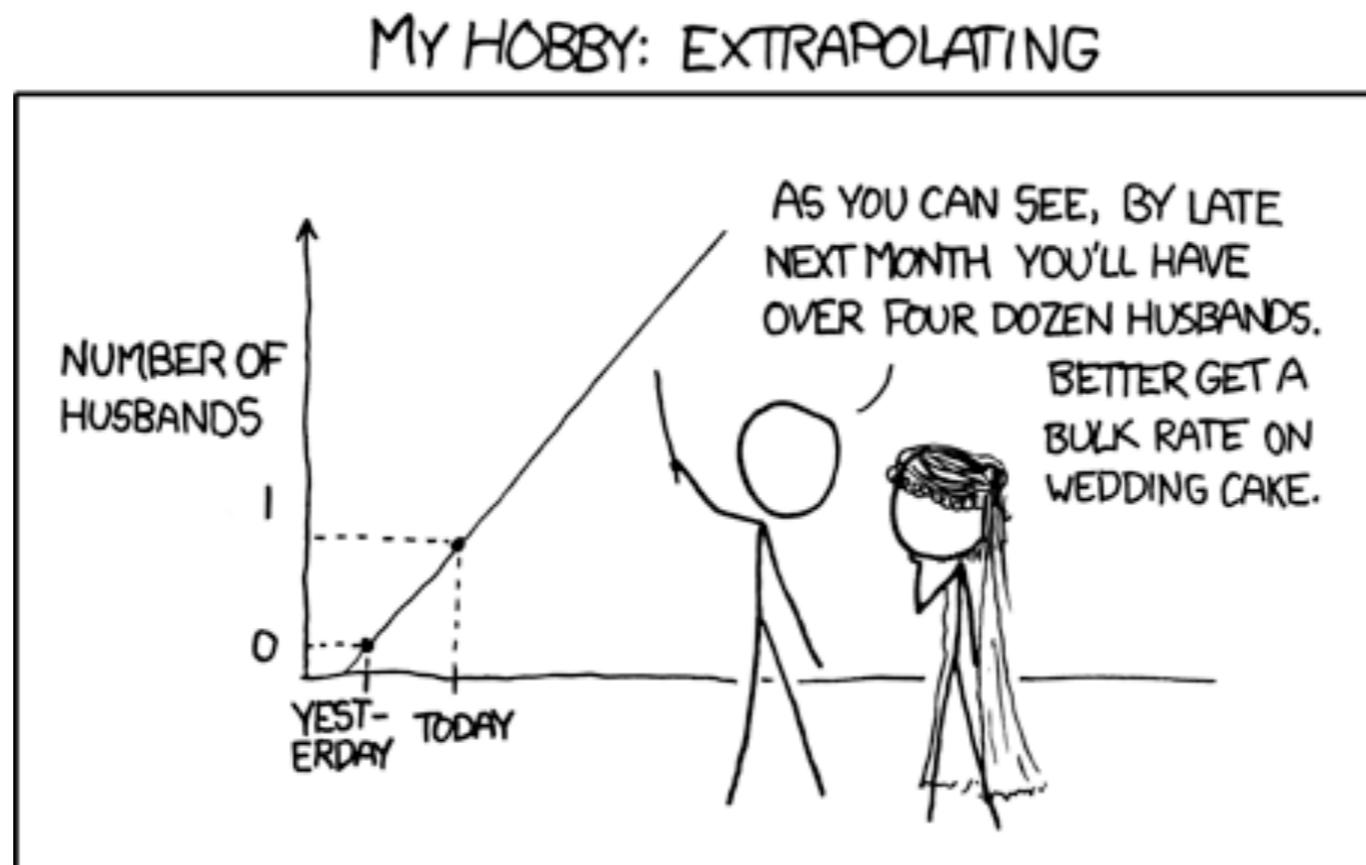


Linear model 1



Logistics

Homework 4

My name goes here

The names of the people I have worked with go here

2022-01-27 21:27:01

3

This homework is due by **Thursday, February 3rd, 8:00pm**. Upload a pdf file to Canvas called `4_modeling_data.pdf`

I. Simple linear regression and prediction

In this section and the next, we'll be revisiting the credit dataset.

```
# Load data
df.credit = read_csv("data/credit.csv") %>%
  clean_names()
```

1. Explore and visualize the data

Use the `ggpairs()` function from the `GGally` package to make scatterplots of all pairwise combinations of the numeric variables.

Tip: use `progress = FALSE` within the function to suppress unwanted progress bars of each plot

```
### YOUR CODE HERE ###
```

```
#####
```

That's perhaps a bit too congested to be useful. Recreate the plot, focusing on just ~5 of the variables that seem interesting to you.

```
### YOUR CODE HERE ###
```

```
#####
```

Does your plot tell you anything interesting about the data? Briefly describe one observation.

YOUR ANSWER HERE

2. Simple linear regression

You decide you want to test the relationship between credit limit (`limit`) and average credit card debt (`balance`).

a) First, let's visualize this relationship. Create a scatterplot of `balance` (on the y-axis) as a function of `limit` (on the x-axis). Set the transparency of the points to 0.3.

```
### YOUR CODE HERE ###
```

II. Multiple linear regression and controls

In psychological research, people often run linear regressions in which the goal is to assess the relationship between two variables while "controlling" for other variables. These control variables could, for example, be age and gender. But how should we decide whether and which variables to control for? In this exercise, we will see what potential effects controlling for variables can have in different situations.

Now you are interested in whether age is a significant predictor of credit limit.

4. Interpreting model parameters

a) Build a simple linear regression model to predict `limit` from `age`. Is age a significant predictor of credit limit?

```
# Simple regression without control variables
### YOUR CODE HERE ###
fit.lm1 =
#####
```

YOUR ANSWER HERE

Then you realize that age is actually related to income, which is a strong predictor of credit limit, so you are interested in seeing whether age is related to credit limit controlling for income.

b) Build a multiple regression model to predict `limit` using both `age` and `income` as predictors. Is age still a significant predictor? What could be an explanation for the change, if there was any?

```
# Multiple regression with control variables
### YOUR CODE HERE ###
fit.lm2 =
fit.lm3 =
#####
```

YOUR ANSWER HERE

III. Interactions

We will be using the following dataset.

`families.csv`:

Data from a study of 68 companies, examining relationships between the quality of family-friendly programs at each company, the percentage of employees with families who use these programs, and employee satisfaction (all continuous variables).

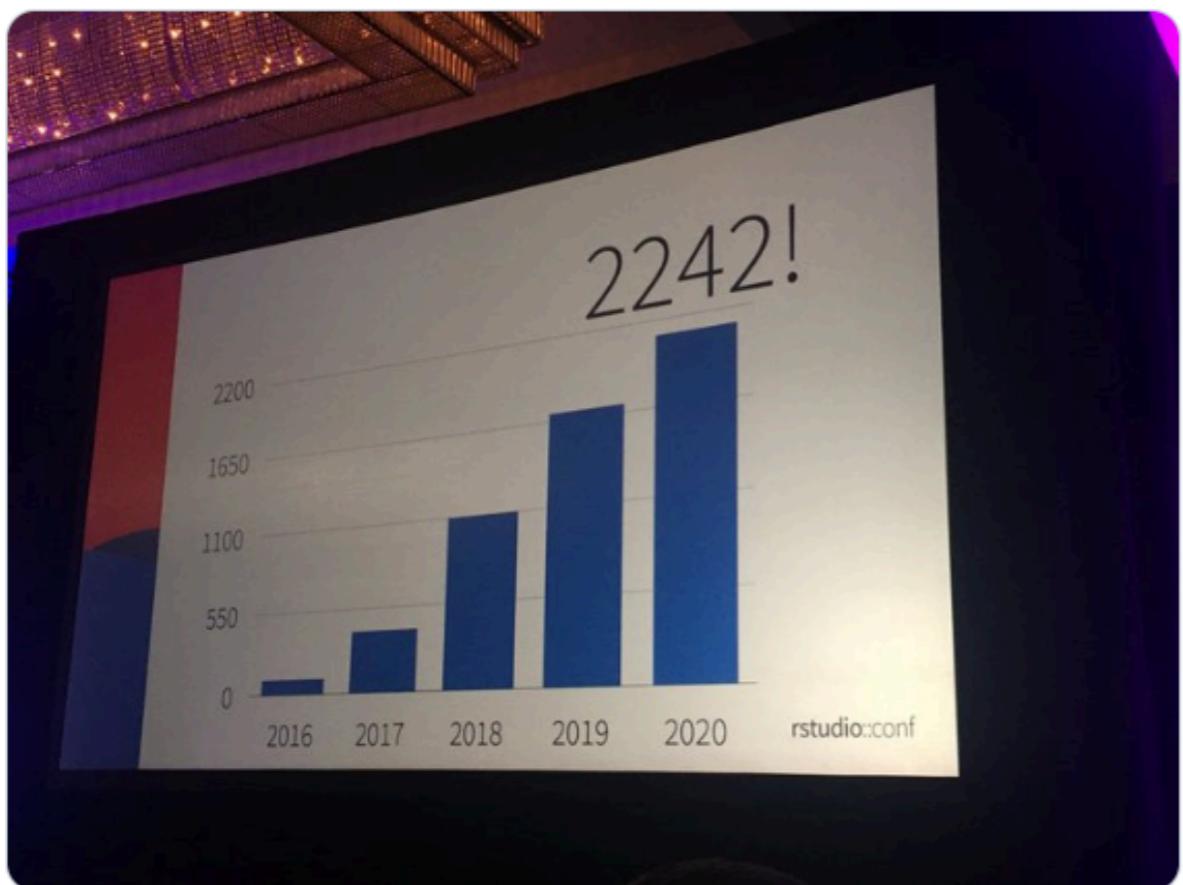
Things that came up



Gabriela de Queiroz @ #rstudioconf
@gdequeiroz

If they say "R is not a popular language and it is not our priority", they are missing out! #rstudioconf2020 has over 2200 attendees this year.

#rstudioconf



RStudio

9:27 AM · Jan 29, 2020 · Twitter for iPhone

29 Retweets 121 Likes



Hadley Wickham ✅
@hadleywickham

The utter shame of presenting a keynote plot where the bar is the wrong height and the axis labels are so irregular 😭 😳 😳



Gabriela de Queiroz @ #rstudioconf @gdequeiroz · 3h

If they say "R is not a popular language and it is not our priority", they are missing out! #rstudioconf2020 has over 2200 attendees this year.
#rstudioconf



9:30 AM · Jan 29, 2020 · Tweetbot for iOS

29 Retweets 435 Likes



Daniel Litt

@littmath

...

You are given an urn containing 100 balls; n of them are red, and $100-n$ are green, where n is chosen uniformly at random in $[0, 100]$. You take a random ball out of the urn—it's red—and discard it. The next ball you pick (out of the 99 remaining) is:

XX

8:50 AM · Jan 28, 2024 · 489.2K Views

291

202

827

702

↑

More likely to be red

0%

More likely to be green

0%

Equally likely

0%



Daniel Litt

@littmath

...

You are given an urn containing 100 balls; n of them are red, and $100-n$ are green, where n is chosen uniformly at random in $[0, 100]$. You take a random ball out of the urn—it's red—and discard it. The next ball you pick (out of the 99 remaining) is:

XX

8:50 AM · Jan 28, 2024 · 489.2K Views

291

202

827

702

↑

guy whose whole personality is duckdb 
@tjmahr

...

```
f <- function(n = 100) {  
    n_red <- sample.int(n, 1)  
    p1 <- n_red / n  
    p2 <- (n_red - 1) / (n - 1)  
    draw <- sample(c("r", "g"), 1, prob = c(p1, 1 - p1))  
    if (draw == "g") {  
        "skip"  
    } else {  
        sample(c("r", "g"), 1, prob = c(p2, 1 - p2))  
    }  
}  
counts <- replicate(100000, f()) |> table()  
counts  
#>  
#>     g      r  skip  
#> 16843 33661 49496  
  
# p(color | first red)  
counts[1:2] |> proportions()  
#>  
#>          g            r  
#> 0.3334983 0.6665017
```

ALT

Created on 2024-01-29 with [reprex v2.1.0](#)

Your feedback

Your feedback

Sometimes I think the recap at the beginning of the class can be shortened so we have more time for the new content!

Plan for today

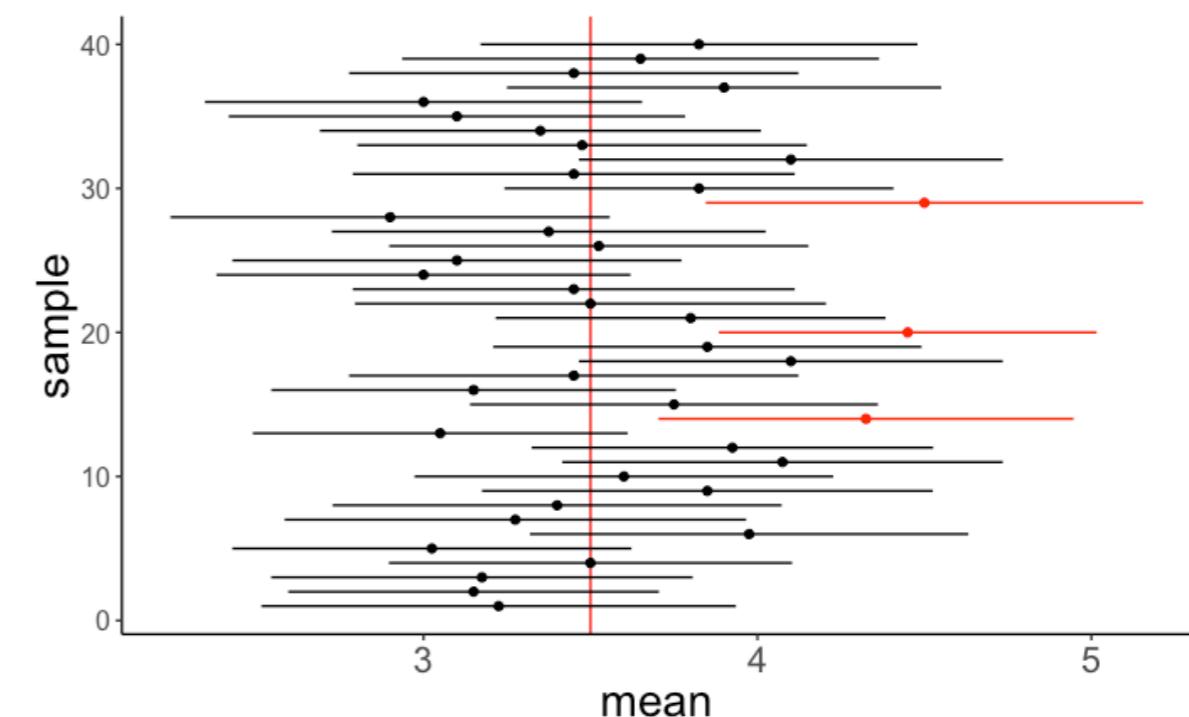
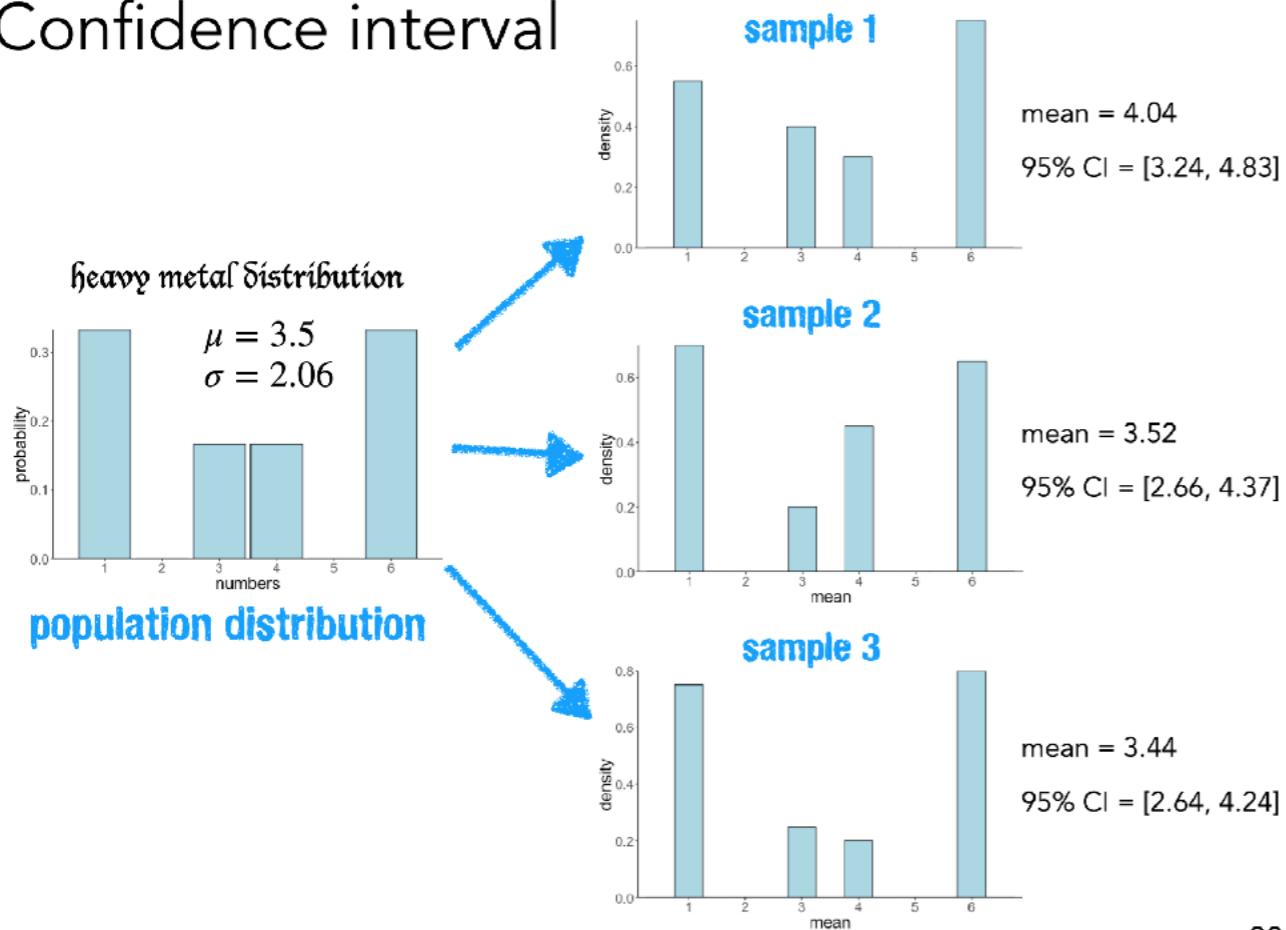
- Quick recap
- Modeling data
- Hypothesis testing as model comparison
- Correlation
 - Pearson's moment correlation
 - Spearman's rank correlation
- Regression

