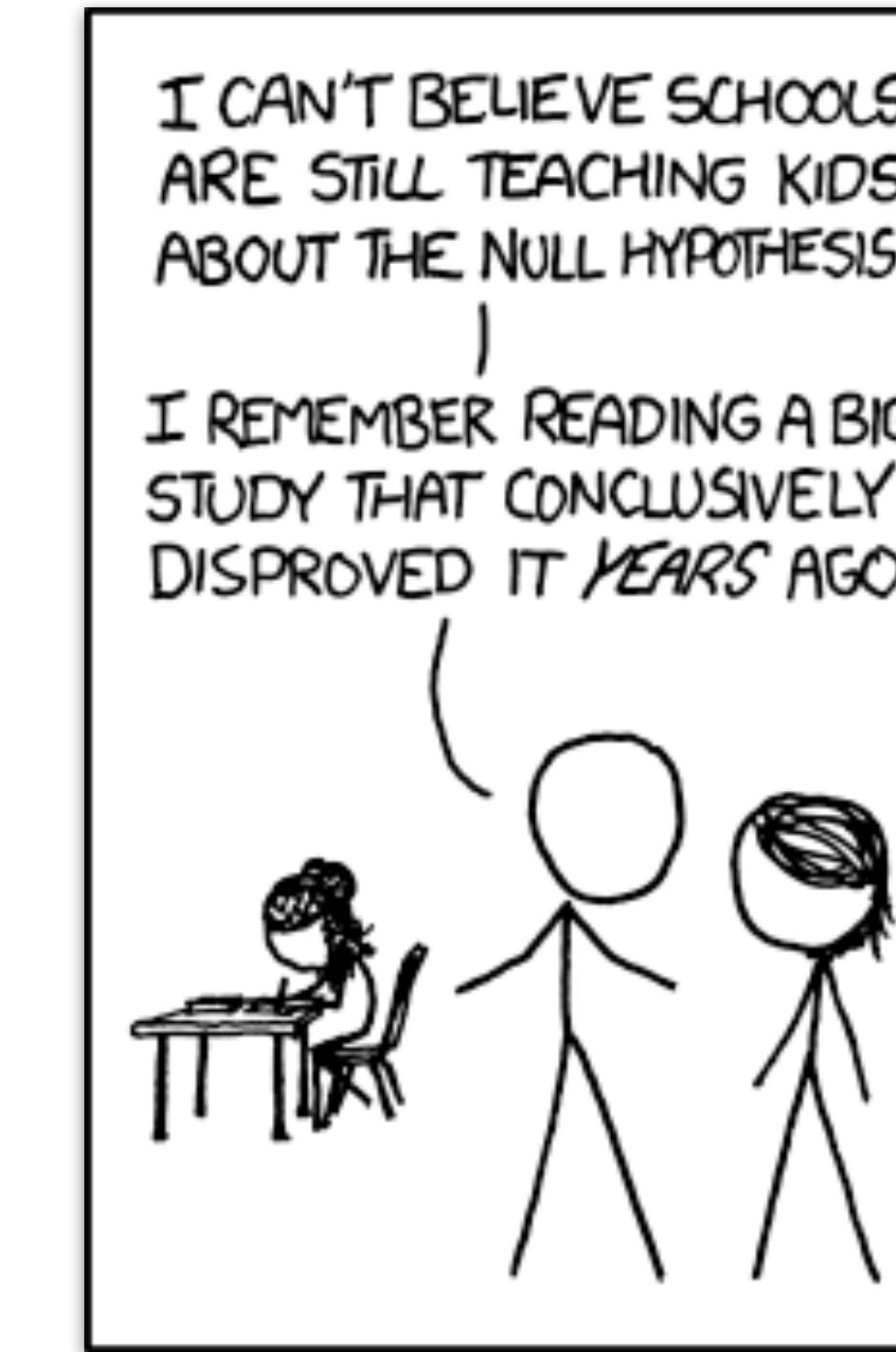


# Modeling data

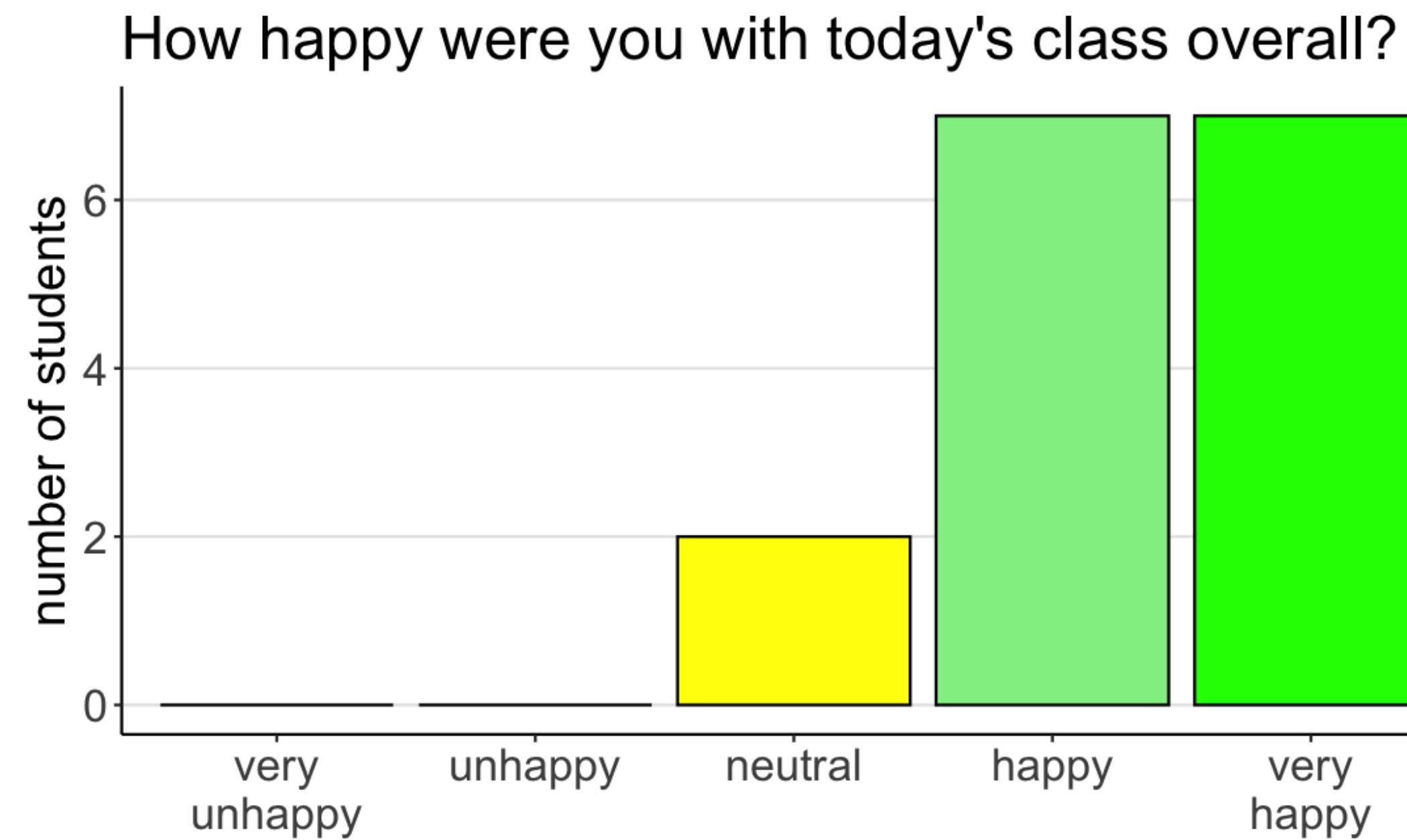
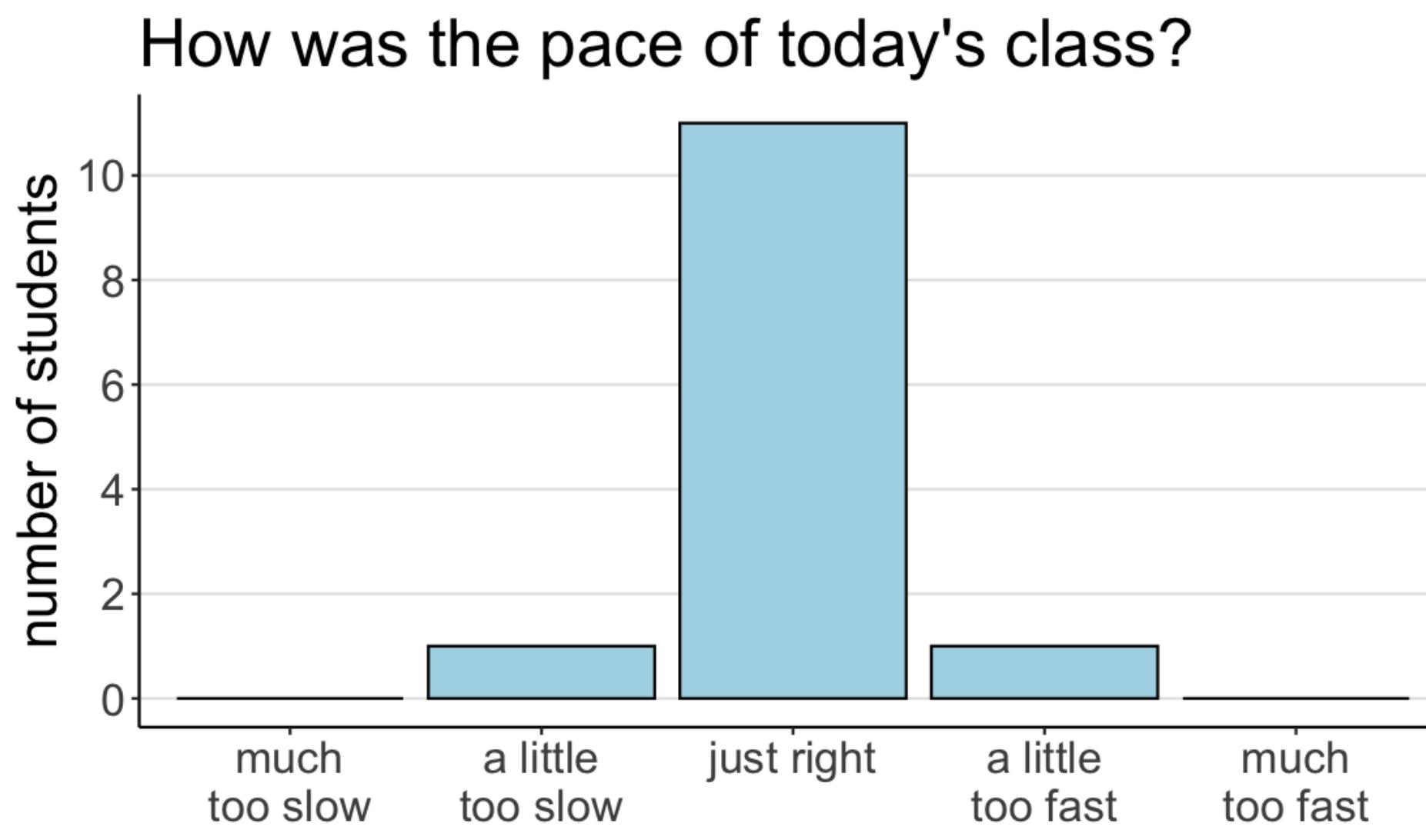


O COLLABORATIVE PLAYLIST  
**psych252**  
<https://tinyurl.com/psych252spotify25>

01/27/2025

# **Feedback**

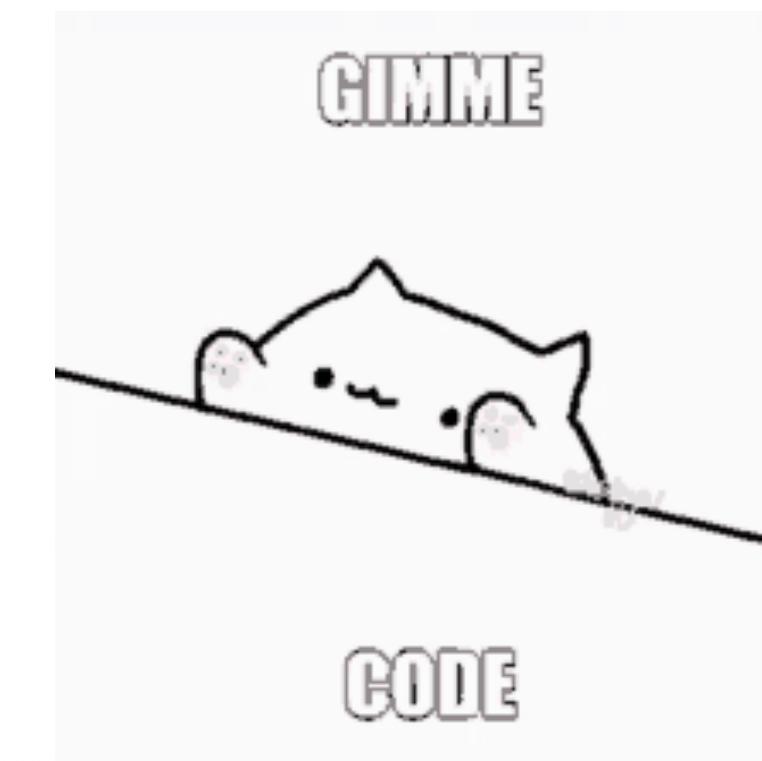
# Your feedback



- Seeing the logic and computations of the permutation test felt like a more concrete/accessible way to understand the logic of p values etc derived analytically!
- very thorough review of distributions and p-values!
- We should talk about p hacking
- Release the files so I can download them the night before class instead of in the morning.
- I would appreciate baking in time for questions. Maybe pause every 3 minutes of lecture and do a sanity check with the audience
- You are responding well to questions. Keep the good work!

**Things that came up ...**

# How to better understand!



08\_simulation2 - master - RStudio

simulation2.Rmd

```
1 ---  
2 title: "Class 8"  
3 author: "Tobias Gerstenberg"  
4 date: "January 24th, 2020"  
5 output:  
6   bookdown::html_document2:  
7   toc: true  
8   toc_depth: 4  
9   theme: cosmo  
10  highlight: tango  
11  pandoc_args: ["--number-offset=7"]  
12 ---  
13  
14 # Simulation 2  
15  
16 In which we figure out some key statistical concepts through simulation and plotting. On the menu we have:  
17 | Sampling distributions  
18 | - p-value  
19 | - Confidence interval  
20  
21 ## Load packages and set plotting theme  
22  
23 ``{r simulation2-01, include=FALSE}  
24 # run this code chunk once to make sure you have all the packages  
25 install.packages(c("janitor"))  
26 ``  
27  
28 ``{r simulation2-02, message=FALSE}  
29 library("knitr") # for knitting RMarkdown  
30 library("kableExtra") # for making nice tables  
31 library("janitor") # for cleaning column names  
32 library("tidyverse") # for wrangling, plotting, etc.  
33 ``  
34  
35 ``{r simulation2-03}  
36 theme_set(theme_classic() + #set the theme  
37   theme(text = element_text(size = 20))) #set the default text size  
38  
39 opts_chunk$set(comment = "",  
40   fig.show = "hold")  
41 ``  
42  
17:1 Simulation 2
```

Console

```
> ggplot(data = tibble(x = c(mean - 3 * sd, mean + 3 * sd),  
+   mapping = aes(x = x)) +  
+   stat_function(fun = ~ dnorm(., mean = mean, sd = sd),  
+     color = "black",  
+     size = 2) +  
+   geom_vline(xintercept = qnorm(c(0.025, 0.975), mean = mean, sd = sd),  
+     linetype = 2)  
> # labs(x = "performance")  
>
```

Environment

confidence_level	0.95
df.condition1	'kableExtra' chr "<table class='table table-striped' style='width: a..."
i	20L
k	3
mean	0
n	10
n_simulations	1000
population_mean	3.5
sample_n	20
sample_size	1000
sd	1

Plots

Packages

Help

Viewer

R: Subset rows using their positions

slice (dplyr)

R Documentation

### Subset rows using their positions

#### Description

slice() lets you index rows by their (integer) locations. It allows you to select, remove, and duplicate rows. It is accompanied by a number of helpers for common use cases:

- slice\_head() and slice\_tail() select the first or last rows.
- slice\_sample() randomly selects rows.
- slice\_min() and slice\_max() select rows with highest or lowest values of a variable.

If .data is a grouped\_df, the operation will be performed on each group, so that (e.g.) slice\_head(df, n = 5) will select the first five rows in each group.

#### Usage

```
slice(.data, ..., .preserve = FALSE)  
slice_head(.data, ..., n, prop)  
slice_tail(.data, ..., n, prop)  
slice_min(.data, order_by, ..., n, prop, with_ties = TRUE)  
slice_max(.data, order_by, ..., n, prop, with_ties = TRUE)  
slice_sample(.data, ..., n, prop, weight_by = NULL, replace = FALSE)
```

#### Arguments

.data A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dplyr). See Methods, below, for more details.

... For slice():<data-masking> Integer row values.

# 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 samples that allows you to make claims about the population of interest. It is important to remember that the

This course is part of these tracks:  
**Intro to Statistics with R**



**Jo Hardin**  
Professor at Pomona College

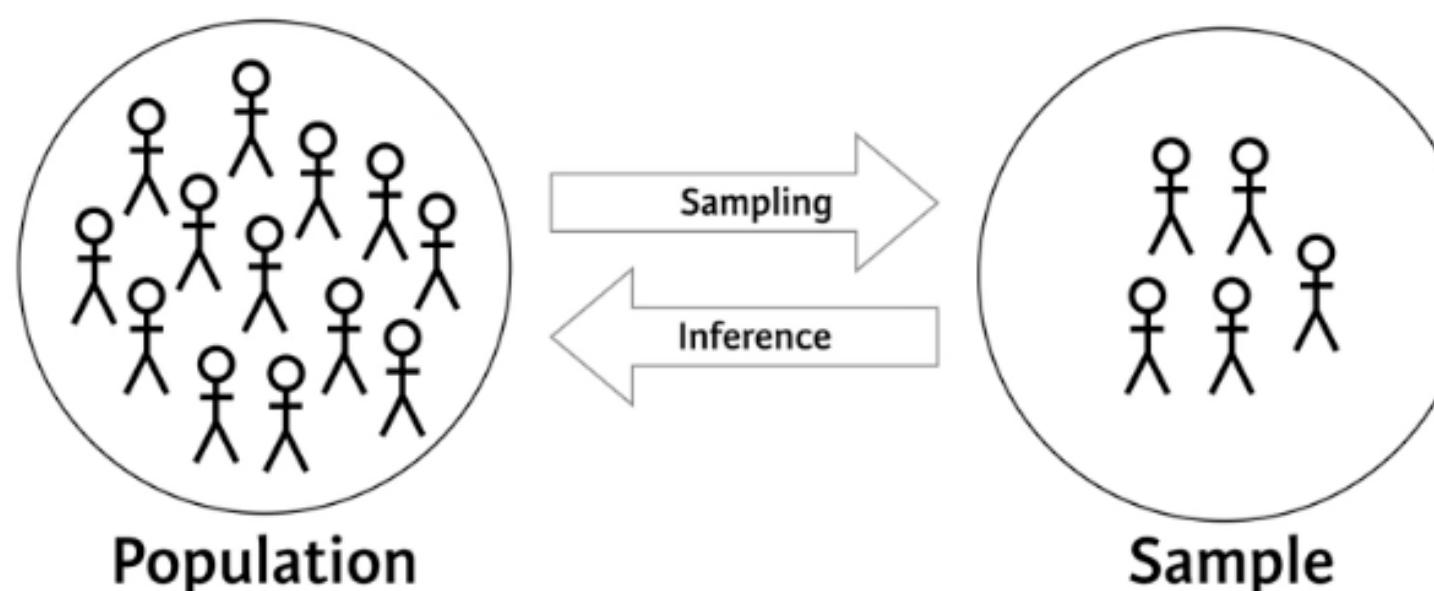
<https://www.datacamp.com/courses-foundations-of-inference>

# Plan for today

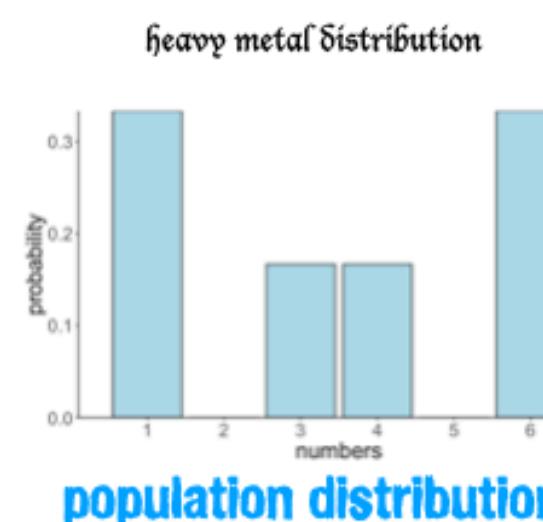
- Quick recap
- Statistical concepts
  - Confidence intervals
  - Bootstrapping
- Cookbook vs. Model Comparison
- Modeling data
- Hypothesis testing as model comparison

# **Quick recap**

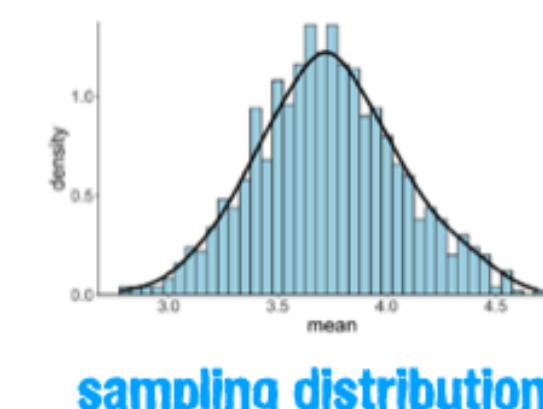
# Quick recap: Inference in frequentist statistics



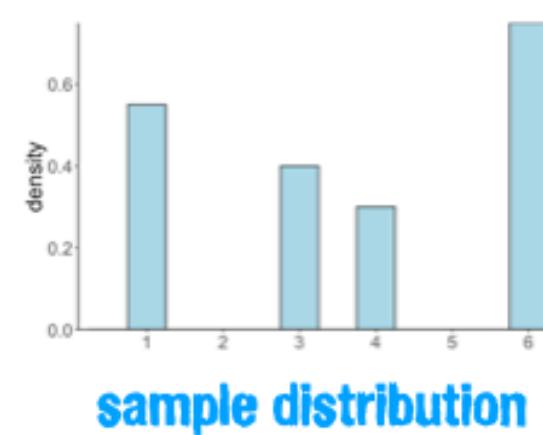
## 3 distributions in statistical inference



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



- 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, ...)
- make an inference about the population via the sampling distribution

# Quick recap: What is a p-value?

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.

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

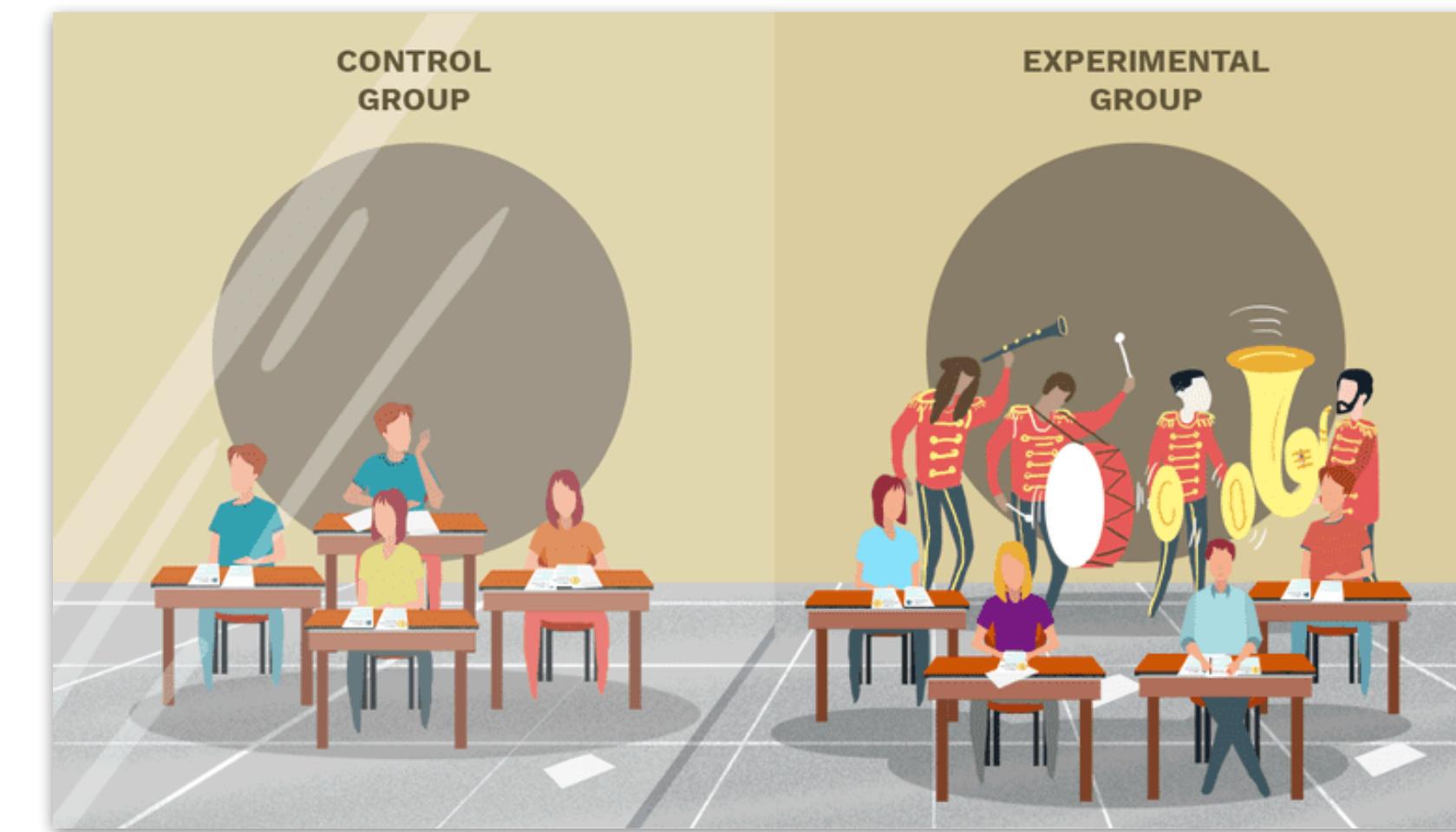
*what we're actually interested in!*

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

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

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

$H$  = Hypothesis  
 $D$  = Data



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

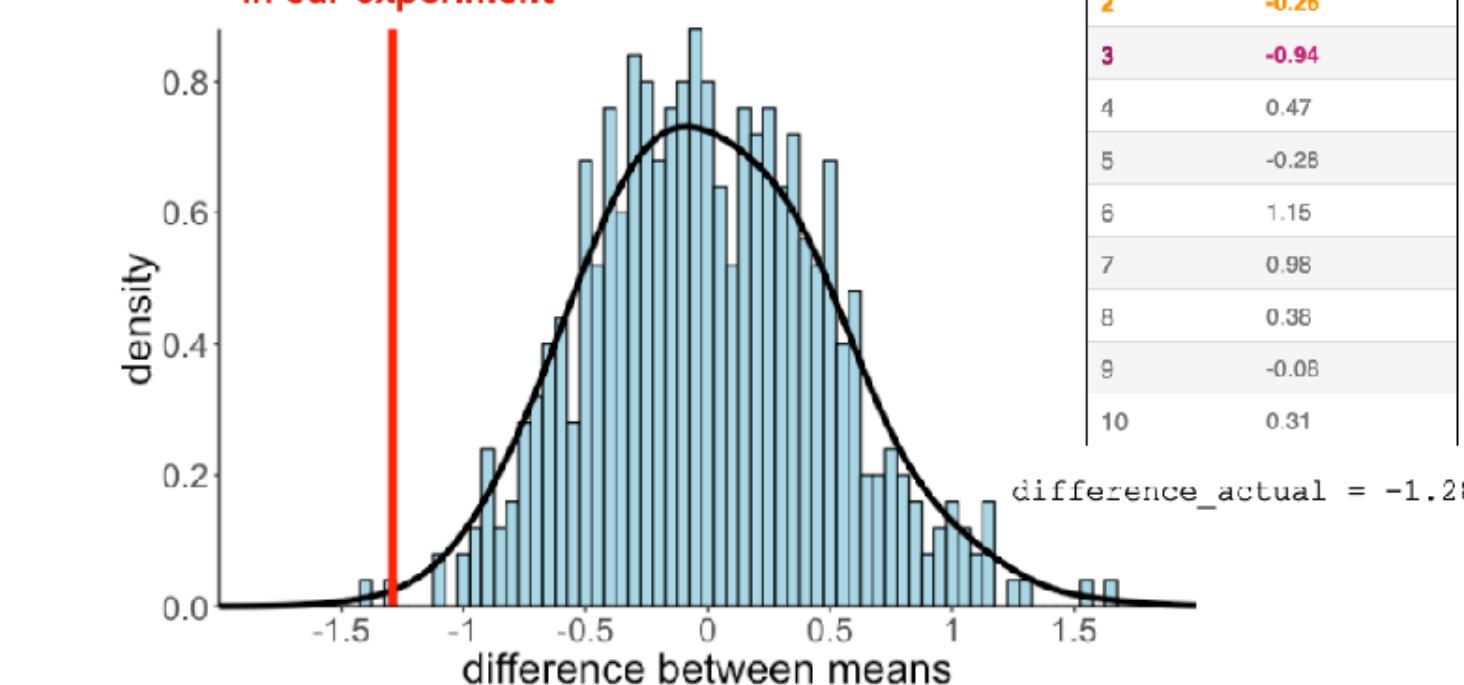
participant	condition	performance
1	experimental	4.65
2	control	5.87
3	control	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	control	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

participant	condition	performance
1	control	4.65
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	control	4.42
27	control	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

Permutation test

observed difference  
in our experiment



```

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

