# Session 10: Fitting models: Central tendency and dispersion

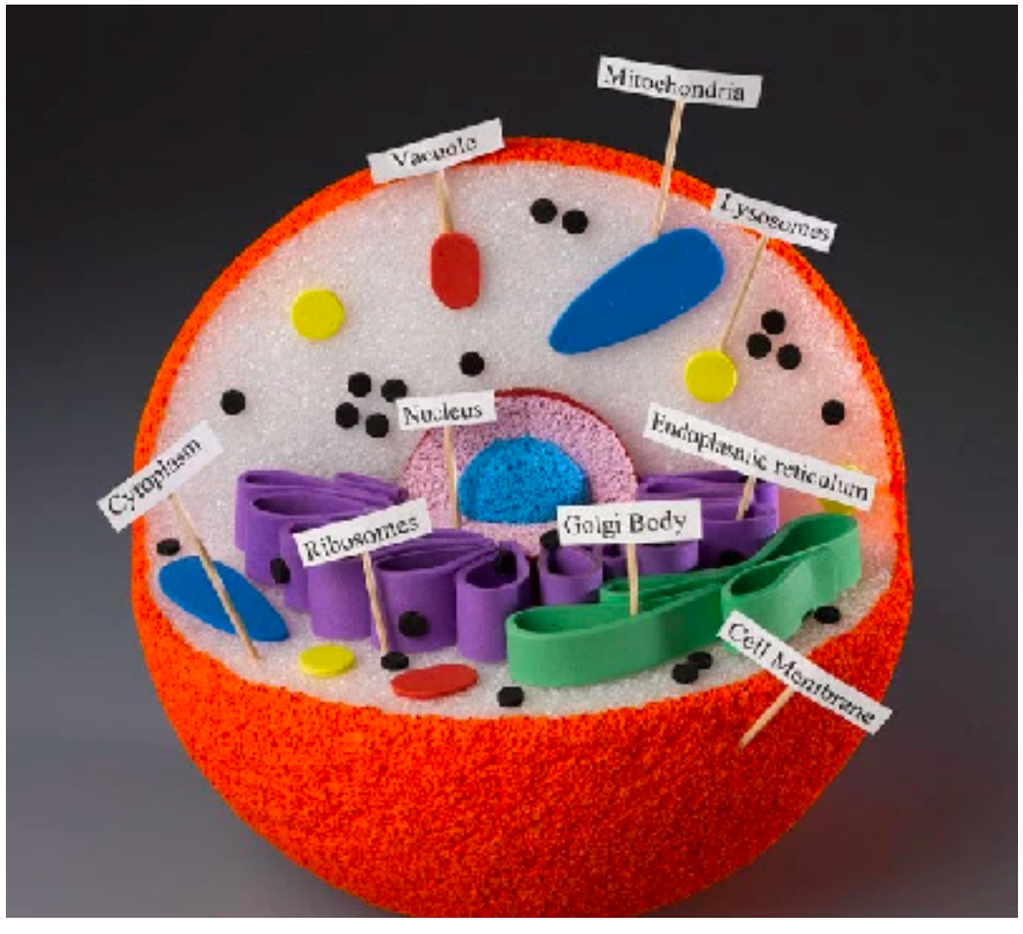
Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

# This time

- Building models to describe data
- Central tendency
- Dispersion and variability

# What is a “model”?

# Models simplify the world for us




# The basic statistical model

outcome = model + error

what we actually observe (the data)

what we expect to observe (our prediction)