Quick recap

Quick recap: Confidence intervals

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

Confidence interval

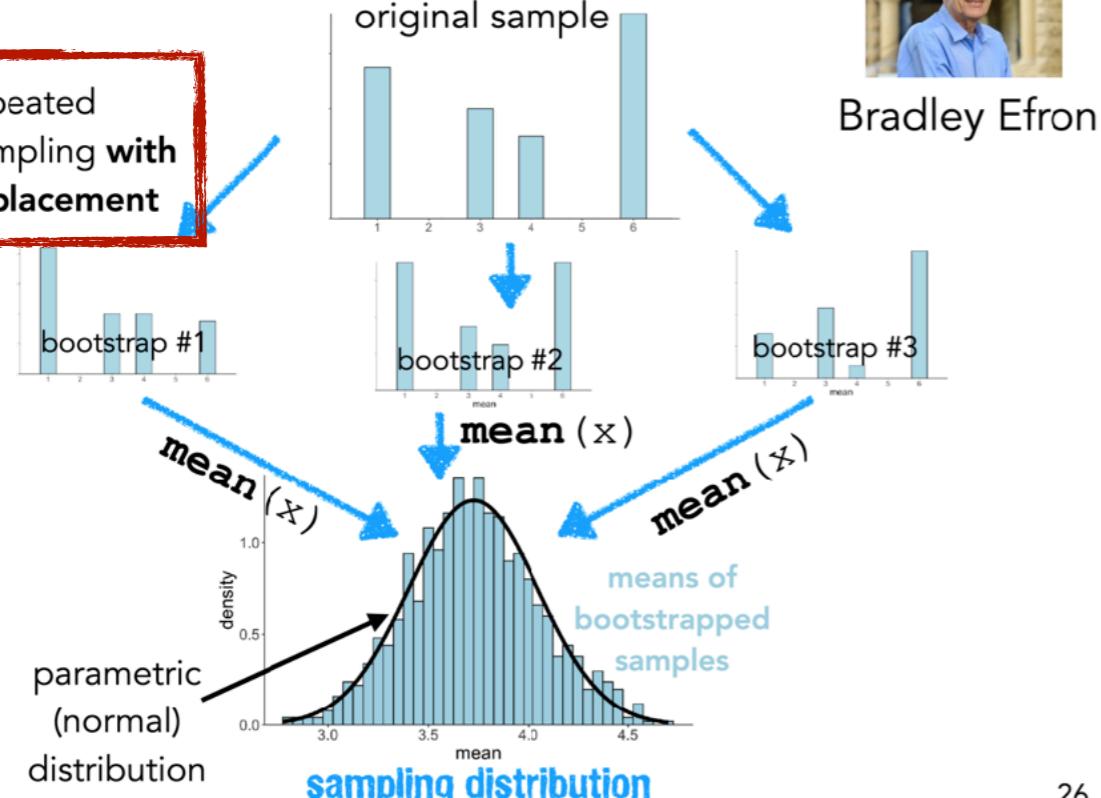


20

Quick recap: Bootstrapping

Bootstrap

repeated sampling **with replacement**



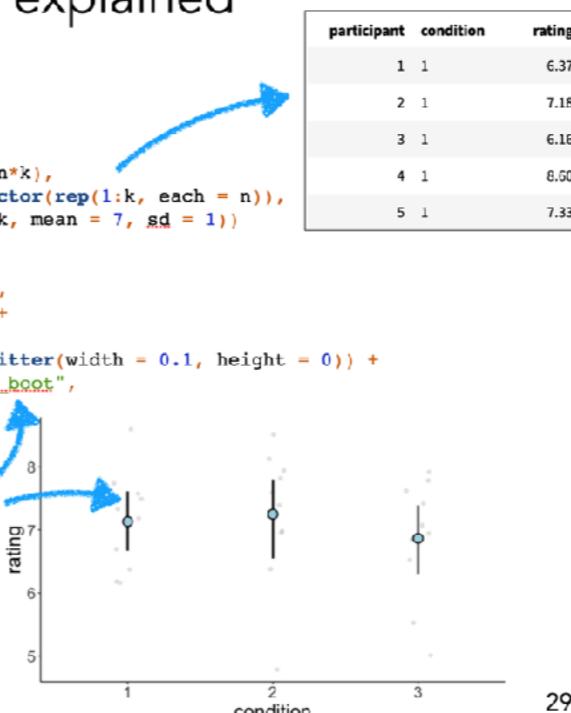
26

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?



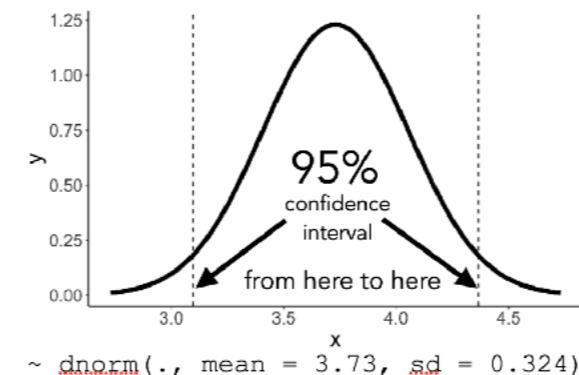
29

Bootstrap

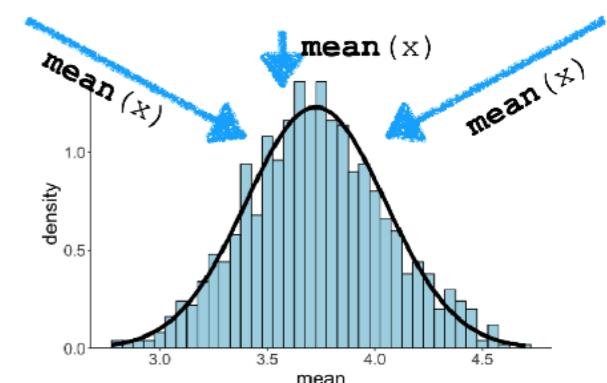
How can I get the confidence interval of a statistical estimate (such as the mean)?

make assumptions

sampling distribution of the mean



bootstrap



27

15

Quick recap: Modeling data

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

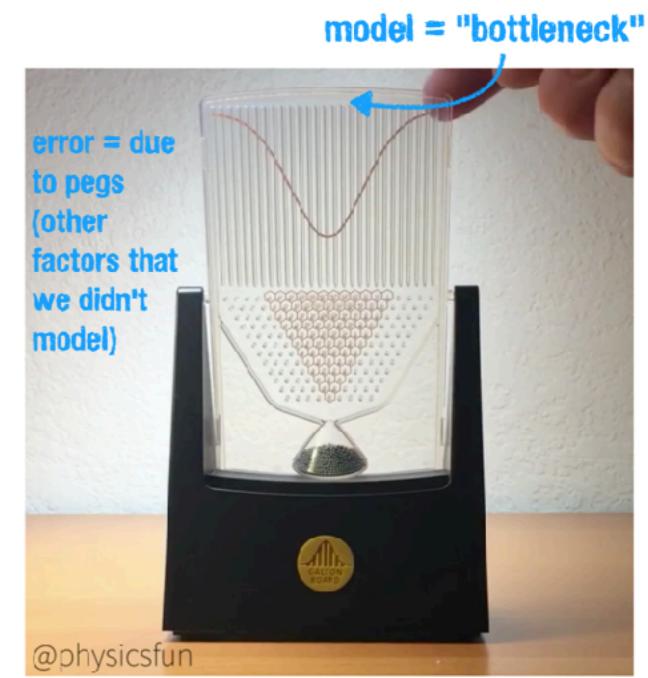
↑
**what makes for
a good model?**

- we build models with parameters, and fit 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**

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

46

Assumption of normal distribution

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

↑
**assumed to be
normally
distributed**

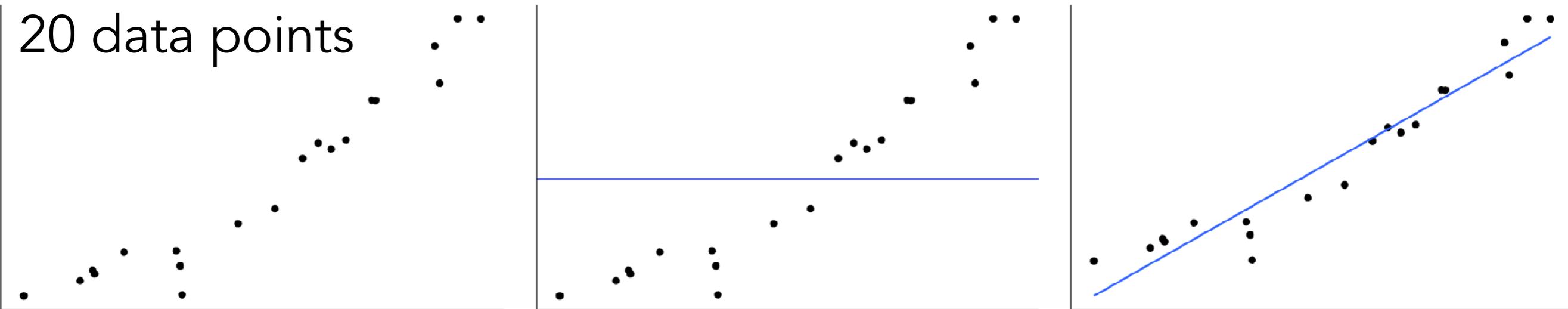
↑
**don't need to
be normally
distributed!!**

very common misconception!!!

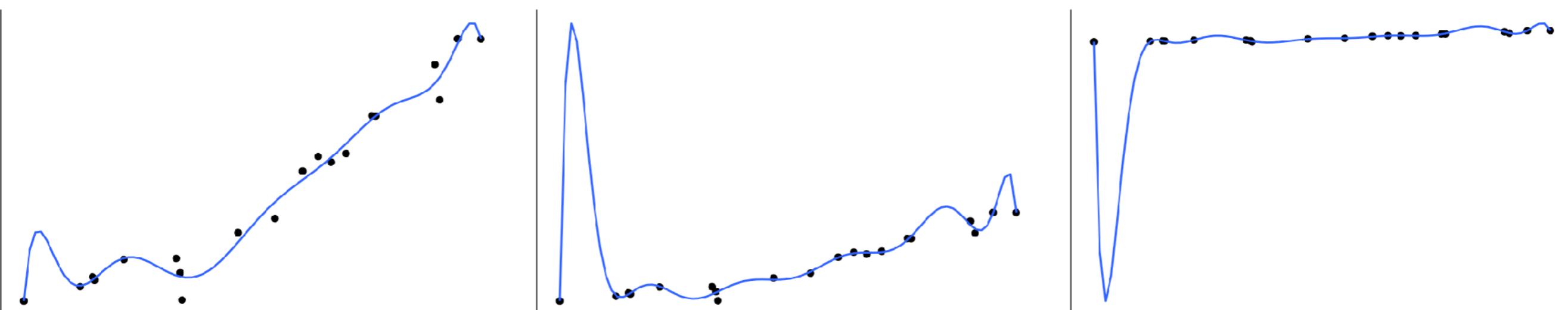
16

Modeling data

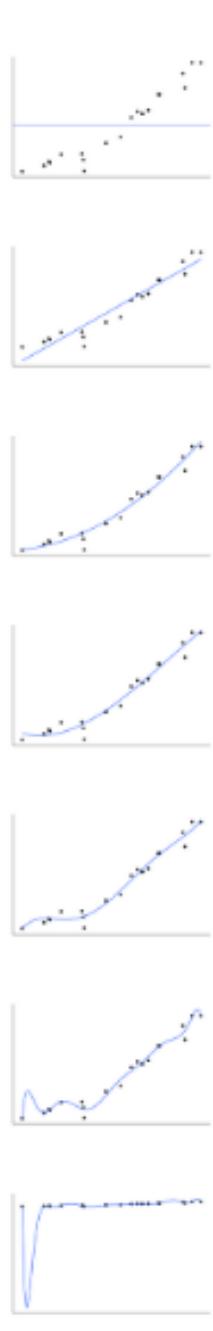
20 data points



Which model describes the data best?



Which model describes the data best





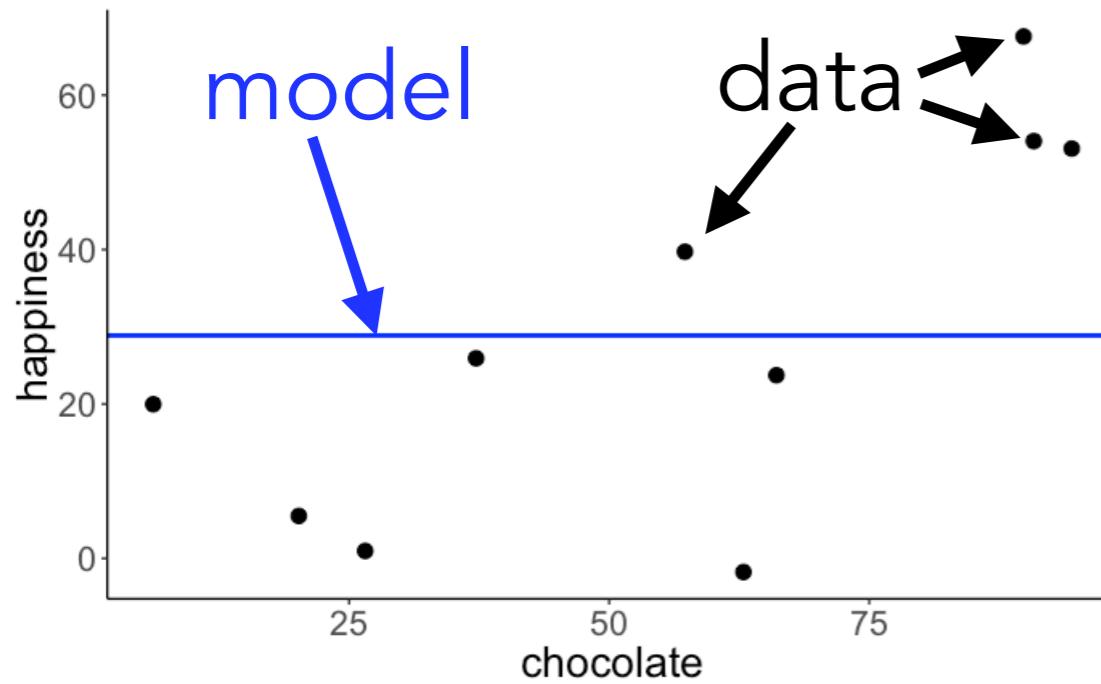
imgflip.com

ADDTXT.COM

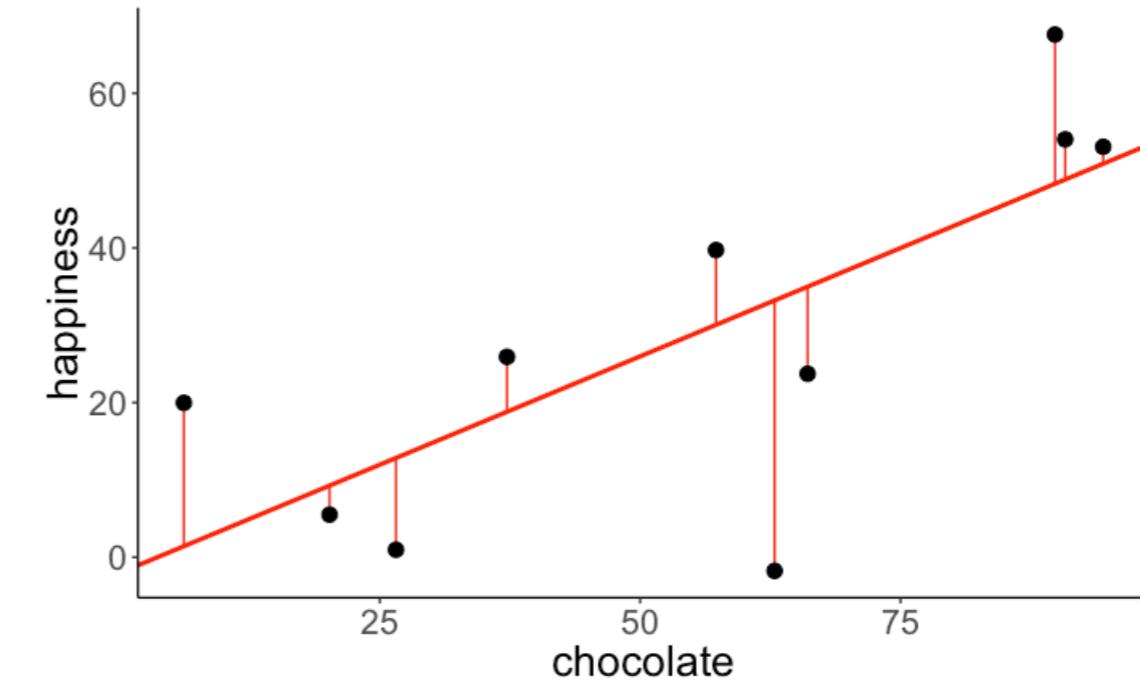
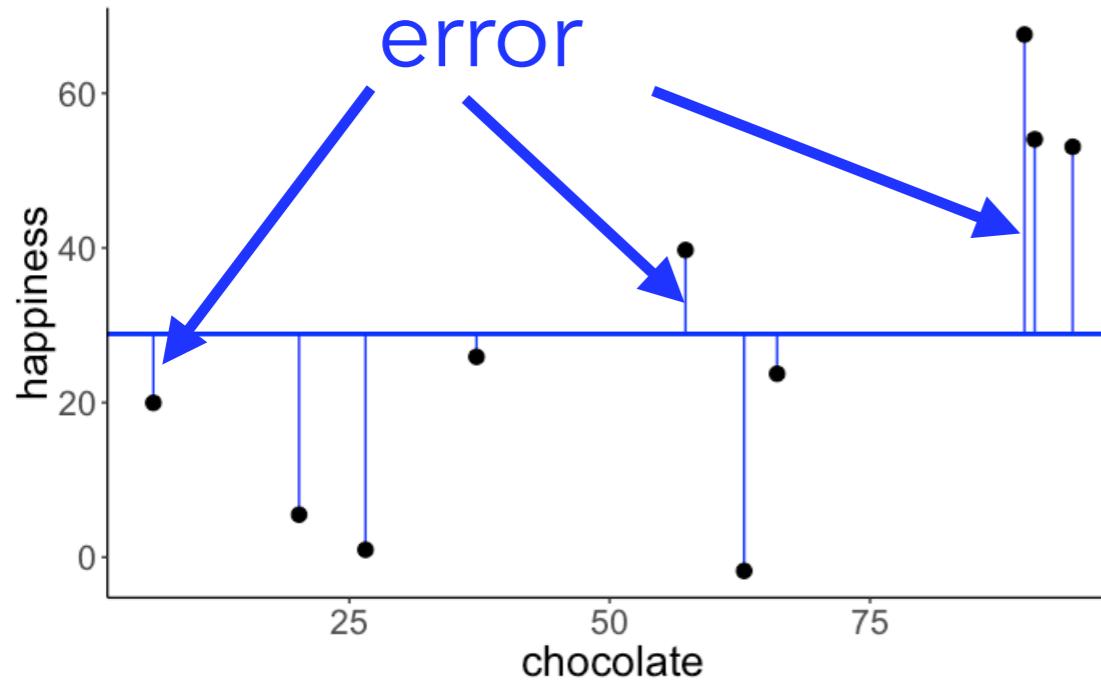
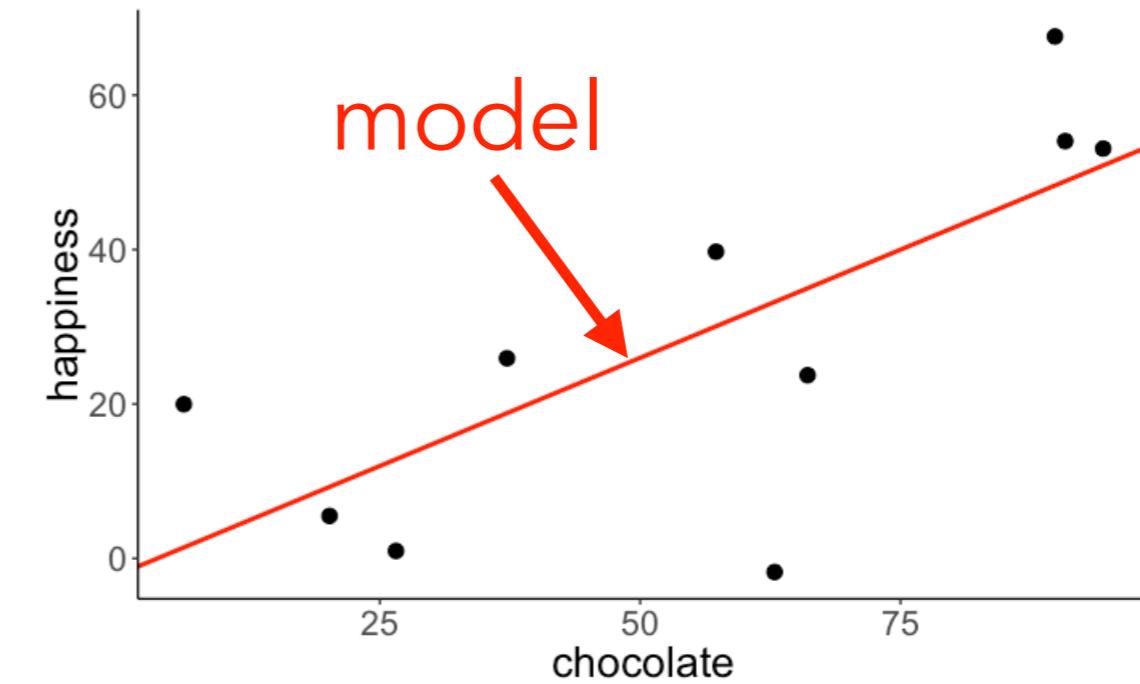


Data = Model + Error

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

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

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

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

Augmented model

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

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

Proportional reduction in error (PRE)

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

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

worth it?

