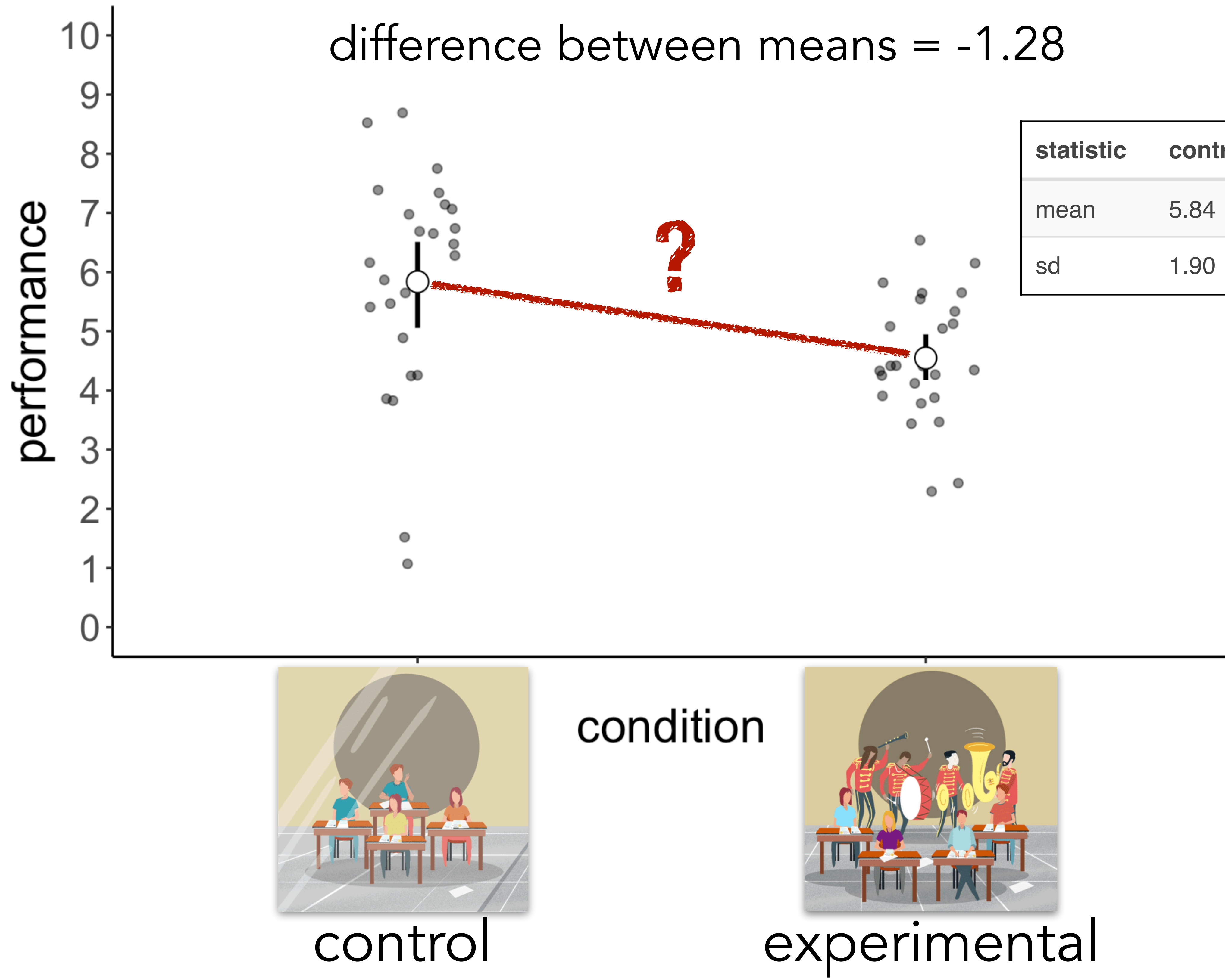


control

experimental

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 the 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)