```

p-value = .002

# **Statistical concepts**

# Confidence intervals



# Confidence interval

Goal: Estimate  $\mu = \text{the mean of the population distribution}$

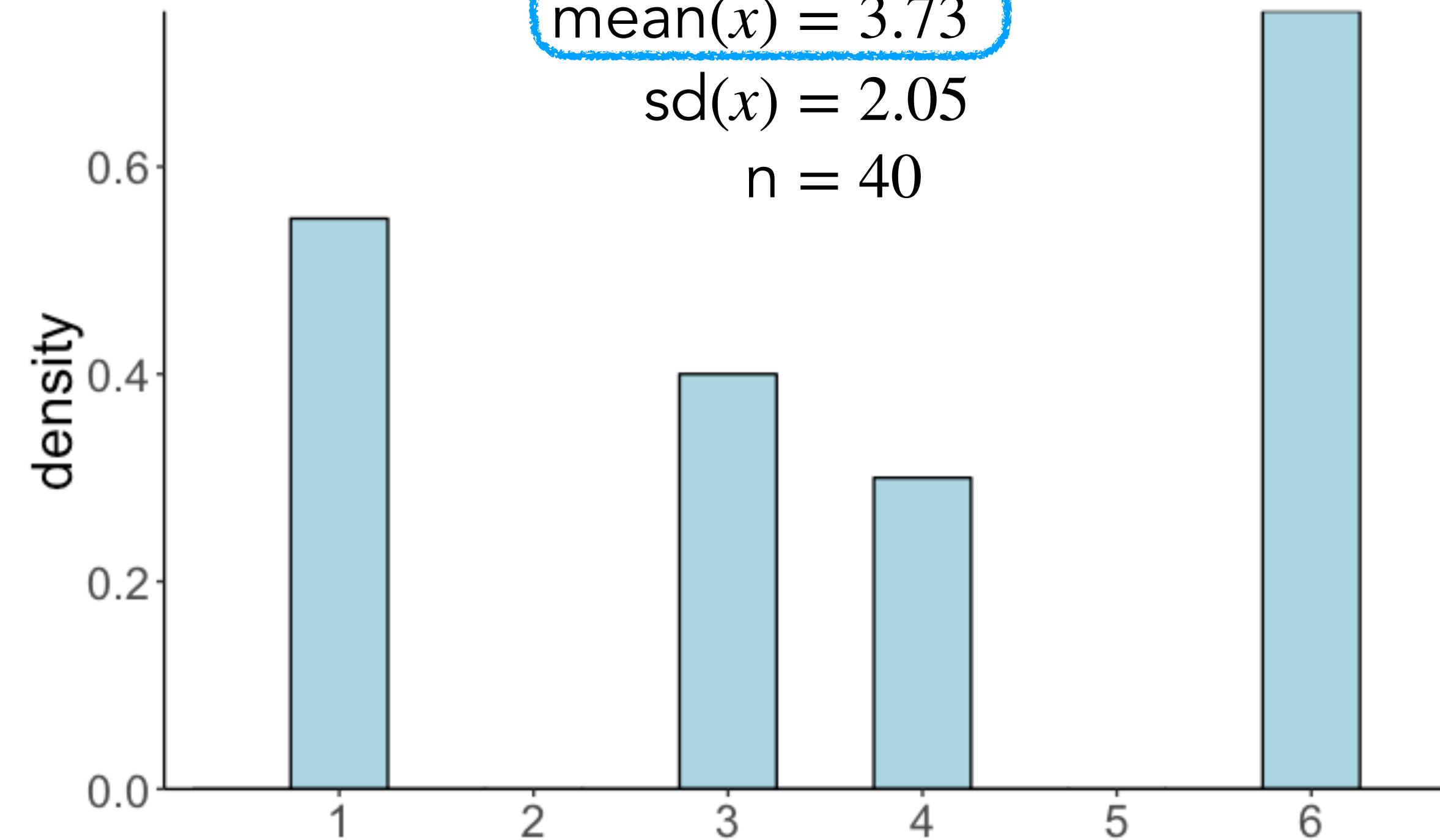
Confidence interval of the mean = point estimate  $\pm$  critical value

the sample mean is our best  
guess of the population mean

$$\begin{aligned} \text{mean}(x) &= 3.73 \\ \text{sd}(x) &= 2.05 \\ n &= 40 \end{aligned}$$

depends on

1. variance in the data
2. number of data points
3. desired level of confidence



# Confidence interval

**Goal: Estimate  $\mu$  = the mean of the population distribution**

**what we need:**

- sample mean
- sample standard deviation
- sample size
- desired level of confidence

# Confidence interval

Confidence interval of the mean = point estimate  $\pm$  critical value

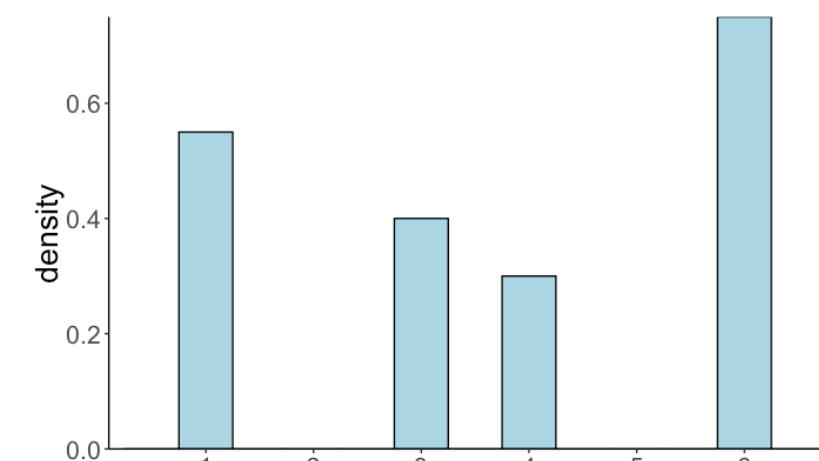
I would like to be 95% confident.

How confident would you like to be that you're correct?



# Confidence interval

**Parametric assumption:** The sampling distribution of the mean is a normal distribution.

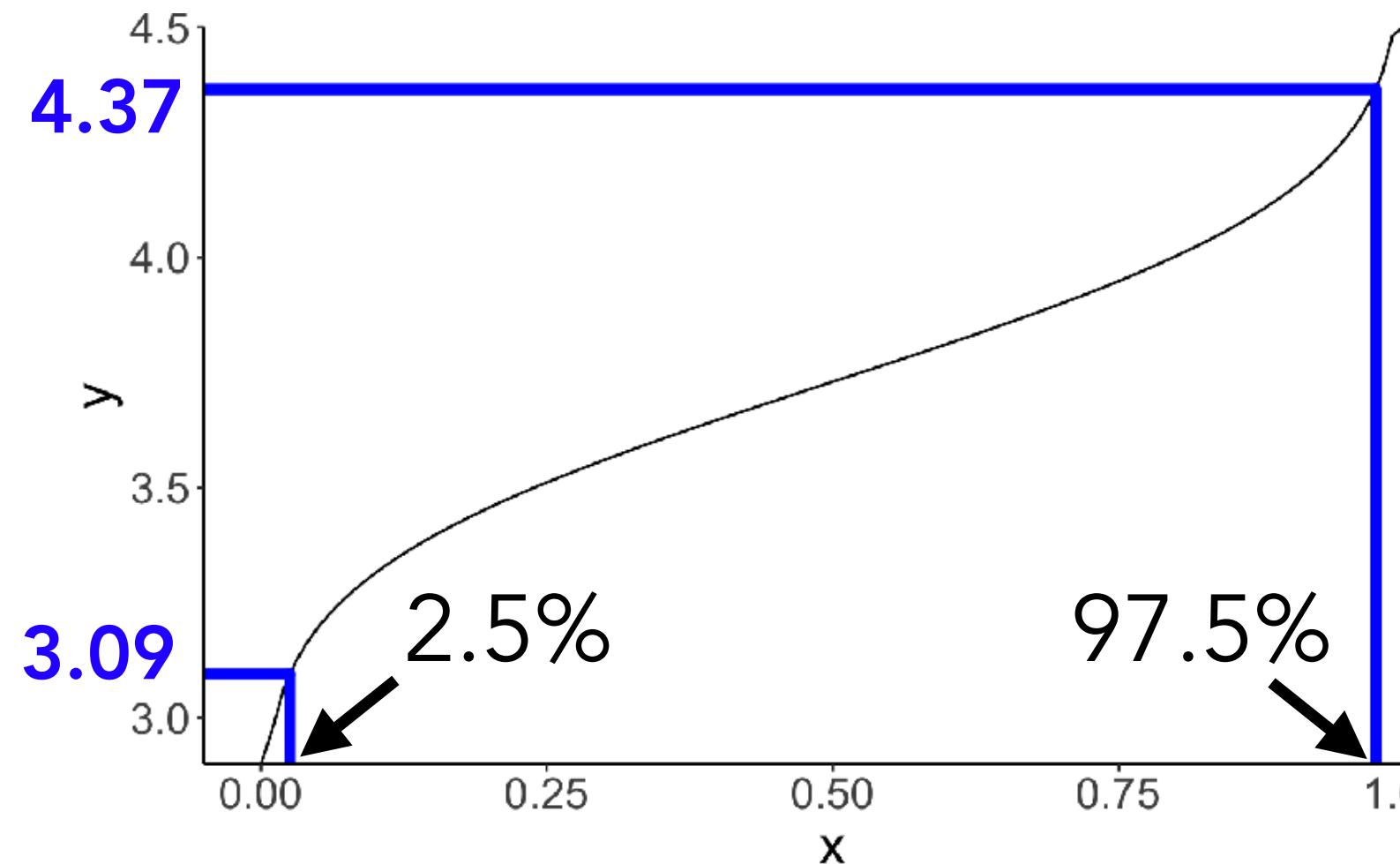


our sample

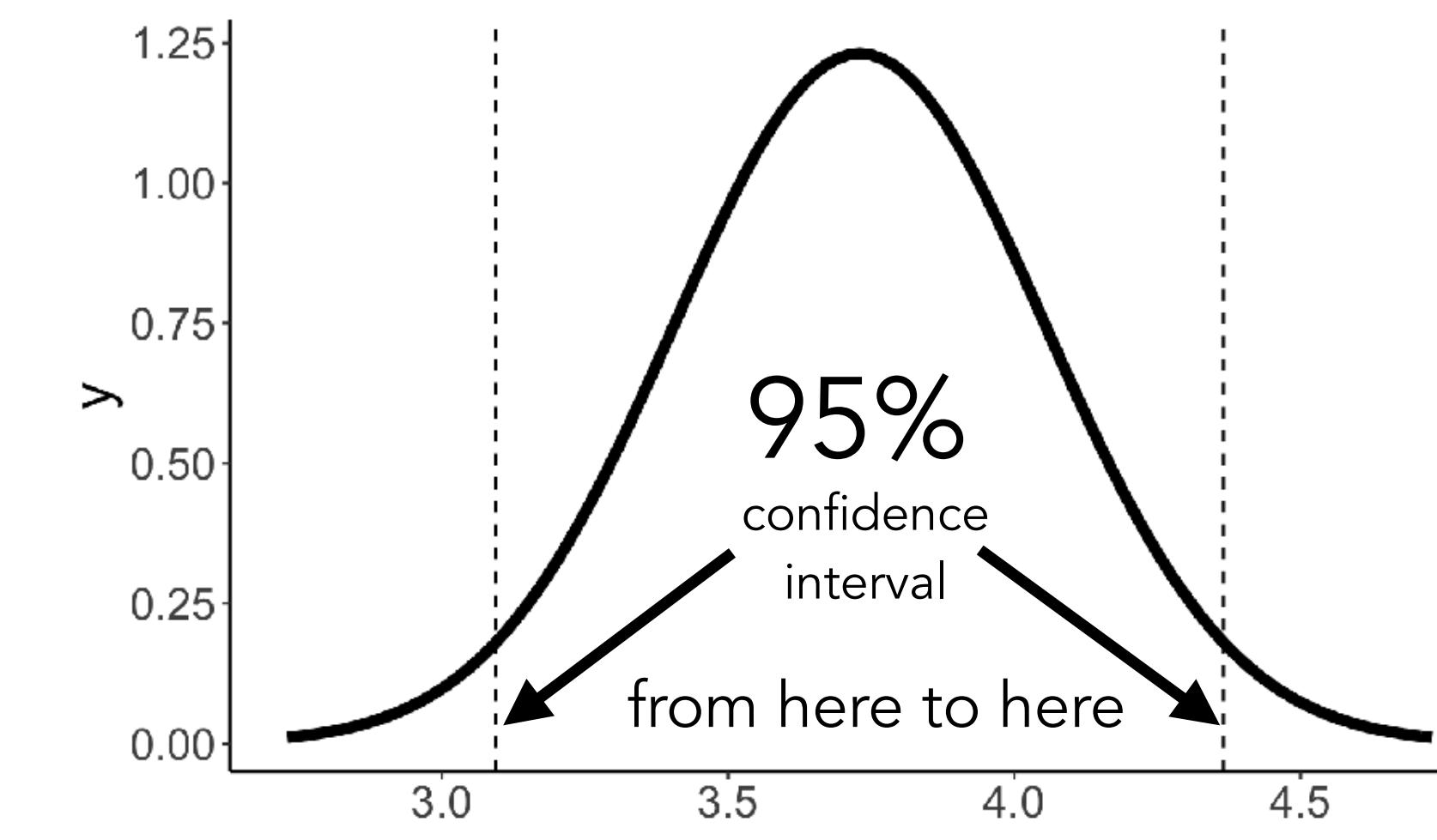
$$\begin{aligned}\text{mean}(x) &= 3.73 \\ \text{sd}(x) &= 2.05 \\ n &= 40\end{aligned}$$

$$\begin{aligned}\text{mean}(\bar{x}) &= 3.73 \\ \text{sd}(\bar{x}) &= \frac{\text{sd}(x)}{\sqrt{n}} = \frac{2.05}{\sqrt{40}} \approx 0.324\end{aligned}$$

sampling distribution of the mean

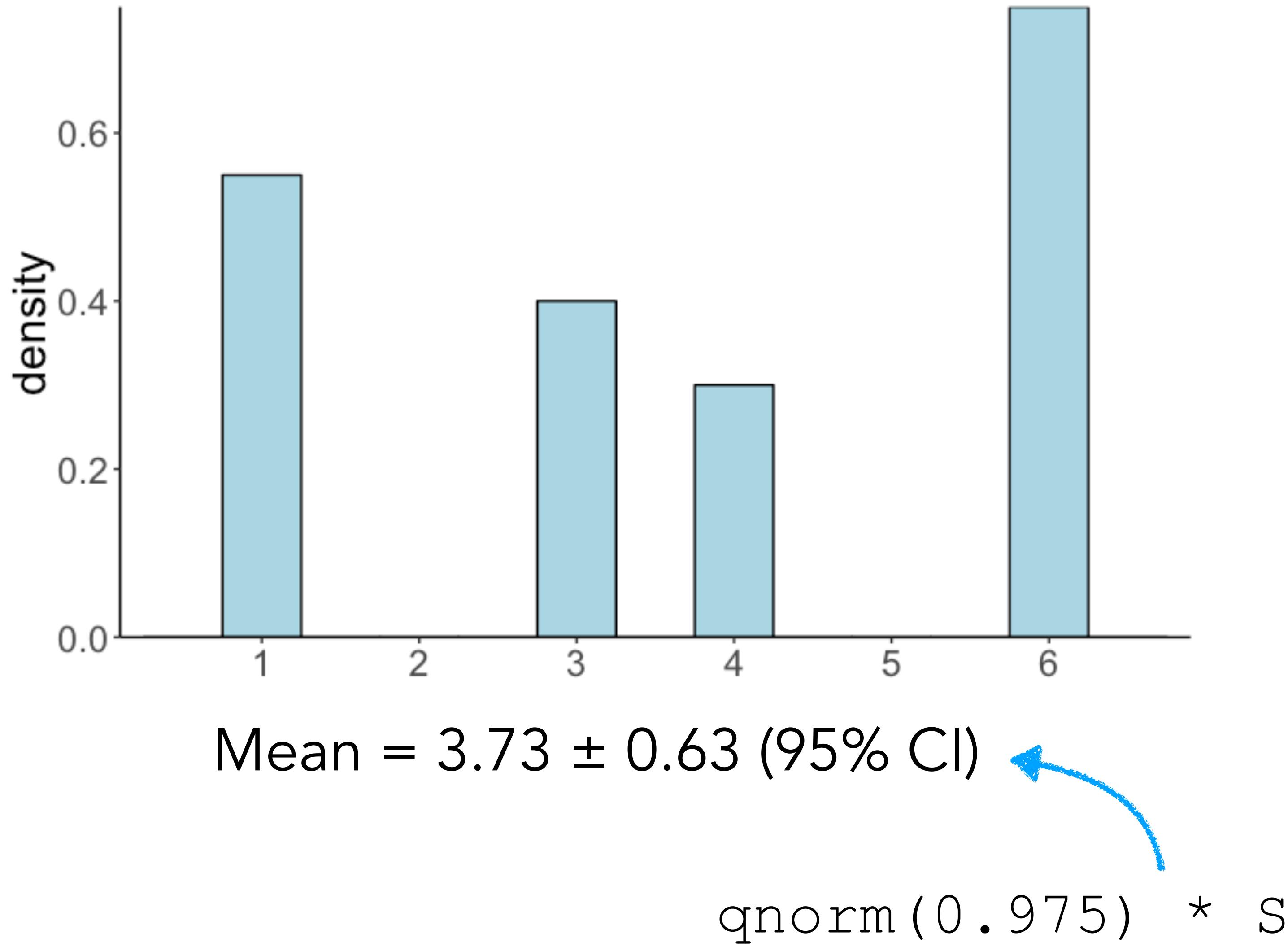


`~ qnorm(., mean = 3.73, sd = 0.324)`



`~ dnorm(., mean = 3.73, sd = 0.324)`

## What does the confidence interval mean?



# Confidence interval

Confidence interval of the mean = point estimate  $\pm$  critical value

I would like to be 95% confident.

How confident would you like to be that you're correct?



**But what does  
95% confident even mean?**



# What can we say based on the result of our sample ( $N = 40$ ): Mean = $3.73 \pm 0.63$ (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.

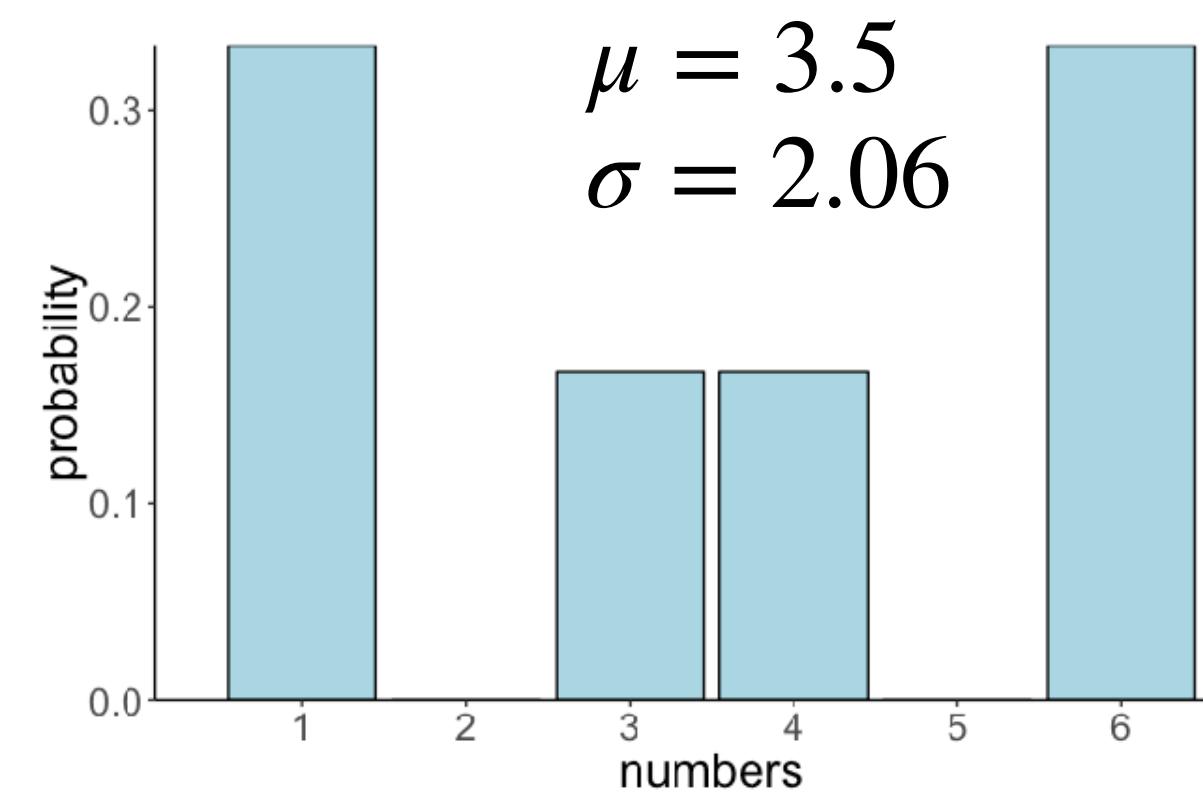
We can be 95% confident that the sample mean is in this interval.

# What is a confidence interval? Your answers

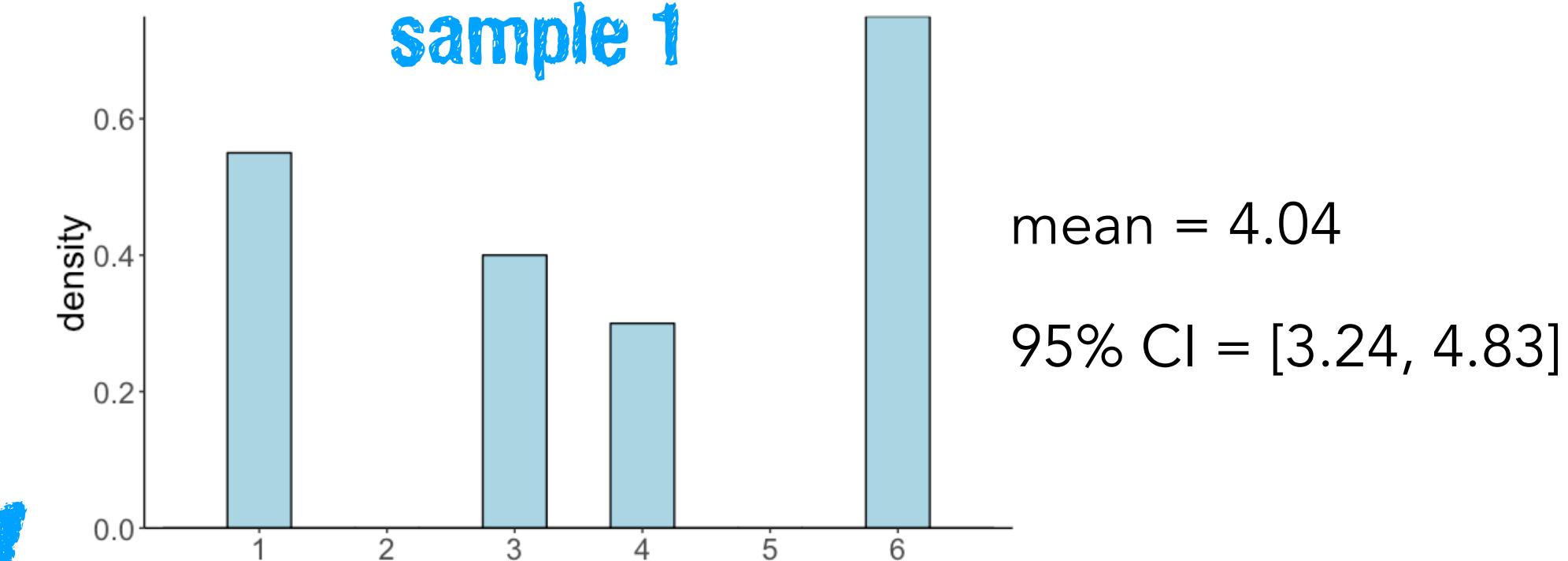
- I'm just extrapolating from general basic knowledge of "confidence," but perhaps this is the range of result values for which one would take their results to be statistically significant. E.g.,  $x$  to  $y$ , where a result less than  $x$  or greater than  $y$  would have  $p > 0.05$  (or some other p-value).
- if calculated repeatedly, a confidence interval is an interval that will contain the true value of an estimate X% (e.g., 95%) of the time
- confidence interval refers to a brackets of values that contains the true parametric under a specific probability.
- the space around the actual result that you got, where you are confident the 'true' effect is??
- Estimated percentage of the time you expect your values to fall within the range.

# Confidence interval

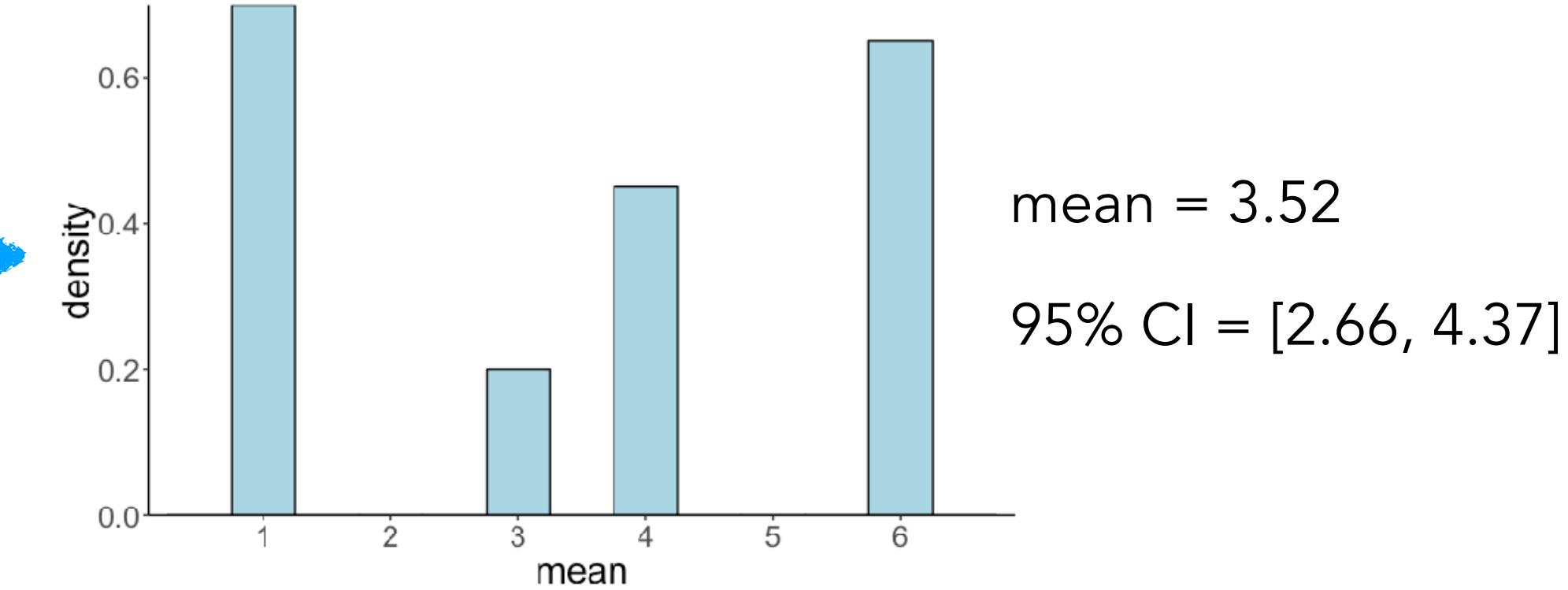
heavy metal distribution



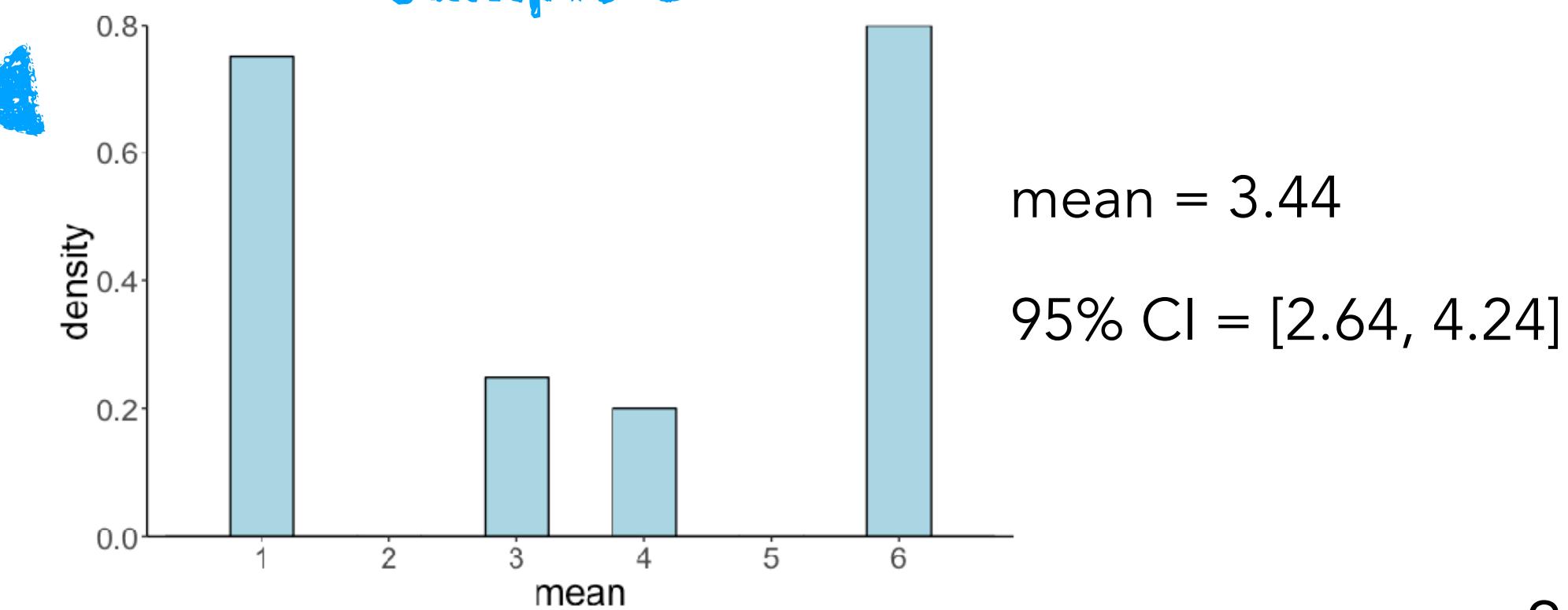
population distribution



sample 2



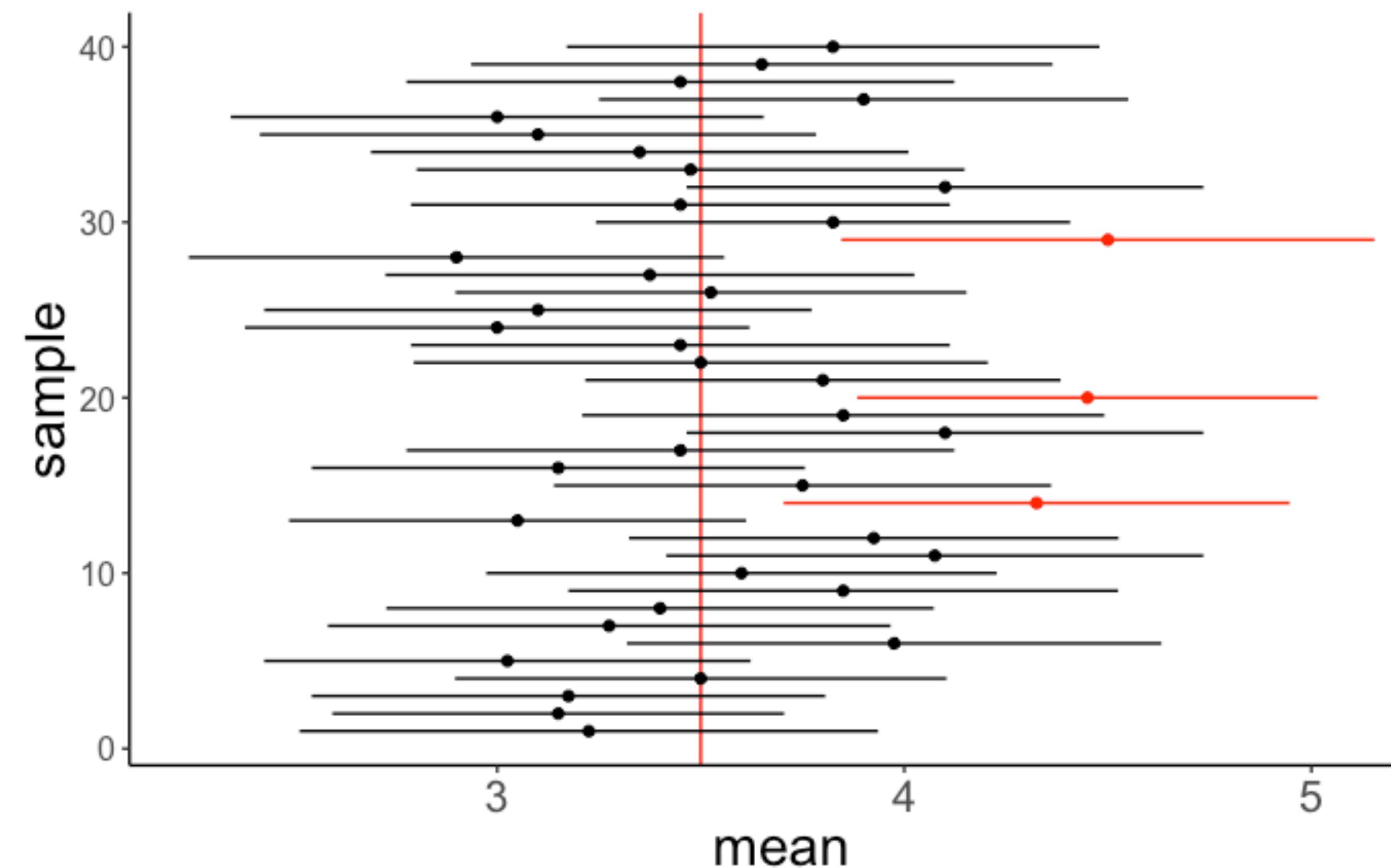
sample 3



# 95% confidence interval

## Definition

"If we were to repeat the experiment over and over, then 95% of the time the confidence interval contains the estimate of interest."



