# Psyc 60: Intro to Statistics

Prof. Judith Fan

Spring 2022

# Due This Week

| | | | | |
|---|---|---|---|---|
| 6 | May 2 | **Sampling distributions** *Before:* Chapter 9 *During:* Lab 3C | **Review Session 2** *Before:* None *During:* Wrap-up Lab 3 | Quiz 3; Project Milestone 3 Due (Preregistration) |

Chapter 9 CourseKata modules are due today.

Note: If you finish modules a few days late, there may be a delay between finishing your CourseKata modules and the Gradebook in Canvas being updated (b/c there are multiple steps involved to correct these). But don't worry, these will be updated!

# Due This Week

| | | | | |
|---|---|---|---|---|
| 6 | May 2 | **Sampling distributions**<br>*Before:* Chapter 9<br>*During:* Lab 3C | **Review Session 2**<br>*Before:* None<br>*During:* Wrap-up Lab 3 | Quiz 3; Project Milestone 3 Due (Preregistration) |

Released Thursday at 5PM & due by 4:59PM on Friday

# Due This Week

| | | | | |
|---|---|---|---|---|
| 6 | May 2 | **Sampling distributions**<br>*Before:* Chapter 9<br>*During:* Lab 3C | **Review Session 2**<br>*Before:* None<br>*During:* Wrap-up Lab 3 | Quiz 3; Project Milestone 3 Due (Preregistration) |

Project Milestone 3 is about getting practice articulating the research question for your final project & thinking about different potential DGPs.

# Today

## Mini-Review Session #2

1

*Modeling data with the mean*

2

*Thinking about variability as model error*

3

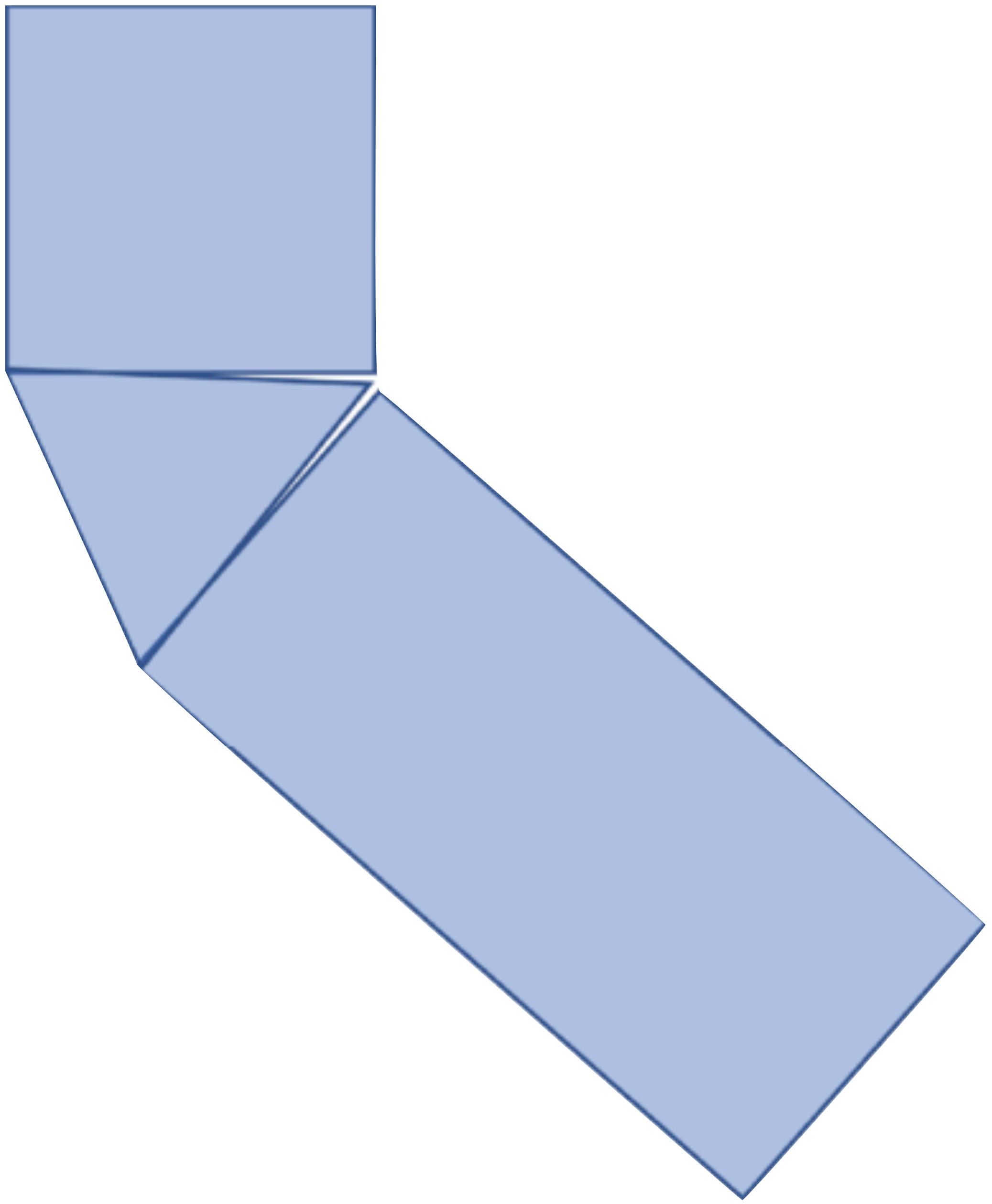
*Estimating variability*

1

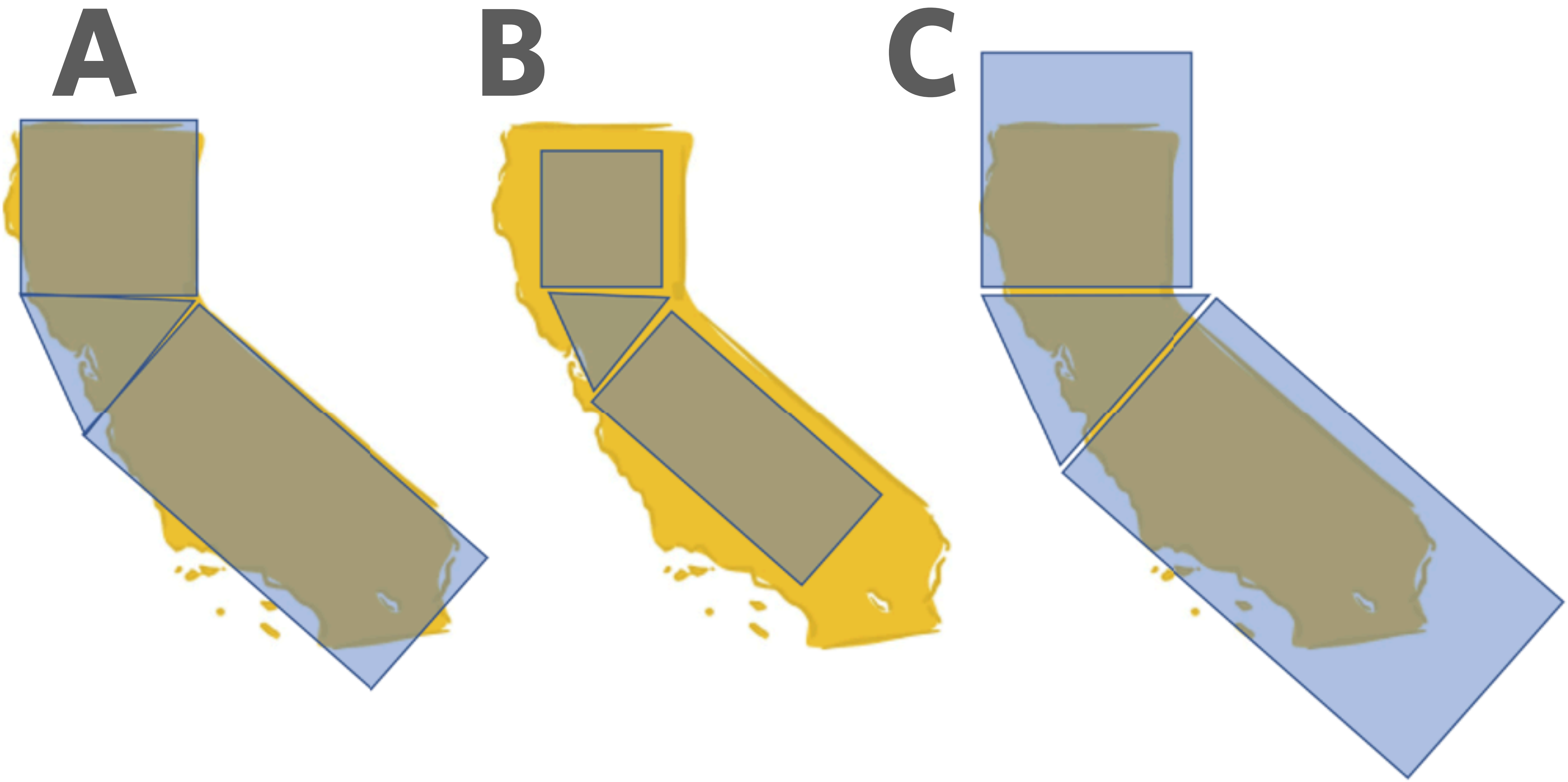
*What is a model? Why do we want one?*

1

*What is a model? Why do we want one?*



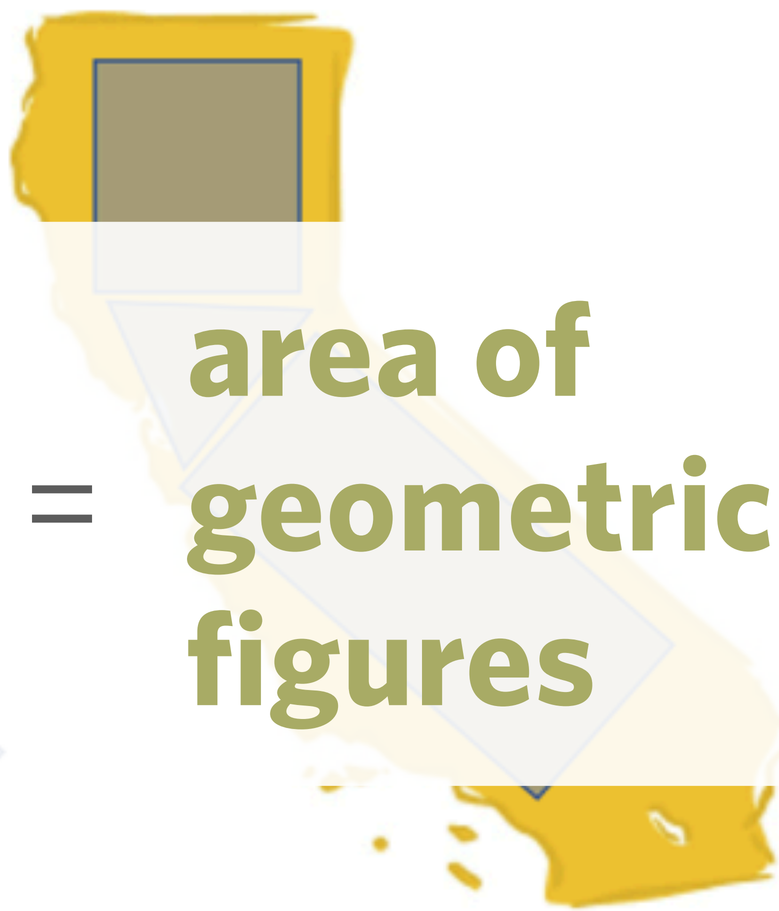
# 1 *What is a model? Why do we want one?*



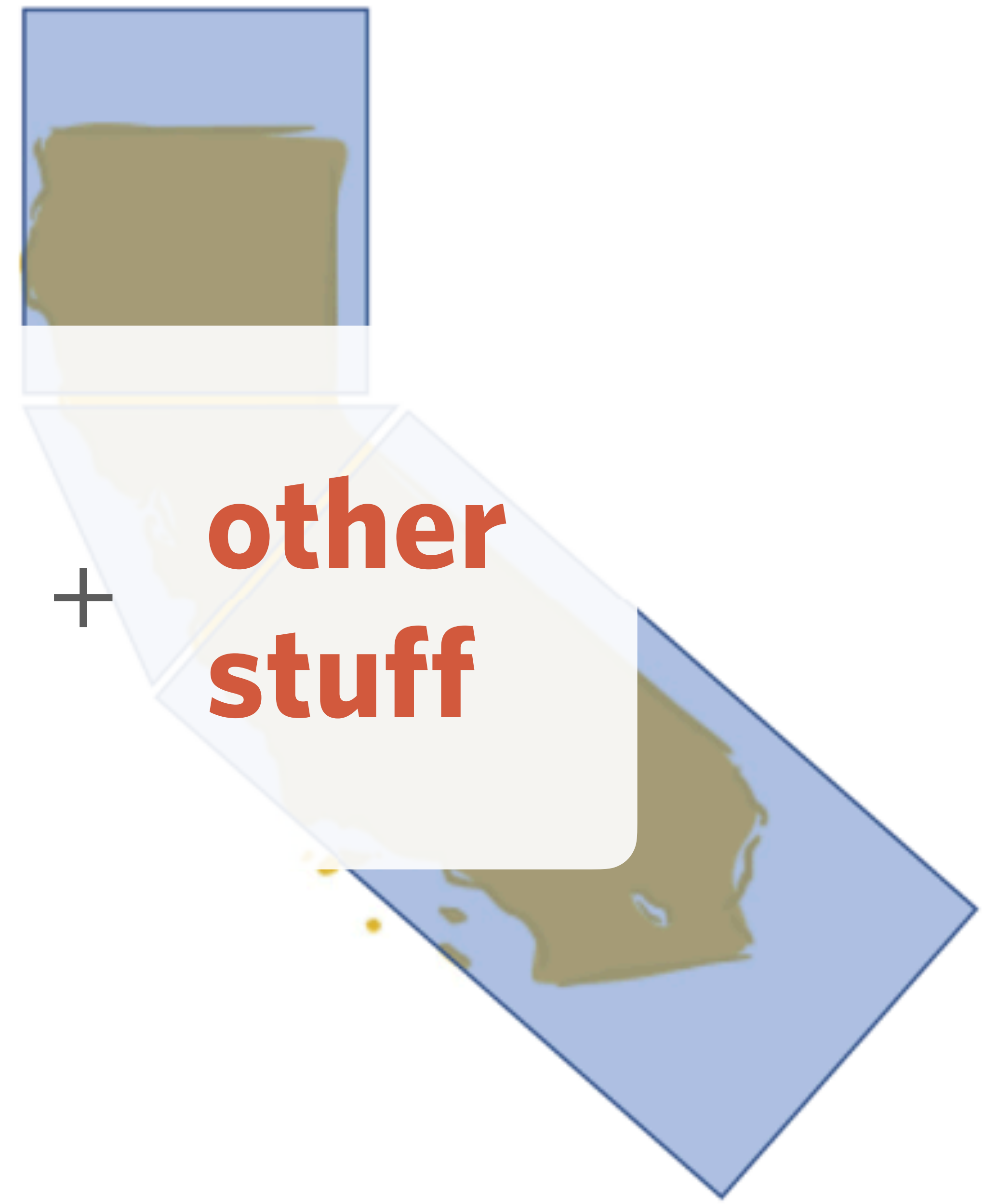
# 1 *What is a model? Why do we want one?*

**A**

**B**



**C**



**area of CA** = **area of geometric figures** + **other stuff**

1

*What is a model? Why do we want one?*

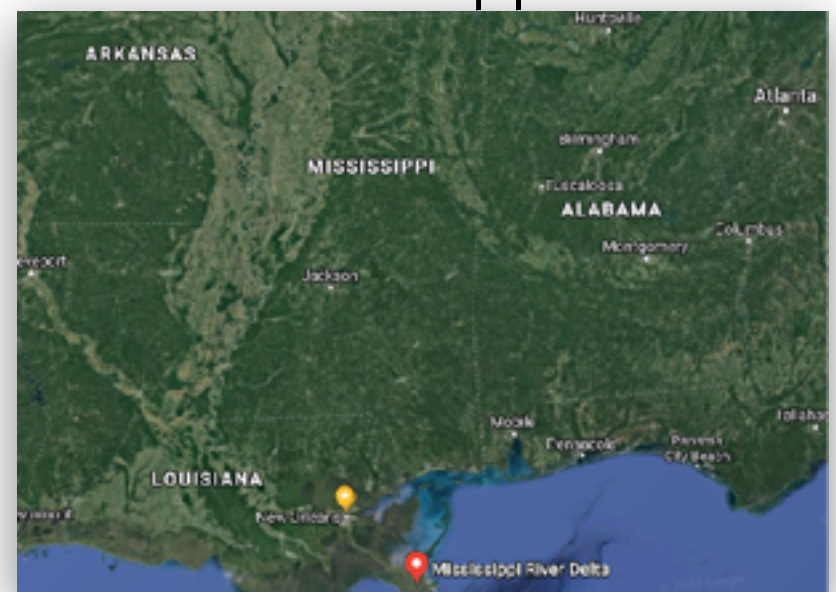
## Models simplify the world for us.

