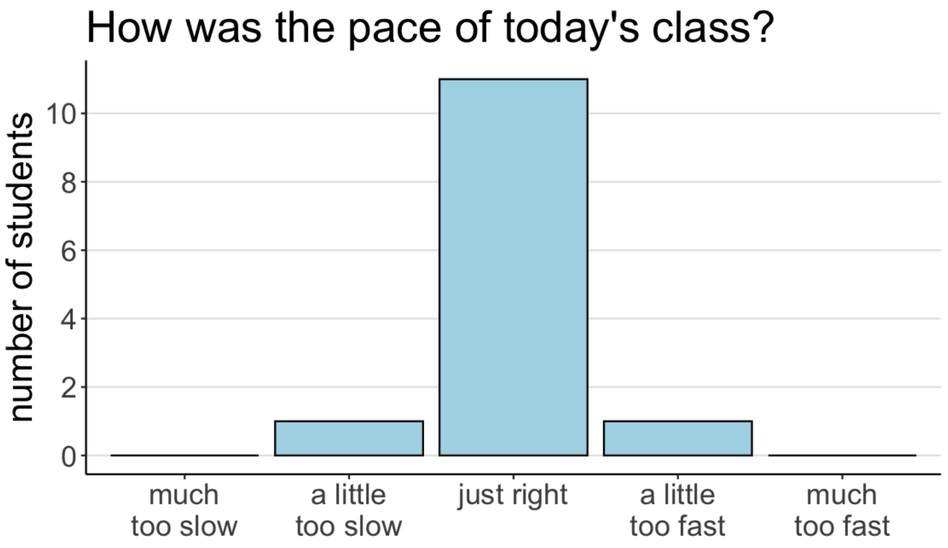
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

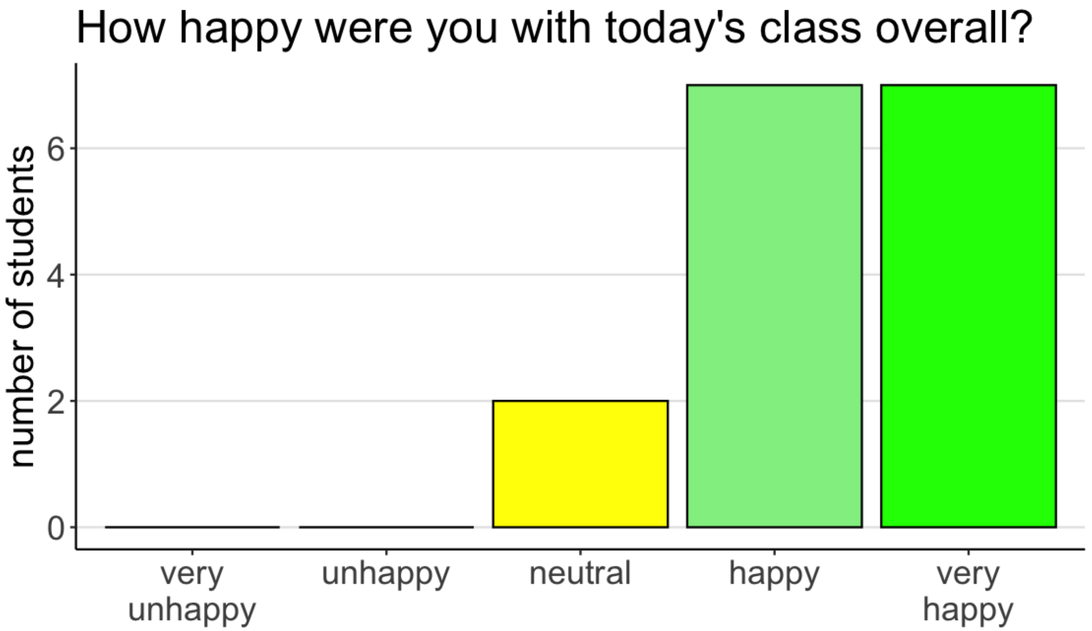
Modeling data

01/27/2025

Feedback

Your feedback





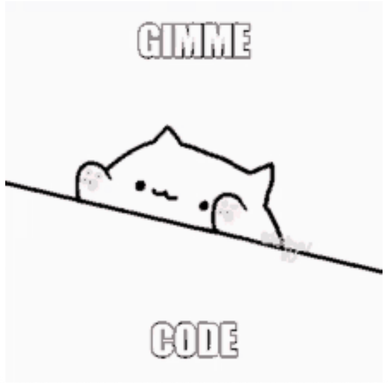
- Seeing the logic and computations of the permutation test felt like a more concrete/accessible way to understand the logic of p values etc derived analytically!
- very thorough review of distributions and p-values!
- We should talk about p hacking

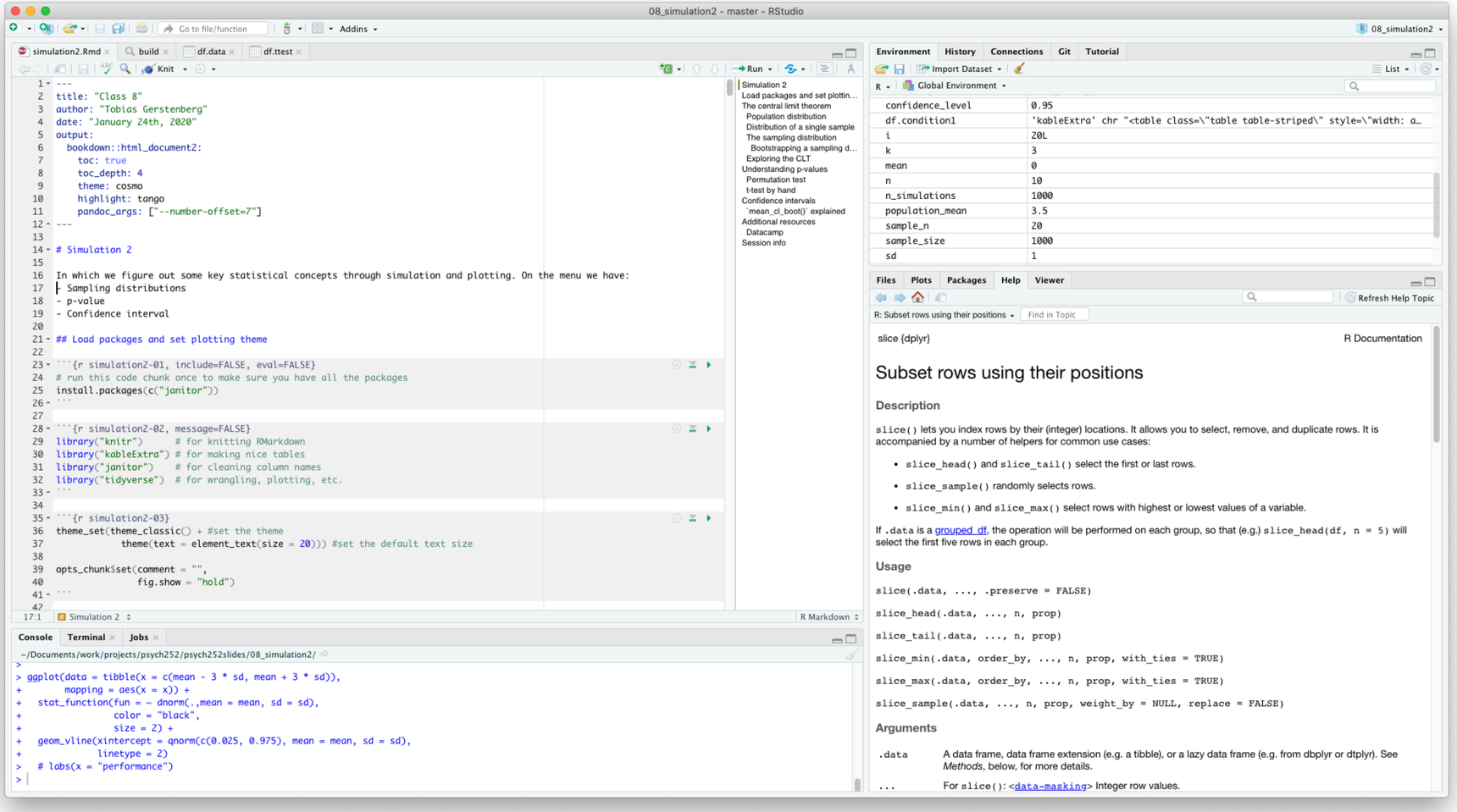
- Release the files so I can download them the night before class instead of in the morning.
- I would appreciate baking in time for questions. Maybe pause every 3 minutes of lecture and do a sanity check with the audience

- You are responding well to questions. Keep the good work!

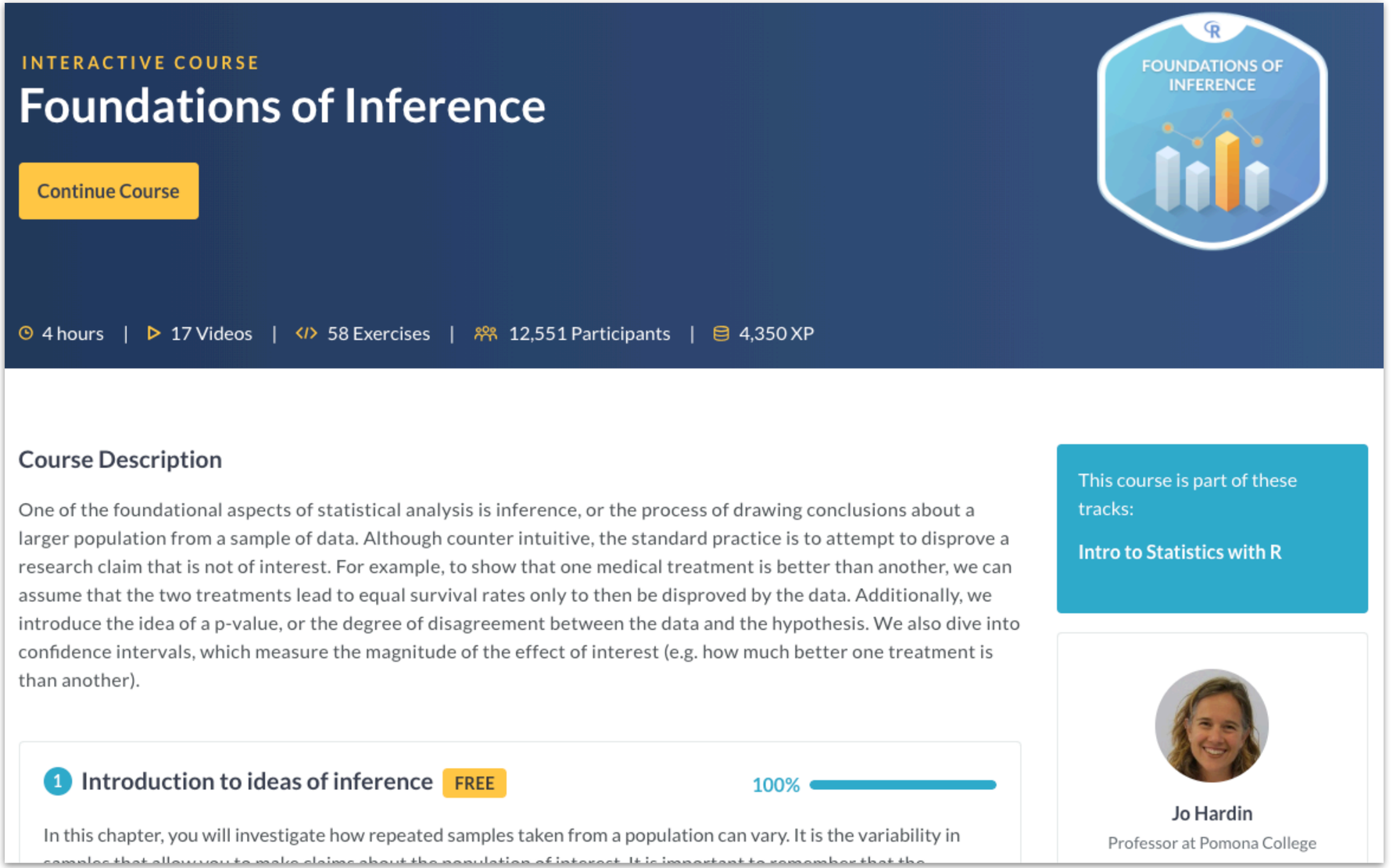
Things that came up ...

How to better understand!





Datacamp



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

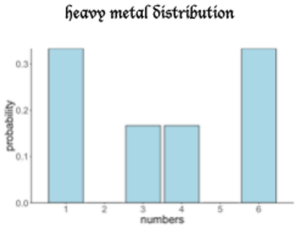
Plan for today

- 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

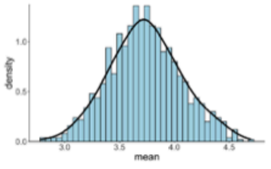
3 distributions in statistical inference



population distribution

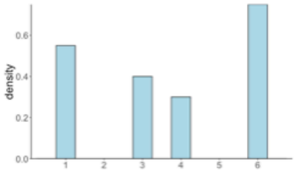
- unknown
- our target for inference
- e.g. we might be interested in the mean of the population distribution

---



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

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.

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

what we're actually interested in!