# What can we say based on the result of our sample ( $N = 40$ ):

Mean =  $3.73 \pm 0.63$  (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.

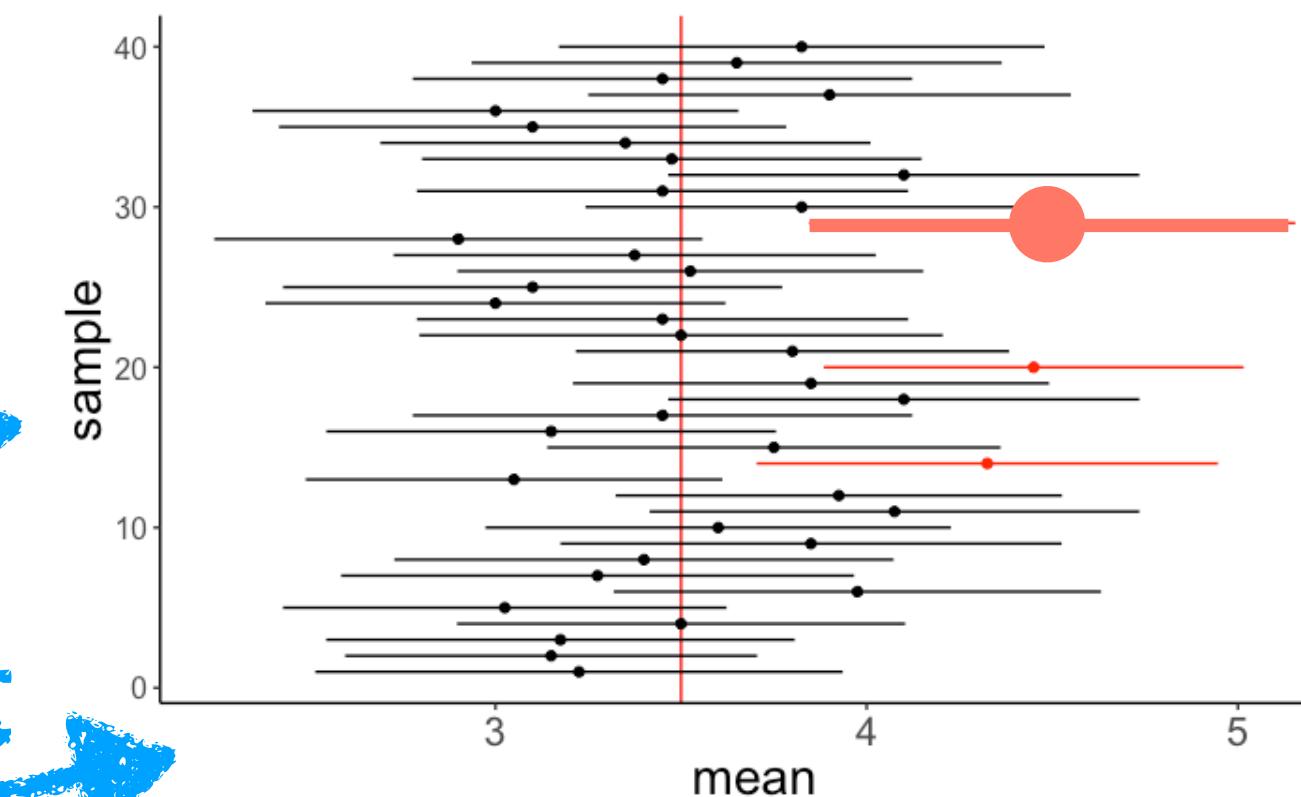
The sample means of 95% of the random samples of size 40 will be in this interval.

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

It either is in this interval or isn't.

correct

incorrect

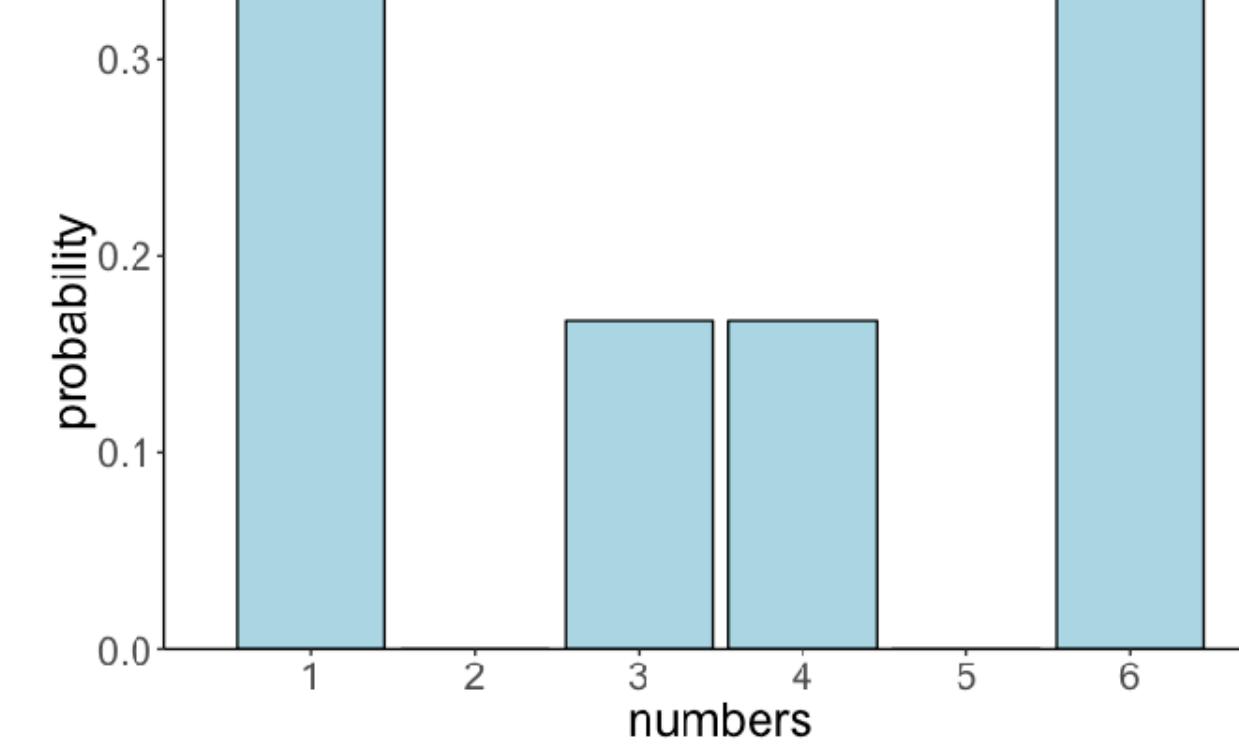


We know what the sample mean is.

# Bootstrapping

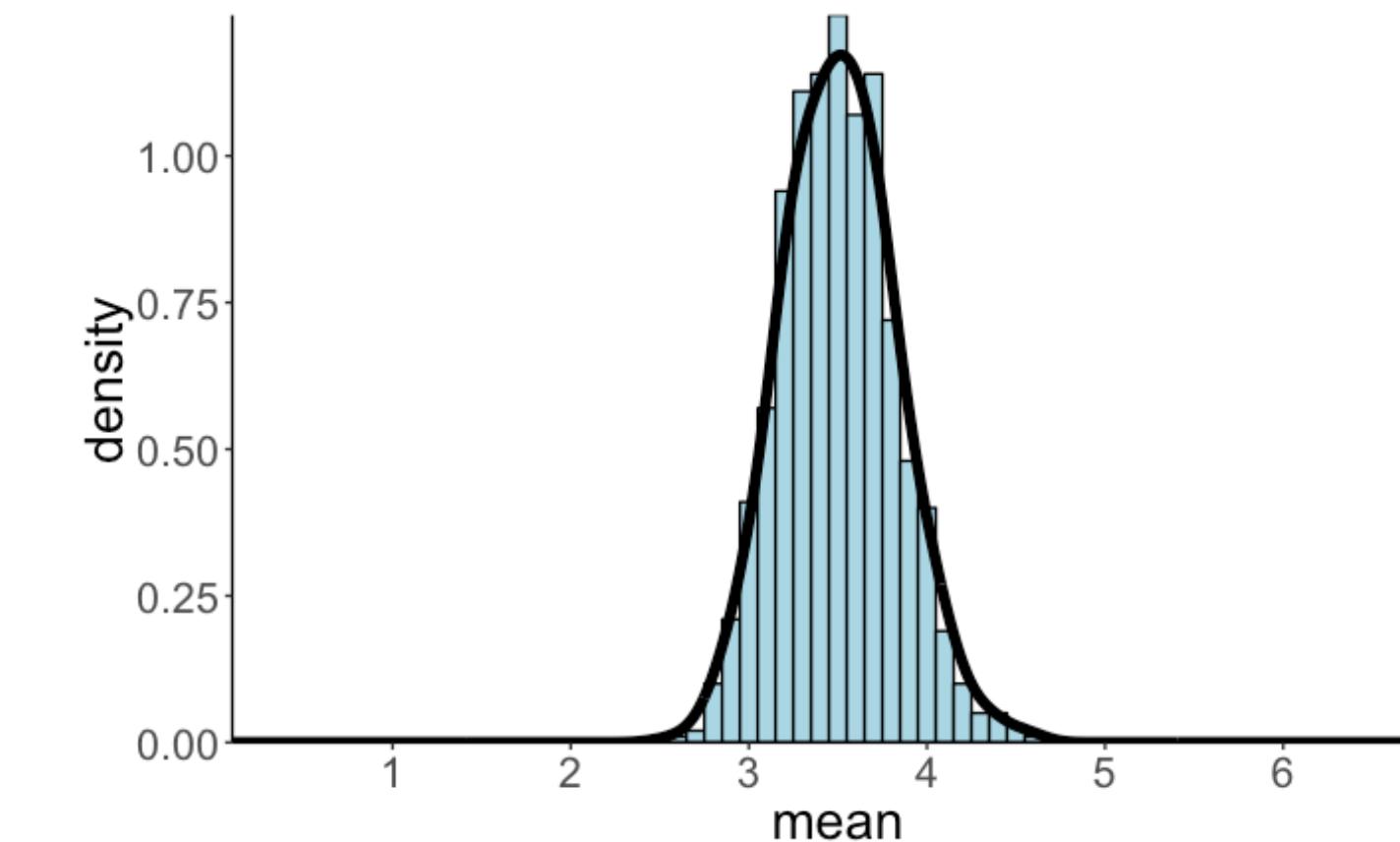


# Bootstrap



**population distribution**

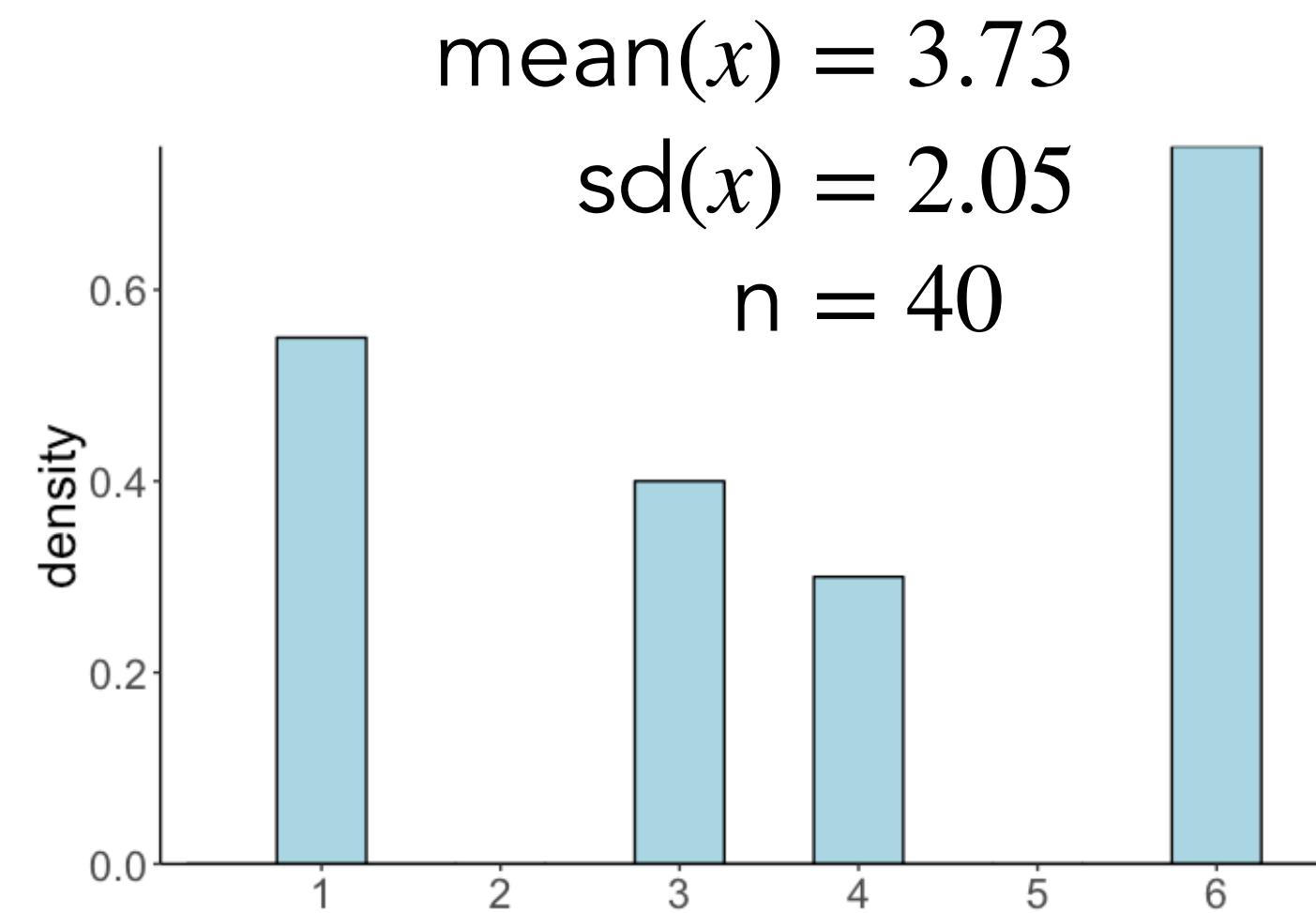
repeated  
sampling



**sampling distribution**

**but we don't know the population distribution!**

# Bootstrap

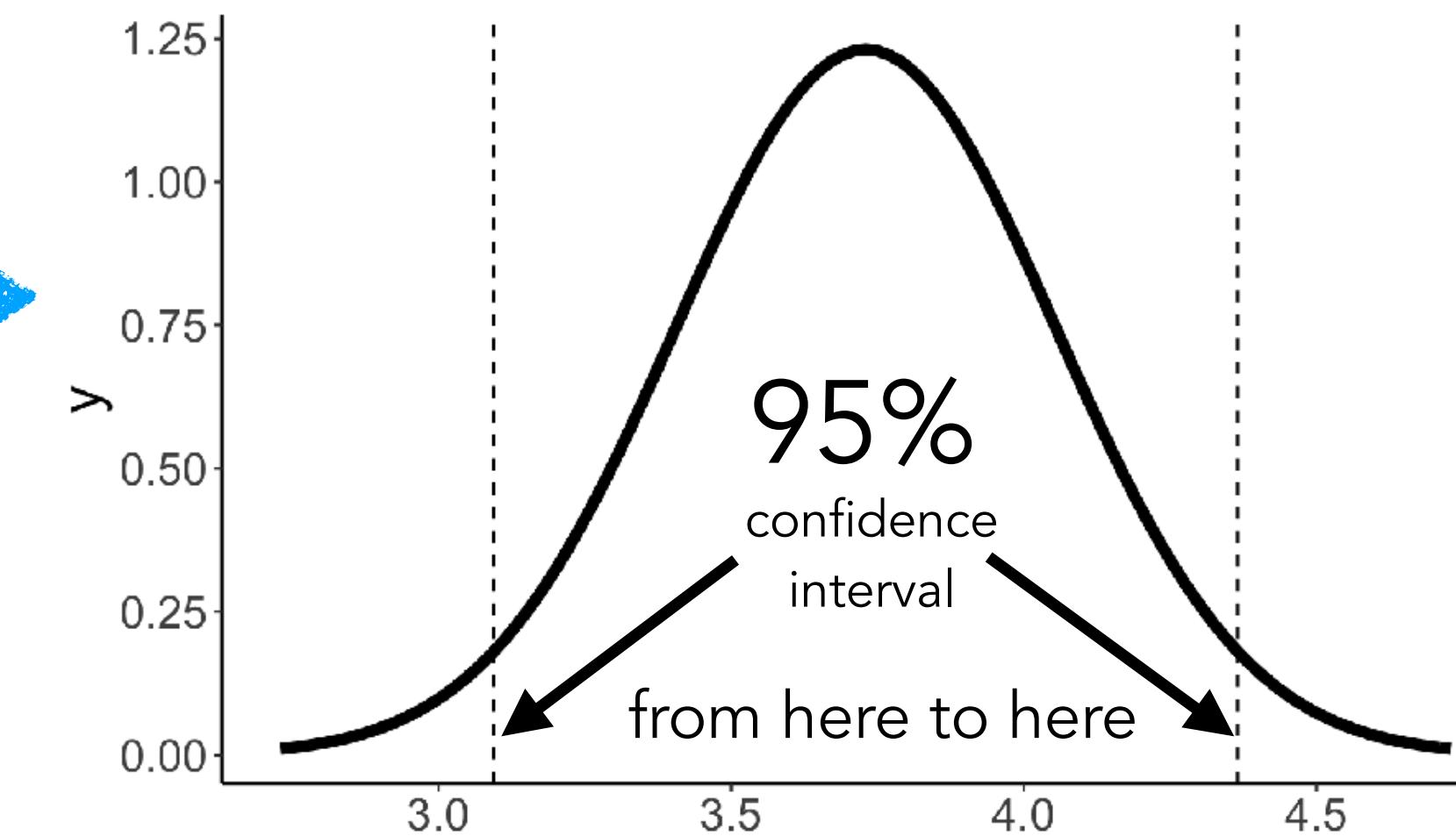


assuming a normal distribution

A blue arrow originates from the text "assuming a normal distribution" and points towards the normal distribution curve shown below.

mean( $\bar{x}$ ) = 3.73  
 $sd(\bar{x}) = \frac{sd(x)}{\sqrt{n}} = \frac{2.05}{\sqrt{40}} \approx 0.324$

sampling distribution of the mean

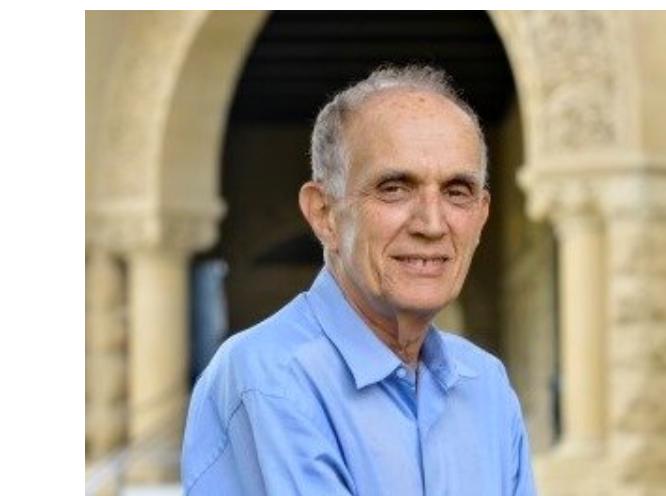


~ dnorm(., mean = 3.73, sd = 0.324)

# Bootstrap

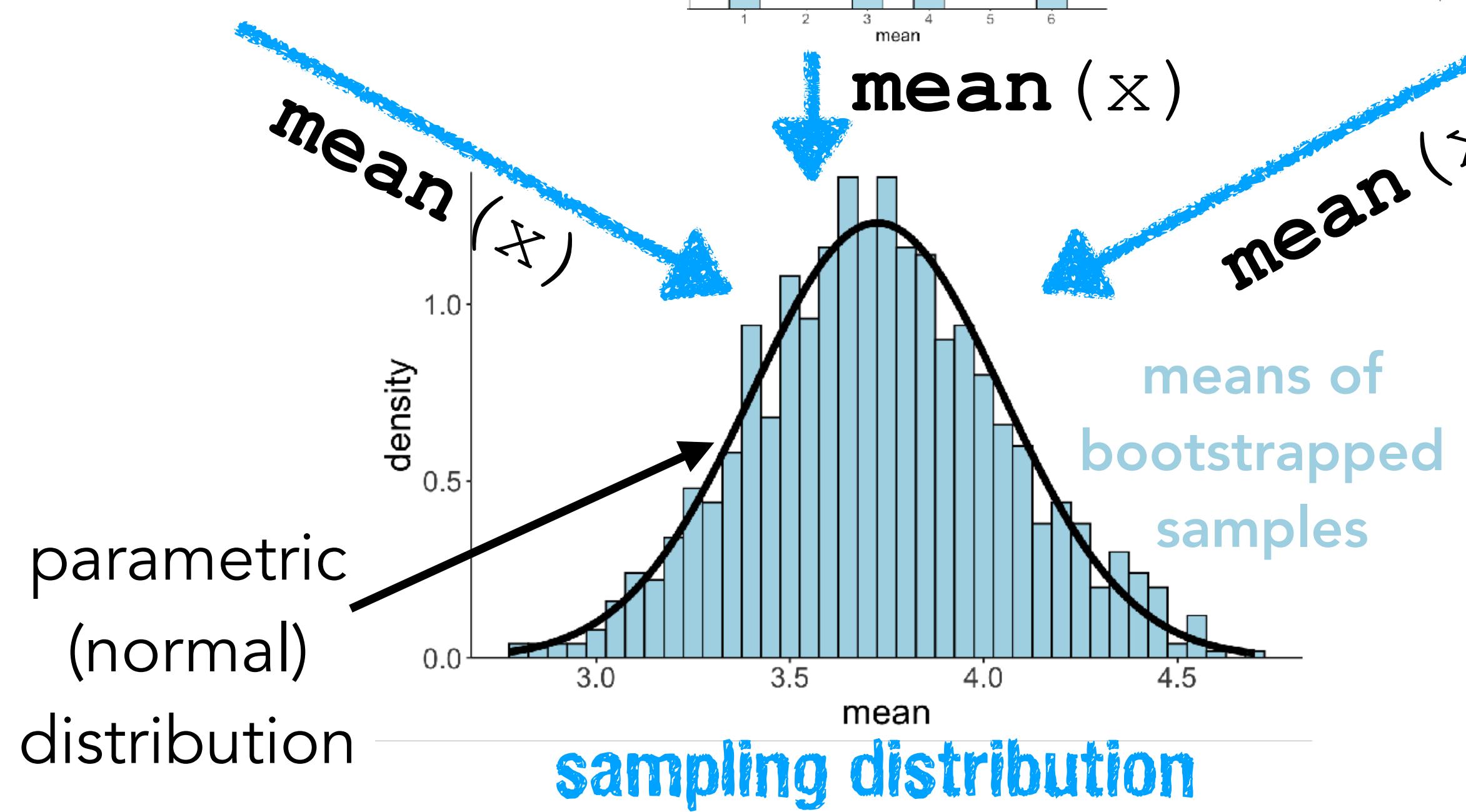
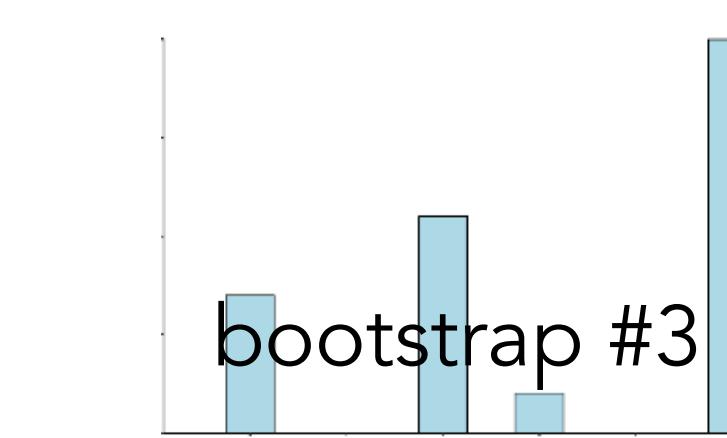
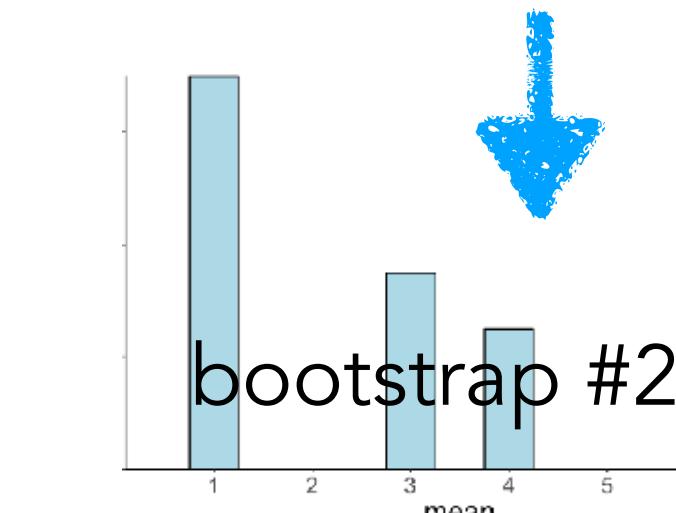
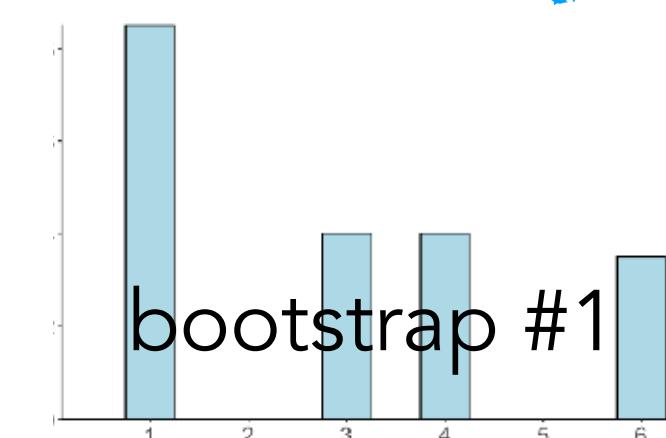
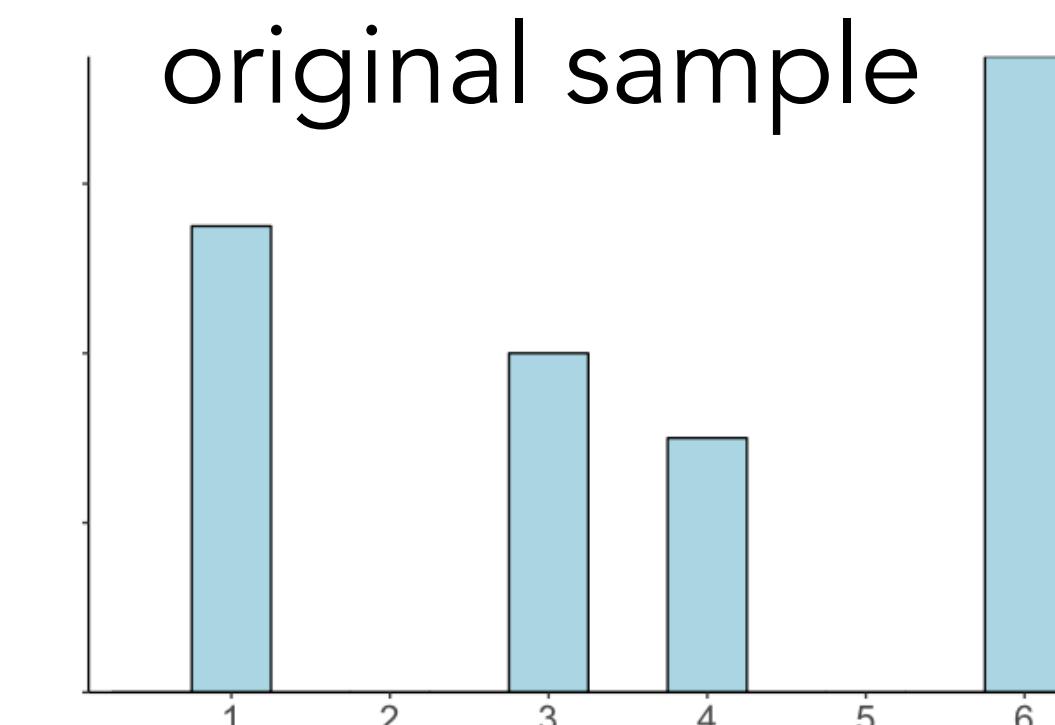