Compact model

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

Augmented model

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

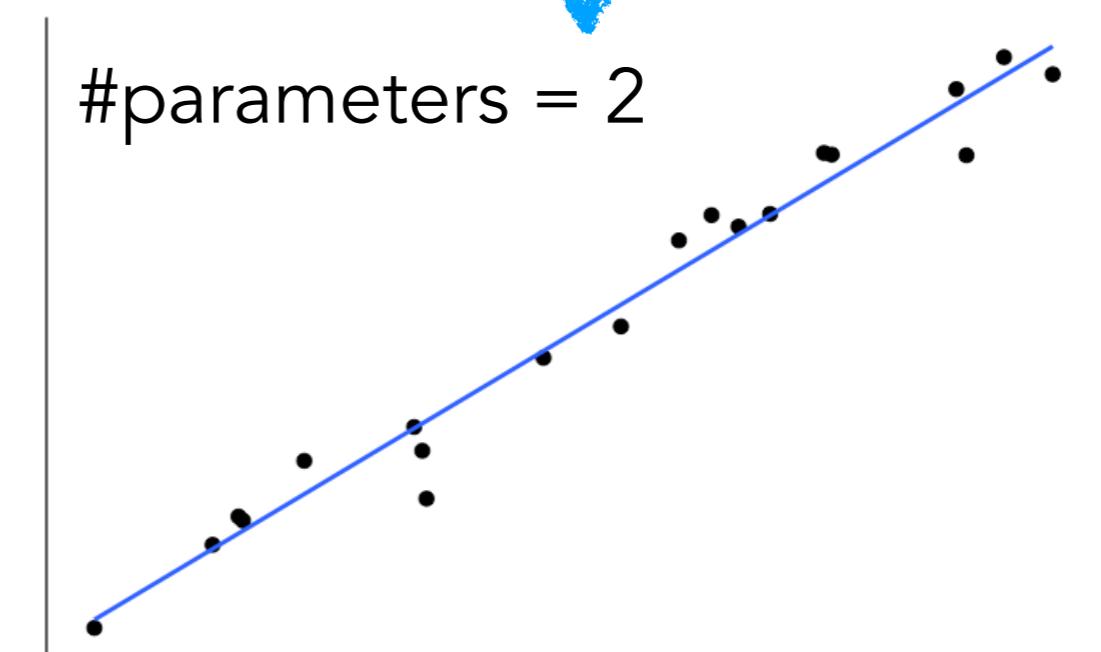
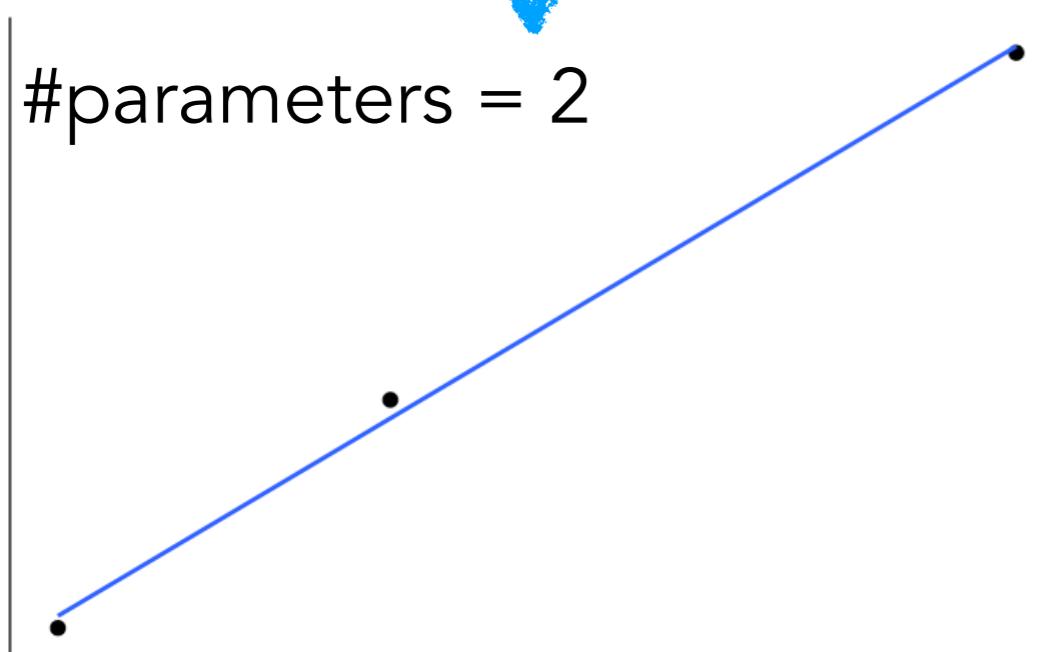
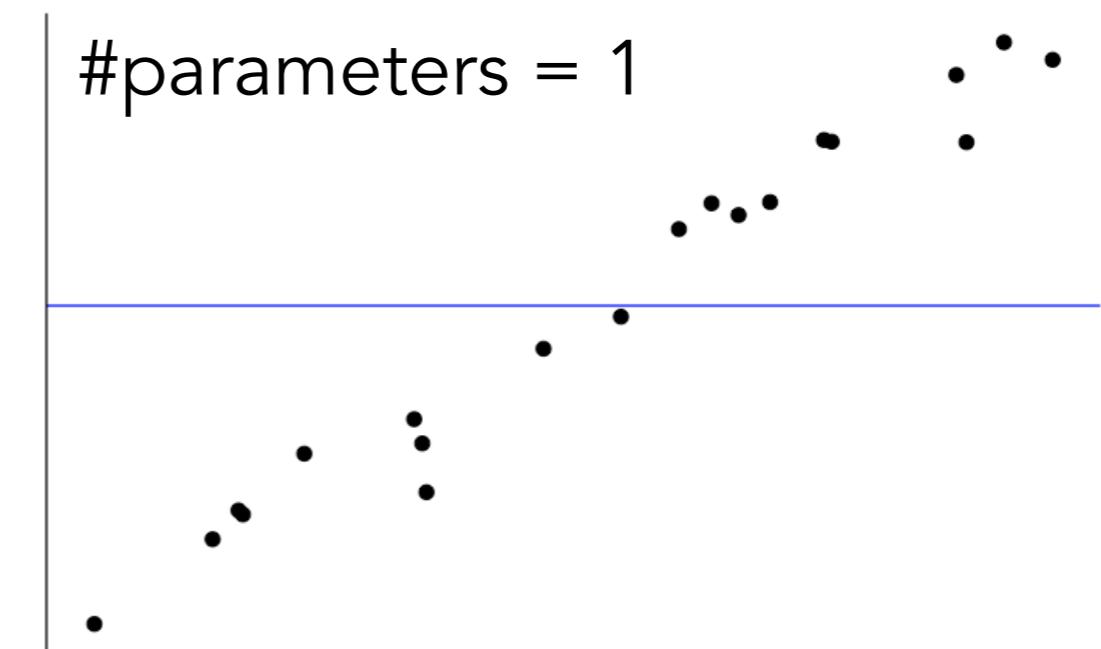
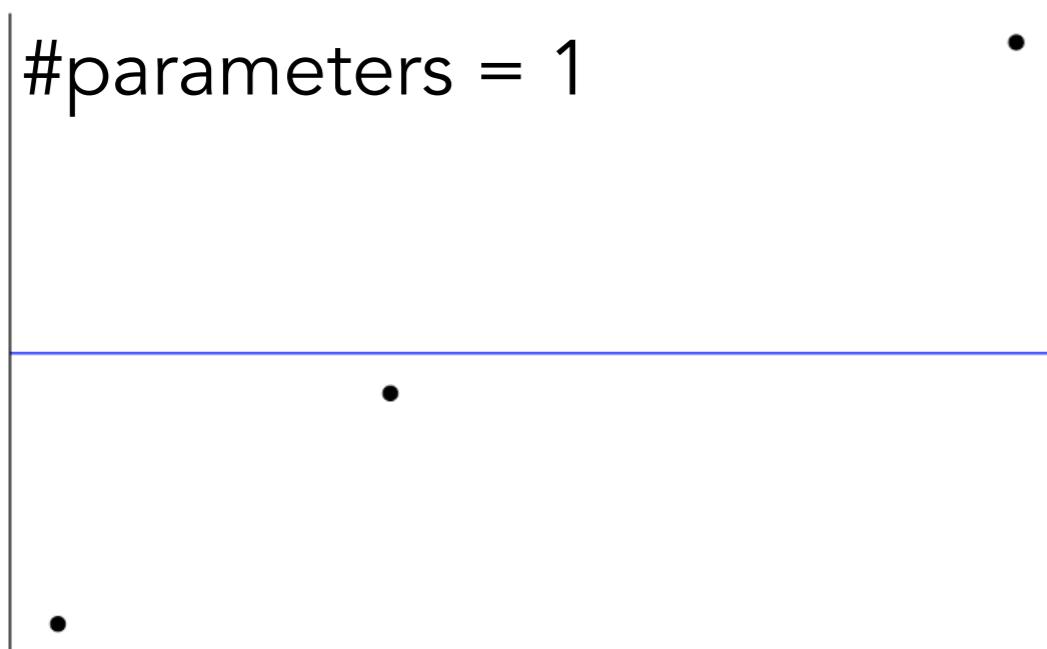
Proportional reduction in error (PRE)

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

- 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_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



neato!

impressive!

General procedure

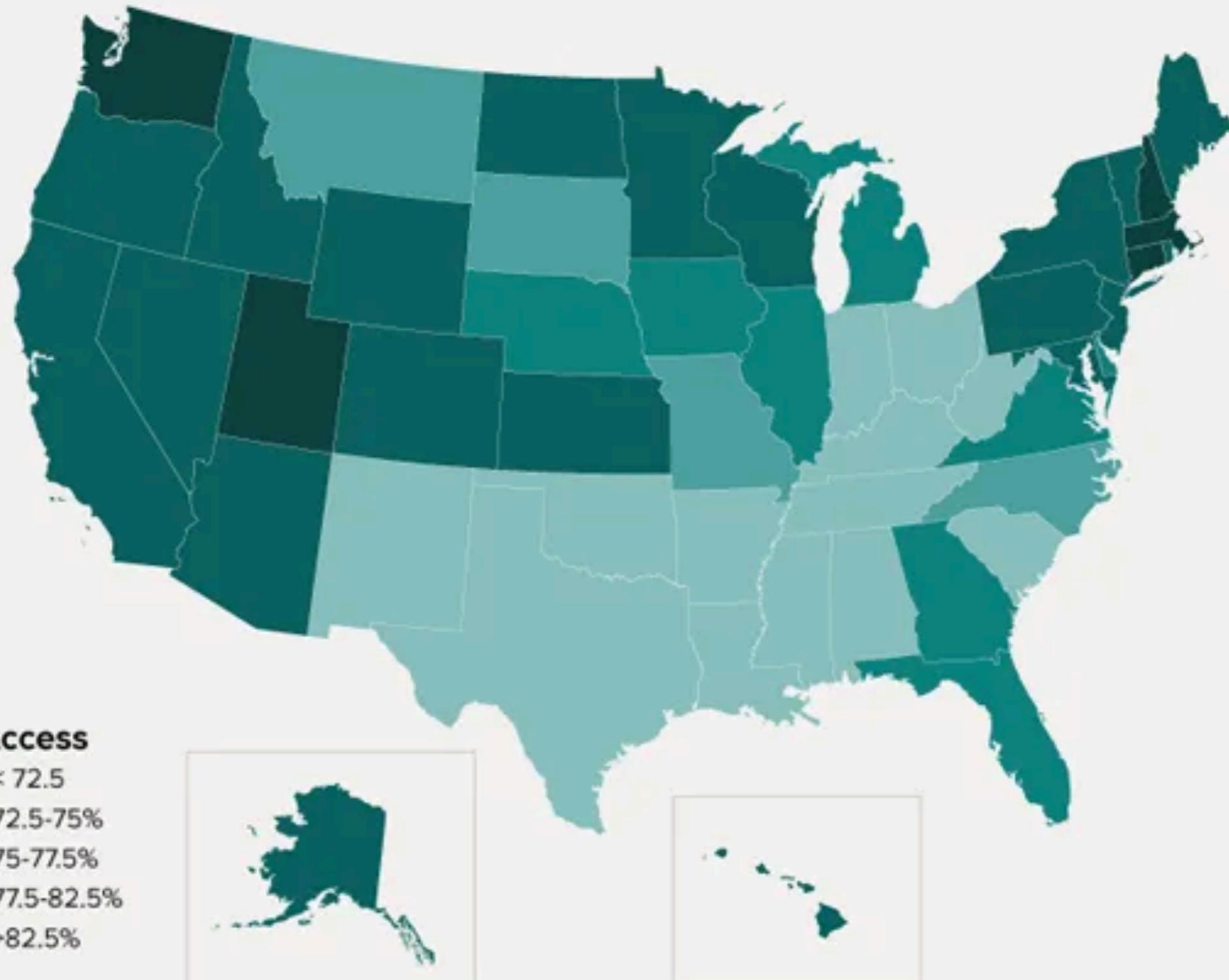
- for any question we want to ask about our DATA
 - we define model_C and model_A
 - compare the models using PRE
 - determine whether PRE is **worth it**
 - in standard frequentist lingo:
 - model_C = H_0 (null hypothesis) 
 - model_A = H_1 (alternative hypothesis) 
 - hypothesis test:
 - H_0 : **all**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

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

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

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

density

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

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 significantly 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

$$Y_i = b_0 + e_i$$

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

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

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

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

$$SSE(C) = 1595 \quad SSE(A) = 1355$$

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

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 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**

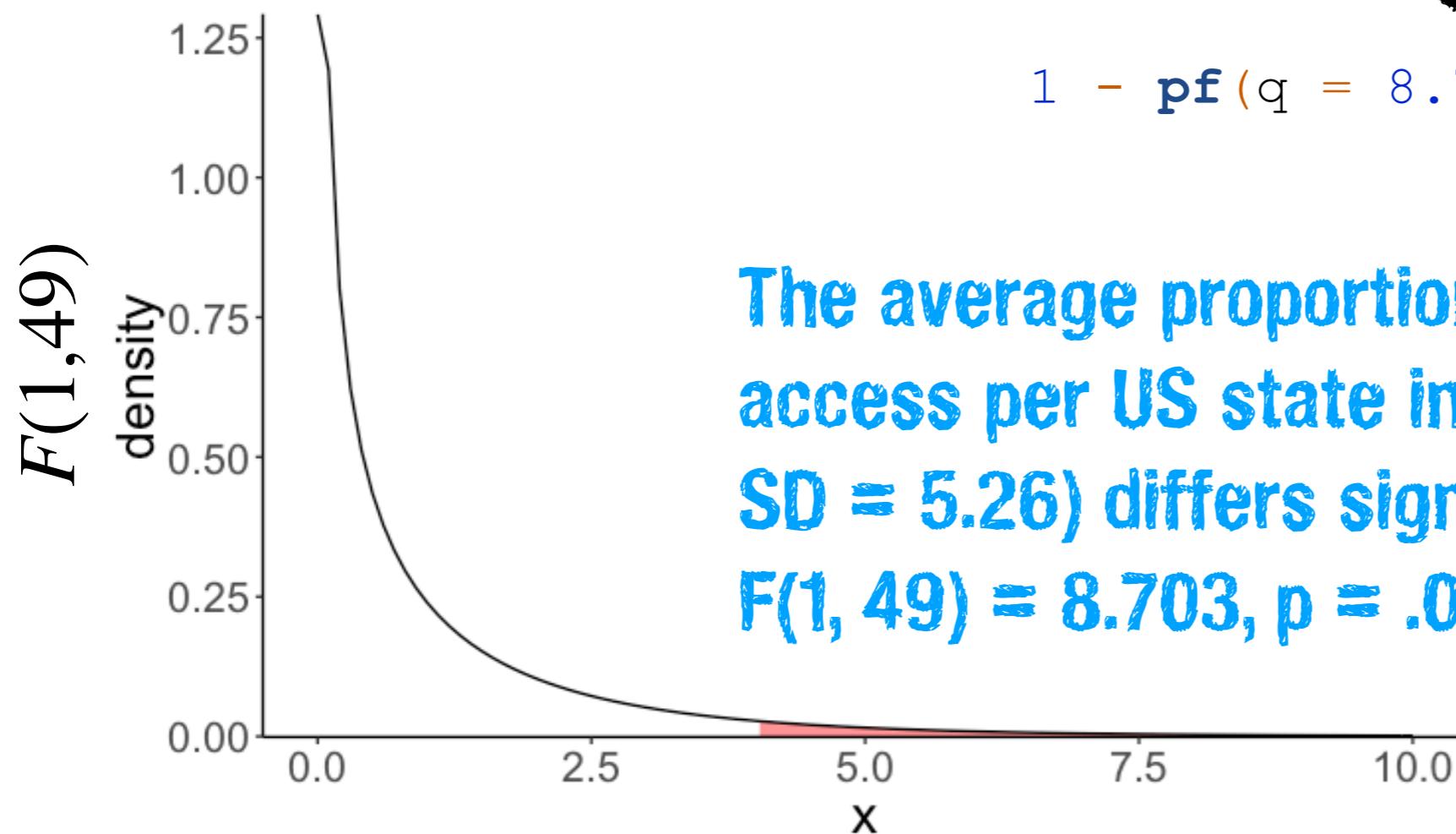
$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