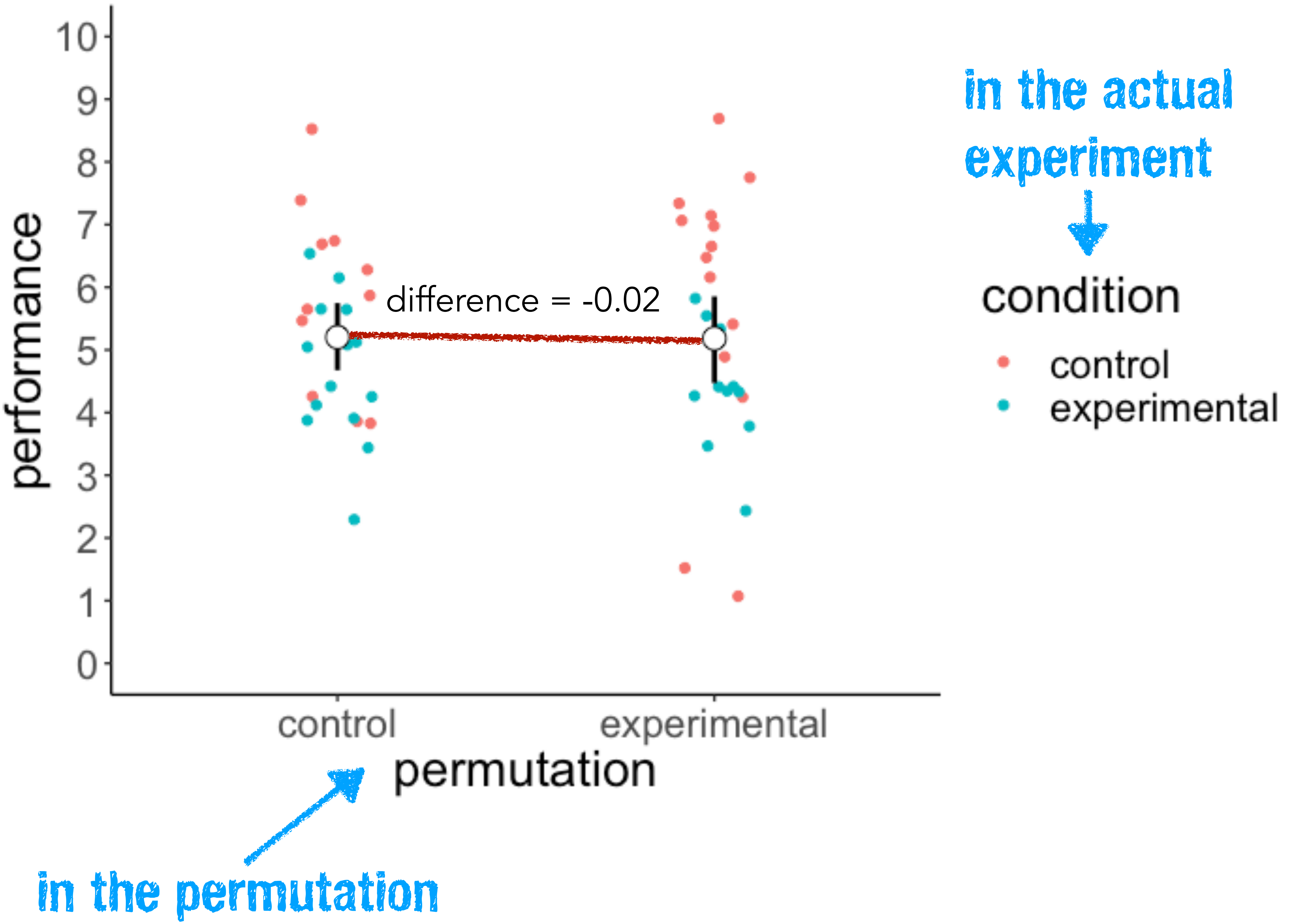
Permutation test

observed data **random permutation**

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

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

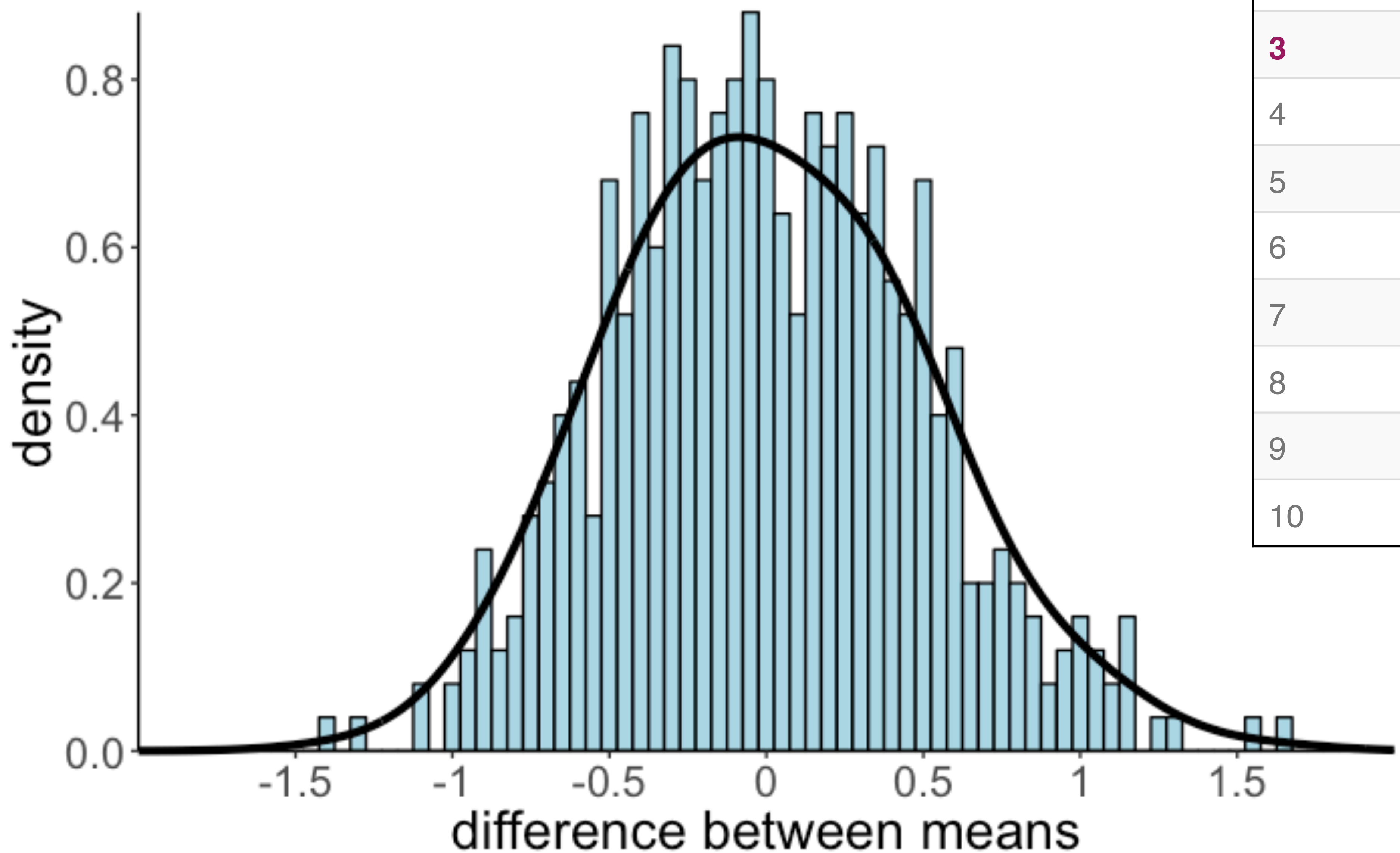
3

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

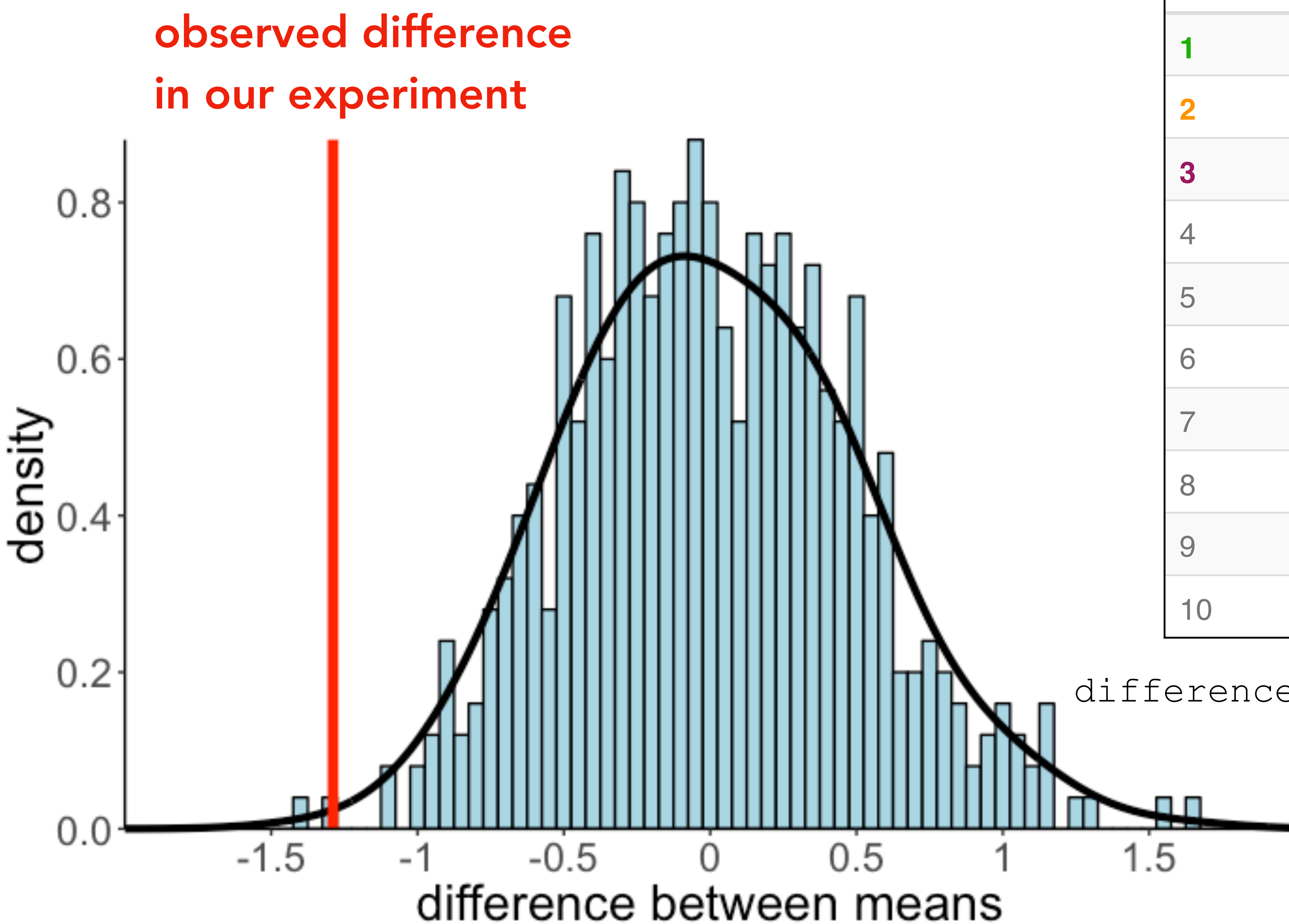


df.permutations

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 is true)