invented at Stanford



Bradley Efron

repeated  
sampling **with  
replacement**

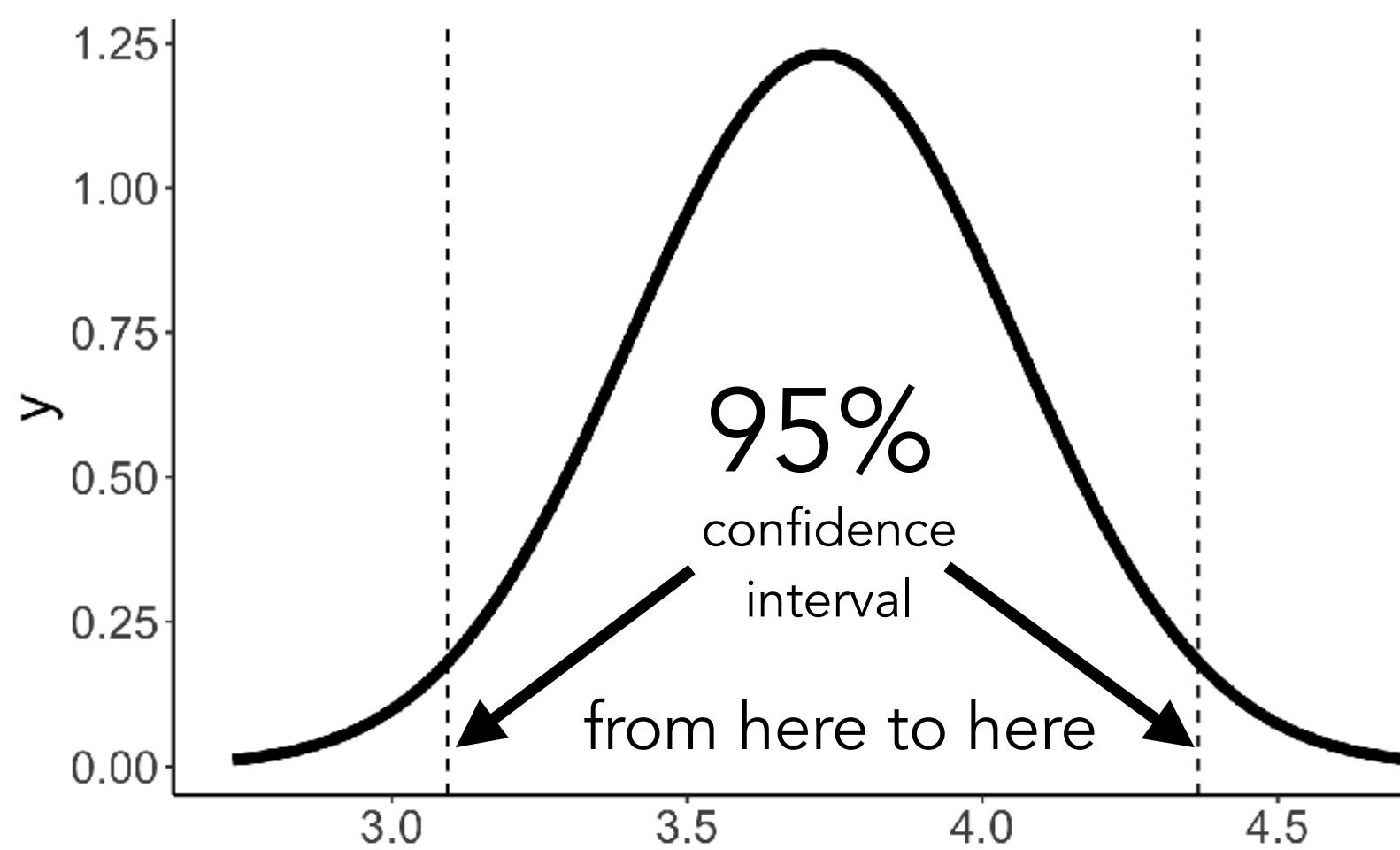


# Bootstrap

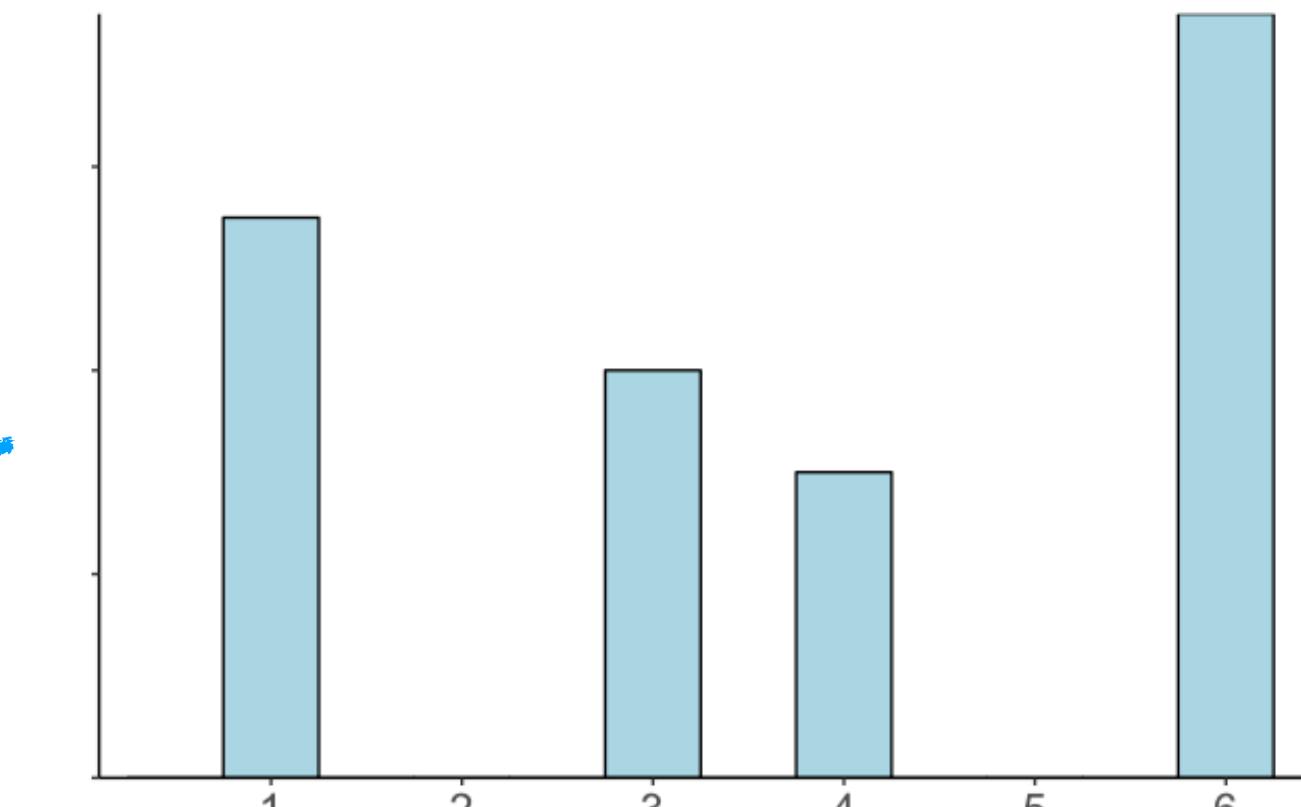
How can I get the confidence interval of a statistical estimate (such as the mean)?

make assumptions

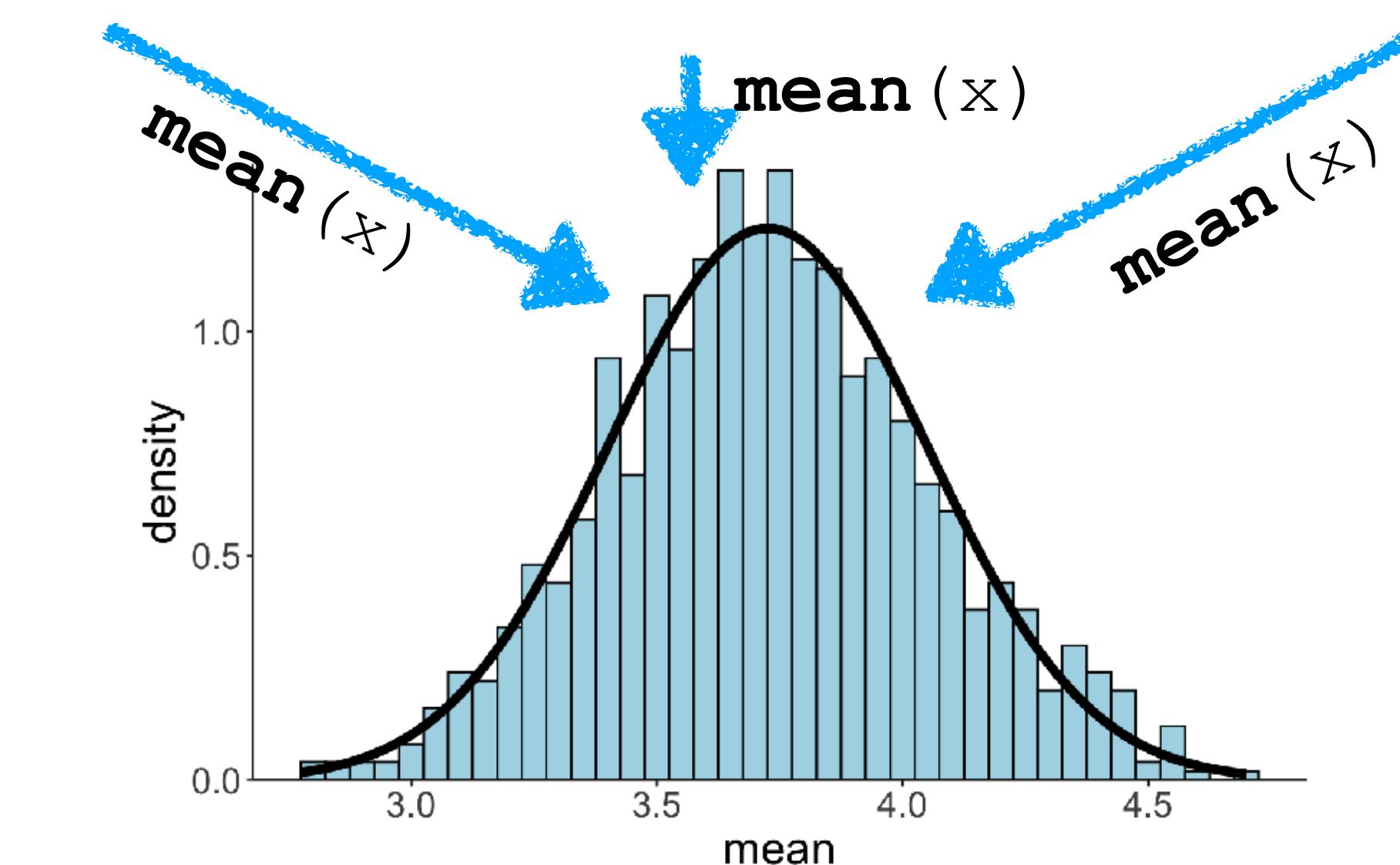
sampling distribution of the mean



$\sim \text{dnorm}(\cdot, \text{mean} = 3.73, \text{sd} = 0.324)$

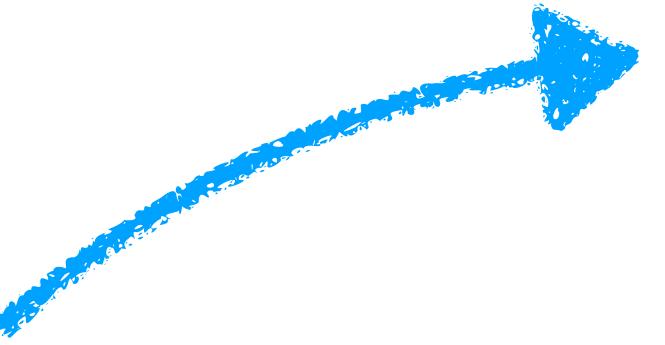


bootstrap



# mean\_cl\_boot() explained

```
1 set.seed(1)
2
3 n = 10 # sample size per group
4 k = 3 # number of groups
5
6 df.data = tibble(participant = 1:(n*k),
7                   condition = as.factor(rep(1:k, each = n))),
8                   rating = rnorm(n*k, mean = 7, sd = 1))
```

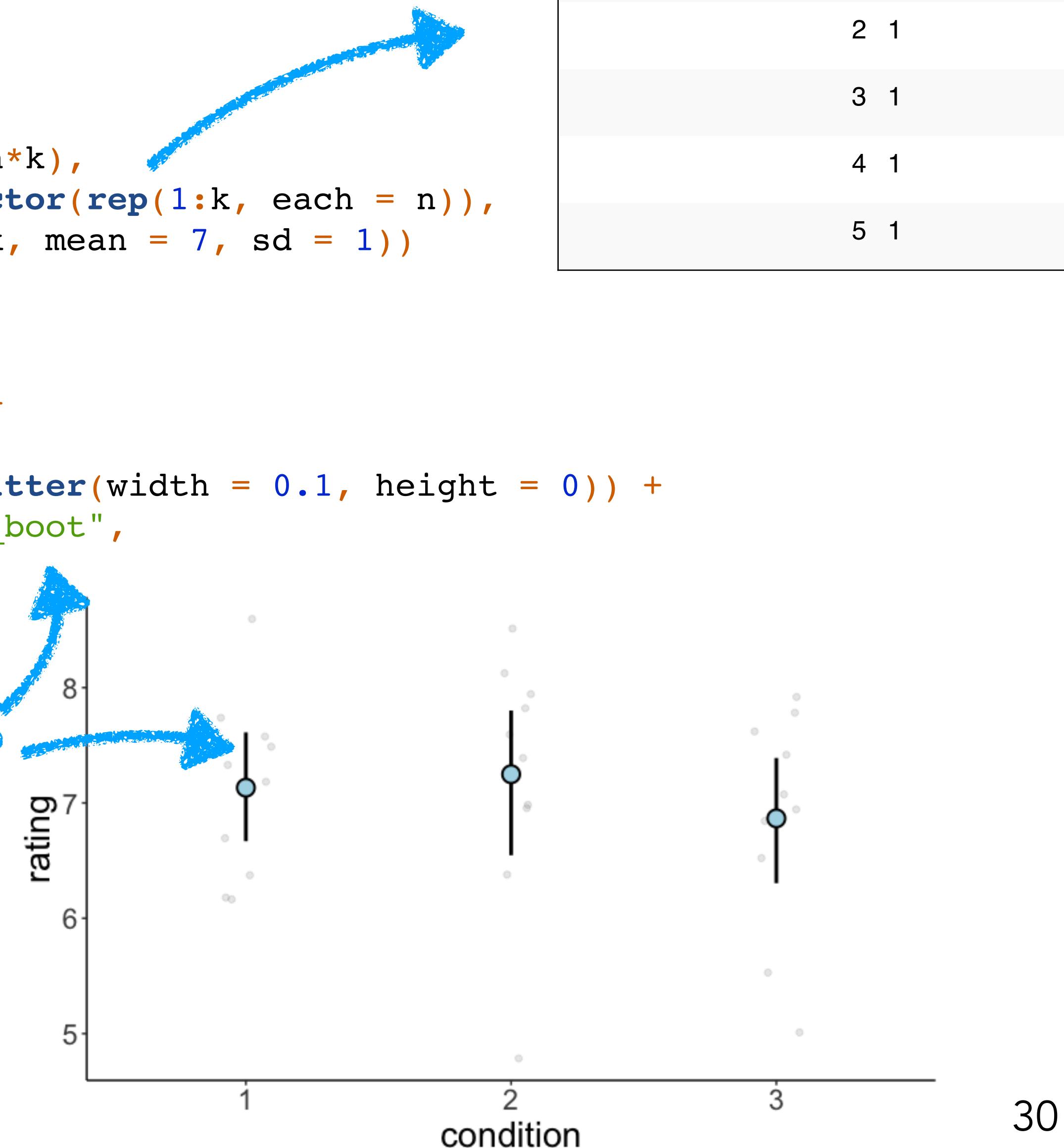


participant	condition
1	1
2	1
3	1
4	1
5	1

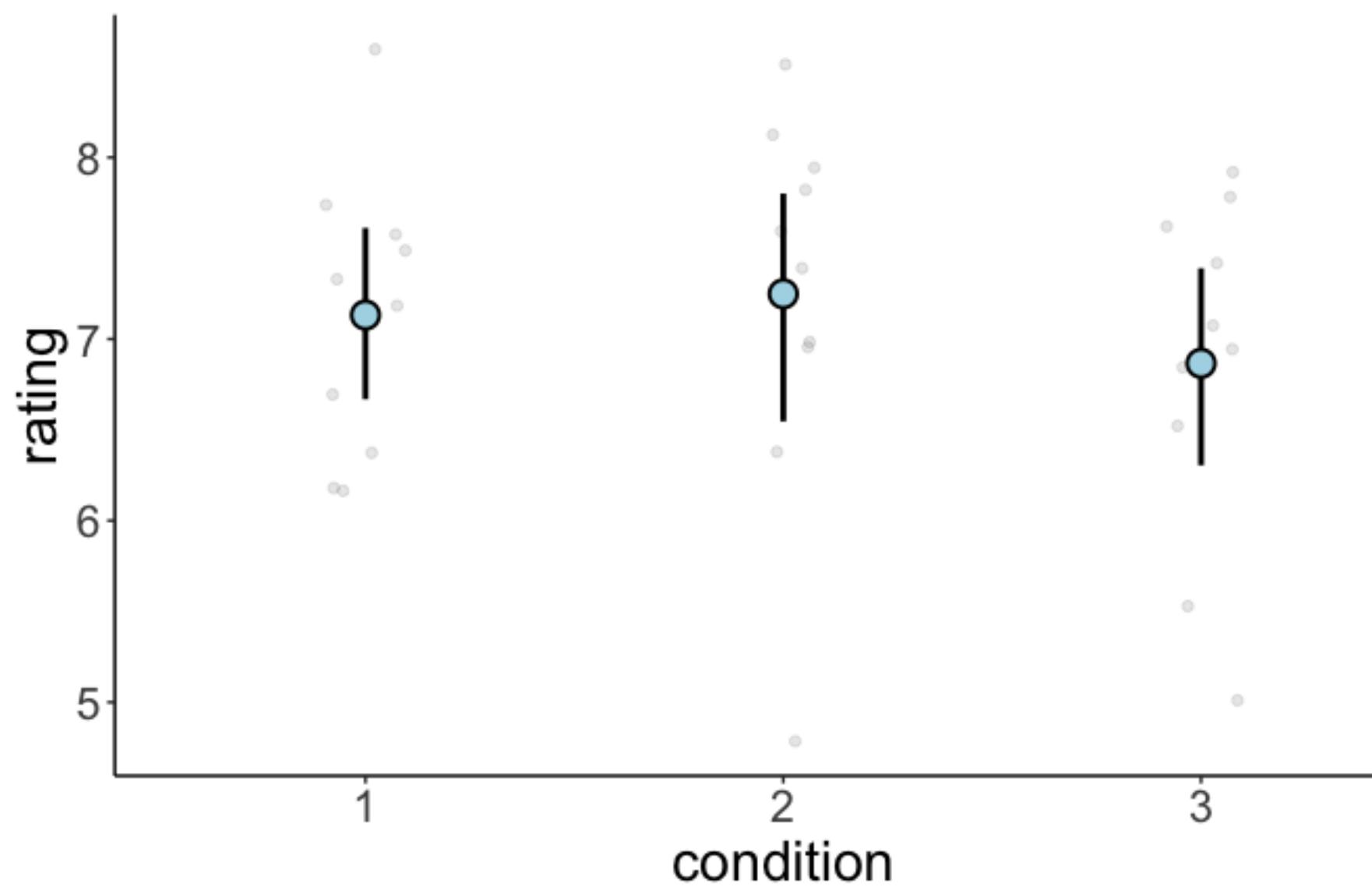
# mean\_cl\_boot() explained

```
1 set.seed(1)
2
3 n = 10 # sample size per group
4 k = 3 # number of groups
5
6 df.data = tibble(participant = 1:(n*k),
7                   condition = as.factor(rep(1:k, each = n)),
8                   rating = rnorm(n*k, mean = 7, sd = 1))
9
10 ggplot(data = df.data,
11           mapping = aes(x = condition,
12                           y = rating)) +
13   geom_point(alpha = 0.1,
14             position = position_jitter(width = 0.1, height = 0)) +
15   stat_summary(fun.data = "mean_cl_boot",
16               shape = 21,
17               size = 1,
18               fill = "lightblue")
```

what is this magic?



# mean\_cl\_boot() explained

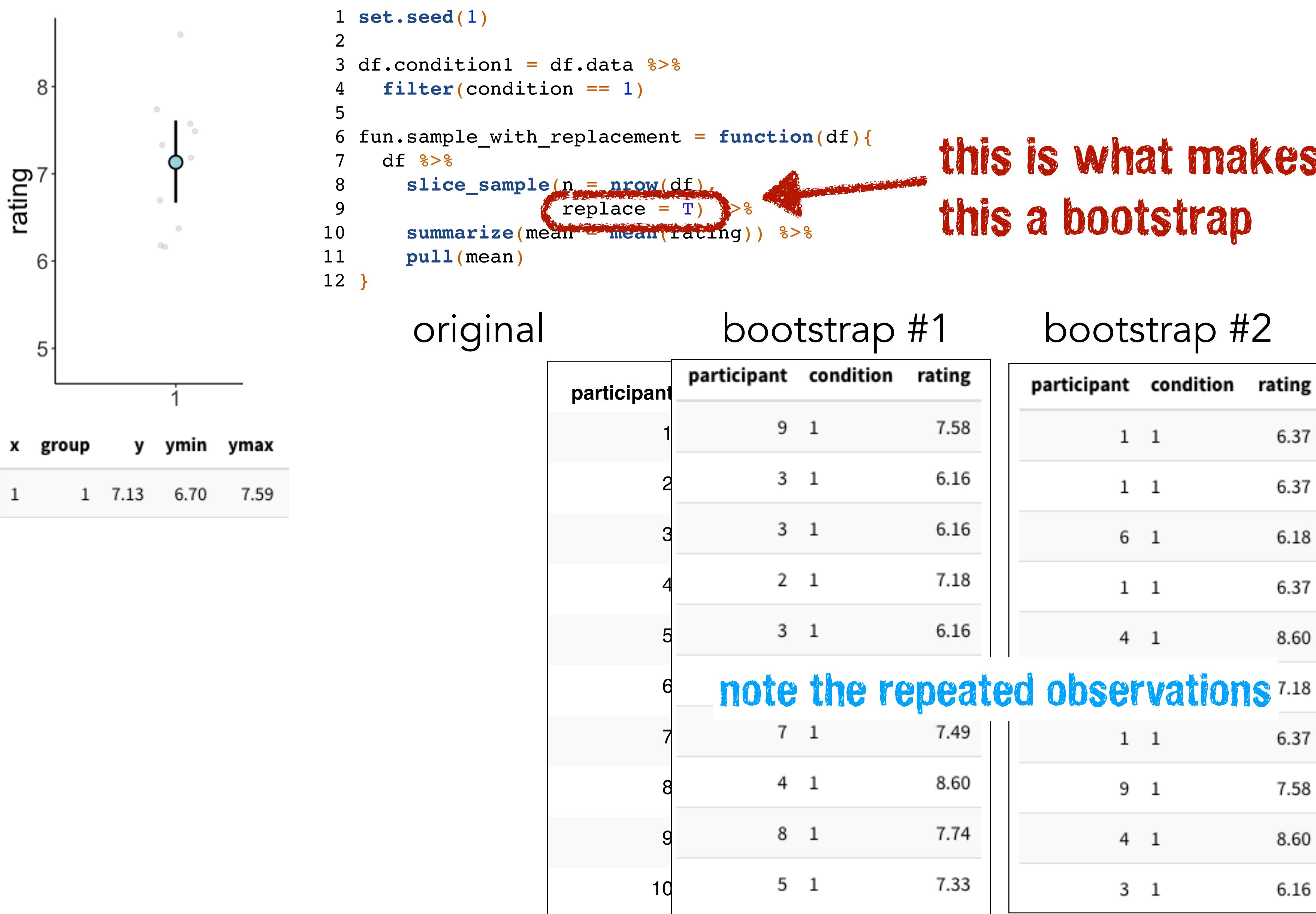


`ggplot_build(p)`

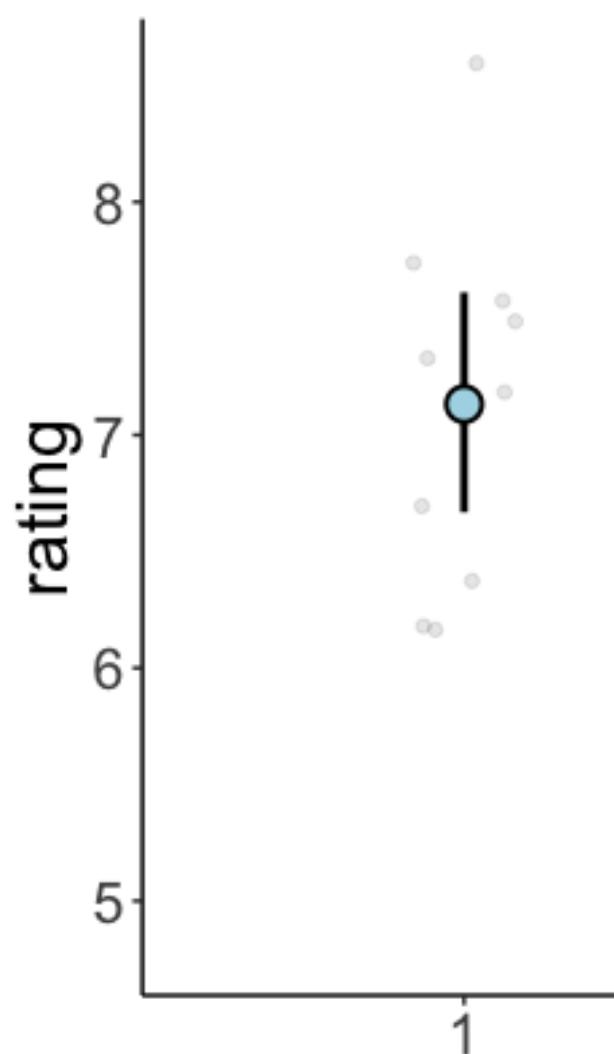
nice function for peeking  
behind the scenes of ggplots

x	group	y	ymin	ymax	PANEL	flipped_aes	colour	size	linetype	shape	fill	alpha	stroke
1	1	7.13	6.70	7.59	1	FALSE	black	1	1	21	lightblue	NA	1
2	2	7.25	6.54	7.83	1	FALSE	black	1	1	21	lightblue	NA	1
3	3	6.87	6.26	7.39	1	FALSE	black	1	1	21	lightblue	NA	1

# mean\_cl\_boot() explained



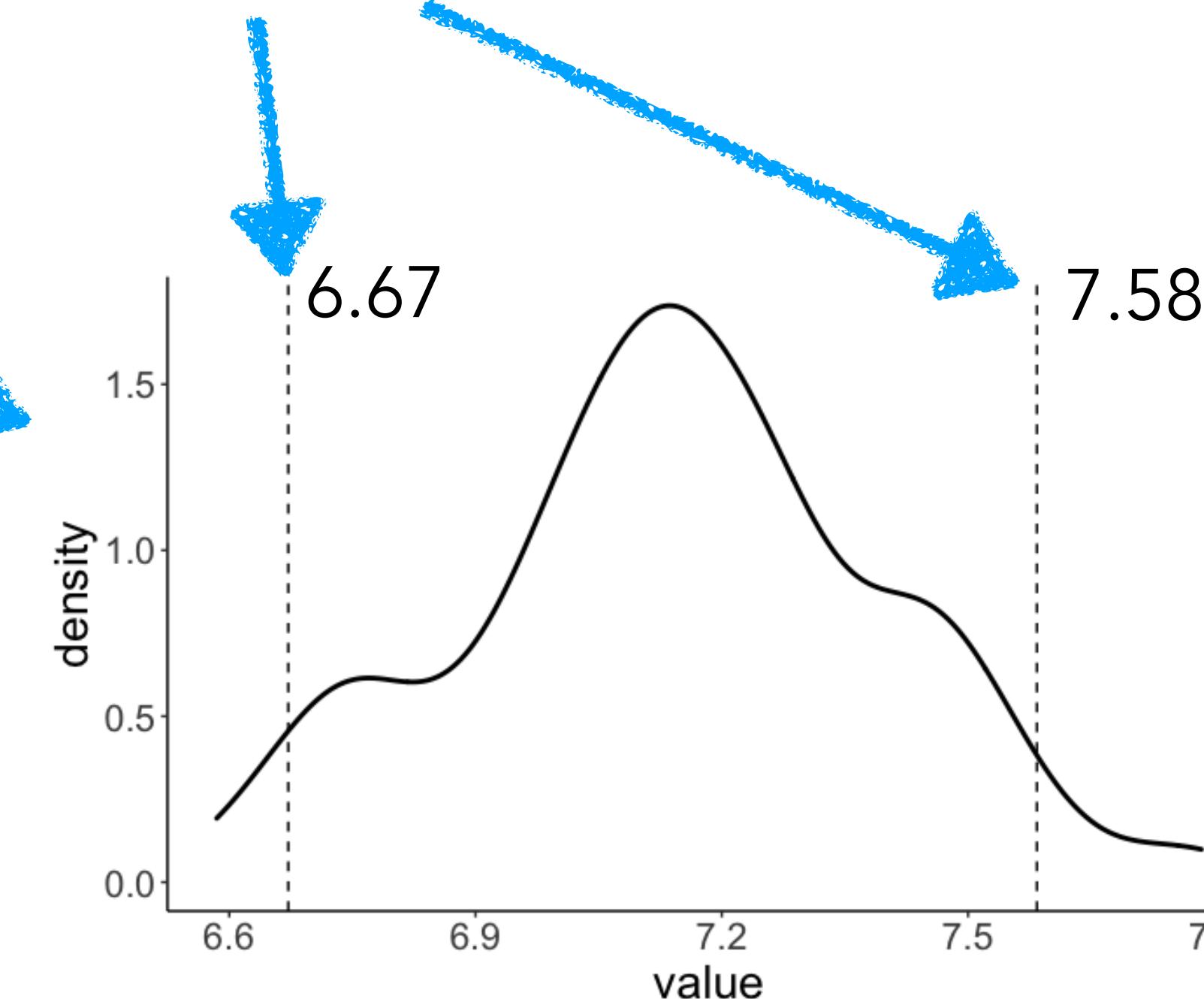
# mean\_cl\_boot() explained



```
1 set.seed(1)
2
3 df.condition1 = df.data %>%
4   filter(condition == 1)
5
6 fun.sample_with_replacement = function(df) {
7   df %>%
8     slice_sample(n = nrow(df),
9                   replace = T) %>%
10    summarize(mean = mean(rating)) %>%
11    pull(mean)
12 }
13
14 bootstraps = replicate(n = 100, fun.sample_with_replacement(df.condition1))
15
16 quantile(bootstraps, prob = c(0.025, 0.975))
```