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

Decide whether it's **worth it**

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

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



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

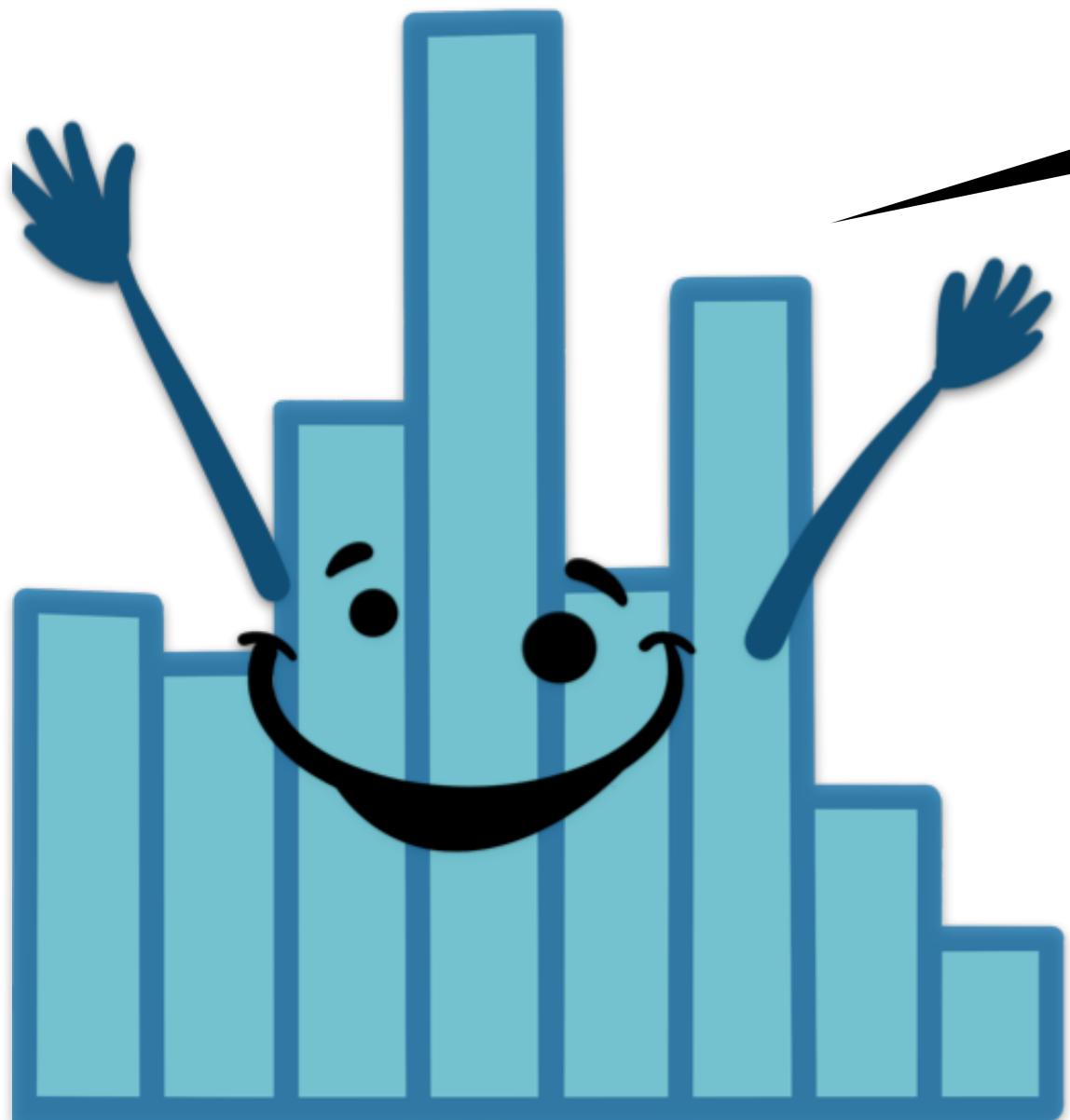
```
data: df.internet$internet  
t = -2.9502, df = 49, p-value = 0.00486  
alternative hypothesis: true mean is not equal to 75
```

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

02:00

stretch break!

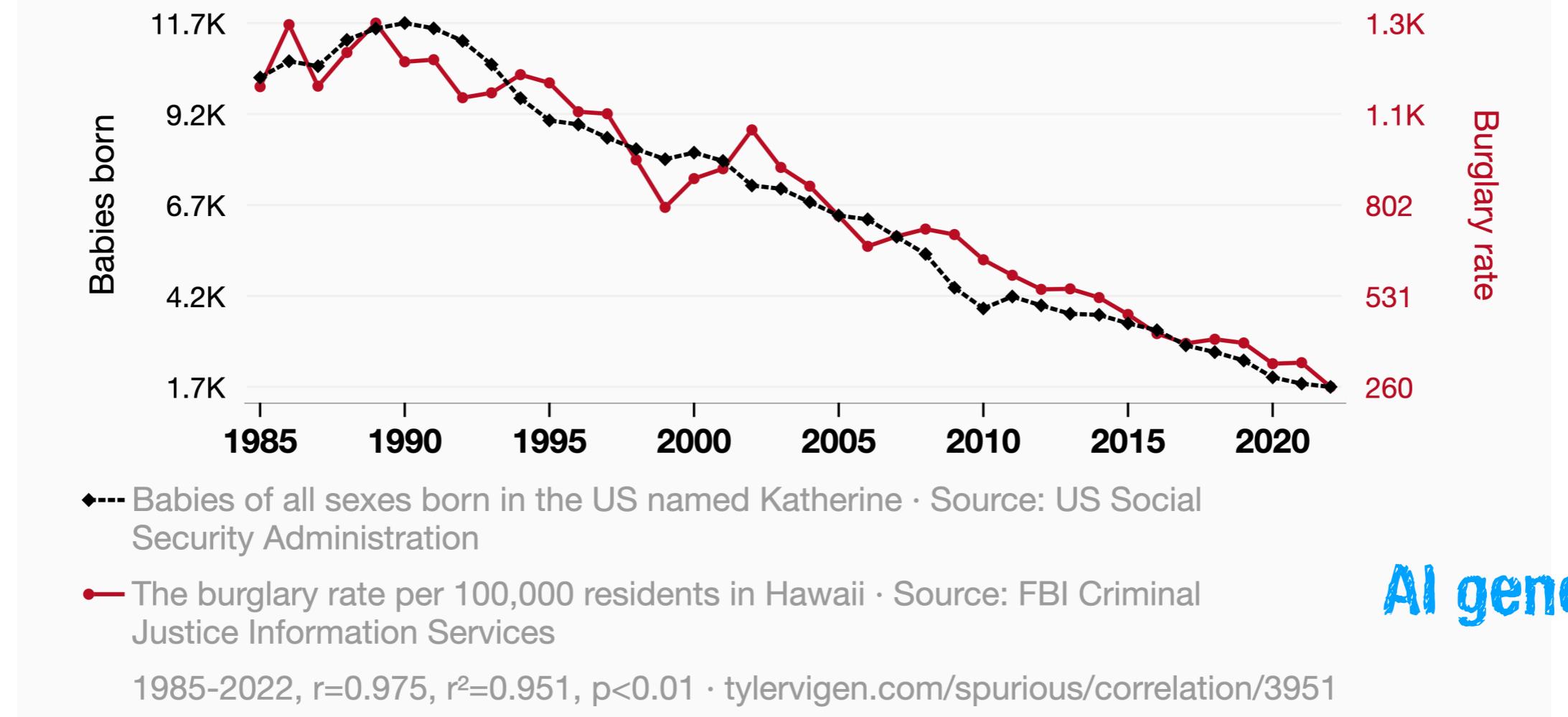


Correlation

Popularity of the first name Katherine

correlates with

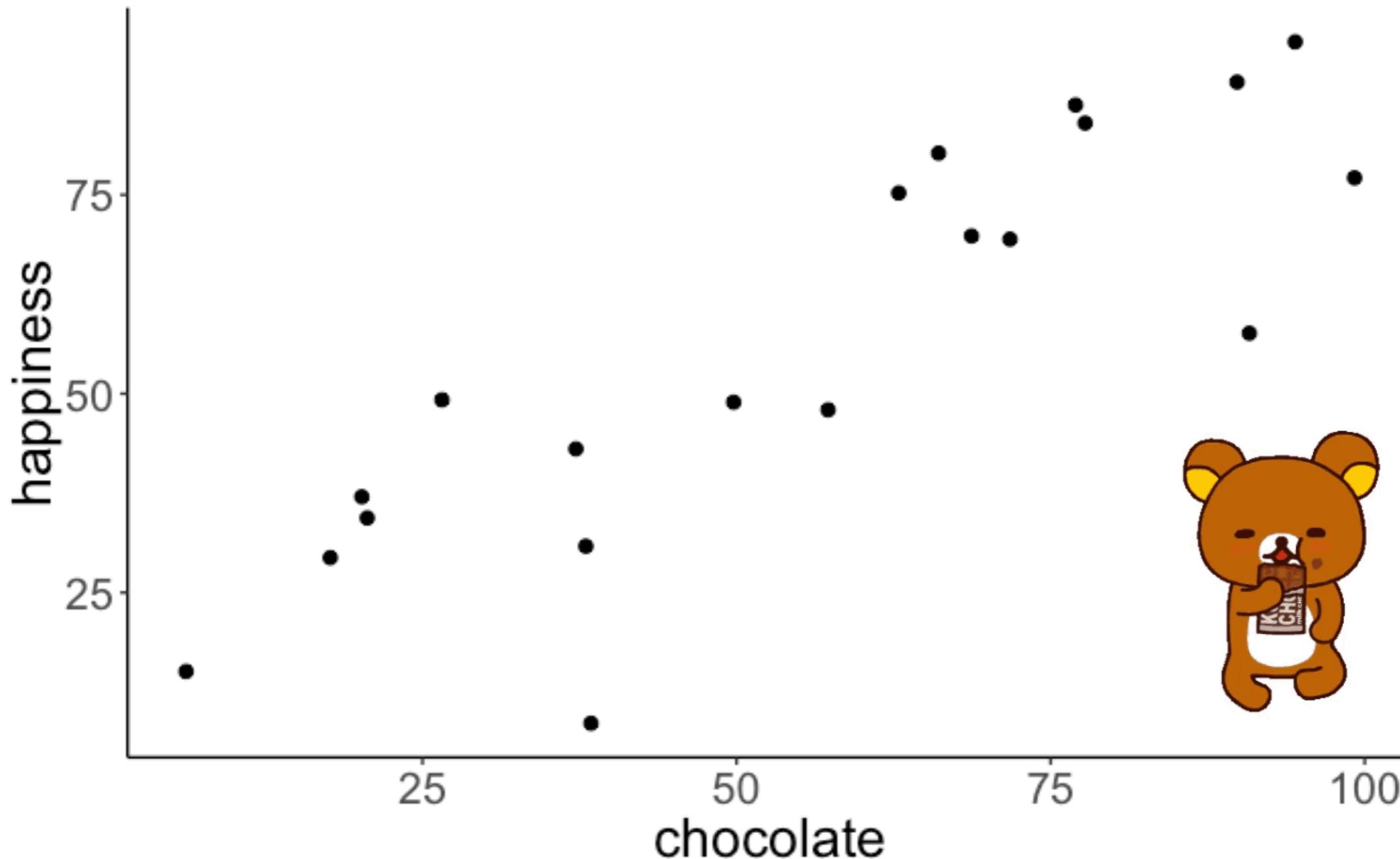
Burglaries in Hawaii



As the saying goes, "Kat's out of the bag," and it seems that also applies to burglars in Hawaii! With fewer Katherines around, there were less Kat burglars trying to pull off heists in the sunny state. It appears that the name Katherine was previously a common alias for cat burglars with a penchant for pilfering pineapples. However, with this name falling out of favor, it seems the purr-petrators have also disappeared, leading to a decrease in burglaries. It's a feline mystery, but it looks like Hawaii can rest easy knowing that the Katherine connection has been pawsitively purr-vented!

<https://tylervigen.com/spurious-correlations>

How to best characterize the relationship between x and y by a single number?

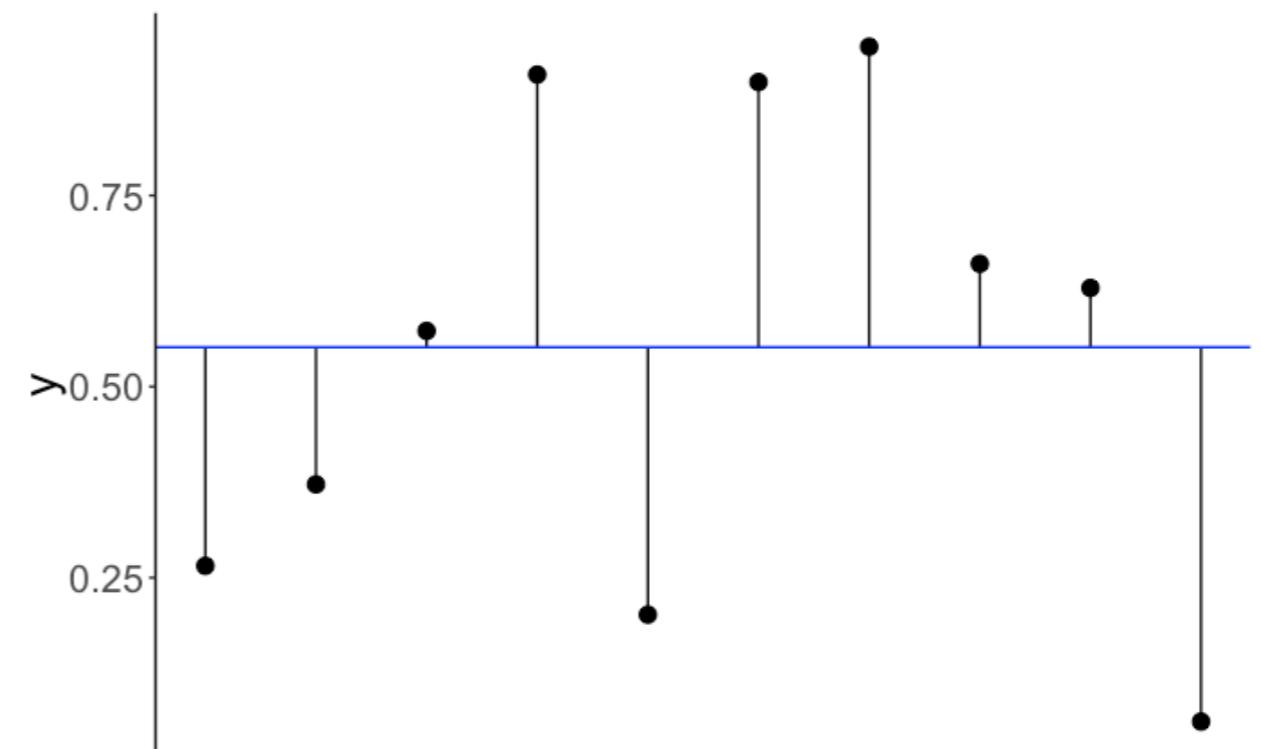
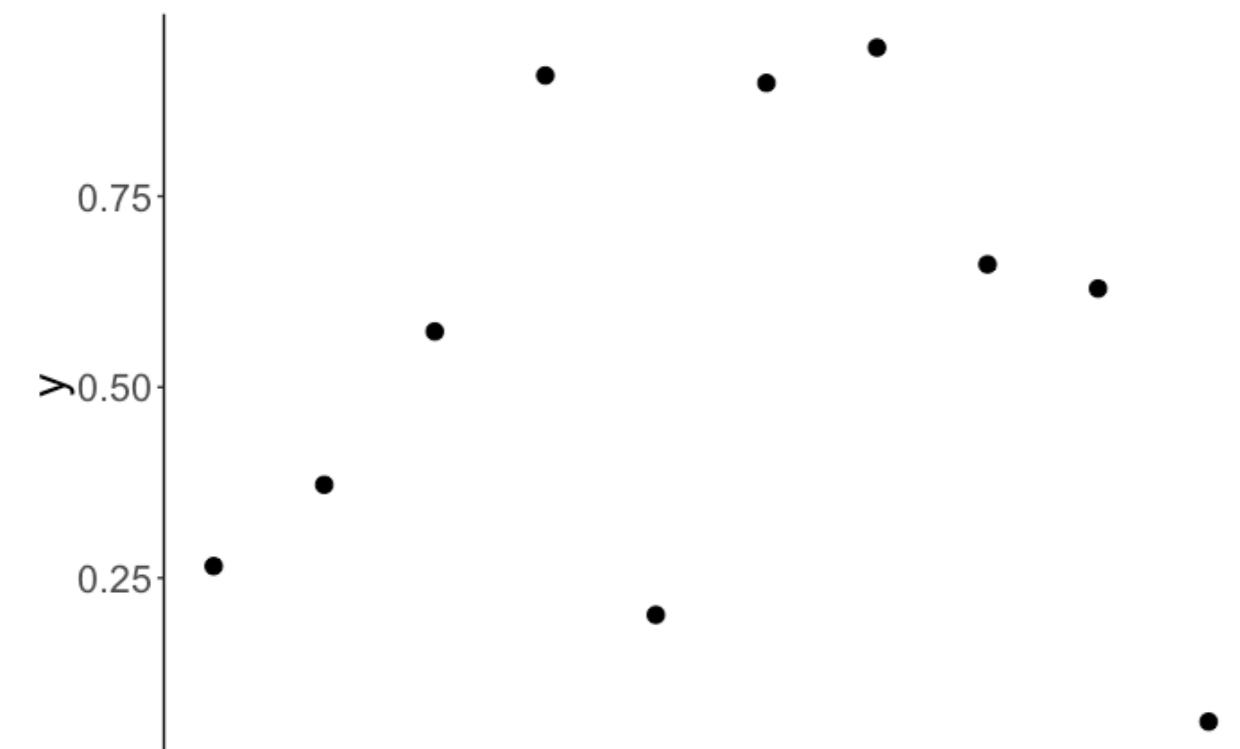