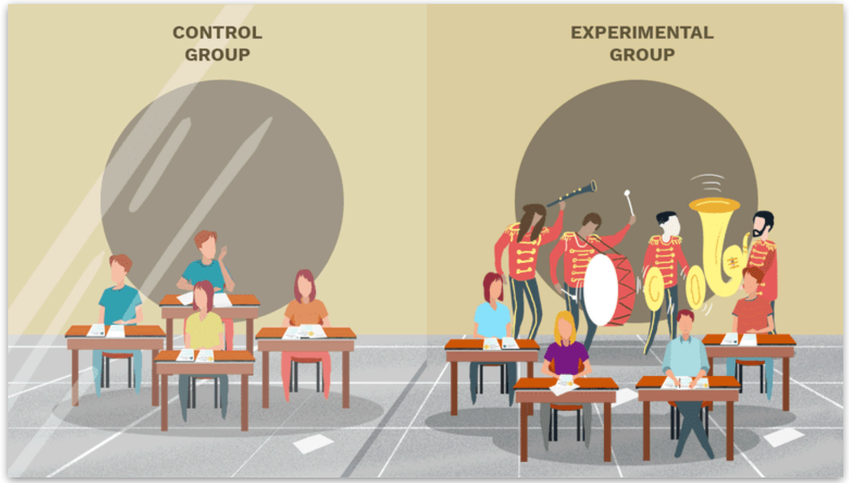
~~$p(H_1 = \text{true} \mid \text{test statistic} \geq \text{observed value})$~~

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

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

H = Hypothesis

D = Data



Permutation test

observed data

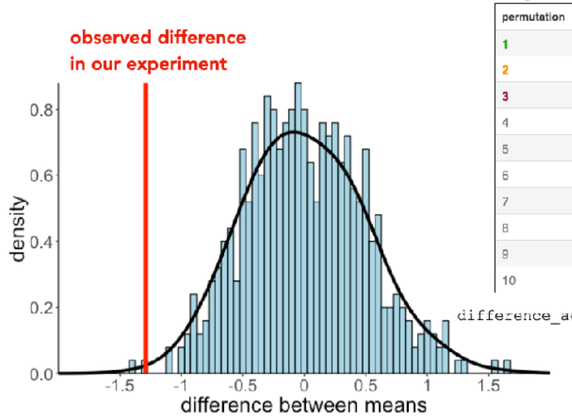
participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

1 2 3

permutation	mean_difference
1	-0.88
2	-0.28
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

observed difference in our experiment



df.permutations

permutation	mean_difference
1	-0.88
2	-0.28
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

Statistical concepts

Confidence intervals

Confidence interval

Goal: Estimate μ = the mean of the population distribution

Confidence interval of the mean = point estimate ± critical value

the sample mean is our best guess of the population mean

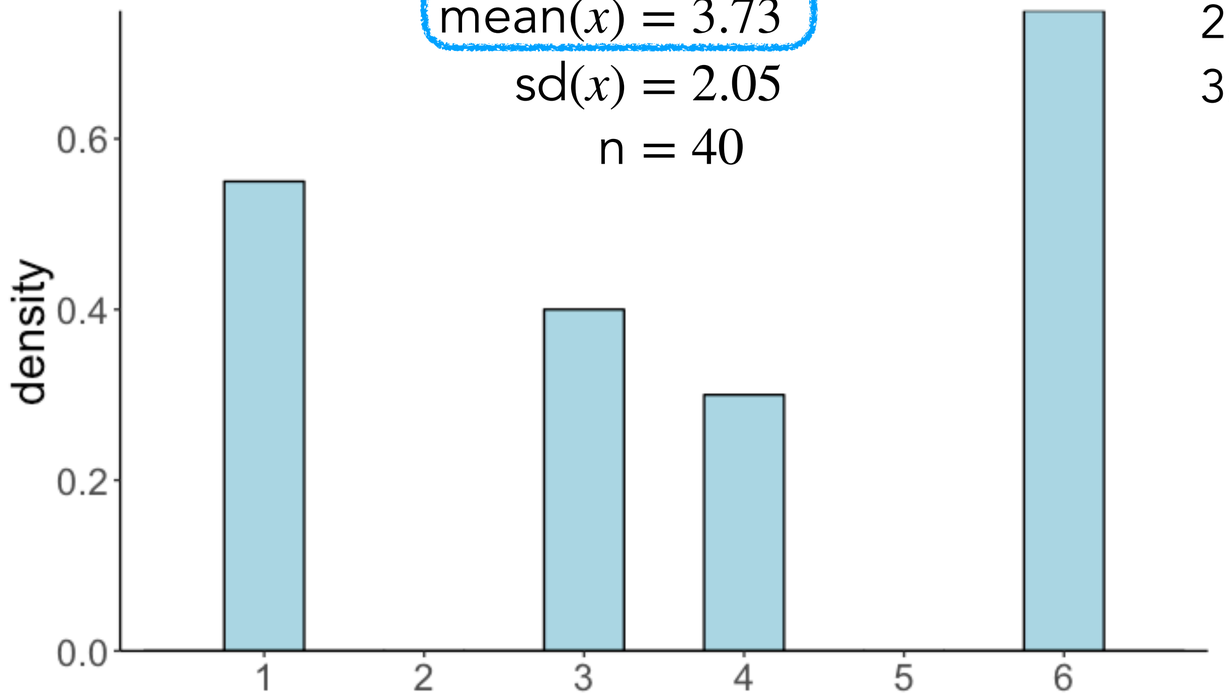
depends on

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

$\text{mean}(x) = 3.73$

$\text{sd}(x) = 2.05$

$\text{n} = 40$



Confidence interval

Goal: Estimate μ = the mean of the population distribution

what we need:

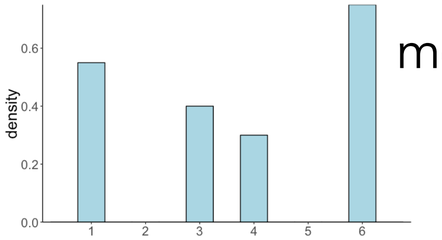
- sample mean
- sample standard deviation
- sample size
- desired level of confidence

Confidence interval

Confidence interval of the mean = point estimate ± critical value

Confidence interval

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

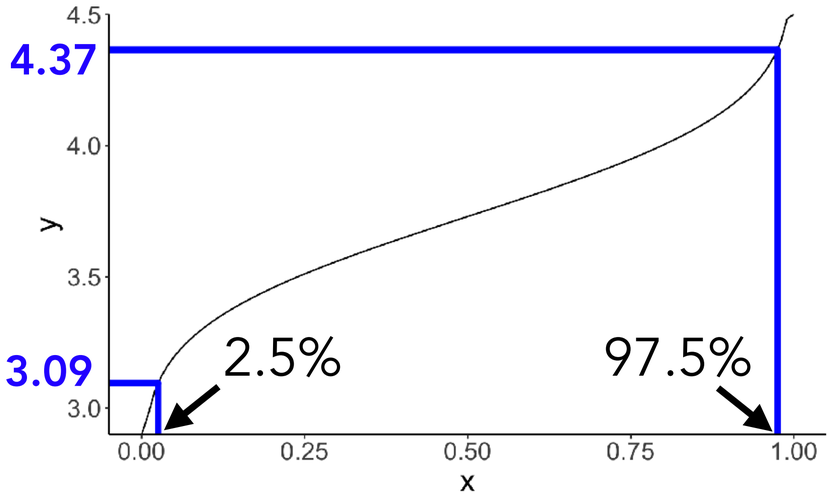


$$\text{mean}(x) = 3.73$$
$$\text{sd}(x) = 2.05$$
$$\text{n} = 40$$

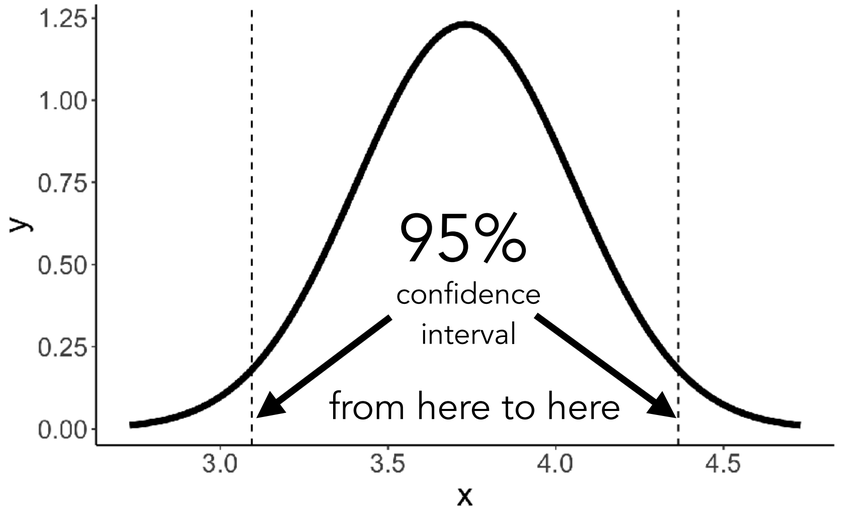
our sample

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

sampling distribution of the mean

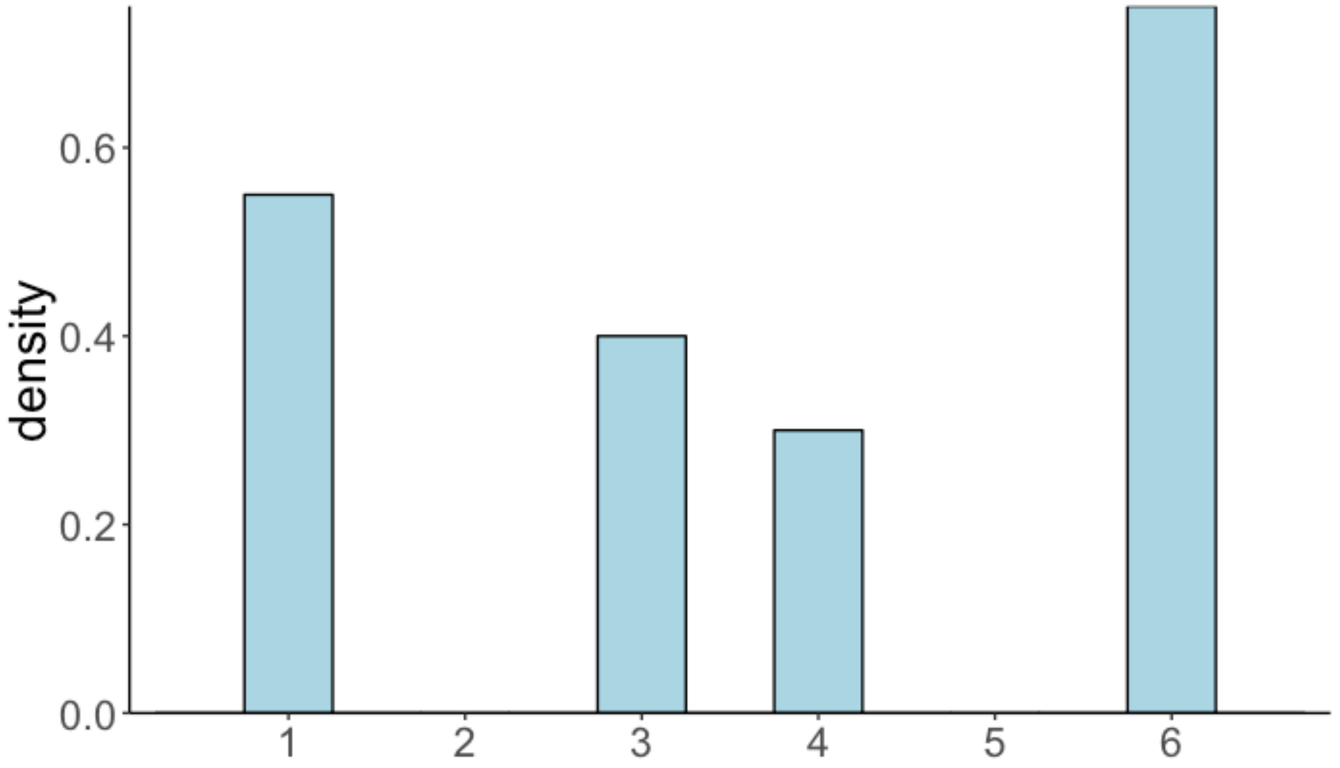


~ qnorm(., mean = 3.73, sd = 0.324)



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

What does the confidence interval mean?



Mean = 3.73 ± 0.63 (95% CI)

`qnorm(0.975) * SE`

Confidence interval

Confidence interval of the mean = point estimate ± critical value

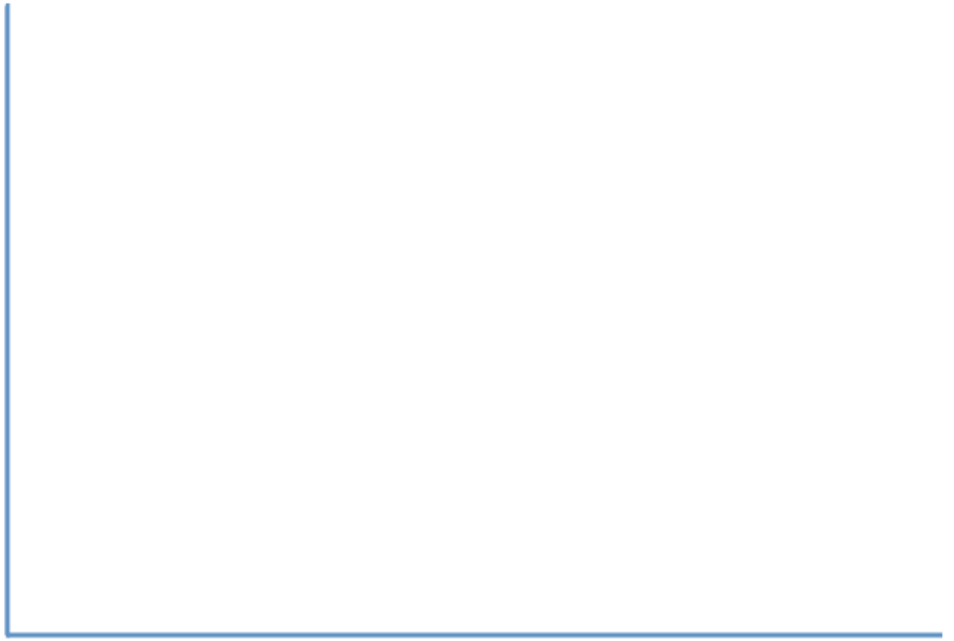
What can we say based on the result of our sample (N = 40): Mean = 3.73 ± 0.63 (95% CI)?

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.

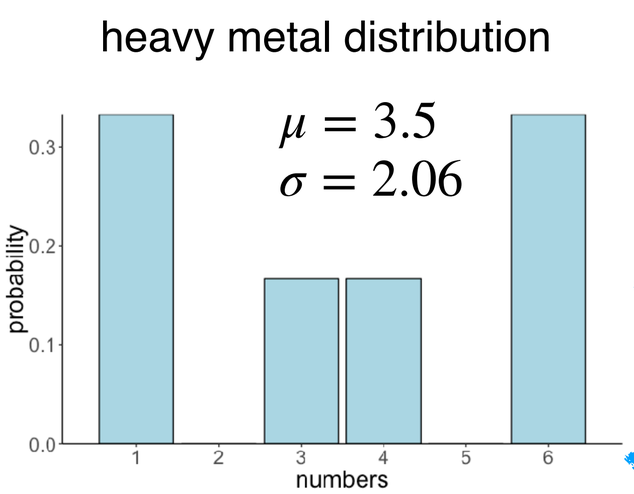
We can be 95% confident that the sample mean is in this interval.

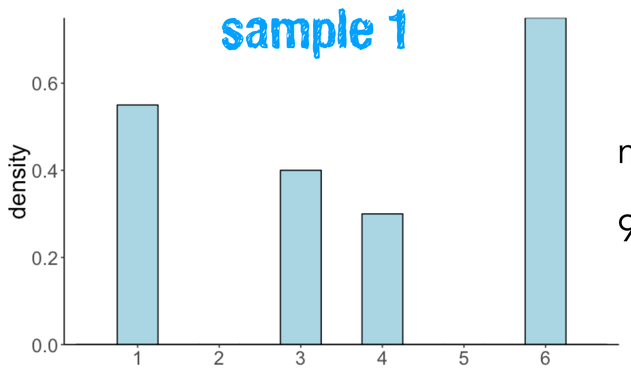


What is a confidence interval? Your answers

- 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.