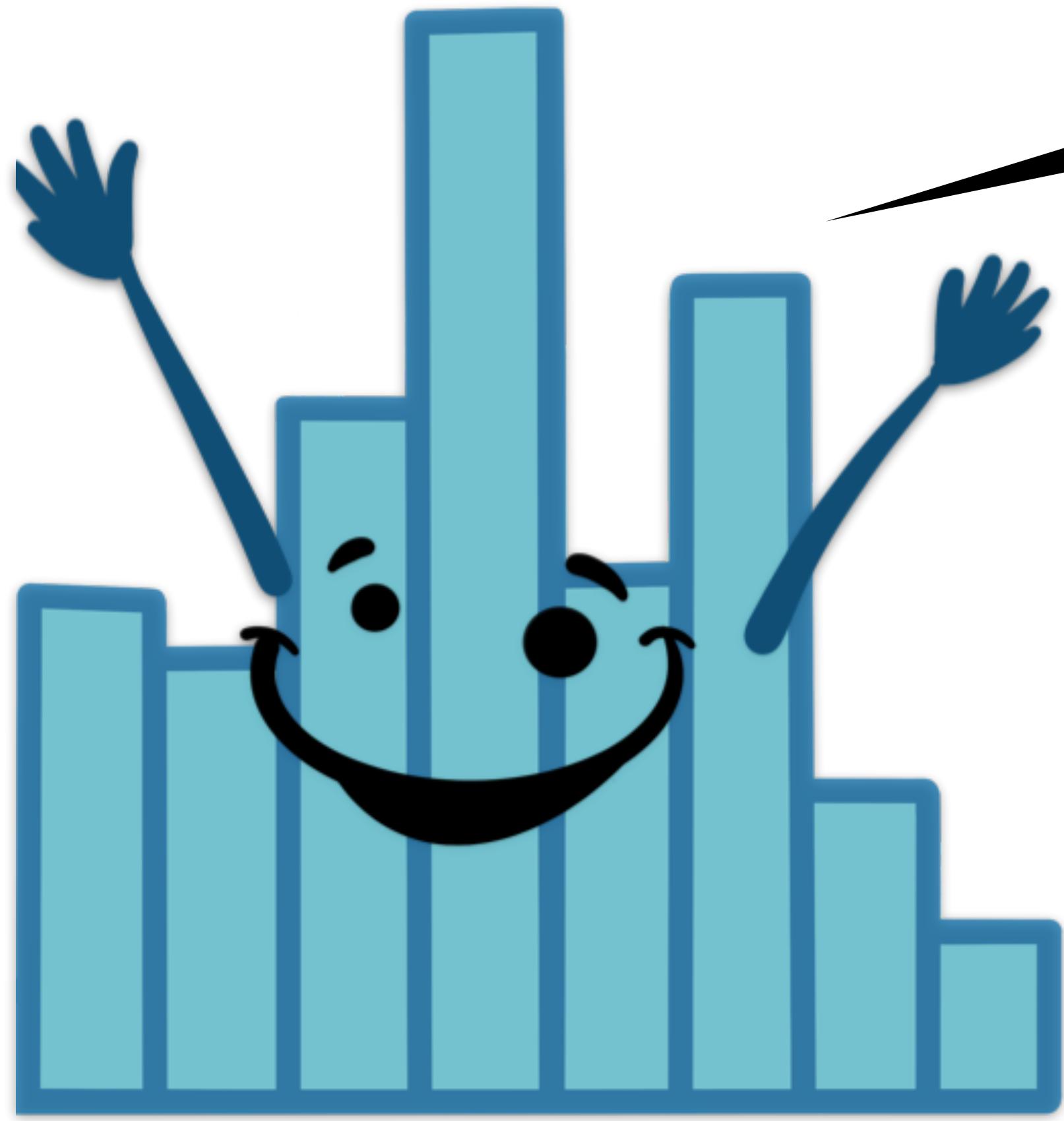
x	group	y	ymin	ymax
1	1	7.13	6.70	7.59

```
1 ggplot(data = as_tibble(bootstraps),
2         mapping = aes(x = value)) +
3   geom_density(size = 1) +
4   geom_vline(xintercept = quantile(bootstraps,
5                                     probs = c(0.025, 0.975)),
6             linetype = 2)
```



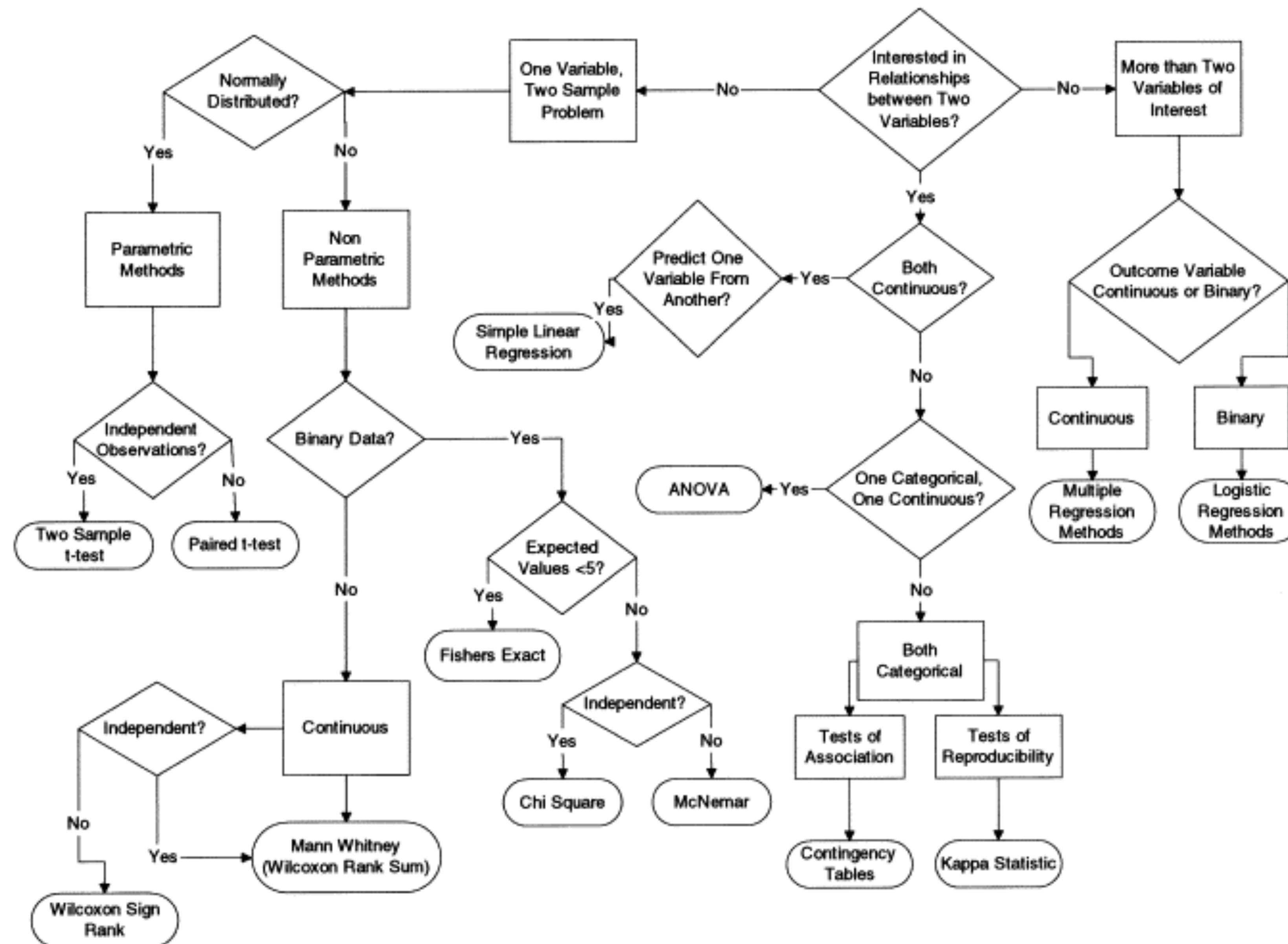
02:00

stretch break!



# **Cookbook vs. Model Comparison**

# The cookbook approach (decision trees)

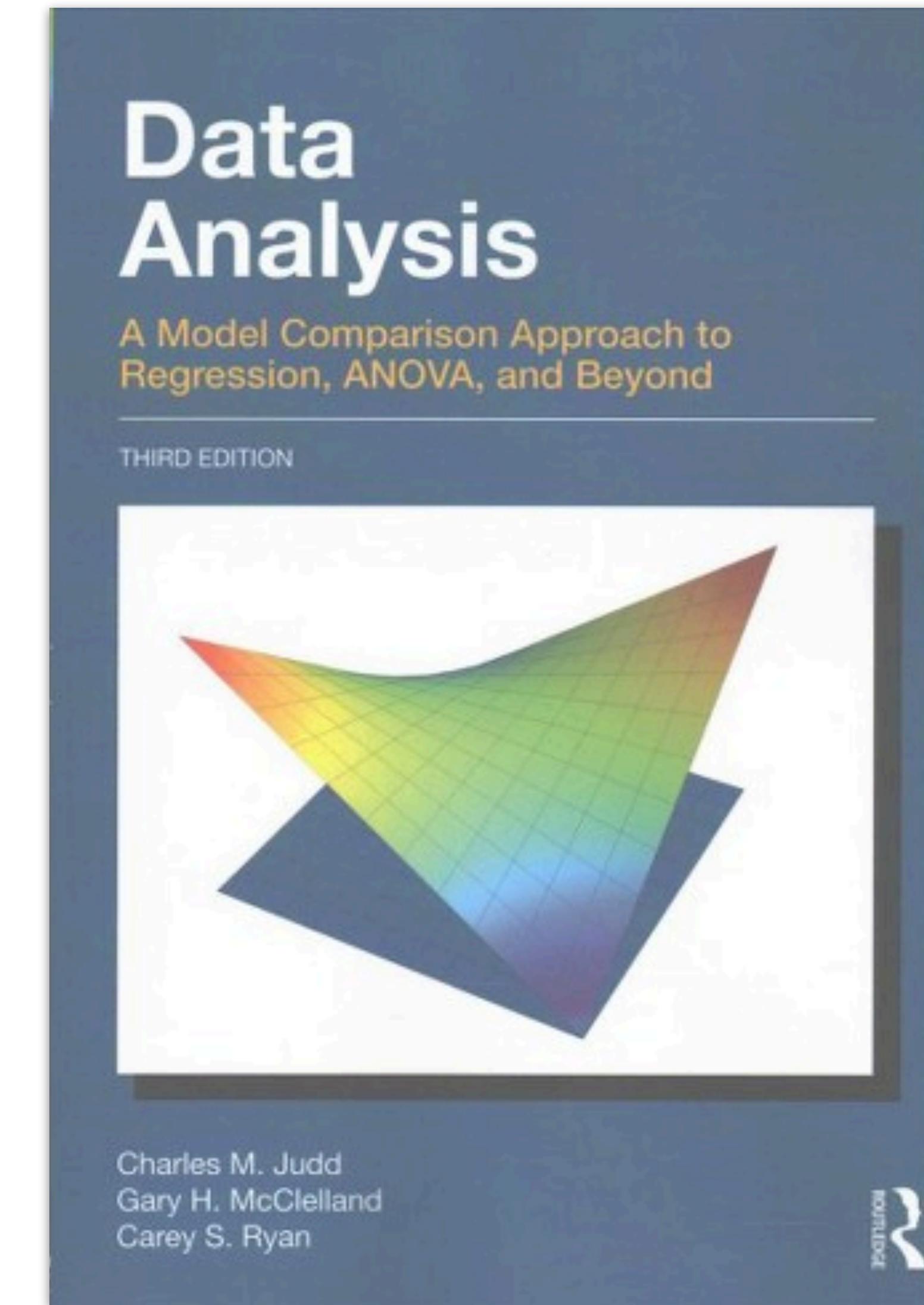
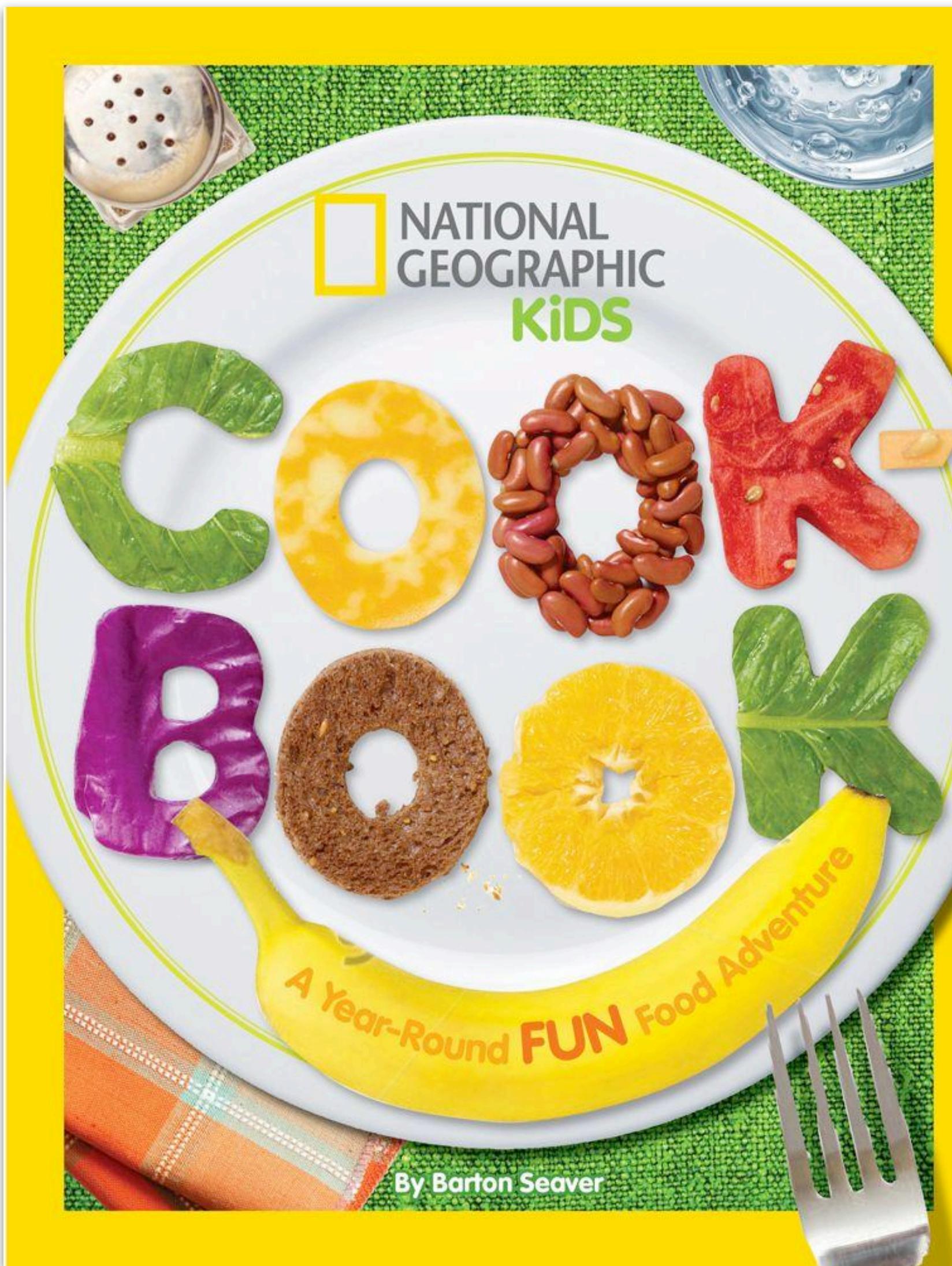


# The cookbook approach



- many statistics textbooks are organized in this way
- works reasonably well if what we want to cook is in the book
- leaves us to infer what to do if we need a new recipe

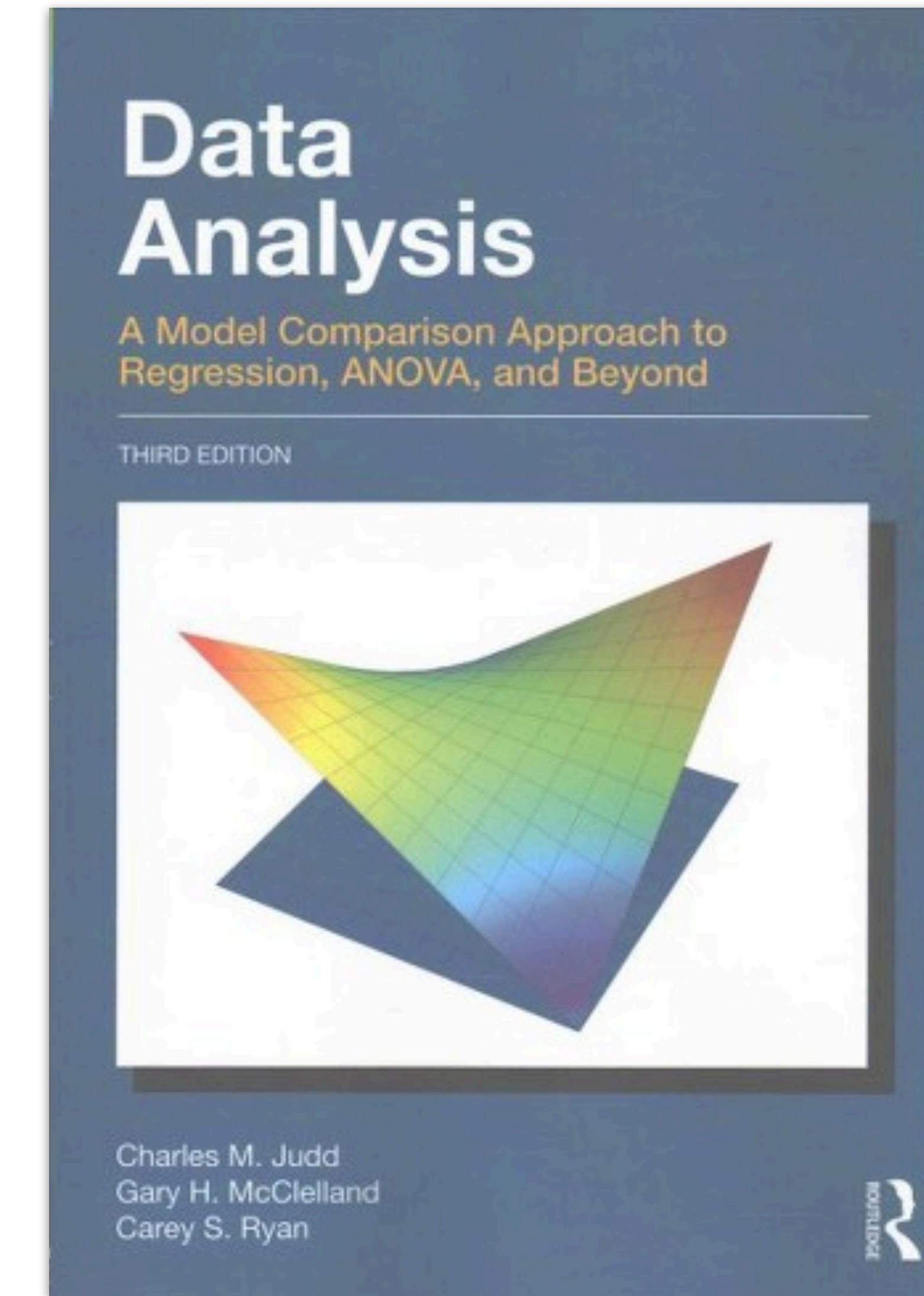
# Model comparison approach



Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

# Model comparison approach

- Flexibility in approach
- helps generate insight into phenomena
- thinking of statistical analysis as modeling
- allows for a smoother transition into Bayesian data analysis, and probabilistic modeling more generally



Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

# Modeling data

$$\text{Data} = \text{Model} + \text{Error}$$



what's a good  
model?



how shall we  
define this?

= residual: the part that's left over  
after we have used the model to  
predict/explain the data



$$\text{Error} = \text{Data} - \text{Model}$$

to reduce  
error we can:

improve the quality  
of the data

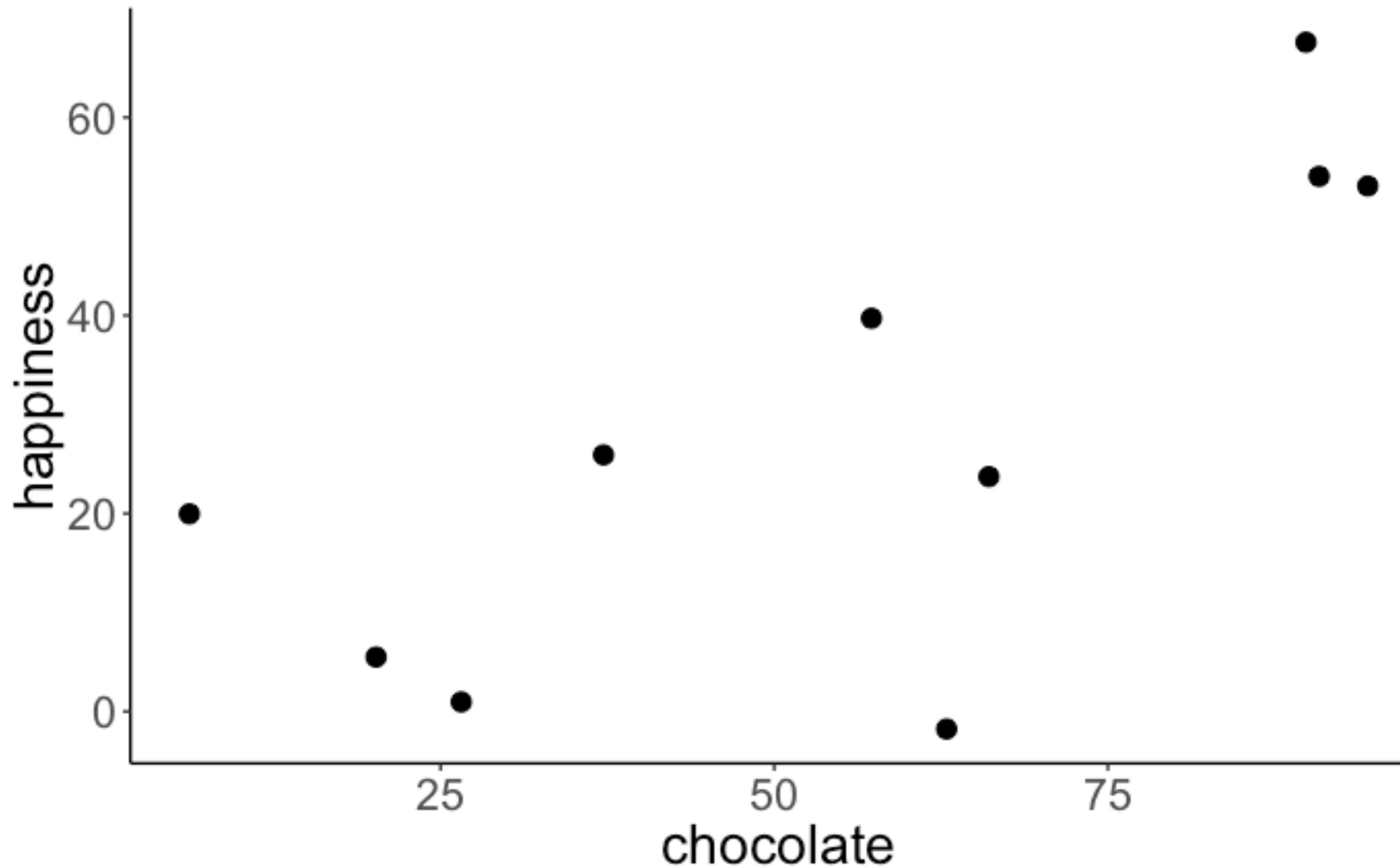
e.g. run good  
experiments



improve the model

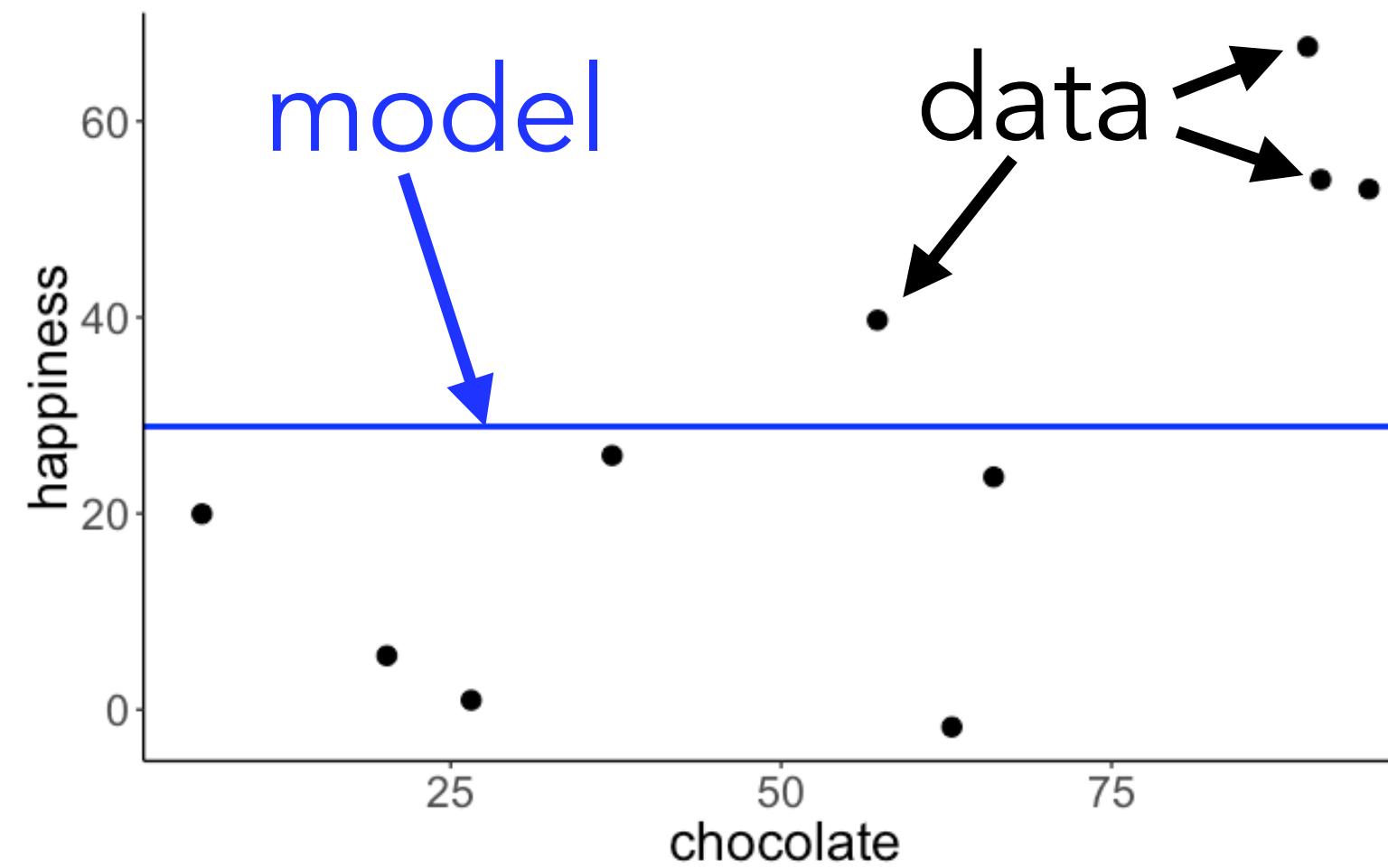
e.g. make predictions  
conditional on  
additional information