correlation = a measure of the relationship
between two variables

sample variance

$$Var(Y) = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

sum of squared errors

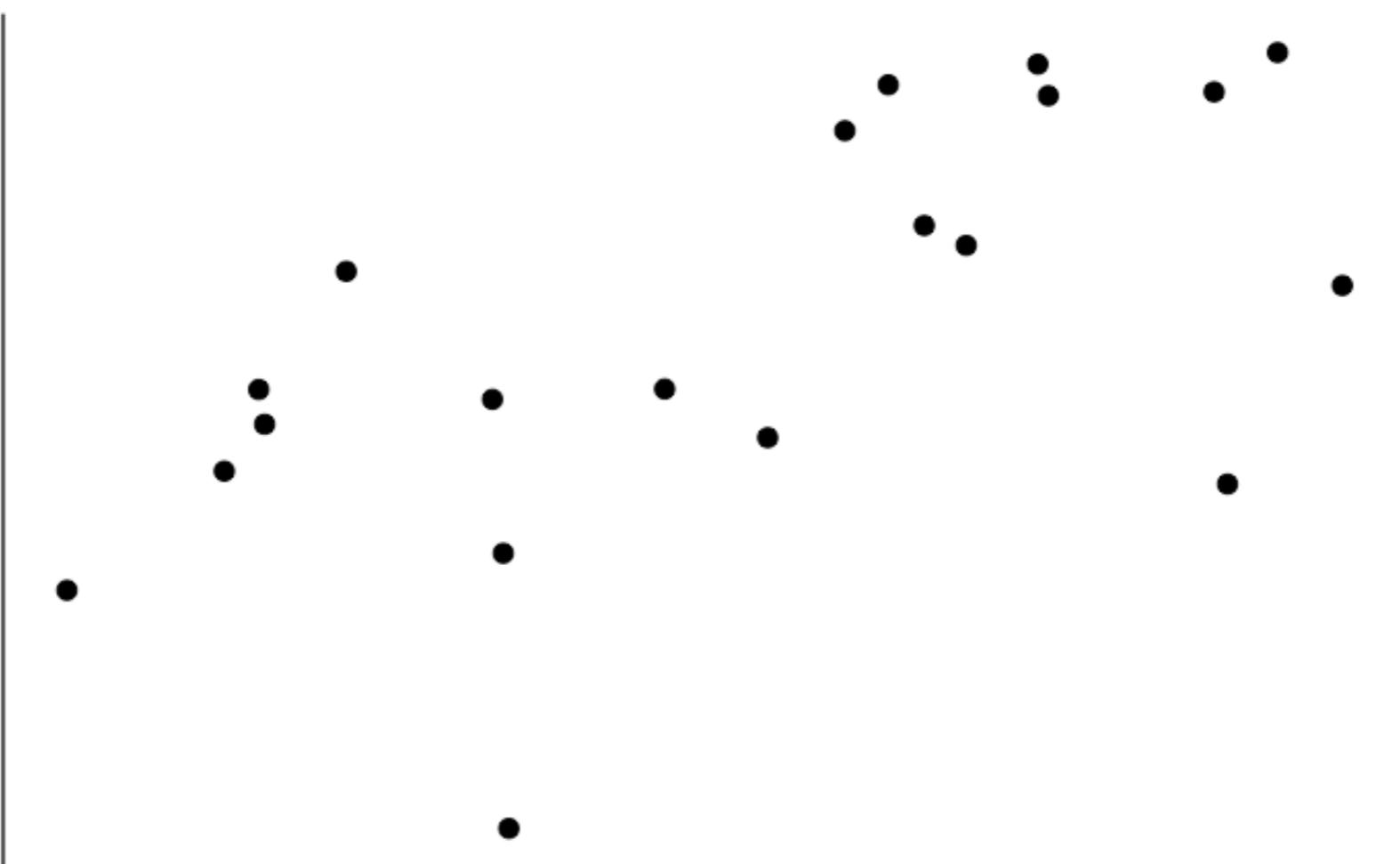


(I was too lazy to draw rectangles ...)

How well does the mean capture the data?

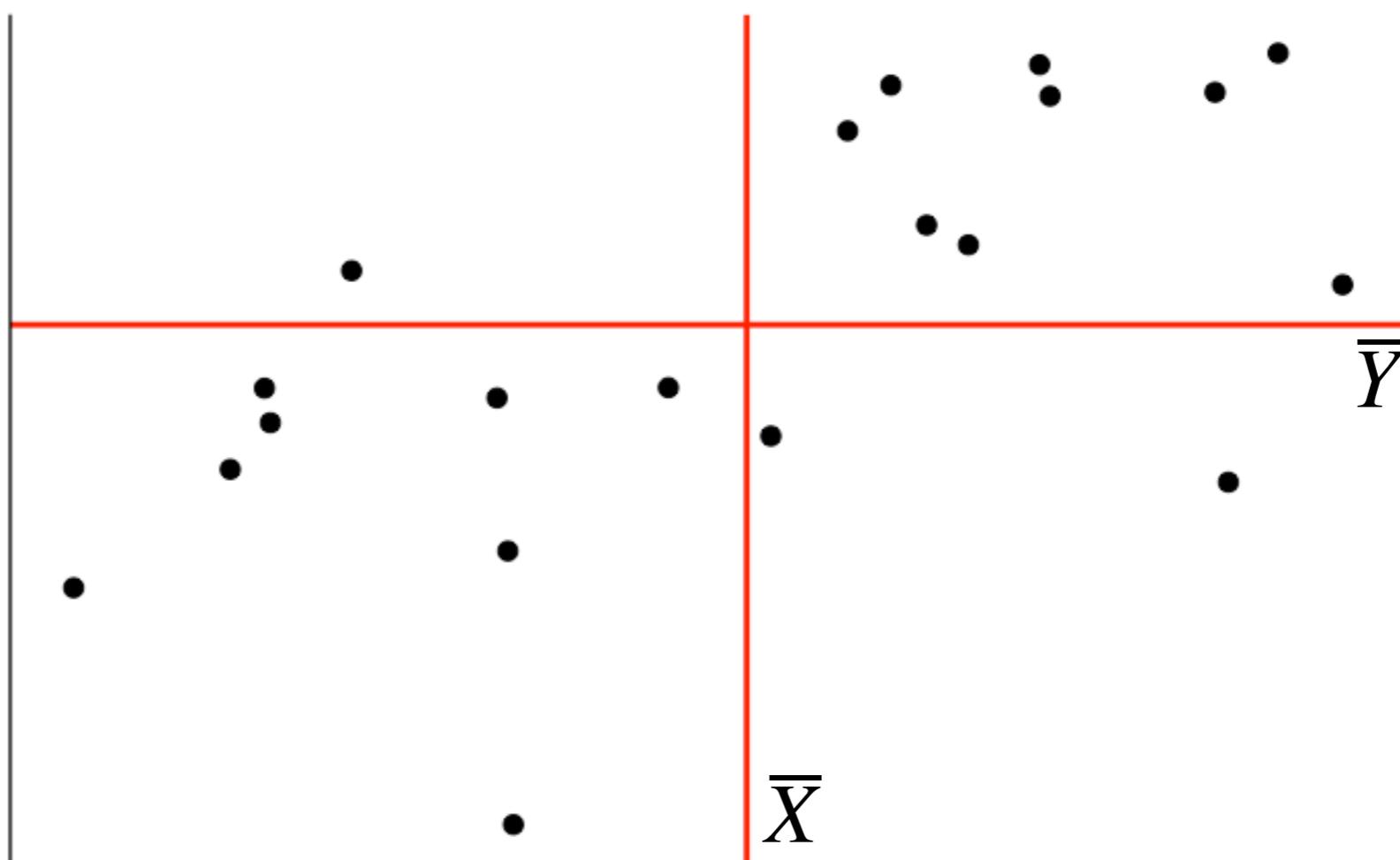
sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



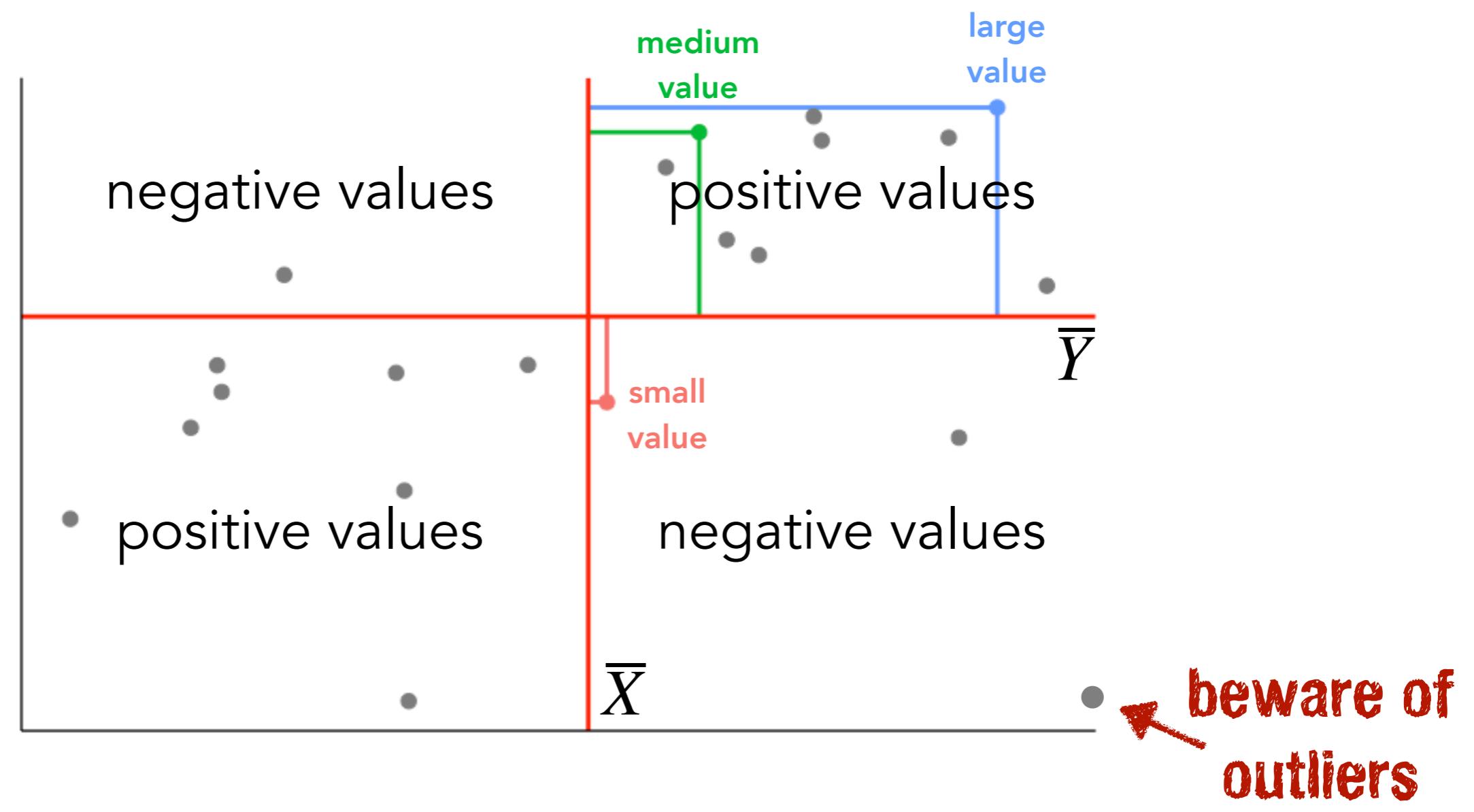
sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



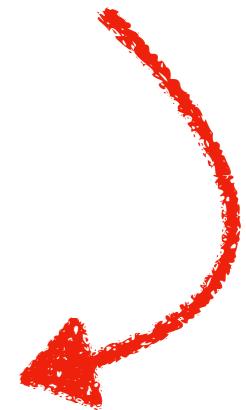
sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

depends on the scale of the variables

the $n - 1$ s cancel out

sample correlation coefficient

$$r(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$


standardized covariation
(dividing by the standard deviations)

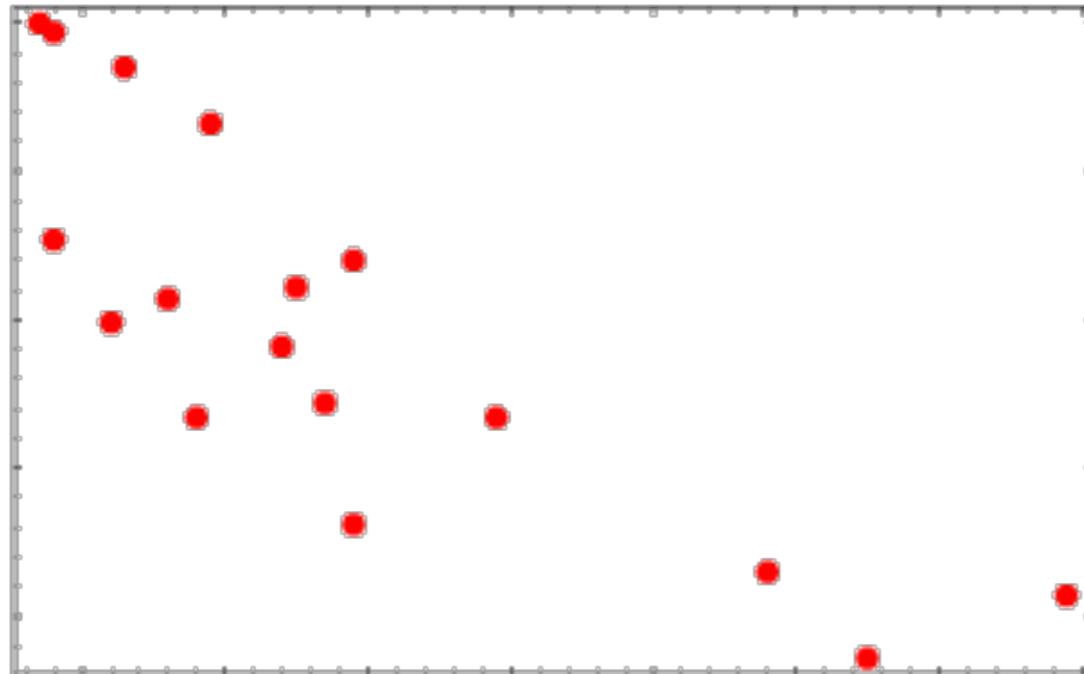
Properties of the Pearson correlation

- standardized: $-1 \leq r \leq 1$
- scale independent (for both X and Y)
- commutativity: $r(X, Y) = r(Y, X)$
- sign determines the direction of dependence
- captures **linear dependence** only

association not
causation



Who is the correlation champion?



Winner gets chocolate!

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

Who is the correlation champion?

Get ready to compete!

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

Leaderboard

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

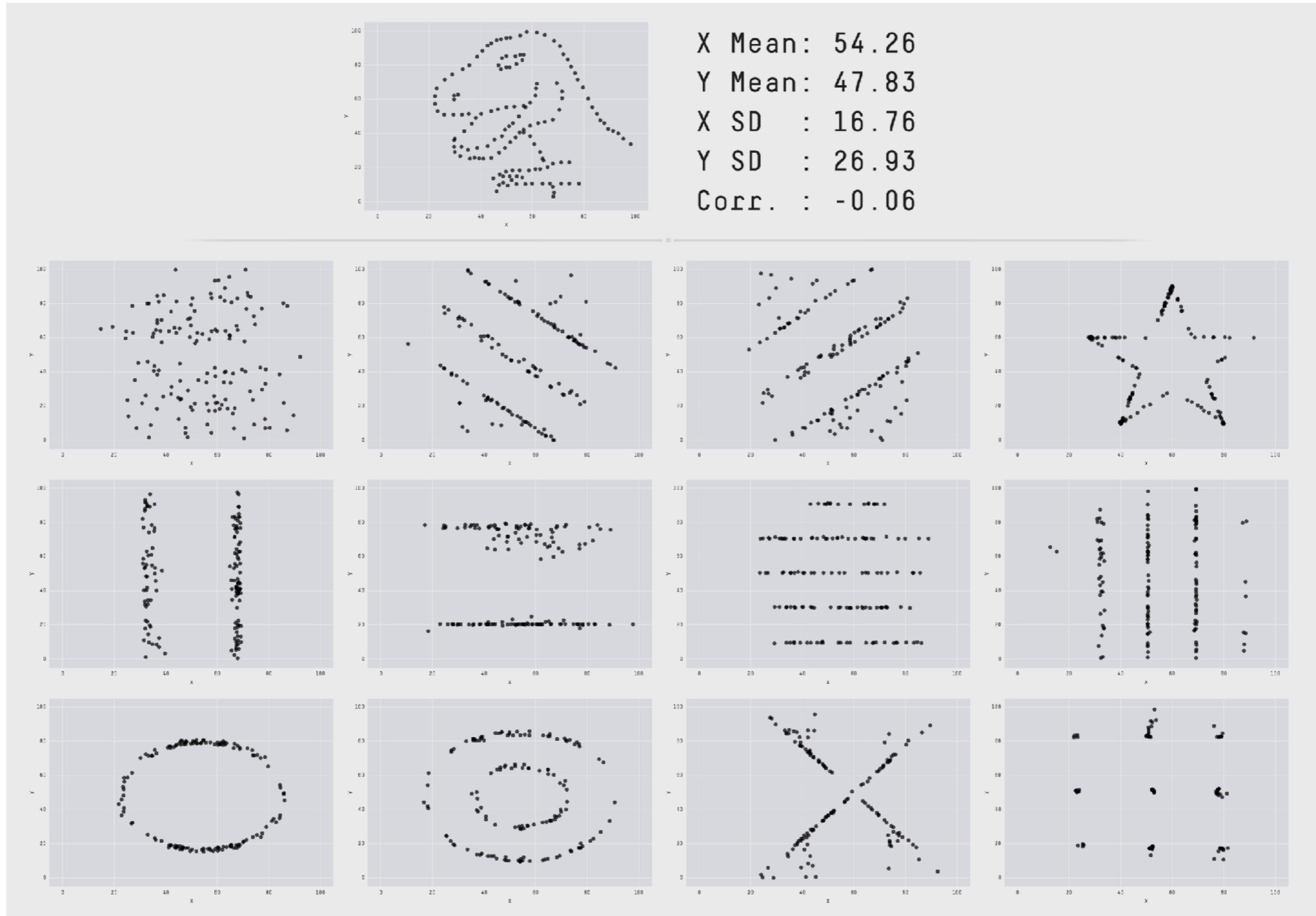
Leaderboard

Start the presentation to see live content. Still no live content? Install the app or get help at PollEv.com/app

Solution

XX

Be careful about interpreting correlations!