Permutation test



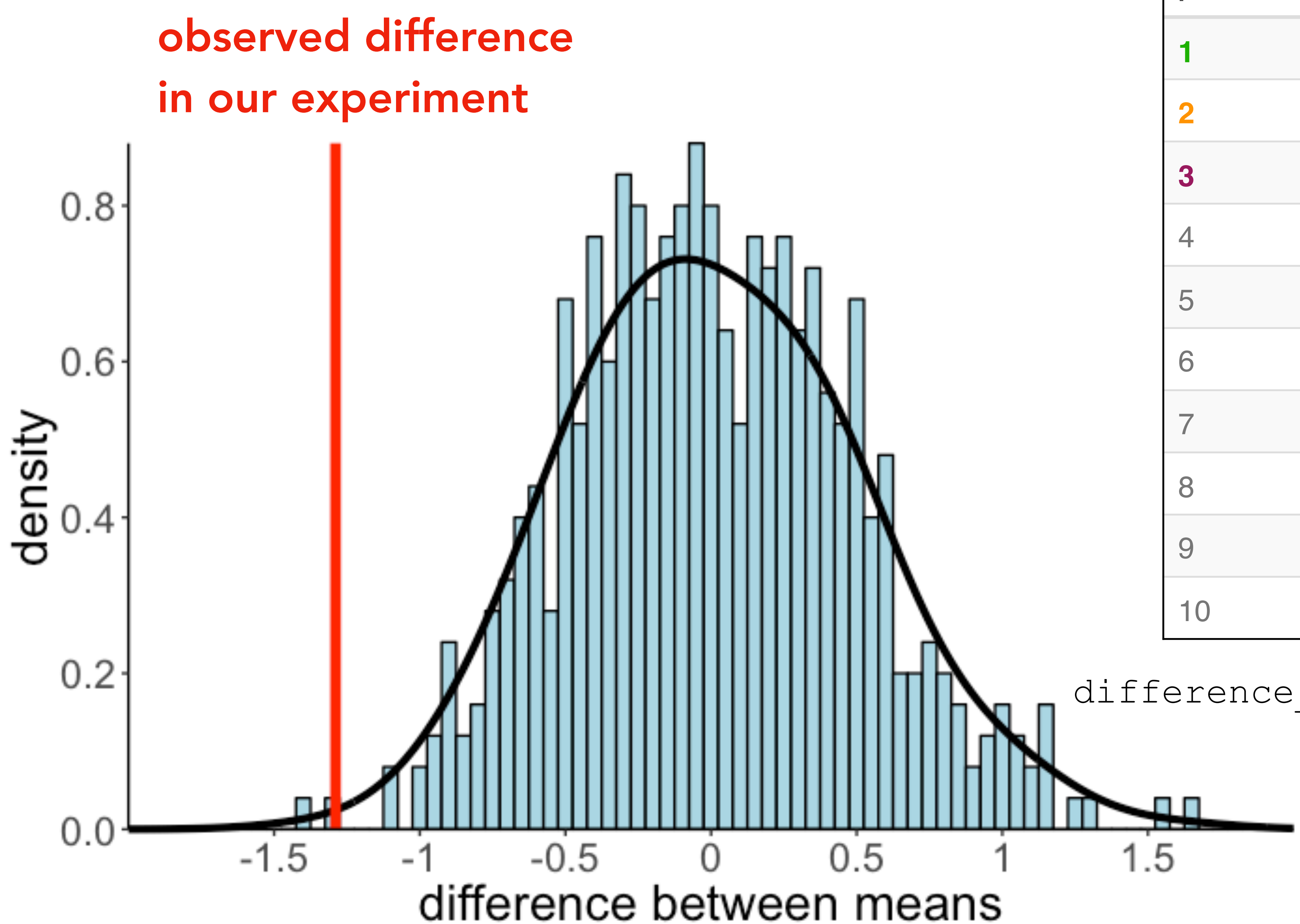
`df.permutations`

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`

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

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	2.29
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	2.29
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

```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))
)
```

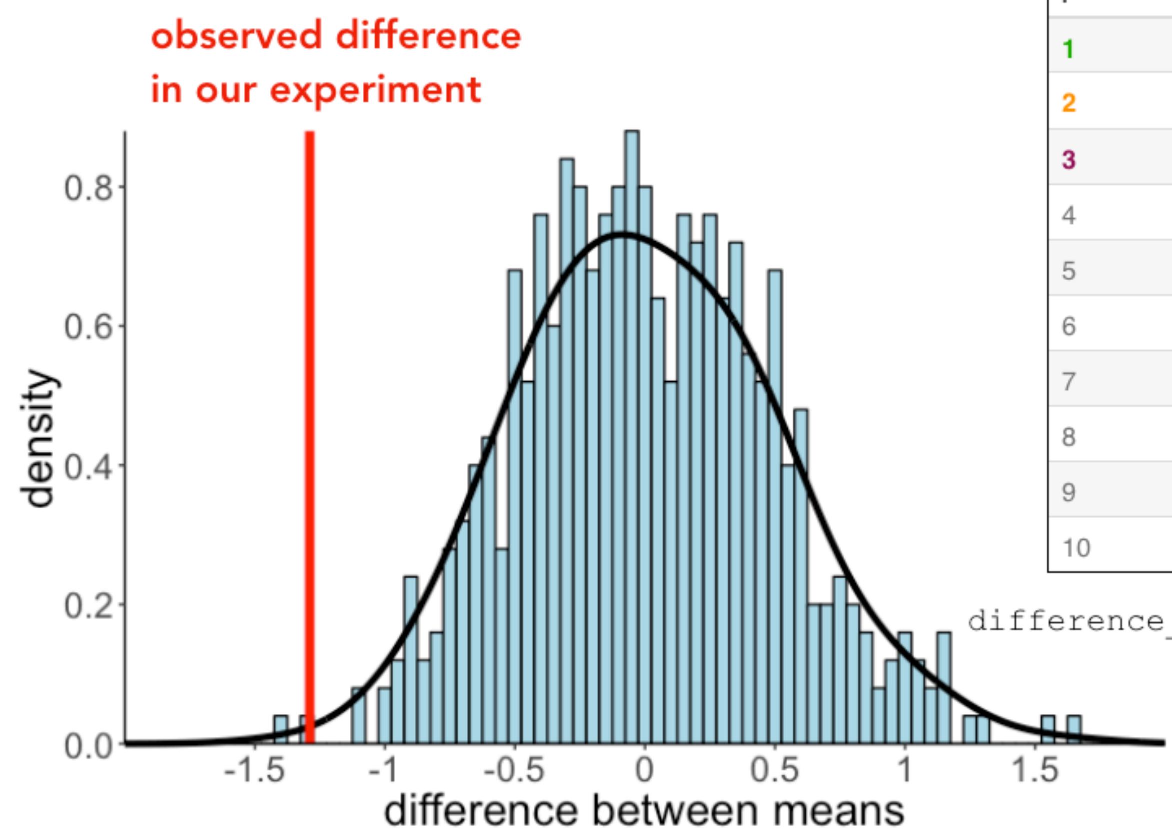
run the `func_permutations()` function many times

(instead of using a for loop)