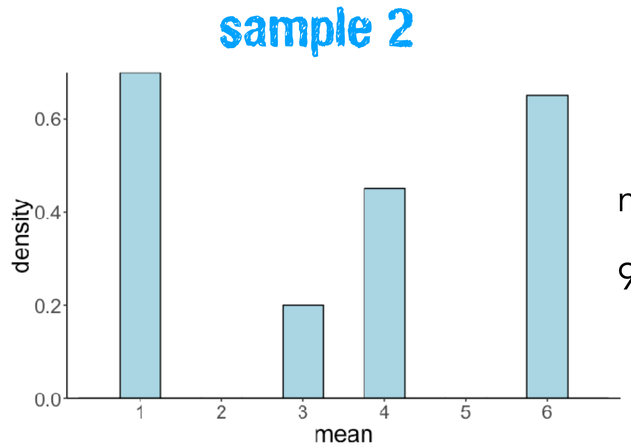
Confidence interval





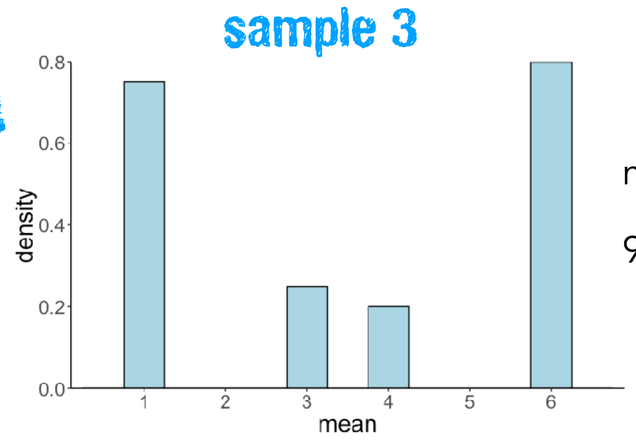
mean = 4.04

95% CI = [3.24, 4.83]



mean = 3.52

95% CI = [2.66, 4.37]



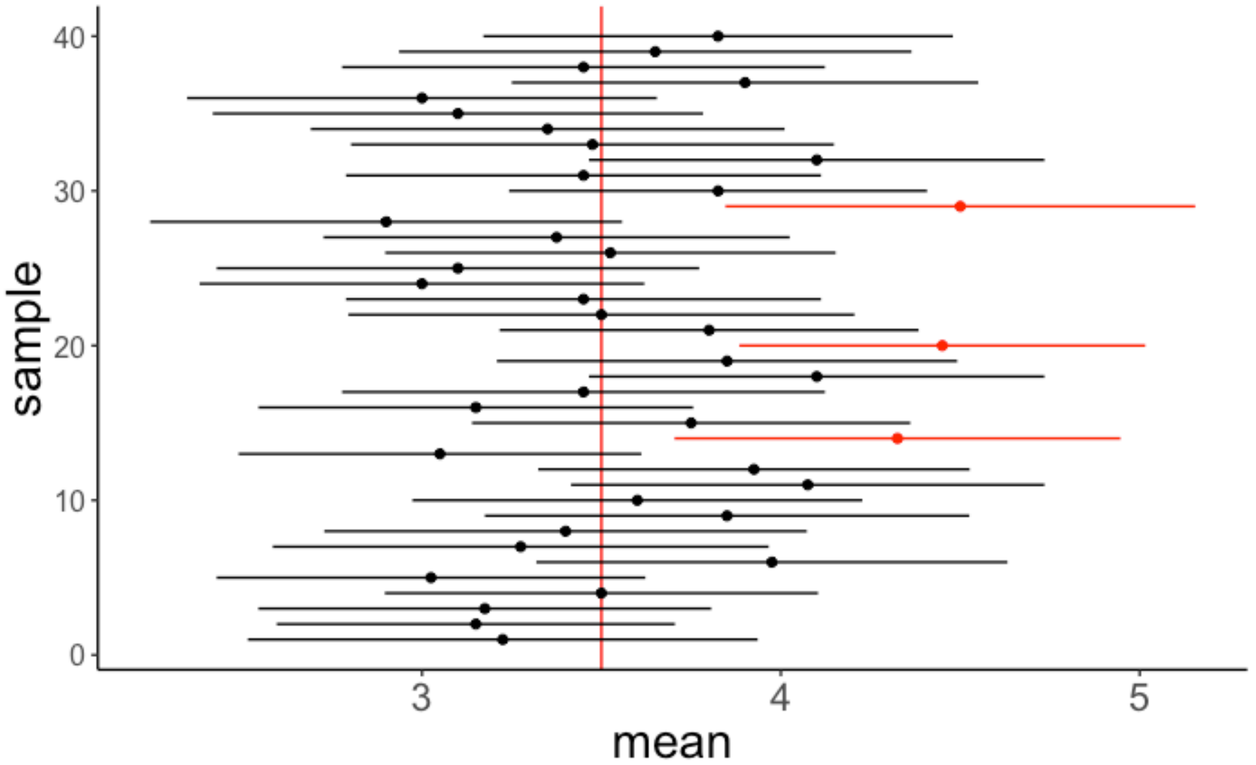
mean = 3.44

95% CI = [2.64, 4.24]

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."



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.

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

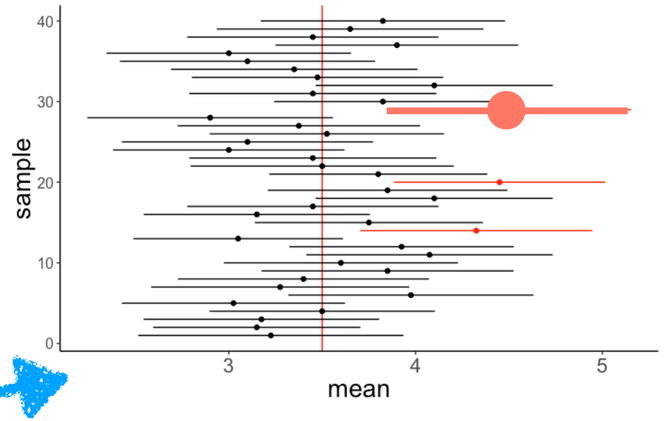
The sample means of 95% of the random samples of size 40 will be in this interval.

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

It either is in this interval or isn't.

correct

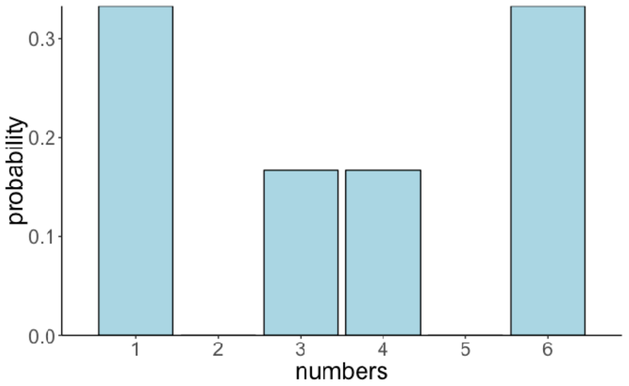
incorrect



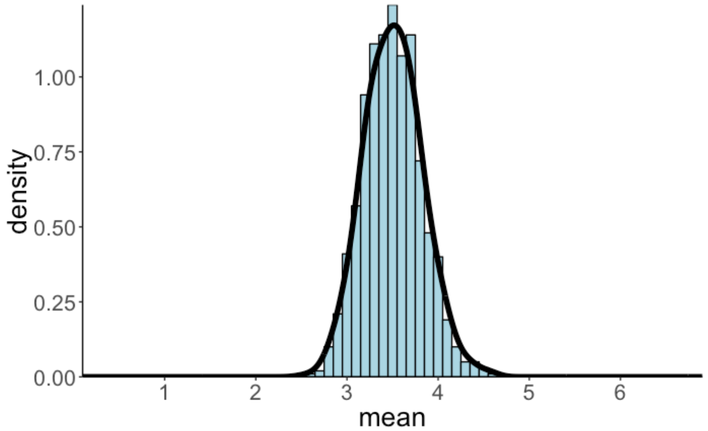
We know what the sample mean is.

Bootstrapping

Bootstrap



repeated sampling



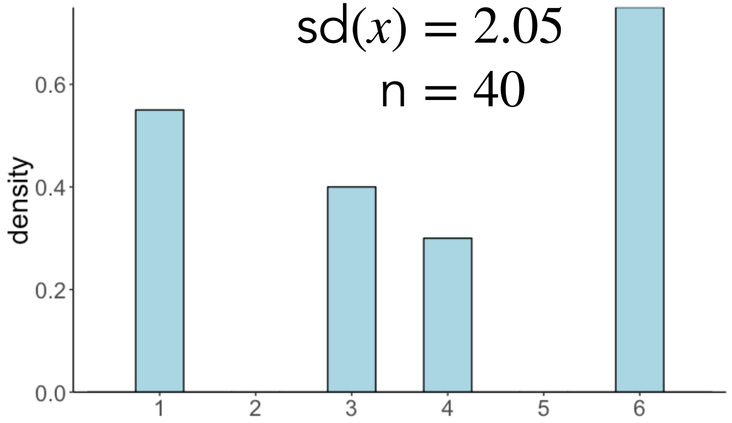
population distribution

sampling distribution

but we don't know the population distribution!

Bootstrap

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

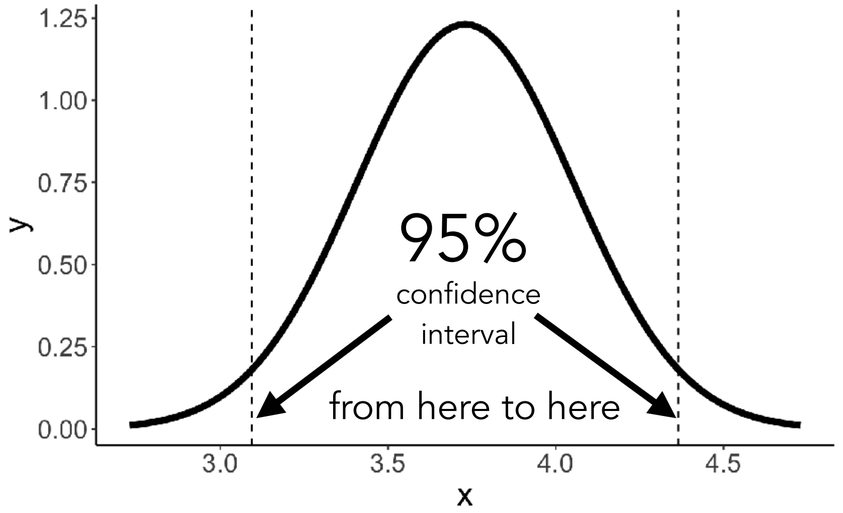


assuming a normal distribution

$\text{mean}(\bar{x}) = 3.73$

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

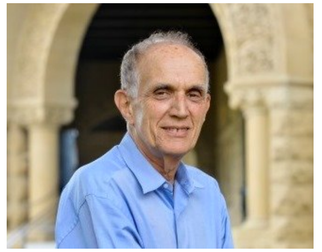
sampling distribution of the mean



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

Bootstrap

invented at Stanford



Bradley Efron

original sample

repeated sampling **with replacement**

bootstrap #1

bootstrap #2

bootstrap #3

mean (x)

mean (x)

mean (x)

means of bootstrapped samples

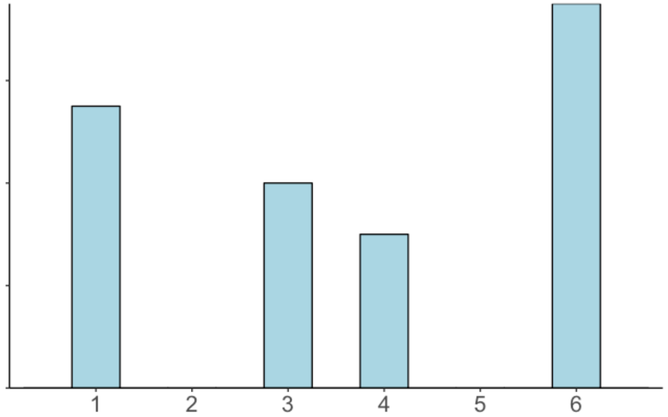
parametric (normal) distribution

sampling distribution

Bootstrap

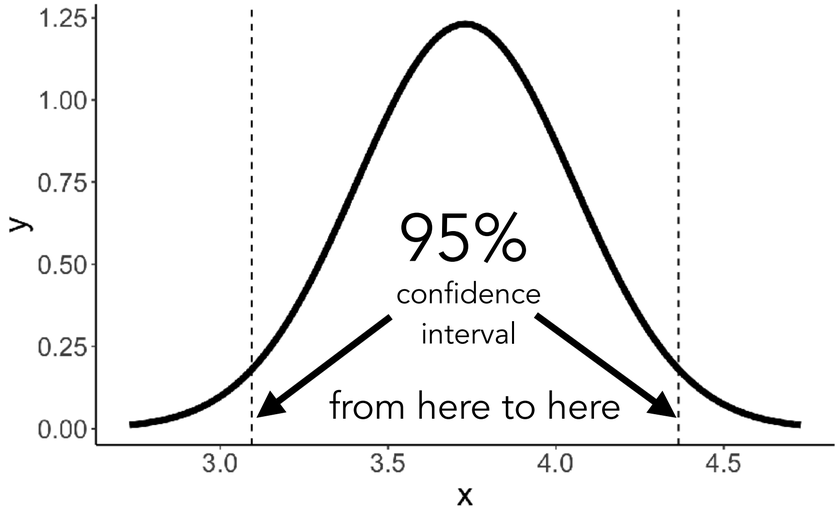
How can I get the confidence interval of a statistical estimate (such as the mean)?

make assumptions

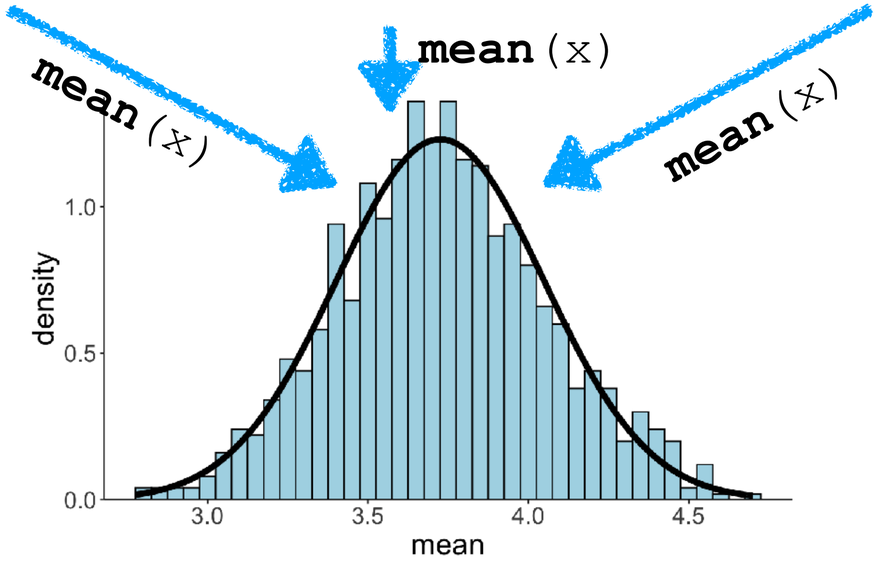


bootstrap

sampling distribution of the mean

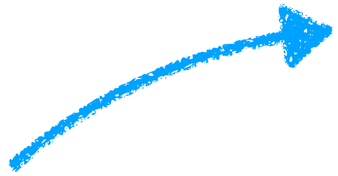


~ dnorm(., mean = 3.73, sd = 0.324)