*What is a model? Why do we want one?*

## Models simplify the world for us.

Mississippi River Basin Model

Actual Mississippi River Basin



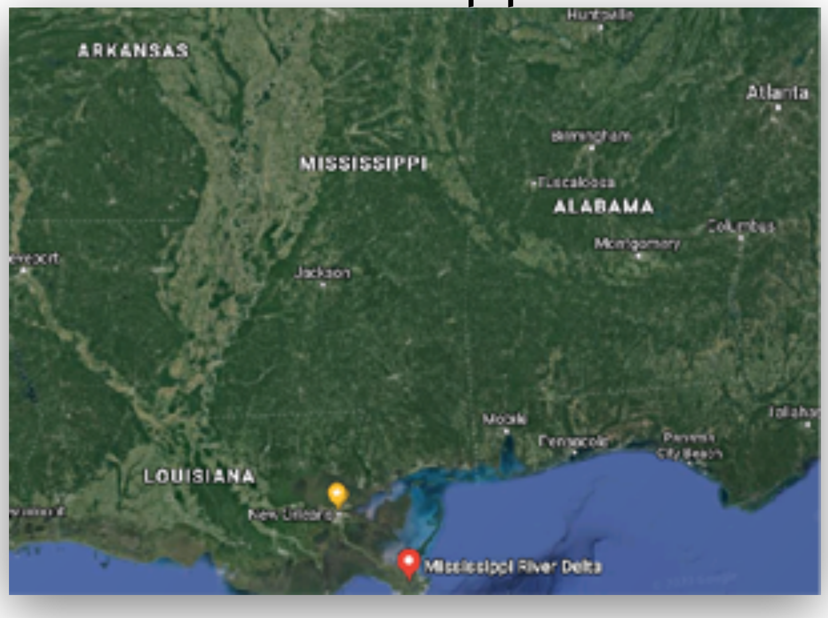
1

*What is a model? Why do we want one?*

# Models simplify the world for us.

Mississippi River Basin Model

Actual Mississippi River Basin




Model of Eukaryotic Cell


Actual Image of Cell

“

. . . In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it.

*-Jorge Luis Borges*

(*from* On Exactitude In Science)

1

*How to model data with a single number*

## Your predictions about the next random observation reveal your intuitions about the best value to ***model*** these distributions!

Best value will depend on the type of variable & shape of distribution

### For quantitative variables

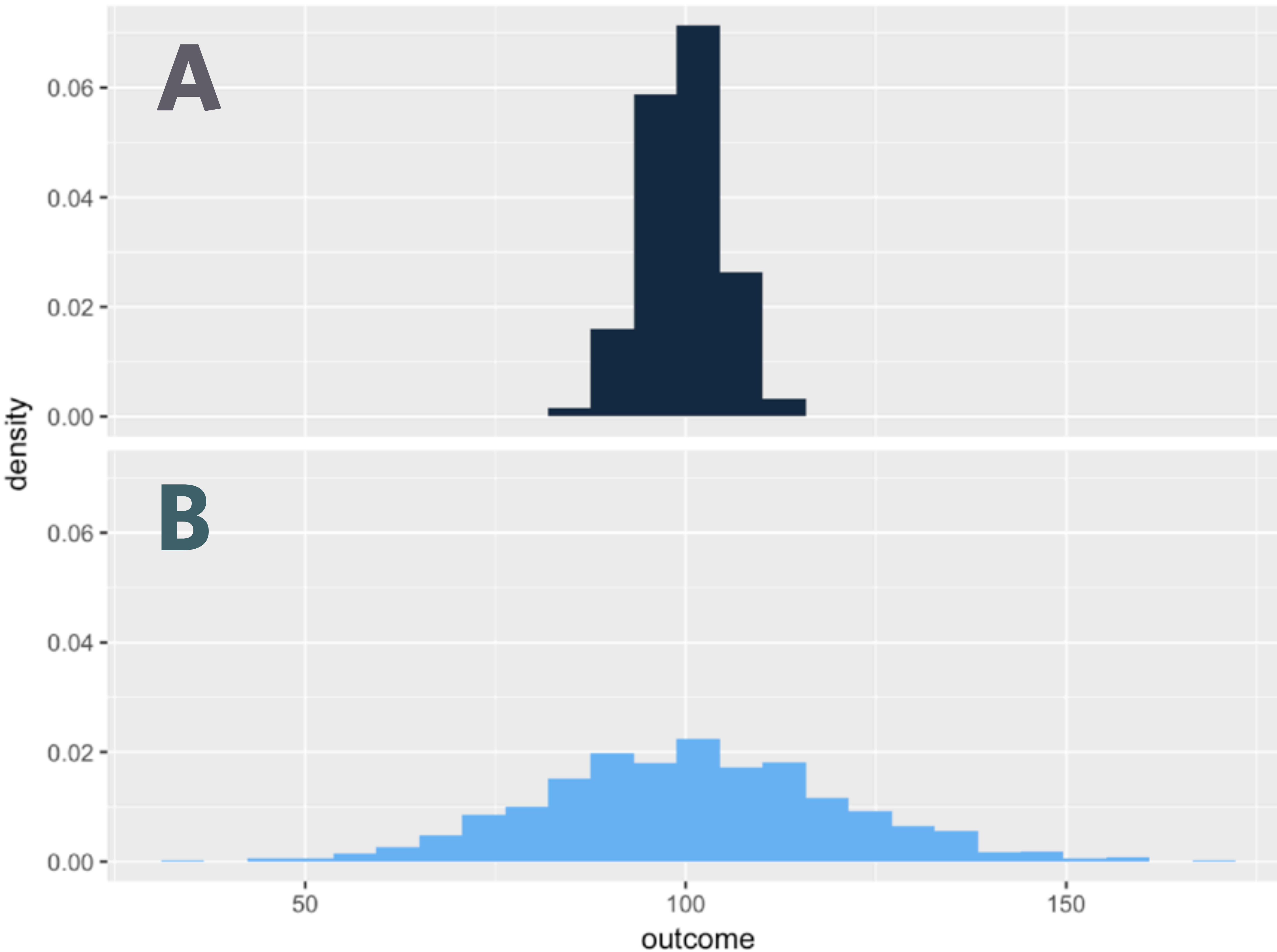
- If roughly symmetric & bell-shaped, a number right in the middle…
- If skewed, a number toward where the middle would be if you ignored the long tail

### For categorical variables

- Generally best value is the category that is most frequent

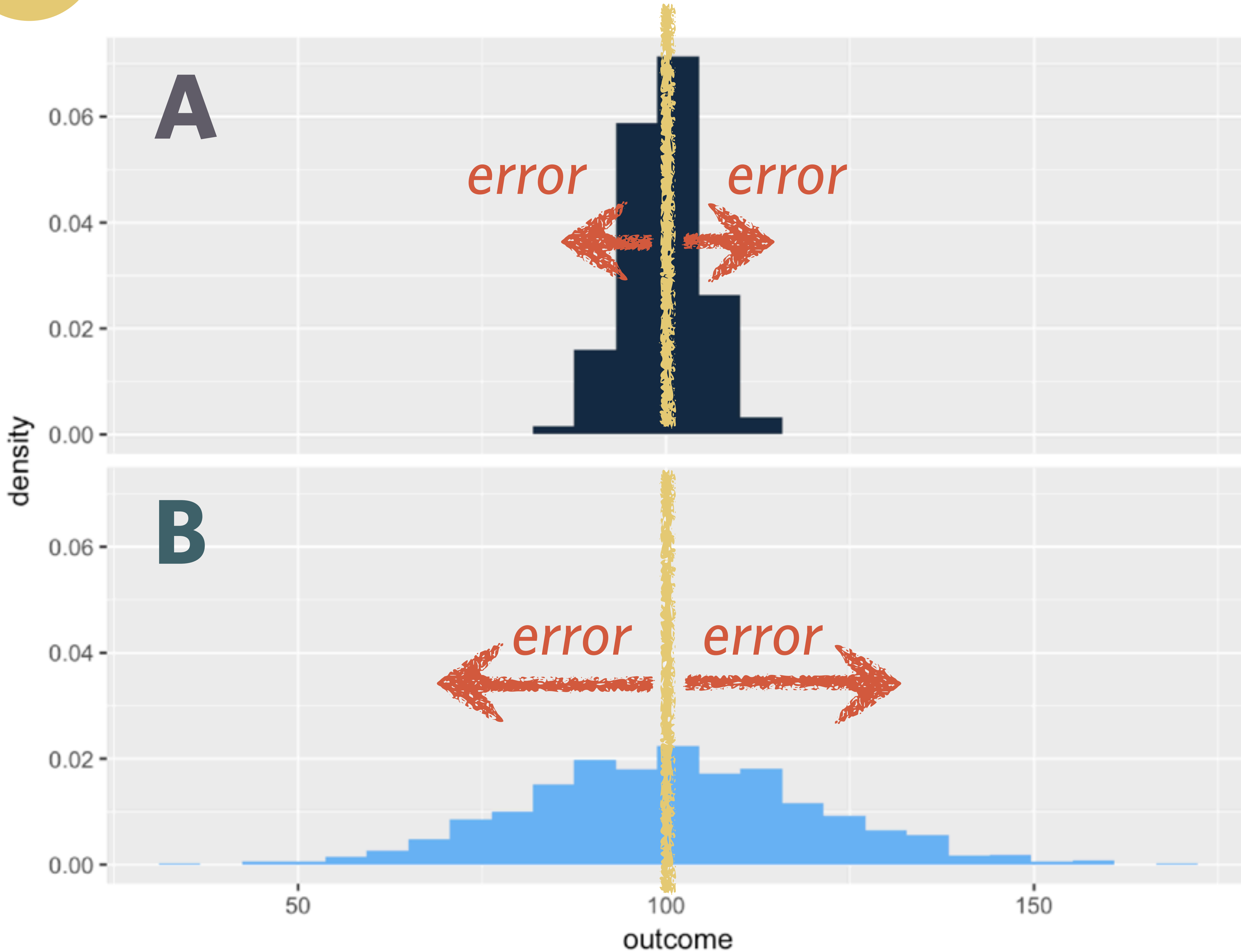
1

*How to model data with a single number*



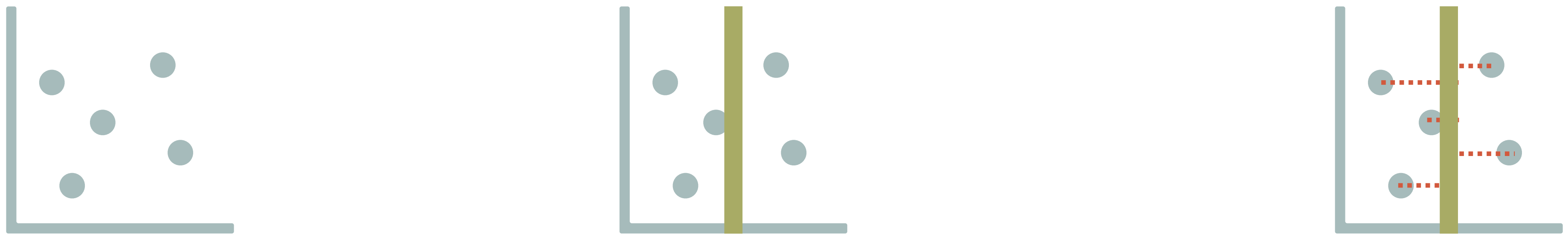
1

*How to model data with a single number*



1

# *How to model data with a single number*

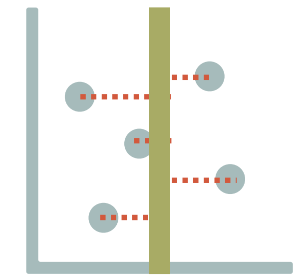


**data = model + error**

( area of CA = area of geometric figures + other stuff )