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`

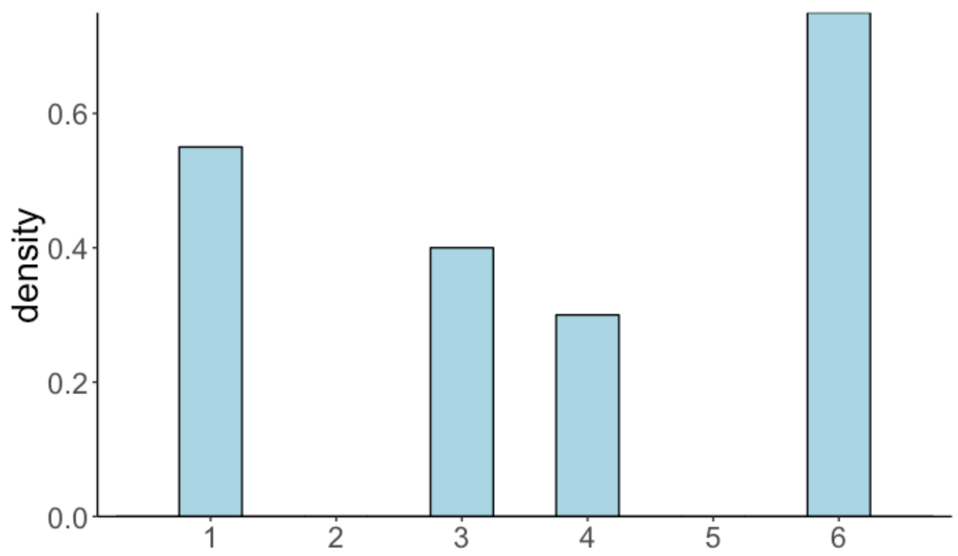
```
#calculate p-value of our observed result
df.permutations %>%
  summarize(p_value = sum(mean_difference <= difference_actual)/n())
```

p-value = .002

57

Confidence intervals

our sample

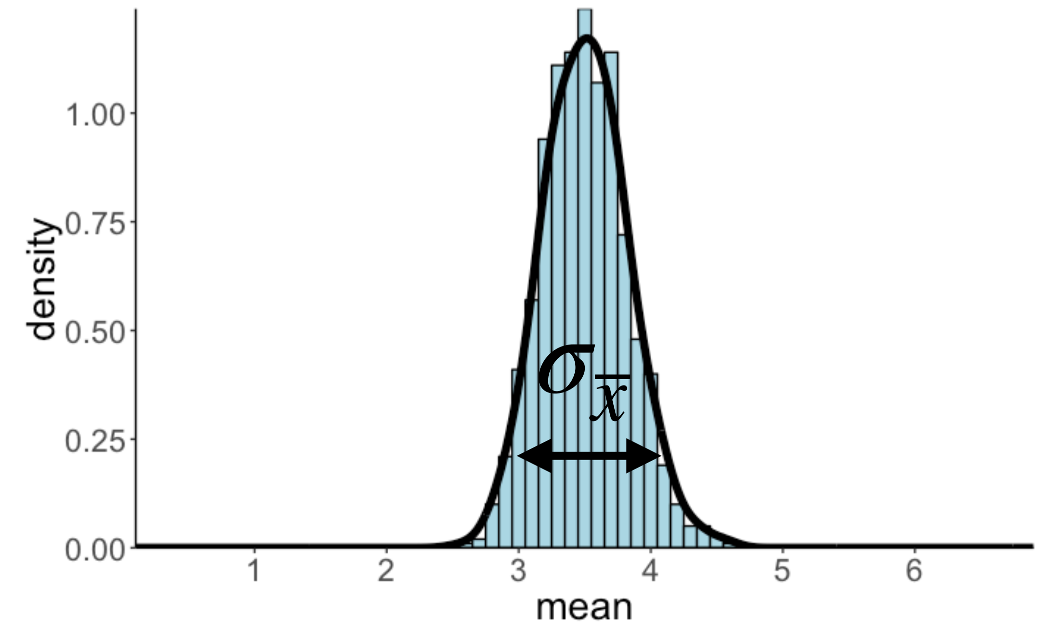


standard deviation

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

gives a sense for how well the mean captures the data

sampling distribution



standard error

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

$\sigma_{\overline{x}}$ decreases as:

σ decreases

N increases

$\hat{\sigma} = s$

for large enough samples (> 30)

we are more confident in our inference with larger samples, and less variance

Confidence interval

Goal: Estimate the mean of the population distribution μ

what we need:

- sample mean
- standard deviation
- sample size
- desired level of confidence (often 95%)

Confidence interval

Goal: Estimate the mean of the population distribution μ

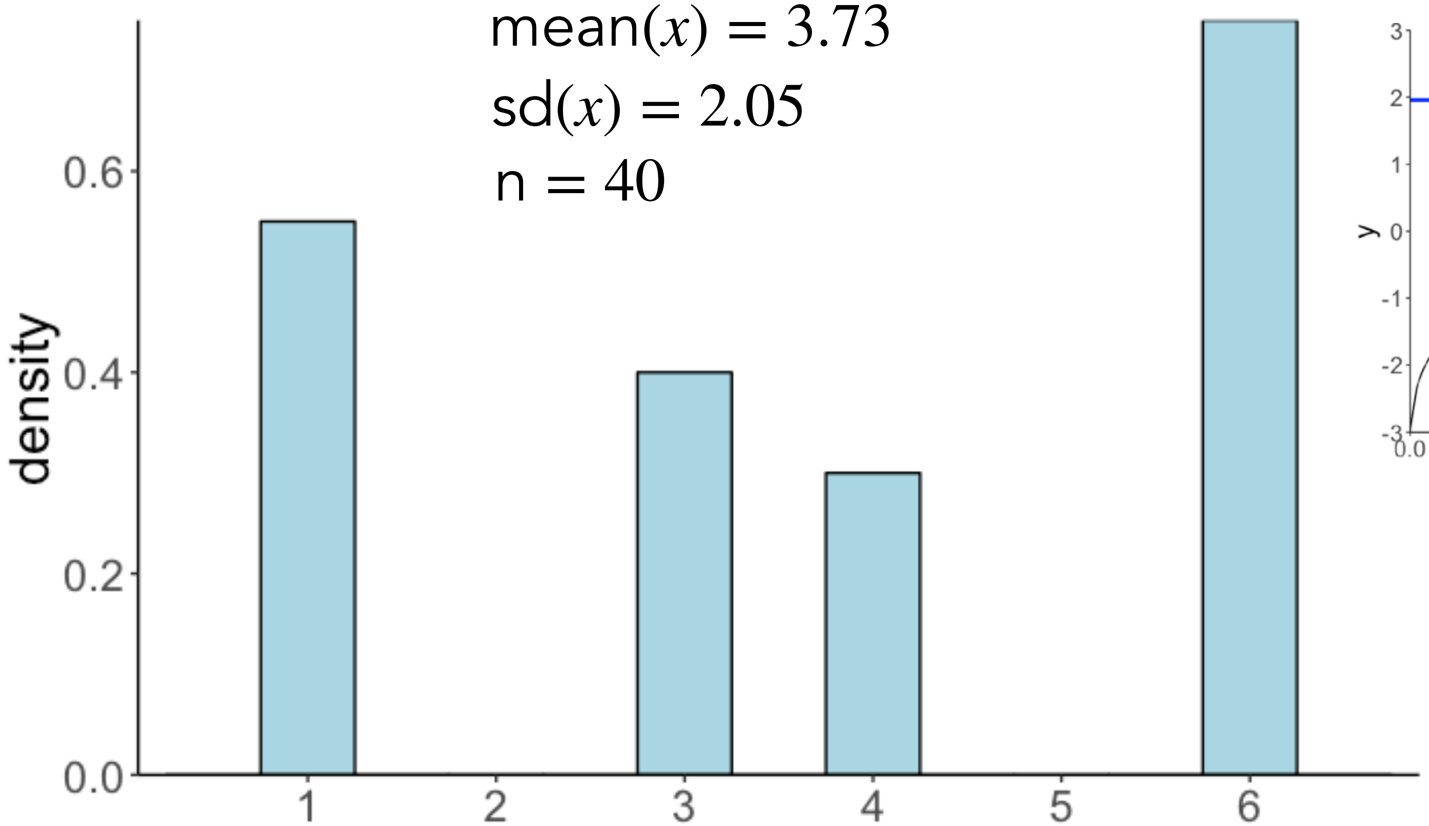
Confidence interval = point estimate $\pm$ critical value