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))
```



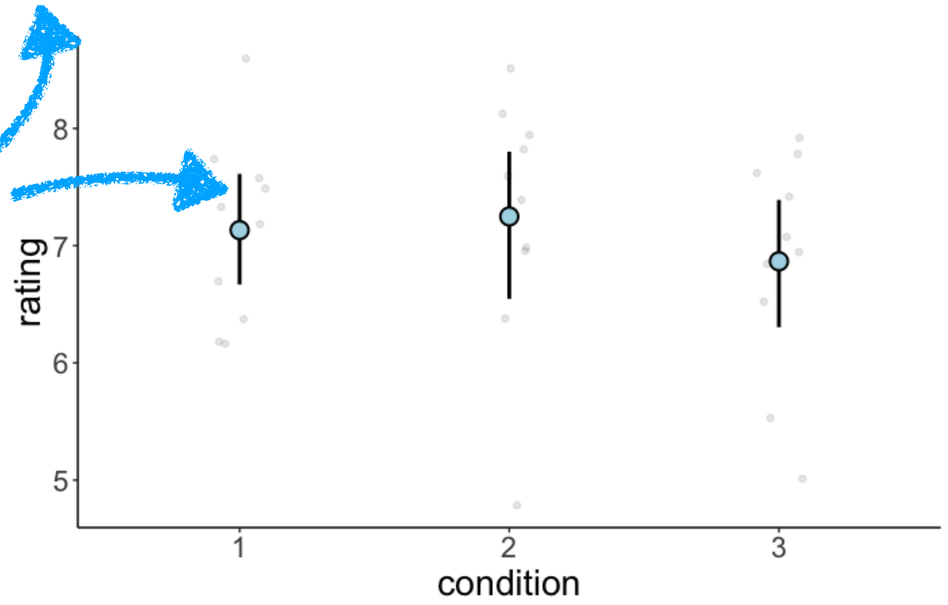
participant	condition
1	1
2	1
3	1
4	1
5	1

mean_cl_boot() explained

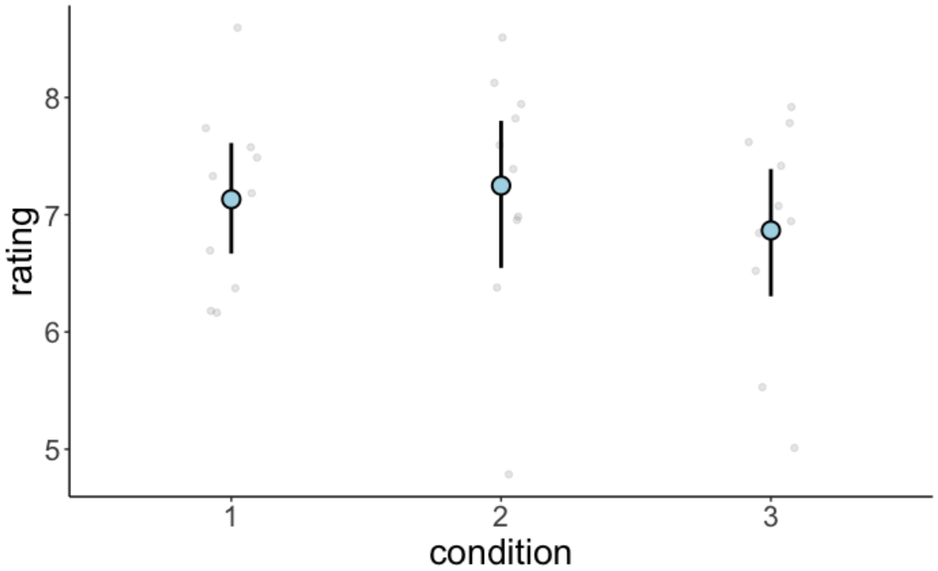
```r
set.seed(1)

n = 10 # sample size per group
k = 3 # number of groups

df.data = tibble(participant = 1:(n*k),
                 condition = as.factor(rep(1:k, each = n)),
                 rating = rnorm(n*k, mean = 7, sd = 1))

ggplot(data = df.data,
       mapping = aes(x = condition,
                     y = rating)) +
  geom_point(alpha = 0.1,
             position = position_jitter(width = 0.1, height = 0)) +
  stat_summary(fun.data = "mean_cl_boot",
               shape = 21,
               size = 1,
               fill = "lightblue")
```

participant	condition	rating
1	1	6.37
2	1	7.18
3	1	6.16
4	1	8.60
5	1	7.33

what is this magic?



`mean_cl_boot()` explained

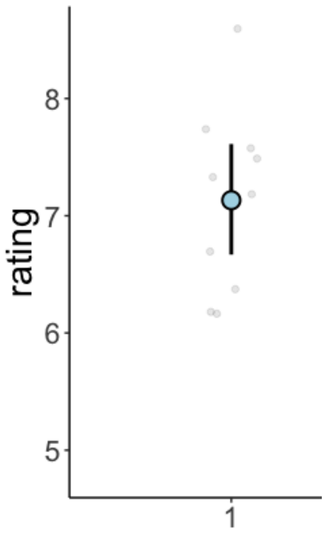



```
ggplot_build(p)
```

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



x	group	y	ymin	ymax
1	1	7.13	6.70	7.59

```
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 }
```

this is what makes this a bootstrap

original

participant
1
2
3
4
5
6
7
8
9
10

bootstrap #1

participant	condition	rating
9	1	7.58
3	1	6.16
3	1	6.16
2	1	7.18
3	1	6.16
7	1	7.49
4	1	8.60
8	1	7.74
5	1	7.33

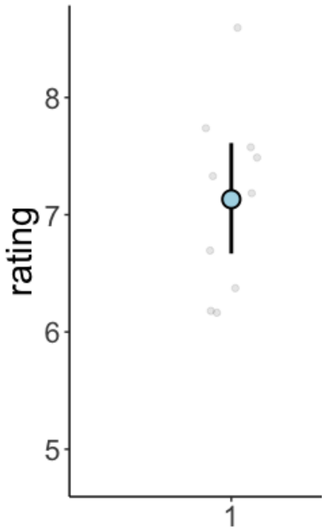
note the repeated observations

bootstrap #2

participant	condition	rating
1	1	6.37
1	1	6.37
6	1	6.18
1	1	6.37
4	1	8.60
1	1	6.37
9	1	7.58
4	1	8.60
3	1	6.16

...

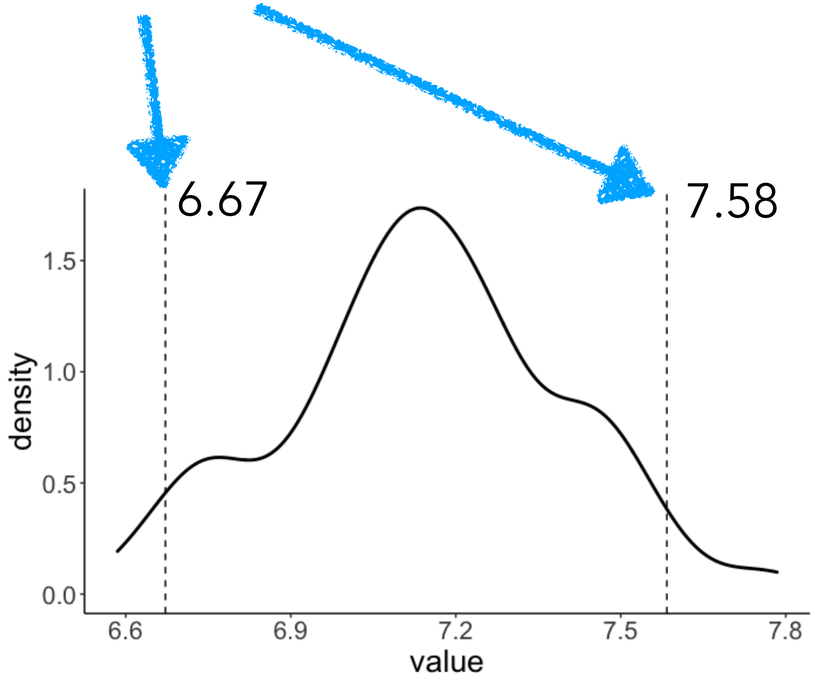
mean_cl_boot() explained

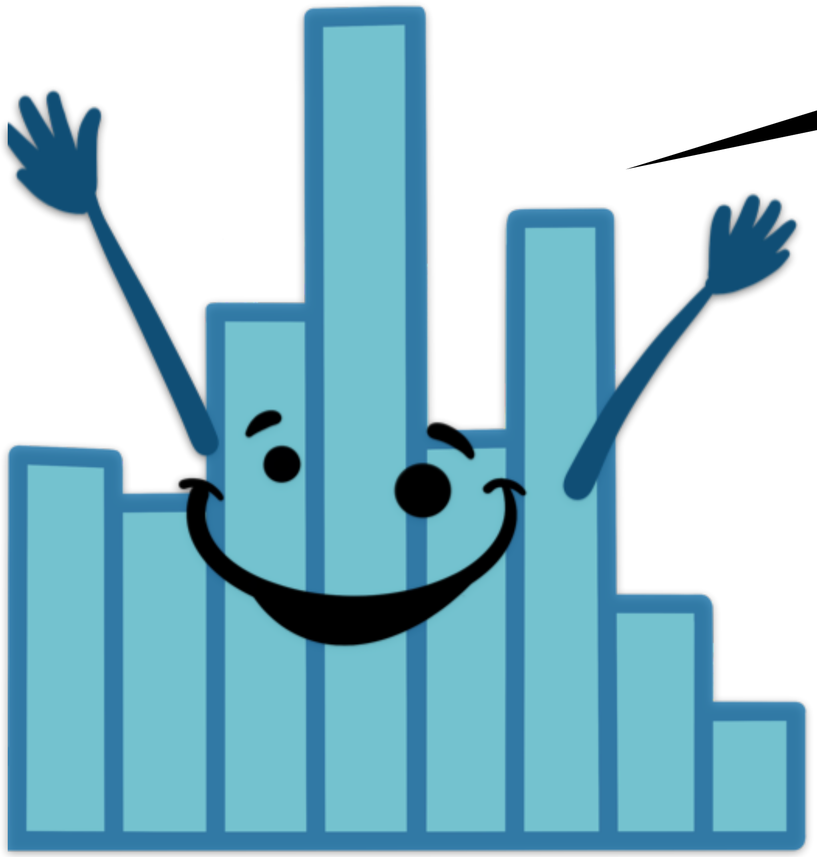



```r
set.seed(1)

df.condition1 = df.data %>%
  filter(condition == 1)

fun.sample_with_replacement = function(df){
  df %>%
    slice_sample(n = nrow(df),
                 replace = T) %>%
    summarize(mean = mean(rating)) %>%
    pull(mean)
}

bootstraps = replicate(n = 100, fun.sample_with_replacement(df.condition1))