1

## *How to model data with a single number*



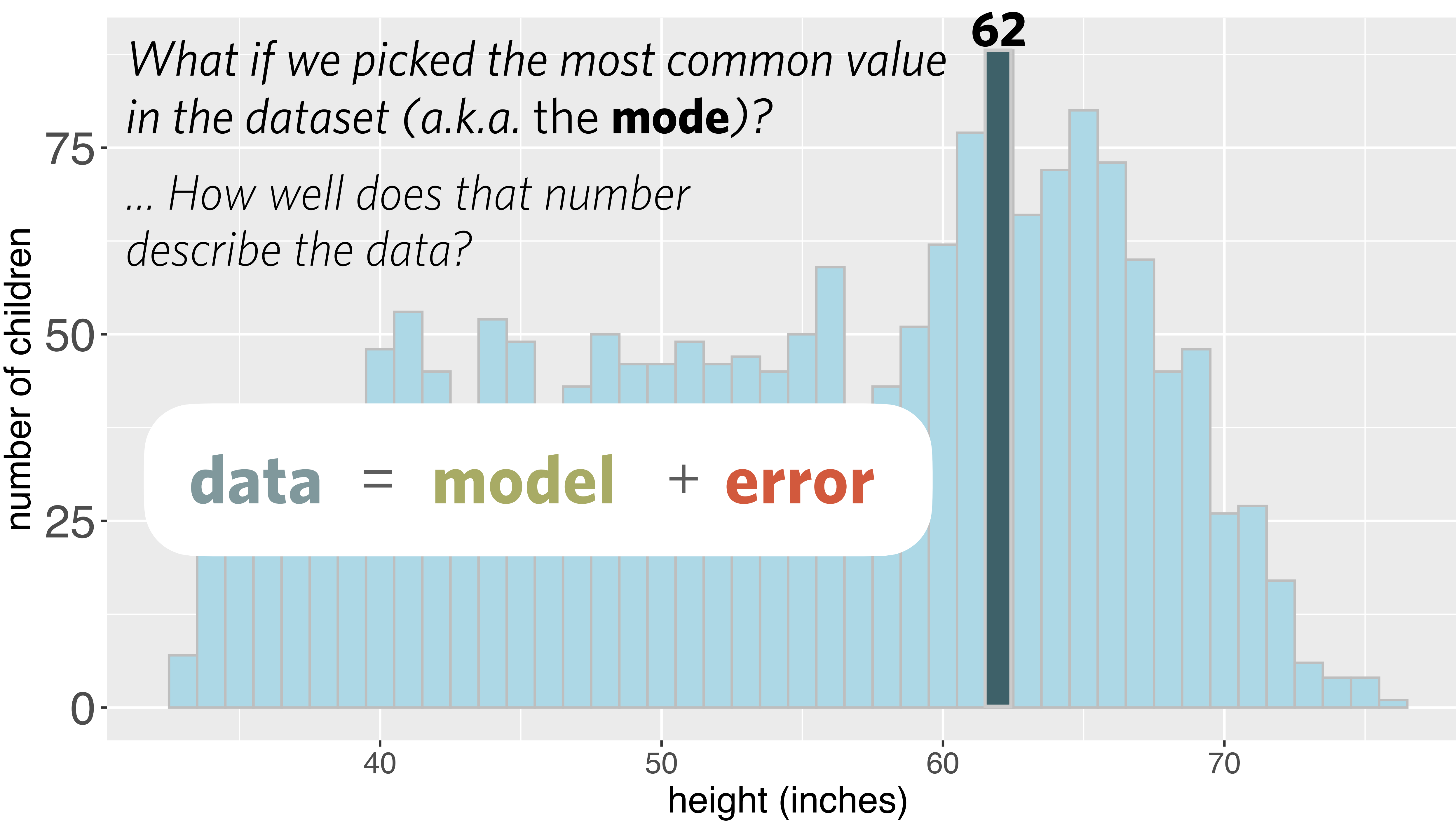
**data = model + error**

what we actually observe

what we expect to observe

difference between expected and observed

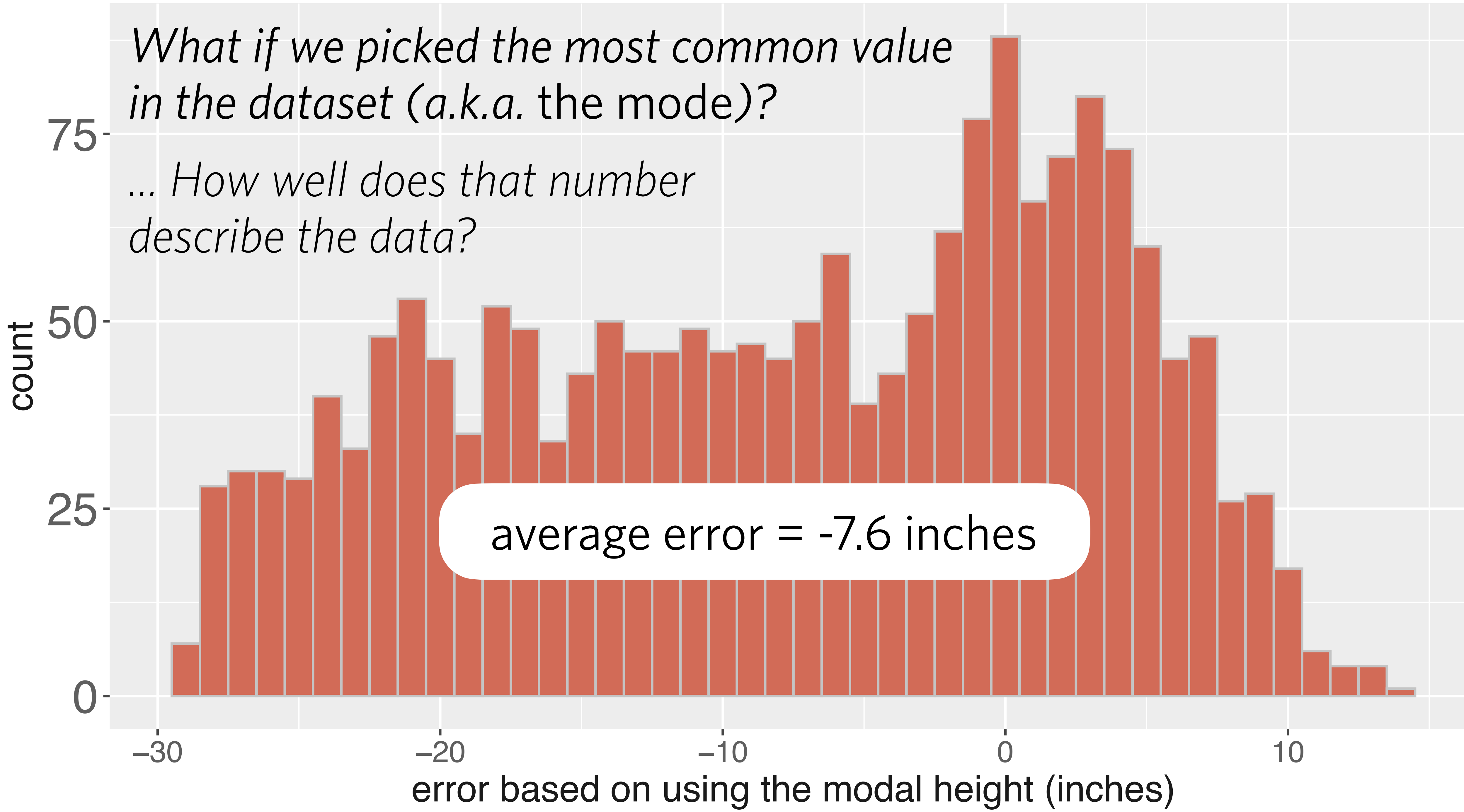
# What is our best guess for random child in NHANES?



1

# How to model data with a single number

## What is our best guess for random child in NHANES?

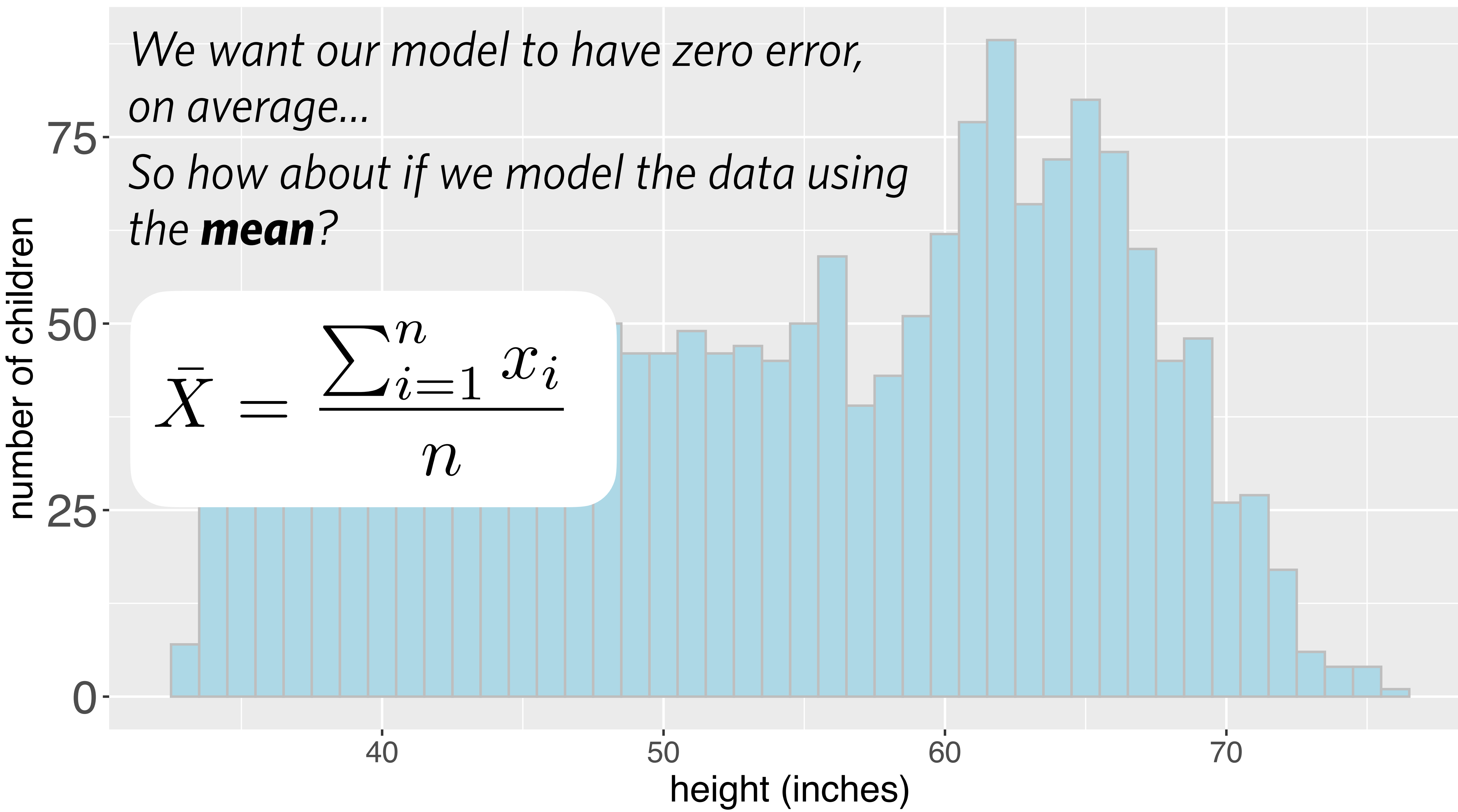


## What is our best guess for random child in NHANES?

*We want our model to have zero error, on average…*

*So how about if we model the data using the* ***mean****?*

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$



How to calculate the **sample mean**:

*"X bar" is symbol used to represent the mean of a sample*

*sum of observed values in sample*

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

*number of observations in the sample*

The sum of the errors from the **sample mean** = zero.

# 1 *How to model data with a single number*

**Note: "average" usually refers to the mean**

How to calculate the **sample mean**:

*"X bar" is symbol used to represent the mean of a sample*

*sum of observed values in sample*

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

*number of observations in the sample*

The sum of the errors from the **sample mean** = zero.

# *How to model data with a single number*

Calculating the **sample mean**:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

And the **population mean**:

*sum of all observed values in population*

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

*"mu" is symbol used to represent the population mean*

*number of observations in the whole population*

*same formula, different symbols*

# How to model data with a single number

**We can easily calculate the sample mean. We often want to infer the population mean.**

Calculating the **sample mean**:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

And the **population mean**:

*sum of all observed values in population*

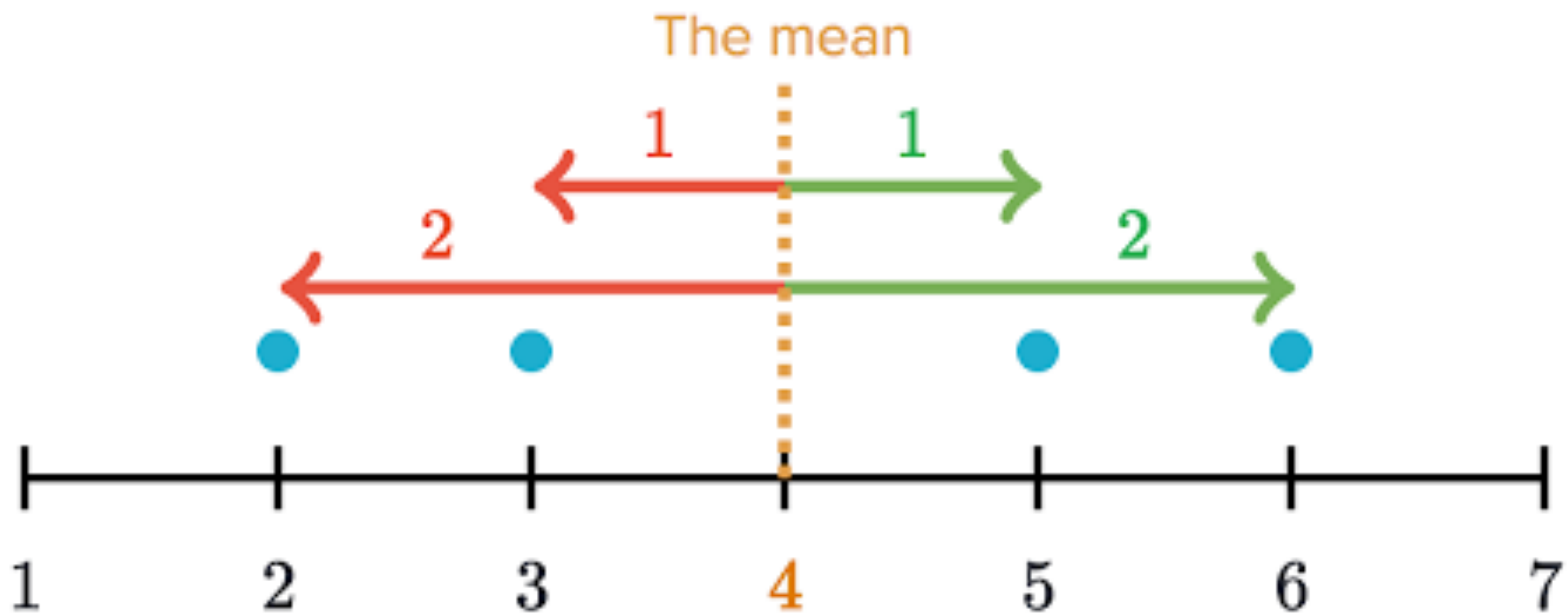
$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

*"mu" is symbol used to represent the population mean*

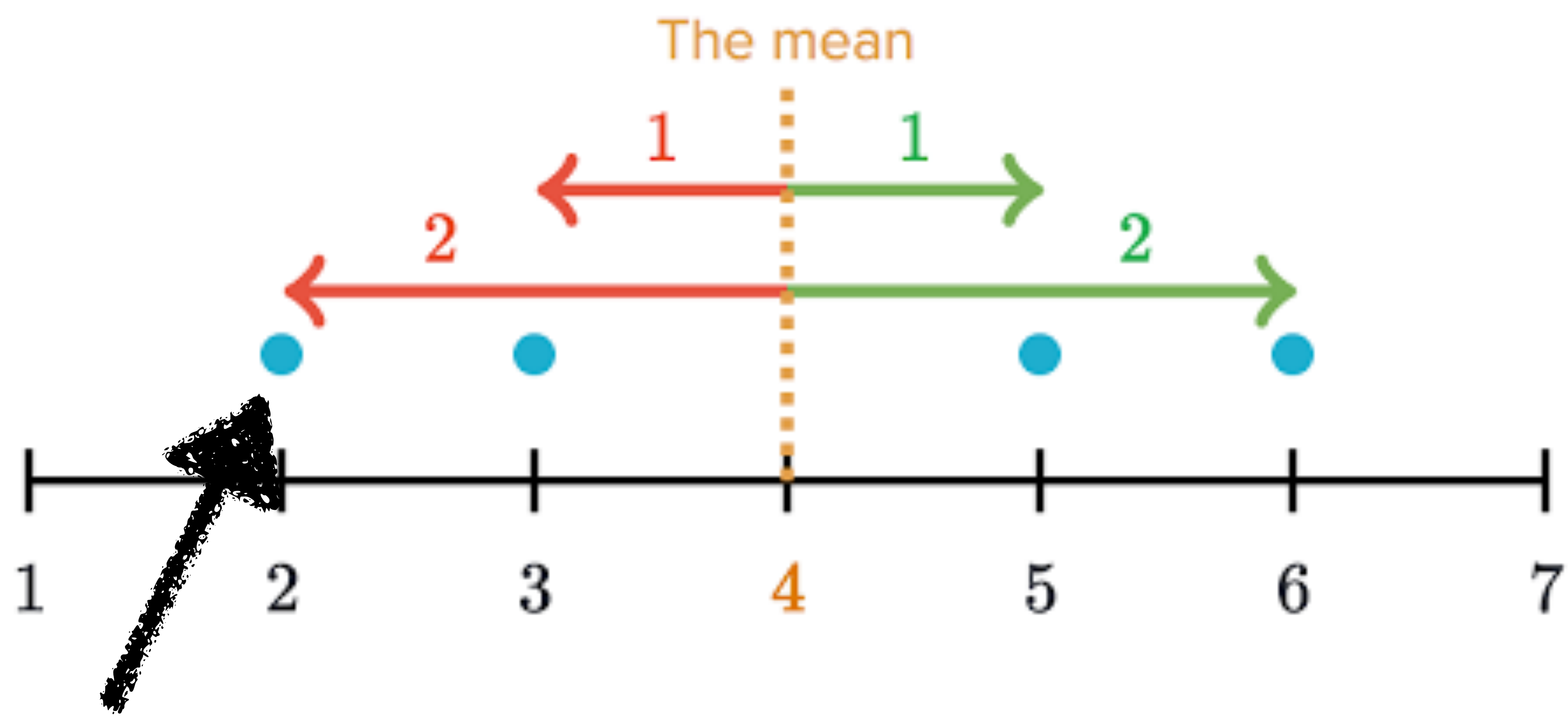
*number of observations in the whole population*

*same formula, different symbols*

The mean is the *balancing point* of the distribution.