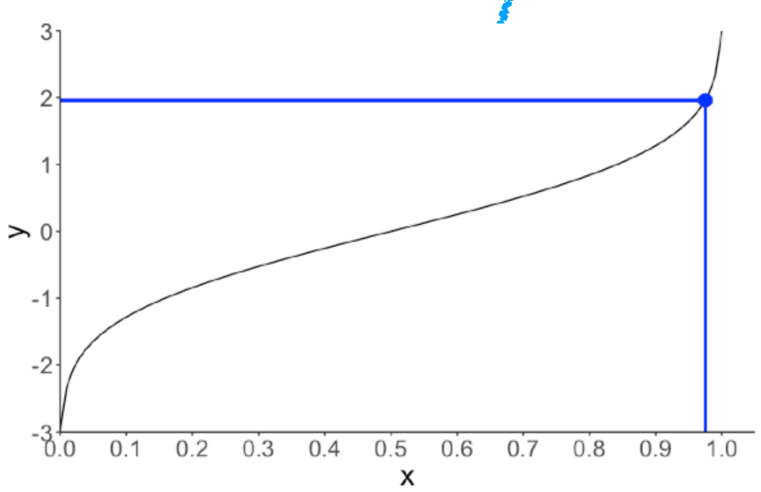
`mean(x)`

`qnorm(.975)×SE`

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

mean(x) = 3.73
sd(x) = 2.05
n = 40





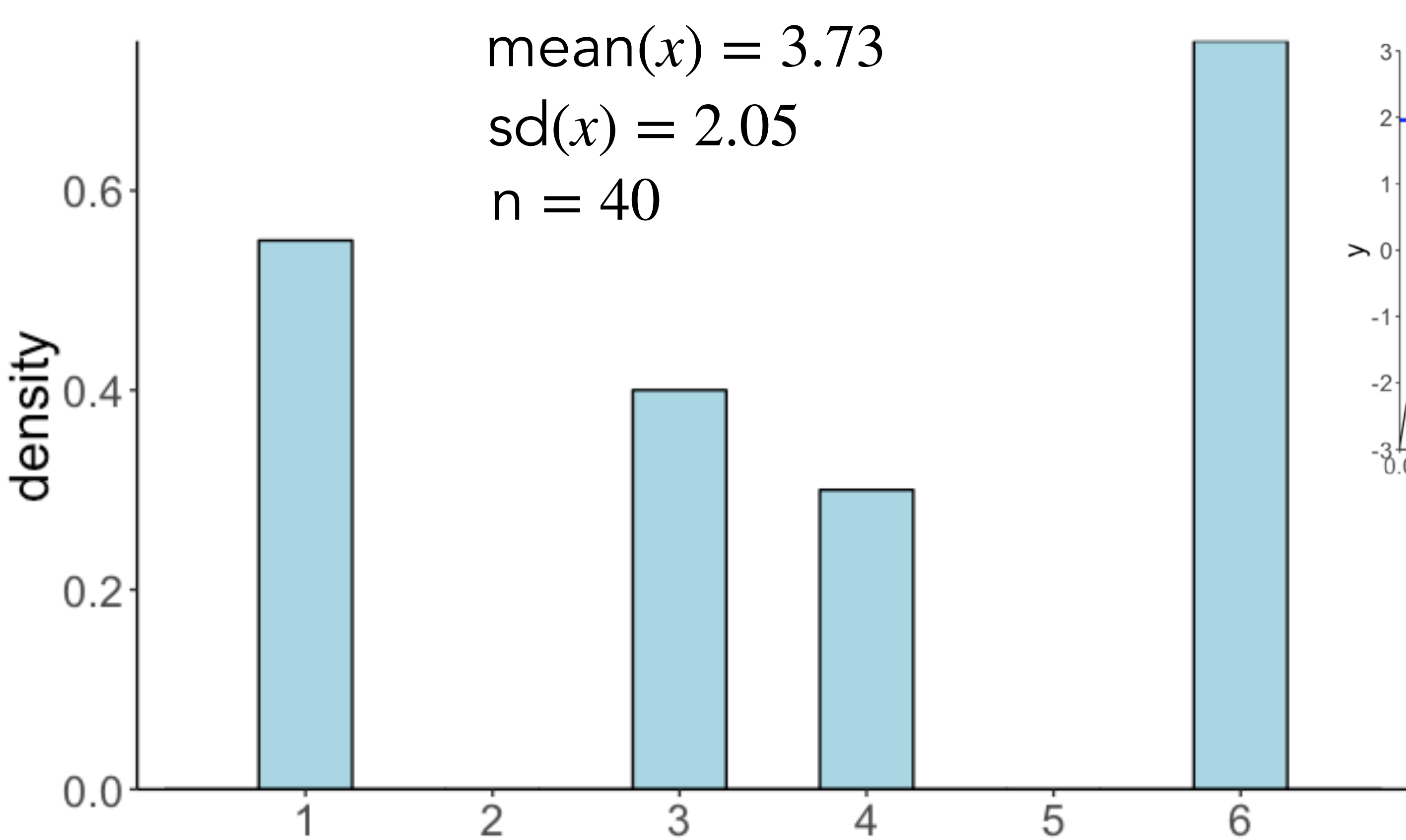
Confidence interval

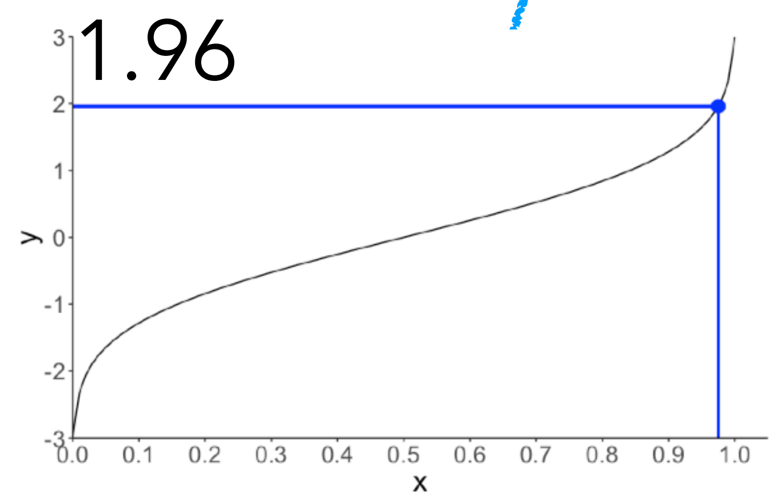
Goal: Estimate the mean of the population distribution μ

