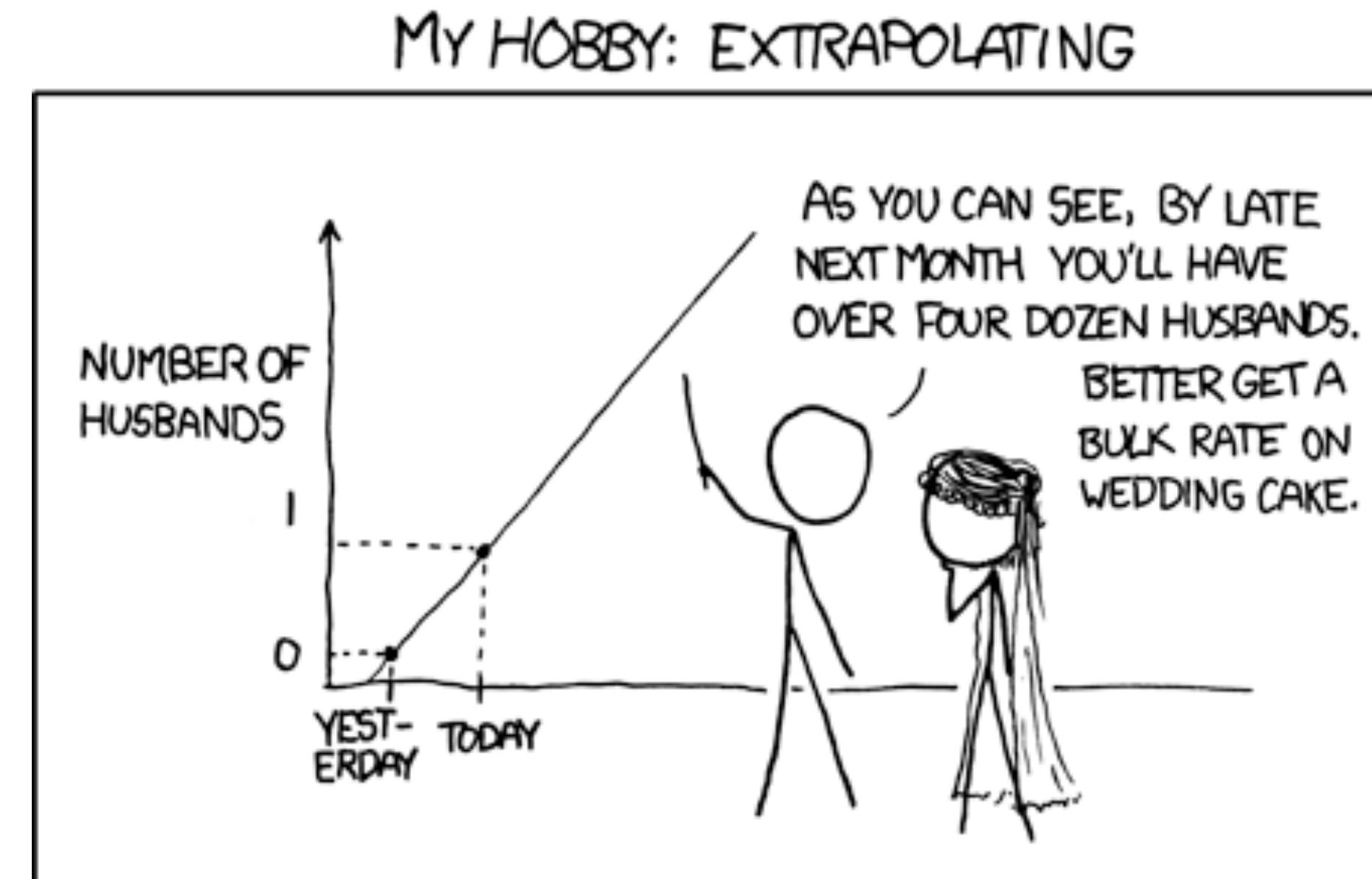


# Linear model 1



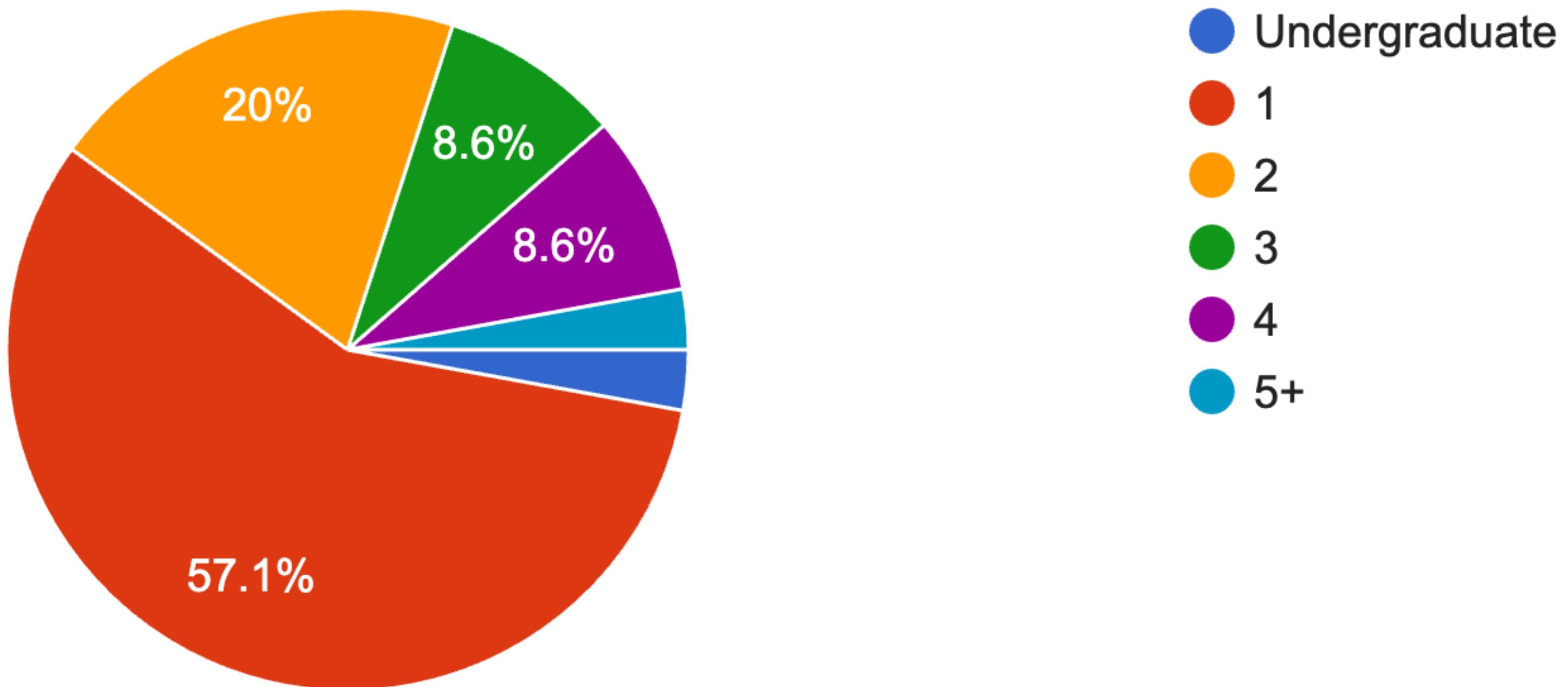
01/29/2025

# **Things that came up**

# Survey

What year of graduate school are you in?

35 responses

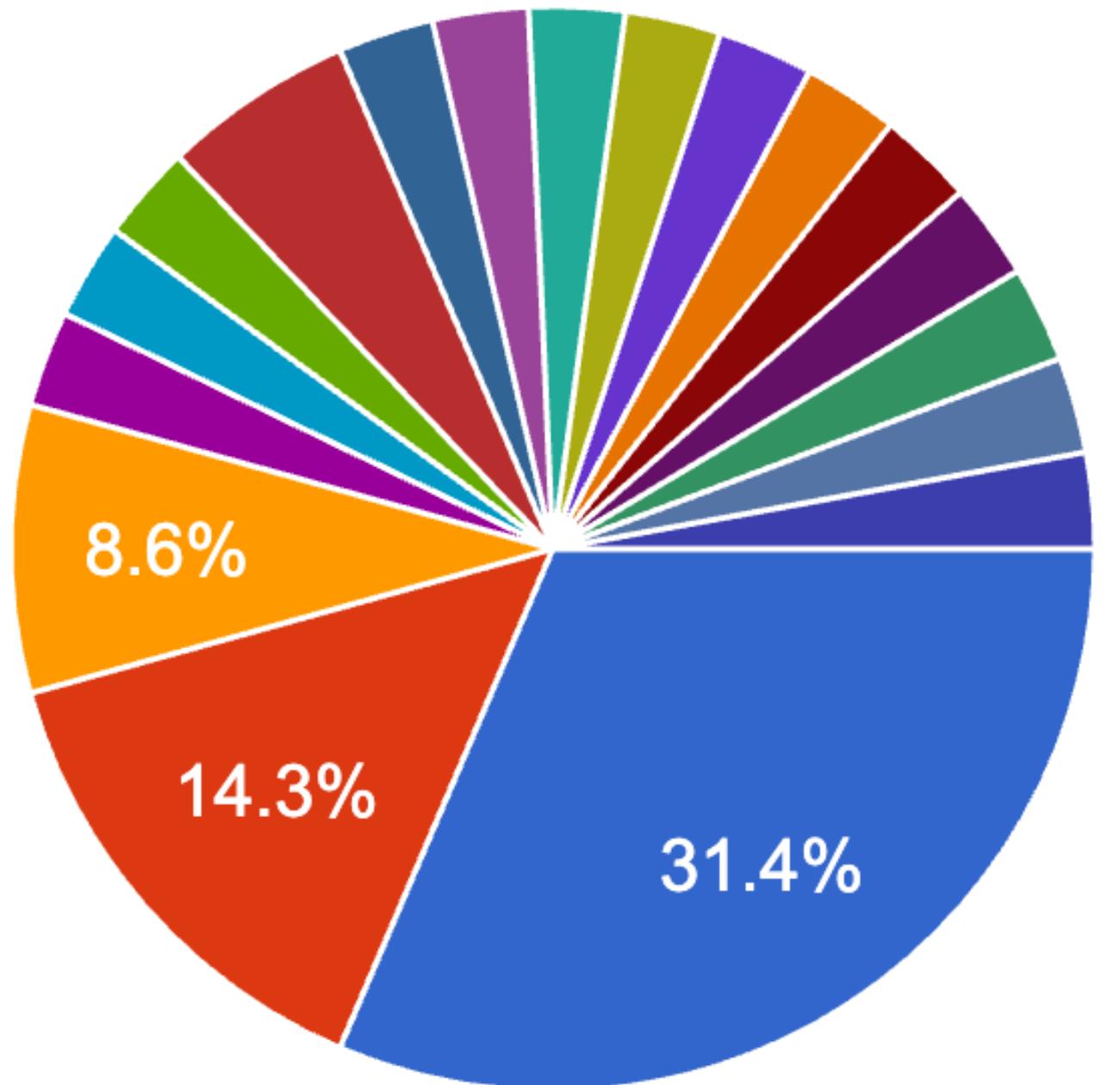


# Survey

What department are you in?

35 responses

- Neuroscience
- Music
- Earth Systems
- Sustainability
- Neuroscience (School of Medicine)
- Engineering: CME, ME
- Communication
- Community Health and Prevention Re...
- MS&E: Organizations, Technology, Entrepreneurship
- Communication PhD
- Symbolic System
- MS&E



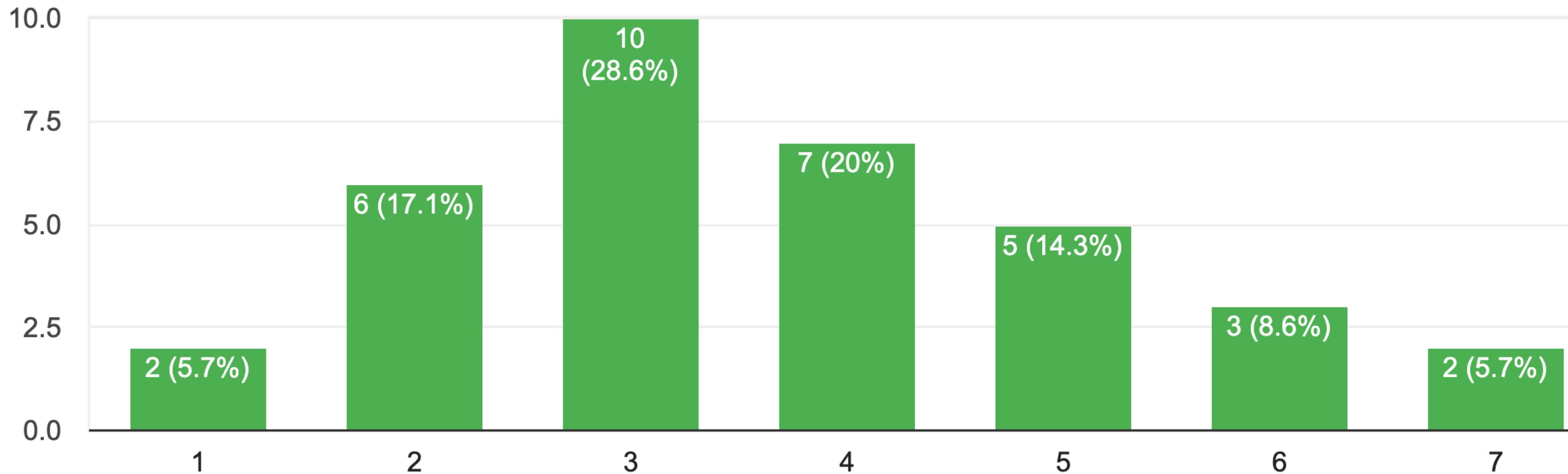
- Psychology
- Education
- GSB: Organizational Behavior
- GSB: Other
- Linguistics
- Computer Science: HCI
- Computer Science: Other
- Economics

▲ 1/3 ▼

# Survey

Please rate your level of experience with computer programming

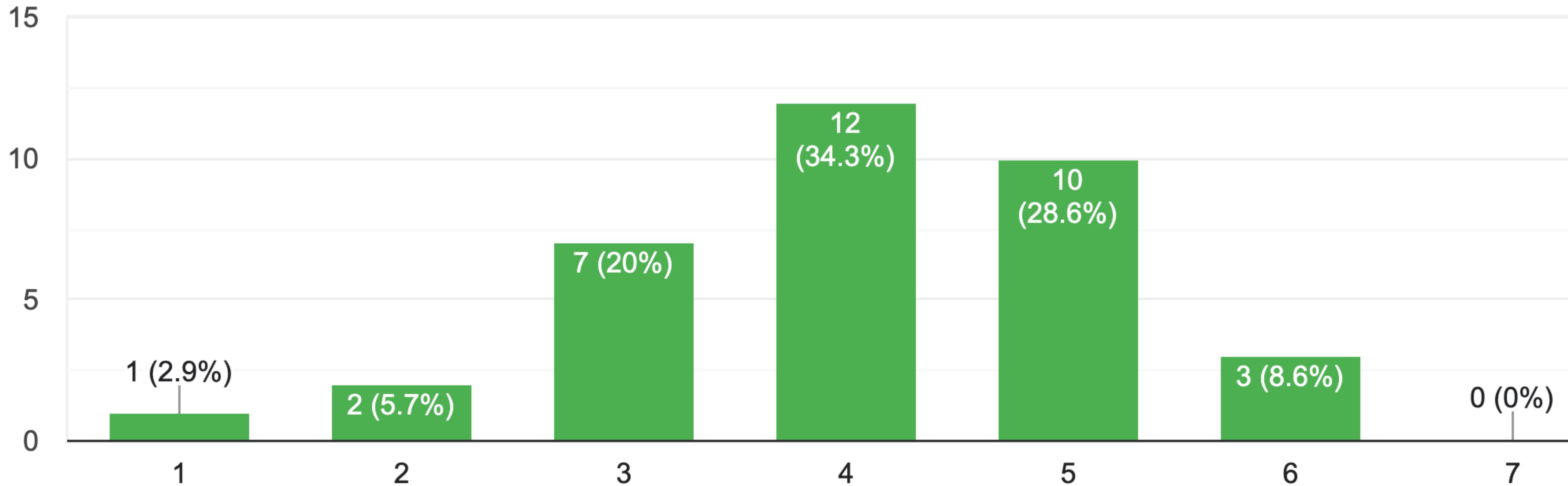
35 responses



# Survey

Please rate your level of experience with statistics

35 responses





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:

More likely to be red

22.6%

**More likely to be green**

**37.1%**

Equally likely

20.9%

Don't know/see results

19.5%

22,563 votes · Final results

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
```

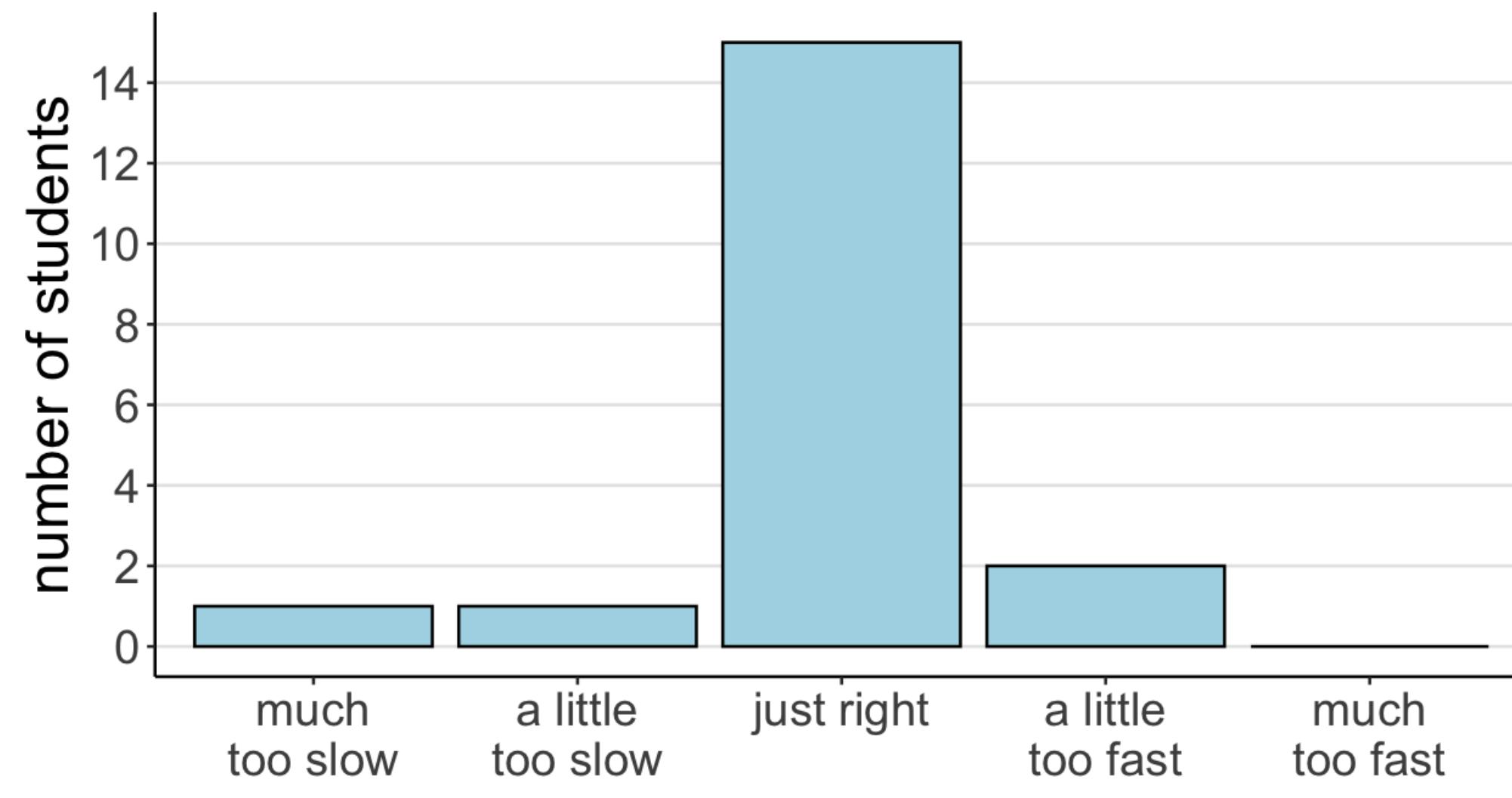
ALT

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

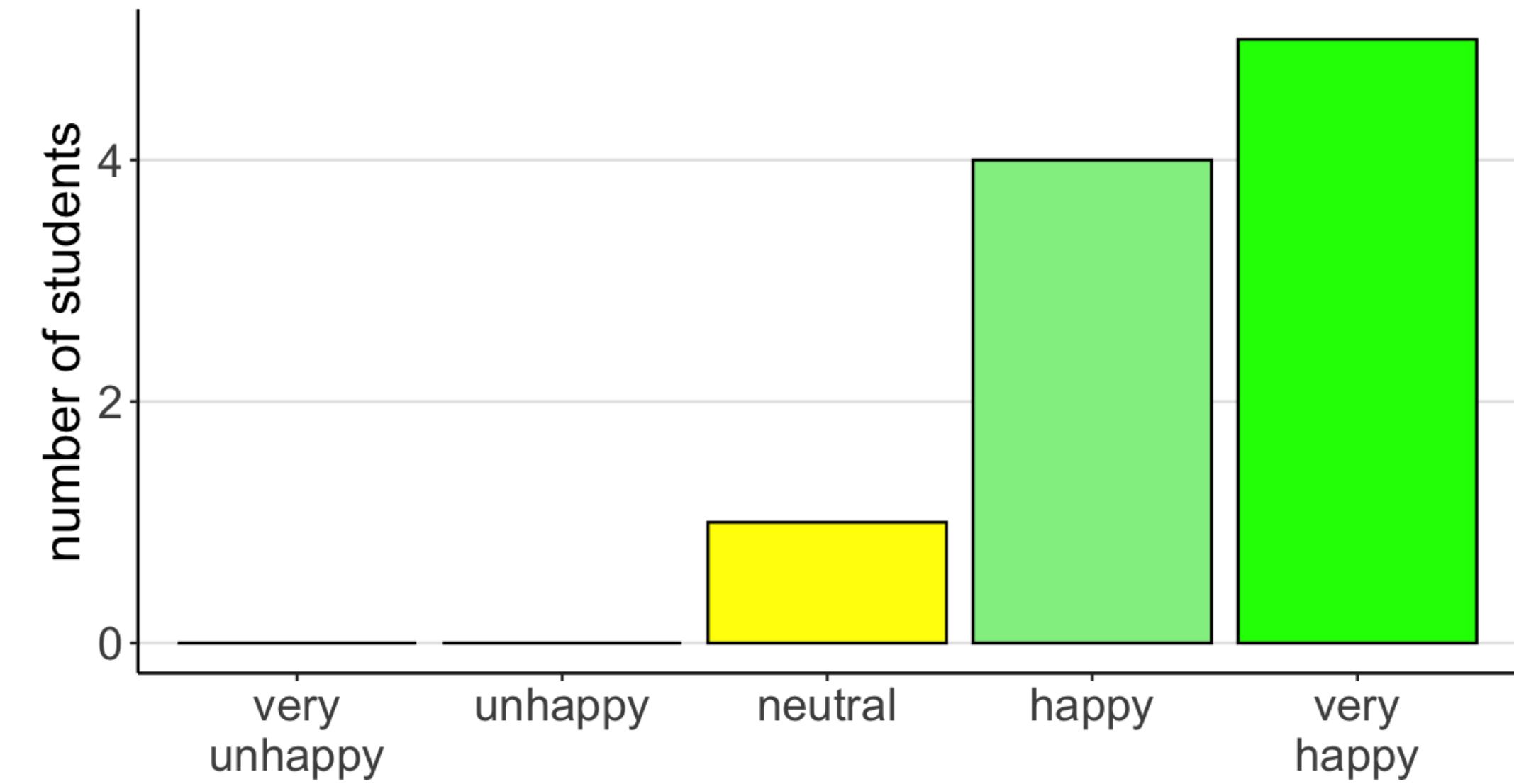
# Your feedback

# Your feedback

How was the pace of today's class?



How happy were you with today's class overall?



- Wonderful
- Showing the code/implementation behind mean\_cl\_boot was very helpful!
- It would be ideal to have a worked example to apply the different concepts throughout the class
- I like that the lecture focuses on intuition building. Sometimes the analogies/jokes are excessive
- I get it more but I'm lost still

# Plan for today

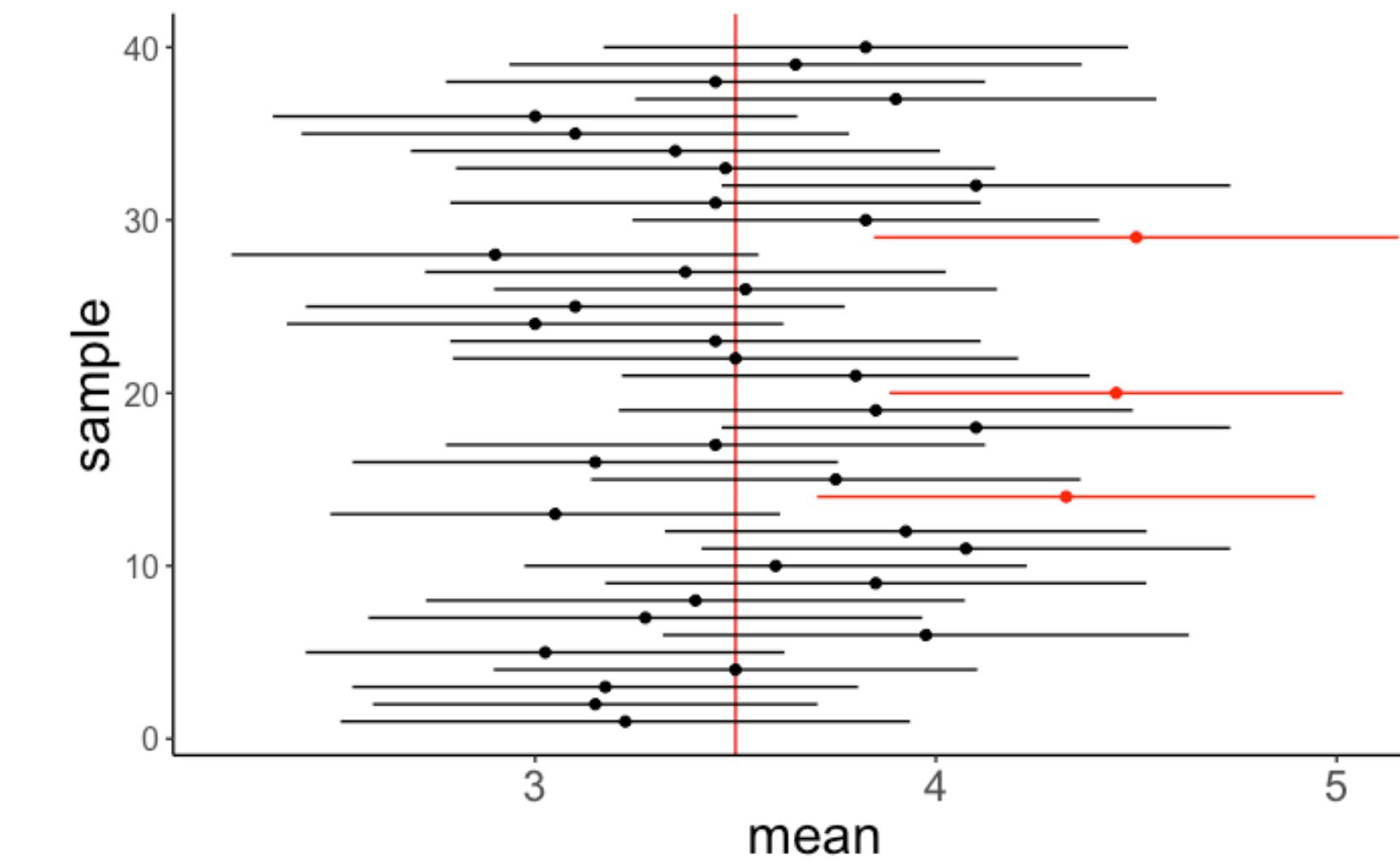
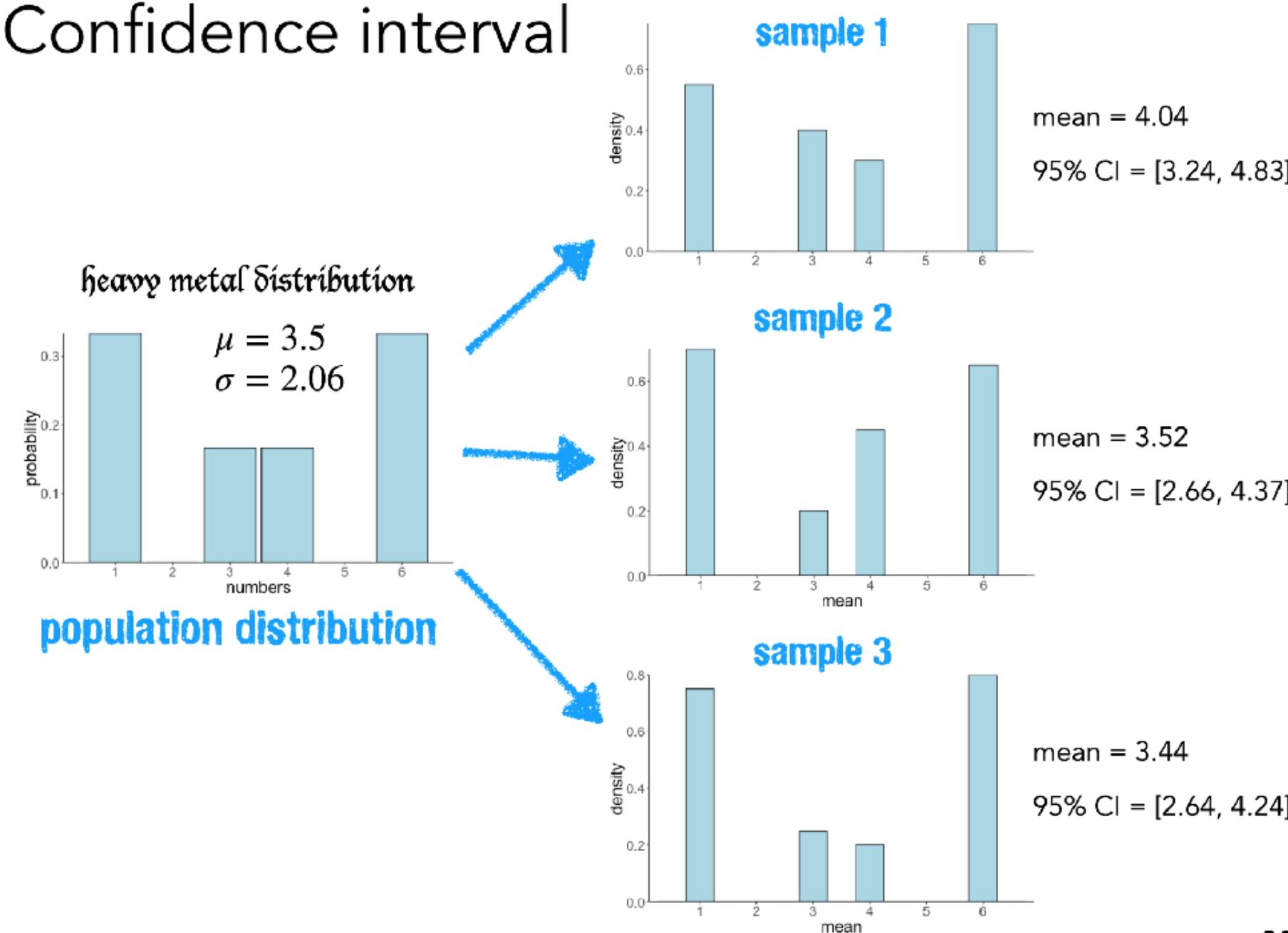
- Quick recap
- Modeling data
- Hypothesis testing as model comparison
- Correlation
  - Pearson's moment correlation
  - Spearman's rank correlation
- Regression

# Quick recap

# Quick recap: Confidence intervals

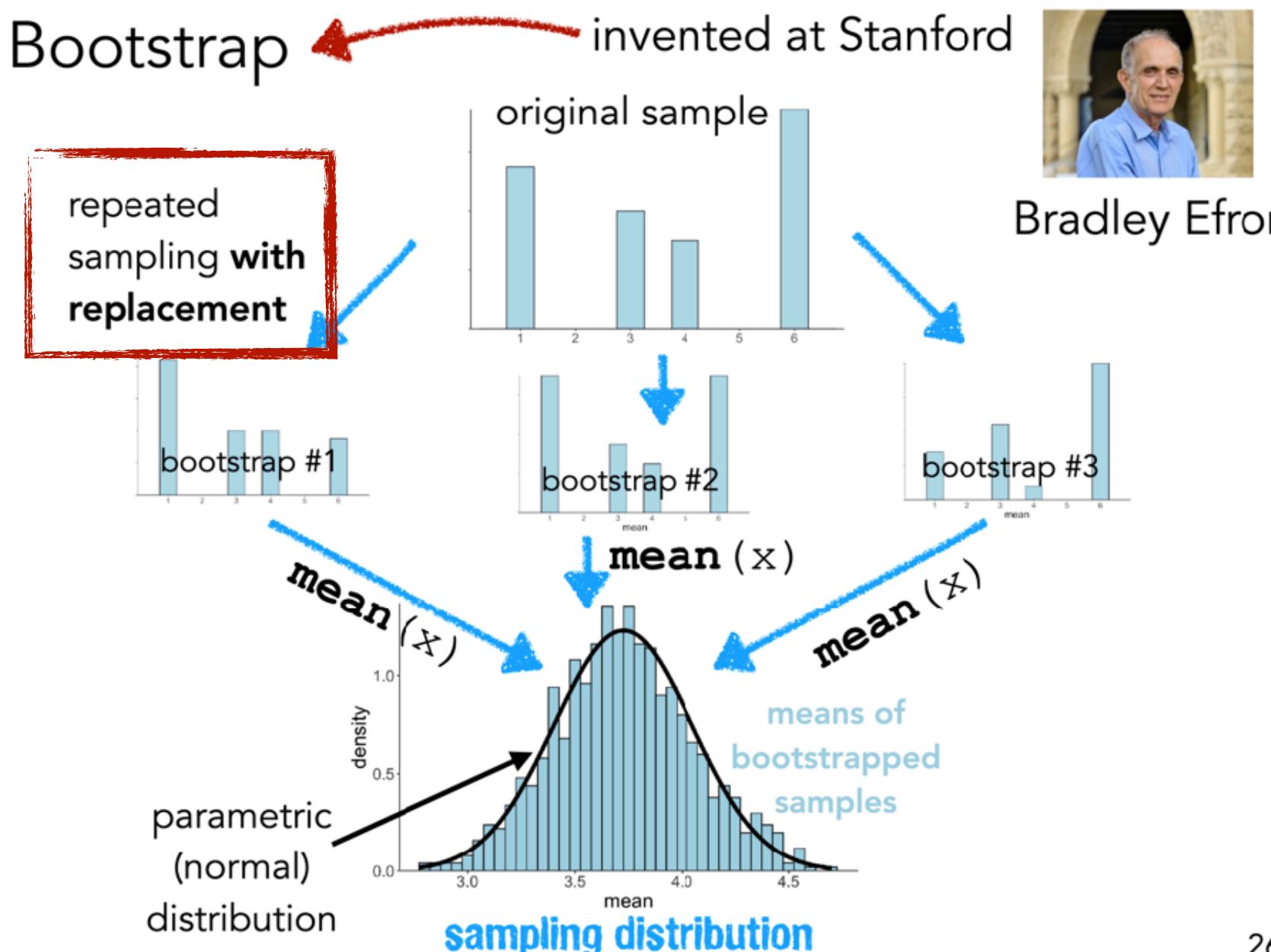
"If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the estimate of interest."

Confidence interval

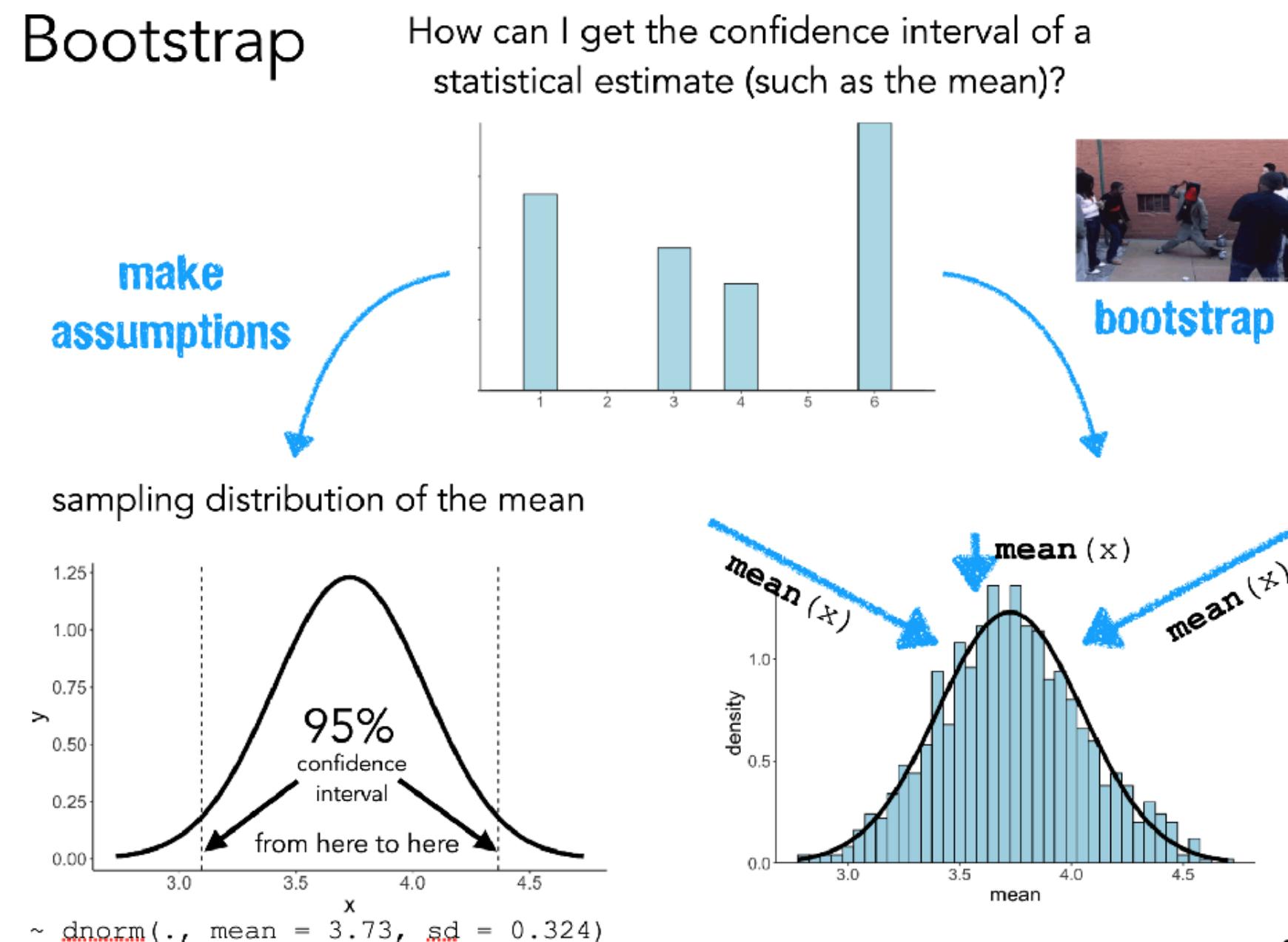


20

# Quick recap: Bootstrapping



26



27

make sure to set the seed!

mean\_cl\_boot() explained

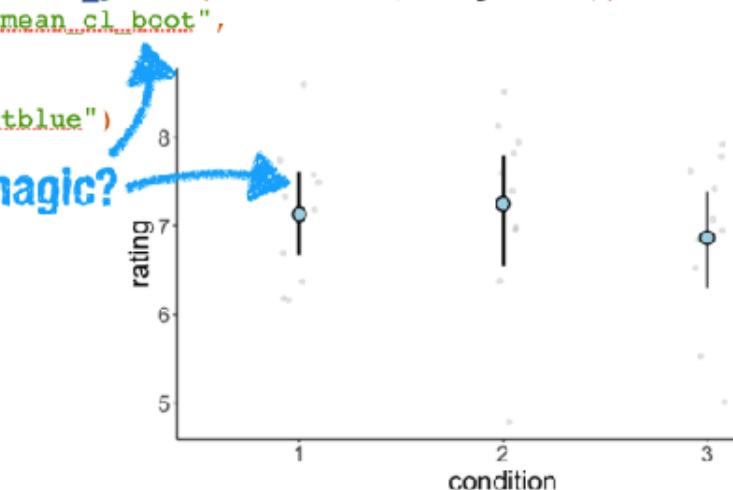
```