The mean is the *balancing point* of the distribution.




You can think of this blue dot as having some "**deviation**" from the mean. The deviation means its distance from the mean and isn't the same thing as **"standard deviation"** (more on that later)

The sum of the errors from the the **sample mean** = zero.

Try it out yourself!

```
d <- c(3,5,6,7,9)
mean(d)
[1] 6

errors=d-mean(d)
print(errors)
[1] -3 -1  0  1  3

print(sum(errors))
[1] 0
```

| x | error |
|---|---|
| 3 | -3 |
| 5 | -1 |
| 6 | 0 |
| 7 | 1 |
| 9 | 3 |

sum=0

1

## *How to model data with a single number*

The mean is the "best" estimate because it minimizes the sum of squared errors (abbreviated SSE below)

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

1

*How to model data with a single number*

The mean is the "best" estimate because it minimizes the sum of squared errors (abbreviated SSE below)

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

*Sum of Squared Errors*

The mean is the "best" estimate because it minimizes the sum of squared errors (abbreviated SSE below)

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

*Sum of Squared Errors*

*the "i-th" observation*

The mean is the "best" estimate because it minimizes the sum of squared errors (abbreviated SSE below)

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

*Sum of Squared Errors*

*the "i-th" observation*

*"x hat" is a symbol representing our prediction*

The mean is the "best" estimate because it minimizes the sum of squared errors (abbreviated SSE below)

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

Sum of Squared Errors

the "i-th" observation

"x hat" is a symbol representing our prediction

$$model\, prediction : \hat{x} = mean(x) = \frac{\sum_{i=1}^{n} x_i}{n}$$

1

*How to model data with a single number*

# One not-so-useful feature of the mean:

| people | income |
|---|---|
| Joe | 48000 |
| Karen | 64000 |
| Mark | 58000 |
| Andrea | 72000 |
| Pat | 66000 |

w/o Beyoncé:

**mean income: $61,600**

1

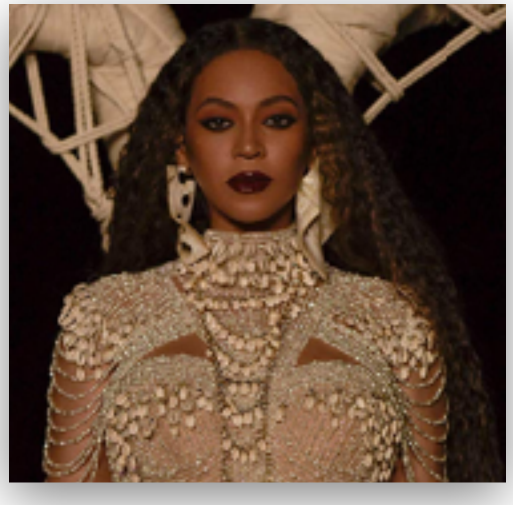
*How to model data with a single number*

# One not-so-useful feature of the mean:

| people | income |
|---|---|
| Joe | 48000 |
| Karen | 64000 |
| Mark | 58000 |
| Andrea | 72000 |
| Pat | 66000 |

→

| people | income |
|---|---|
| Joe | 48000 |
| Karen | 64000 |
| Mark | 58000 |
| Andrea | 72000 |
| Beyonce | 54,000,000 |



w/o Beyoncé:

**mean income: $61,600**

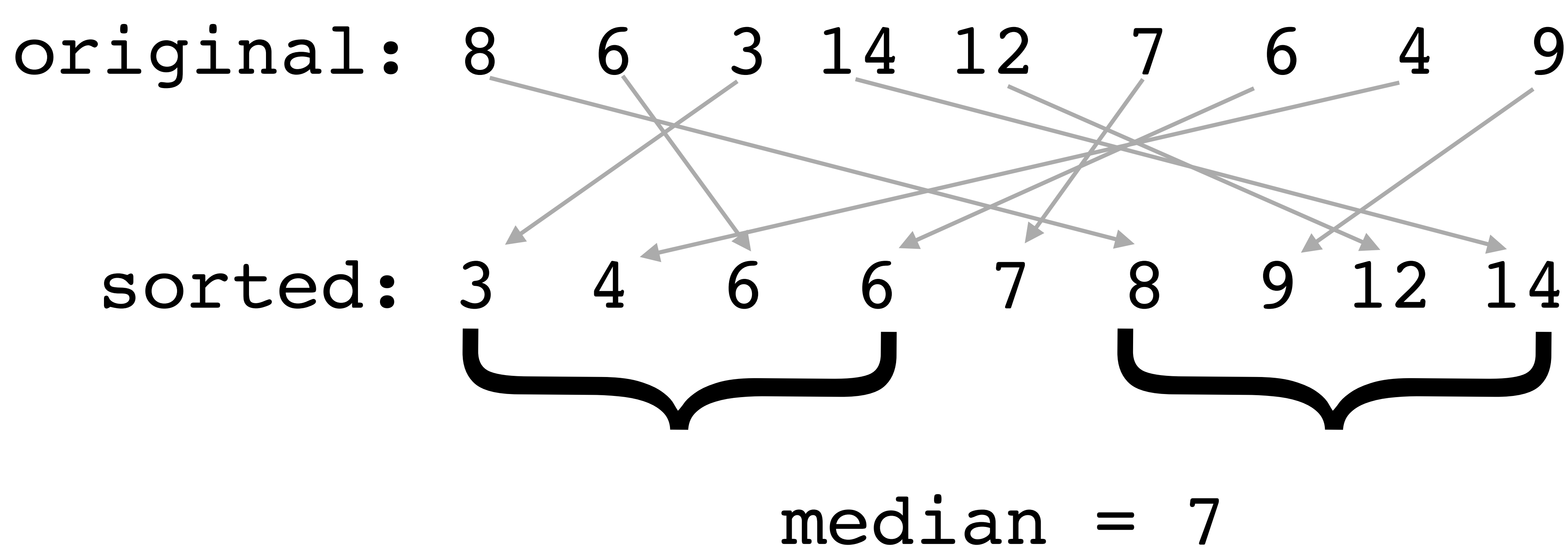
w/ Beyoncé:

**mean income: $10,848,400**

## Introducing the **median:**

When the scores are ordered from smallest to largest, the median is the middle score

When there is an even number of scores, the median is the average between the middle two scores



**The median minimizes the *sum of absolute errors*:**

$$SAE = \sum_{i=1}^{n} |x_i - \hat{x}|$$

**The mean minimizes the *sum of squared errors*:**

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

When might that difference matter?

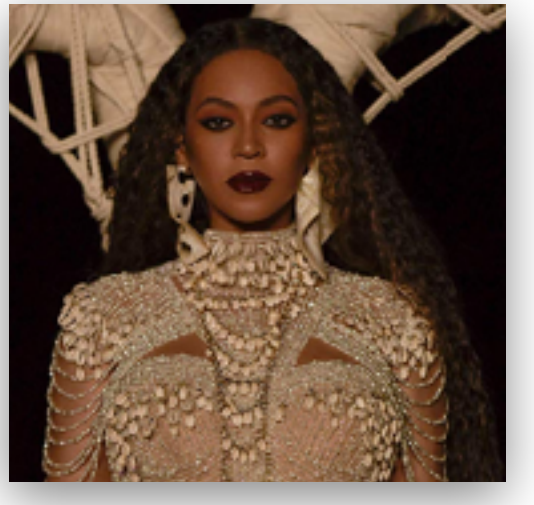
1

*How to model data with a single number*

# One not-so-useful feature of the mean:

| people | income |
|---|---|
| Joe | 48000 |
| Karen | 64000 |
| Mark | 58000 |
| Andrea | 72000 |
| Pat | 66000 |

| people | income |
|---|---|
| Joe | 48000 |
| Karen | 64000 |
| Mark | 58000 |
| Andrea | 72000 |
| Beyonce | 54,000,000 |



w/o Beyoncé:
mean income: $61,600
**median income: $64,000**

w/ Beyoncé:
mean income: $10,848,400
**median income: $64,000**

So why would we ever use the **mean** instead of the **median**?

The mean is the "best" estimator

***It bounces around less from sample to sample than any other estimator.***

But the median is more robust to outliers.

Such tradeoffs are unavoidable in statistics.

# Today

## Mini-Review Session #2

1 → 2 → 3

1. *Modeling data with the mean*
2. *Thinking about variability as model error*
3. *Estimating variability*

2

## *How to know how well a model fits*

The mean is the "best" estimate because it minimizes the sum of squared errors (abbreviated SSE below)

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

*Sum of Squared Errors*

*the "i-th" observation*

*"x hat" is a symbol representing our prediction*

$$model\, prediction : \hat{x} = mean(x) = \frac{\sum_{i=1}^{n} x_i}{n}$$

2

# How to know how well a model fits

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

Sum of Squared Errors

the "i-th" observation

"x hat" is a symbol representing our prediction

2

# *How to know how well a model fits*

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

*Sum of Squared Errors*

*the "i-th" observation*

*"x hat" is a symbol representing our prediction*

To obtain a measure of model error that does not depend on the number of observations, you can compute the **Root Mean Squared Error,** which you calculate by dividing SSE by the number of observations, then taking the square root:

2

## *How to know how well a model fits*

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

*Sum of Squared Errors*

*the "i-th" observation*

*"x hat" is a symbol representing our prediction*

To obtain a measure of model error that does not depend on the number of observations, you can compute the **Root Mean Squared Error,** which you calculate by dividing SSE by the number of observations, then taking the square root:

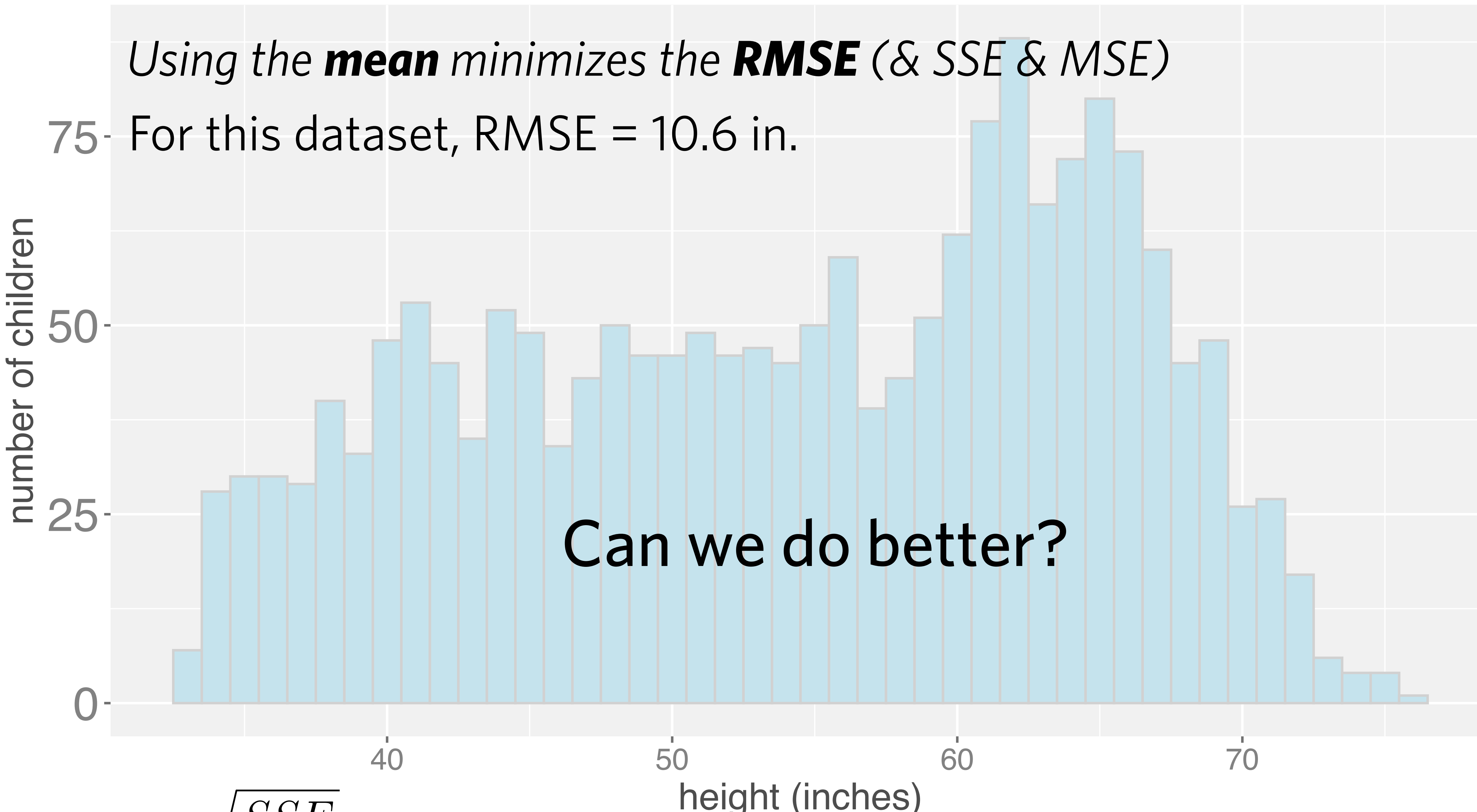
$$RMSE = \sqrt{\frac{SSE}{n}}$$

*Root-Mean-Squared Error*

## What is our best guess for random child in NHANES?

*Using the* ***mean*** *minimizes the* ***RMSE*** *(& SSE & MSE)*

For this dataset, RMSE = 10.6 in.




Can we do better?