# Is there a relationship between chocolate consumption and happiness?

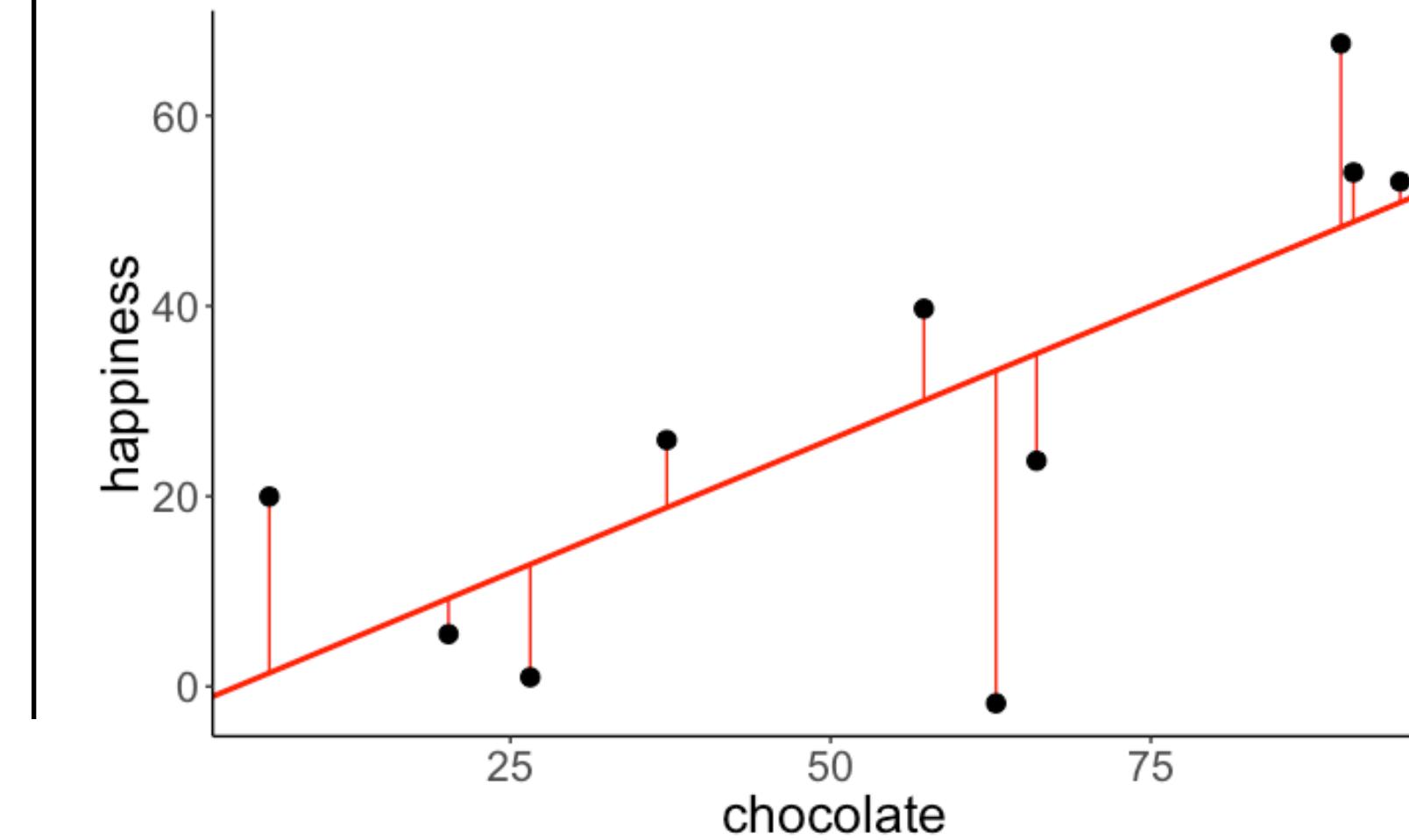
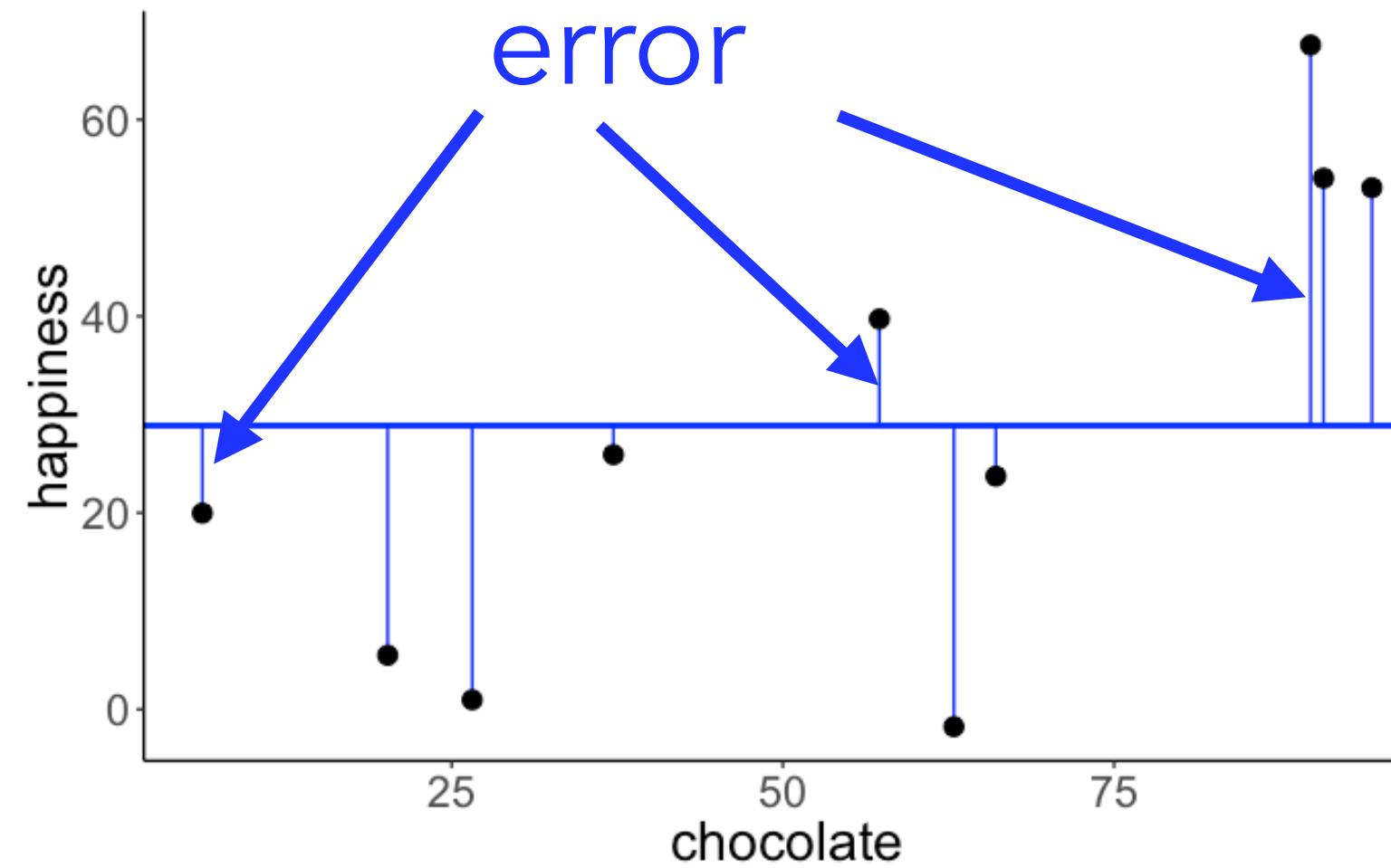
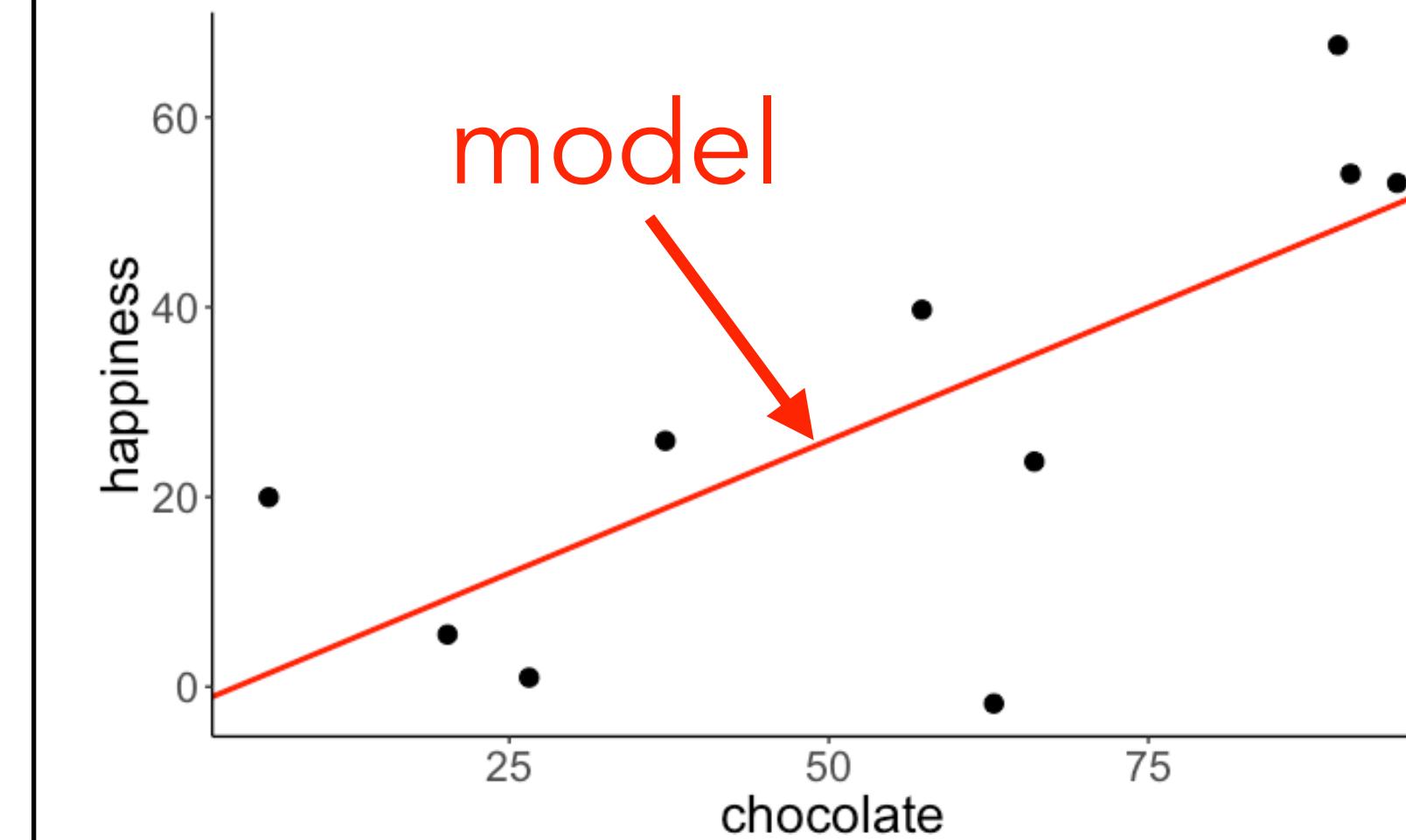


# Data = Model + Error

$H_0$ : Chocolate consumption and happiness are unrelated.



$H_1$ : Chocolate consumption and happiness are related.



# ERROR

$$\text{Error} = \text{Data} - \text{Model}$$



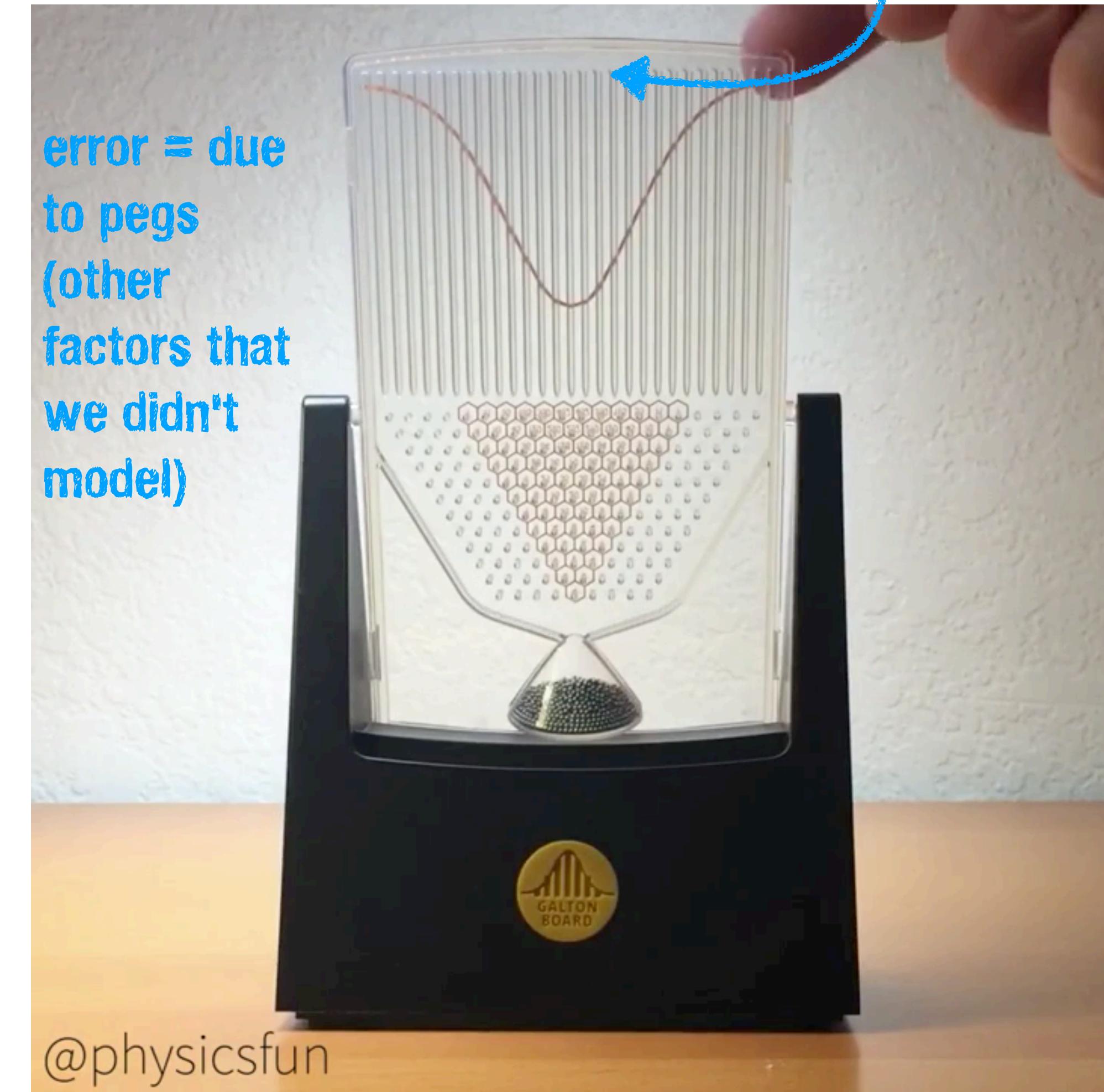
how shall we  
define this?

# ERROR

1. We assume that the error between model and data is due to (a potentially large number of) factors that we didn't take into account.
2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).

# ERROR

1. We assume that the error between model and data is due to (a potentially large number of) factors that we didn't take into account.
2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).



model = "bottleneck"

error = due  
to pegs  
(other  
factors that  
we didn't  
model)

data = where the balls land

**Result:** normal distribution

# ERROR

$$\text{Error} = \text{Data} - \text{Model}$$



how shall we  
define this?

concretely: we will fit our models such that they minimize  
the **sum of squared errors**



why squared error?

- we can sum up all the error terms  
(positive and negative prediction  
errors don't cancel out)
- larger errors are weighed more

# Assumption of normal distribution

$$\text{Error} = \text{Data} - \text{Model}$$



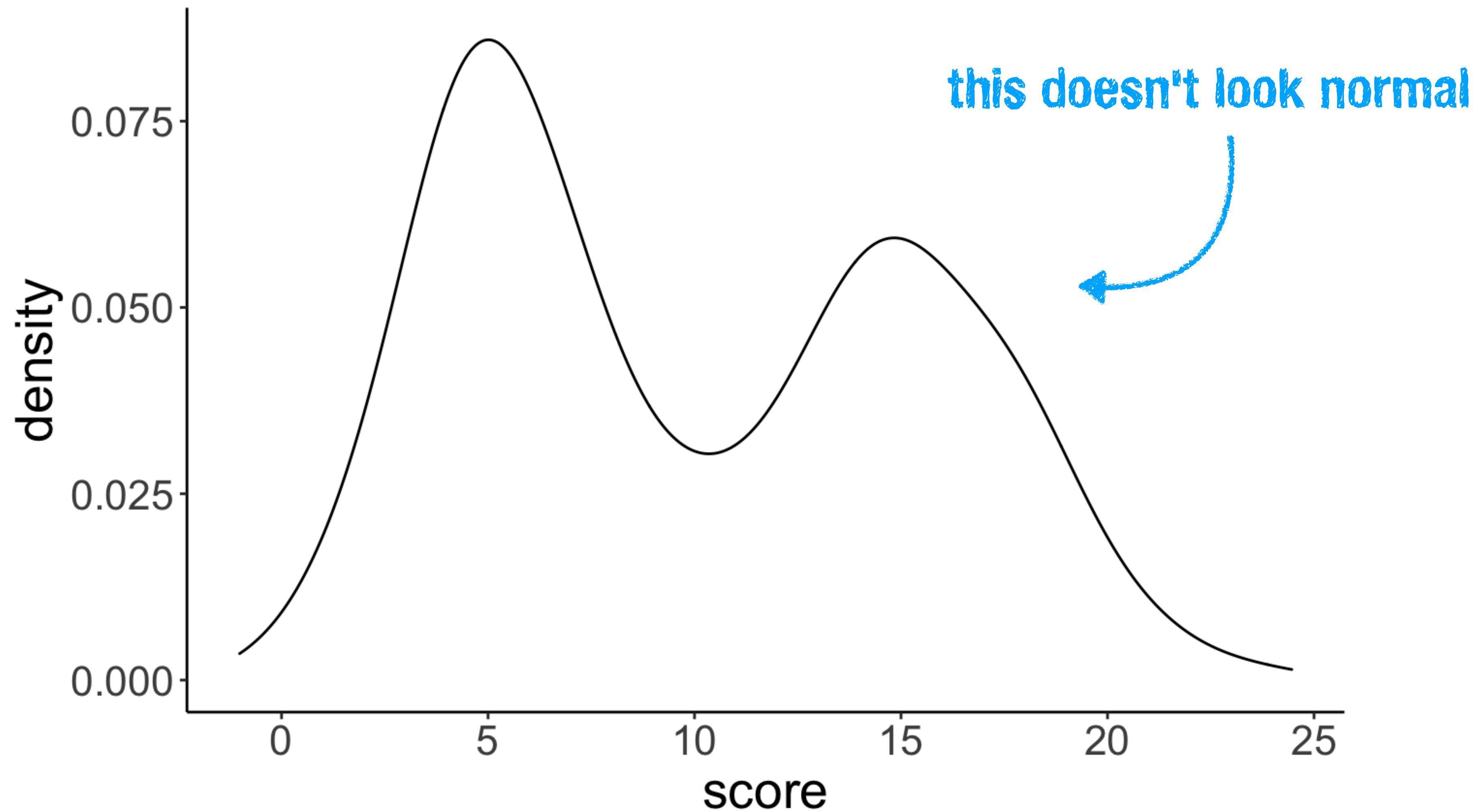
**assumed to be  
normally  
distributed**



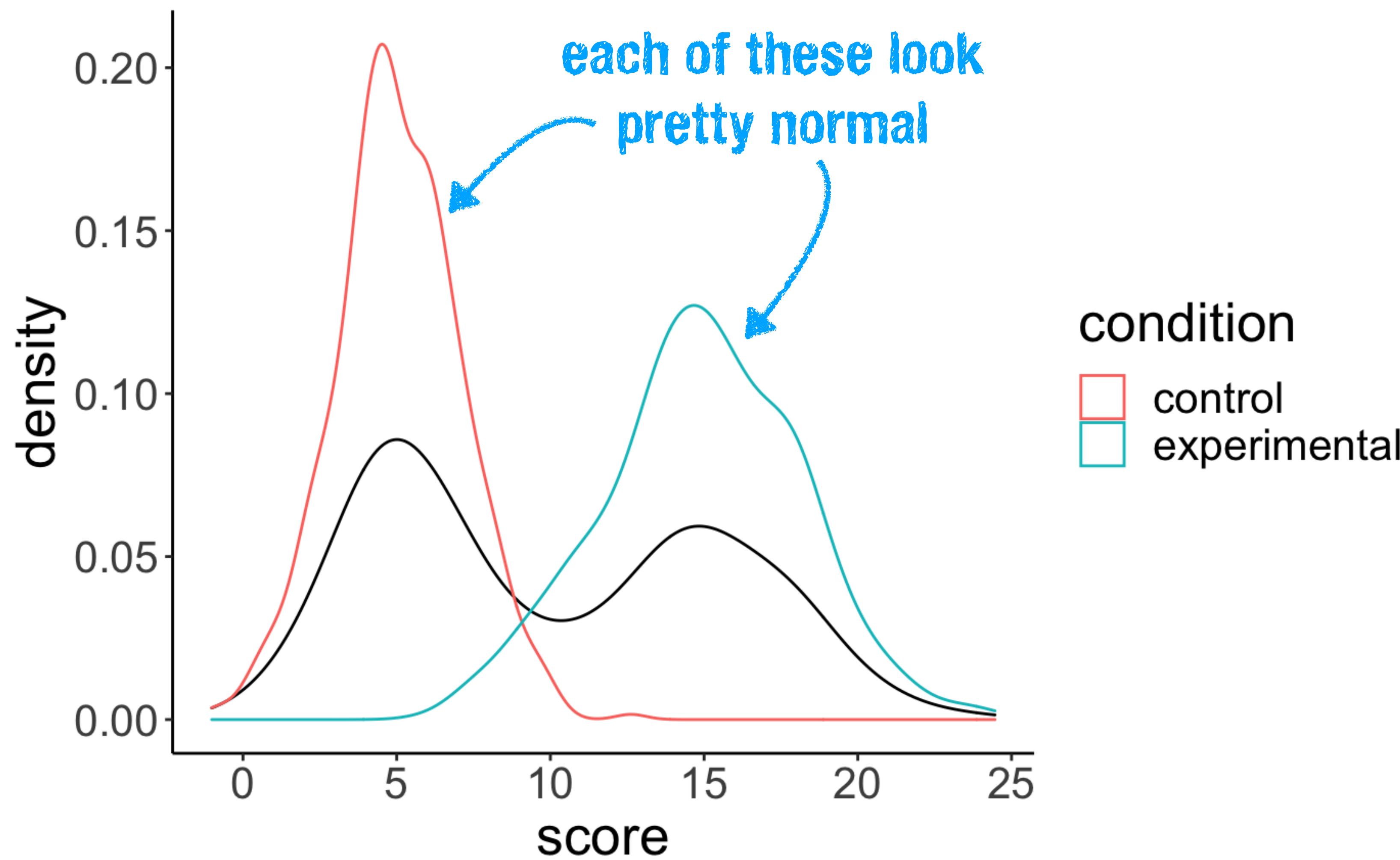
**don't need to  
be normally  
distributed!!**

very common misconception!!!

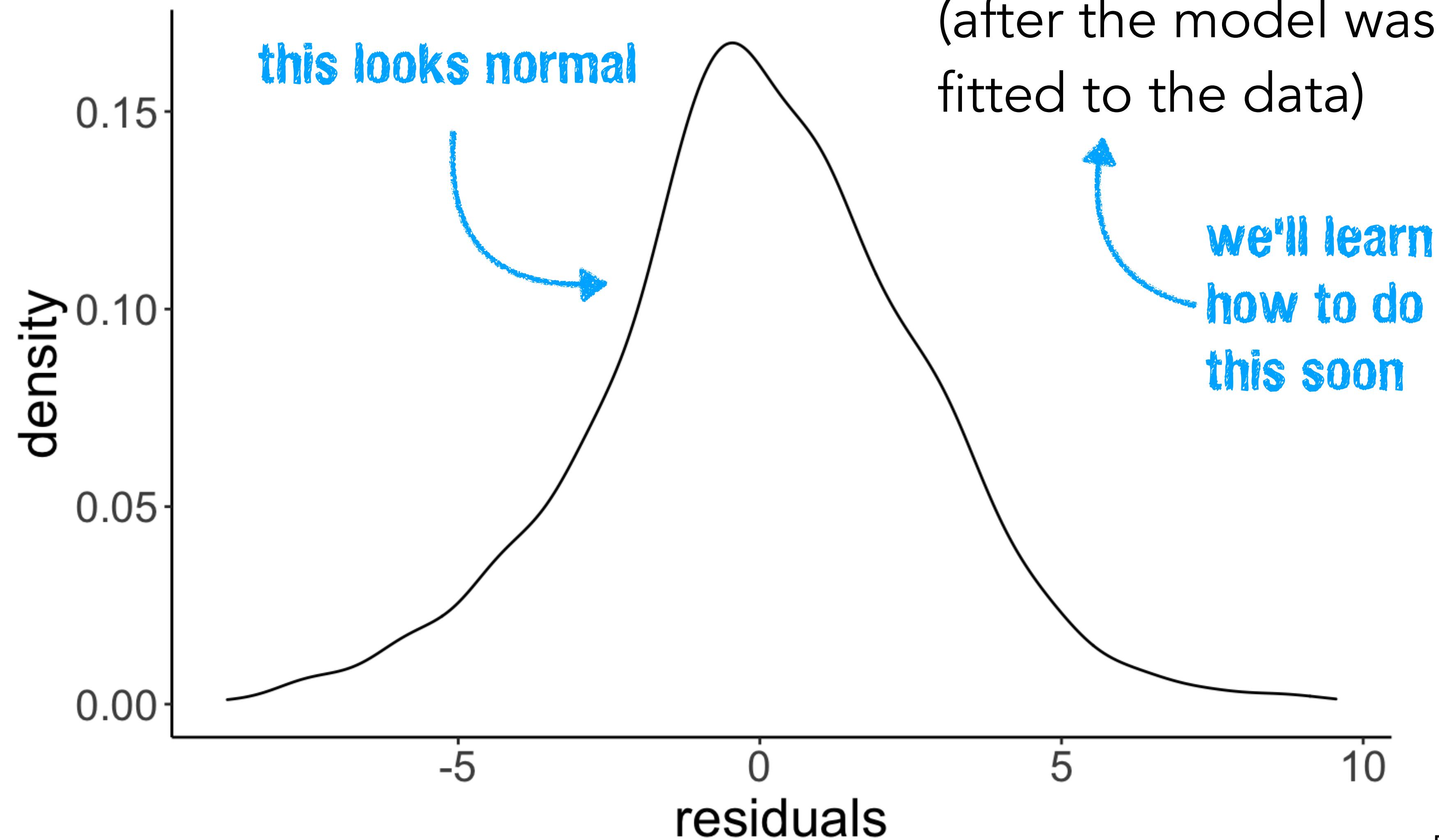
## Distribution of **test scores**



## Distribution of test scores



## Distribution of the **residuals**



Data = Model + Error

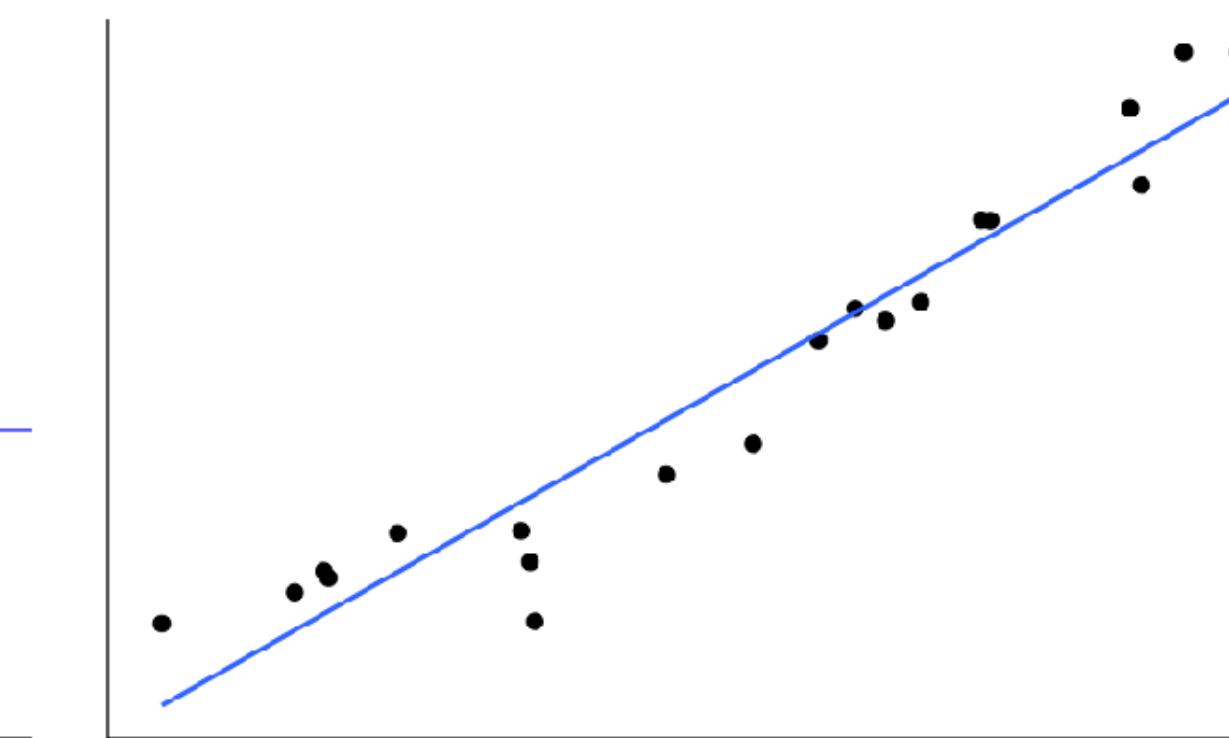
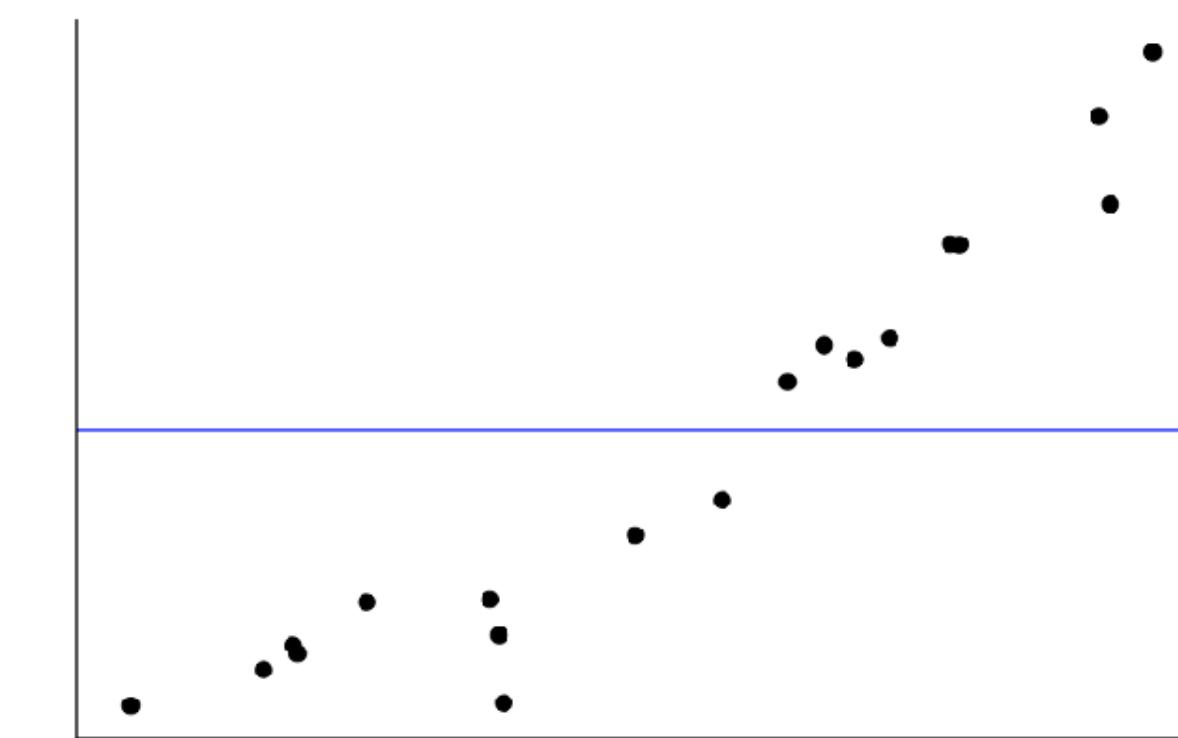
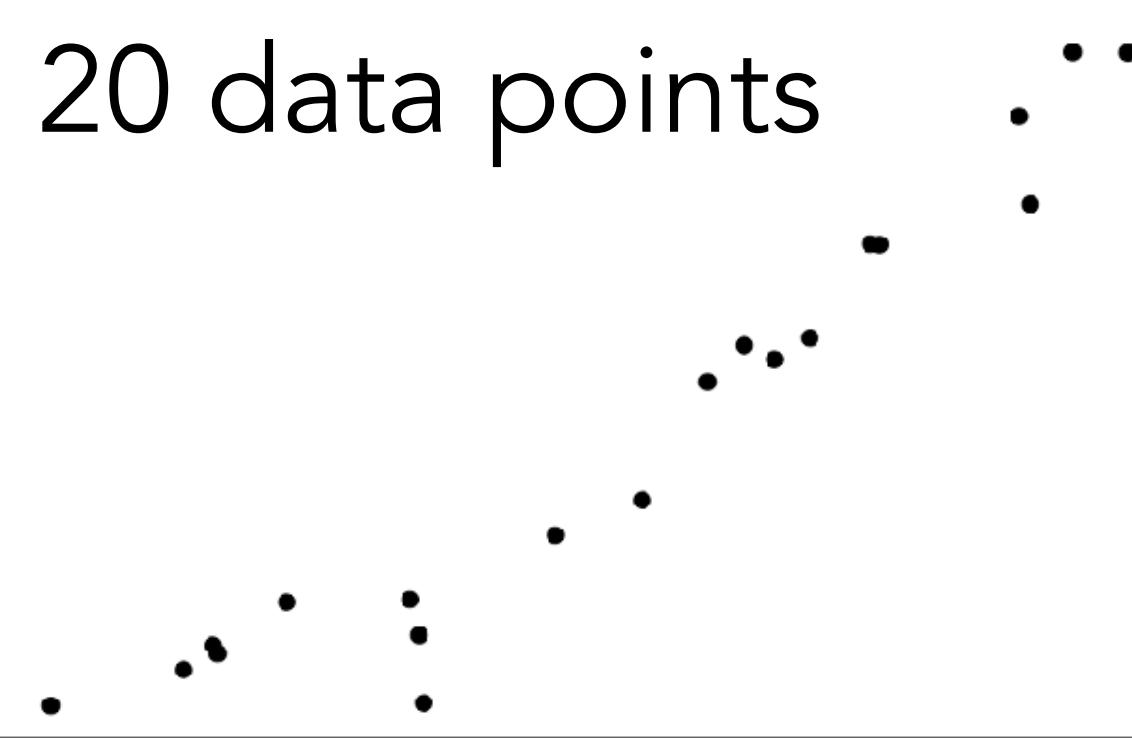


what makes for  
a good model?