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")

```

participant	condition	rating
1	1	6.37
2	1	7.18
3	1	6.16
4	1	6.60
5	1	7.33

what is this magic?



29

16

# 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**

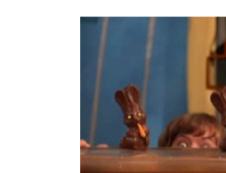
Assumption of normal distribution

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

↑  
assumed to be  
normally  
distributed

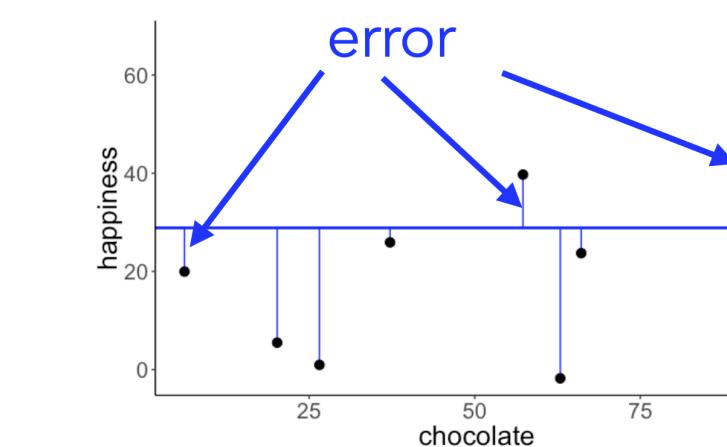
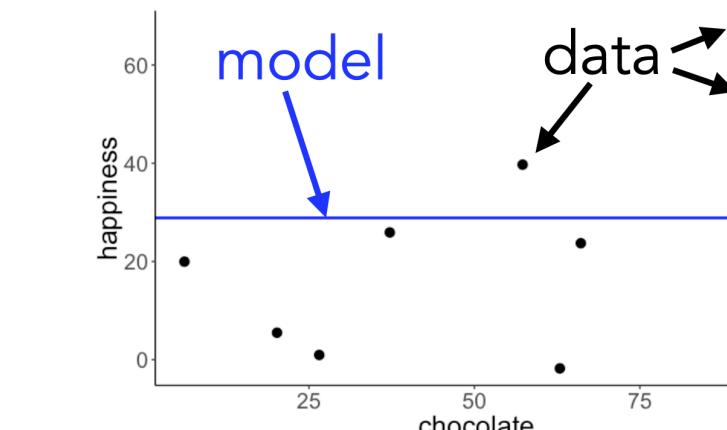
↑  
don't need to  
be normally  
distributed!!

very common misconception!!!



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

$H_0$ : Chocolate consumption and happiness are unrelated.

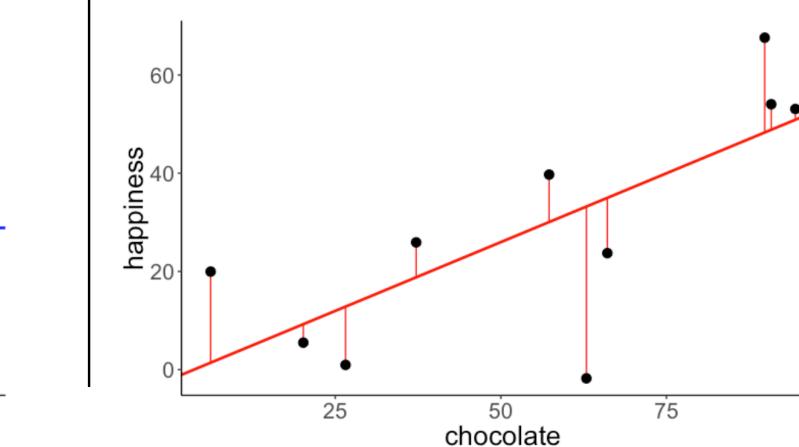
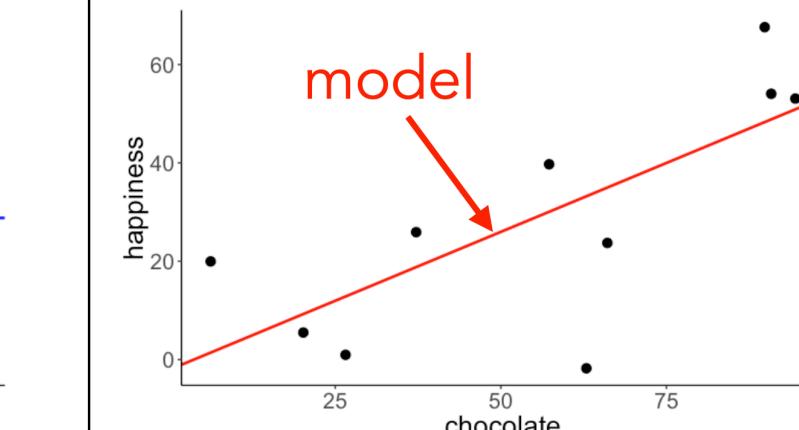


ERROR

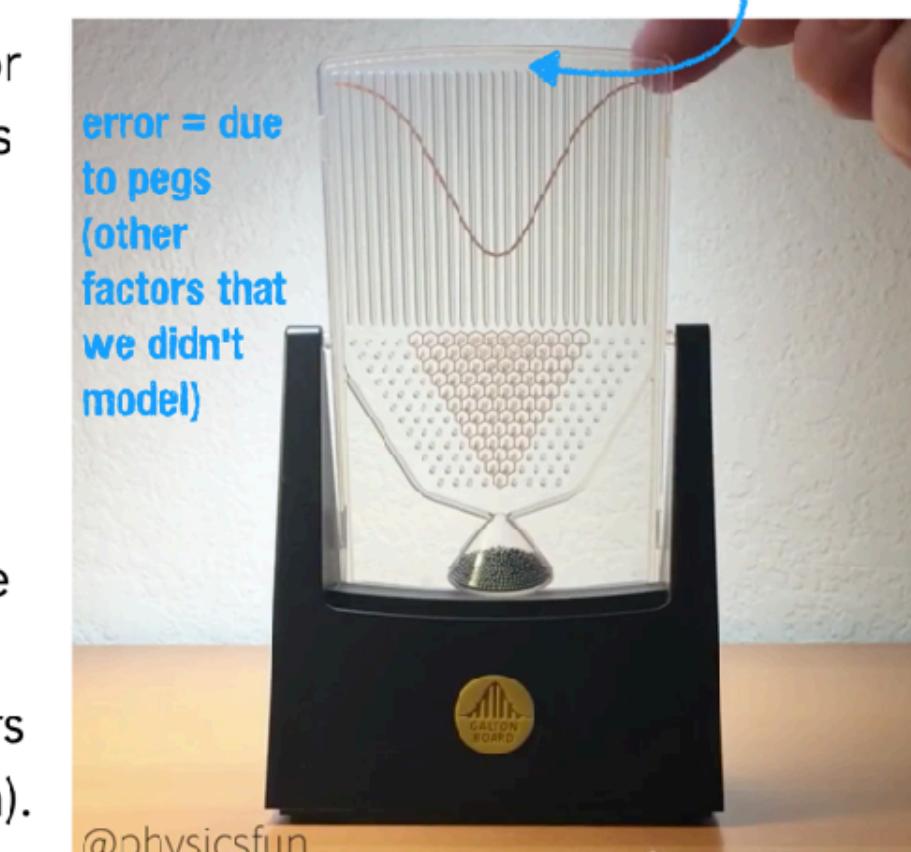
1. We assume that the error between model and data is due to (a potentially large number of) factors that we didn't take into account.

2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).

$H_1$ : Chocolate consumption and happiness are related.

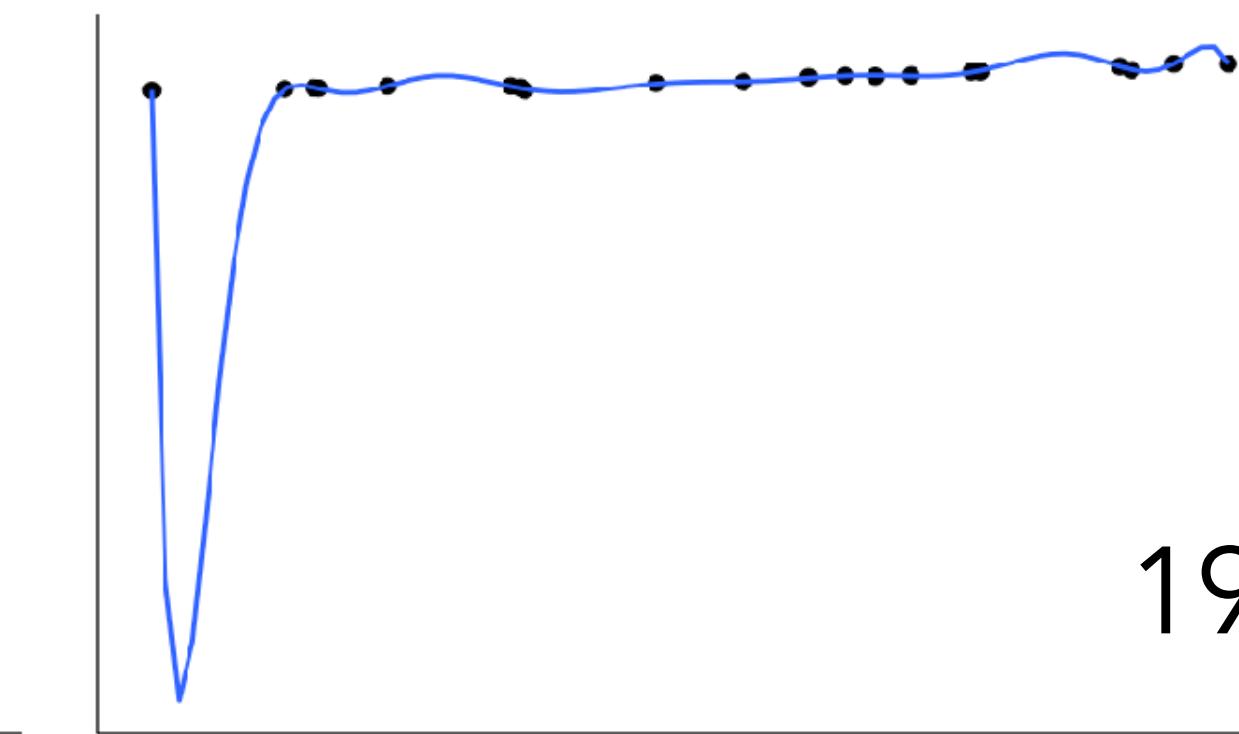
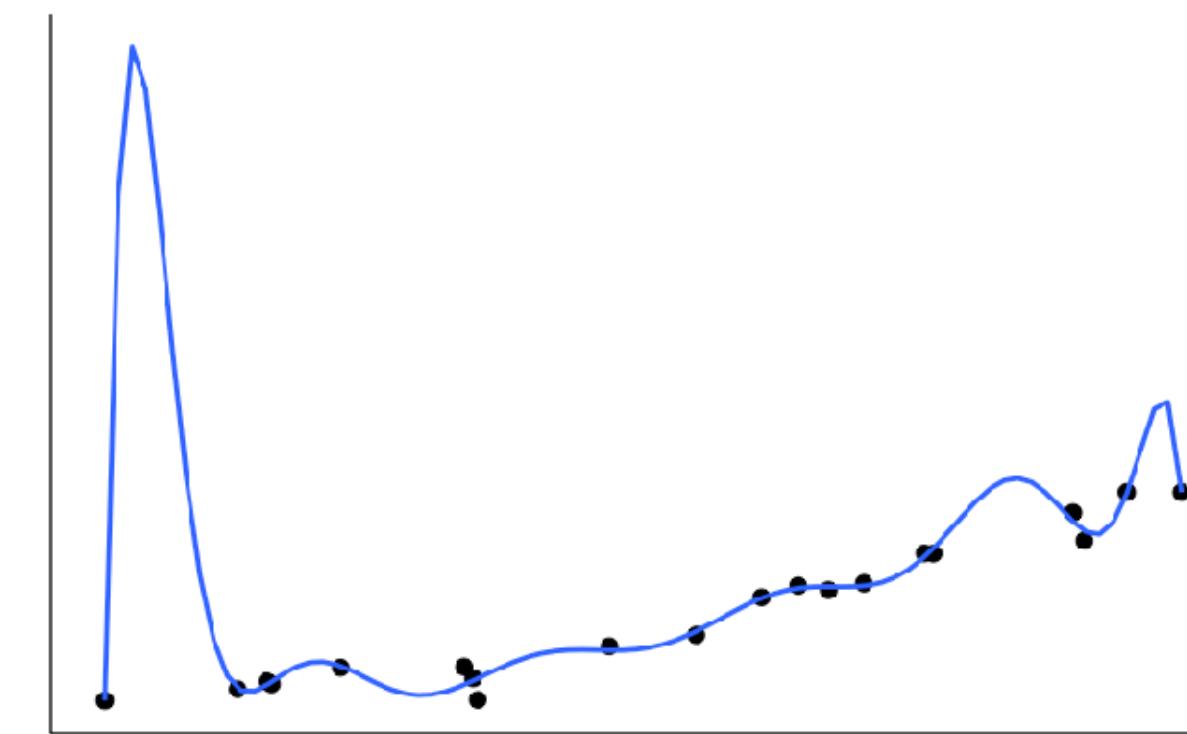
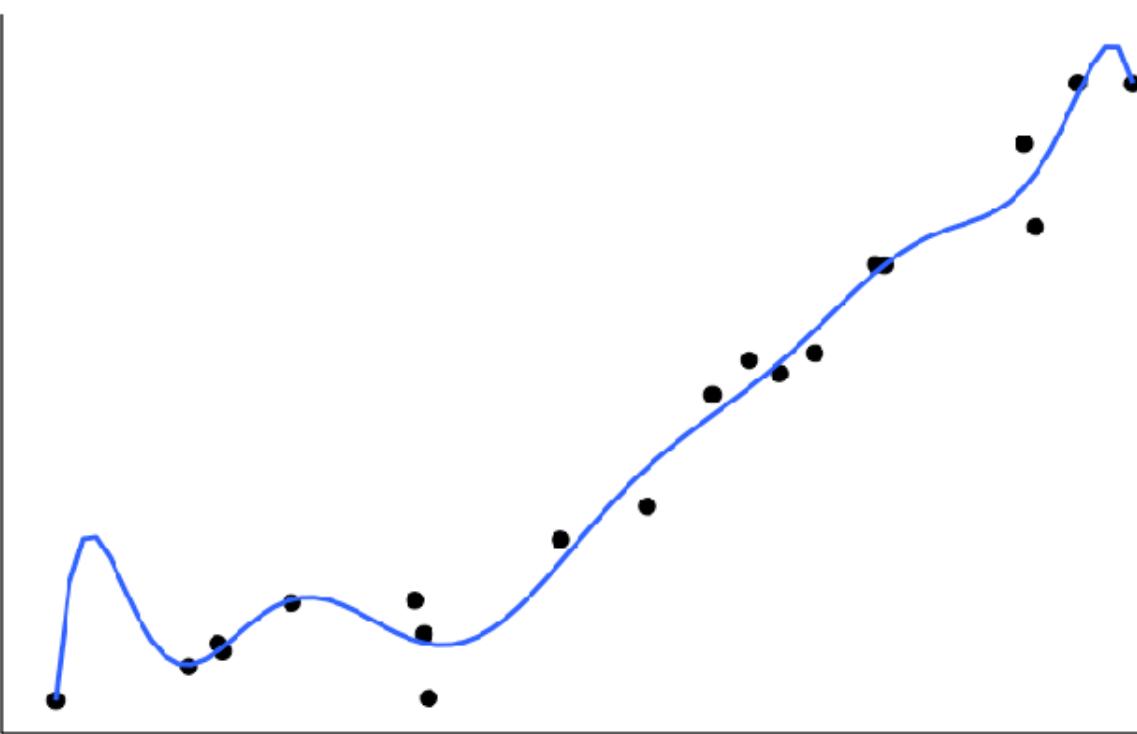
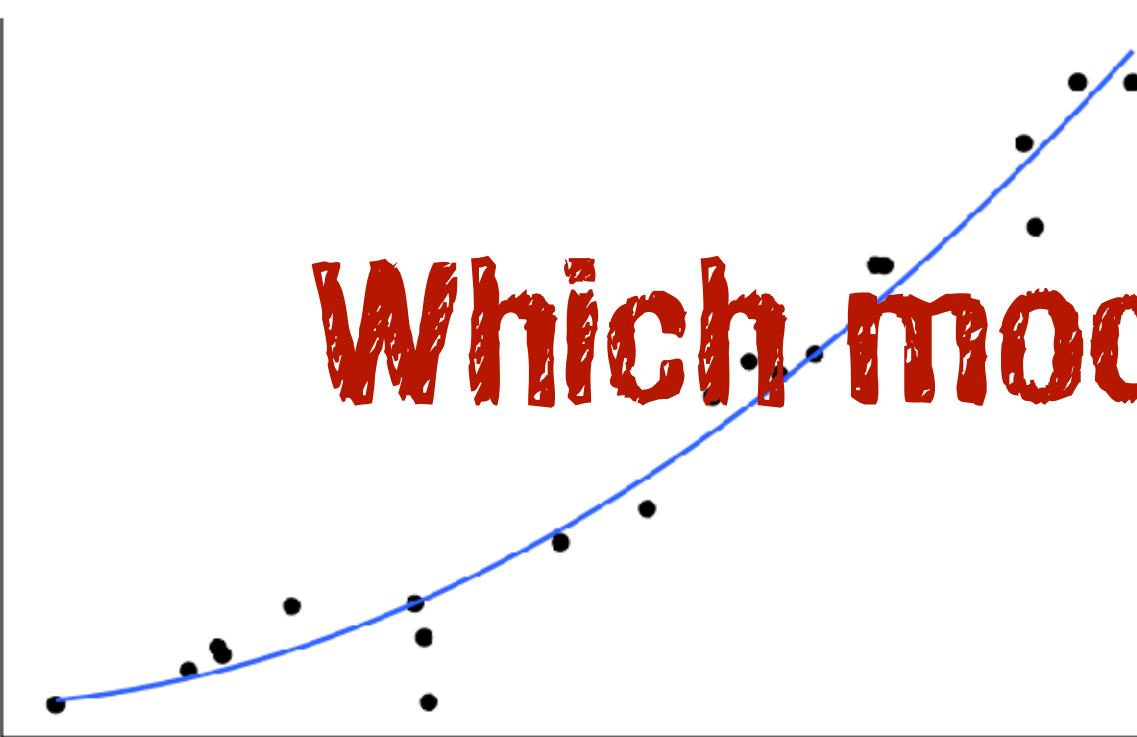
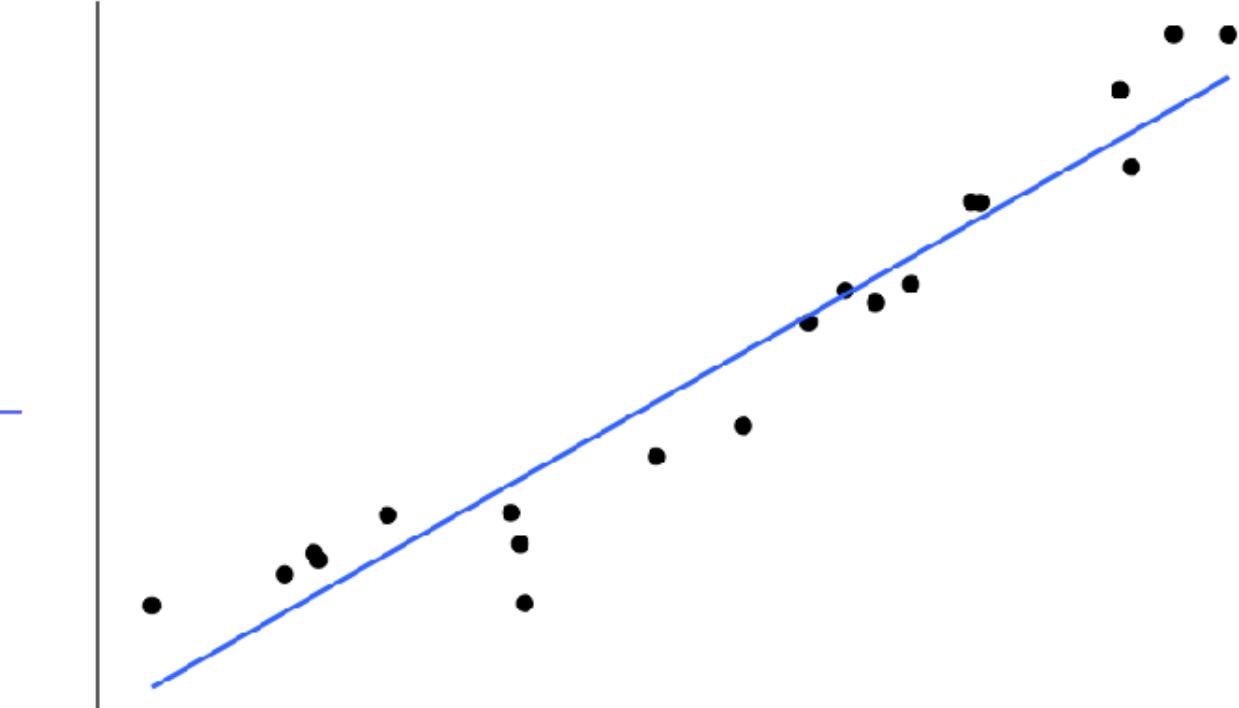
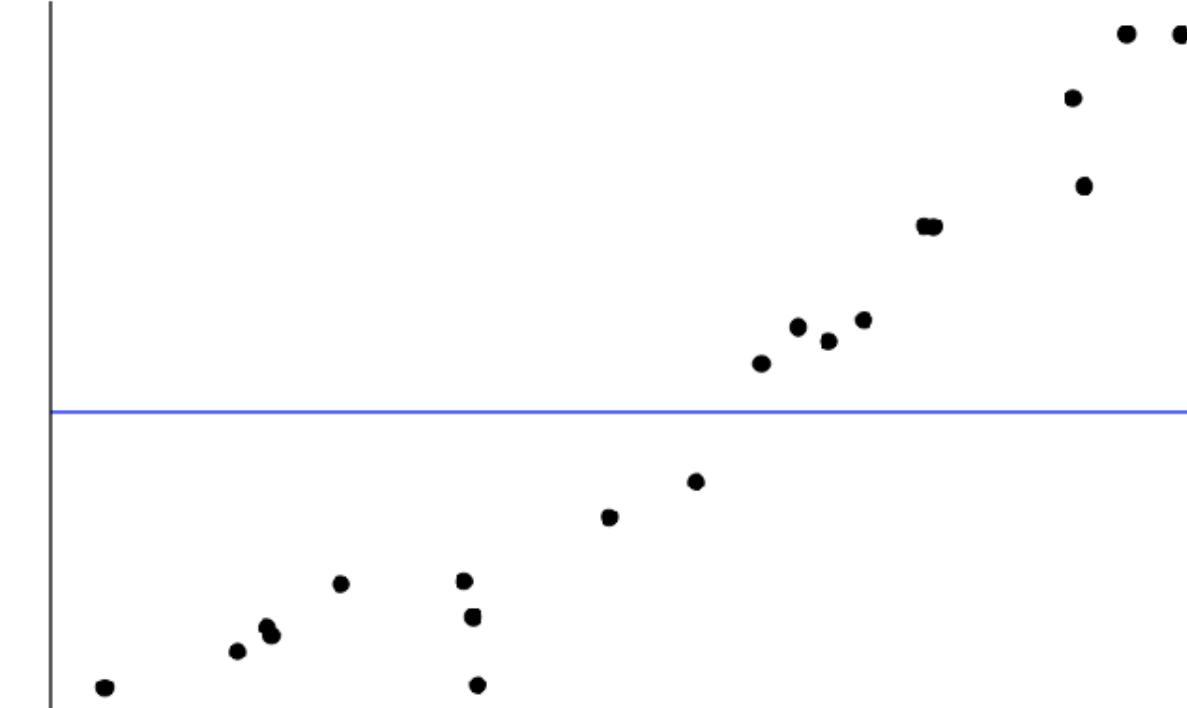
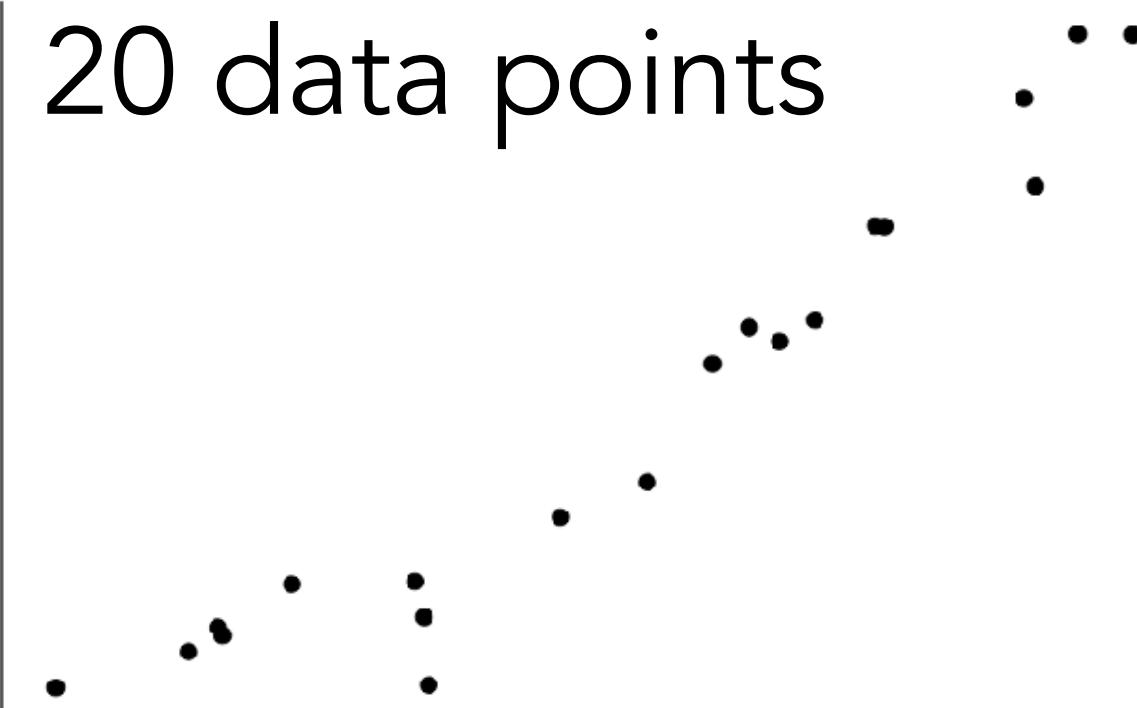


model = "bottleneck"



Result: normal distribution

# Modeling data



19 parameters



**THE BEST WAY TO  
EXPLAIN OVERFITTING**

# 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?**

# worth it?

## Compact model

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

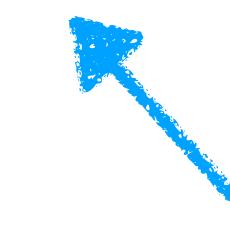
## Augmented model

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

## Proportional reduction in error (PRE)

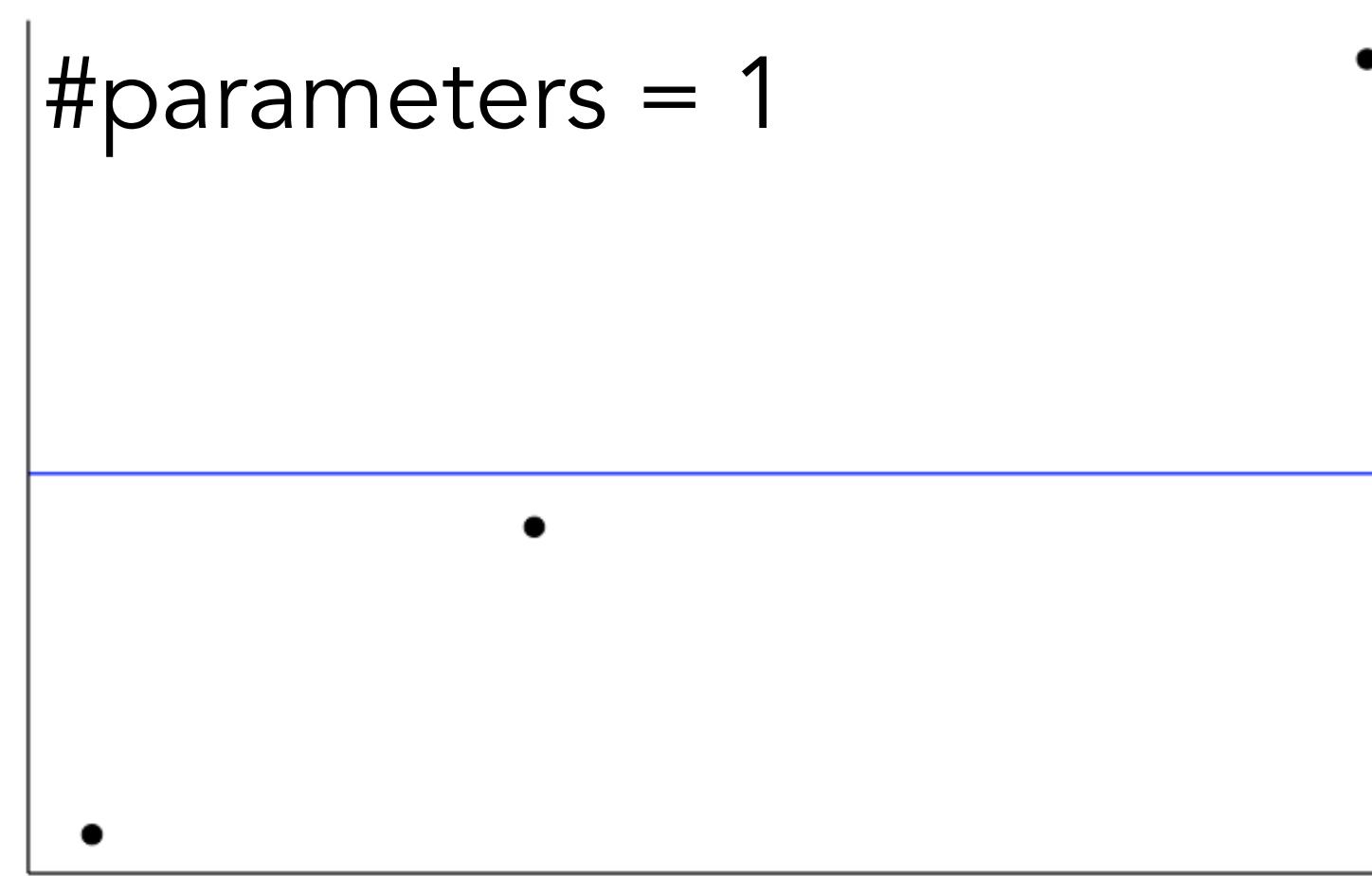
$$\begin{aligned}\text{PRE} &= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)} \\ &= 1 - \frac{30}{50} = .40\end{aligned}$$