Confidence interval $= 3.73 \pm 0.63$

Confidence interval = point estimate ± critical value

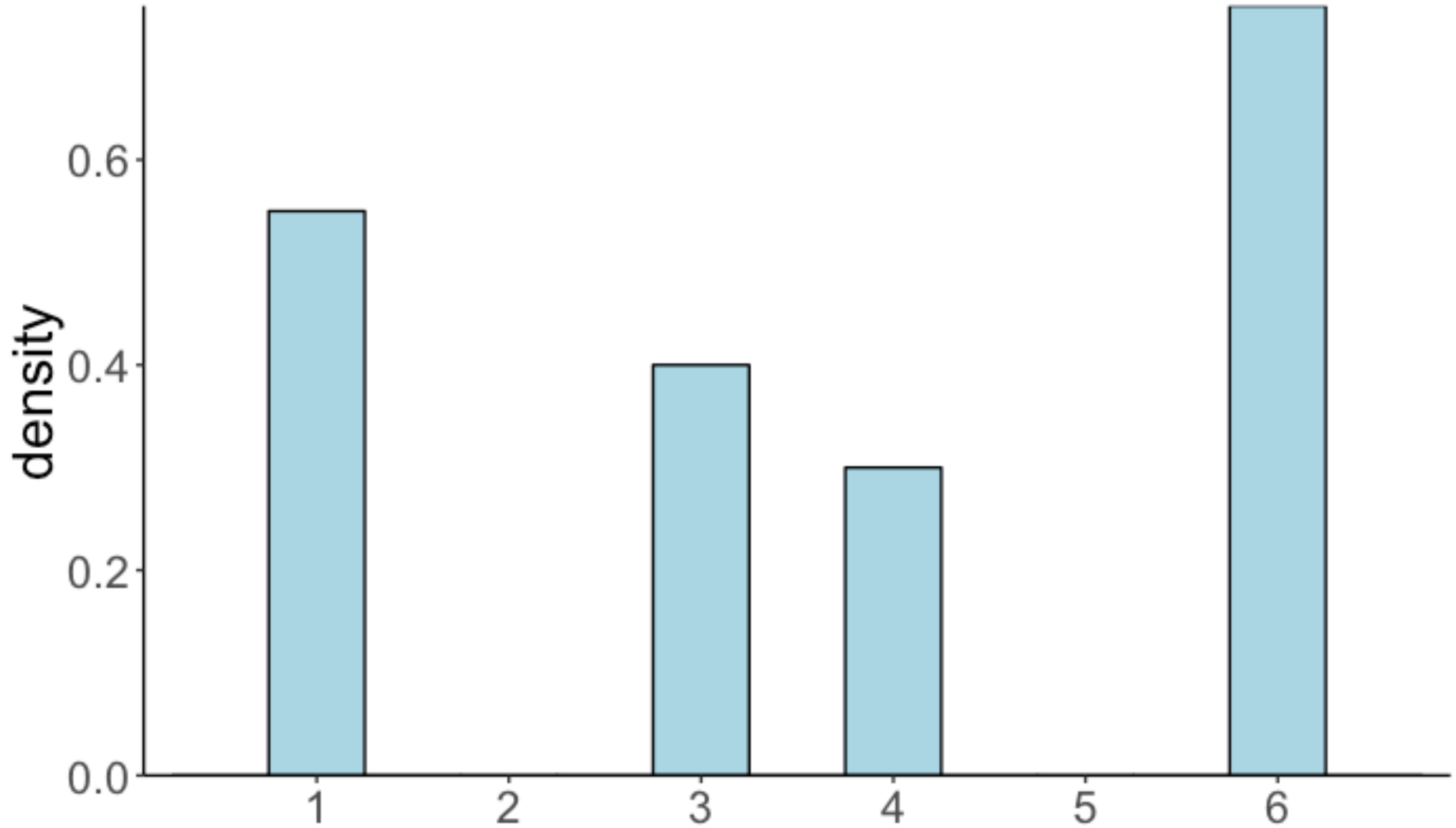
`mean(x)` `qnorm(.975)×SE`





$$\sigma_{\bar{x}} = \frac{2.05}{\sqrt{40}}$$

What does the confidence interval mean?



Mean = 3.73 ± 0.63 (95% CI)

What can we say based on the result of our sample (N = 40): Mean = 3.73 ± 0.63 (95% CI)?

95% of the time, the true population mean will be in this interval.

95% of random samples of size 40 will yield confidence intervals that contain the population mean.

The sample means of 95% of the random samples of size 40 will be in this interval.

We can be 95% confident that the sample mean is in this interval.

What is a confidence interval?

- TBH, not super **confident** on this one. I think it has to do with how strong an effect is, but that is effect size so I'm not actually sure.
- I don't know the definition – but I know it (at least in part) reflects the precision of our results

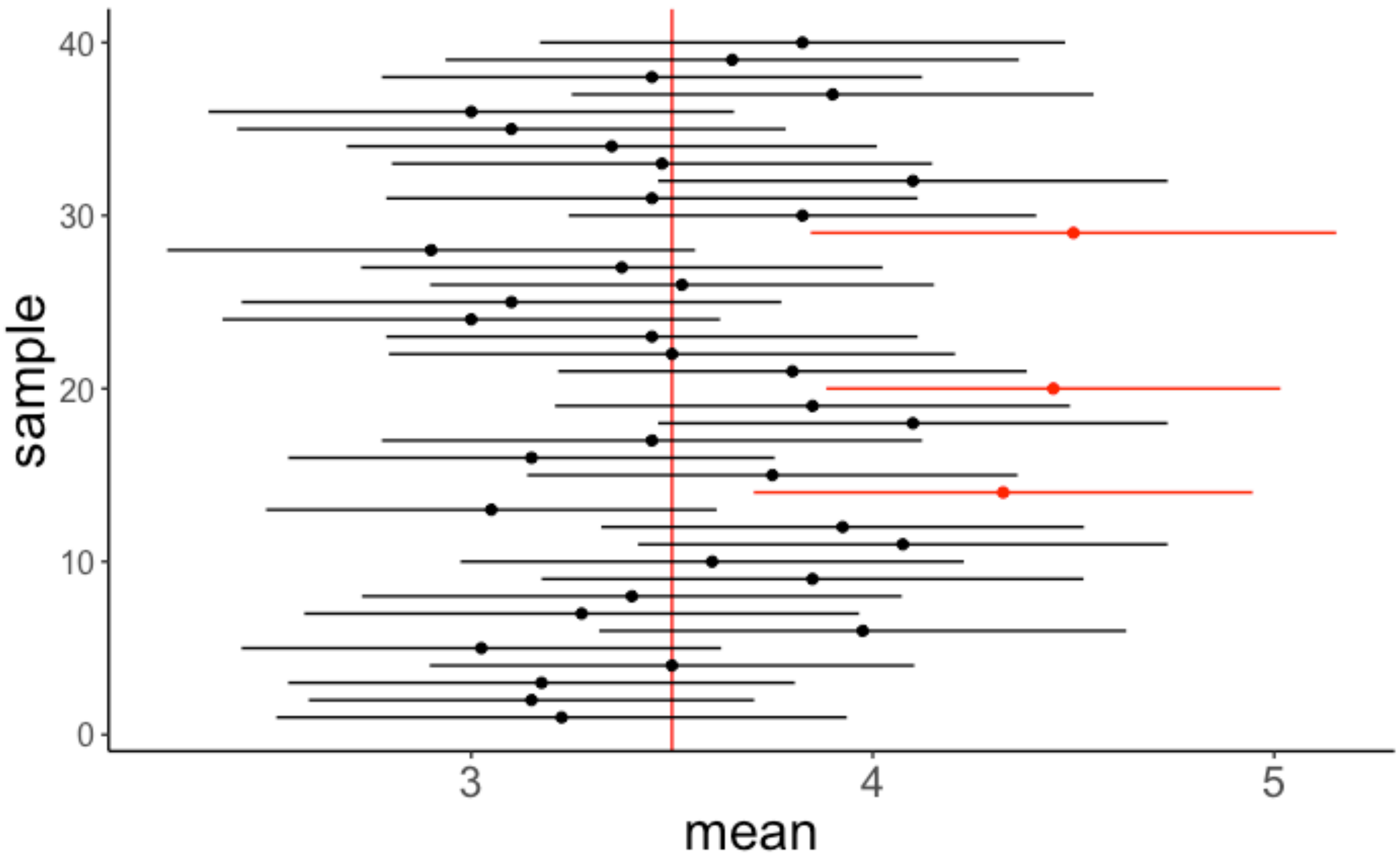
- The chance that your data lies within the interval.
- the likelihood that the true p-value is within the range of numbers contained in the CI

- Interval within which a certain percentage (typically 95%) of means would fall if the data was resampled using an identical sampling procedure/ population.
- For some confidence level X (usually X=95%), a confidence interval represents the range of values that the statistic of interest would fall inside X% of the time if you repeated the experiment multiple times.