quantile(bootstraps, prob = c(0.025, 0.975))
```

x	group	y	ymin	ymax
1	1	7.13	6.70	7.59

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



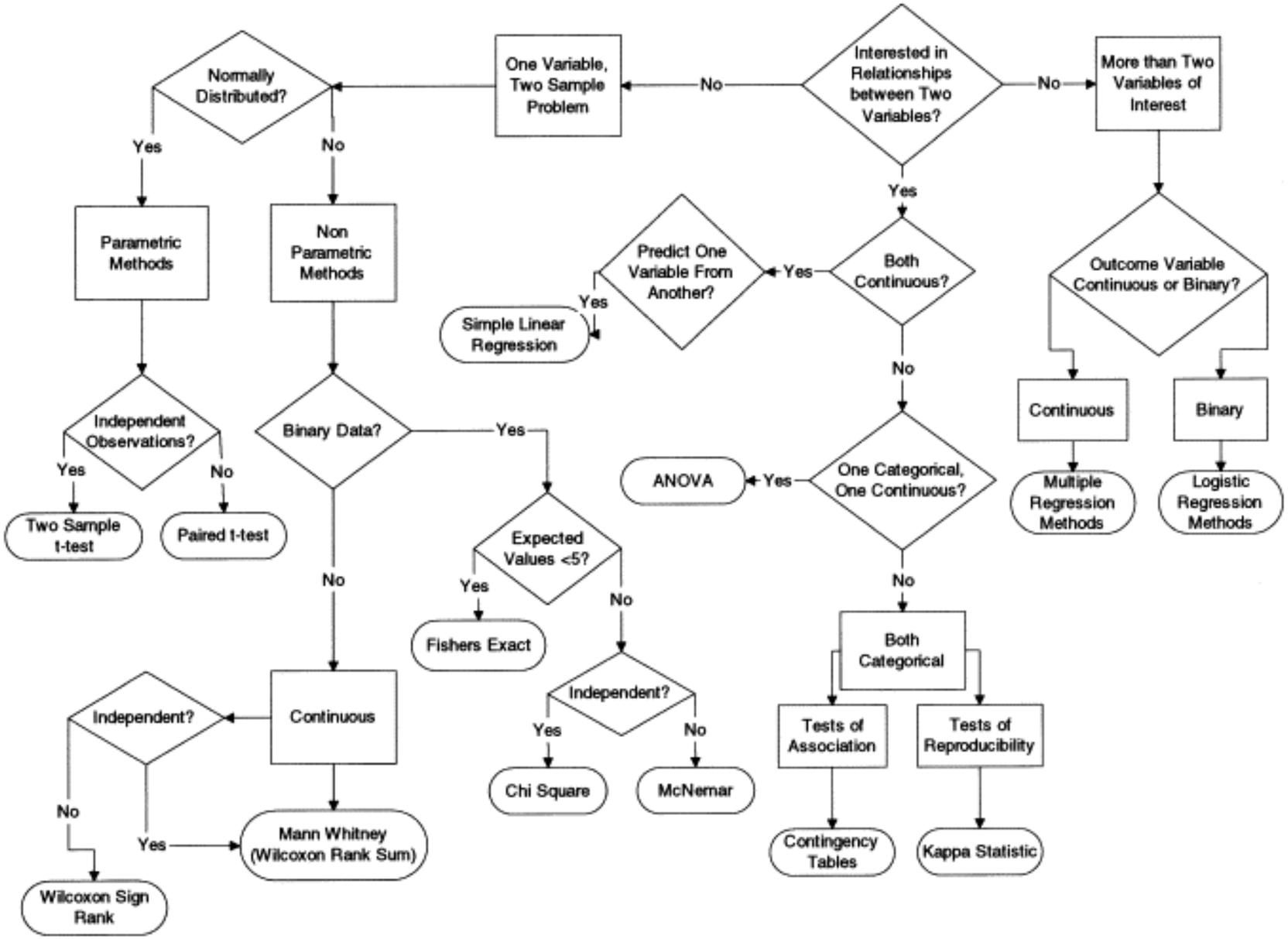


02:00

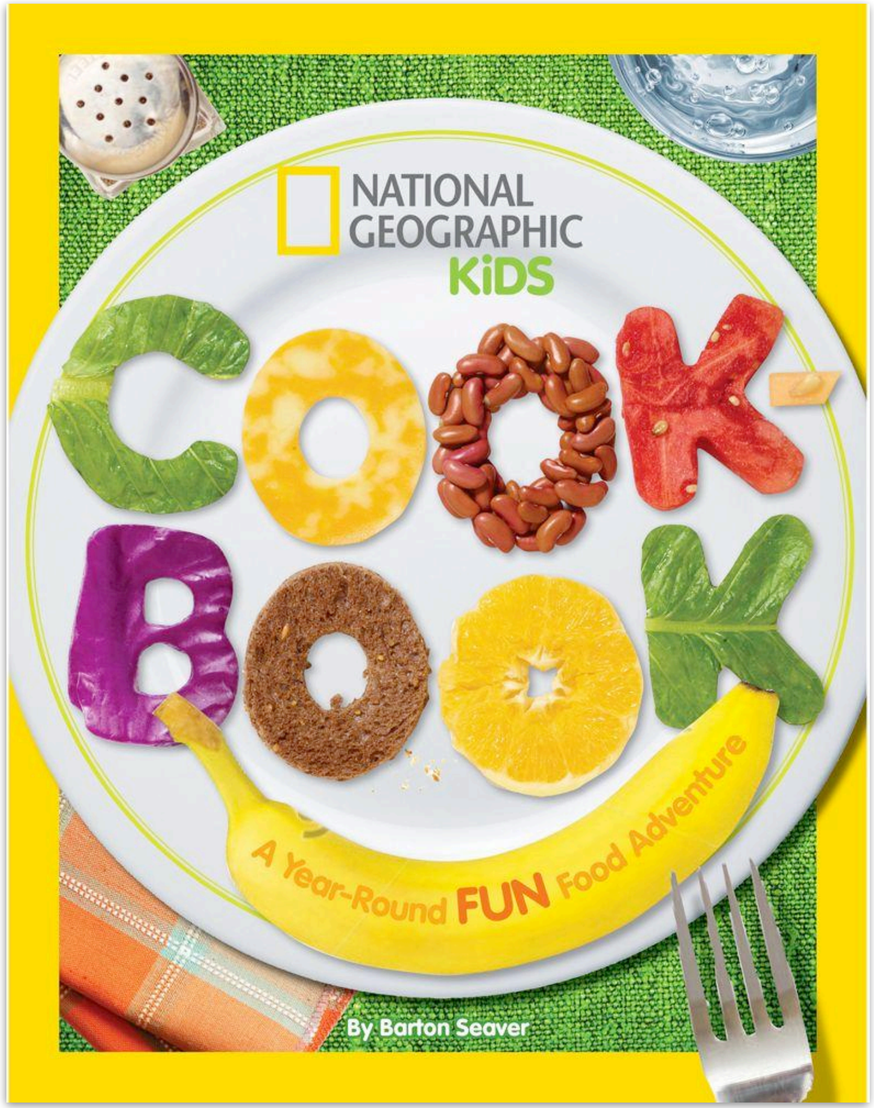
stretch break!

Cookbook vs. Model Comparison

The cookbook approach (decision trees)

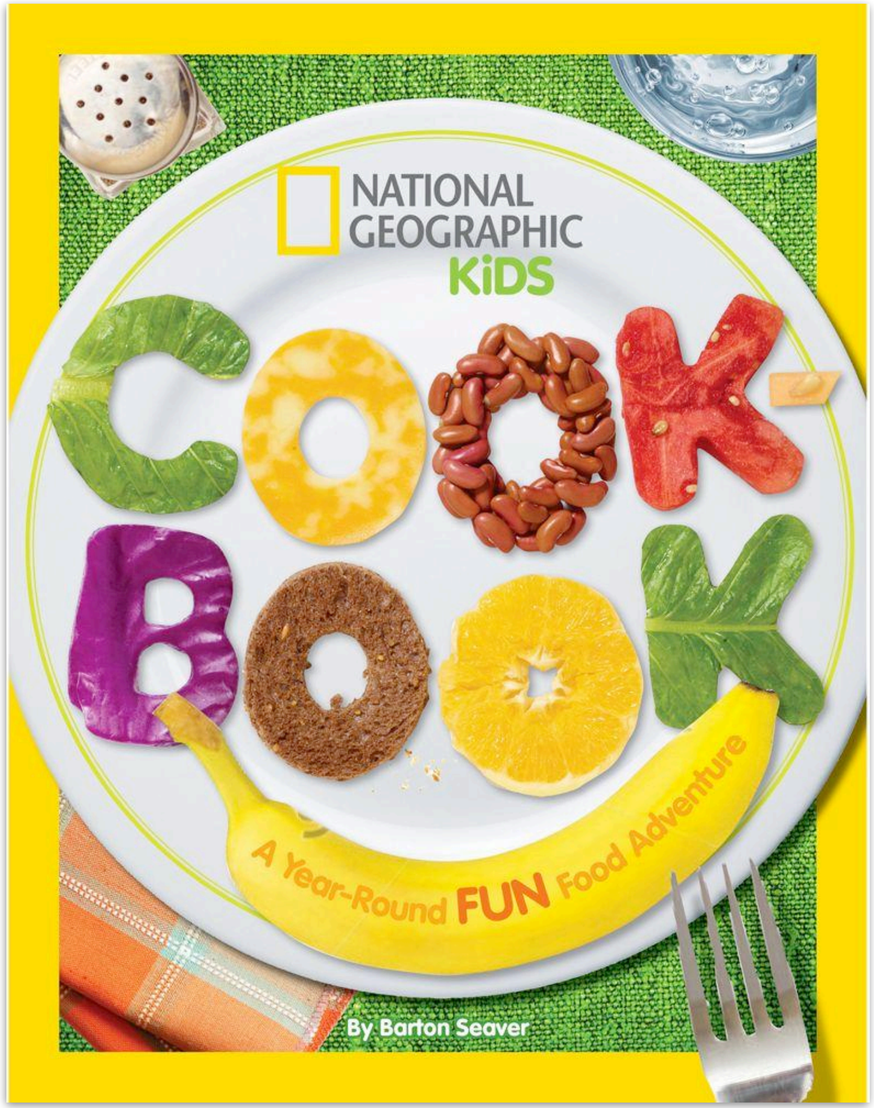


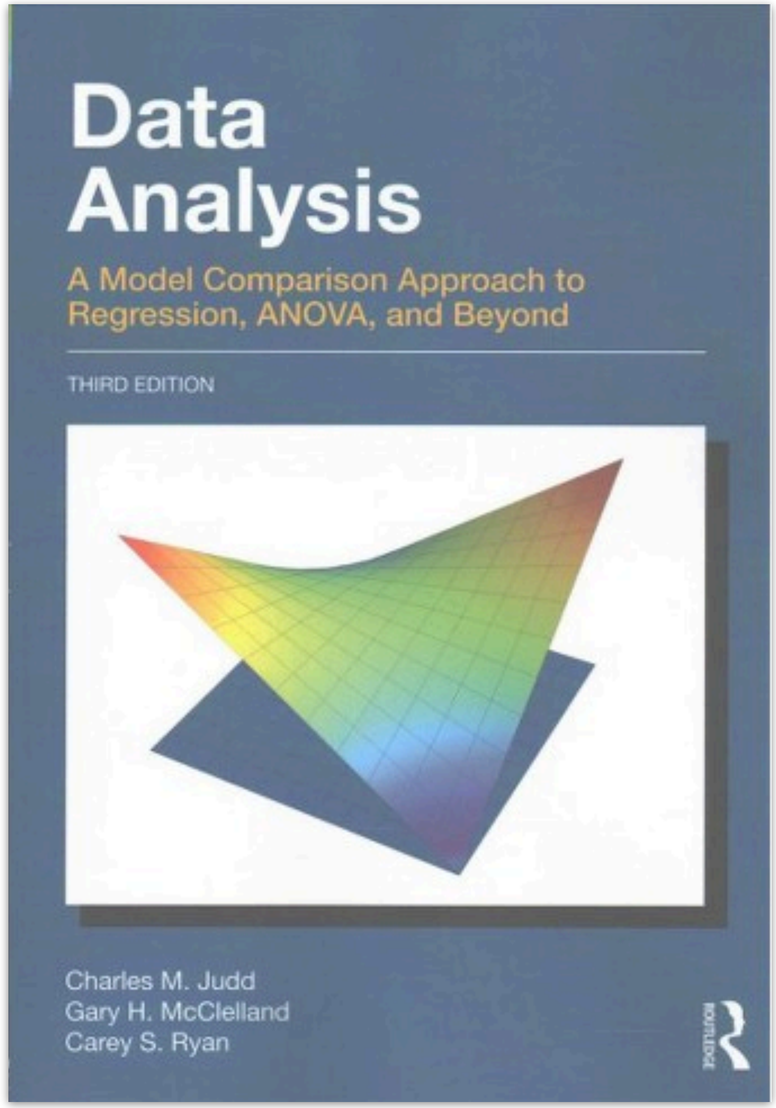
The cookbook approach



- 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

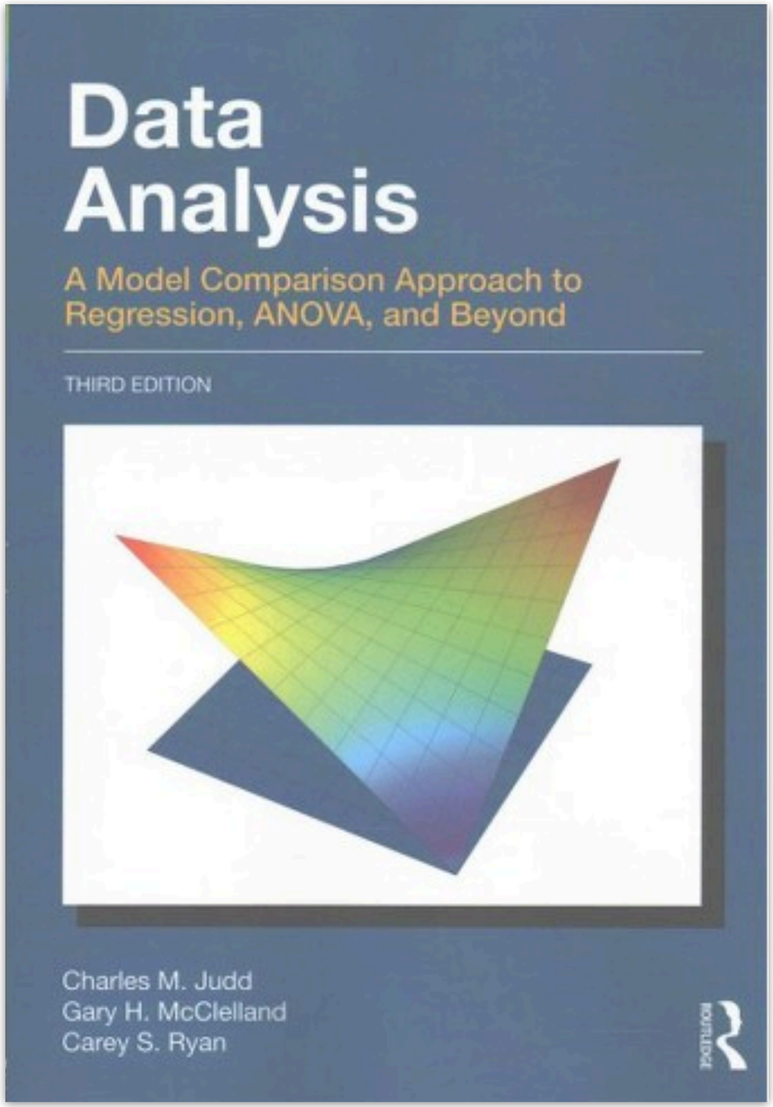




Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

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

$$Error = Data - Model$$

to reduce error we can:

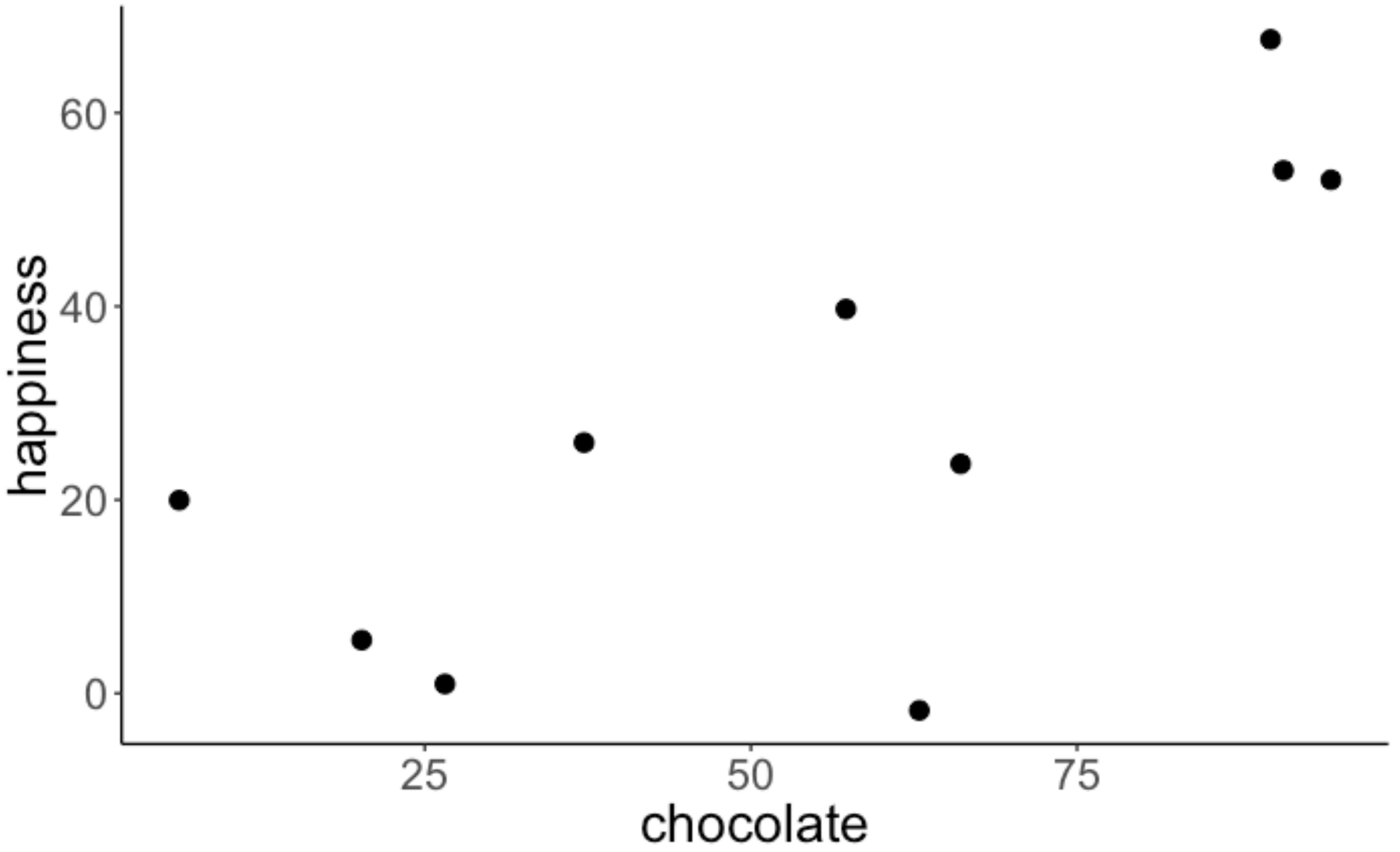
improve the quality of the data

e.g. run good experiments

improve the model

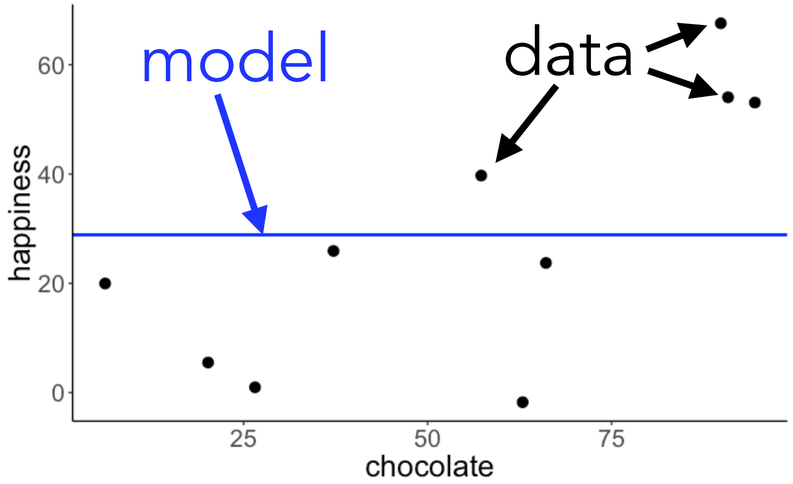
e.g. make predictions conditional on additional information

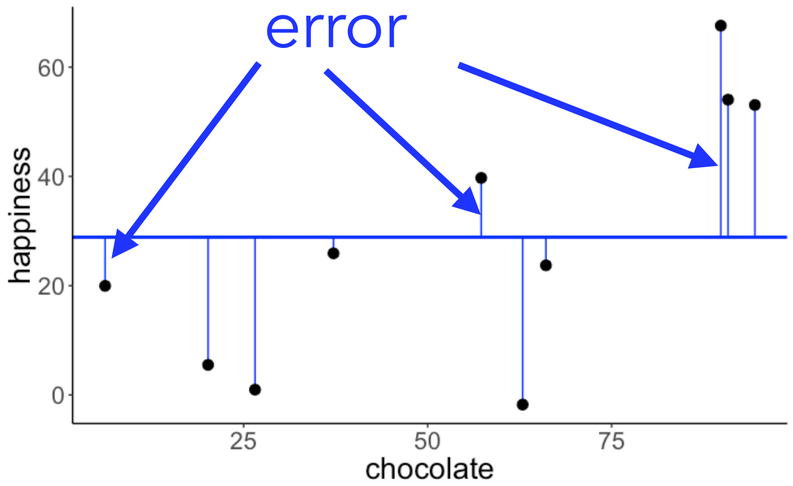
Is there a relationship between chocolate consumption and happiness?



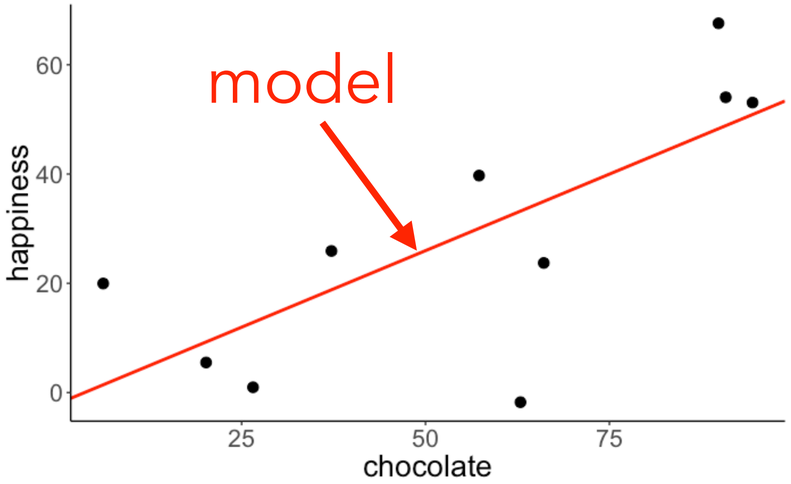
Data = Model + Error

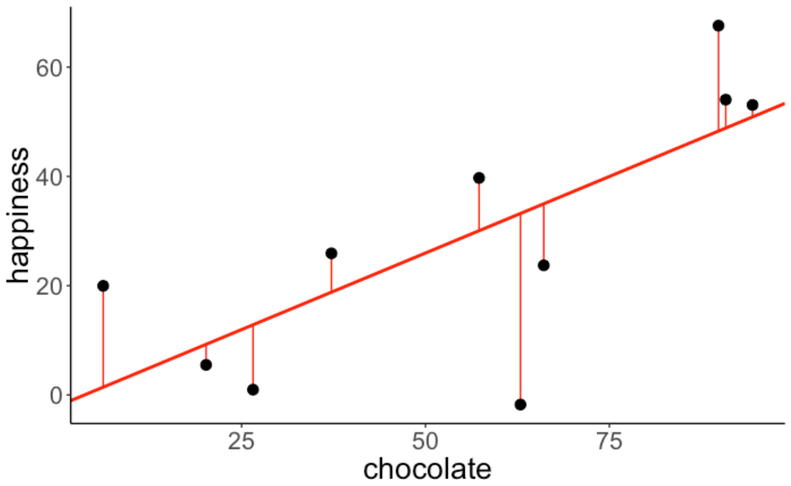
H_0: Chocolate consumption and happiness are unrelated.





H_1: Chocolate consumption and happiness are related.





ERROR

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

how shall we define this?

ERROR

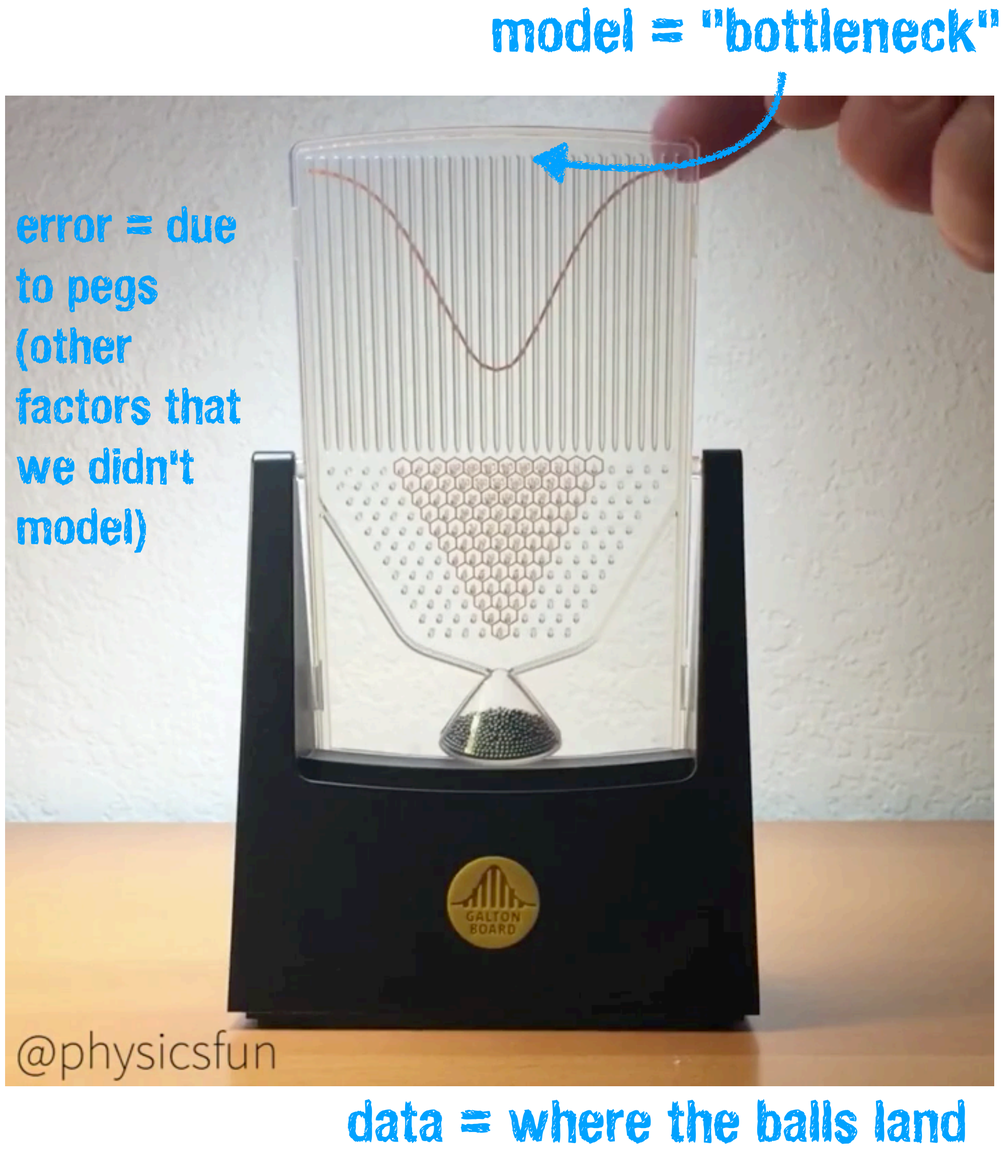
1. We assume that the error between model and data is due to (a potentially large number of) factors that we didn't take into account.

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).



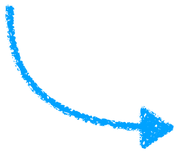
Result: normal distribution

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

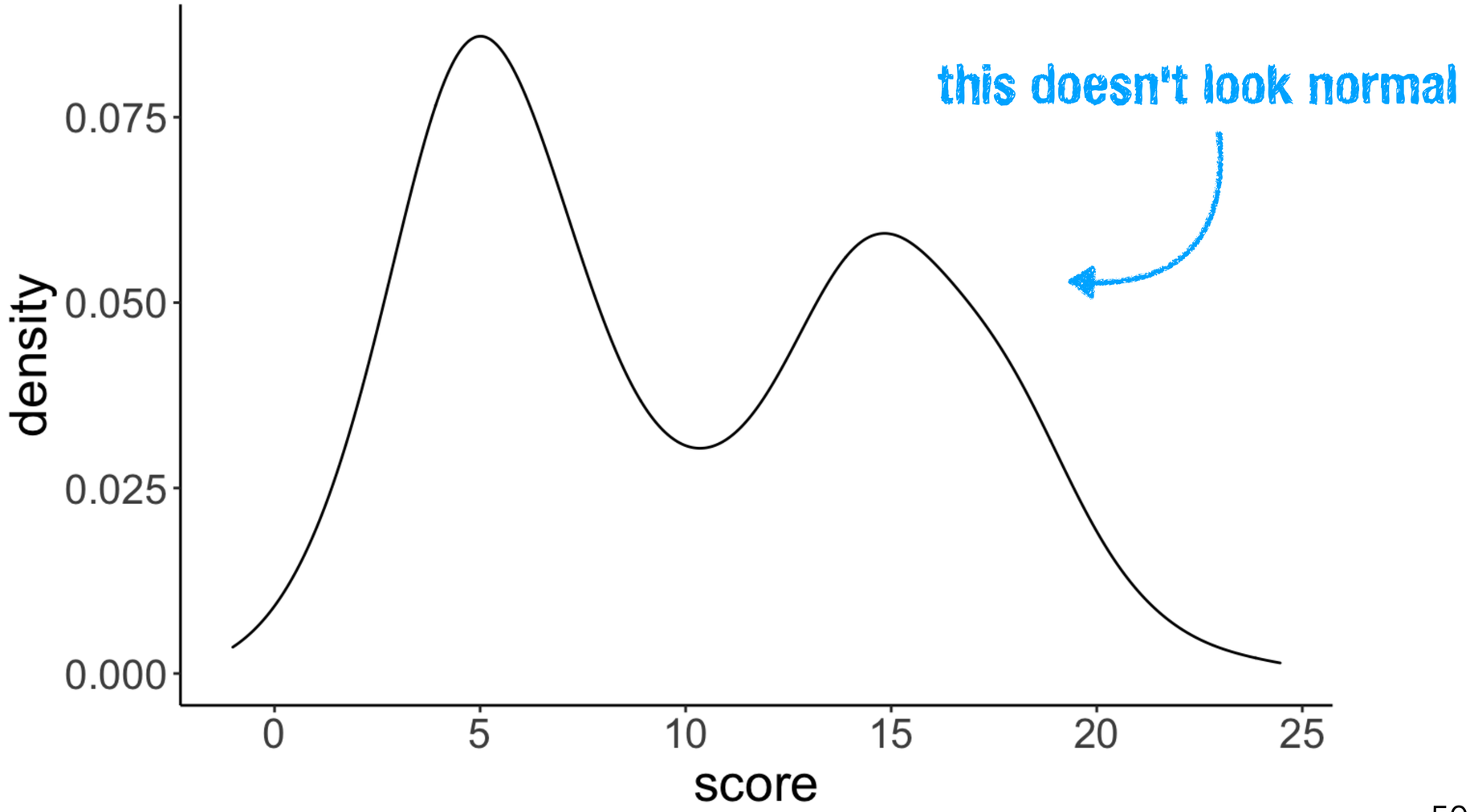
$$\mathrm{Error} = \mathrm{Data} - \mathrm{Model}$$

assumed to be normally distributed

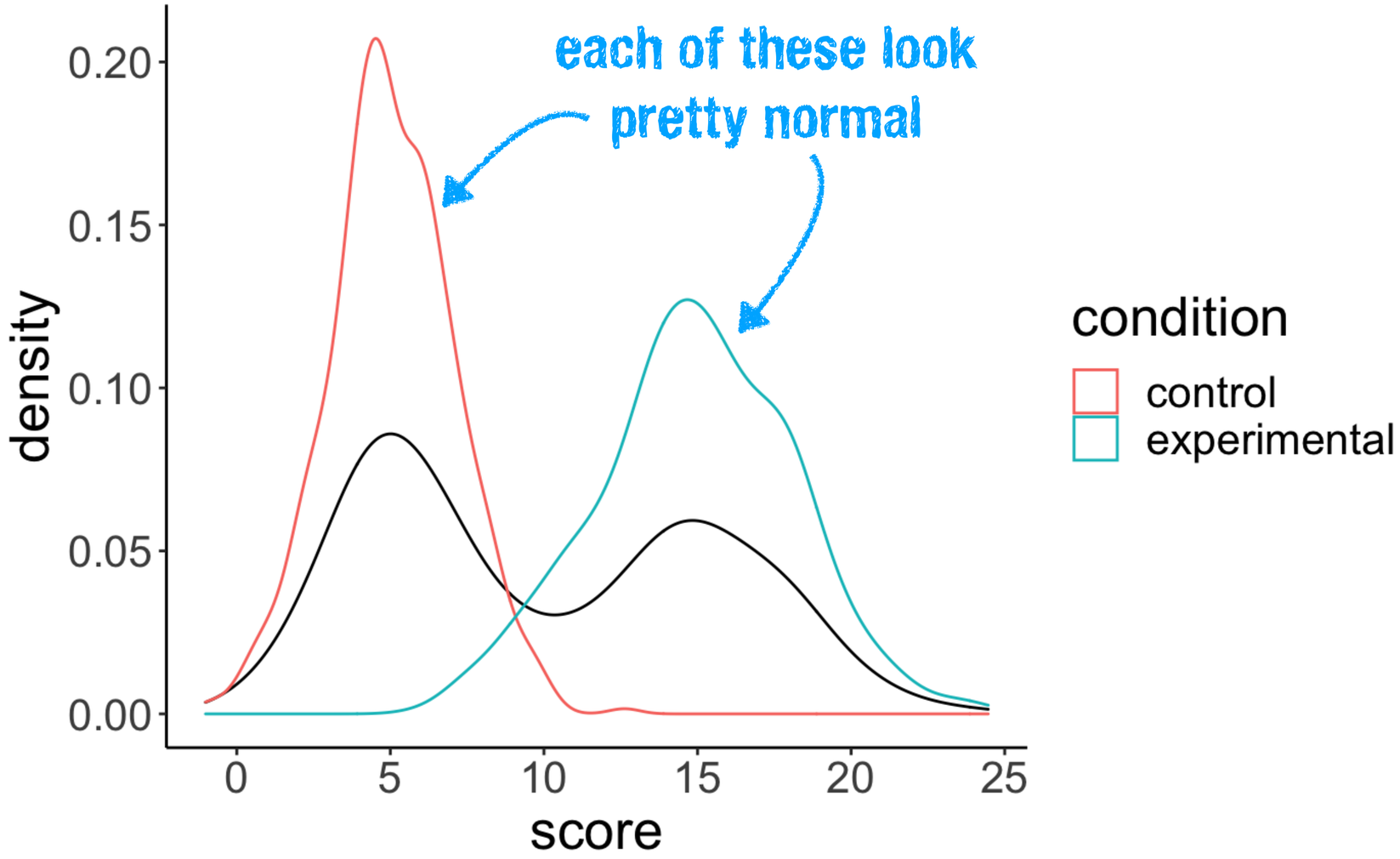
don't need to be normally distributed!!

very common misconception!!!

Distribution of **test scores**

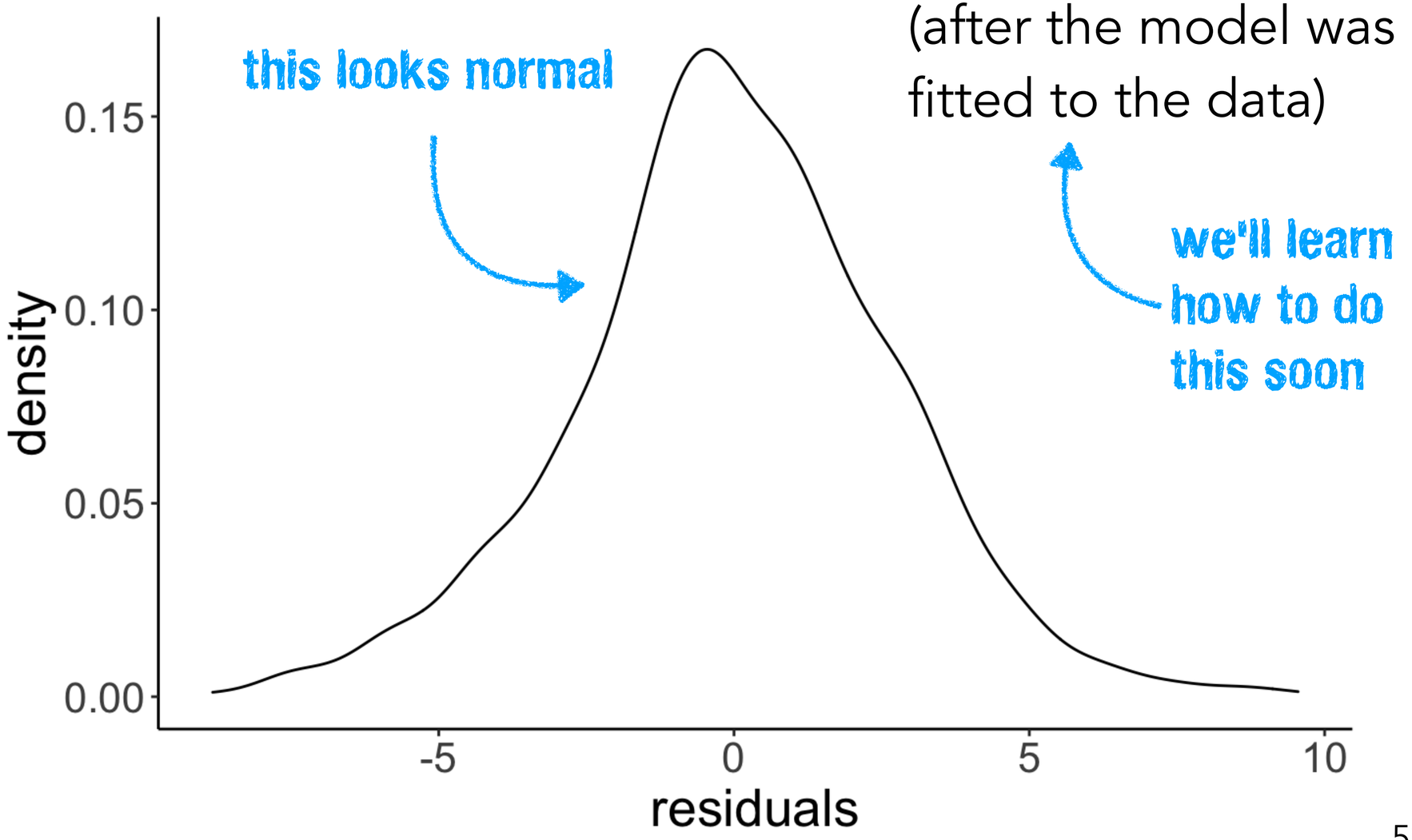


Distribution of **test scores**