- 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

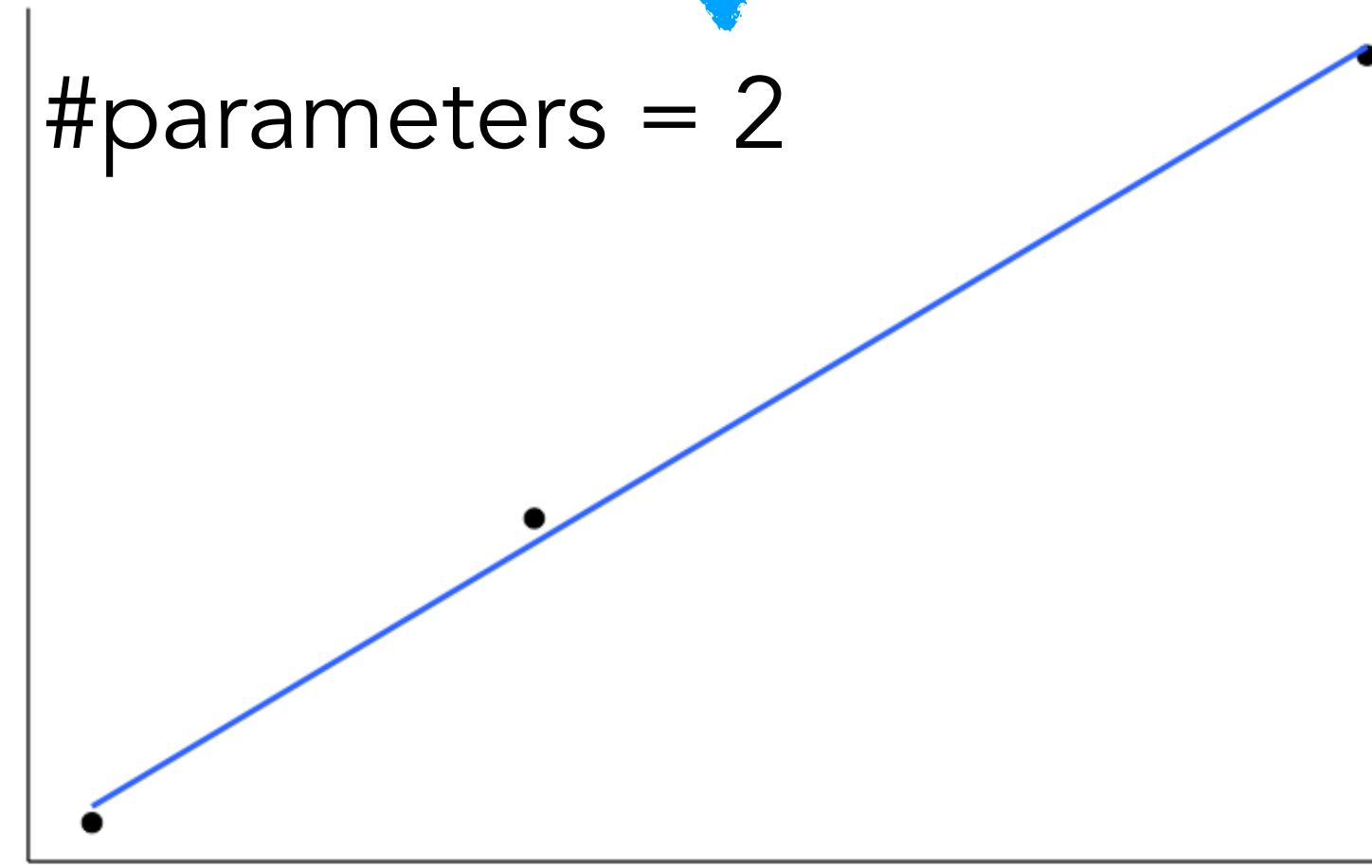


more impressed if the number of observations n is much greater than the number of parameters

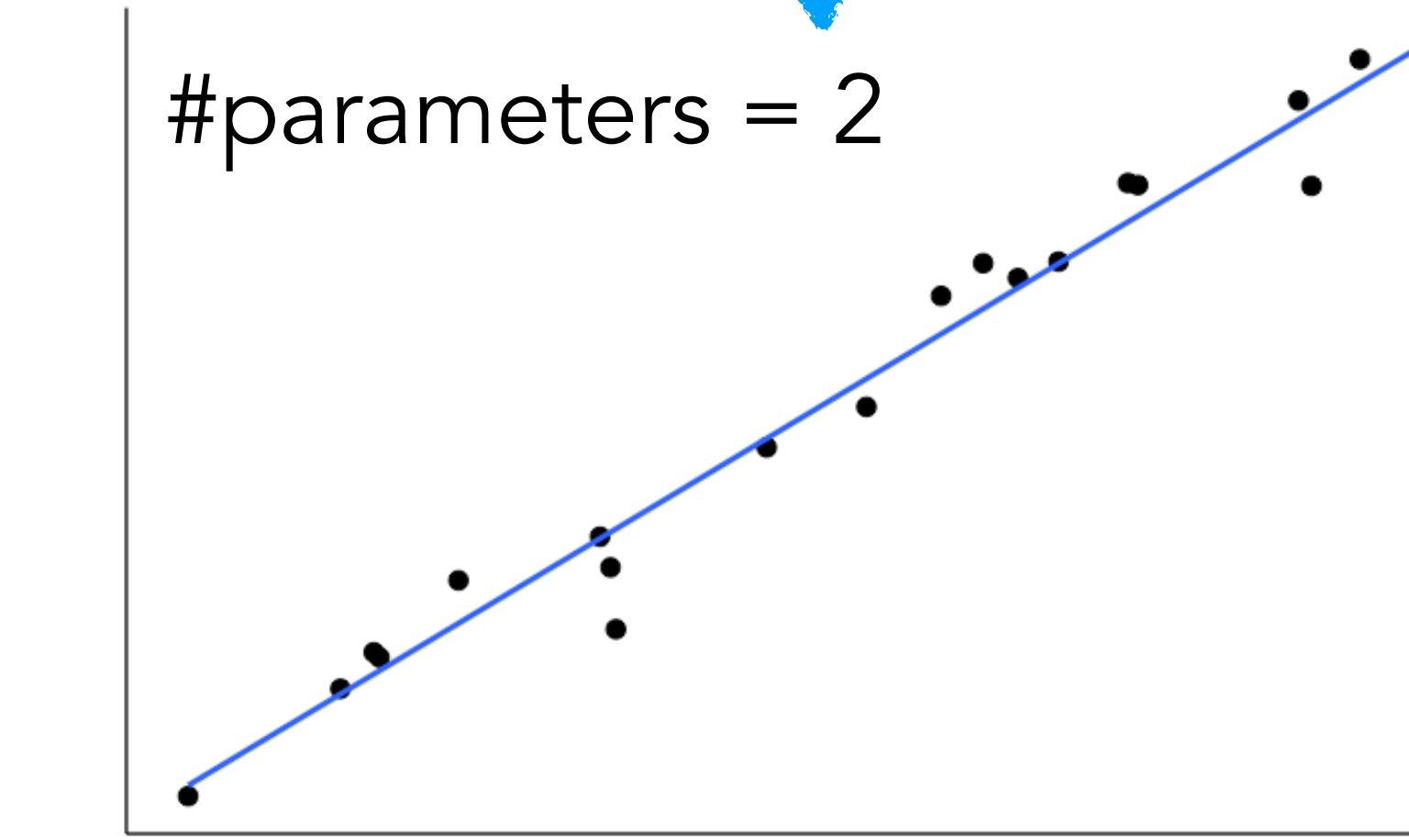
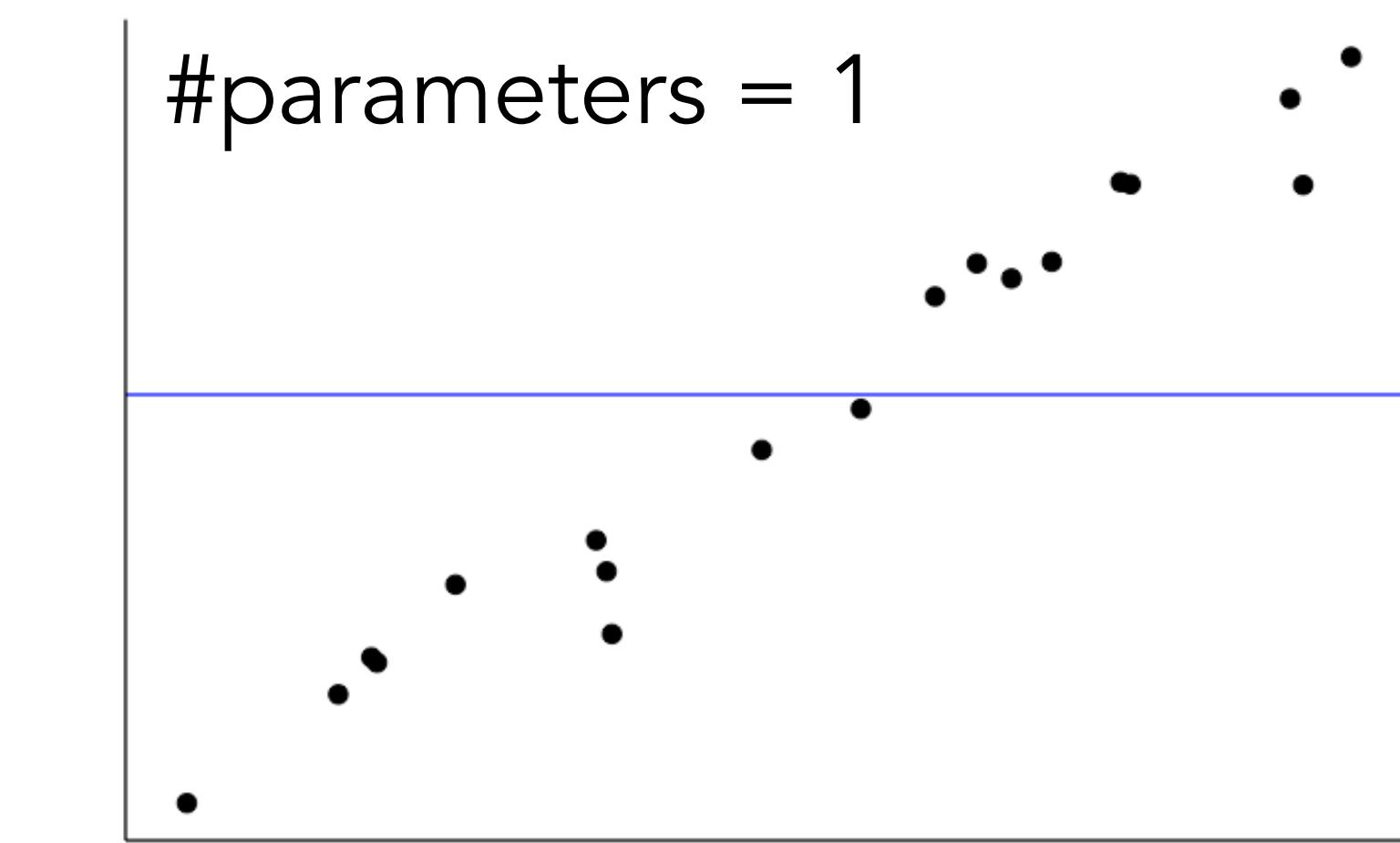
# PRE per parameter for different $n$



↓ neato!



↓ impressive!



# General procedure

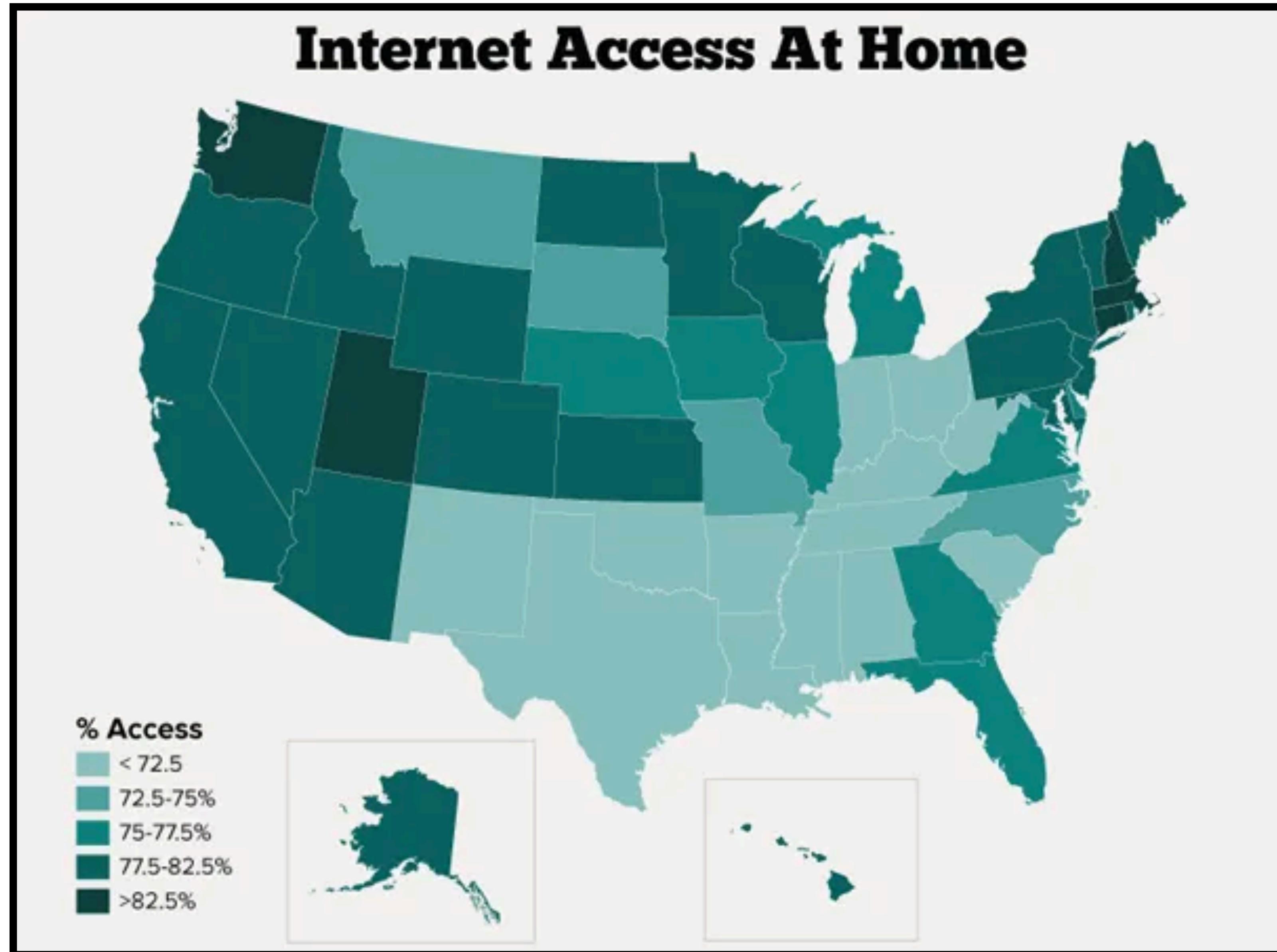
- for any question we want to ask about our DATA
    - we define model<sub>C</sub> and model<sub>A</sub>
    - compare the models using PRE
    - determine whether PRE is **worth it**
  - in standard frequentist lingo:
    - model<sub>C</sub> =  $H_0$  (null hypothesis)
    - model<sub>A</sub> =  $H_1$  (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

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

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

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

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



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.

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

# Research question and hypotheses

Is the average percentage of internet users per state significantly 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**

- we need a sampling distribution of PRE
  - a distribution of what PRE would look like if Model C (our  $H_0$ ) were true
- PRE is closely related to the *F* statistic!

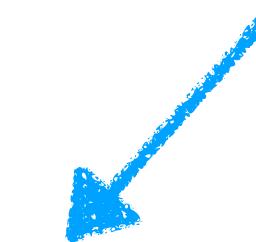
# Decide whether it's **worth it**

- 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<sub>C</sub> to create model<sub>A</sub> but were not

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



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

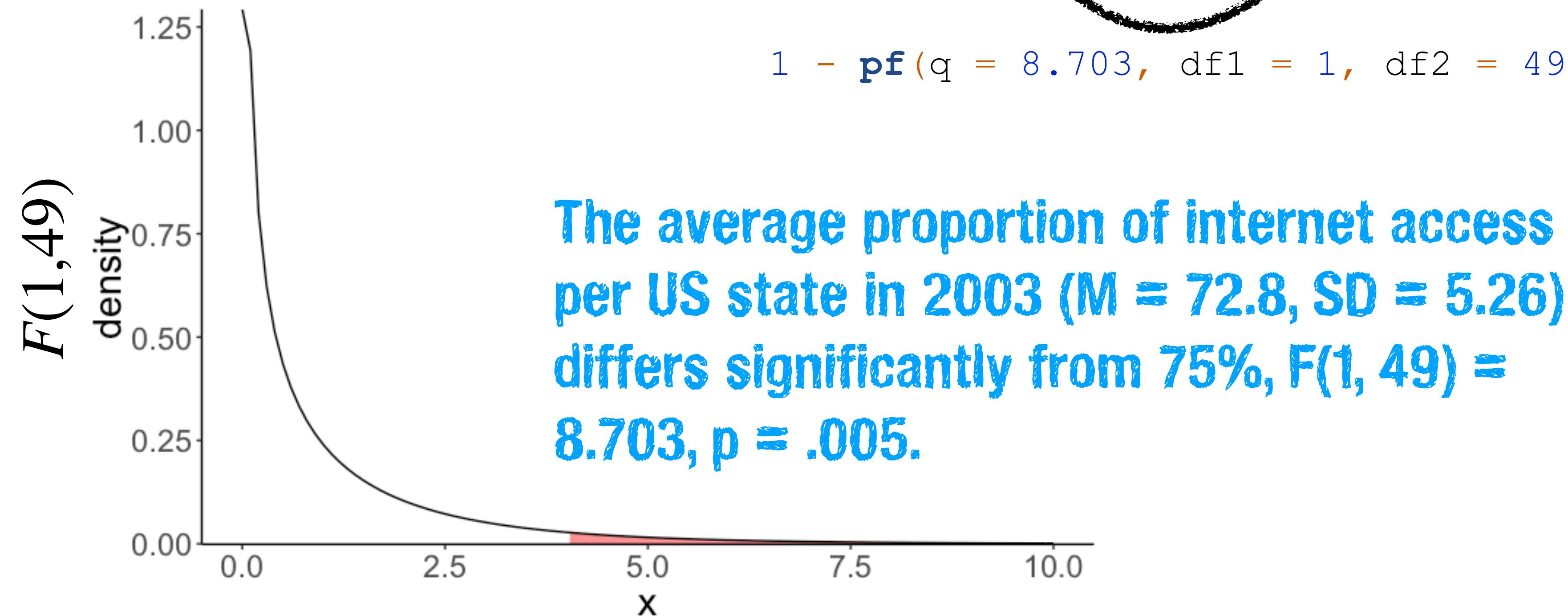


**number of observations vs.  
parameters in Model A**