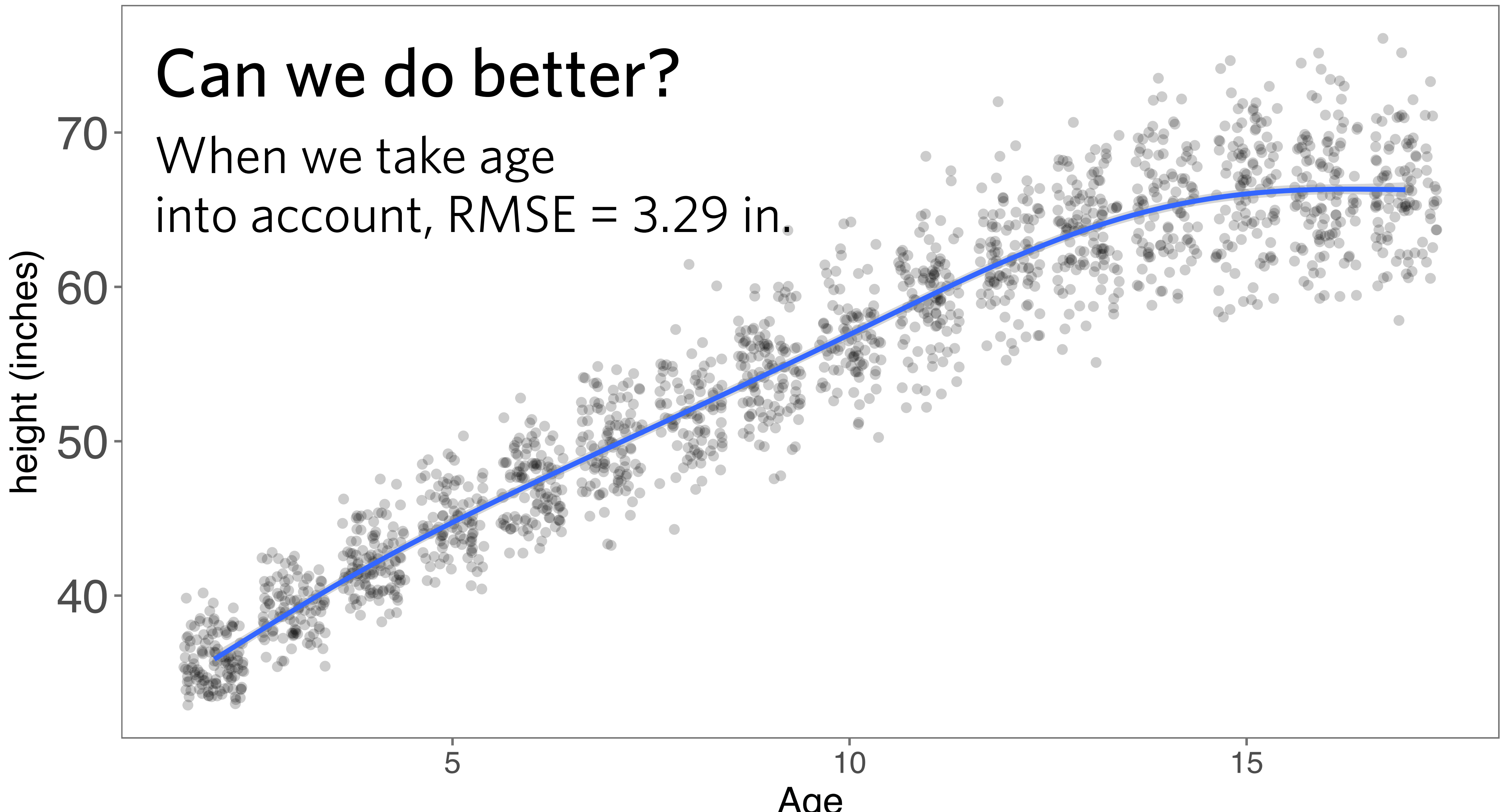
$$RMSE = \sqrt{\frac{SSE}{n}}$$

2

*How to know how well a model fits*

## What is our best guess for random child in NHANES?

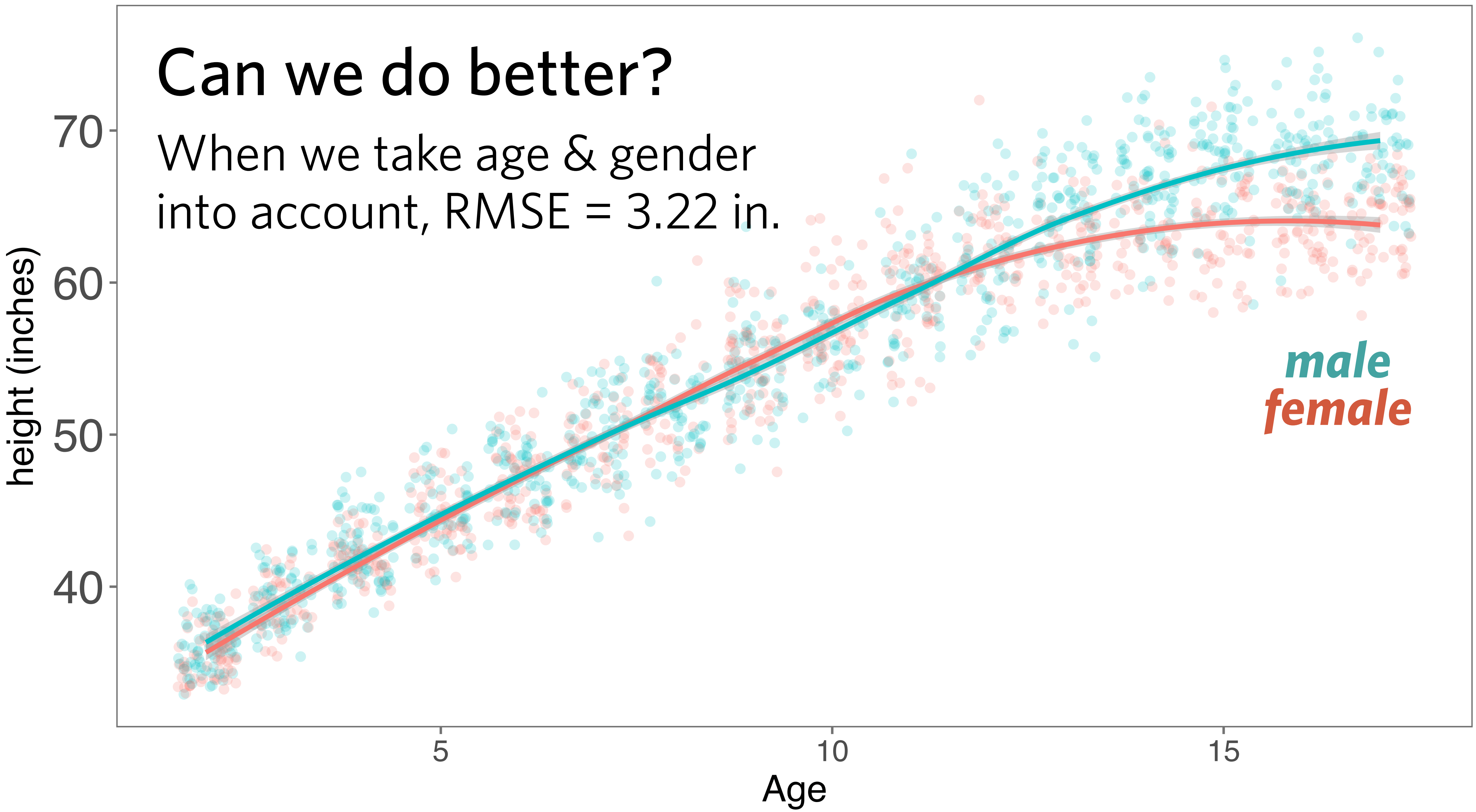
What about their age? Let's plot height vs. age and see how they are related.



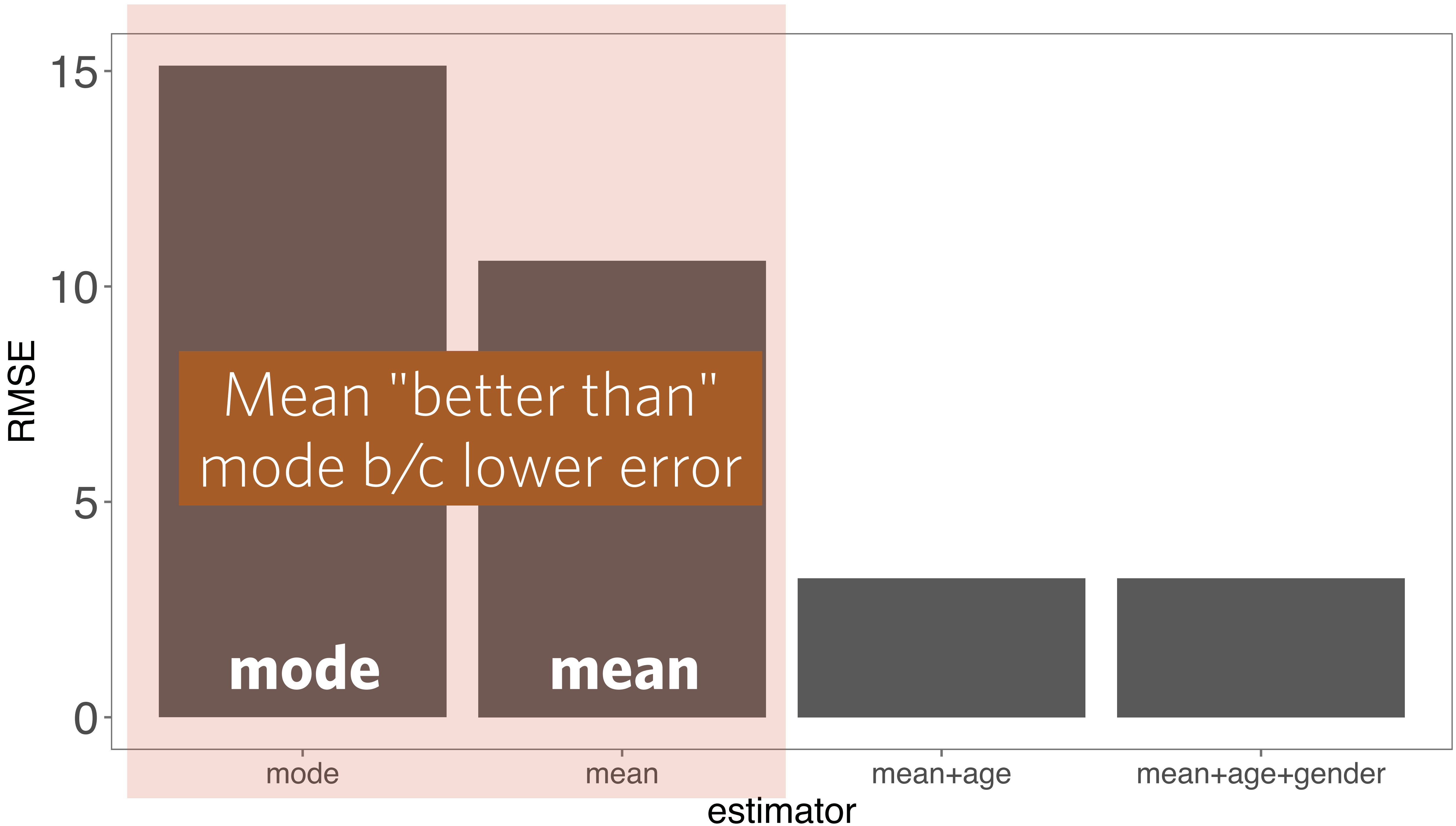
2 *How to know how well a model fits*

## What is our best guess for random child in NHANES?

What about their age? Let's plot height vs. age and see how they are related.



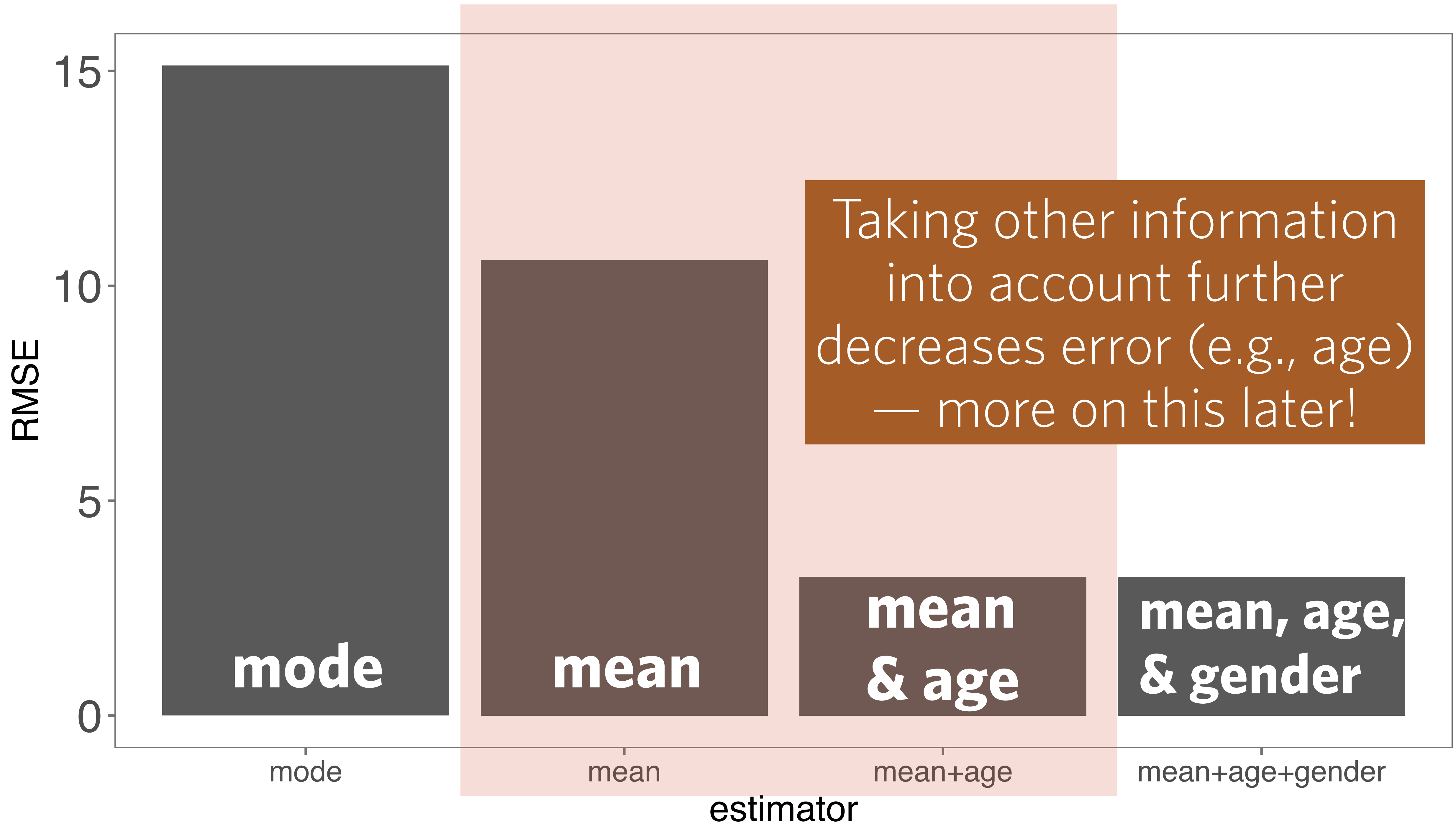
## What is our best guess for random child in NHANES?