Distribution of the **residuals**

(after the model was fitted to the data)

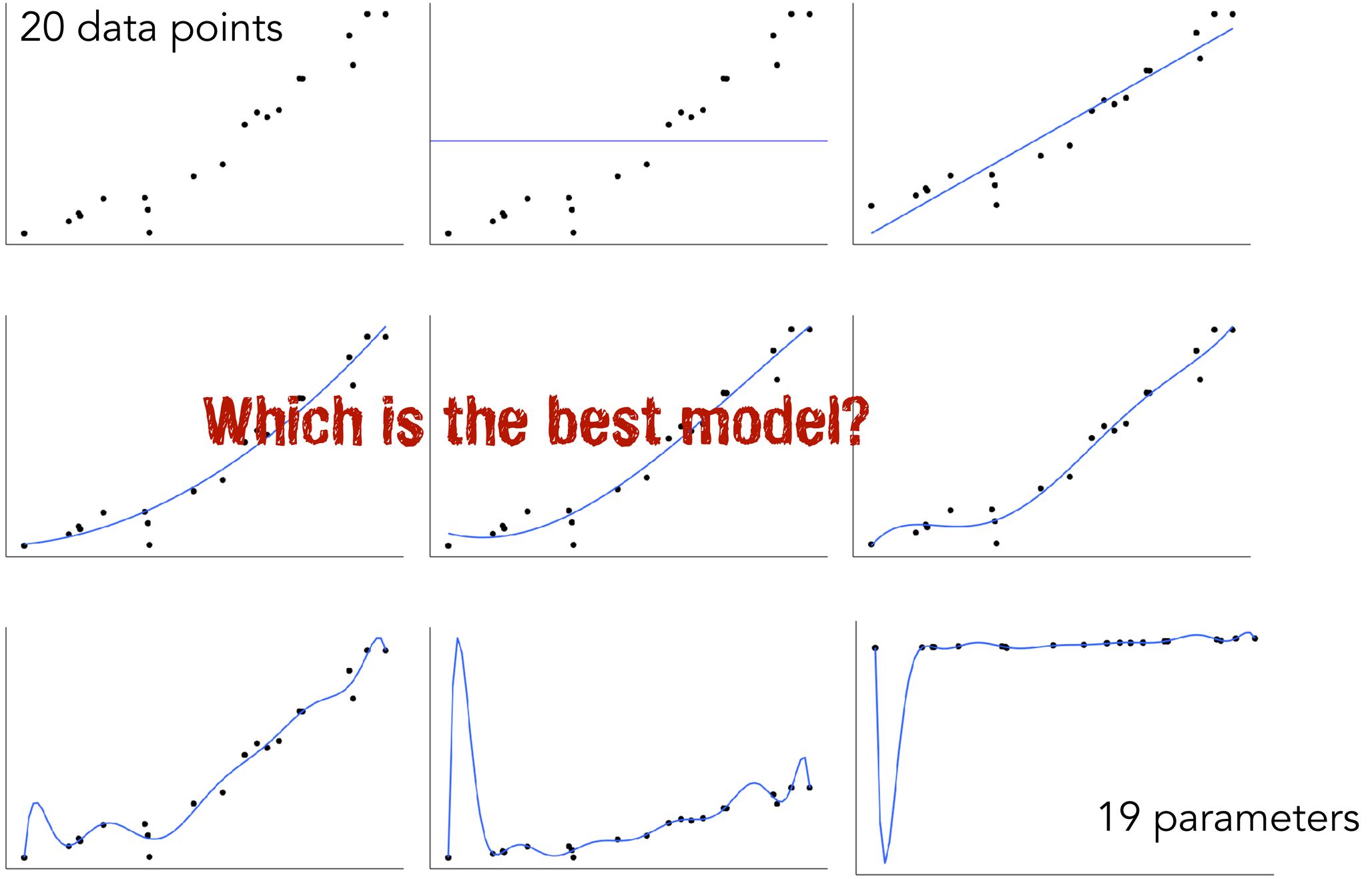


Data = Model + Error

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

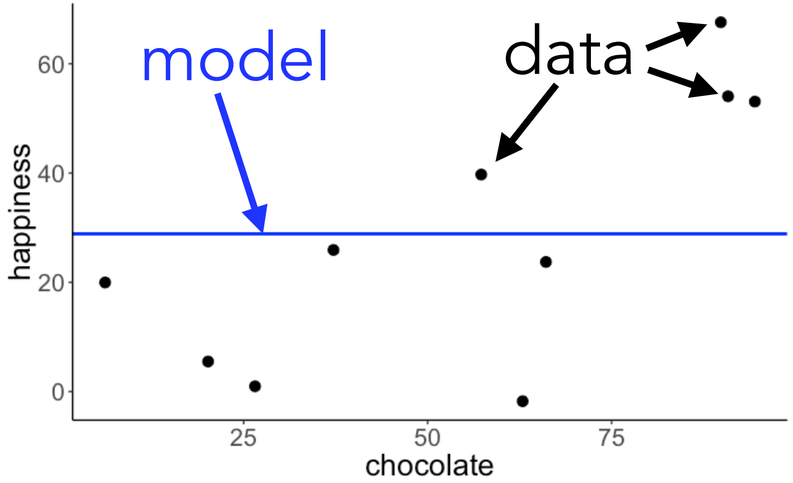

20 data points
Which is the best model?
19 parameters

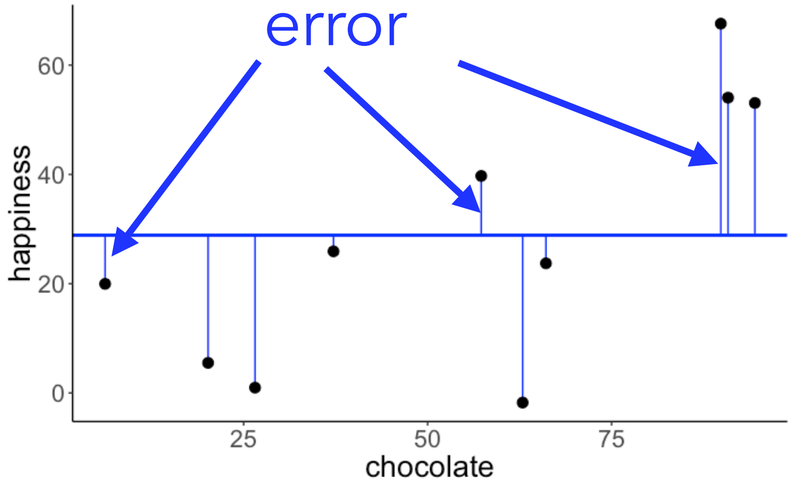
General Procedure

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**

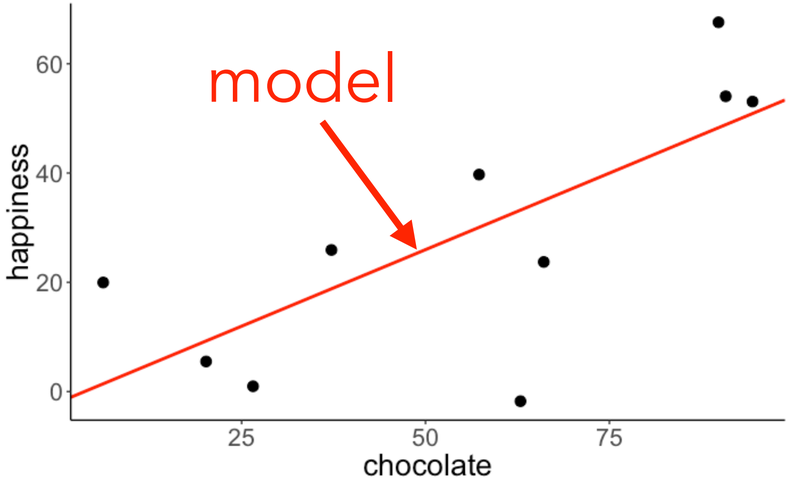
Data = Model + Error

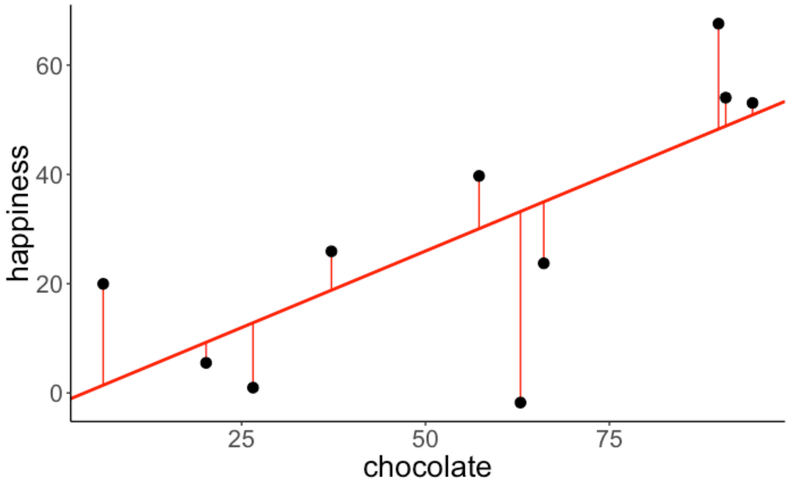
H_0: Chocolate consumption and happiness are unrelated.





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

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

worth it?

fit to the data

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

worth it?

additional predictor

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

fit to the data

Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

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

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

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

Augmented model

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
imgflip.com

worth it?

Compact model

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

Augmented model

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

Proportional reduction in error (PRE)

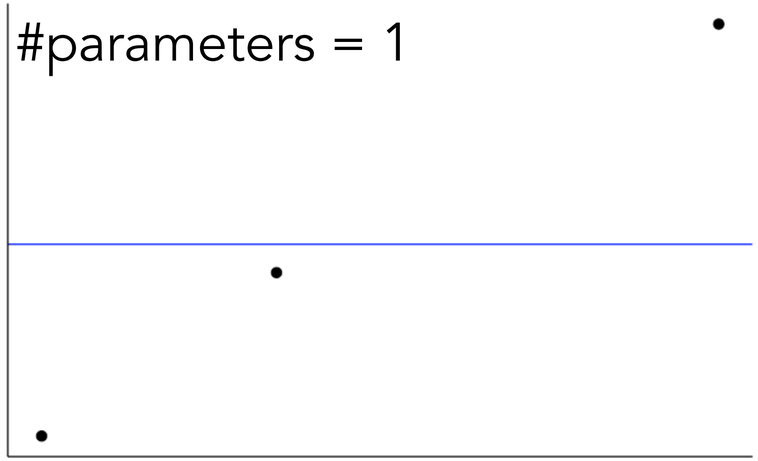
$$\text{PRE} = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

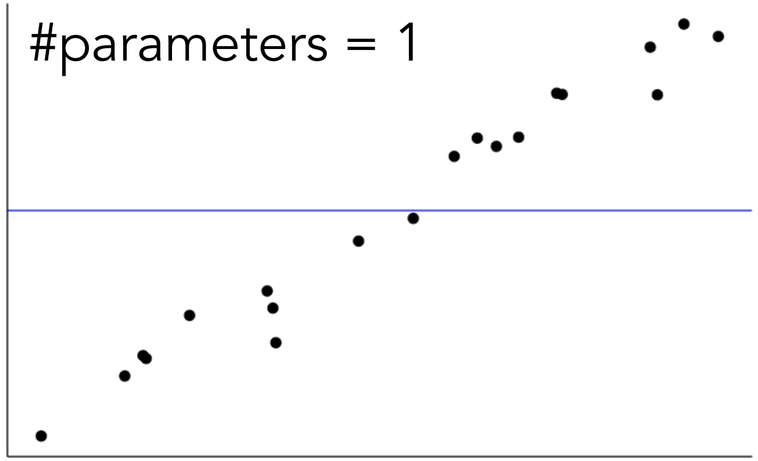
$$= 1 - \frac{30}{50} = .40$$

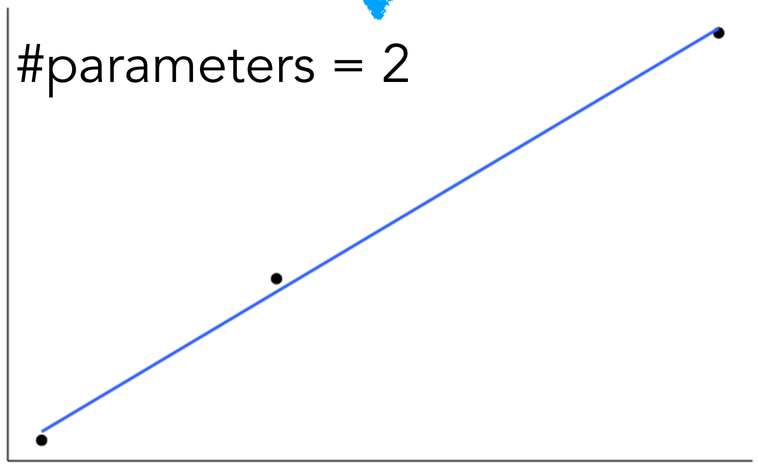
- to answer the **worth it?** question we need inferential statistics (define ERROR, sampling distributions, etc.)
- more likely to be **worth it** if:
  1. **PRE** is high
  2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
  3. the number of parameters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

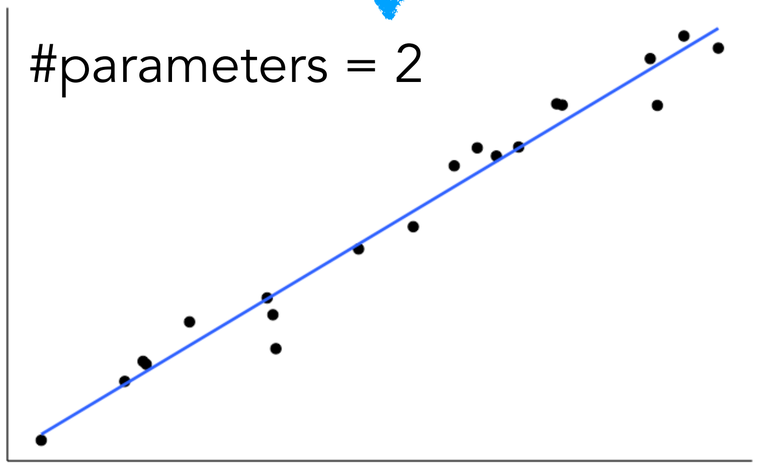
more impressed if the number of observations n is much greater than the number of parameters

PRE per parameter for different *n*