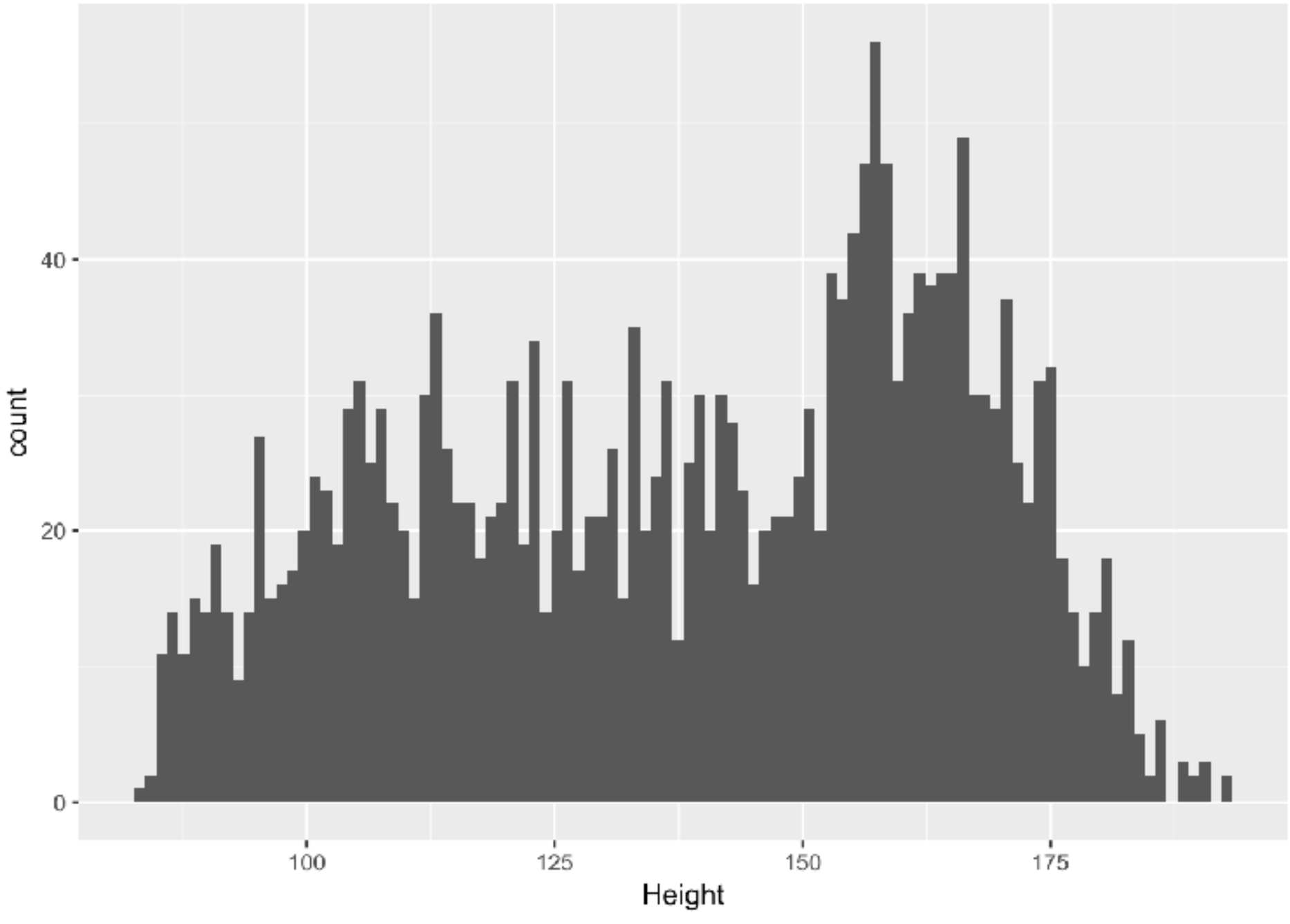
difference between expected and observed

The model is should be much simpler than the thing it is modeling!