2

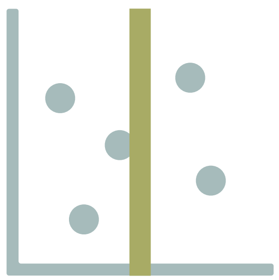
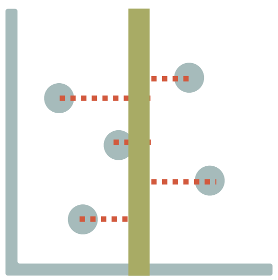
*How to know how well a model fits*

## What is our best guess for random child in NHANES?



2

# How to know how well a model fits

**data = model + error**

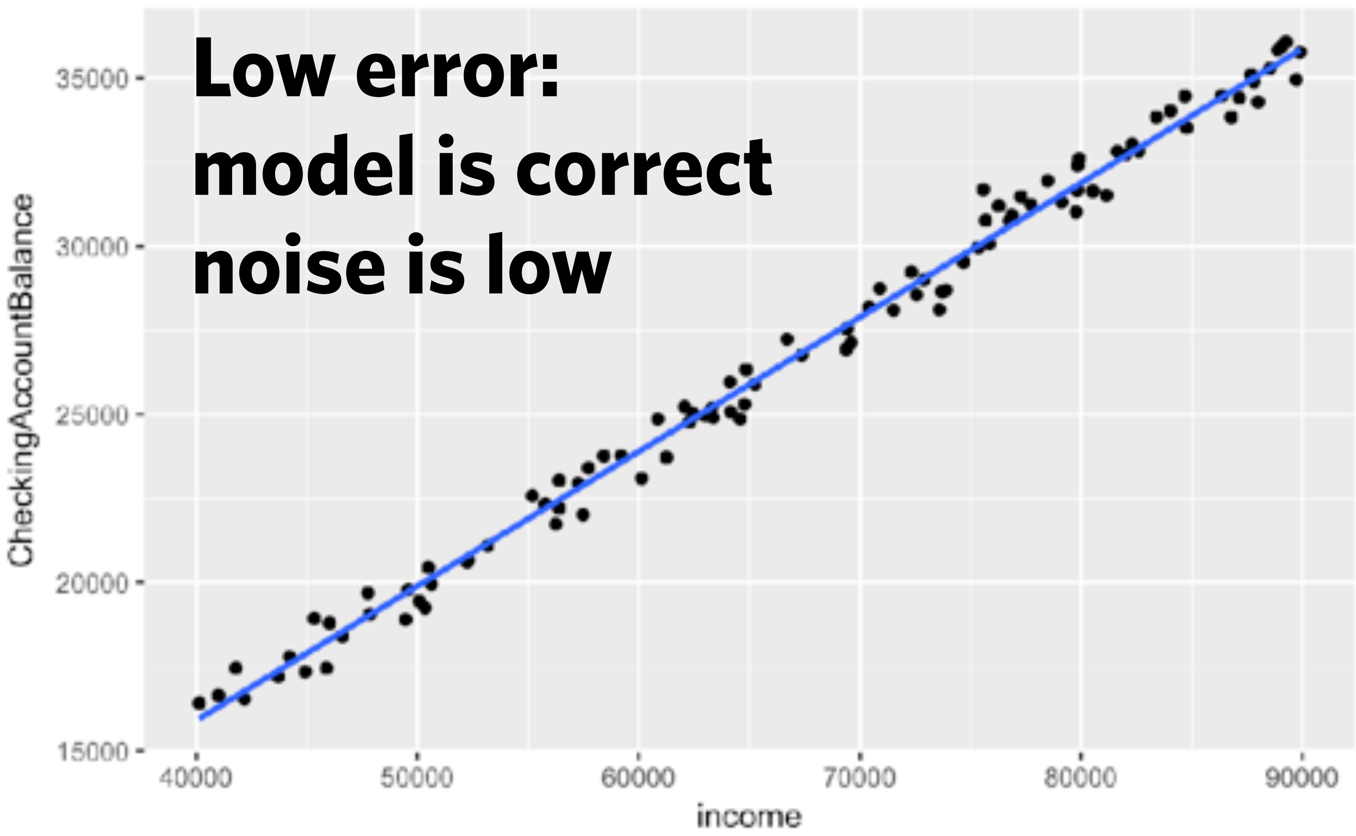
| data | model | error |
|---|---|---|
| what we actually observe | what we expect to observe | difference between expected and observed |

***Error can come from two sources:***

***(1) The model is incorrect***

***(2) The measurements have random error ("noise")***

# 2 *How to know how well a model fits*



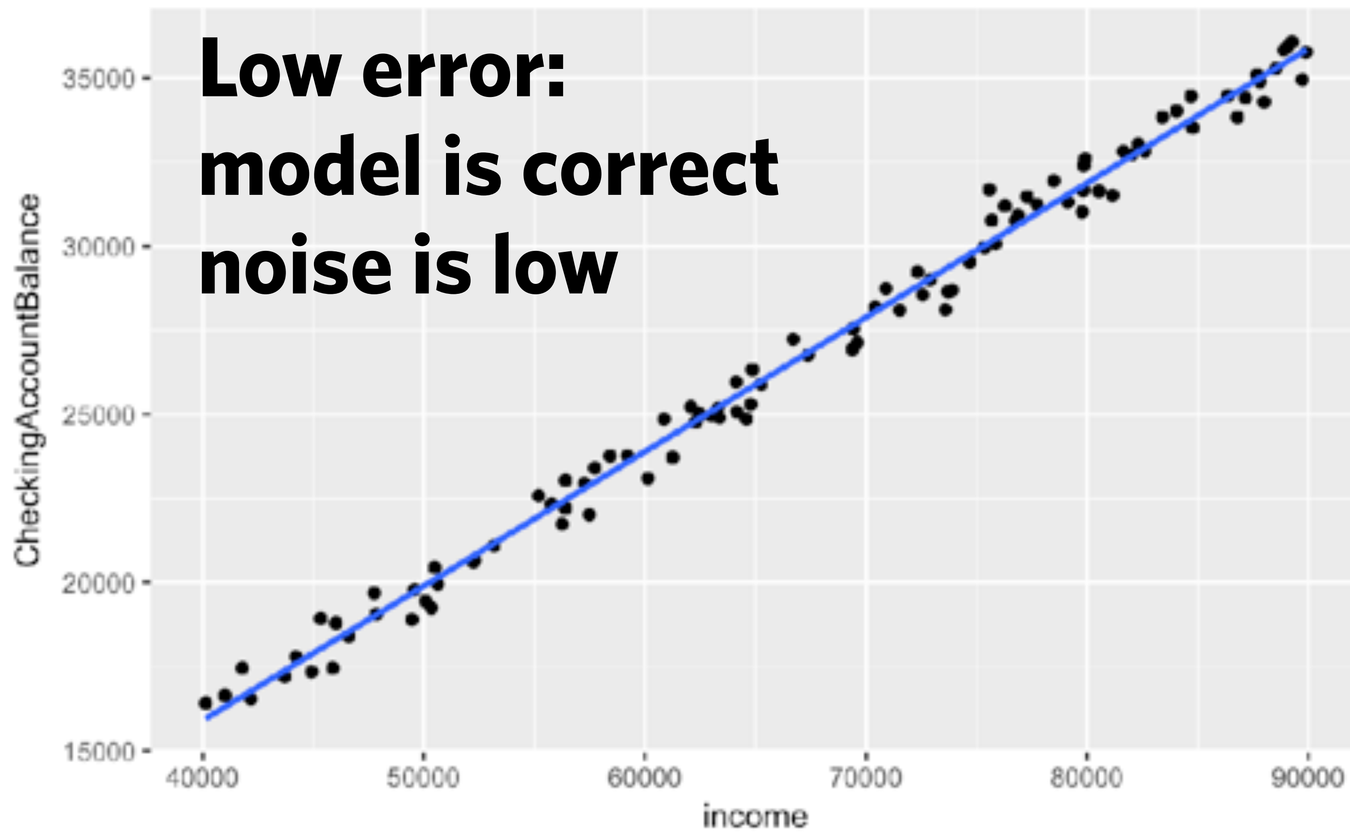

Error can come from two sources:

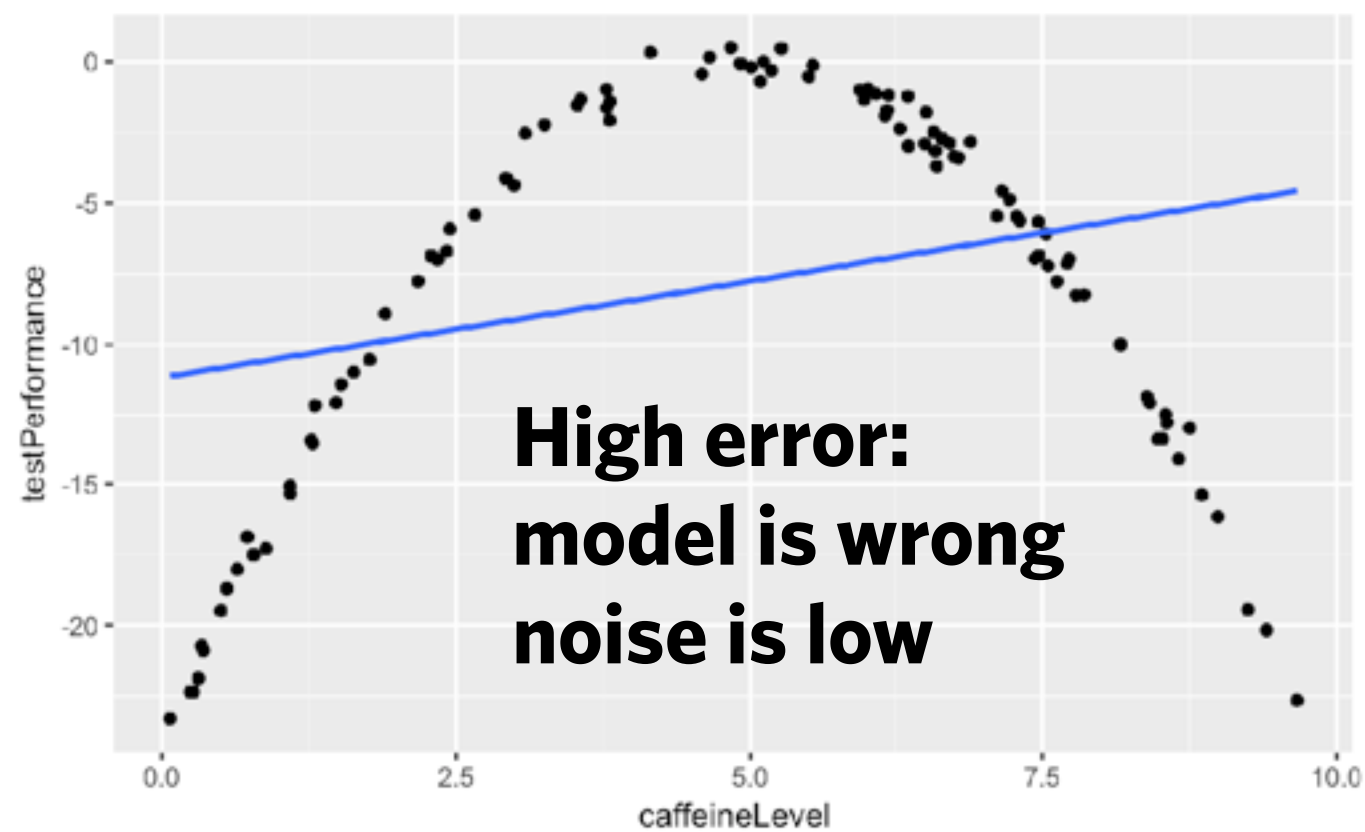
- incorrect model
- noisy data

# 2 *How to know how well a model fits*

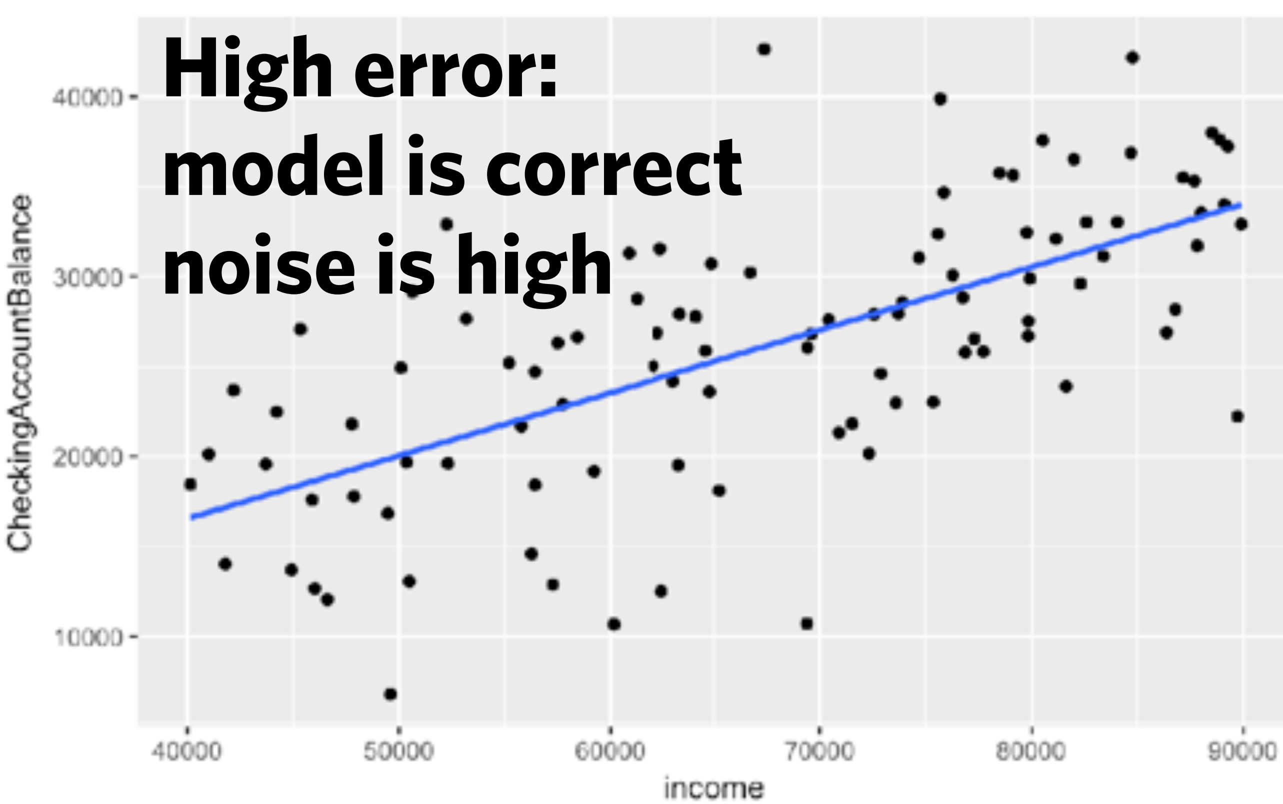
Error can come from two sources:

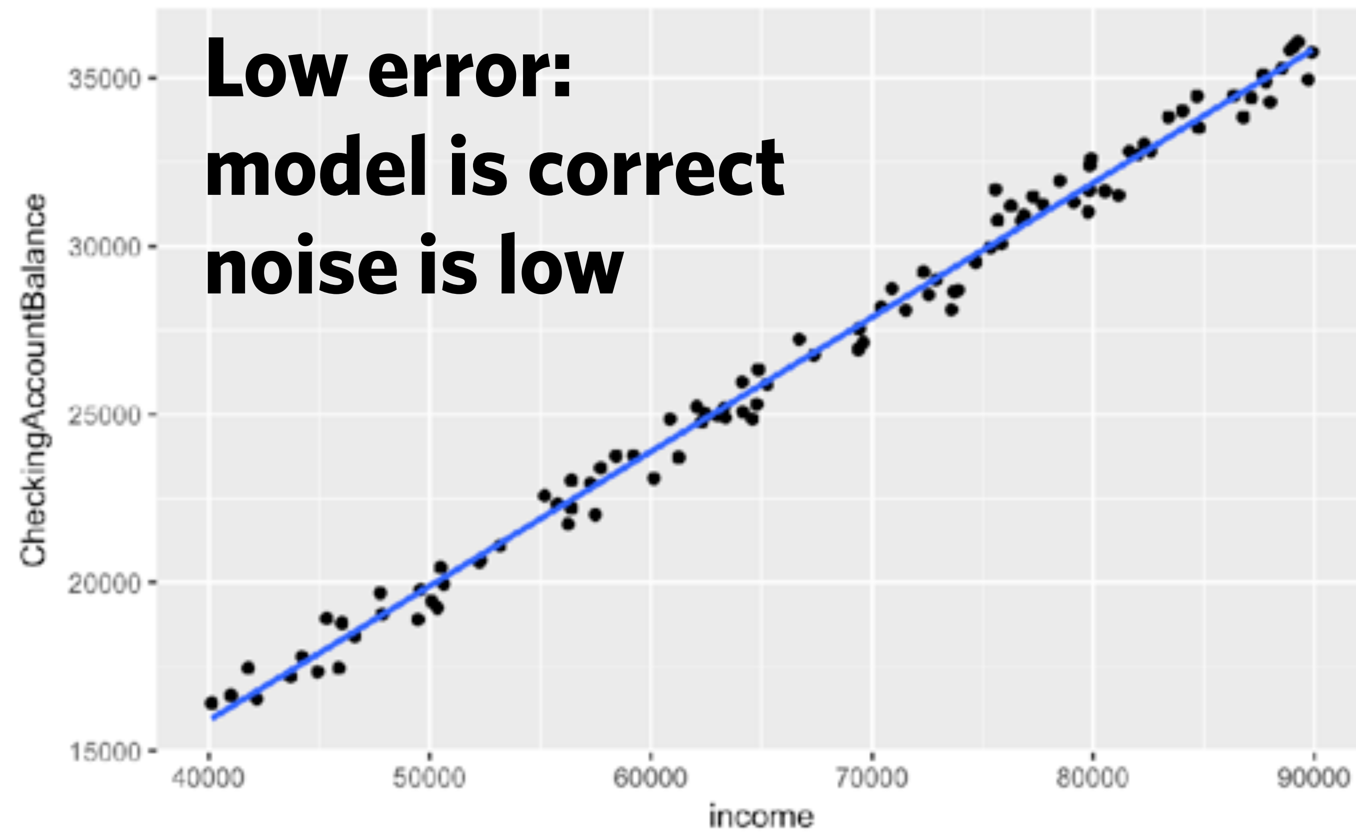
- incorrect model
- noisy data



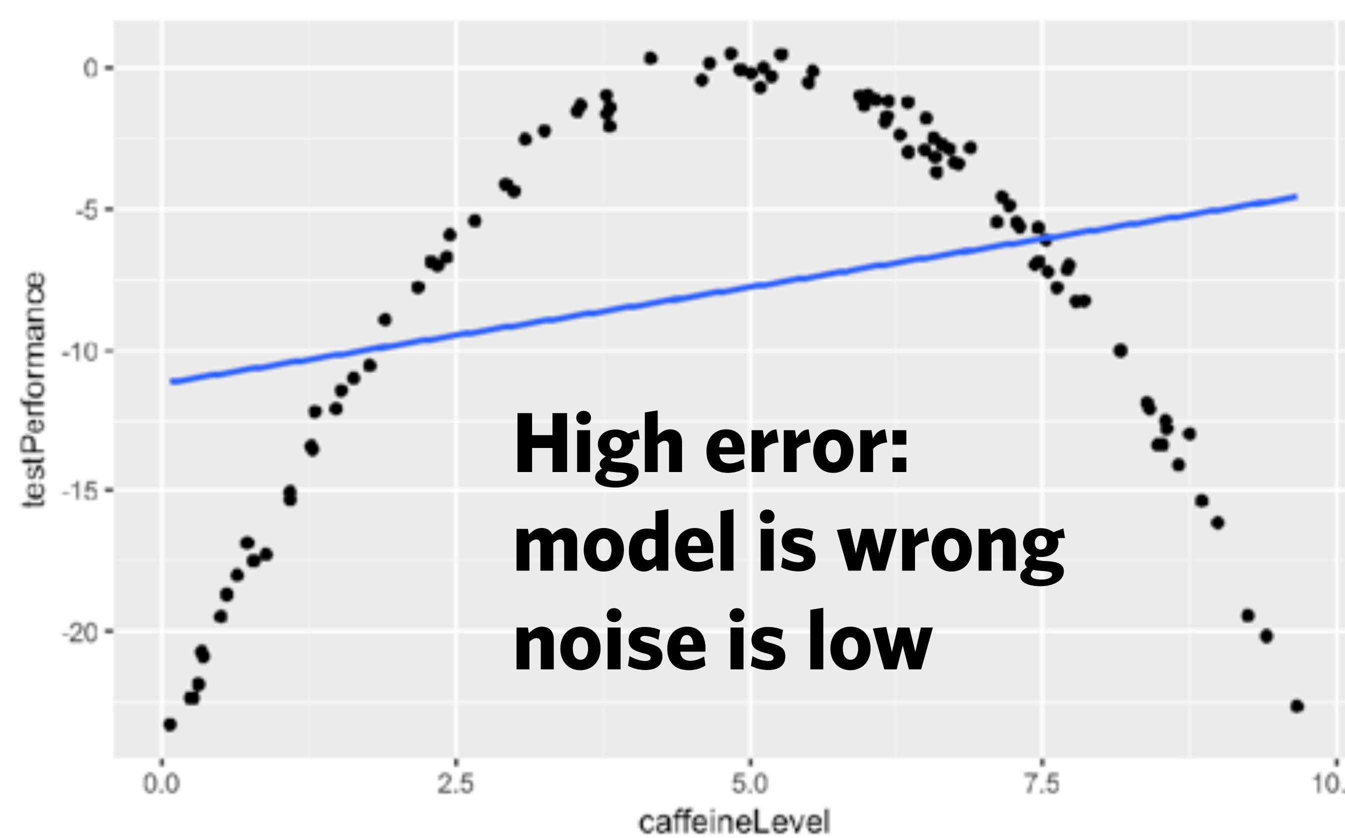
# 2 *How to know how well a model fits*




Error can come from two sources:

- incorrect model
- noisy data



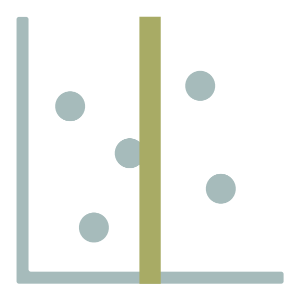
# 2 *How to know how well a model fits*

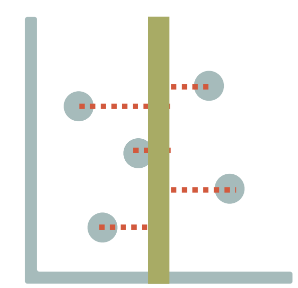
## What makes a model "good"?

**Describes** current dataset well: the error for the fitted data is low

**Generalizes** to new data well: the error for new data is low

*These two are often in conflict!*





**data = model + error**

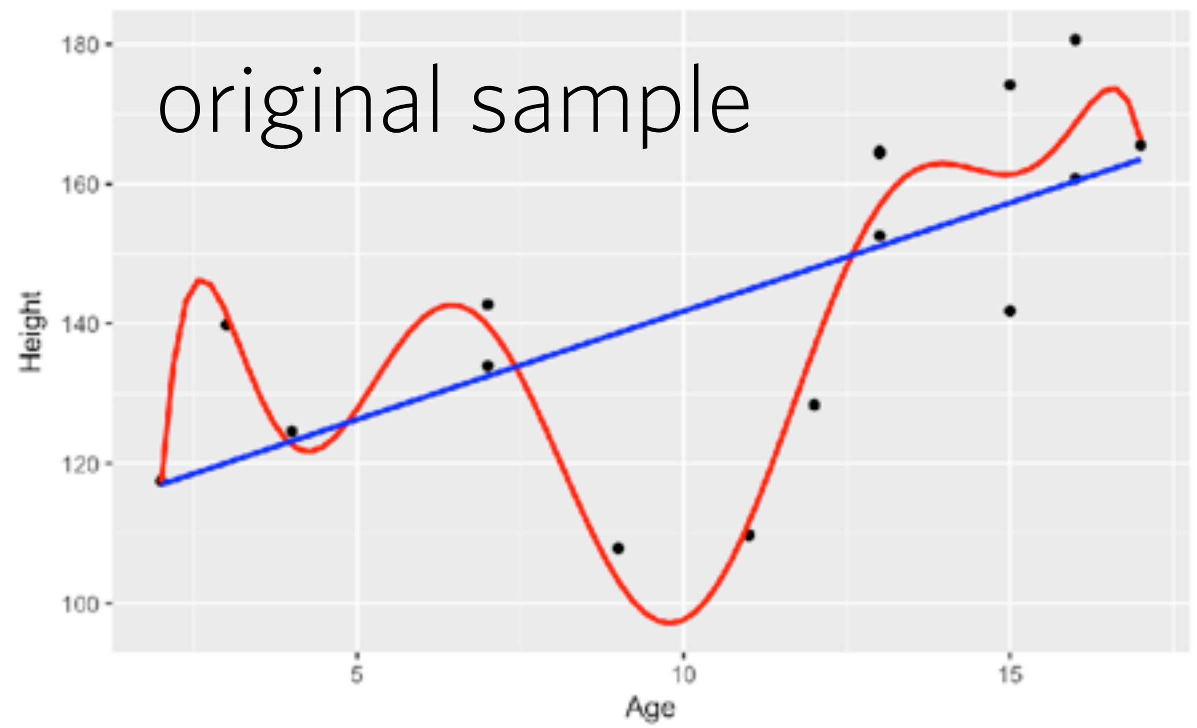
what we actually observe

what we expect to observe

difference between expected and observed

# Overfitting

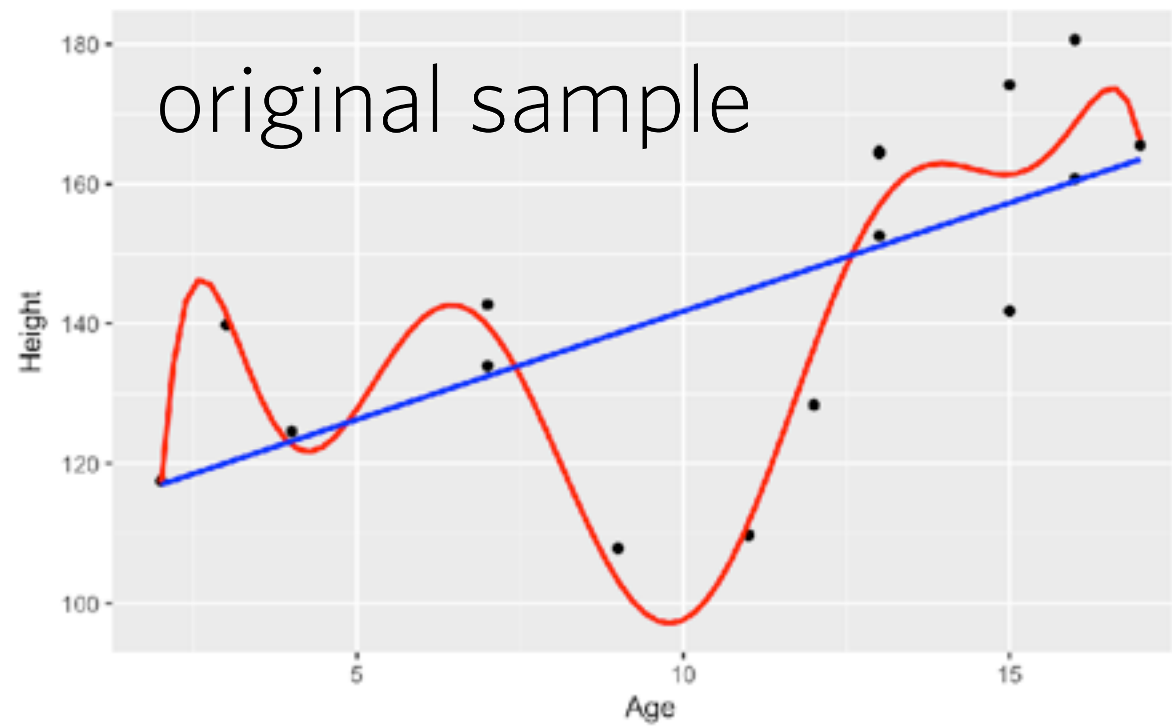
- A more complex model will always fit the data better than a simpler model
  - The model fits the underlying signal as well as the random noise in the data

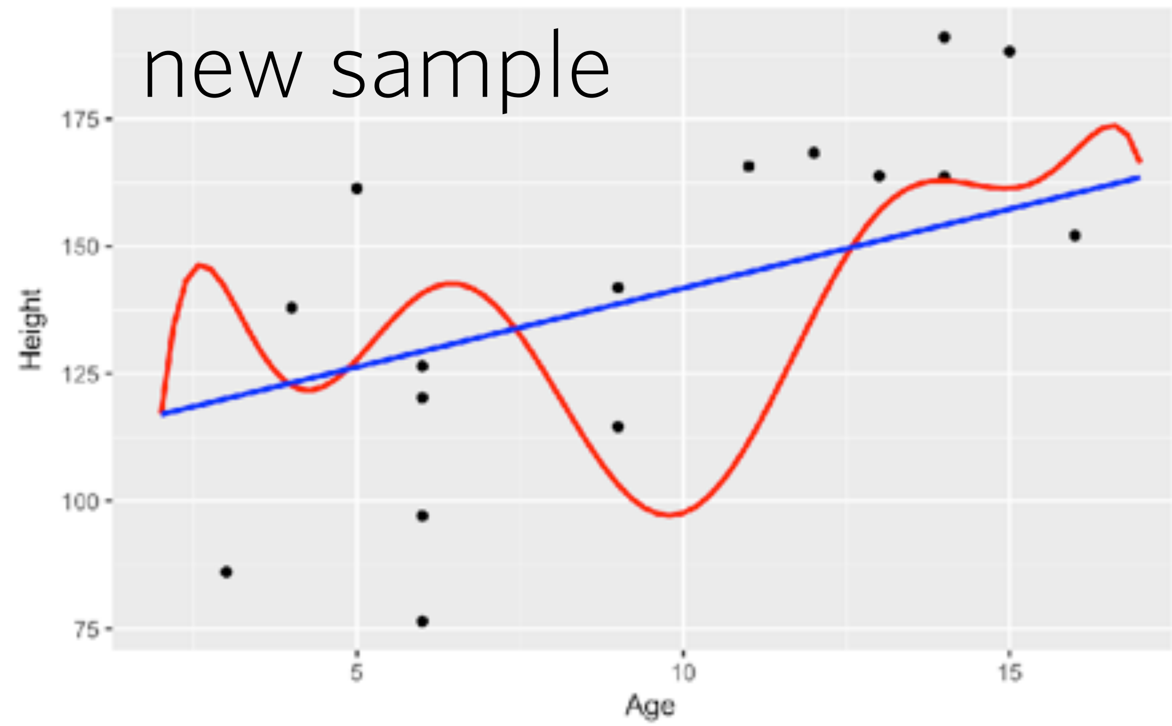


2 *How to know how well a model fits*

# Overfitting

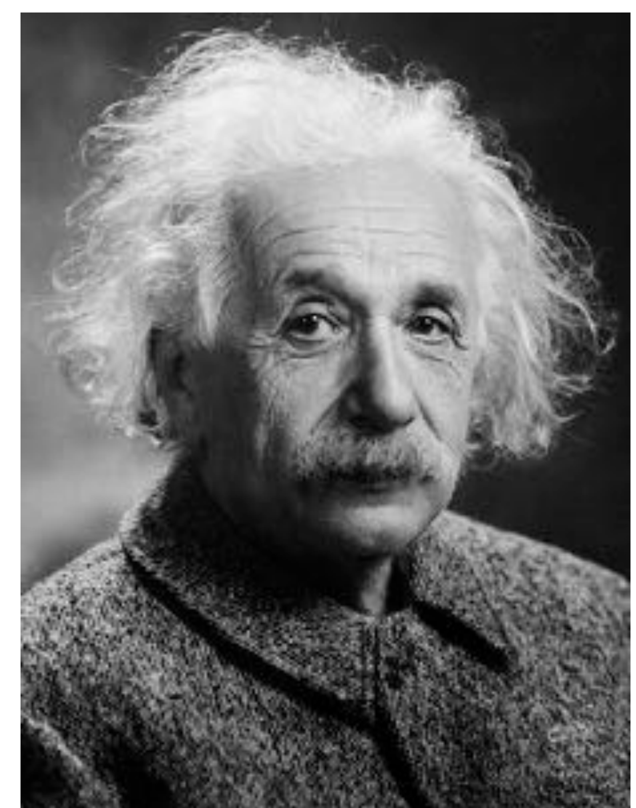
- A more complex model will always fit the data better than a simpler model
  - The model fits the underlying signal as well as the random noise in the data
- But a simpler model often does a better job of explaining a new sample from the same population



“It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.”

*-Albert Einstein*

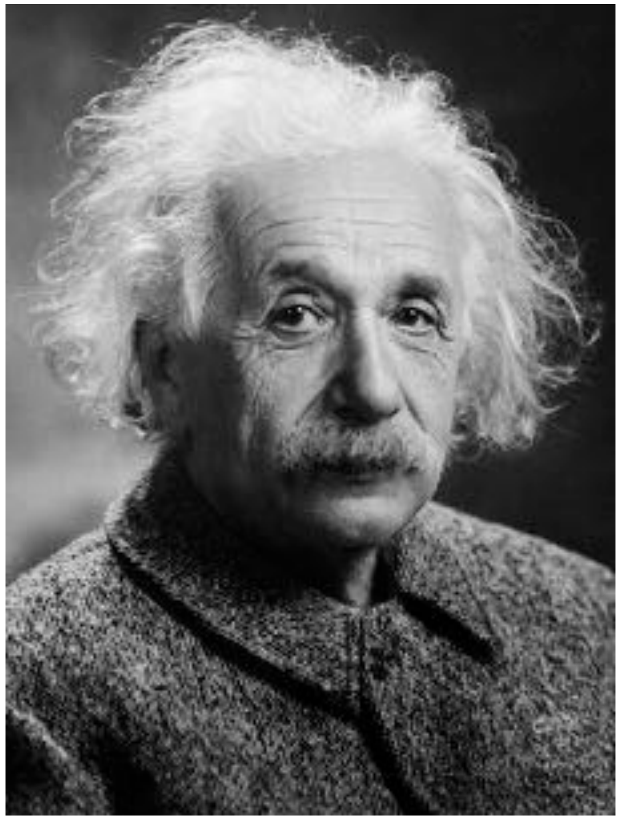


“It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.”