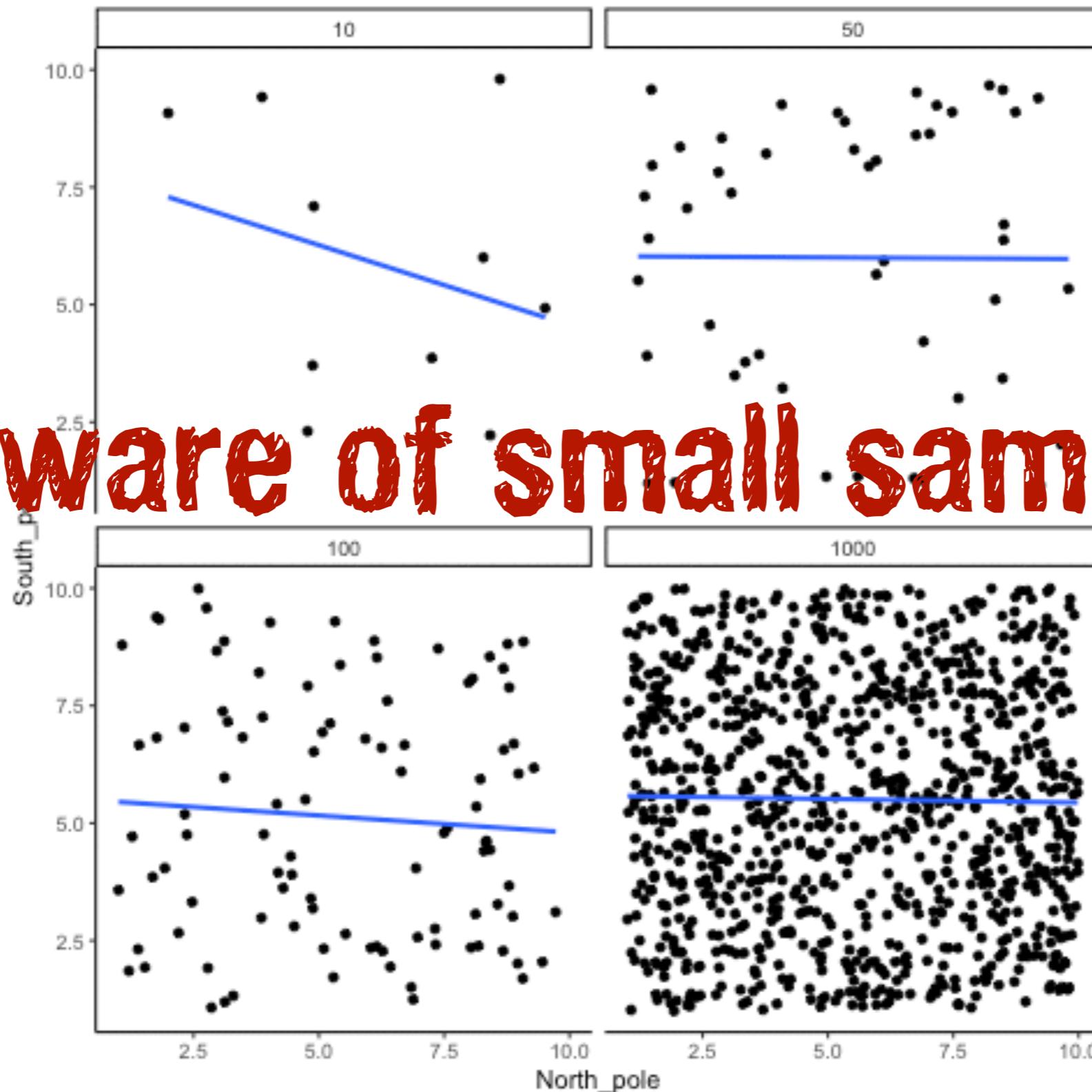


always visualize the data ...

$n = [10, 50, 100, 1000]$

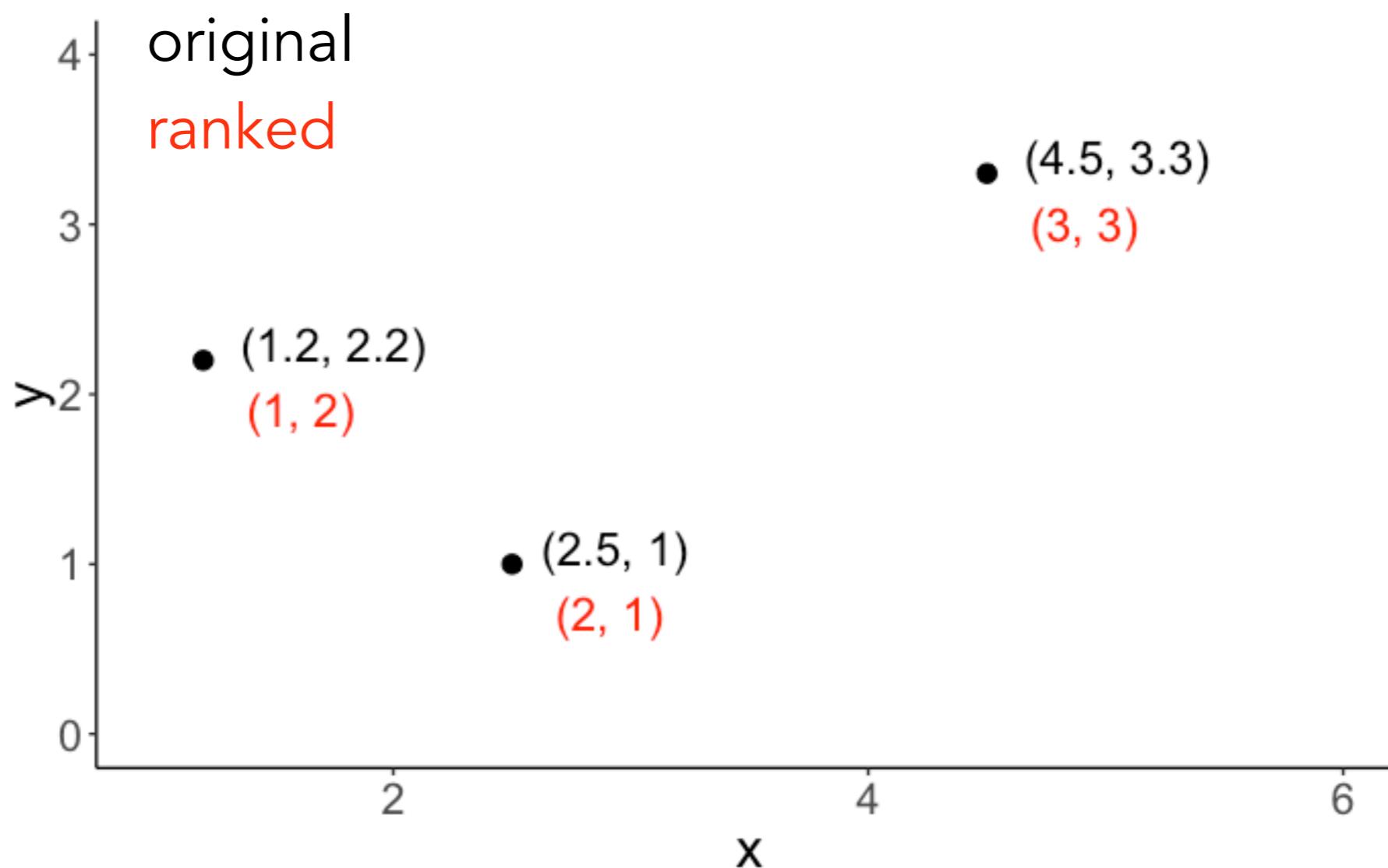
$X \sim \mathcal{U}(\min = 0, \max = 10)$
 $Y \sim \mathcal{U}(\min = 0, \max = 10)$

Beware of small samples!



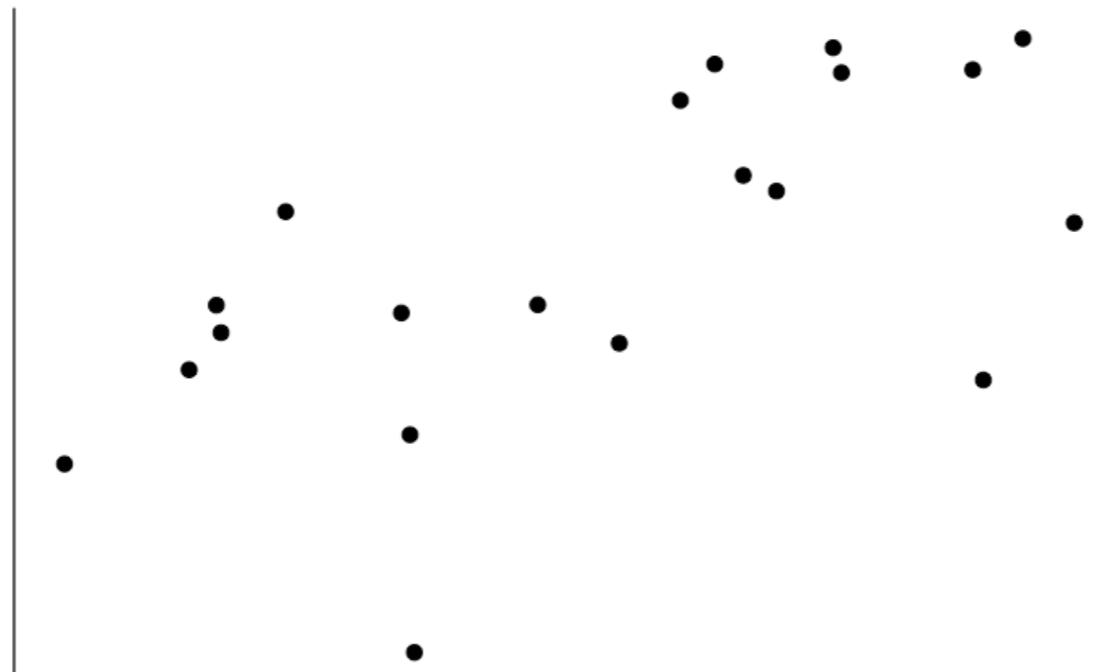
Spearman rank order correlation

- transform original data into ranks
- calculate correlation on the ranked data



Spearman rank order correlation

- transform original data into ranks
- calculate correlation on the ranked data



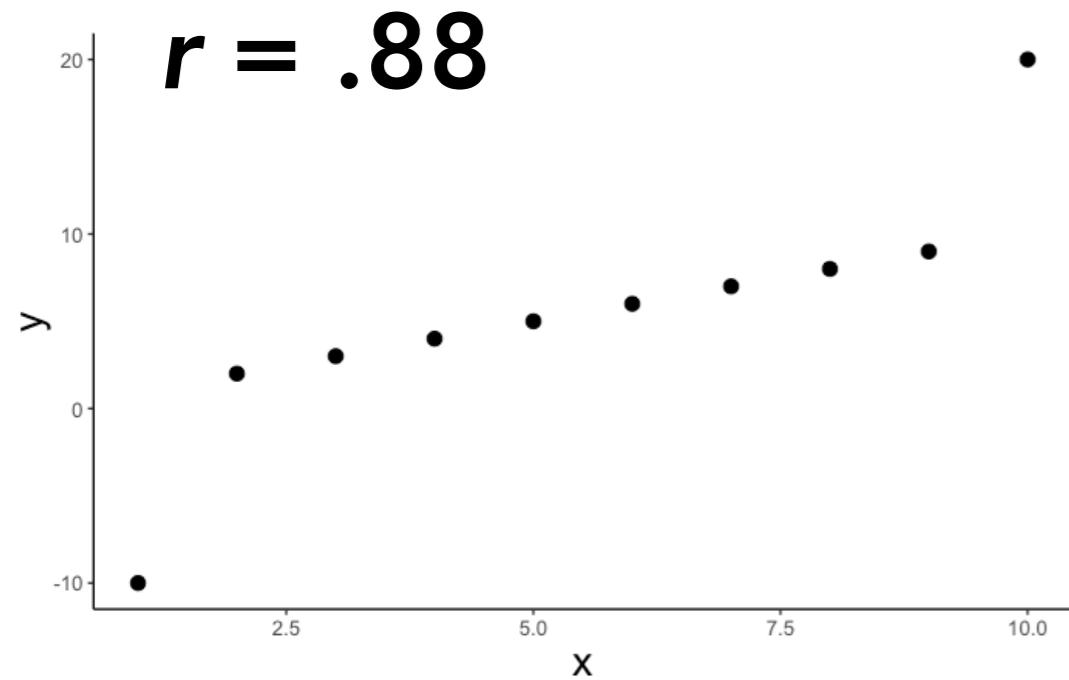
x	y	x_rank	y_rank
0.27	1.14	5	12
0.37	0.97	6	8
0.57	0.92	10	6
0.91	0.85	18	4
0.20	0.98	3	9
0.90	1.39	17	17
0.94	1.44	19	20
0.66	1.40	12	18
0.63	1.33	11	15
0.06	0.71	1	2

r	spearman	r_ranks
0.609	0.595	0.595

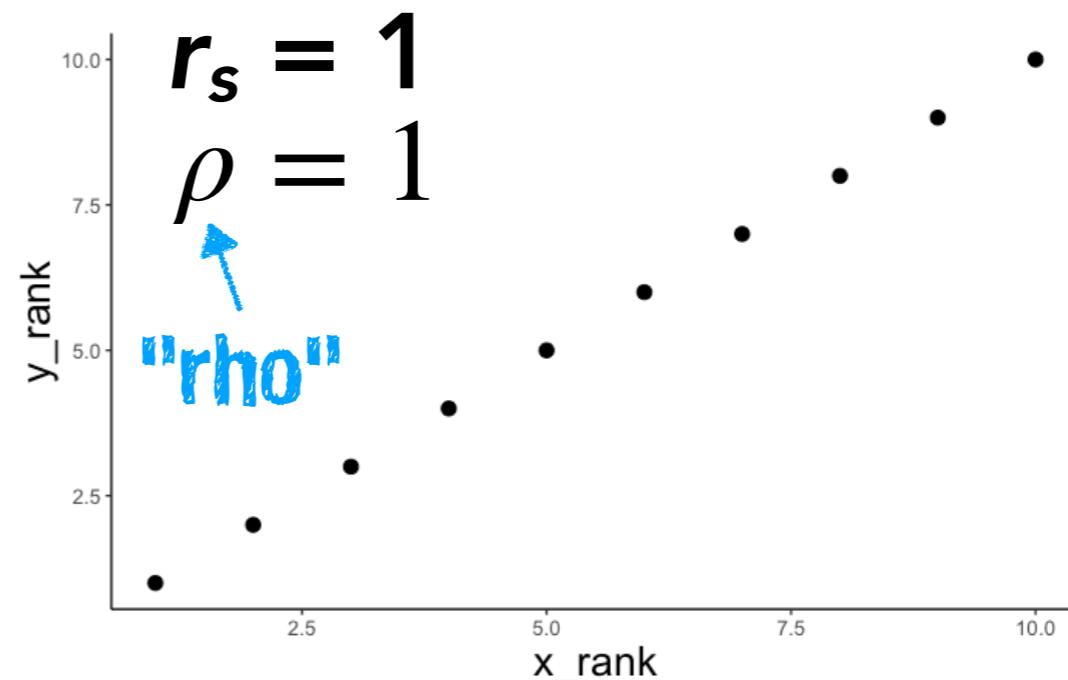
```
1 # correlation
2 df.spearman %>%
3   summarize(r = cor(x, y, method = "pearson"),
4             spearman = cor(x, y, method = "spearman"),
5             r_ranks = cor(x_rank, y_rank, method = "pearson"))
```

Spearman rank order correlation

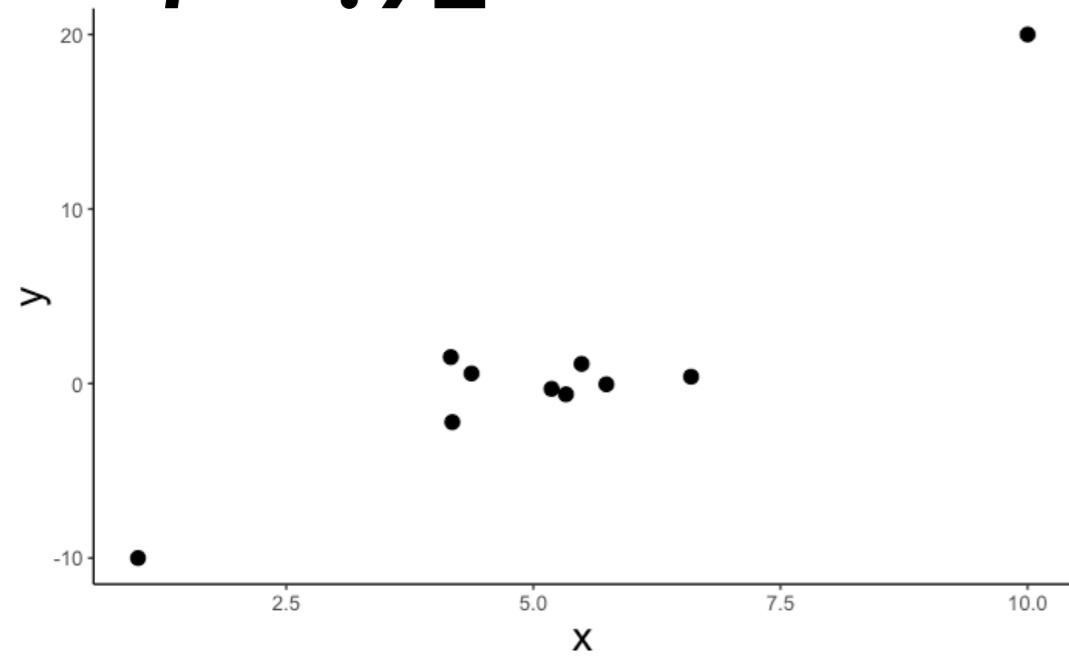
original



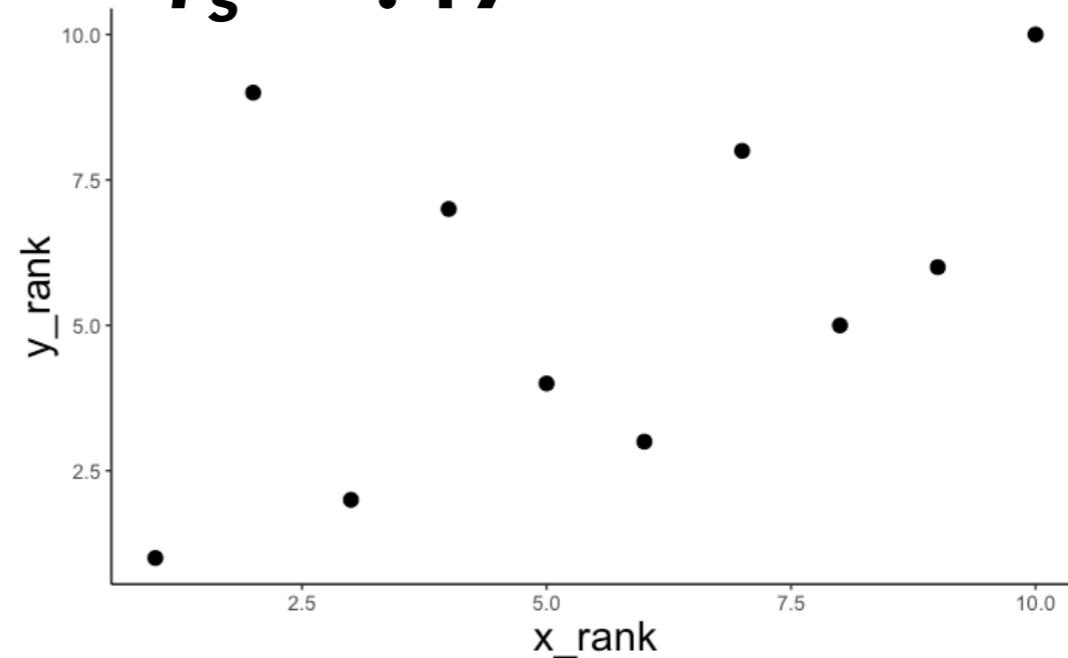
ranked



$r = .92$



$r_s = .47$



Pearson vs. Spearman

- Pearson's r captures the extent to which the relationship between two variable is **linear**
- Spearman's ρ captures the extent to which the relationship between two variables is **monotonic**
- What's better?
 - depends on the context
 - Spearman is robust to outliers, but it throws away (potentially useful) information

CORRELATION IS NOT CAUSATION



NYT Health
@NYTHealth

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.



Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

[nytimes.com](#)

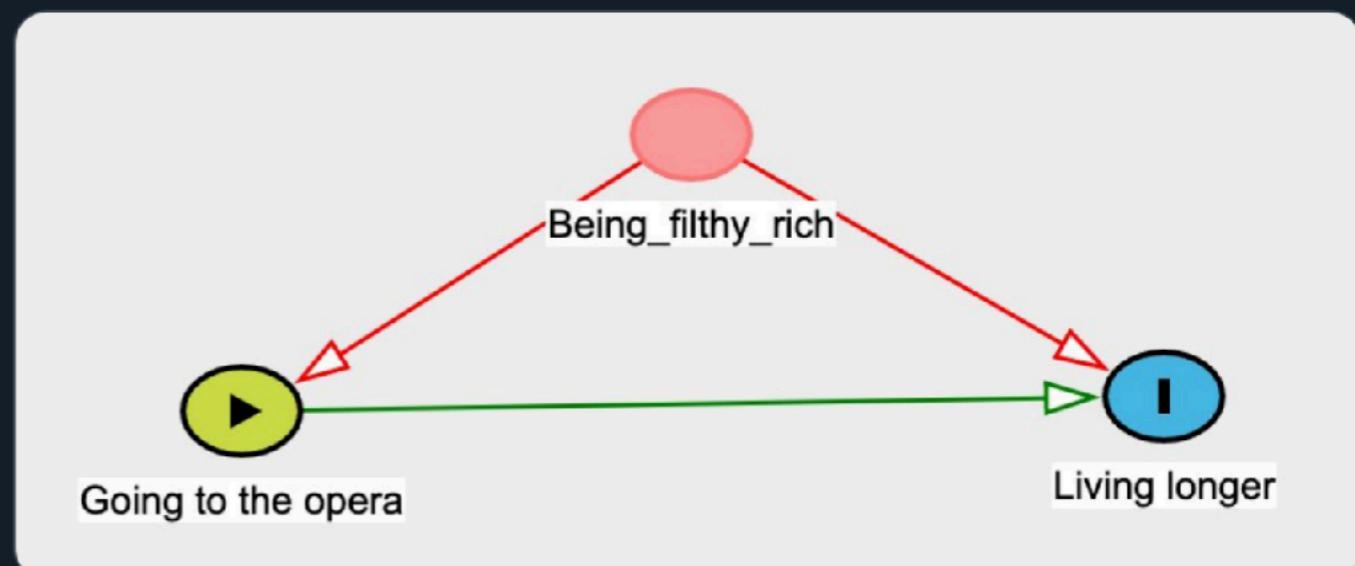
9:19 AM · Dec 22, 2019 · SocialFlow

336 Retweets 1.3K Likes



Andrew Heiss
@andrewheiss

ooh ooh i can draw the dag for this one!



NYT Health @NYTHealth · Dec 22, 2019

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't. [nyti.ms/2Q9AmZV](#)

2:47 PM · Dec 22, 2019 · Twitter Web App

[View Tweet activity](#)

837 Retweets 3.9K Likes

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



- correlations suggest that there is some causal relationship
- but this relationship need not be a direct causal relationship from A to B (or from B to A)

more about causation in a later class

Regression

The conceptual tour

Linear model: Simple regression

Data = Model + Error

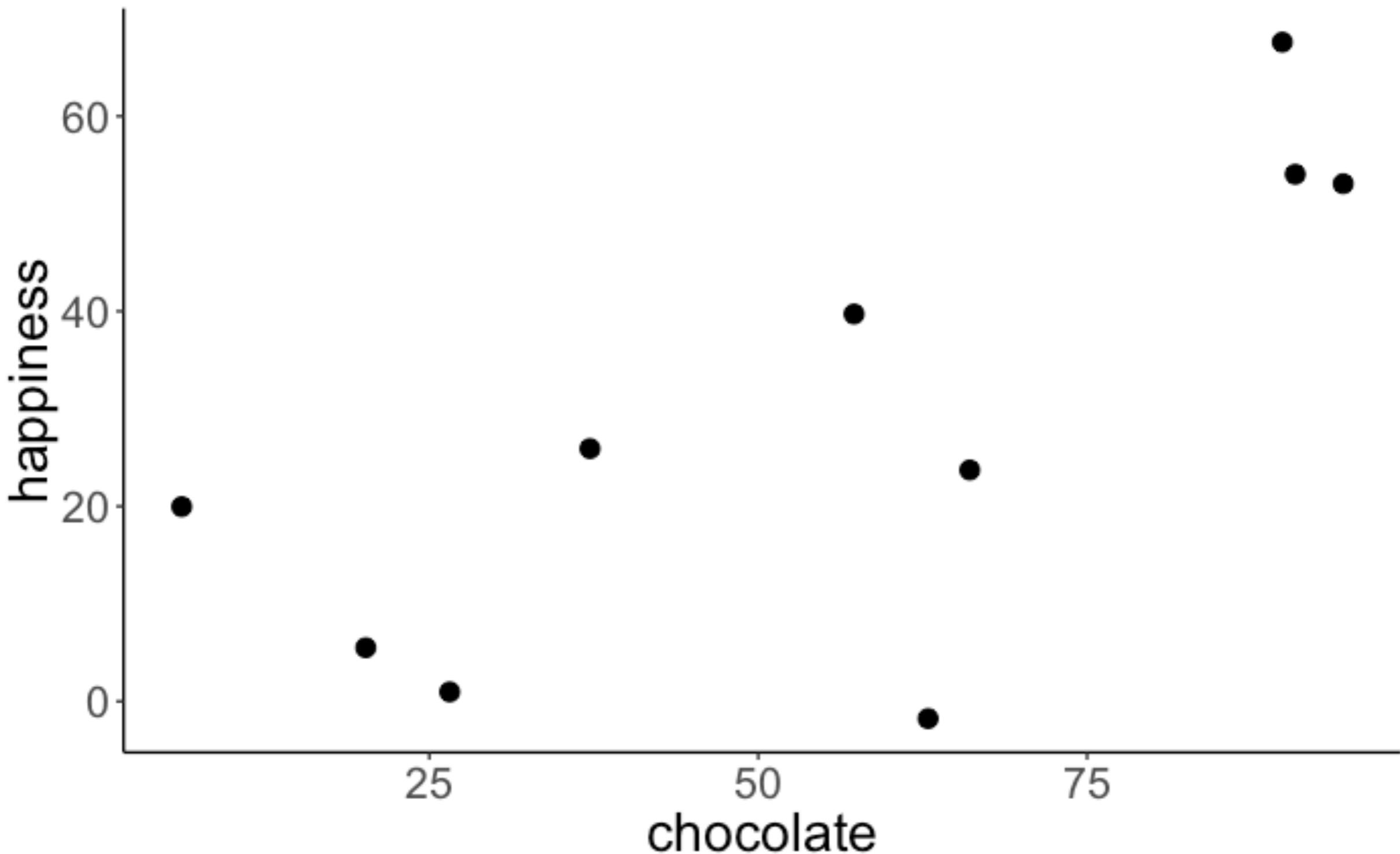
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

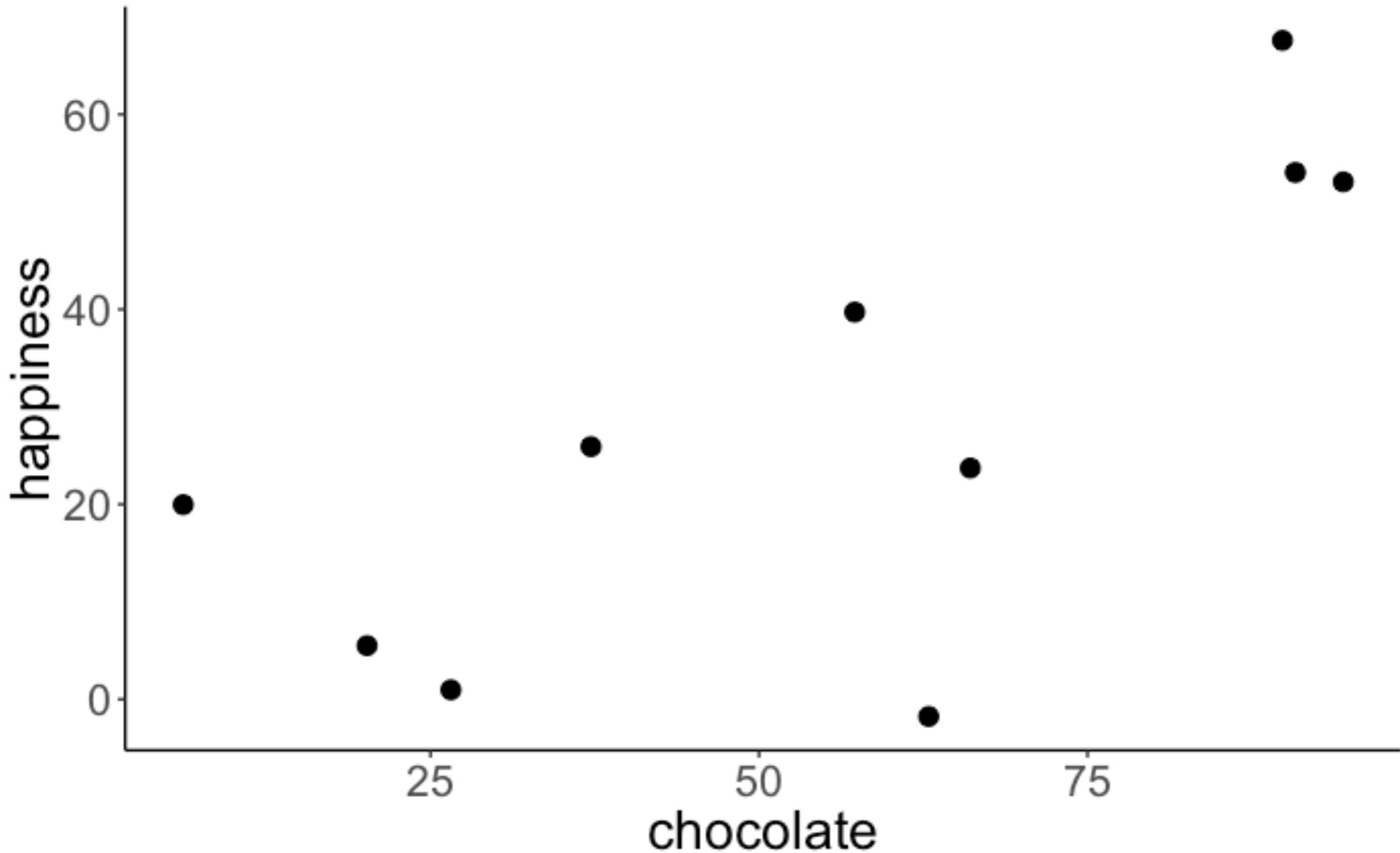


the model is a linear
combination of predictors

Does chocolate make us happy?



Is there a relationship between chocolate consumption and happiness?



The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

H_0 : Chocolate consumption and happiness are unrelated.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and

$$\beta_1 = 0$$

H_1 : Chocolate consumption and happiness are related.

Model A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



chocolate
consumption

The general procedure

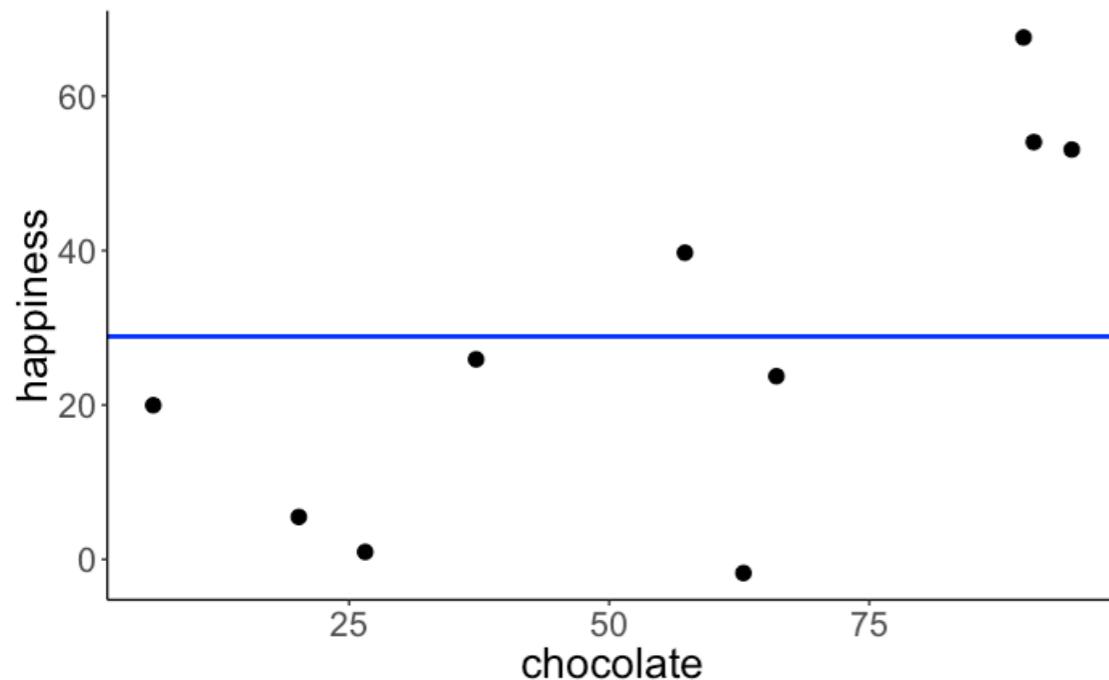
1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
- 2. Fit model parameters to the data**
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

H_0 : Chocolate consumption and happiness are unrelated.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 28.88 + e_i$$

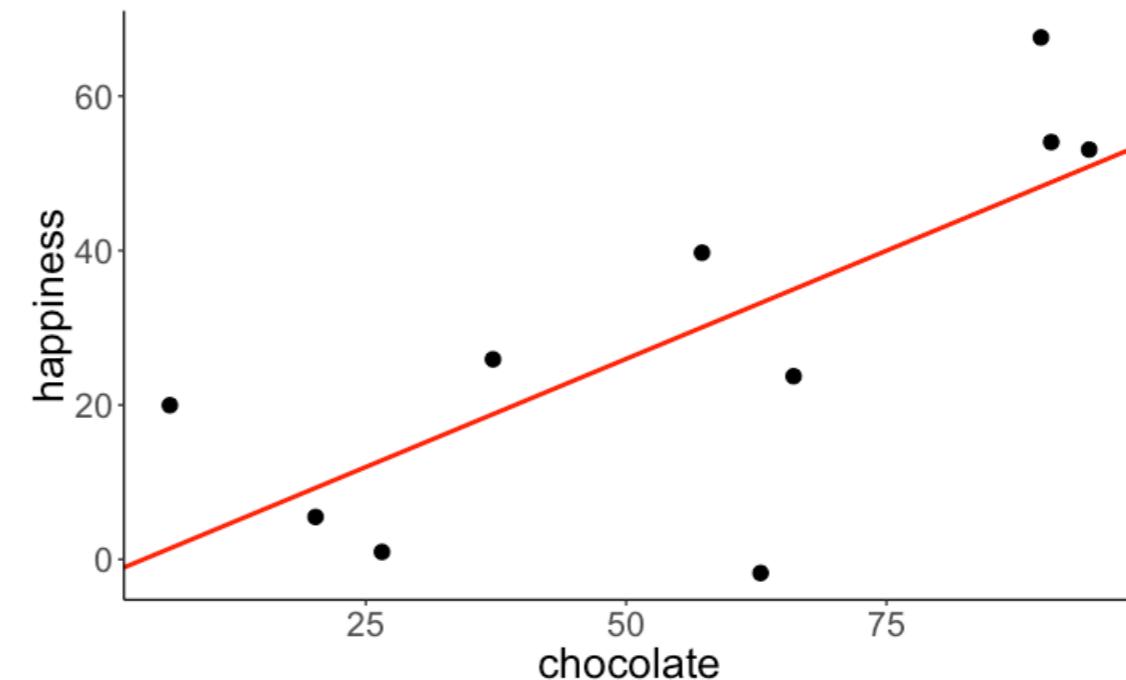
H_1 : Chocolate consumption and happiness are related.