# A simple example

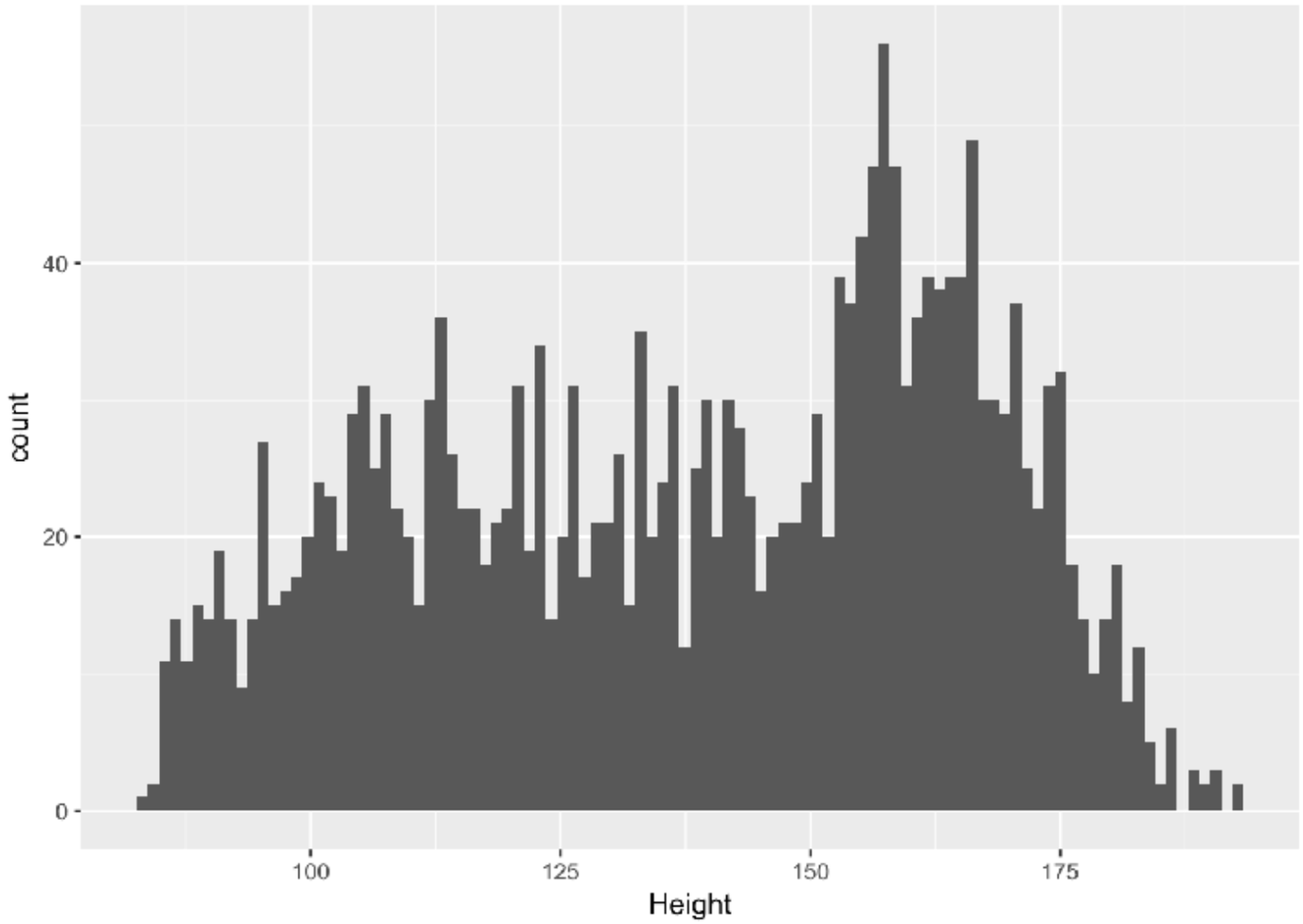
- What is the height of children in the NHANES sample?

```r
NHANES <- NHANES %>%
  mutate(isChild = Age<18)

NHANES_child <- NHANES %>%
  subset(subset=isChild & Height!='NA')

ggplot(data=NHANES_child,aes(Height)) +
  geom_histogram(bins=100)
```