*-Albert Einstein*

*Paraphrased as:*

***“Everything should be as simple as it can be, but not any simpler.”***



# Today

## Mini-Review Session #2

1 → 2 → 3

*Modeling data with the mean*

*Thinking about variability as model error*

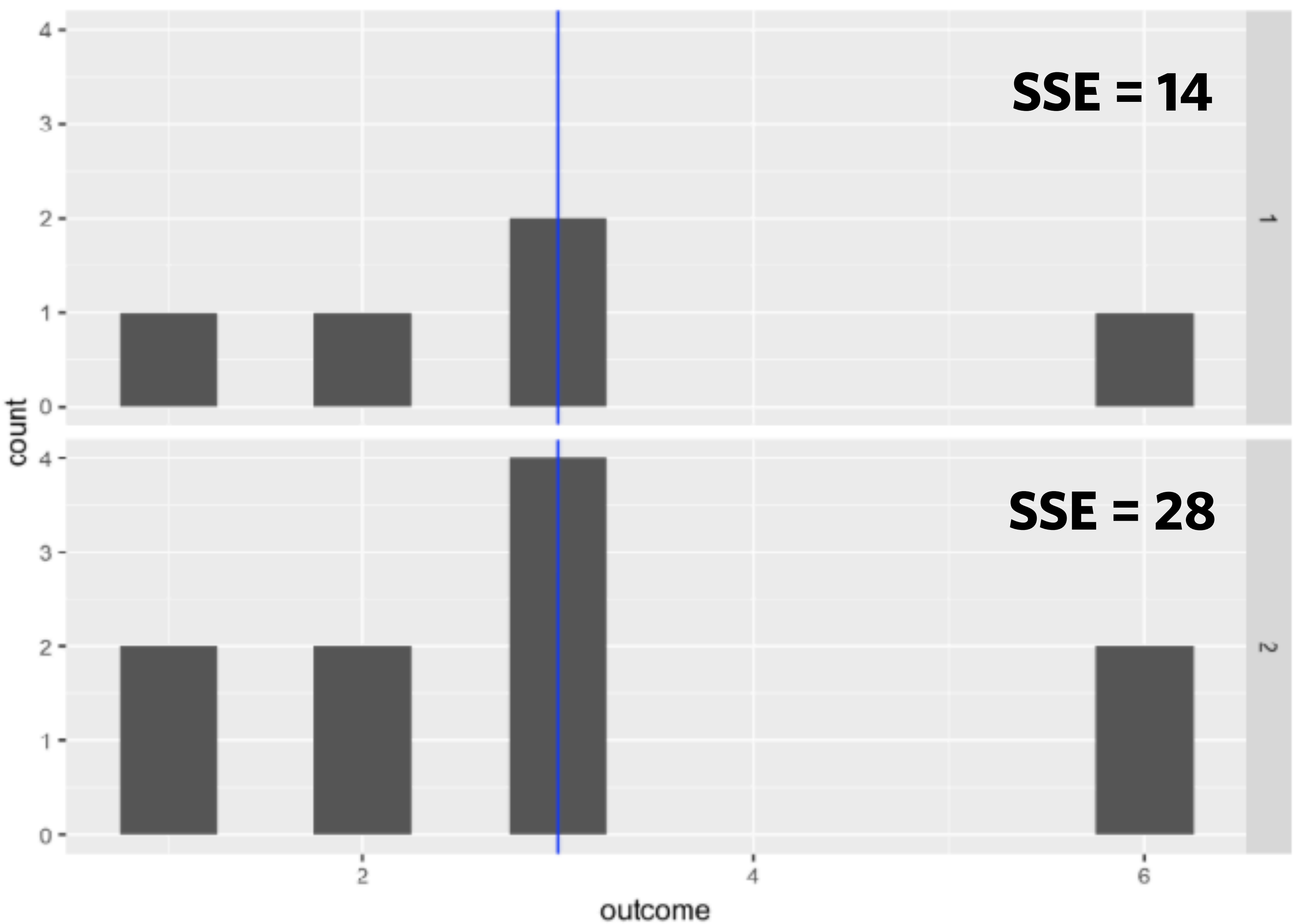
*Estimating variability*

3

## *How do we estimate variability?*

Sum of Squared Error (SSE) is a good measure of total variability if we are using the mean as a model. But, it does have one important disadvantage:

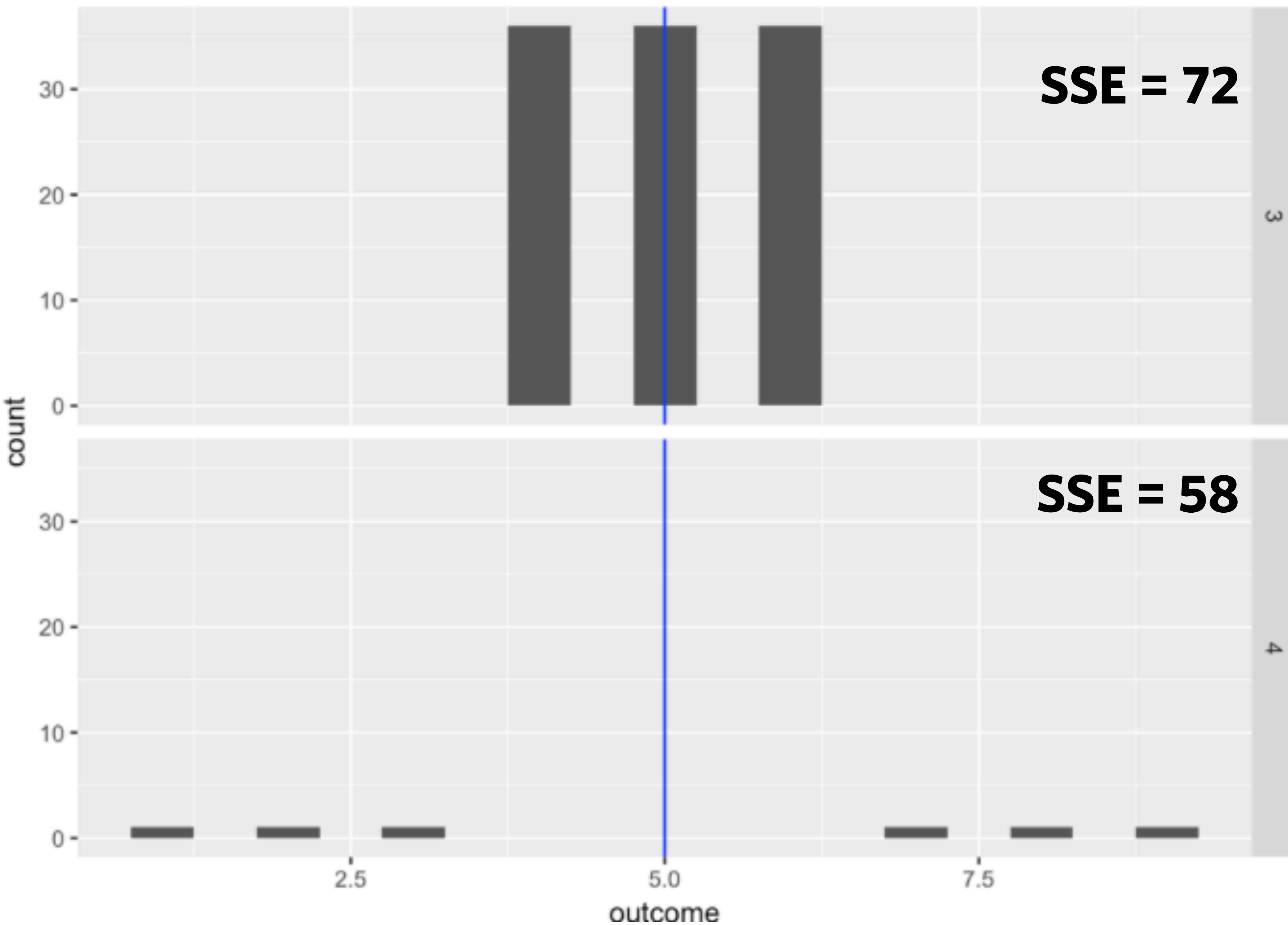
**Which distribution looks more spread out?**



# 3 *How do we estimate variability?*

Sum of Squared Error (SSE) is a good measure of total variability if we are using the mean as a model. But, it does have one important disadvantage:

**Which distribution looks more spread out?**



## 3 *How do we estimate variability?*

**Sum of Squared Error (SSE)** works fine when two distributions have the same sample size (i.e., number of observations).

$$SSE = \sum_{i=1}^{n}(x_i - \hat{x})^2$$

But SSE is hard to interpret if sample sizes are different.
This is b/c SSE always increases as sample size increases, even if the distribution isn't getting "more spread out."

Meet the **sample variance** (kind of like "SSE per data point"):

$$\text{sample } variance = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

3

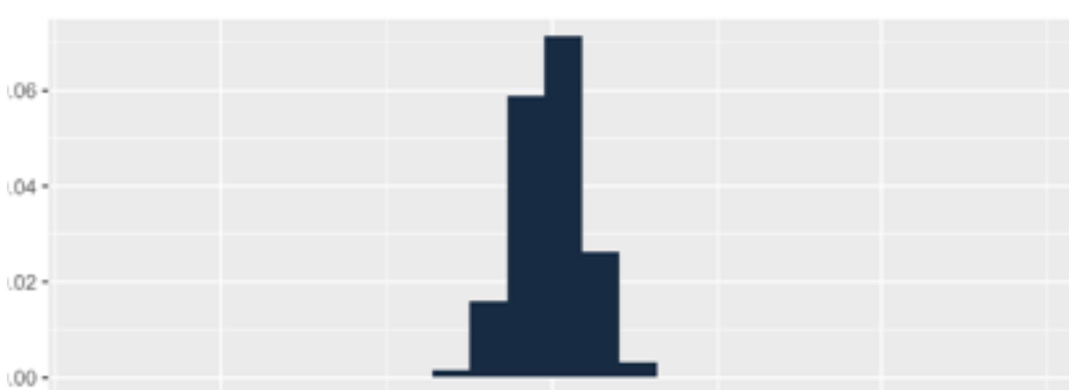
# How do we estimate variability?

Meet the **sample variance** (kind of like "SSE per data point"):

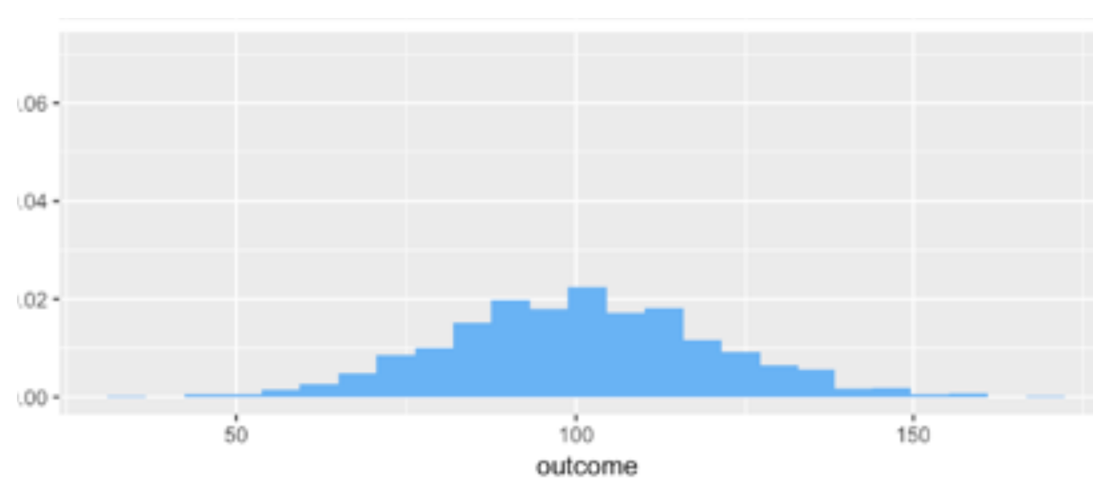
$$\text{sample}\ variance = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

**Variance** is a single number that summarizes how spread out a distribution is.

**low variance**




**high variance**



# 3 How do we estimate variability?

**Variance** is a single number that summarizes how spread out a distribution is.

Notice the symbols!

**sample variance** ("$s^2$")

**population variance**

$$\text{sample } variance = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

lowercase sigma

Notice that the denominators are different!

We divide by **n-1** to get unbiased estimate of population variance from our sample. This is because there are **n-1** degrees of freedom when computing sample variance: once we compute the mean, there are only **n-1** degrees of freedom.

3

## *How do we estimate variability?*

Variance is a single number that summarizes how spread out a distribution is.

**sample variance**

$$\text{sample } variance = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

**population variance**

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

Meet the **standard deviation**

$$SD = \sqrt{variance}$$

square root of the variance

in the same units as
the underlying measurement

often abbreviated s.d.

built-in R function is: "sd"

3

# How do we estimate variability?

02:00

$$variance = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

$$SD = \sqrt{variance}$$

| x | error | error^2 |
|---|---|---|
| 3 | -3 | 9 |
| 5 | -1 | 1 |
| 6 | 0 | 0 |
| 7 | 1 | 1 |
| 9 | 3 | 9 |

Calculate the sample variance of x:

Calculate the sample s.d. of x:

3

## How do we estimate variability?

$$variance = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

$$SD = \sqrt{variance}$$

| x | error | error^2 |
|---|---|---|
| 3 | -3 | 9 |
| 5 | -1 | 1 |
| 6 | 0 | 0 |
| 7 | 1 | 1 |
| 9 | 3 | 9 |

Calculate the sample variance of x:

SSE= 20

variance ($s^2$)=20/4=5

Calculate the sample s.d. of x:

SD=sqrt(5)=2.24

# Today

## Mini-Review Session #2

1. *Modeling data with the mean*
2. *Thinking about variability as model error*
3. *Estimating variability*

**Please complete the daily feedback survey before leaving class!**

Go to: https://psyc60.github.io/syllabus

Before leaving class, please complete daily feedback survey!

## Student Daily Feedback Survey

...se complete the linked daily feedback survey. The purpose of this ... better understand how things are going for you in this class, and ... reflect on what you have been learning.

## Feedback

We welcome student fe... d your TA a Slack message, or ... nline form.

## Acknowledgements

Many thanks to Prof. Ji Son, Prof. James Stigler, everyone in the UCLA Teaching and Learning Lab, Prof. Russ Poldrack and Prof. Tobias Gerstenberg for generously sharing their instructional materials.

# PSYC 60: How was class today?

Hi there!

I would love to know about your experience in today's class. Could you please take 2 minutes to answer the following few questions? It will be hugely useful for helping me know what is working well, what isn't, and how to keep improving this class.

Best,
Prof. Fan

jefan@ucsd.edu Switch account

Your email will be recorded when you submit this form

* Required

How are you finding the pace of this class so far? *

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Much too slow | ○ | ○ | ○ | ○ | ○ | ○ | ○ | Much too fast |

Do you feel like you are learning new things? *

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Not learning anything new | ○ | ○ | ○ | ○ | ○ | ○ | ○ | Learning lots of new things |