Confidence interval

Definition

"If we were to repeat the experiment over and over, then 95 % of the time the confidence intervals contain the true mean."



Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust Misinterpretation of Confidence Intervals. *Psychonomic Bulletin & Review*, 21(5), 1157-1164.

What can we say based on the result of our sample (N = 40): Mean = 3.73 ± 0.63 (95% CI)?

95% of the time, the true population mean will be in this interval.

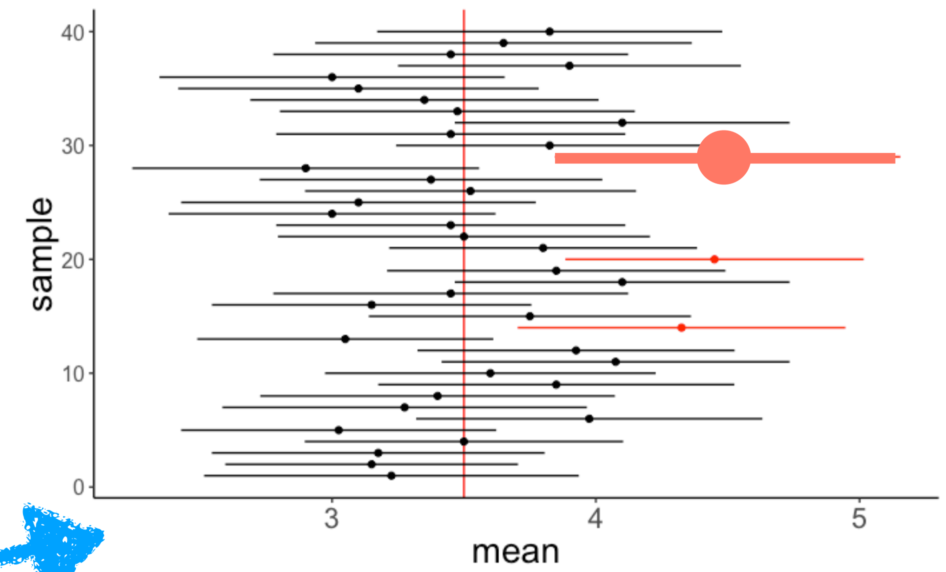
It either is in this interval or isn't.

95% of random samples of size 40 will yield confidence intervals that contain the population mean.

correct

The sample means of 95% of the random samples of size 40 will be in this interval.

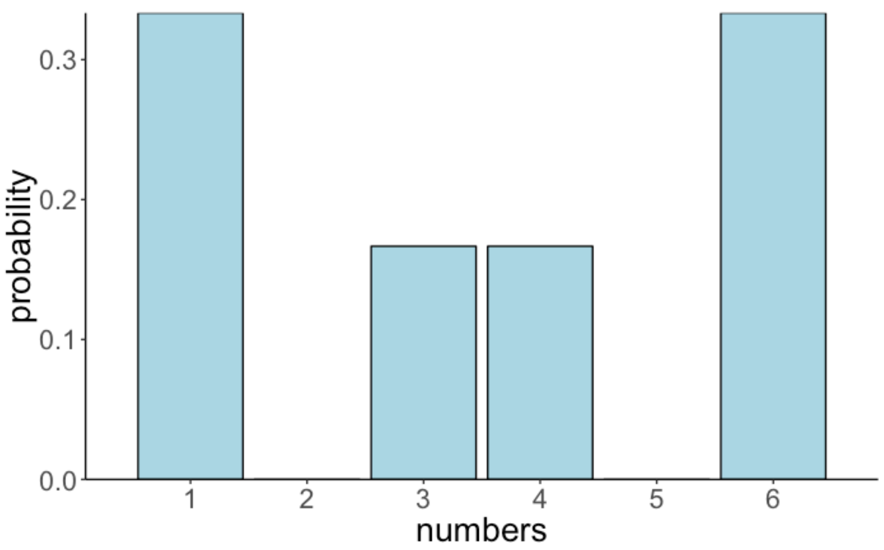
incorrect



We can be 95% confident that the sample mean is in this interval.

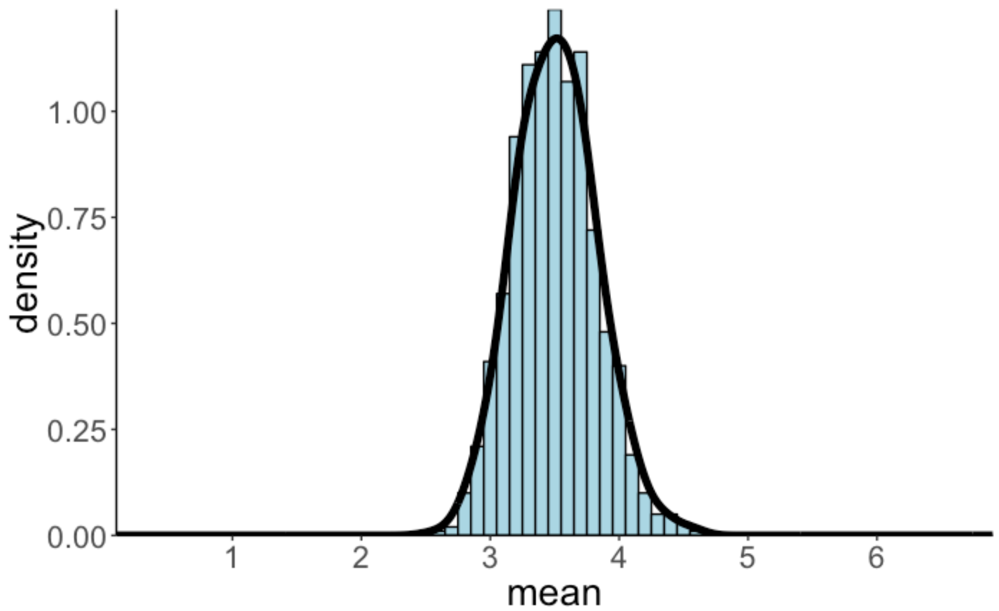
We know what the sample mean is.