# 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

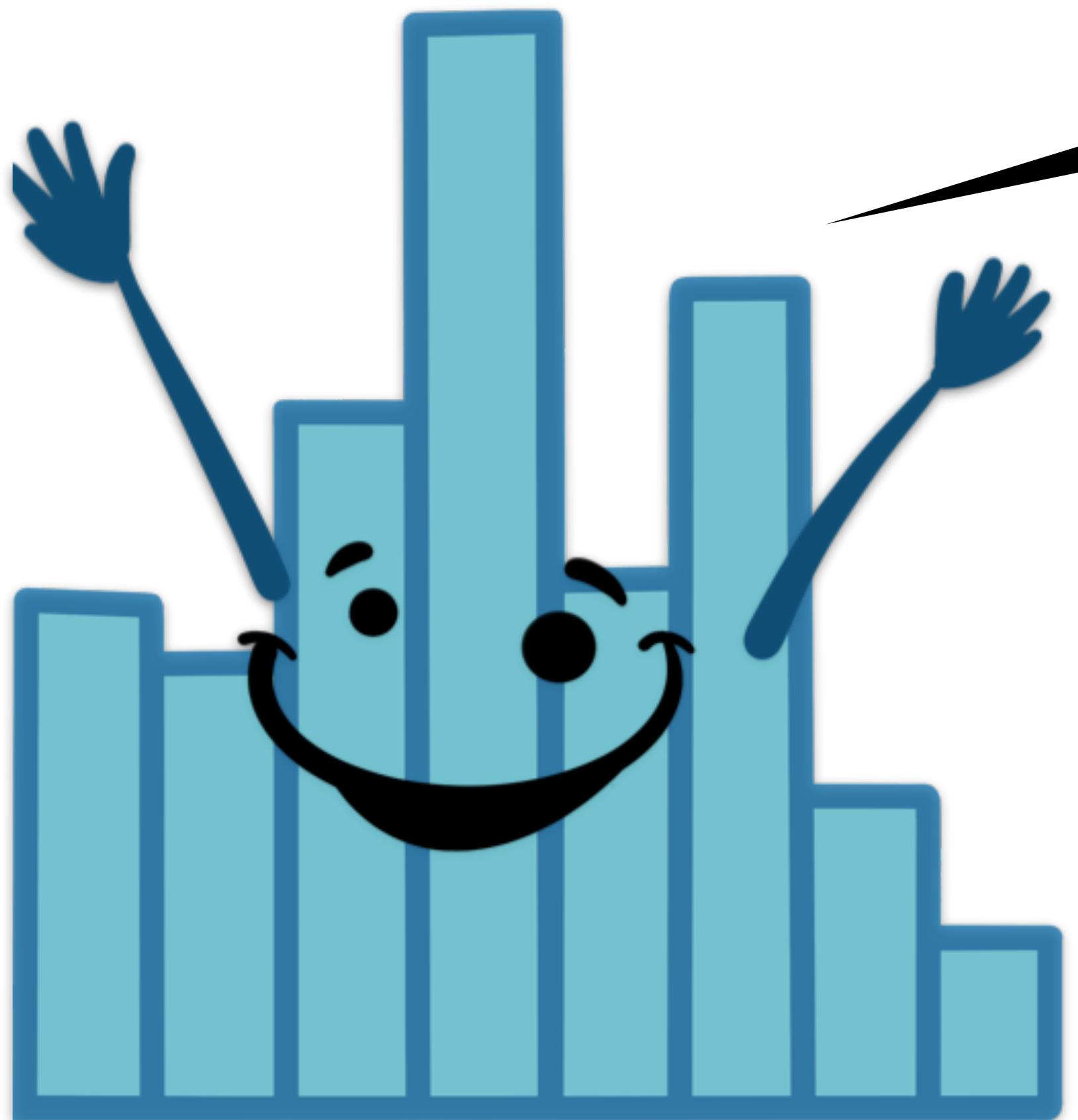
```
data: df.internet$internet  
t = -2.9502, df = 49, p-value = 0.00486  
alternative hypothesis: true mean is not equal to 75
```

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!

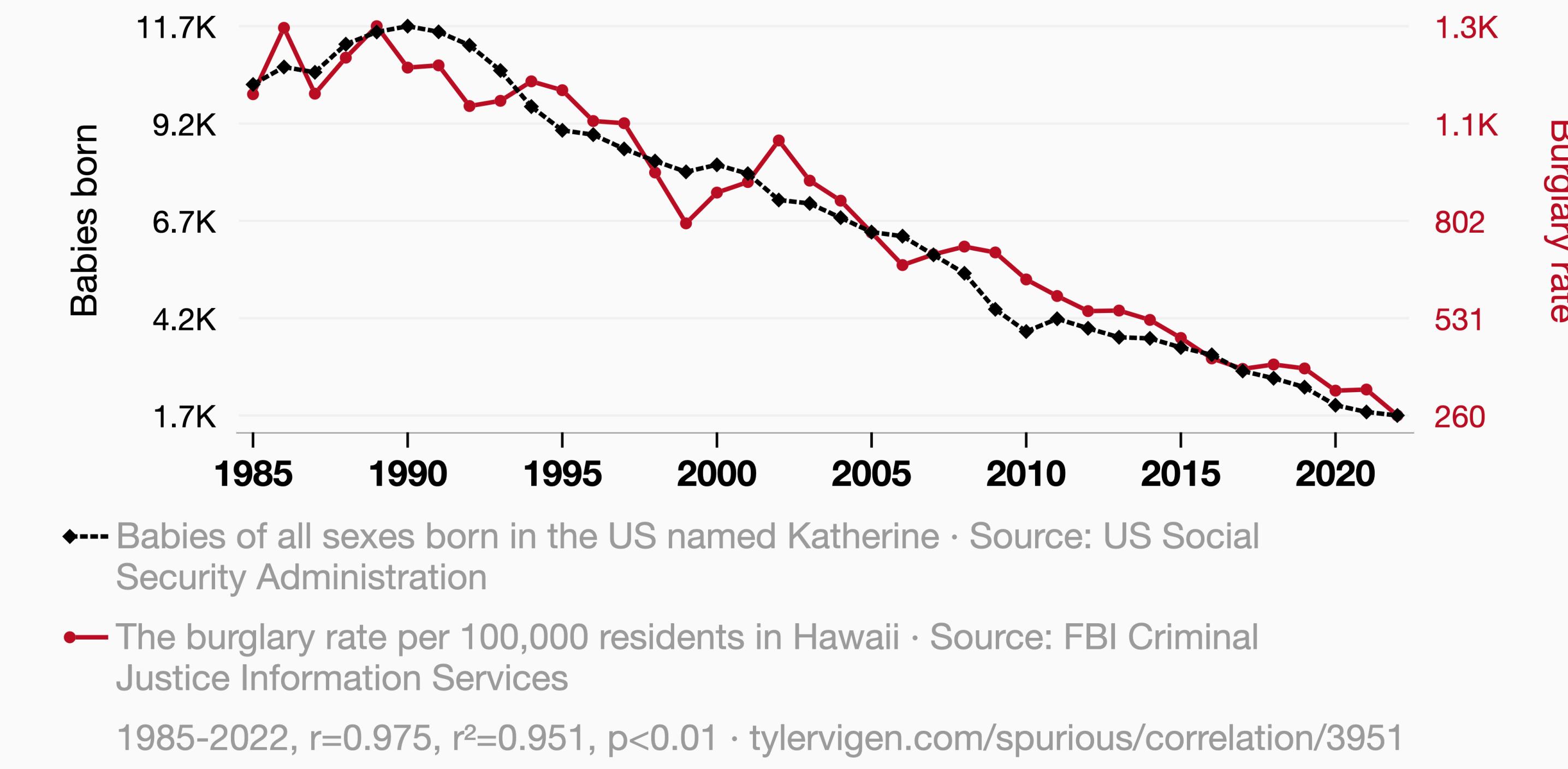


# **Correlation**

# Popularity of the first name Katherine

correlates with

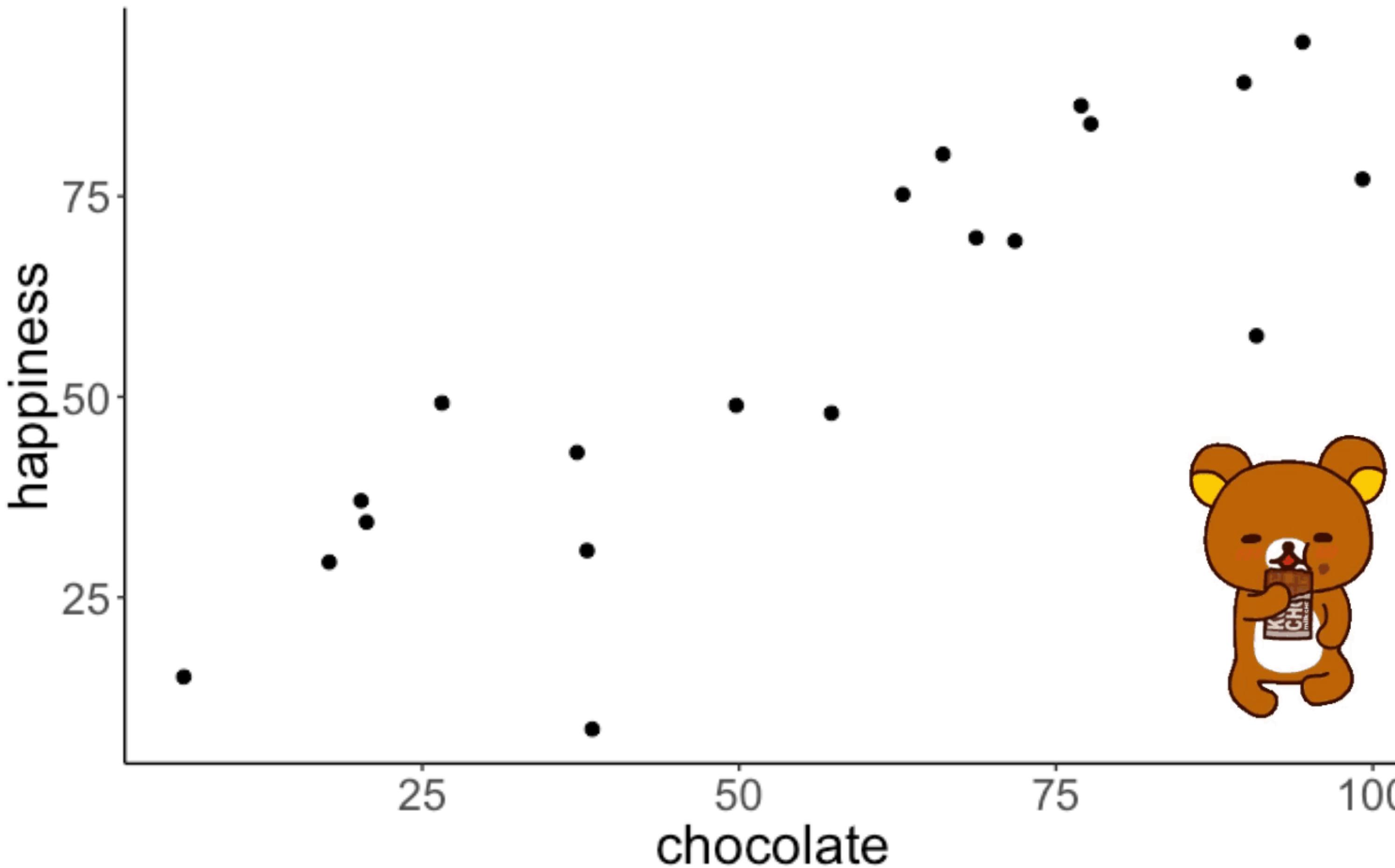
## Burglaries in Hawaii



AI generated

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 perpetrators 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 purvented!

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

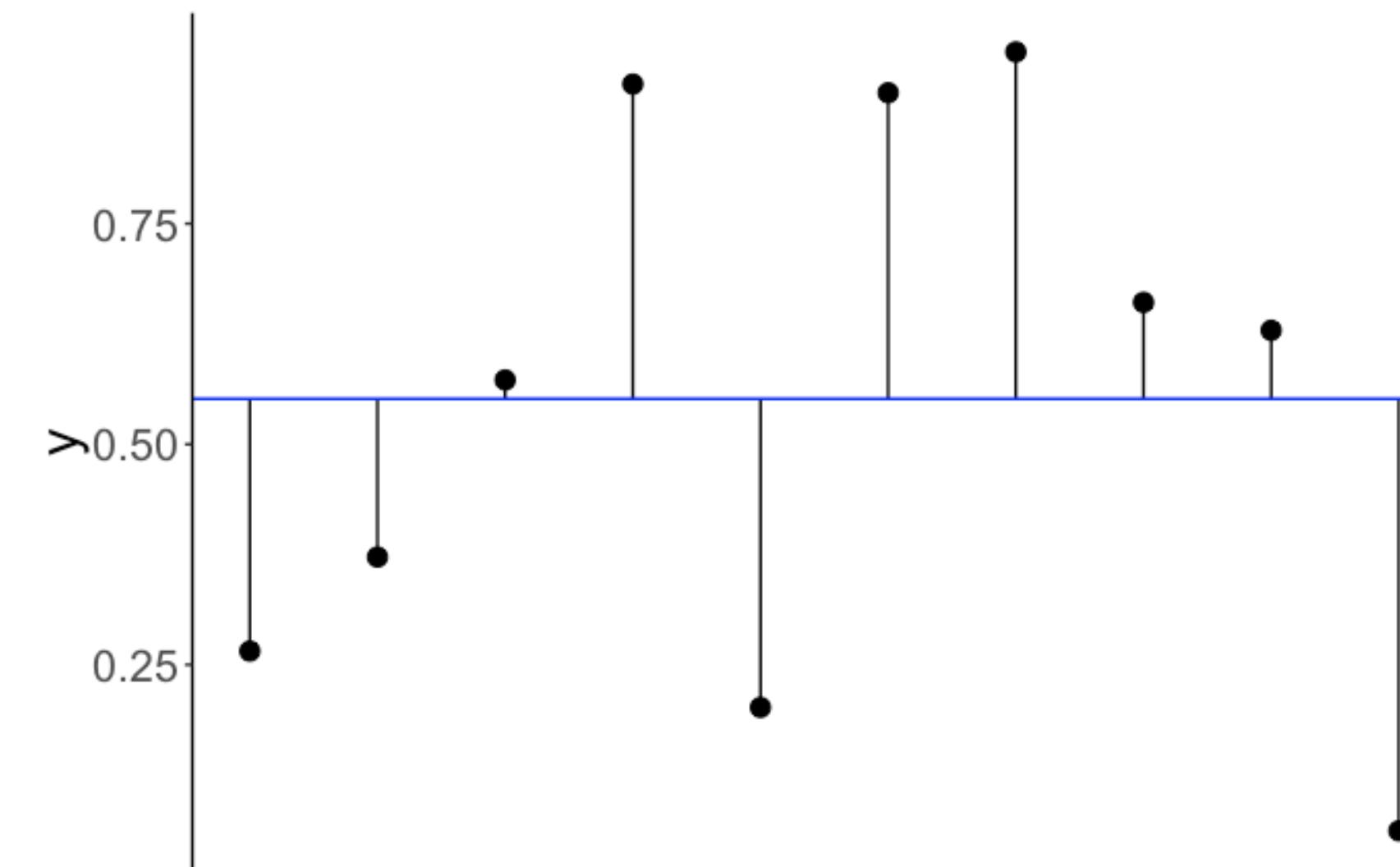
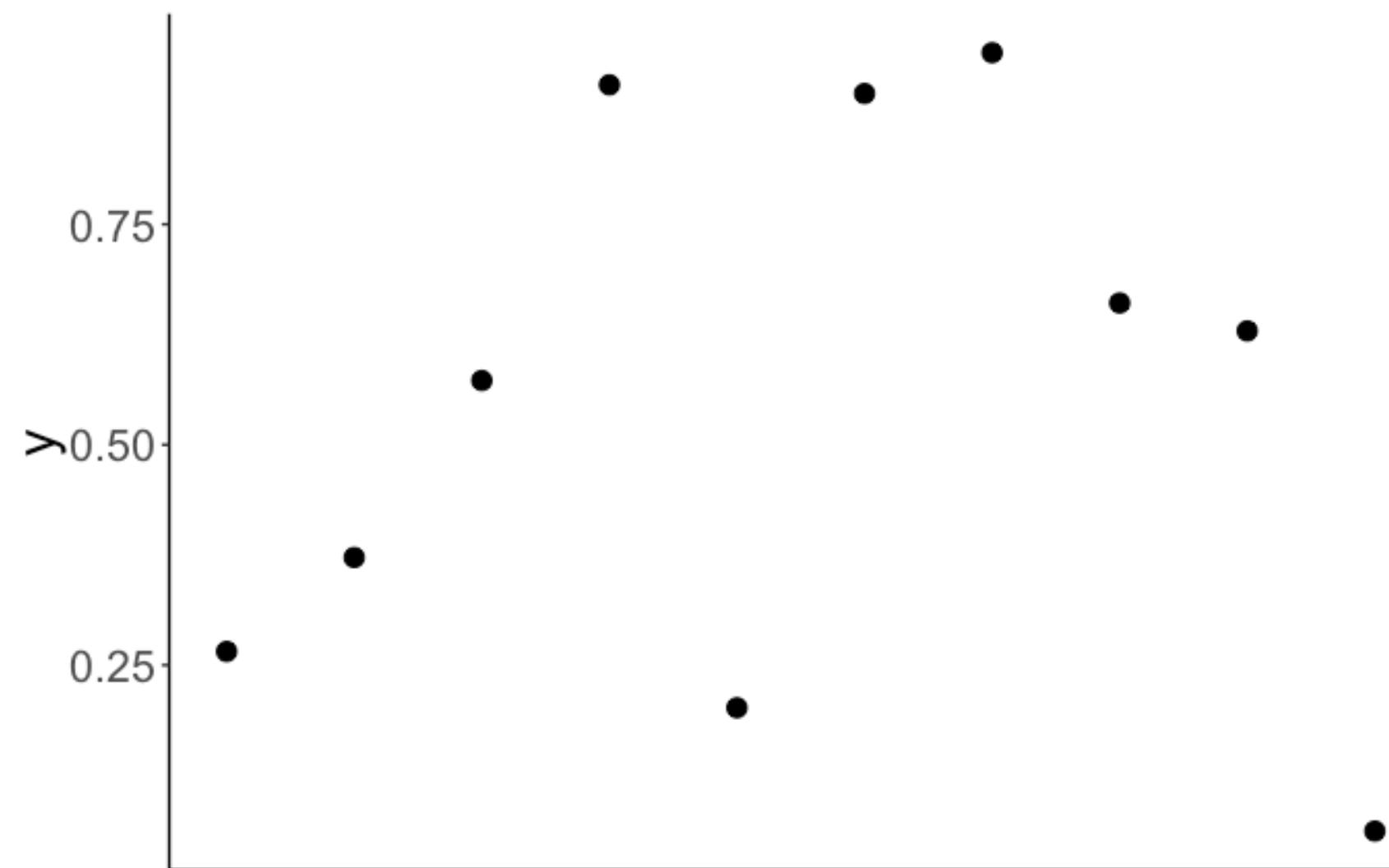


correlation = a measure of the relationship  
between two variables

## sample variance

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

sum of squared errors

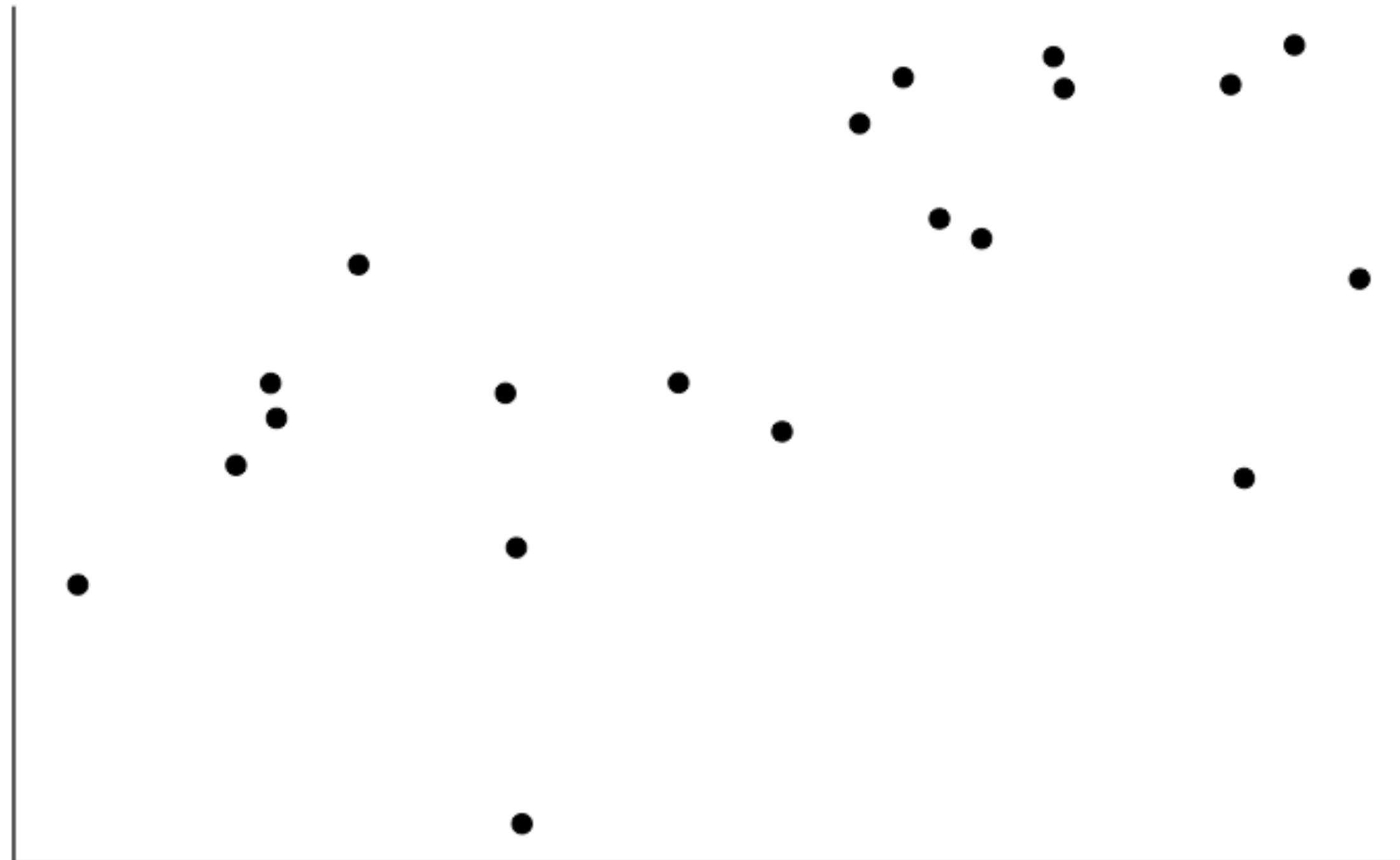


(I was too lazy to draw rectangles ...)

How well does the mean capture the data?

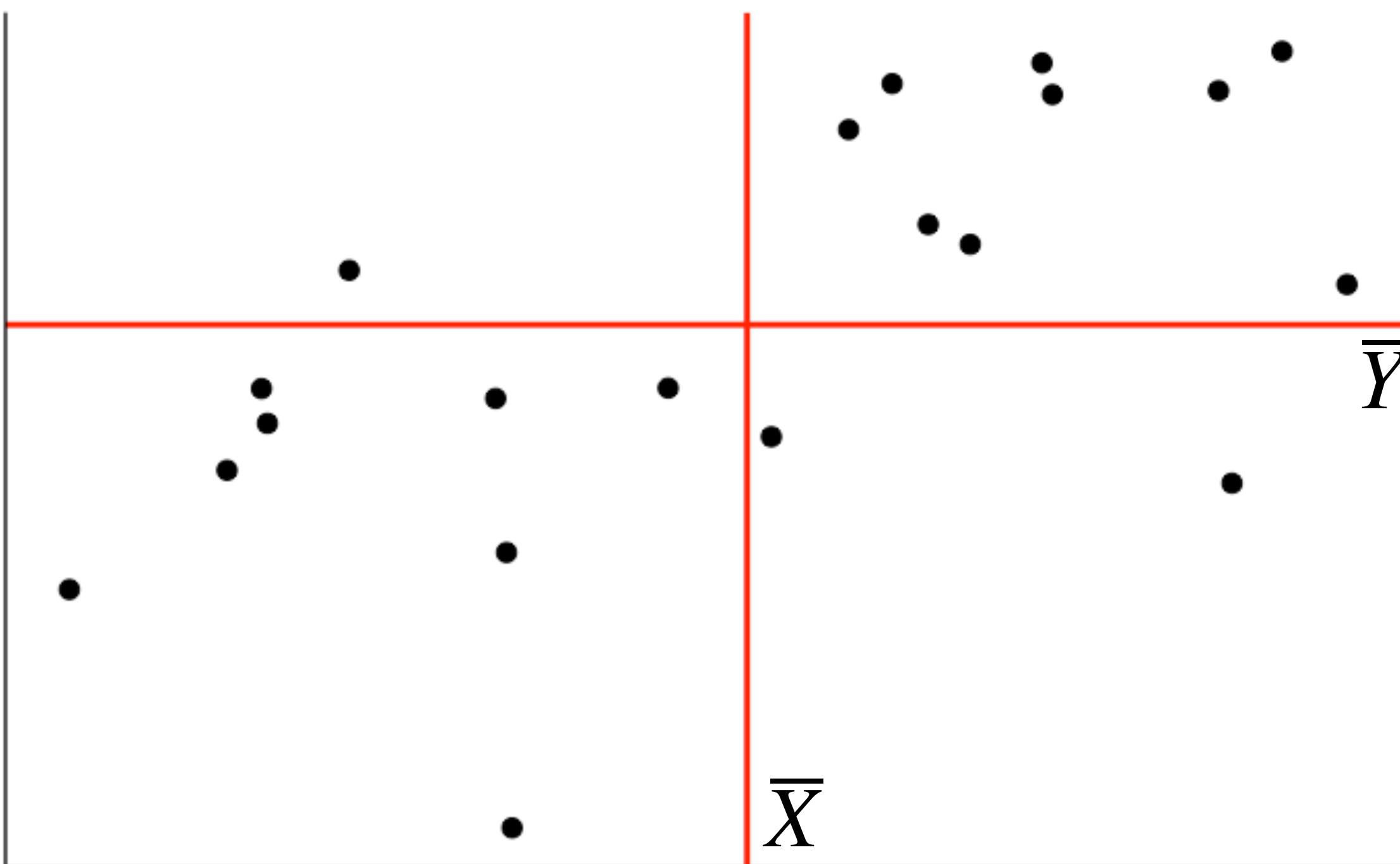
## sample covariance

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



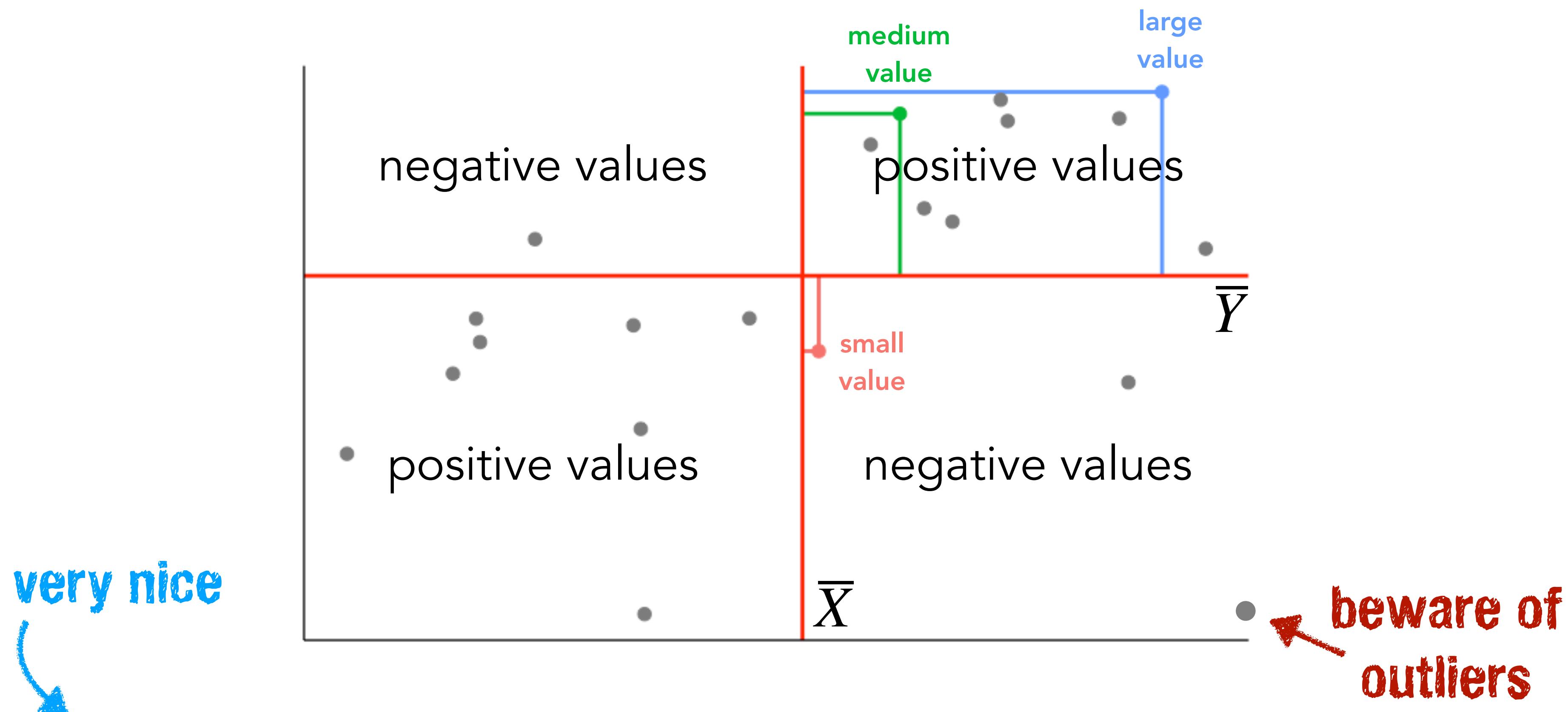
## sample covariance

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



## 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

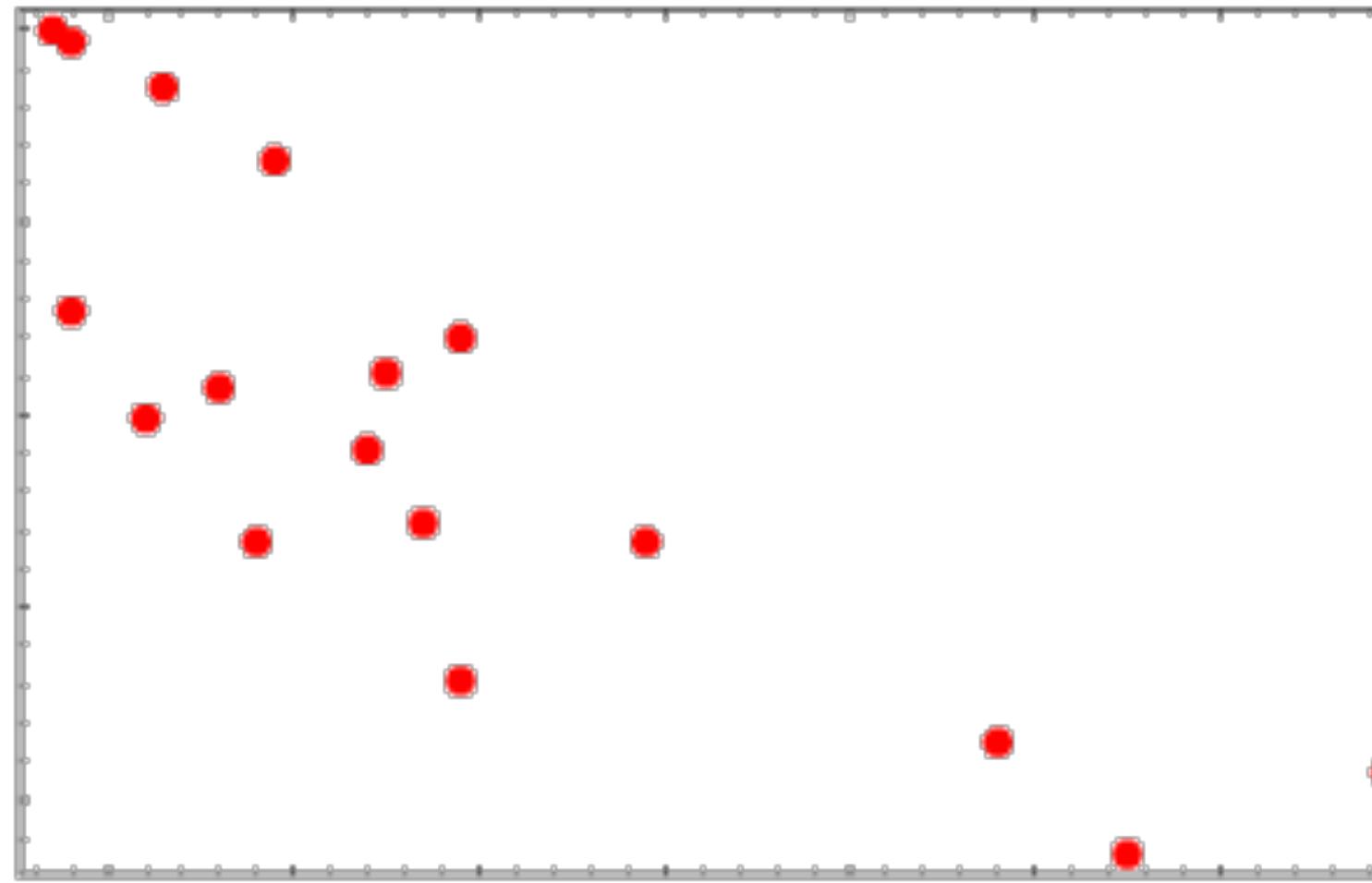
$$r(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

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:  $r(X, Y) = r(Y, X)$    
**association not causation**
- sign determines the direction of dependence
- captures **linear dependence** only

# 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

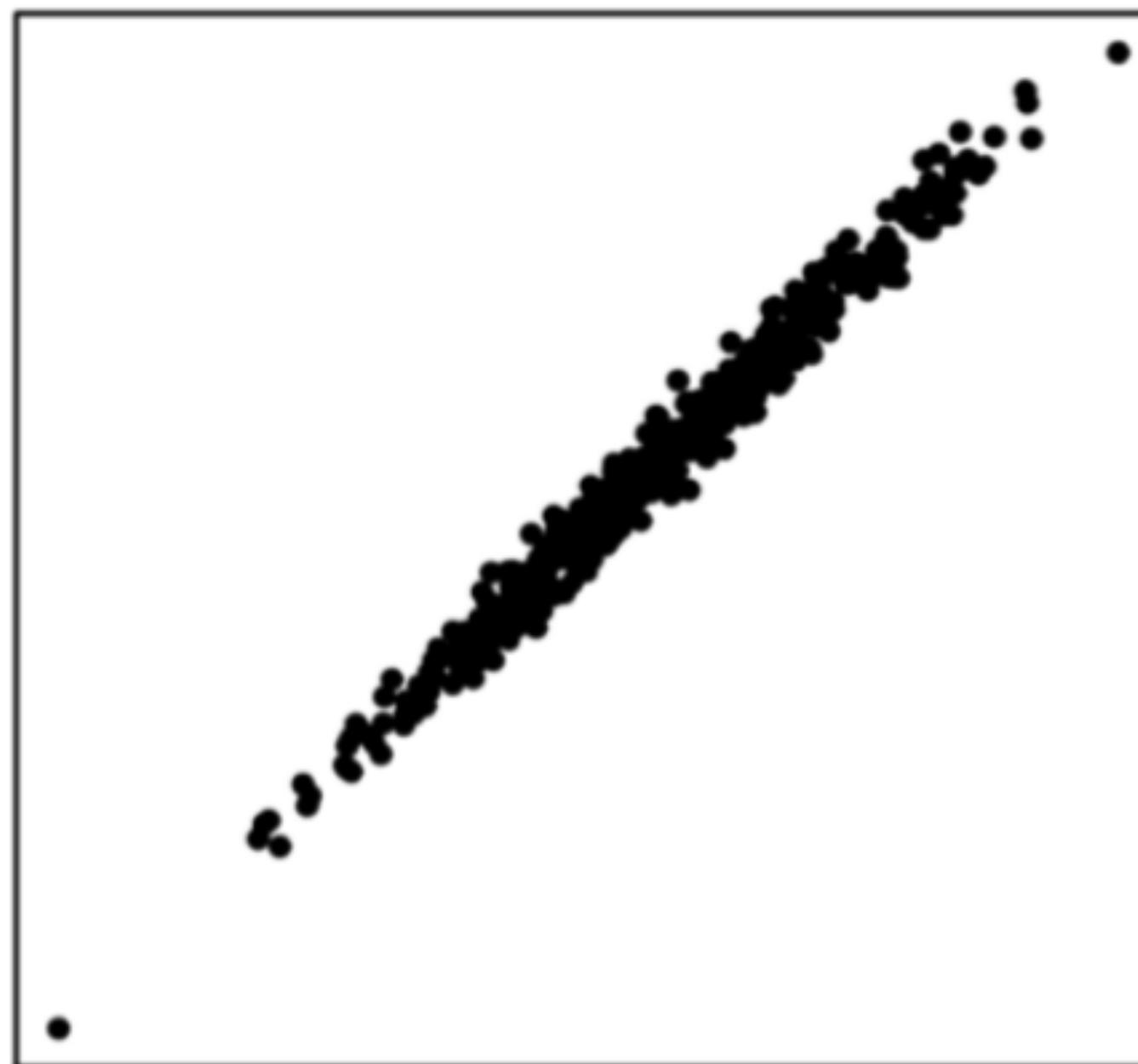
0.5 : 0.75

0.75 : 1

# Who is the correlation champion?

Win up to 1,000 points per answer

# 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

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