Model A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

chocolate consumption

Model prediction



Fitted model

$$Y_i = -2.04 + 0.56X_i + e_i$$

The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
- 3. Calculate the proportional reduction of error (PRE) in our sample**
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

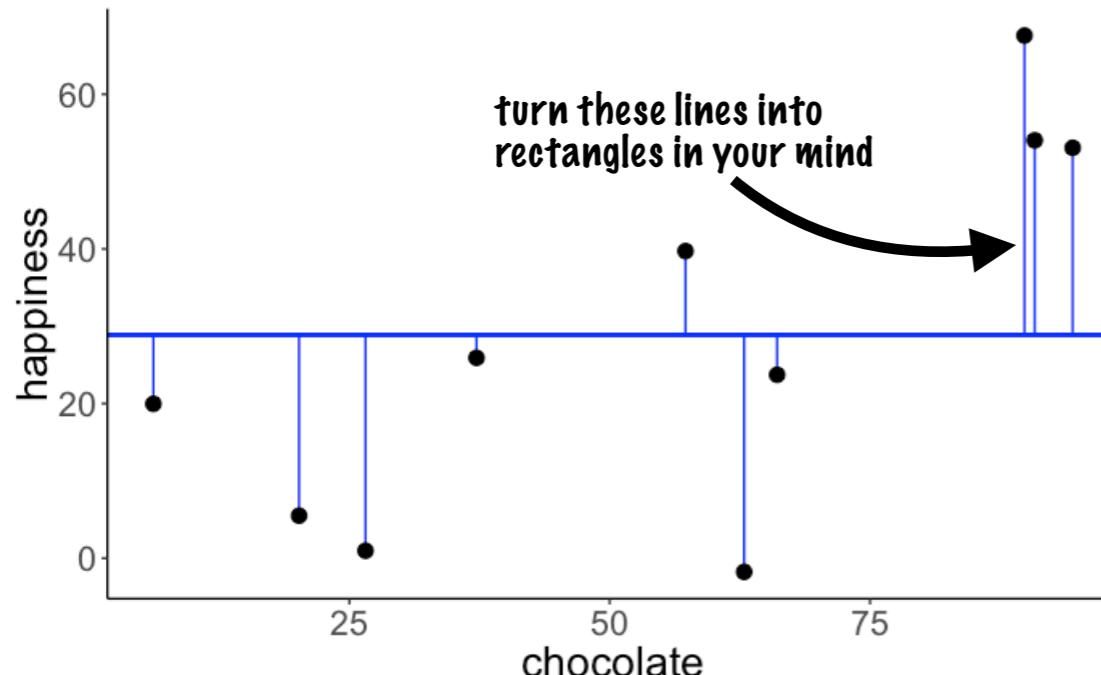
Calculate PRE

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

Both models were fit to minimize the sum of squared errors

OLS = Ordinary **least squares** regression

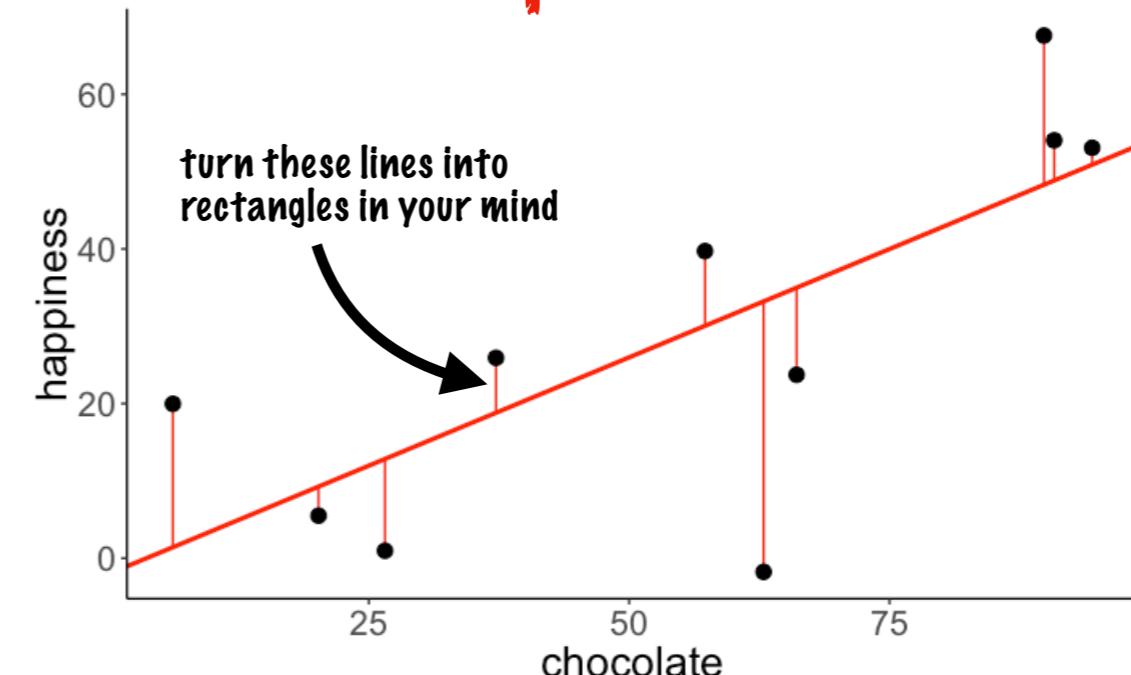
Sum of squared errors



$$\text{SSE}(C) = 5215.016$$

$$PRE = 1 - \frac{2396.946}{5215.016} \approx 0.54$$

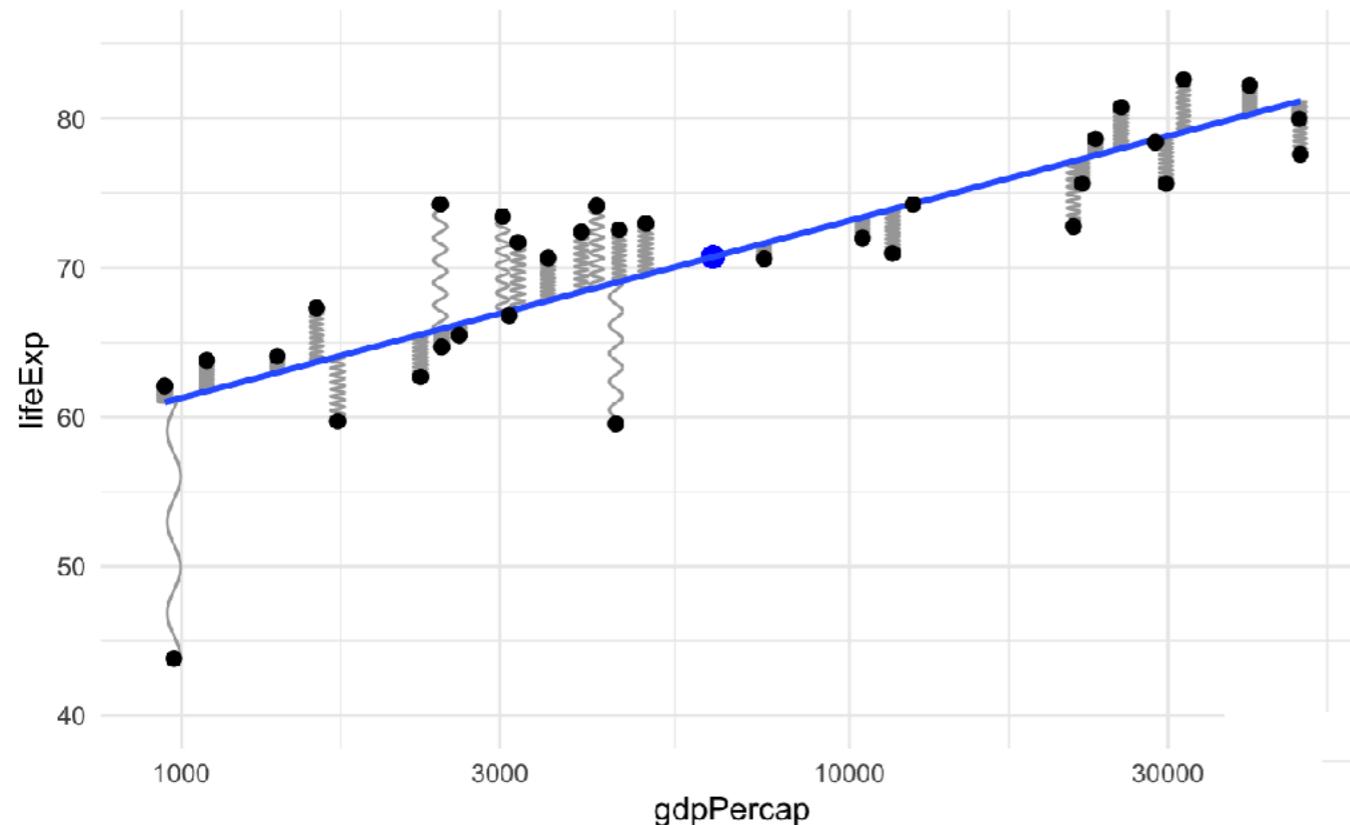
Sum of squared errors



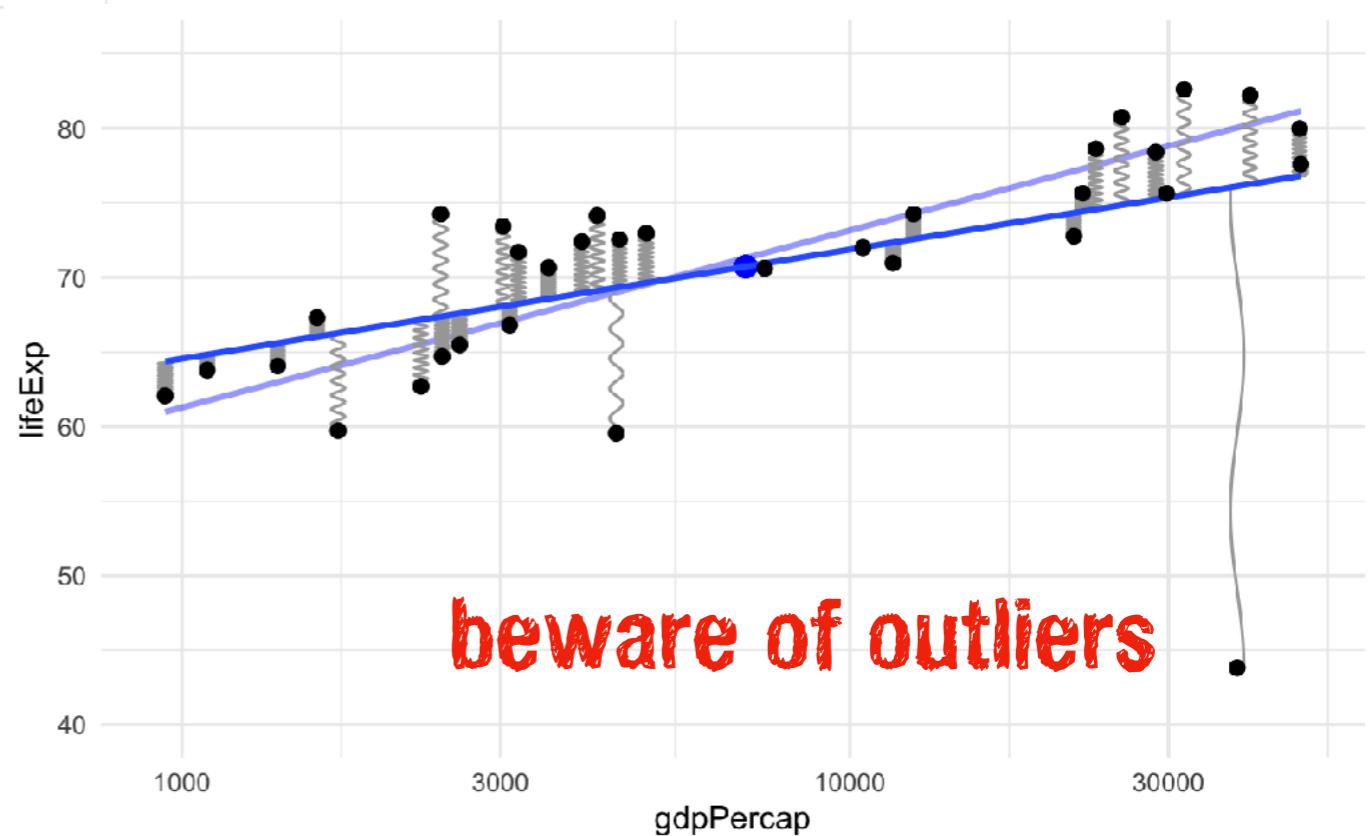
$$\text{SSE}(A) = 2396.946$$

The augmented model
reduces the error by 54%.

Least squares as springs



each point is
attached to the
line with an
identical spring



The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

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**

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

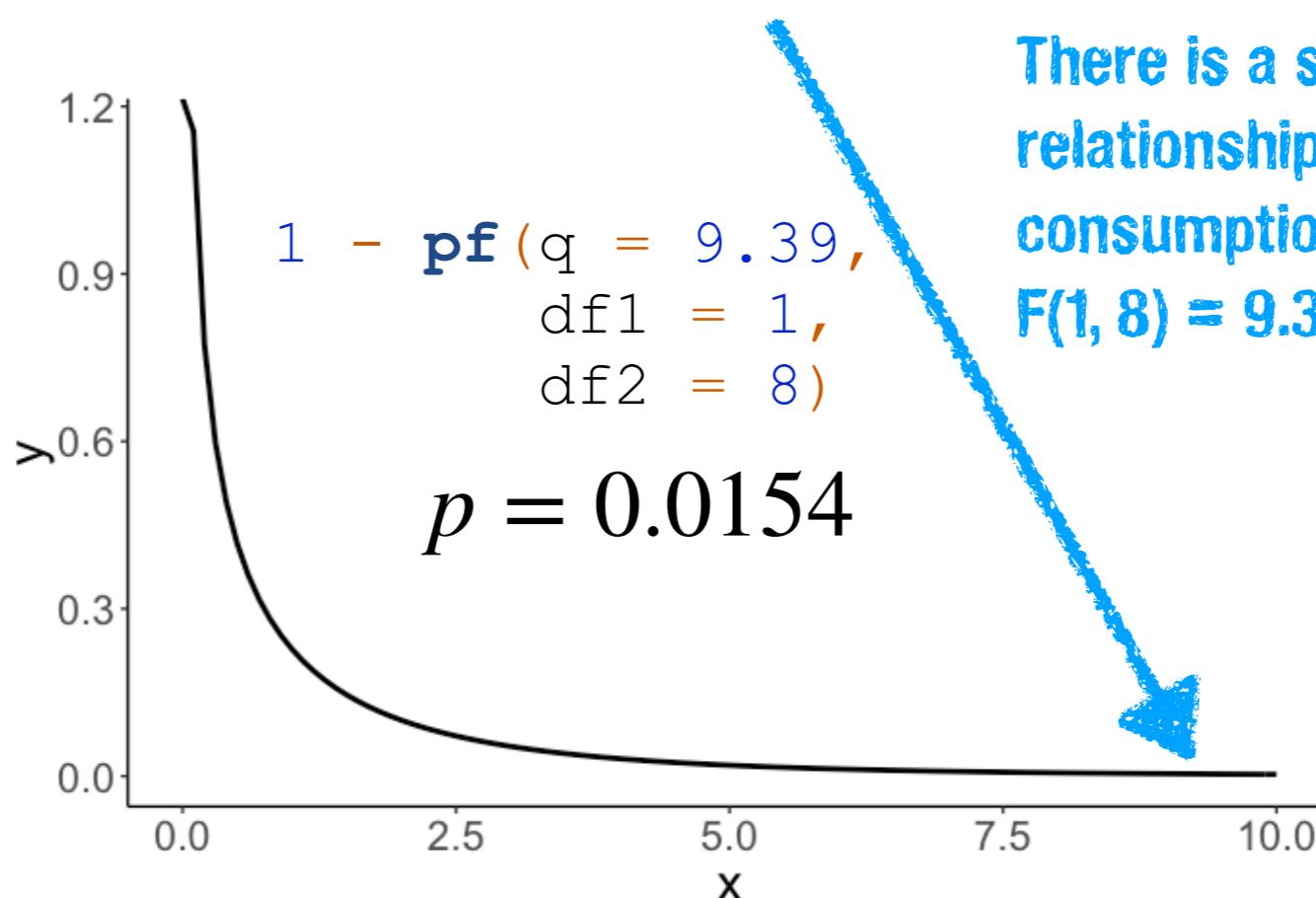
**number of observations
vs. parameters in Model A**

Decide whether it's **worth it**

- To compute the F statistic, we need:

- PRE = 0.54
- PC = 1
- PA = 2
- $n = 10$

$$\begin{aligned} F &= \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} \\ &= \frac{0.54/(2 - 1)}{(1 - 0.54)/(10 - 2)} \\ &= 9.39 \end{aligned}$$

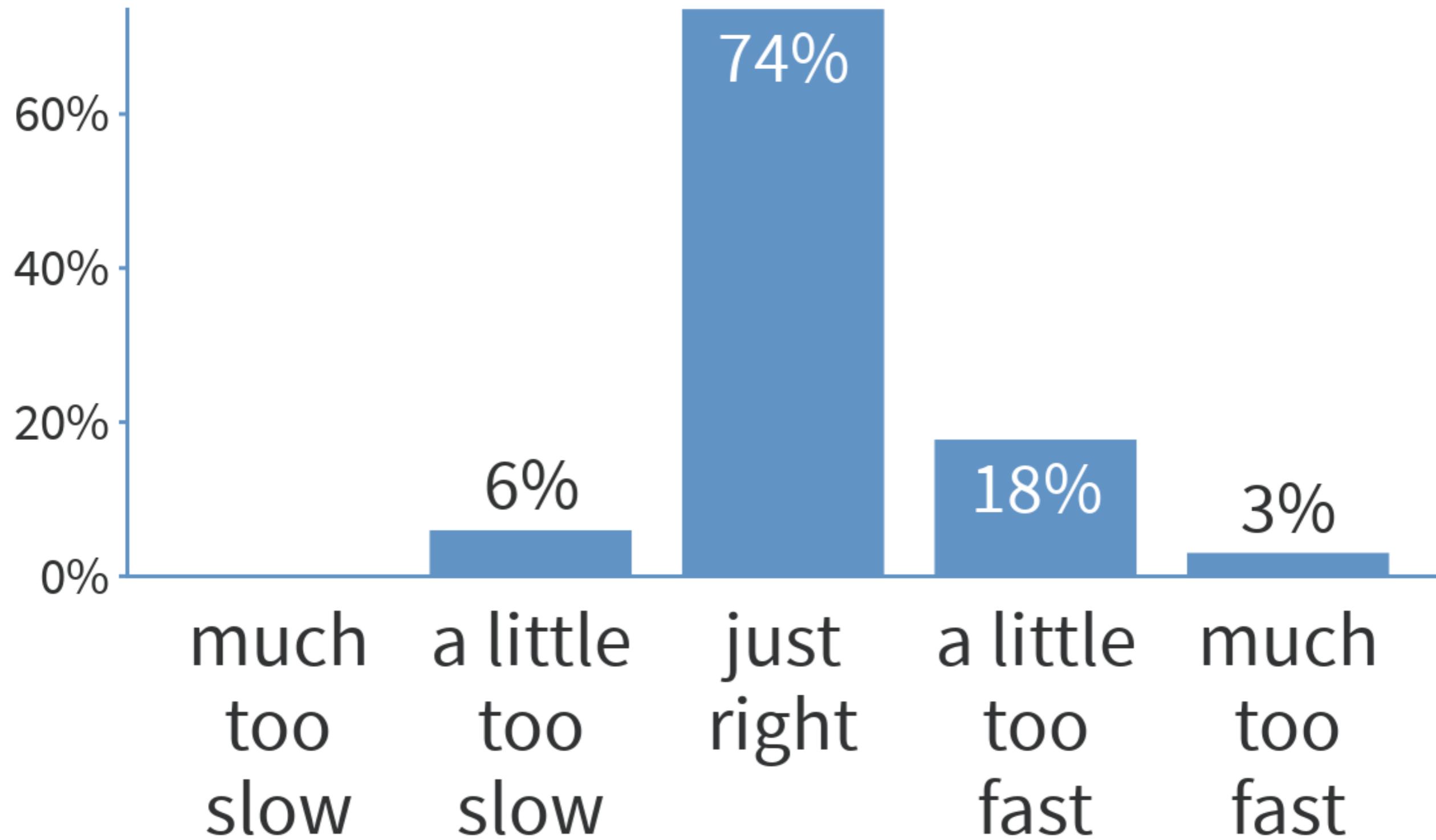


Summary

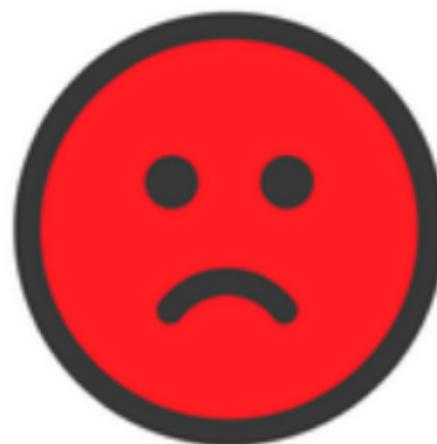
- Quick recap
- Modeling data
- Hypothesis testing as model comparison
- Correlation
 - Pearson's moment correlation
 - Spearman's rank correlation
- Regression

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?

cover
process
foreground
important
limit
lead
junk
understand
explain
interaction
better
poll
questions
concepts
class
really
lecture
interaction
time
fully
us
fix
assume
low
clear
well
tobi
things

Thank you!