General procedure

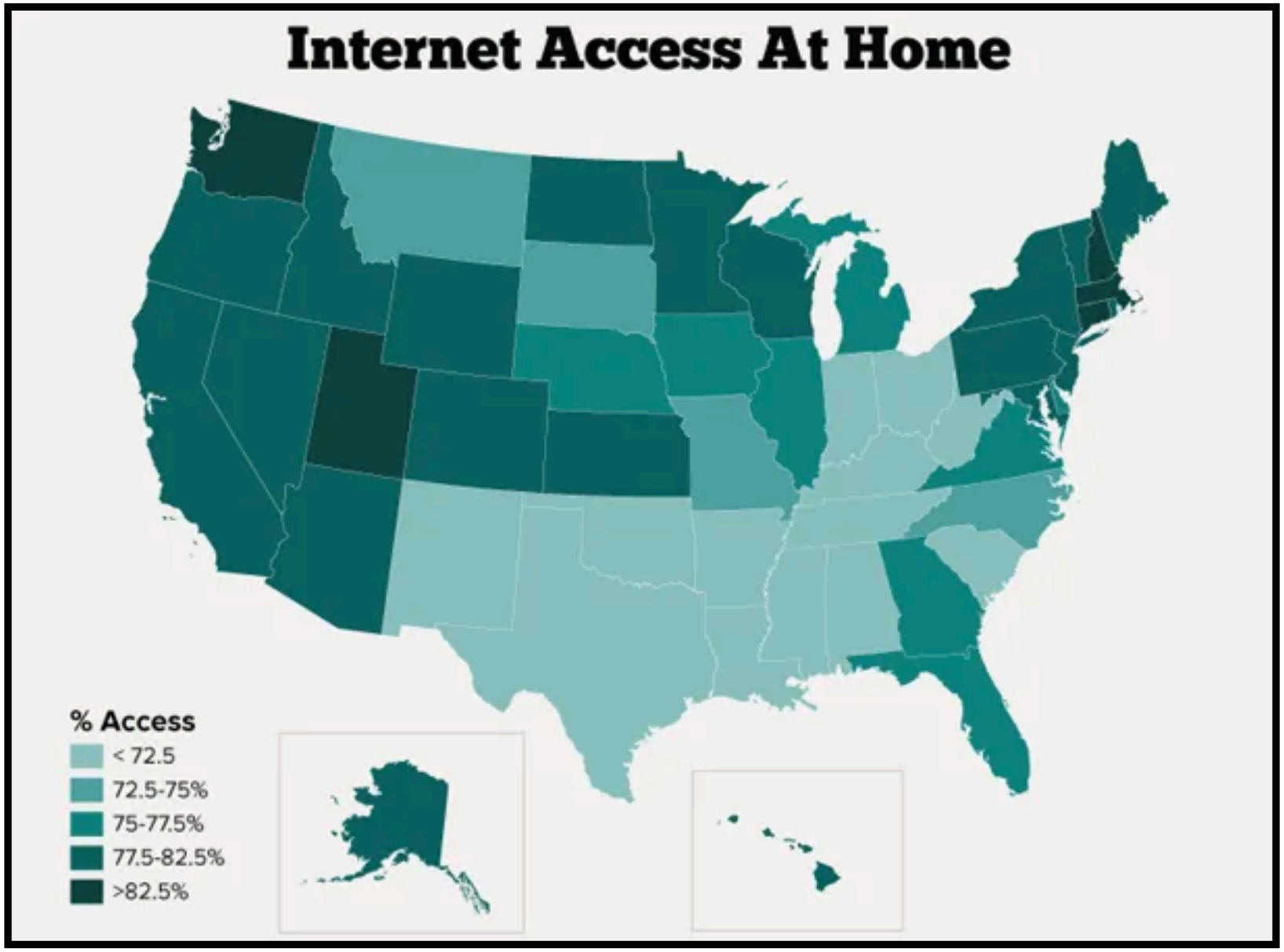
- for any question we want to ask about our DATA
  - we define $model_C$ and $model_A$
  - compare the models using PRE
  - determine whether PRE is **worth it**
- in standard frequentist lingo:
  - $model_C$ = H_0 (null hypothesis)
  - $model_A$ = H_1 (alternative hypothesis)

- hypothesis test:
  - H_0: **all** the parameters that are included in $model_A$ but not in $model_C$ are 0
  - H_1: **not all** the parameters that are included in $model_A$ but not in $model_C$ are 0

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**



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

Notation

DATA = MODEL + ERROR

$$Y_i = \beta_0 + \epsilon_i$$ simple model (true parameters)

$$Y_i = b_0 + e_i$$ simple model (estimated parameters)

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

density

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

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 β or ε 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$_C$: $Y_i = B_0 + \epsilon_i$

0 parameters

$$Y_i = 75 + e_i$$

Model$_A$: $Y_i = \beta_0 + \epsilon_i$

1 parameter

$$Y_i = b_0 + e_i$$

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

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

C: $Y_i = 75 + e_i$ **A**: $Y_i = \overline{Y} + e_i$

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

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

$$= 1 - \frac{1355}{1595} \approx .15$$

Model A has 15% less error than Model C.

SSE(C) = 1595 SSE(A) = 1355

Decide whether it's **worth it**

- PRE is the estimate of an unknown true reduction of error η^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 $model_C$ to create $model_A$ 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

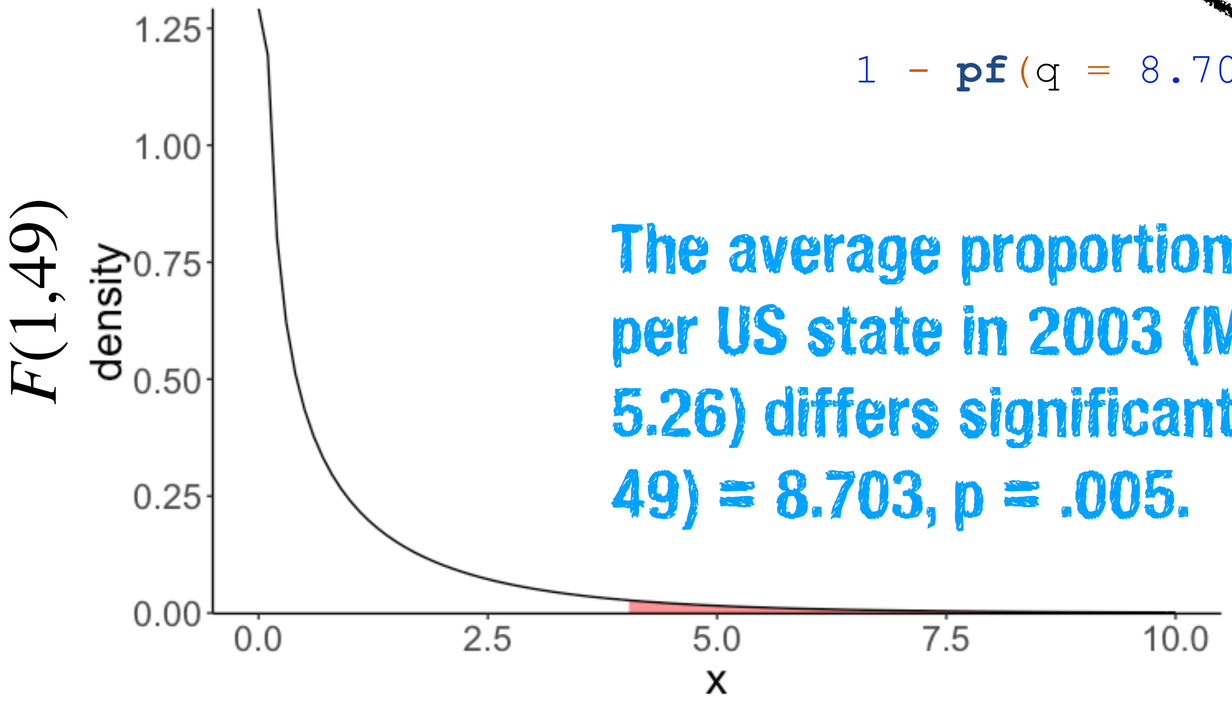
Decide whether it's worth it

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

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

$$= \frac{.15/(1 - 0)}{(1 - .15)/(50 - 1)} = 8.703 \quad p = .00486$$