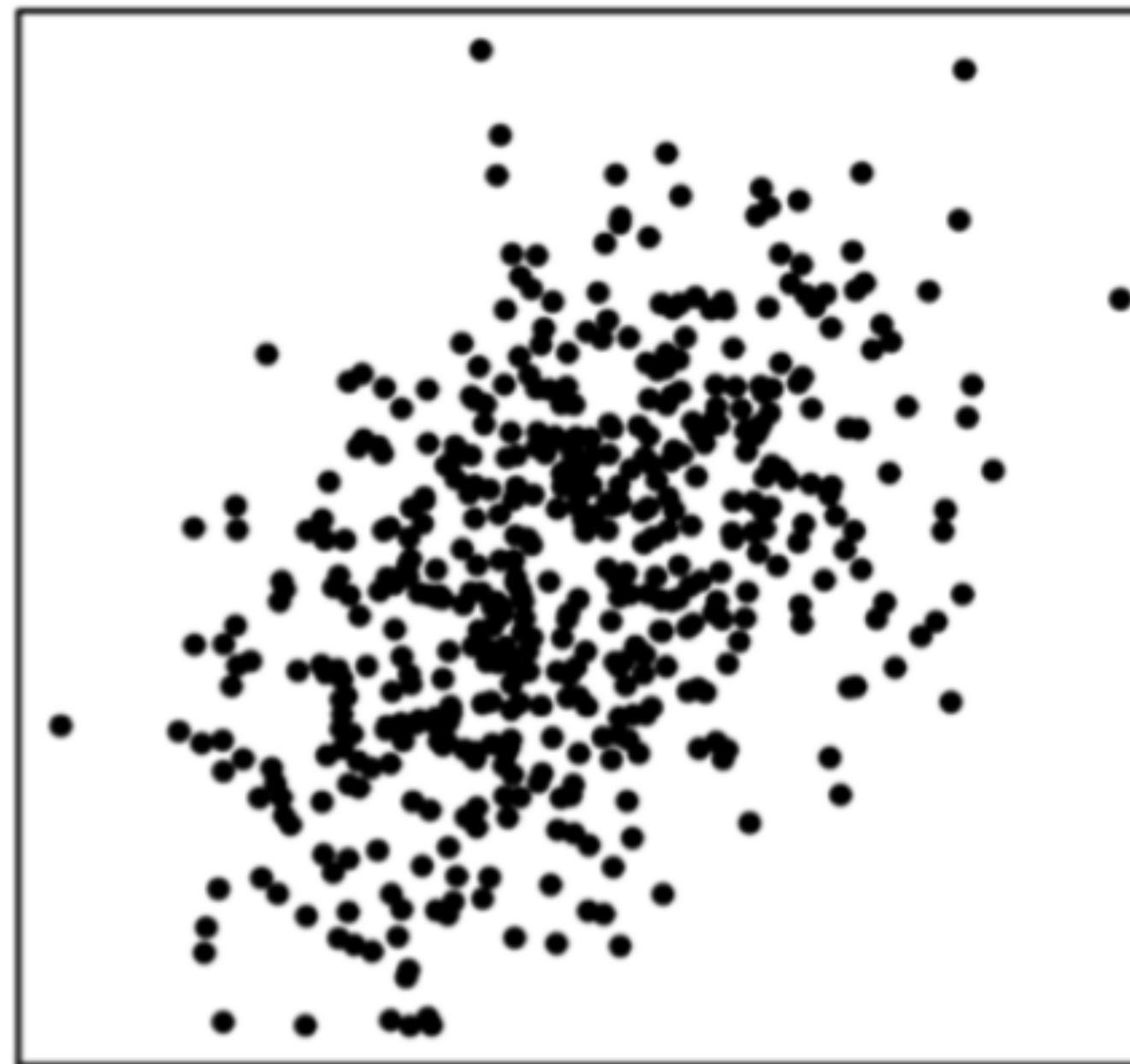
-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1



-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

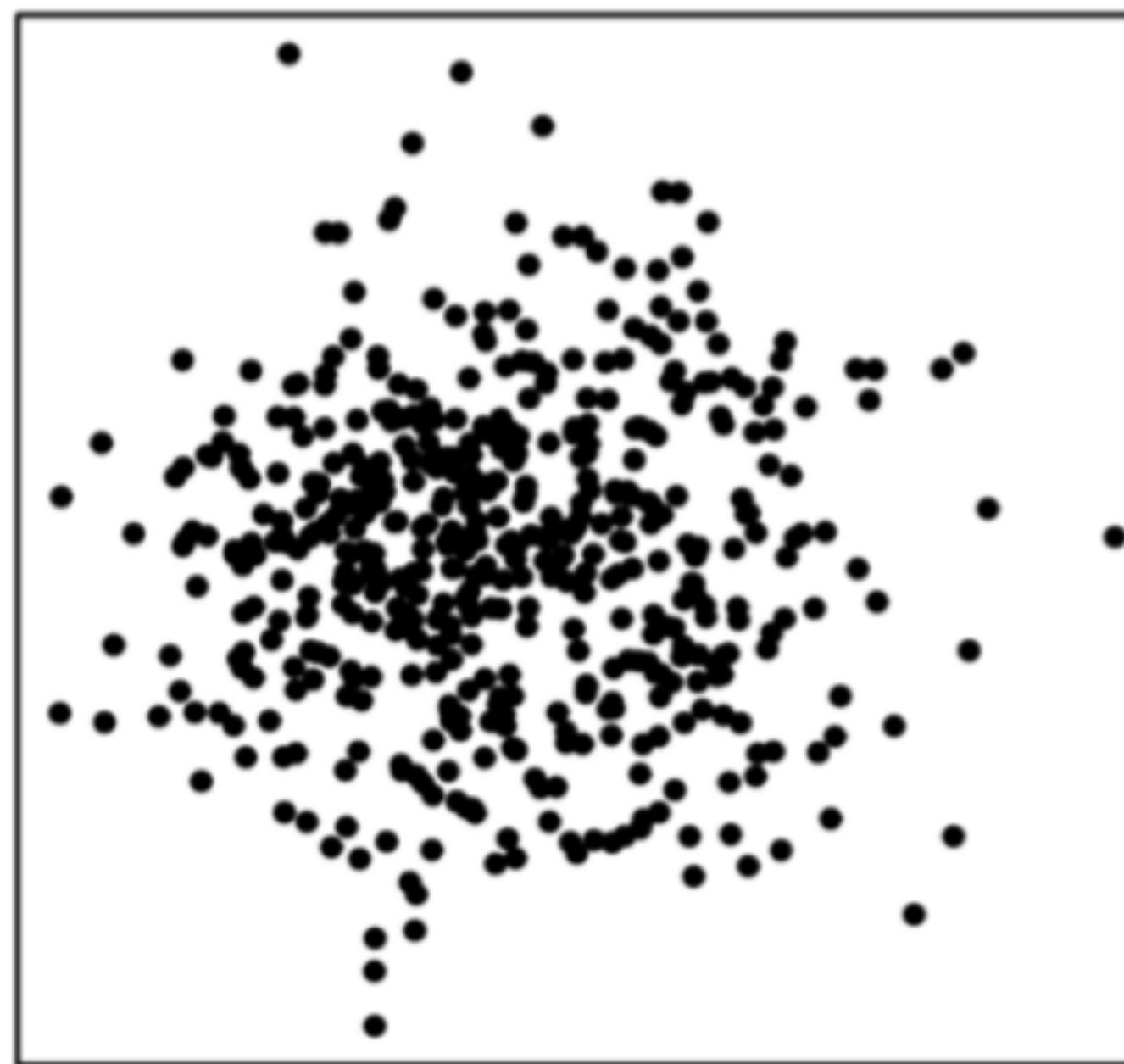
0.25 : 0.5

0.5 : 0.75

0.75 : 1

# Leaderboard

Nobody has responded yet.



-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



-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



-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

Nobody has responded yet.

# Solution

XX

# Be careful about interpreting correlations!

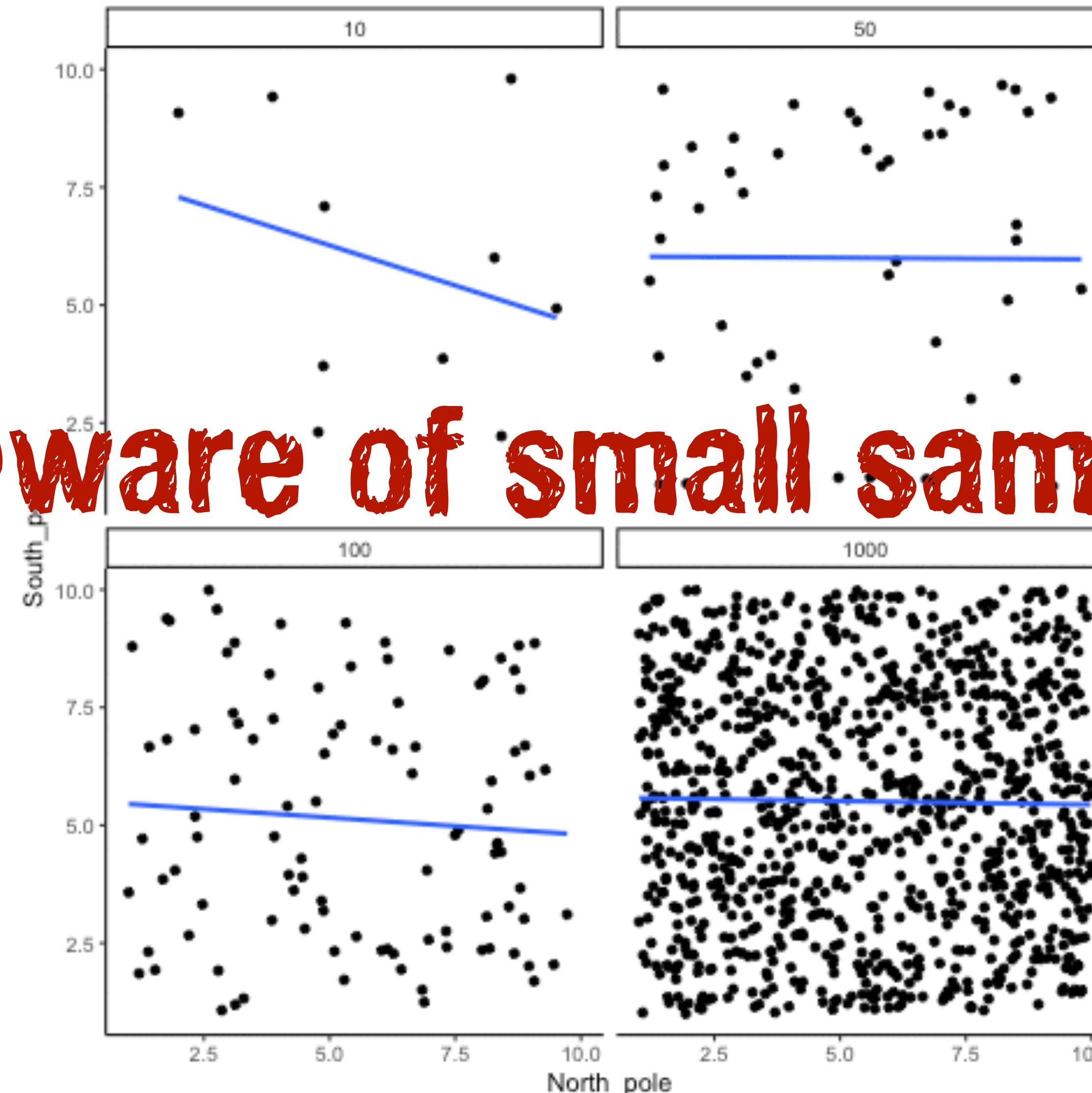


always visualize the data ...

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

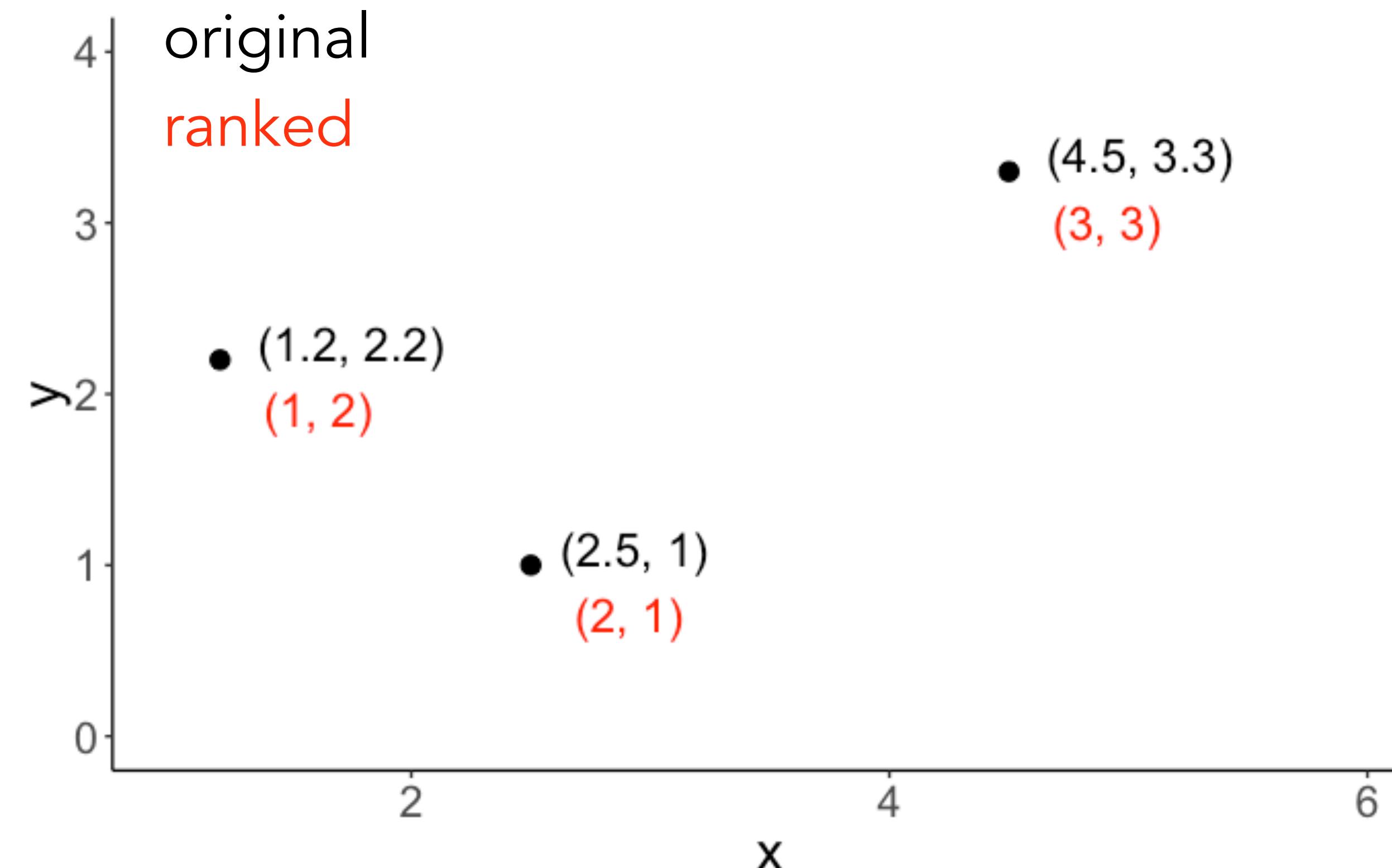
$$X \sim \mathcal{U}(\min = 0, \max = 10)$$

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



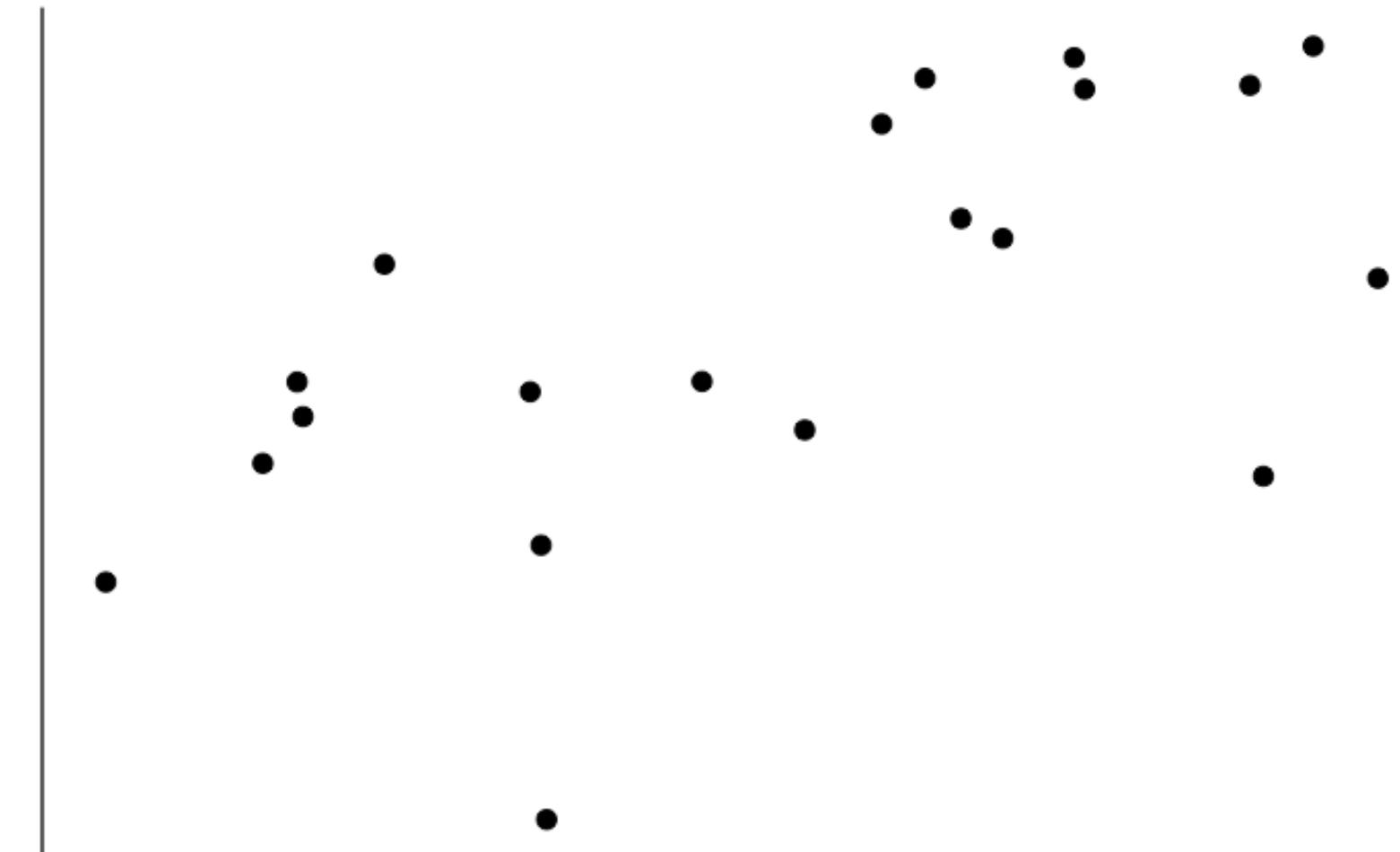
# Spearman rank order correlation

- 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



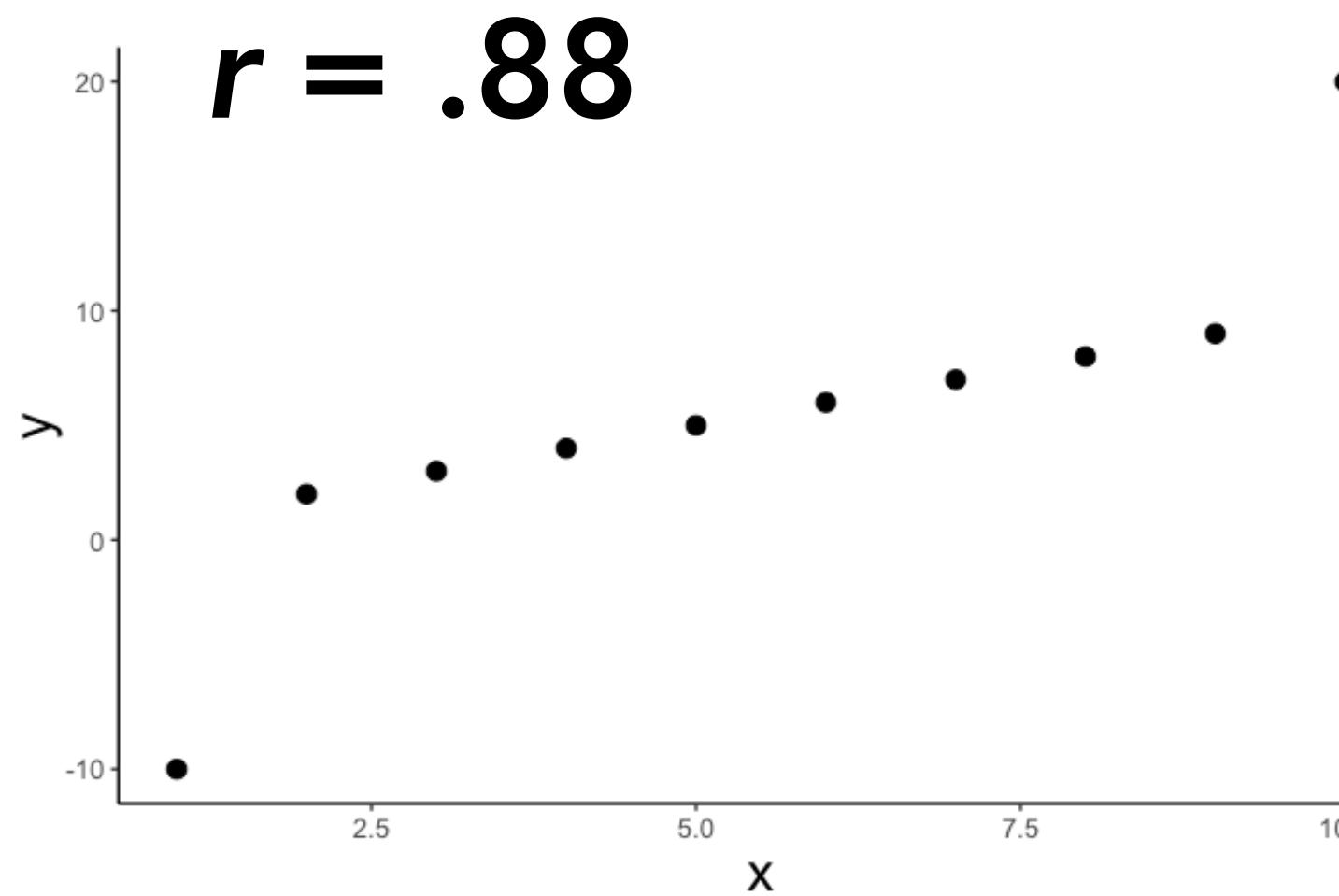
x	y	x_rank	y_rank
0.27	1.14	5	12
0.37	0.97	6	8
0.57	0.92	10	6
0.91	0.85	18	4
0.20	0.98	3	9
0.90	1.39	17	17
0.94	1.44	19	20
0.66	1.40	12	18
0.63	1.33	11	15
0.06	0.71	1	2

r	spearman	r_ranks
0.609	0.595	0.595

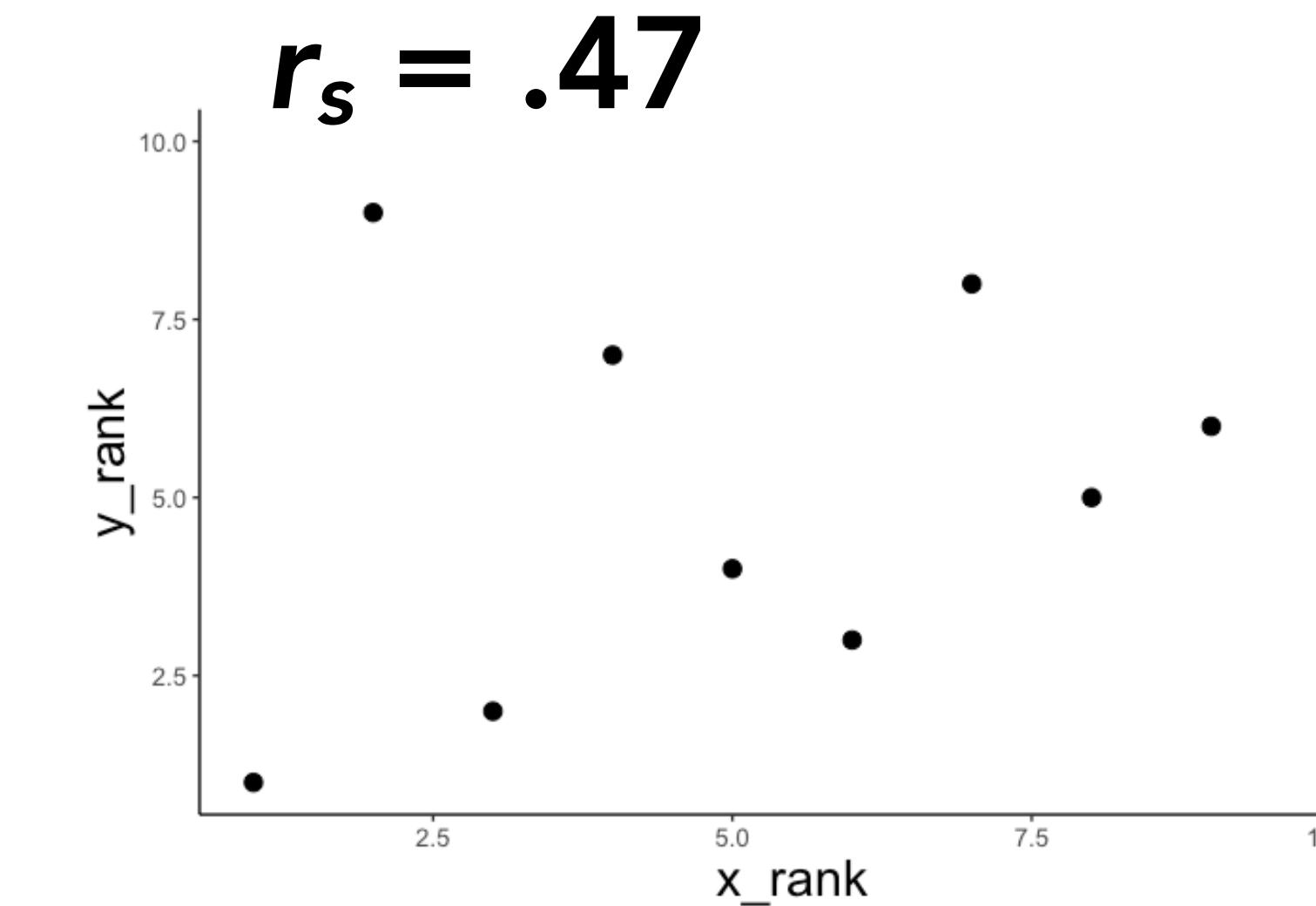
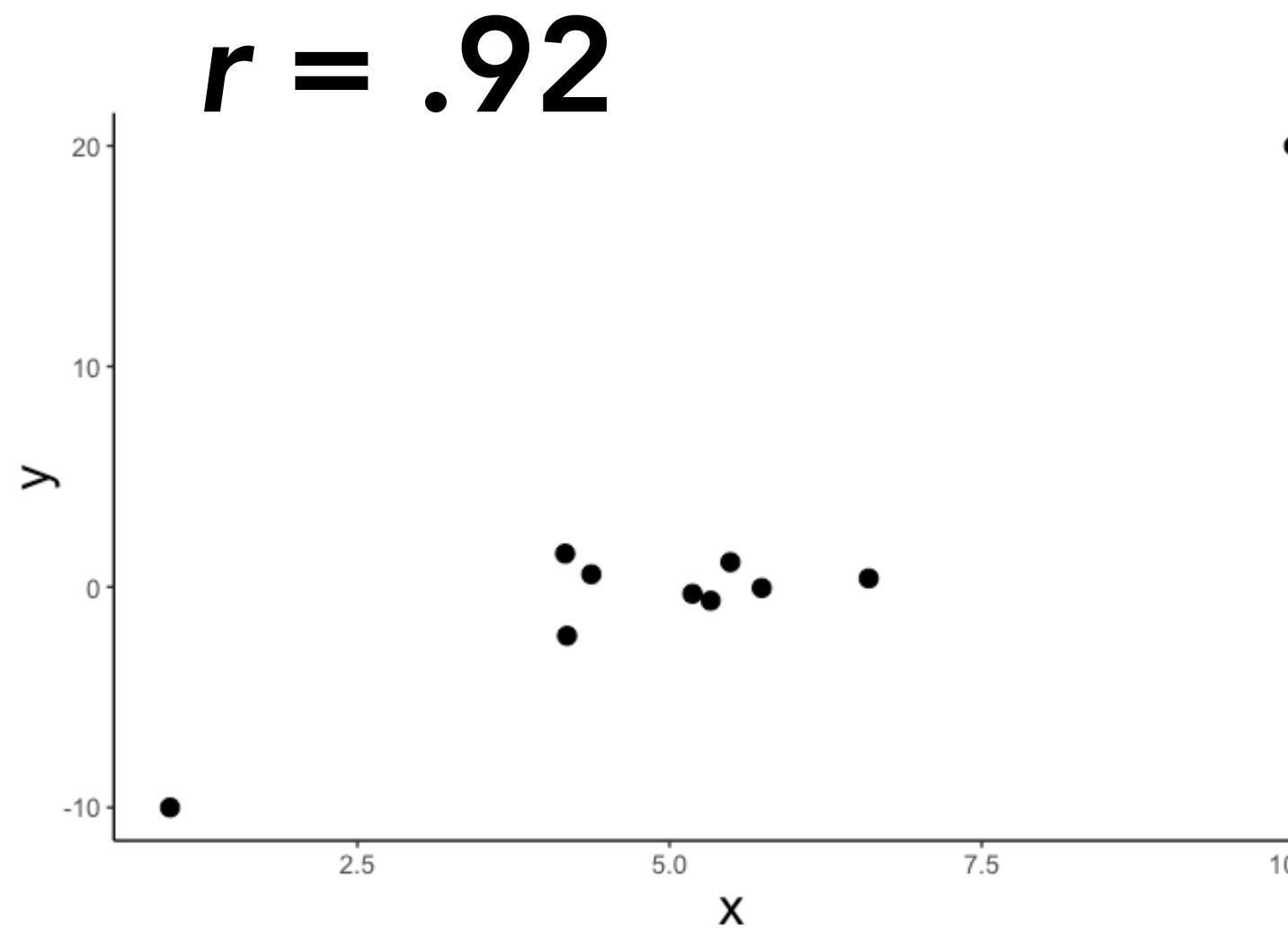
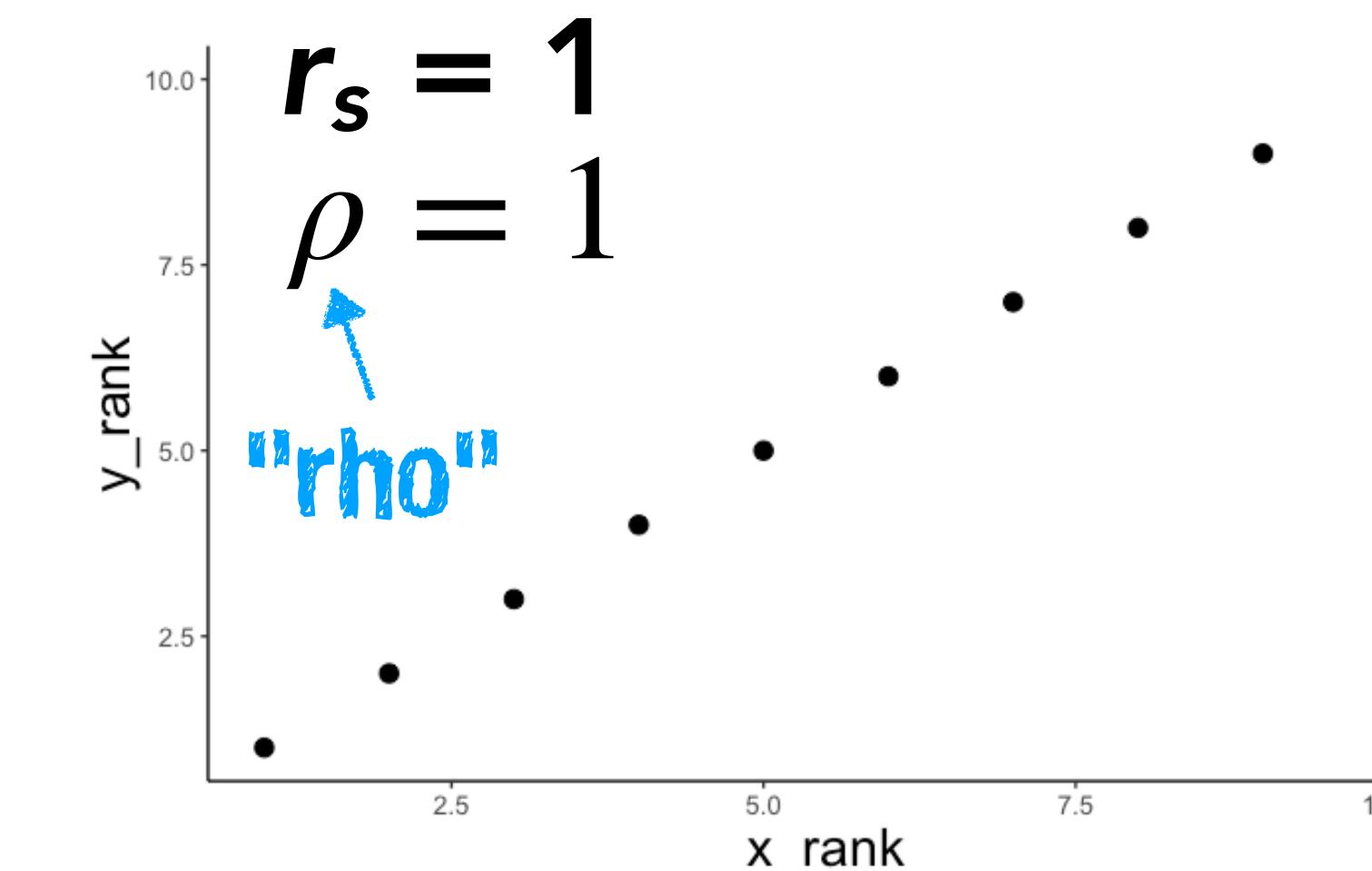
```
1 # correlation
2 df.spearman %>%
3   summarize(r = cor(x, y, method = "pearson"),
4             spearman = cor(x, y, method = "spearman"),
5             r_ranks = cor(x_rank, y_rank, method = "pearson"))
```

# Spearman rank order correlation

original



ranked



# Pearson vs. Spearman

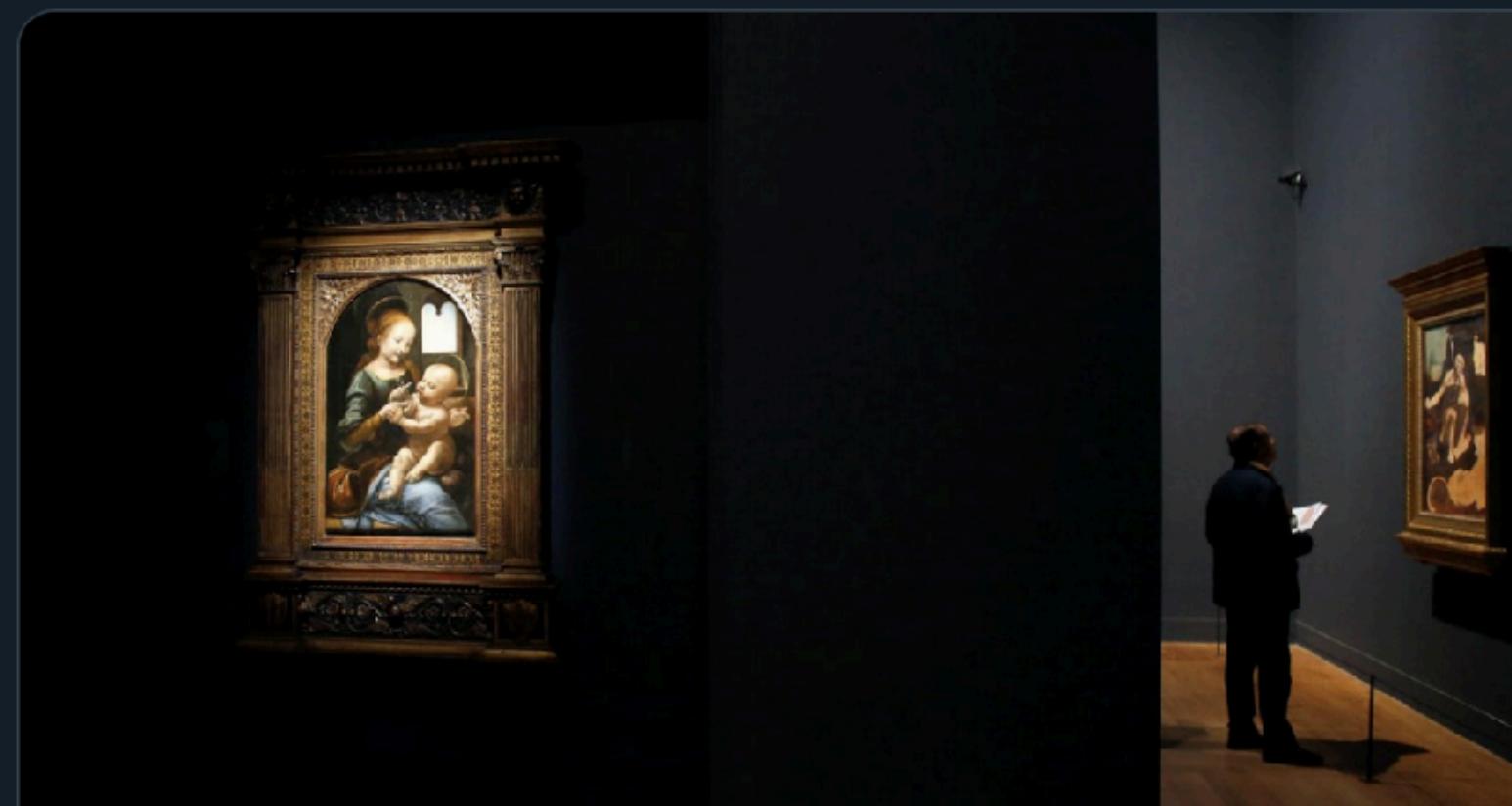
- Pearson's  $r$  captures the extent to which the relationship between two variable is **linear**
- Spearman's  $\rho$  captures the extent to which the relationship between two variables is **monotonic**
- What's better?
  - depends on the context