Bootstrap



population distribution

repeated
sampling



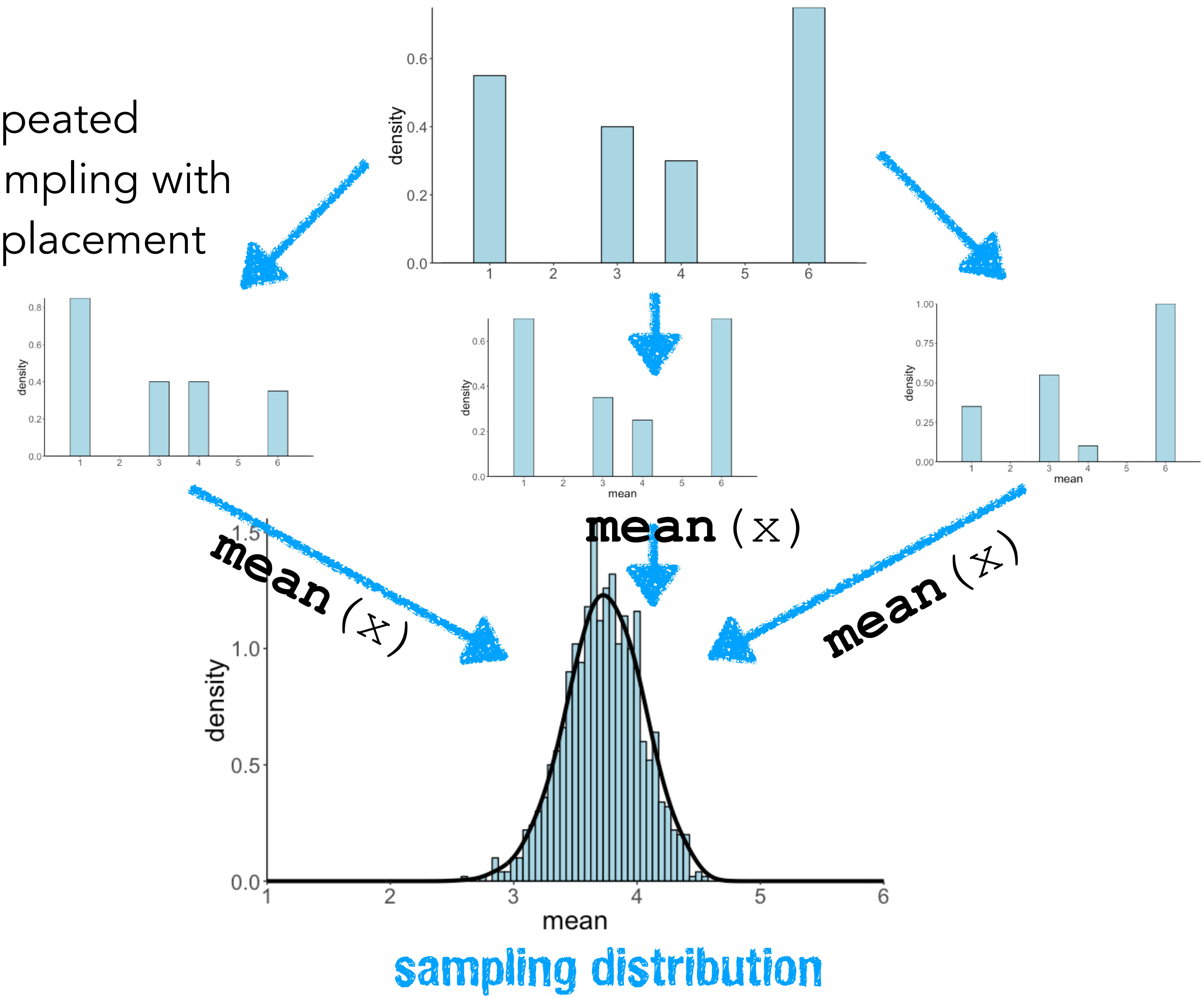
sampling distribution

but we don't know the population distribution!

Bootstrap

all we have is our sample

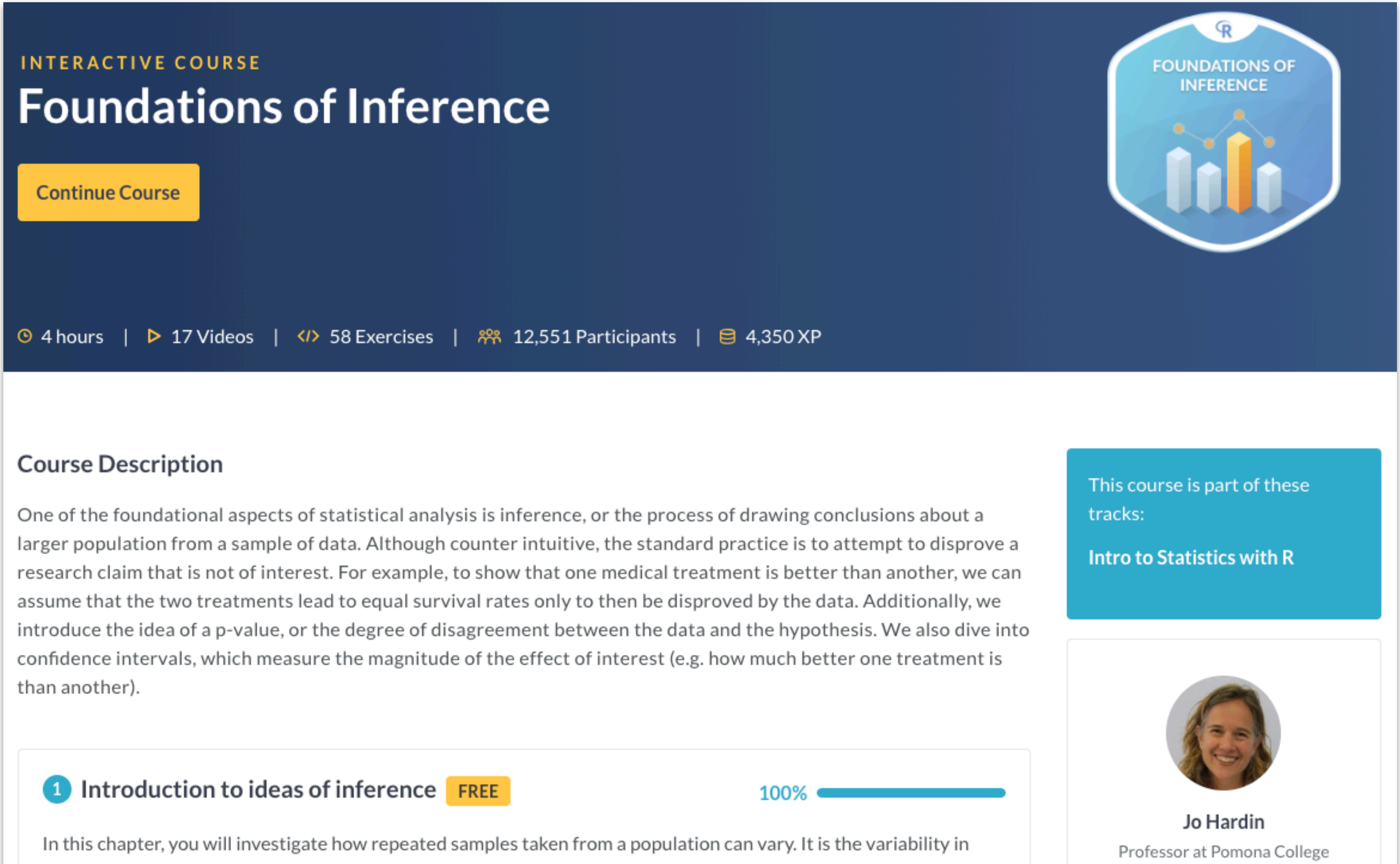
repeated sampling with replacement



Outline

- **Statistical inference**
  - drawing inferences about the population from our sample
- **Central limit theorem**
  - normal distribution of the mean (under certain conditions)
- **Sampling distributions**
  - the bridge between sample and population
  - theoretical/hypothetical distribution
  - simulate via permutation (or bootstrap)
- **p-values**
  - used for hypothesis testing
- **Confidence intervals**
  - parameter estimation

Datacamp



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

Feedback

How was the pace of today's class?

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!