```
1 - pf(q = 8.703, df1 = 1, df2 = 49)
```



The average proportion of internet access per US state in 2003 (M = 72.8, SD = 5.26) differs significantly from 75%, F(1, 49) = 8.703, p = .005.

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 ...

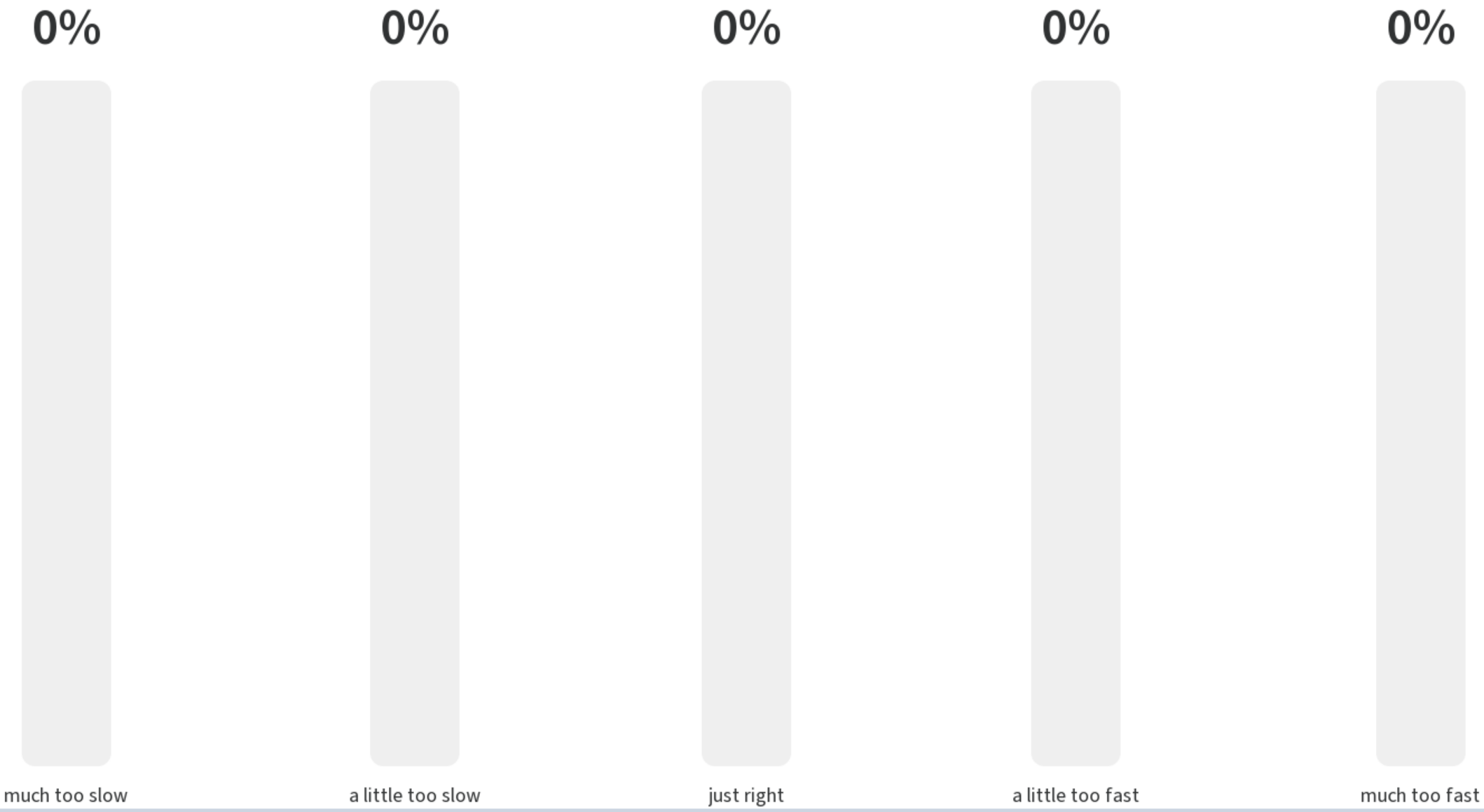
- we will apply the general procedure we went through to day to a large range of different situations
- thinking of hypothesis testing as model comparison is (hopefully more) intuitive
- this approach still works even if we can't find a recipe in our favorite stats cook book

Summary

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

Feedback

0



Thank you!