  - Spearman is robust to outliers, but it throws away (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](#)

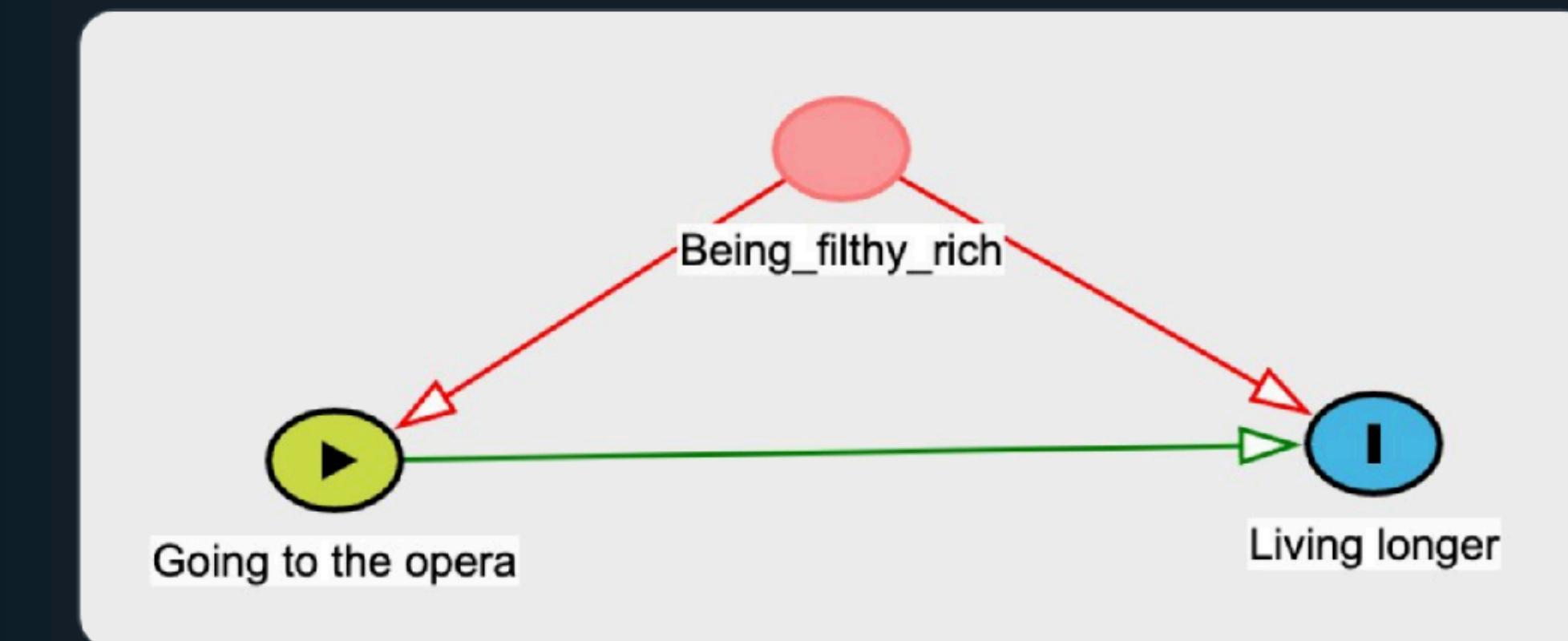
9:19 AM · Dec 22, 2019 · SocialFlow

336 Retweets 1.3K Likes



Andrew Heiss  
@andrewheiss

ooh ooh i can draw the dag for this one!



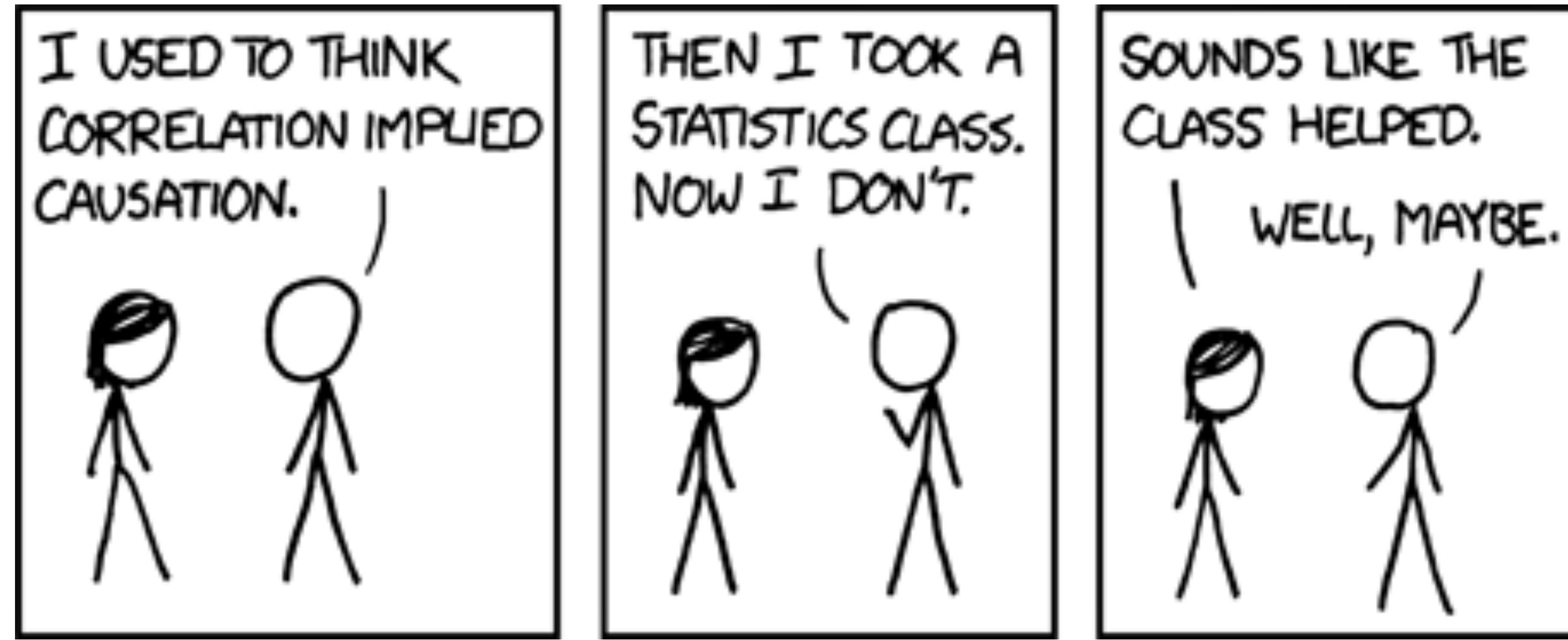
NYT Health @NYTHealth · Dec 22, 2019

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



- 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

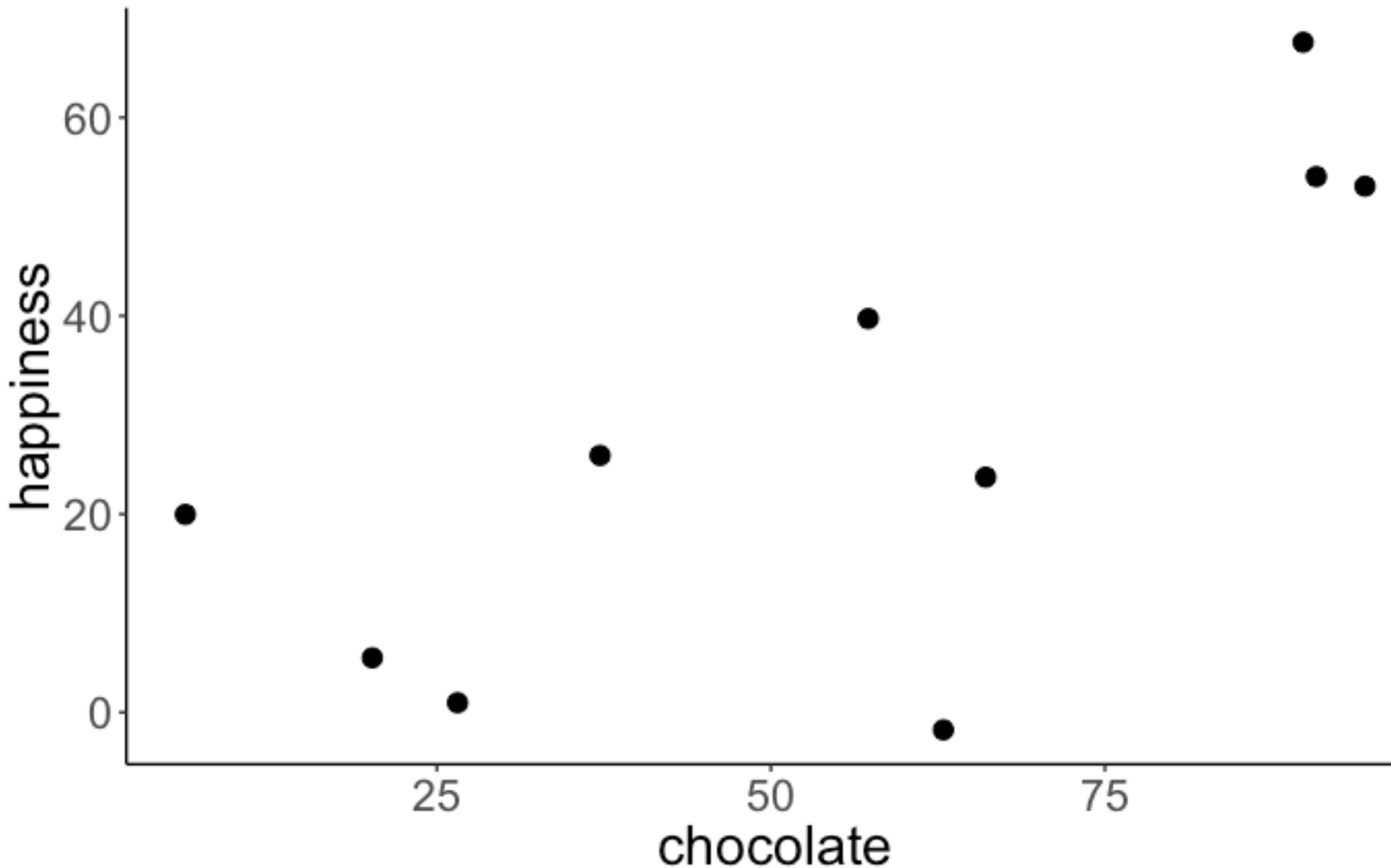
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

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

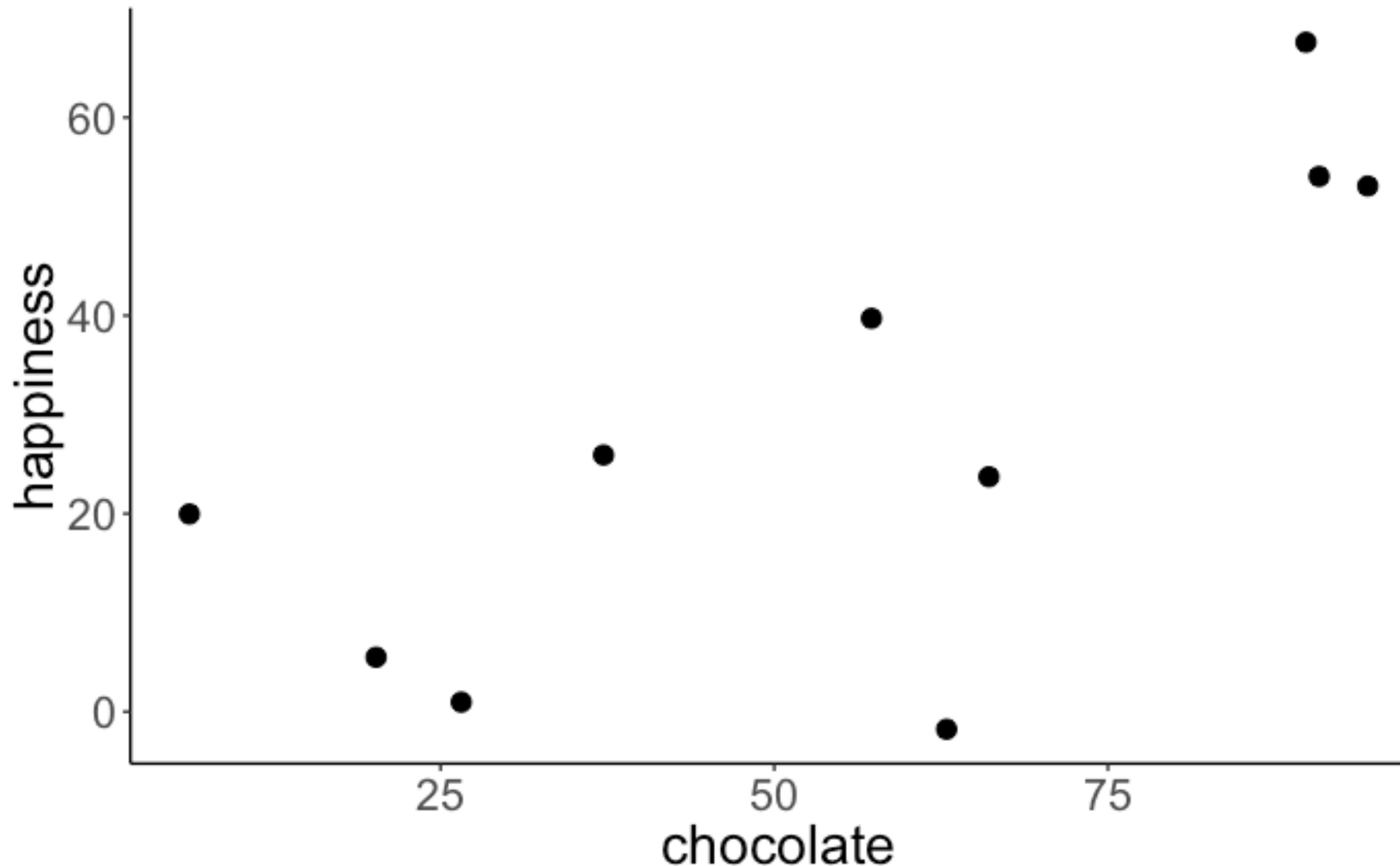


the model is a linear  
combination of predictors

# 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

$$Y_i = \beta_0 + \epsilon_i$$

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

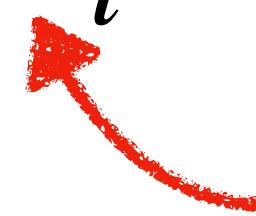
and

$$\beta_1 = 0$$

$H_1$ : Chocolate consumption and happiness are related.

### Model A

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



chocolate  
consumption

# The general procedure

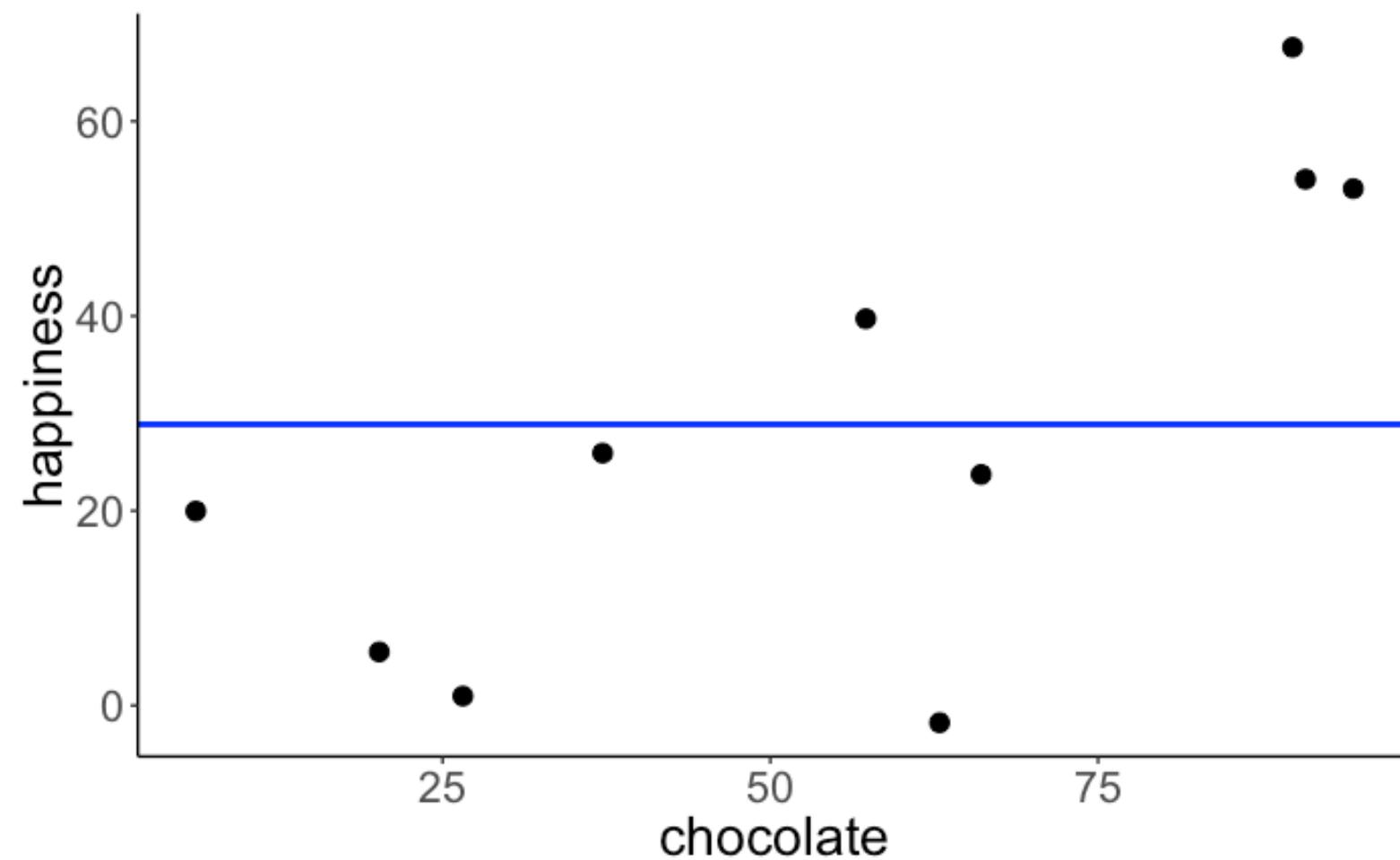
1. Define  $H_0$  as Model C (compact) and  $H_1$  as Model A (augmented)
- 2. Fit model parameters to the data**
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution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

$$Y_i = 28.88 + e_i$$

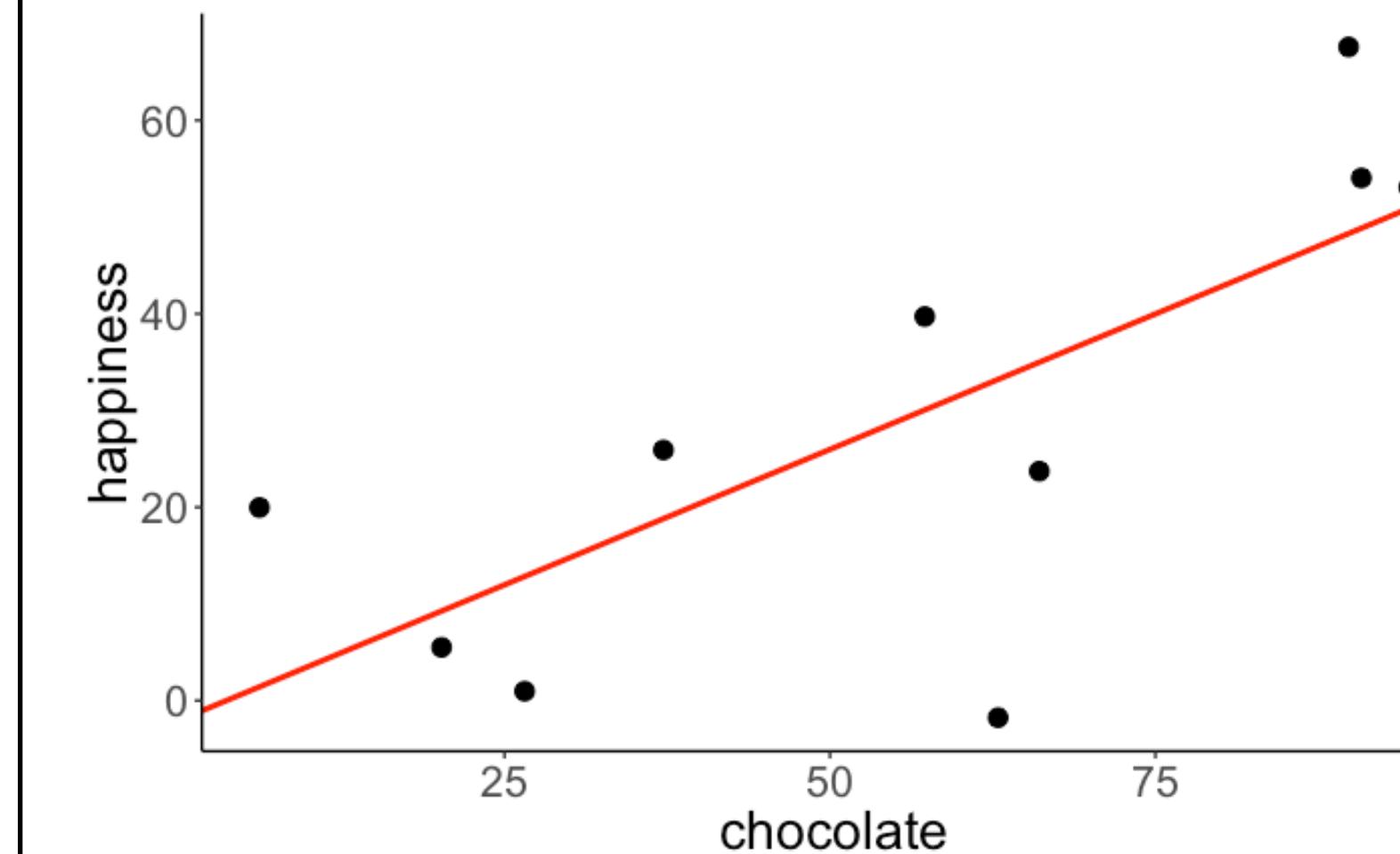
$H_1$ : Chocolate consumption and happiness are related.

### 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

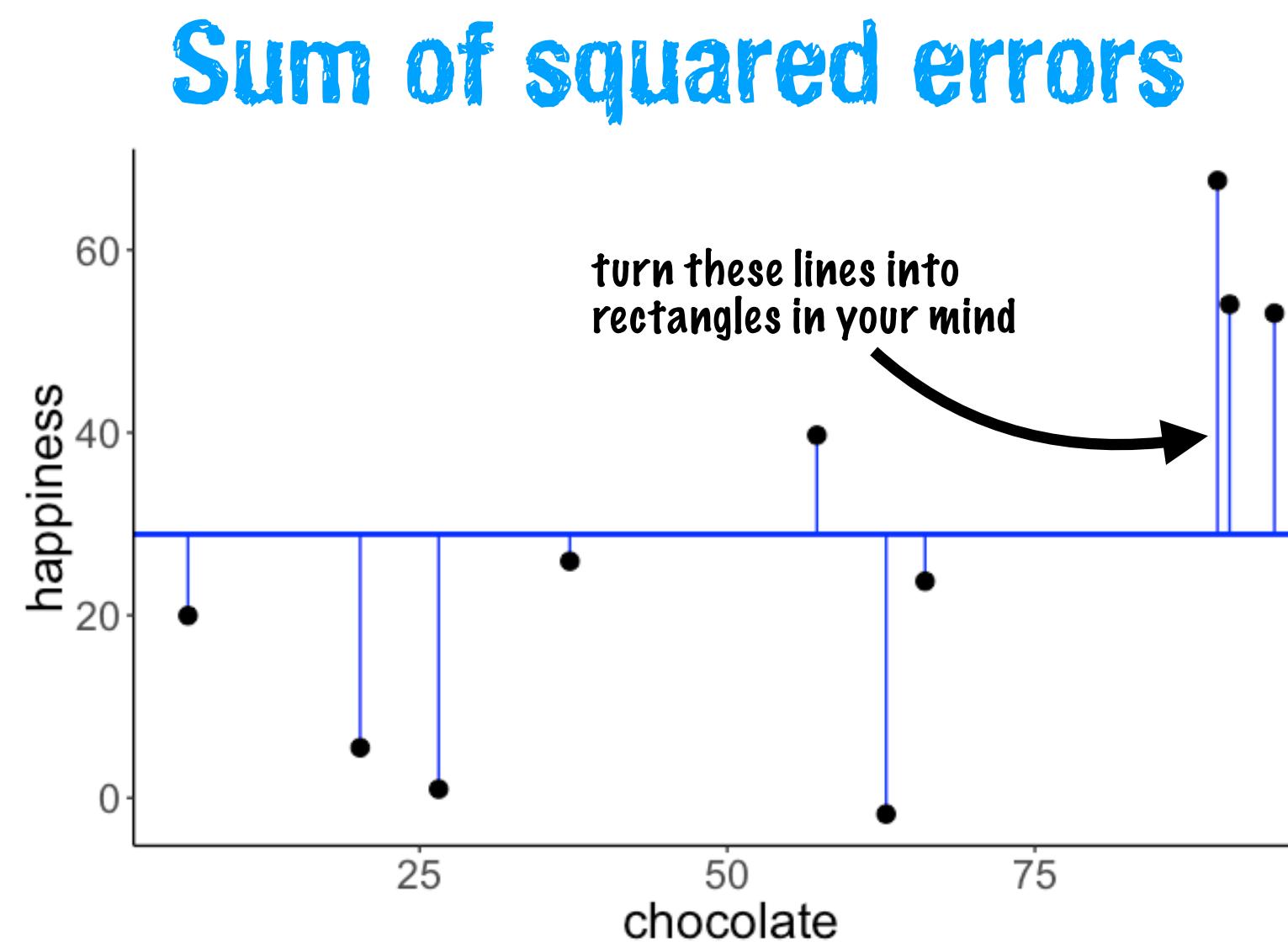
1. Define  $H_0$  as Model C (compact) and  $H_1$  as Model A (augmented)
2. Fit model parameters to the data
3. **Calculate the proportional reduction of error (PRE) in our sample**
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution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

# Calculate PRE

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