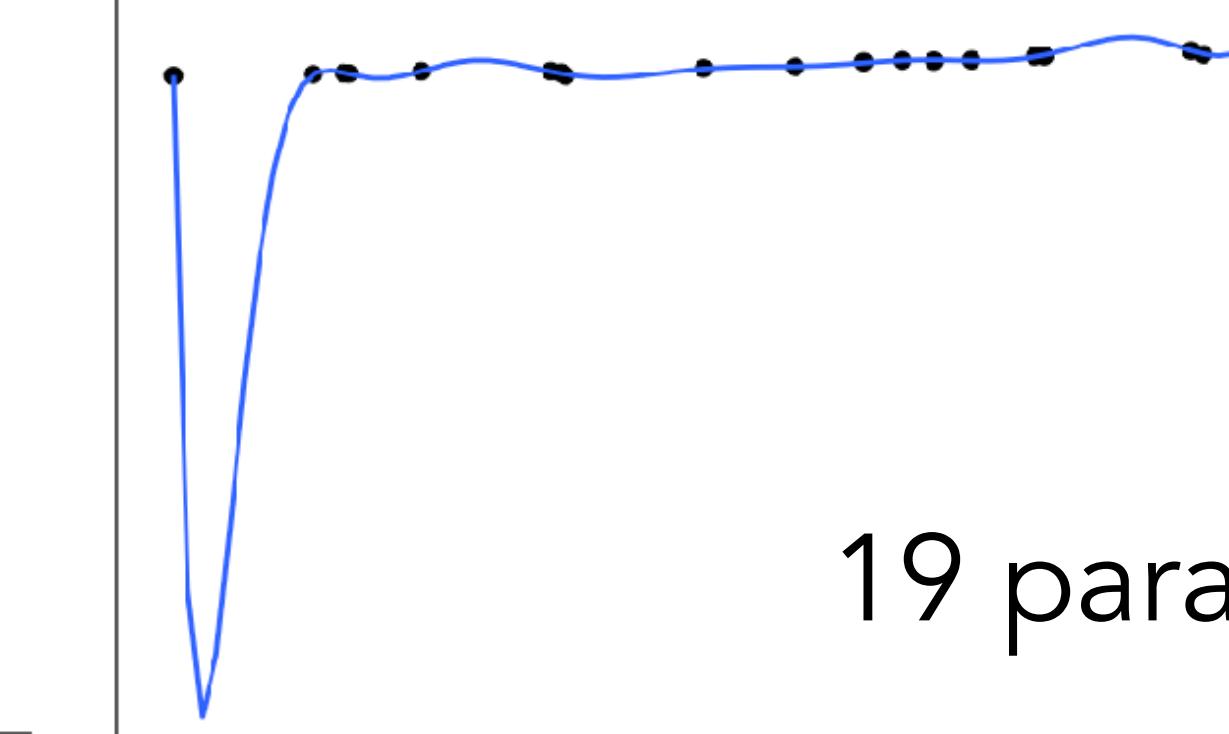
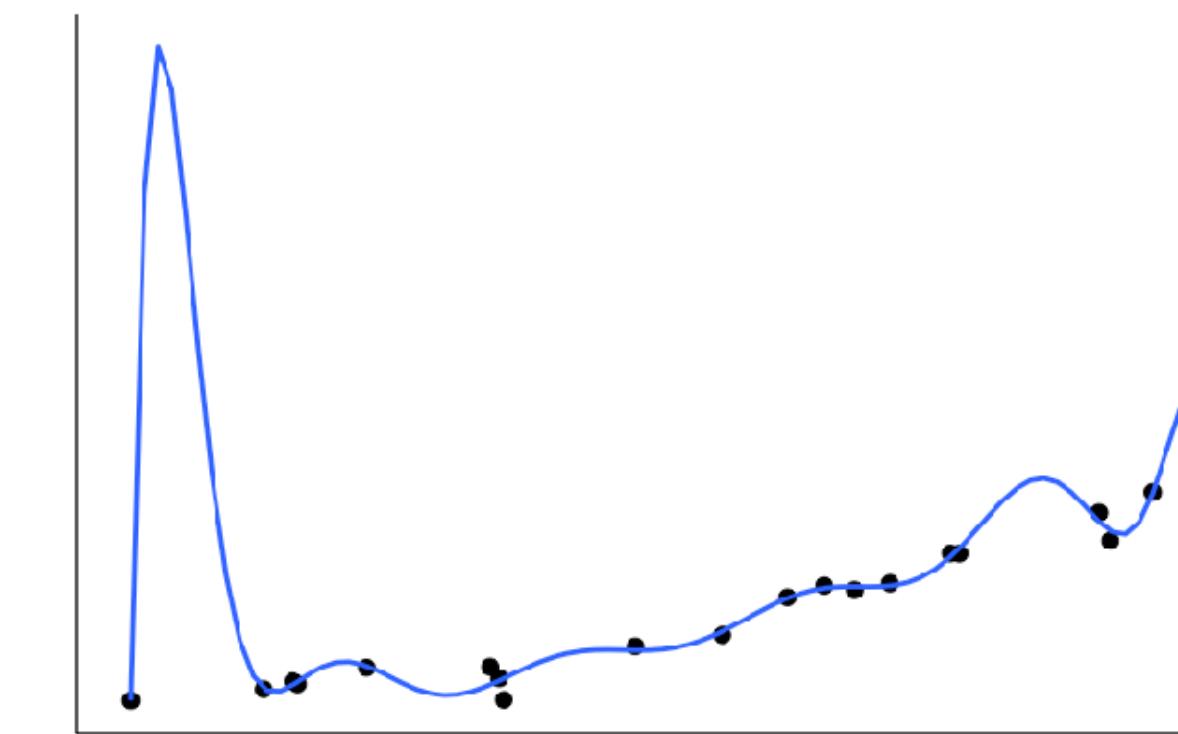
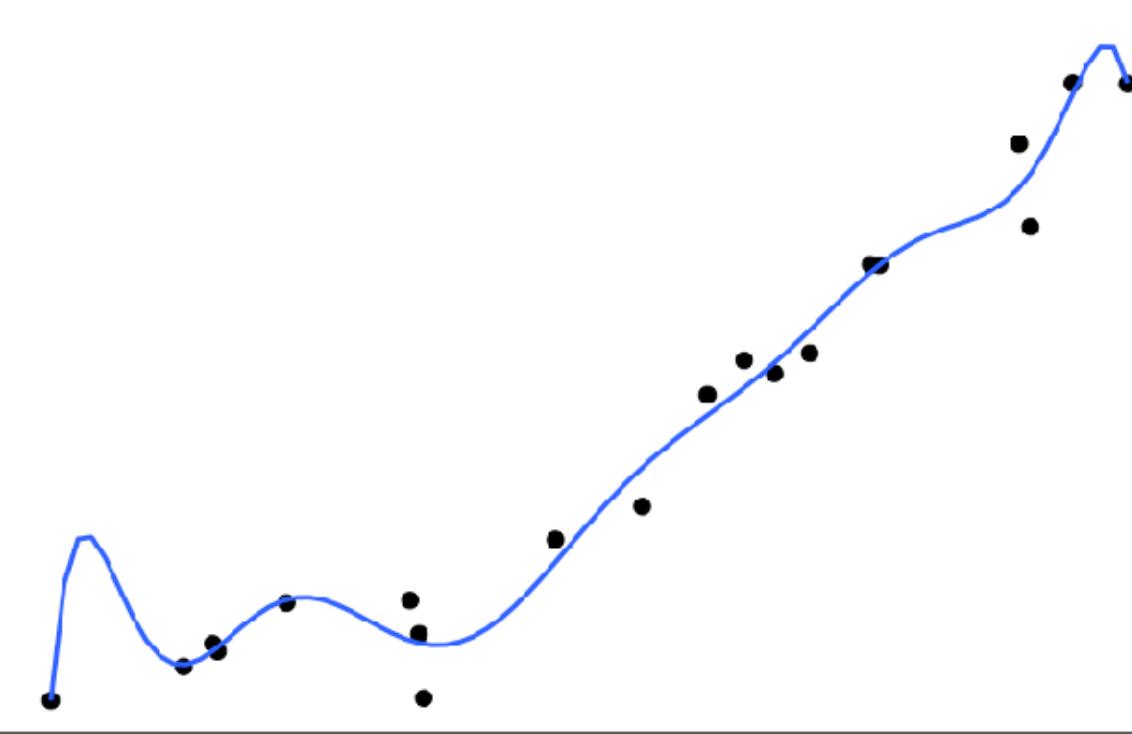
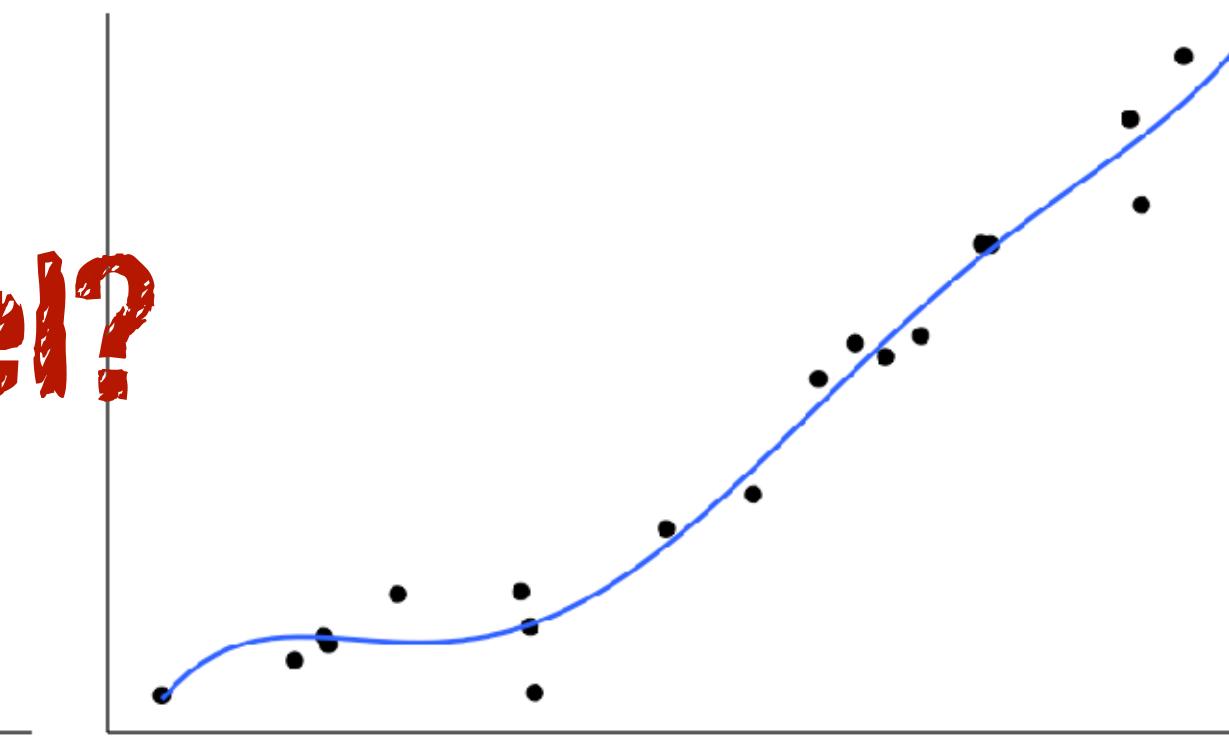
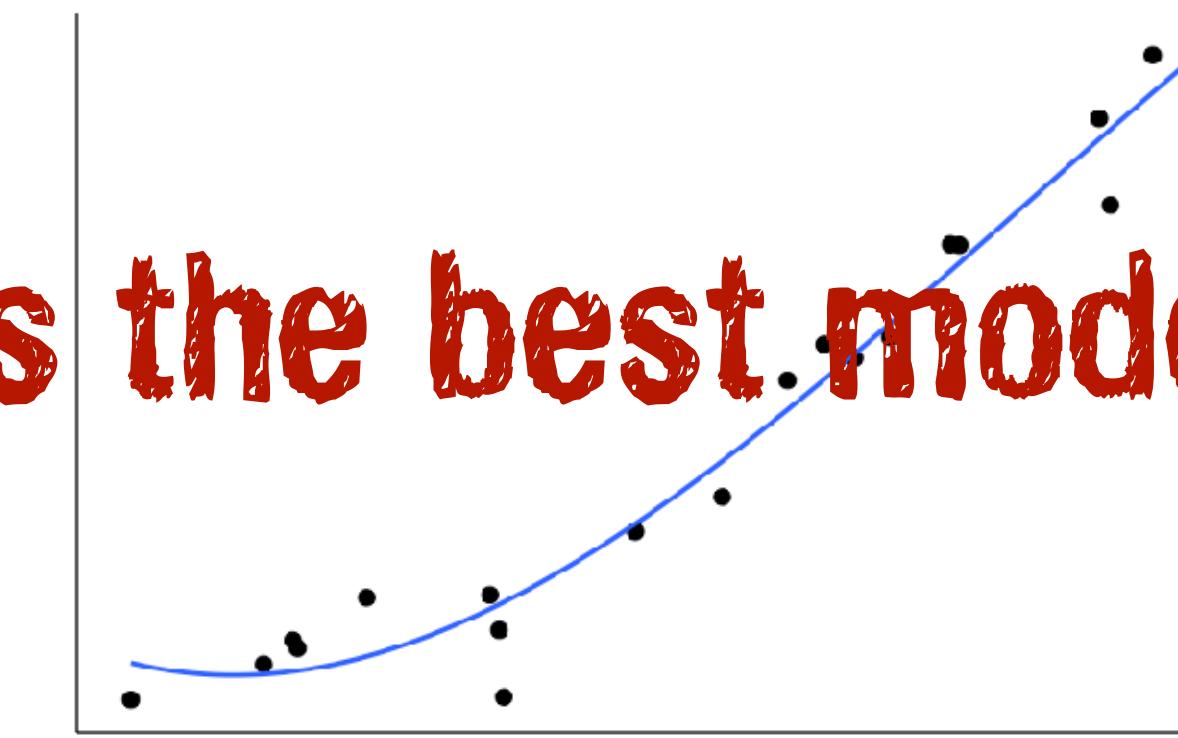
- we build models with parameters, and estimate those parameters to **minimize error**
- adding additional parameters to the model will always improve the model fit and reduce error
- fundamental trade-off between **simplicity** and **accuracy**

# Hypothesis testing as model comparison

20 data points



Which is the best model?



19 parameters

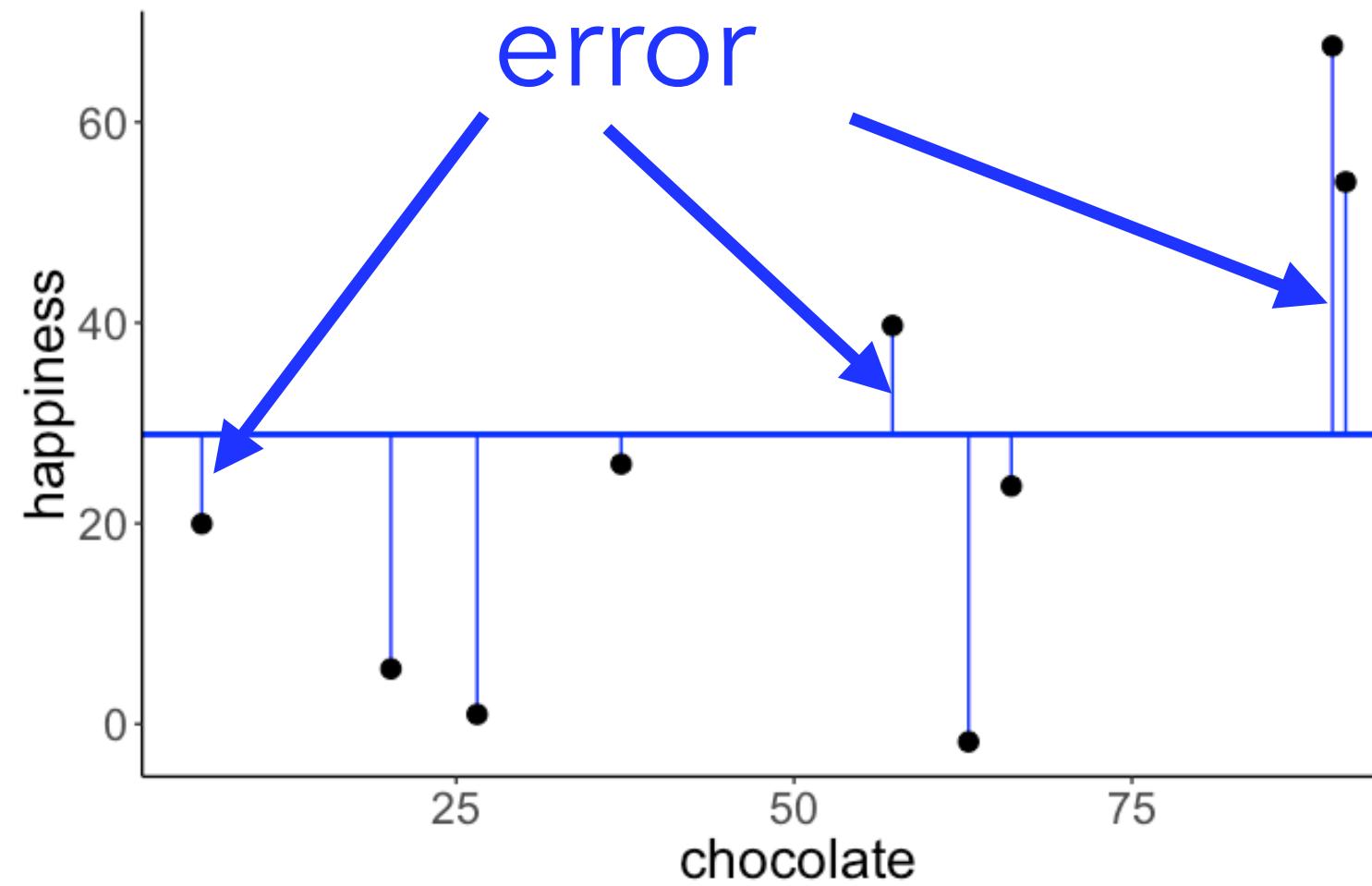
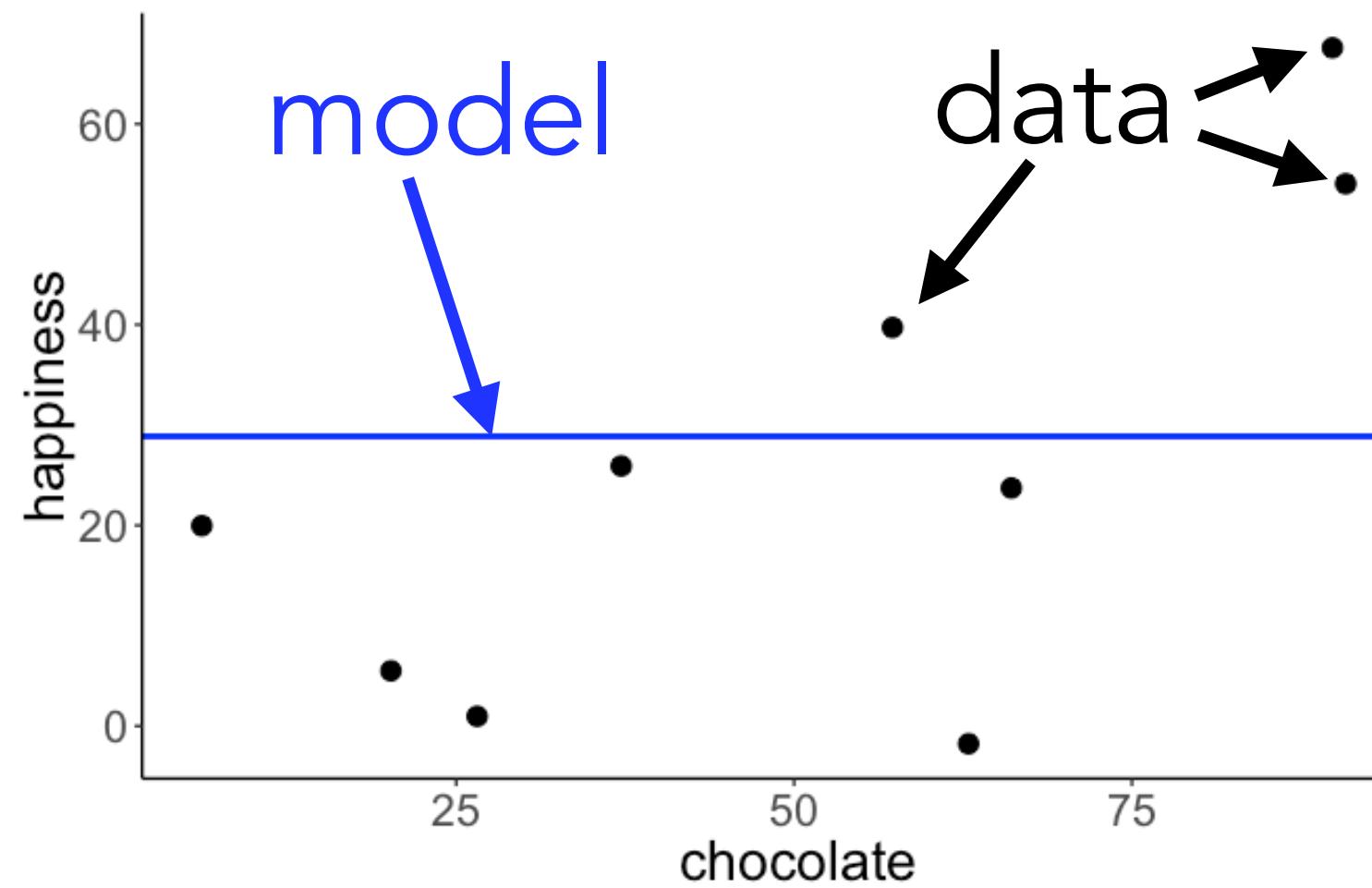
# General Procedure

1. Start with research question
2. Formulate hypothesis as a comparison between compact and augmented model
3. Estimate parameters in each model from data
4. Calculate the proportional reduction of error (PRE)
5. Decide whether PRE is **worth it**

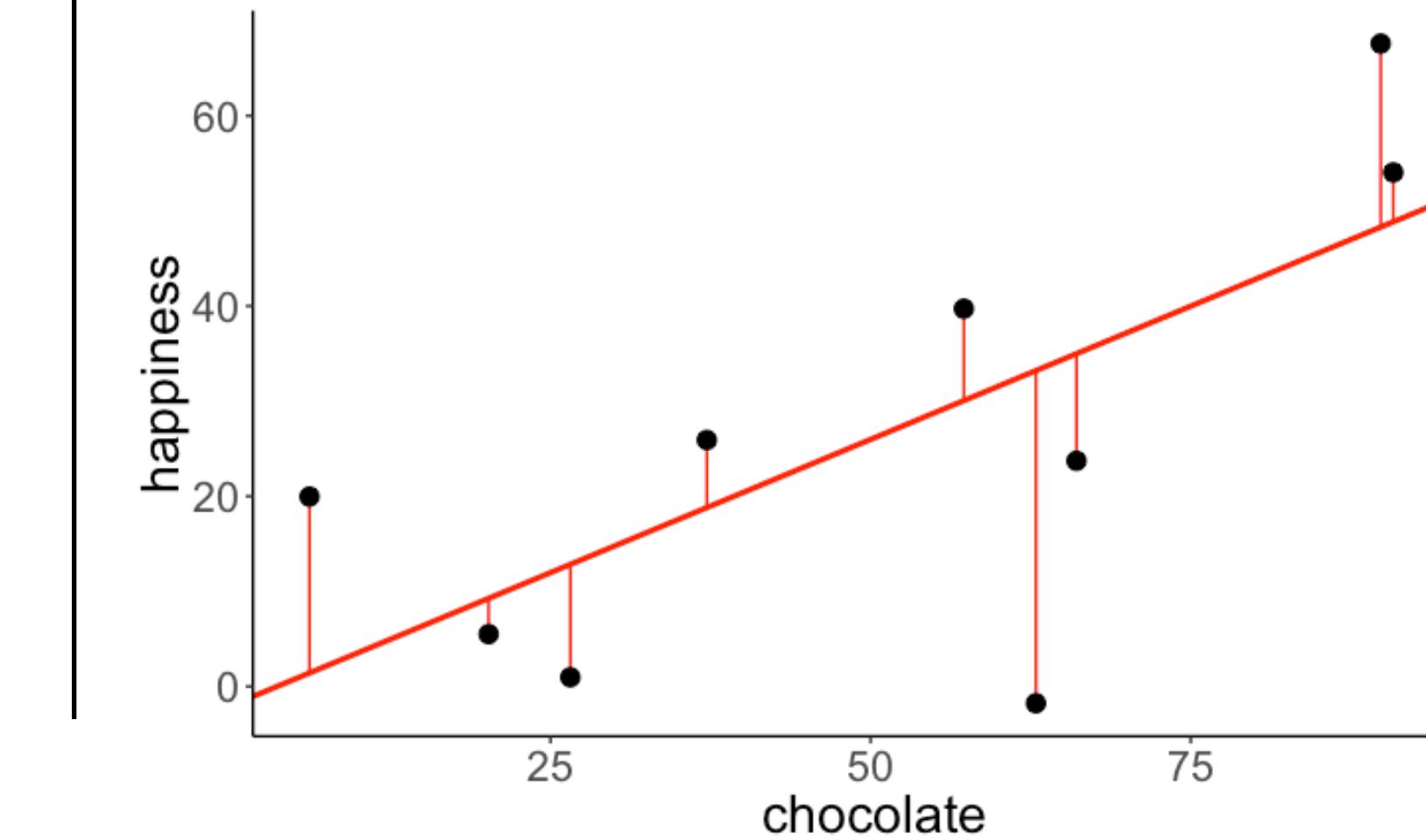
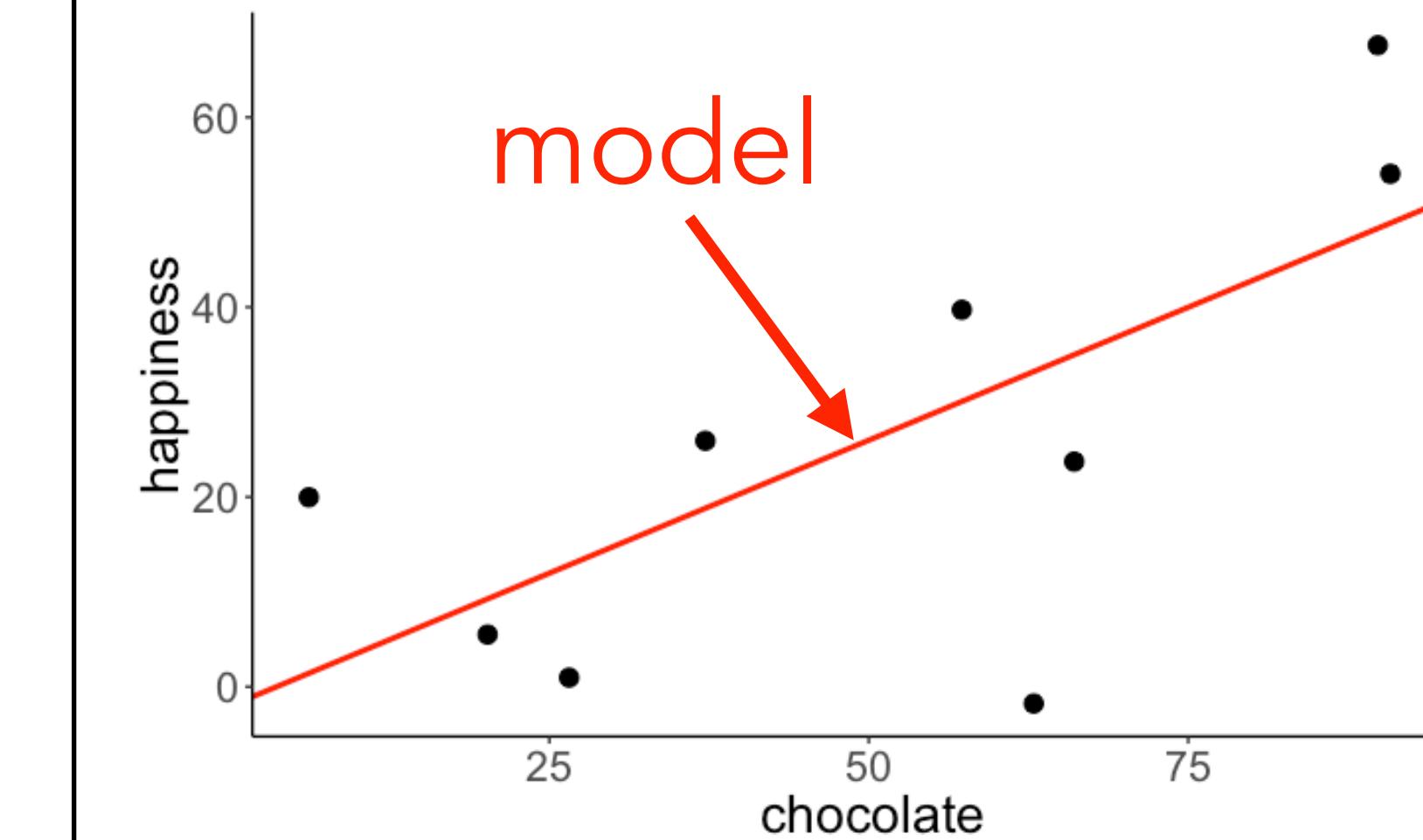


# Data = Model + Error

$H_0$ : Chocolate consumption and happiness are unrelated.