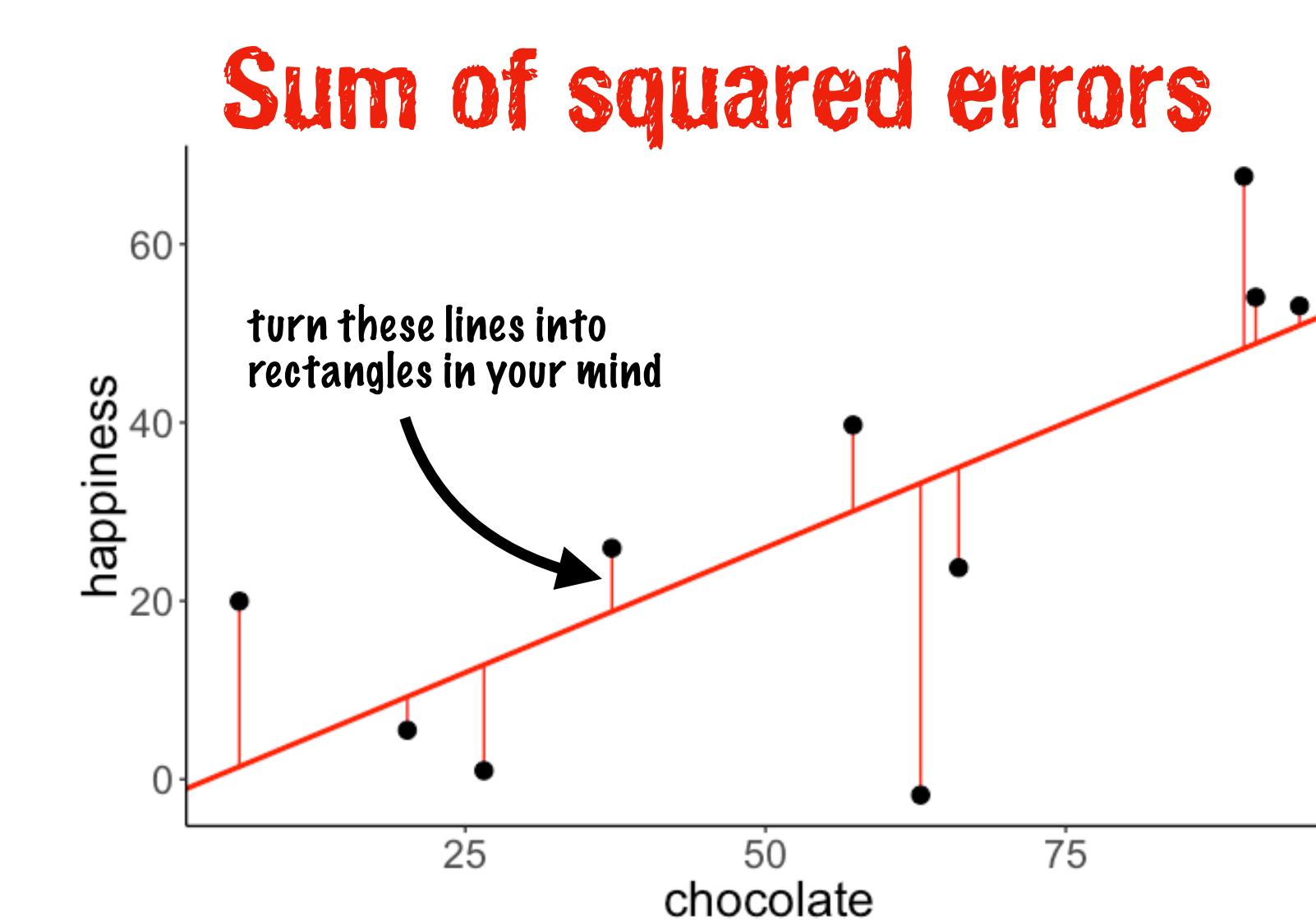
Both models were fit to minimize the sum of squared errors

OLS = Ordinary **least squares** regression



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

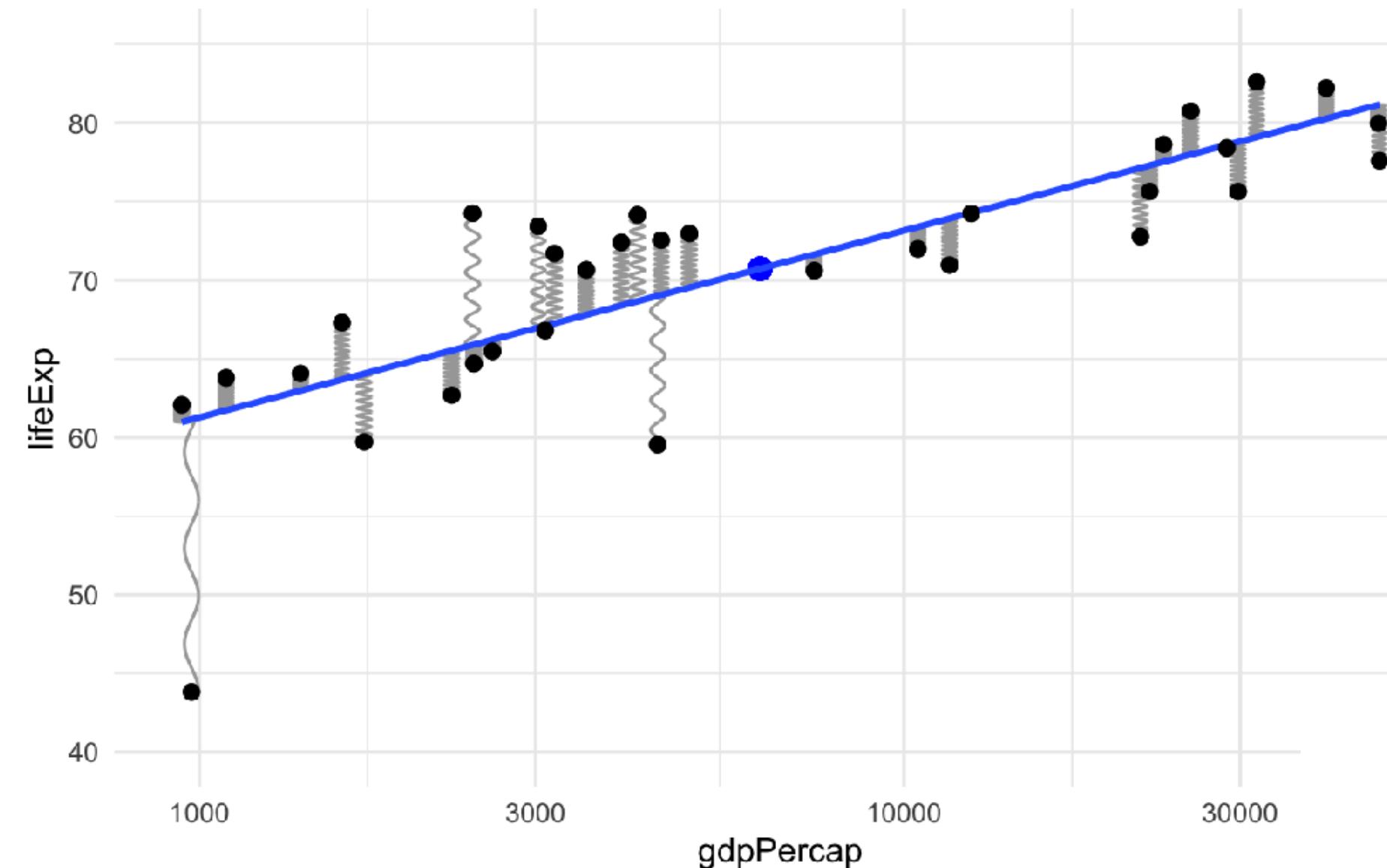
$$\text{PRE} = 1 - \frac{2396.946}{5215.016} \approx 0.54$$



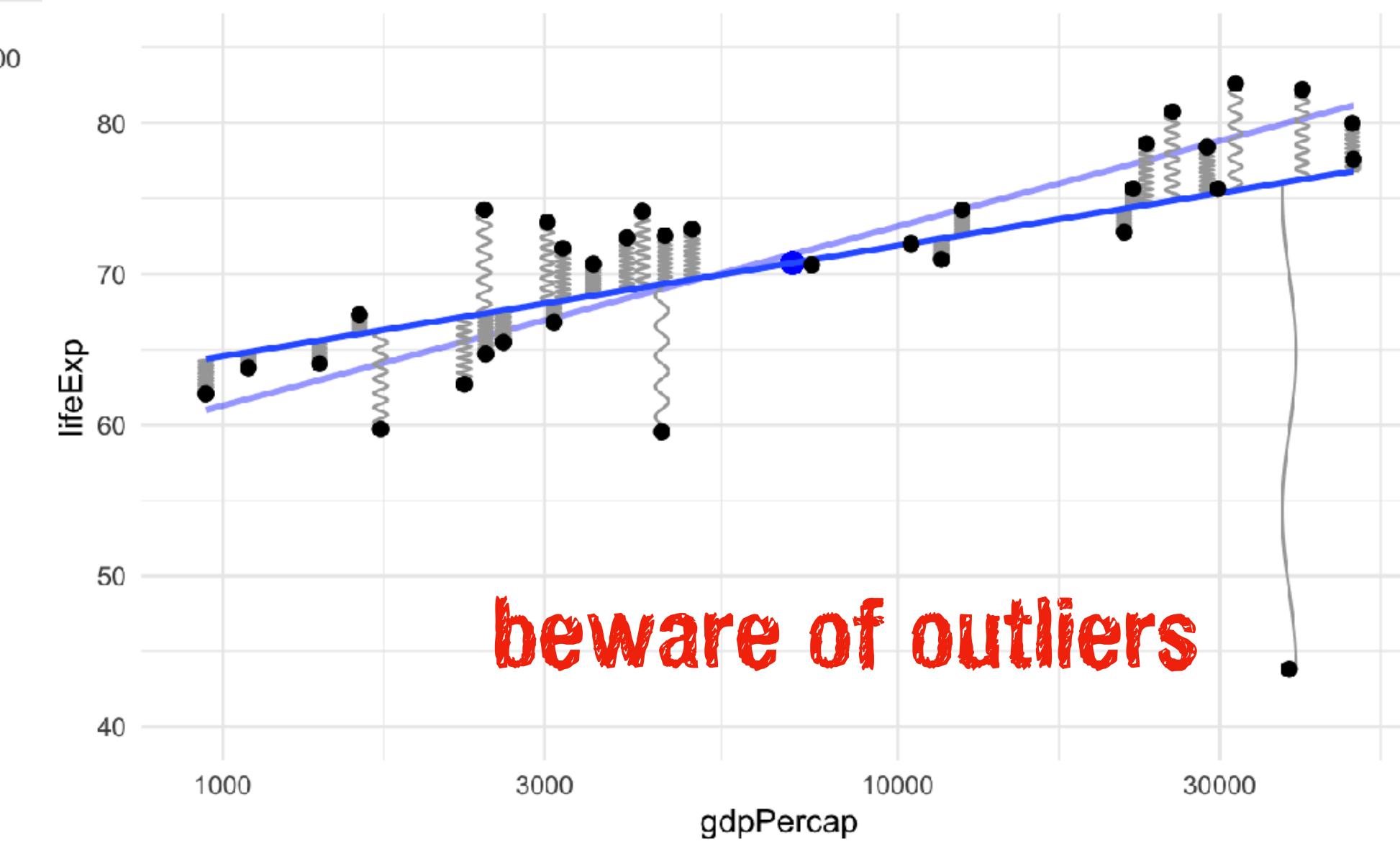
$$\text{SSE}(A) = 2396.946$$

**The augmented model  
reduces the error by 54%.**

# Least squares as springs



each point is  
attached to the  
line with an  
identical spring



beware of outliers

# 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<sub>C</sub> to create model<sub>A</sub> but were not

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

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

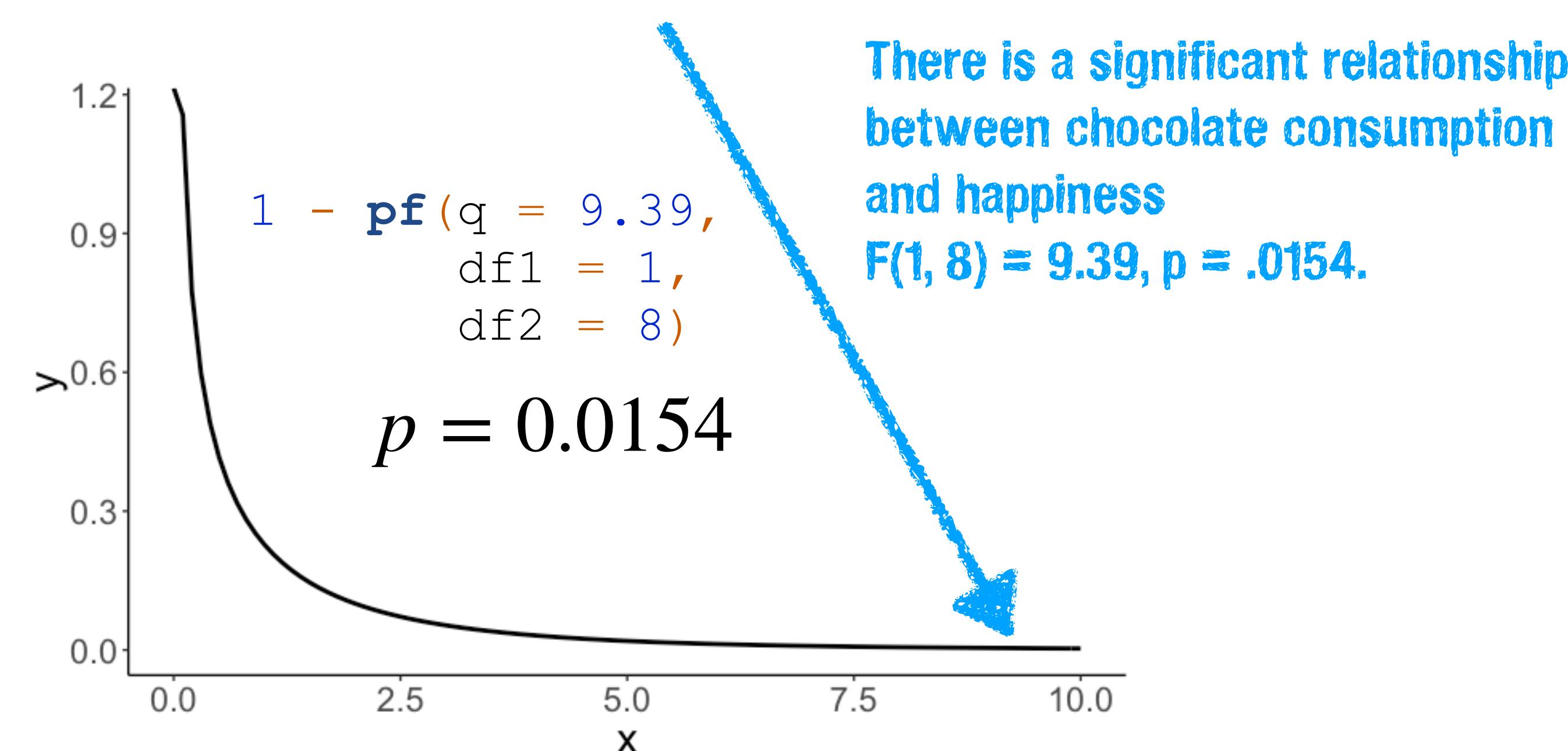
**number of observations vs.  
parameters in Model A**

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

- To compute the  $F$  statistic, we need:

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

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

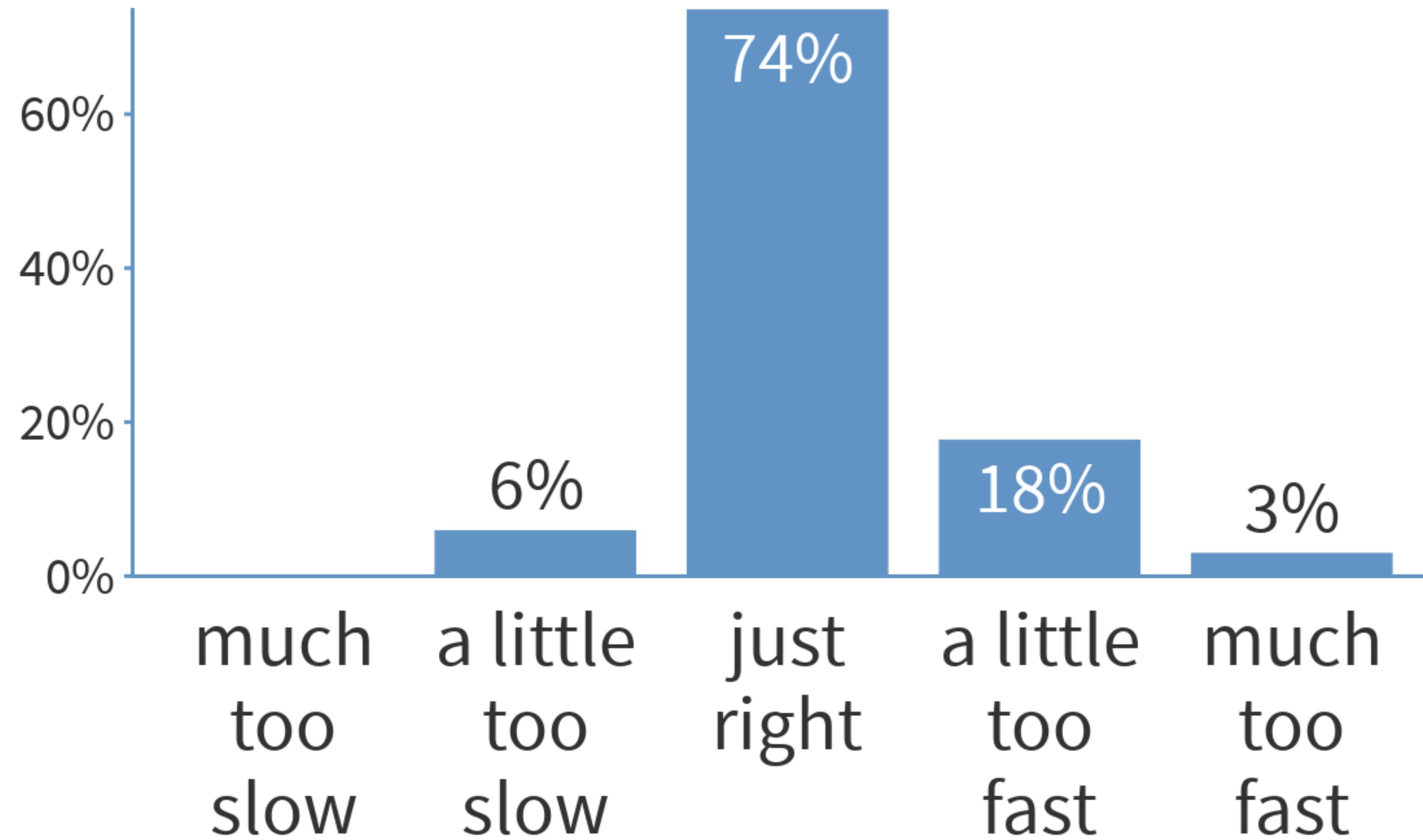


# Summary

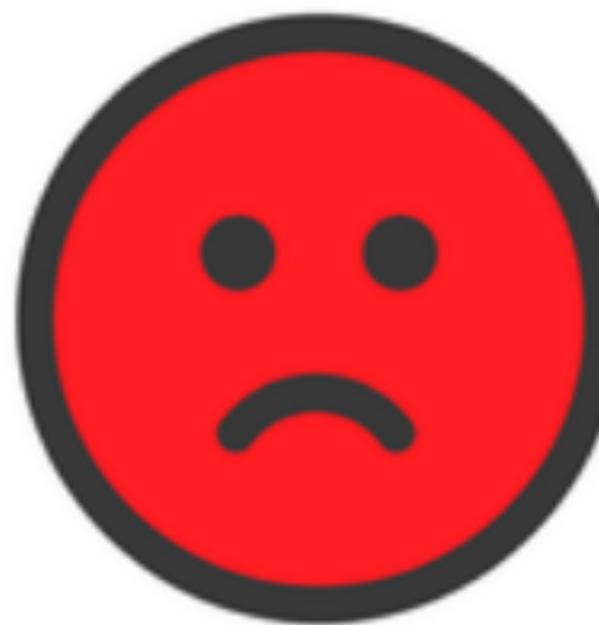
- Quick recap
- Modeling data
- Hypothesis testing as model comparison
- Correlation
  - Pearson's moment correlation
  - Spearman's rank correlation
- Regression

# **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?

A word cloud visualization showing student feedback on today's class. The words are arranged in a large, overlapping cluster. The most prominent words are 'understand', 'really', 'clear', 'well', 'tobi', 'things', 'class', 'concepts', 'background', 'foreground', 'important', 'limit', 'second', 'third', 'class.', 'record', 'explaination', 'lecture', 'interaction', 'time', 'fully', 'fix', 'us', 'assume', 'low', 'may', 'might', 'better', 'didn't', 'let', 'process', 'cover', and 'process'.

**Thank you!**