$H_1$ : Chocolate consumption and happiness are related.



# Example

Percentage of households that had internet access in the year 2013 by US state

i	internet	state	college	auto	density
1	79.0	AK	28.0	1.2	1.2
2	63.5	AL	23.5	1.3	94.4
3	60.9	AR	20.6	1.7	56.0
4	73.9	AZ	27.4	1.3	56.3
5	77.9	CA	31.0	0.8	239.1
6	79.4	CO	37.8	1.0	48.5
7	77.5	CT	37.2	1.0	738.1
8	74.5	DE	29.8	1.1	460.8
9	74.3	FL	27.2	1.2	350.6
10	72.2	GA	28.3	1.1	168.4

set before looking at the data

$$\text{model}_1: Y_i = 75 + \text{ERROR}$$

worth it?

fit to the data

$$\text{model}_2: Y_i = \beta_0 + \text{ERROR}$$

worth it?

additional predictor

$$\text{model}_3: Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$$

fit to the data

## Compact model

$\text{model}_C: Y_i = \beta_0 + \text{ERROR}$

## Augmented model

$\text{model}_A: Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(C) \geq \text{ERROR}(A)$$

## Proportional reduction in error (PRE)

$$\text{PRE} = \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

## Compact model

$$\text{model}_C: Y_i = \beta_0 + \text{ERROR}$$

$$\text{ERROR}(C) = 50$$

## Augmented model

$$\text{model}_A: Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$$

$$\text{ERROR}(A) = 30$$

## Proportional reduction in error (PRE)

$$\text{PRE} = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

$$= 1 - \frac{30}{50} = .40$$

Increasing the complexity of the model reduced the error by 40%. **worth it?**



**THE BEST WAY TO  
EXPLAIN OVERFITTING**

# worth it?

## Compact model

model<sub>C</sub>:  $Y_i = \beta_0 + \text{ERROR}$

## Augmented model

model<sub>A</sub>:  $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

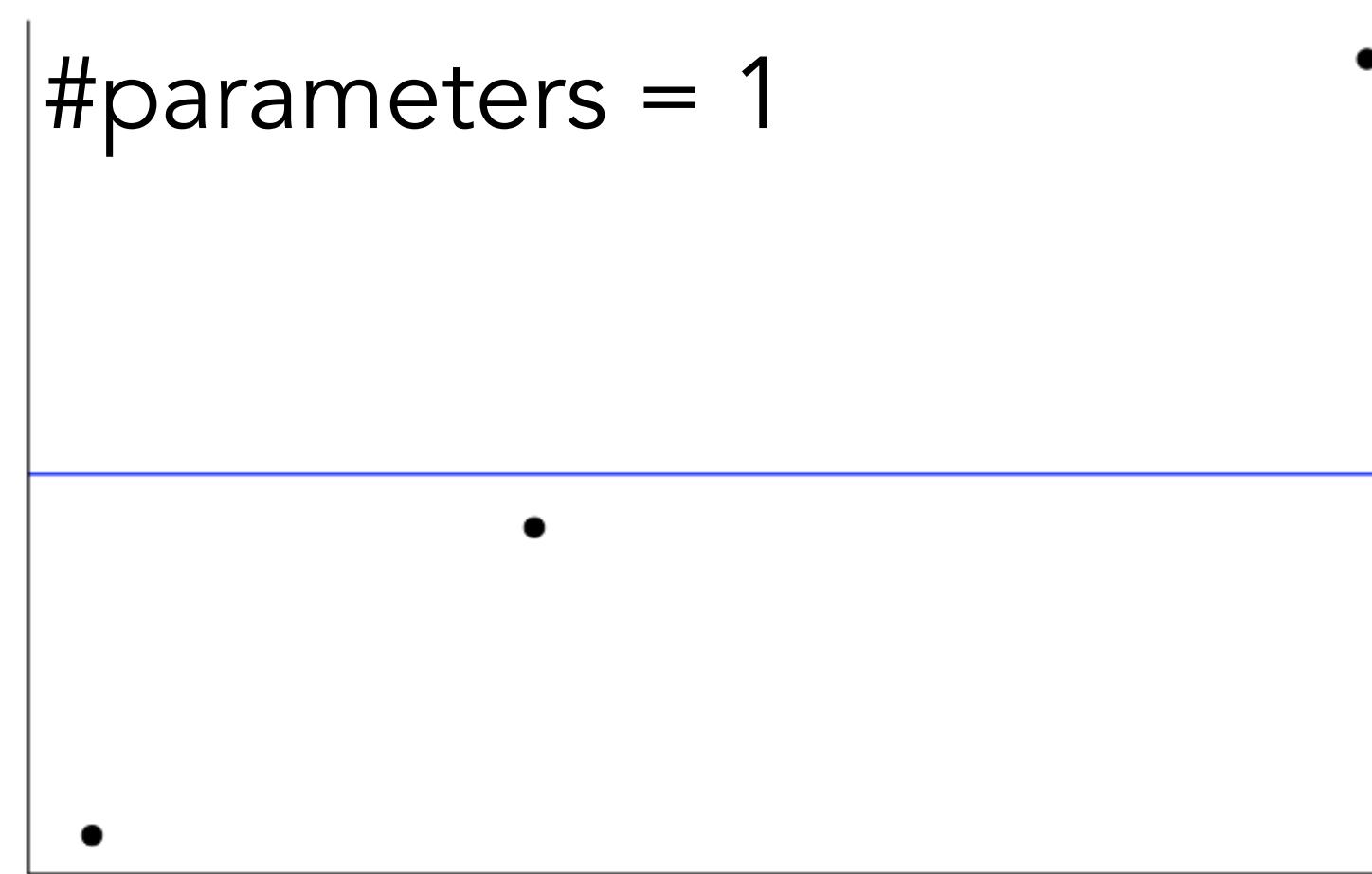
## Proportional reduction in error (PRE)

$$\begin{aligned}\text{PRE} &= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)} \\ &= 1 - \frac{30}{50} = .40\end{aligned}$$

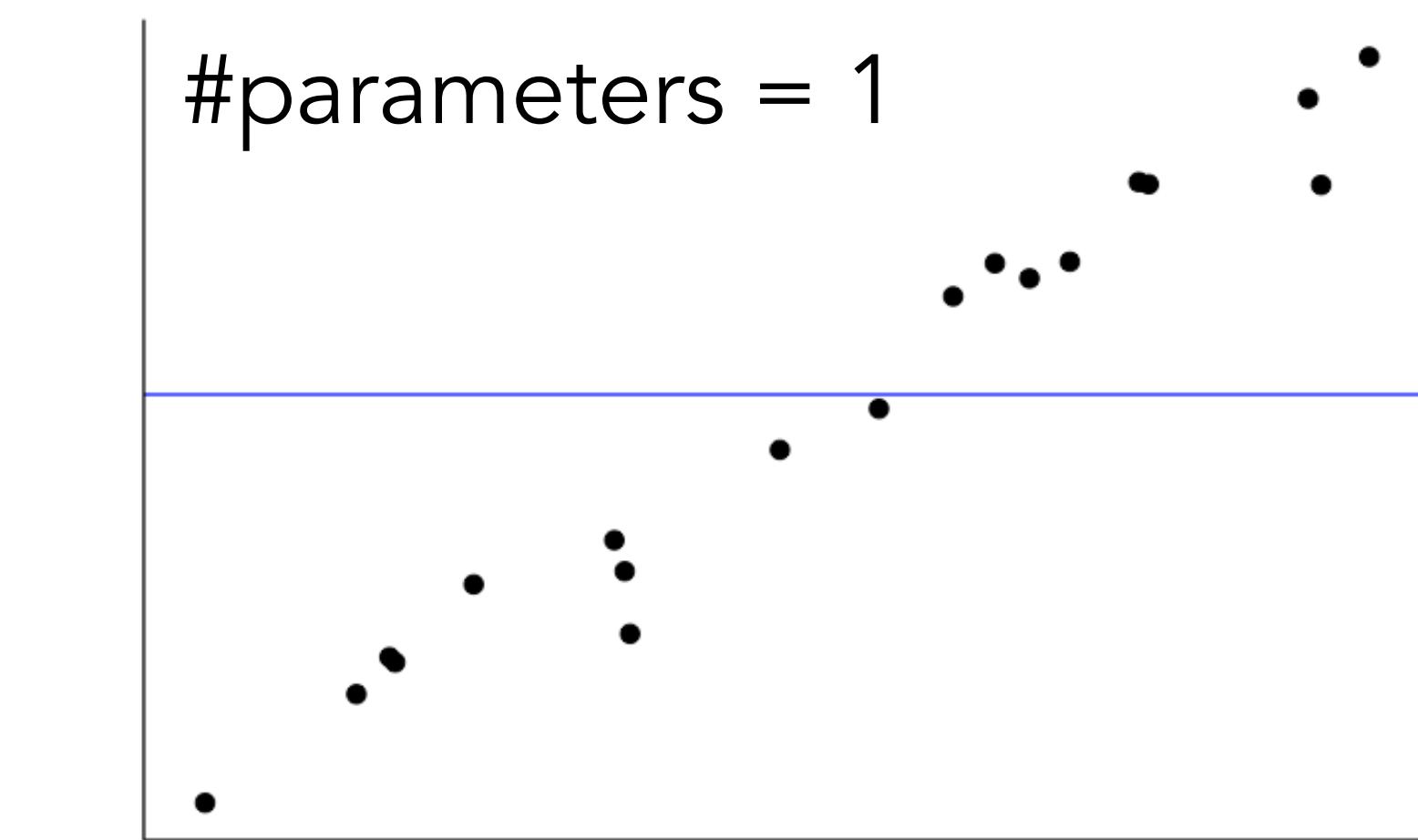
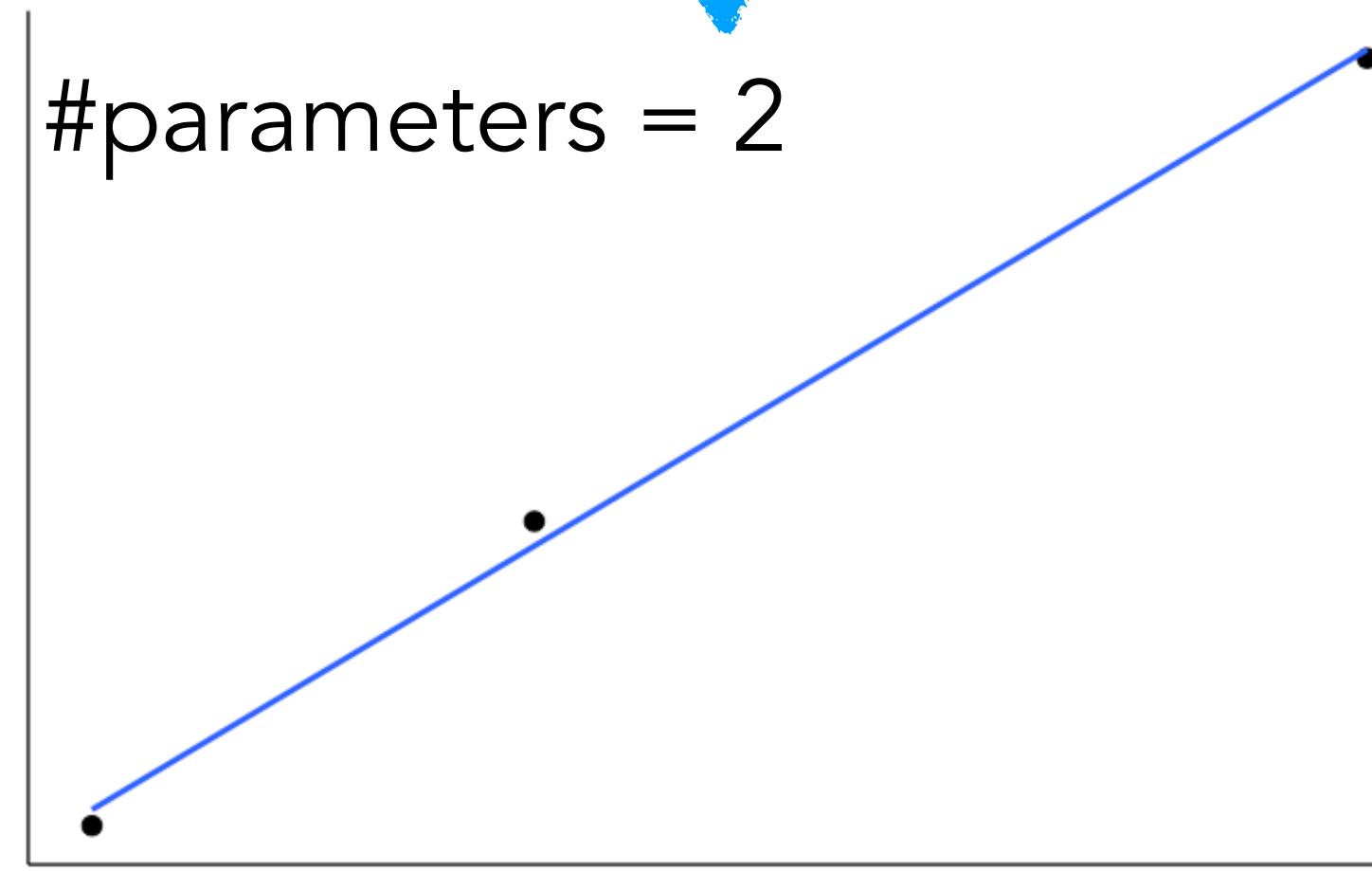
- to answer the **worth it?** question we need inferential statistics (define ERROR, sampling distributions, etc.)
- more likely to be **worth it** if:
  1. **PRE** is high
  2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
  3. the number of parameters that could have been added to model<sub>C</sub> to create model<sub>A</sub> but were not is high

more impressed if the number of observations n is much greater than the number of parameters

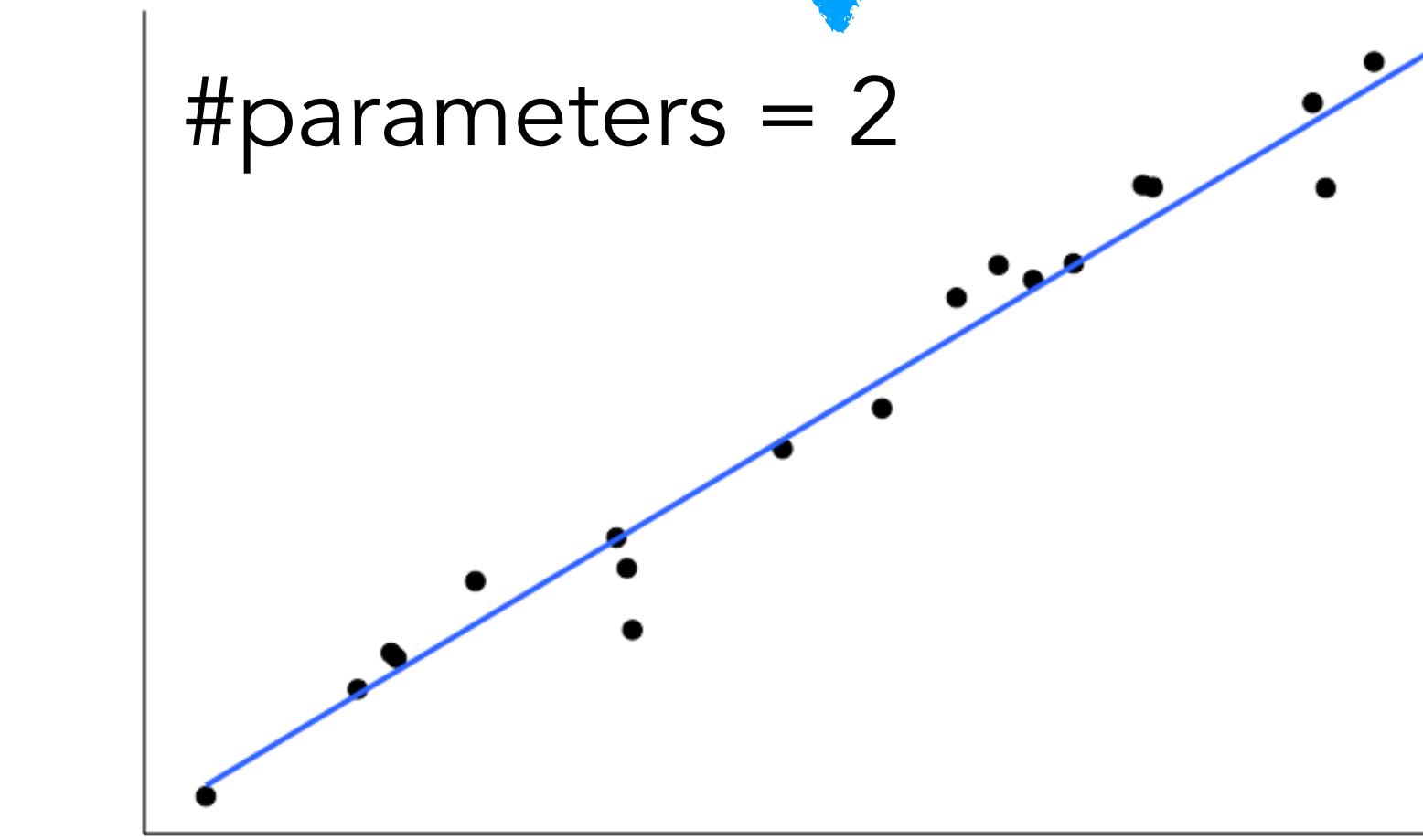
# PRE per parameter for different $n$



↓ neato!



↓ impressive!



# General procedure

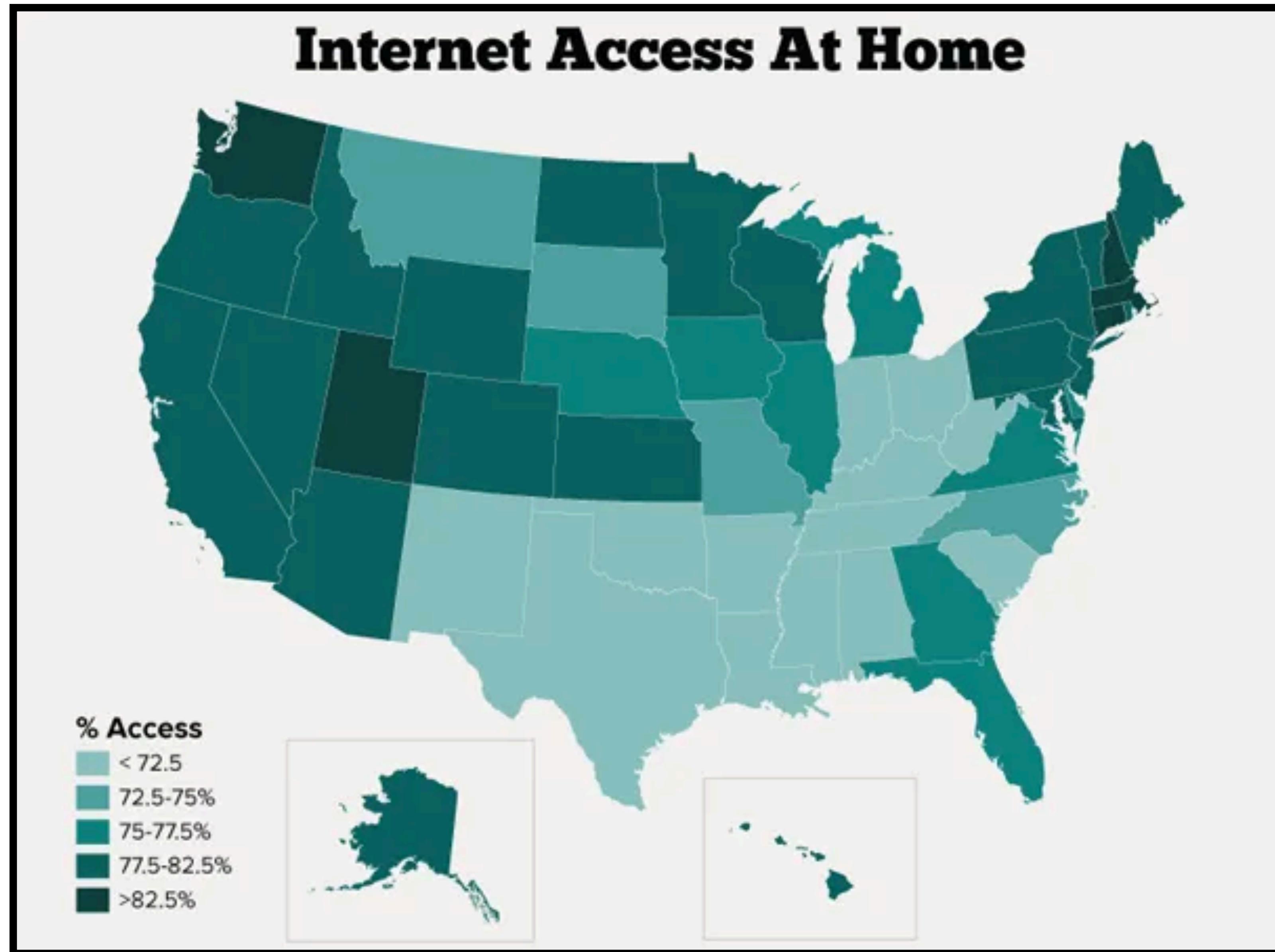
- for any question we want to ask about our DATA
    - we define model<sub>C</sub> and model<sub>A</sub>
    - compare the models using PRE
    - determine whether PRE is **worth it**
  - in standard frequentist lingo:
    - model<sub>C</sub> =  $H_0$  (null hypothesis)
    - model<sub>A</sub> =  $H_1$  (alternative hypothesis)
  - hypothesis test:
    - $H_0$ : **all** the parameters that are included in model<sub>A</sub> but not in model<sub>C</sub> are 0
    - $H_1$ : **not all** the parameters that are included in model<sub>A</sub> but not in model<sub>C</sub> are 0
- 
- model comparison**

# Hypothesis testing as model comparison

# Procedure

1. Start with research question
2. Formulate hypothesis as a comparison between compact and augmented model
3. Fit parameters in each model
4. Calculate the proportional reduction of error (PRE)
5. Decide whether PRE is **worth it**

# Internet Access At Home



<http://thedataviz.com/2012/07/19/mapping-internet-access-in-u-s-homes/>

# Notation

DATA = MODEL + ERROR

$$Y_i = \beta_0 + \epsilon_i \quad \text{simple model (true parameters)}$$

$$Y_i = b_0 + e_i \quad \text{simple model (estimated parameters)}$$

$$\hat{Y}_i = b_0$$

$$Y_i = b_0 + b_1 X_{i1} + e_i \quad \text{more complex model}$$

density

Percentage of households that had internet access in the year 2013 by US state

i	internet	state	college	auto	density
1	79.0	AK	28.0	1.2	1.2
2	63.5	AL	23.5	1.3	94.4
3	60.9	AR	20.6	1.7	56.0
4	73.9	AZ	27.4	1.3	56.3
5	77.9	CA	31.0	0.8	239.1
6	79.4	CO	37.8	1.0	48.5
7	77.5	CT	37.2	1.0	738.1
8	74.5	DE	29.8	1.1	460.8
9	74.3	FL	27.2	1.2	350.6
10	72.2	GA	28.3	1.1	168.4



Greek letters  $\beta$  or  $\epsilon$  represent the true but unknowable parameters in the population.

Roman letters  $b$  or  $e$  represent estimates of these parameters using our DATA.

# Research question and hypotheses

Is the average percentage of internet users per state different from 75%?

Model<sub>C</sub>:  $Y_i = B_0 + \epsilon_i$   
**0 parameters**

$$Y_i = 75 + e_i$$

Model<sub>A</sub>:  $Y_i = \beta_0 + \epsilon_i$   
**1 parameter**

$$Y_i = b_0 + e_i$$

$$= \bar{Y} + e_i$$

i	internet	state	college	auto	density
1	79.0	AK	28.0	1.2	1.2
2	63.5	AL	23.5	1.3	94.4
3	60.9	AR	20.6	1.7	56.0
4	73.9	AZ	27.4	1.3	56.3
5	77.9	CA	31.0	0.8	239.1
6	79.4	CO	37.8	1.0	48.5
7	77.5	CT	37.2	1.0	738.1
8	74.5	DE	29.8	1.1	460.8
9	74.3	FL	27.2	1.2	350.6
10	72.2	GA	28.3	1.1	168.4

# Fit parameters and calculate PRE

$$\mathbf{C: } Y_i = 75 + e_i \quad \mathbf{A: } Y_i = \bar{Y} + e_i$$

i	state	internet	compact_b	compact_se	augmented_b	augmented_se
1	AK	79.0	75	16.00	72.81	38.37
2	AL	63.5	75	132.25	72.81	86.60
3	AR	60.9	75	198.81	72.81	141.75
4	AZ	73.9	75	1.21	72.81	1.20
5	CA	77.9	75	8.41	72.81	25.95
6	CO	79.4	75	19.36	72.81	43.48
7	CT	77.5	75	6.25	72.81	22.03
8	DE	74.5	75	0.25	72.81	2.87
9	FL	74.3	75	0.49	72.81	2.23
10	GA	72.2	75	7.84	72.81	0.37

$$\text{SSE(C)} = 1595 \quad \text{SSE(A)} = 1355$$

$$\begin{aligned}\text{PRE} &= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)} \\ &= 1 - \frac{1355}{1595} \approx .15\end{aligned}$$

Model A has  
15% less error  
than Model C.

# Decide whether it's **worth it**

- PRE is the estimate of an unknown true reduction of error  $\eta^2$
- we need a sampling distribution of PRE
  - a distribution of what PRE would look like if Model C (our  $H_0$ ) were true
  - we could just simulate such a sampling distribution ...
- PRE is closely related to the  $F$  statistic!

# Decide whether it's **worth it**

- To compute the  $F$  statistic, we need:
  - PRE
  - number of parameters in Model C (PC) and Model A (PA)
  - number of observations  $n$

- more likely to be **worth it** if:
  1. PRE is high
  2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
  3. the number of parameters that could have been added to  $\text{model}_C$  to create  $\text{model}_A$  but were not

**difference in parameters  
between models A and C**

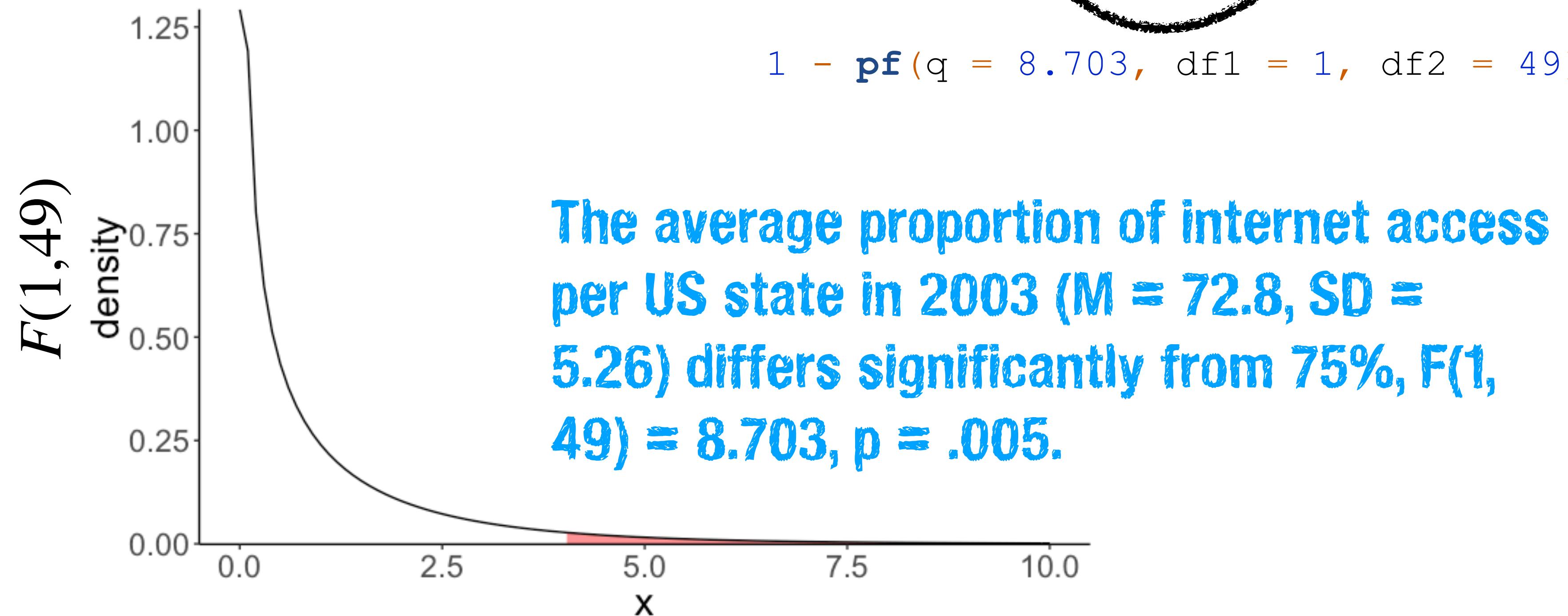
$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

**number of observations vs.  
parameters in Model A**

# Decide whether it's **worth it**

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$
$$= \frac{.15/(1 - 0)}{(1 - .15)/(50 - 1)} = 8.703 \quad p = .00486$$

**Note:** I've used the approximated PRE here (PRE = 0.15), but I'm reporting the F value for the non-approximated PRE value.



we just performed a one sample t-test ...

```
t.test(df.internet$internet, mu = 75)
```

## One Sample t-test

```
data: df.internet$internet  
t = -2.9502, df = 49, p-value = 0.00486  
alternative hypothesis: true mean is not equal to 75
```

but ...

- we will apply the general procedure we went through to day to a large range of different situations
- thinking of hypothesis testing as model comparison is (hopefully more) intuitive
- this approach still works even if we can't find a recipe in our favorite stats cook book

# Summary

- Quick recap
- Statistical concepts
  - Confidence intervals
  - Bootstrapping
- Cookbook vs. Model Comparison
- Modeling data
- Hypothesis testing as model comparison

# **Feedback**



0%

much too slow

0%

a little too slow

0%

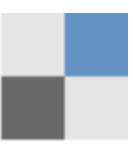
just right

0%

a little too fast

0%

much too fast



Start the presentation to see live content. For screen share software, share the entire screen. Get help at [pollev.com/app](https://pollev.com/app)



**Thank you!**