# What is the simplest model we can image?

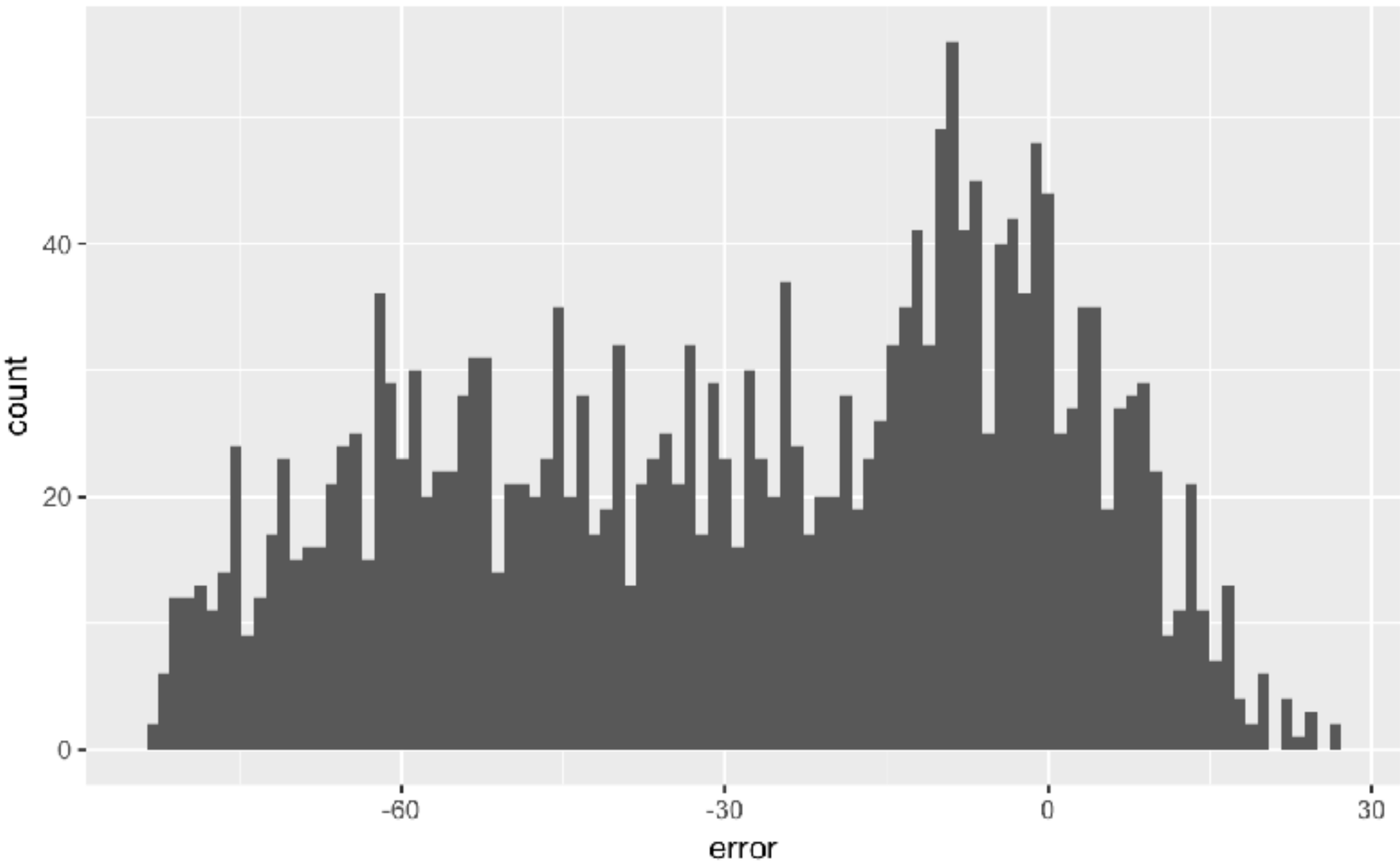
- One guess: what about the most common value in the dataset (the *mode*)?
  - height(i) = 166.5 + error(i)
  - Summarizes 2,223 data points in terms of a single number
- How well does that describe the data?
- Computing the error:
  - error = outcome - model



```
error <- NHANES_child$Height - 166.5
```

```
ggplot(NULL,aes(error)) +
  geom_histogram(bins=100)
```

average error: -27.94 inches



# A better model?

- We would like for our model to have zero error, on average
- If we use the mean of the data as our model, then that will be the case

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

$$error = \sum_{i=1}^{n} (x_i - \bar{X}) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{X} = 0$$

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

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

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i$$

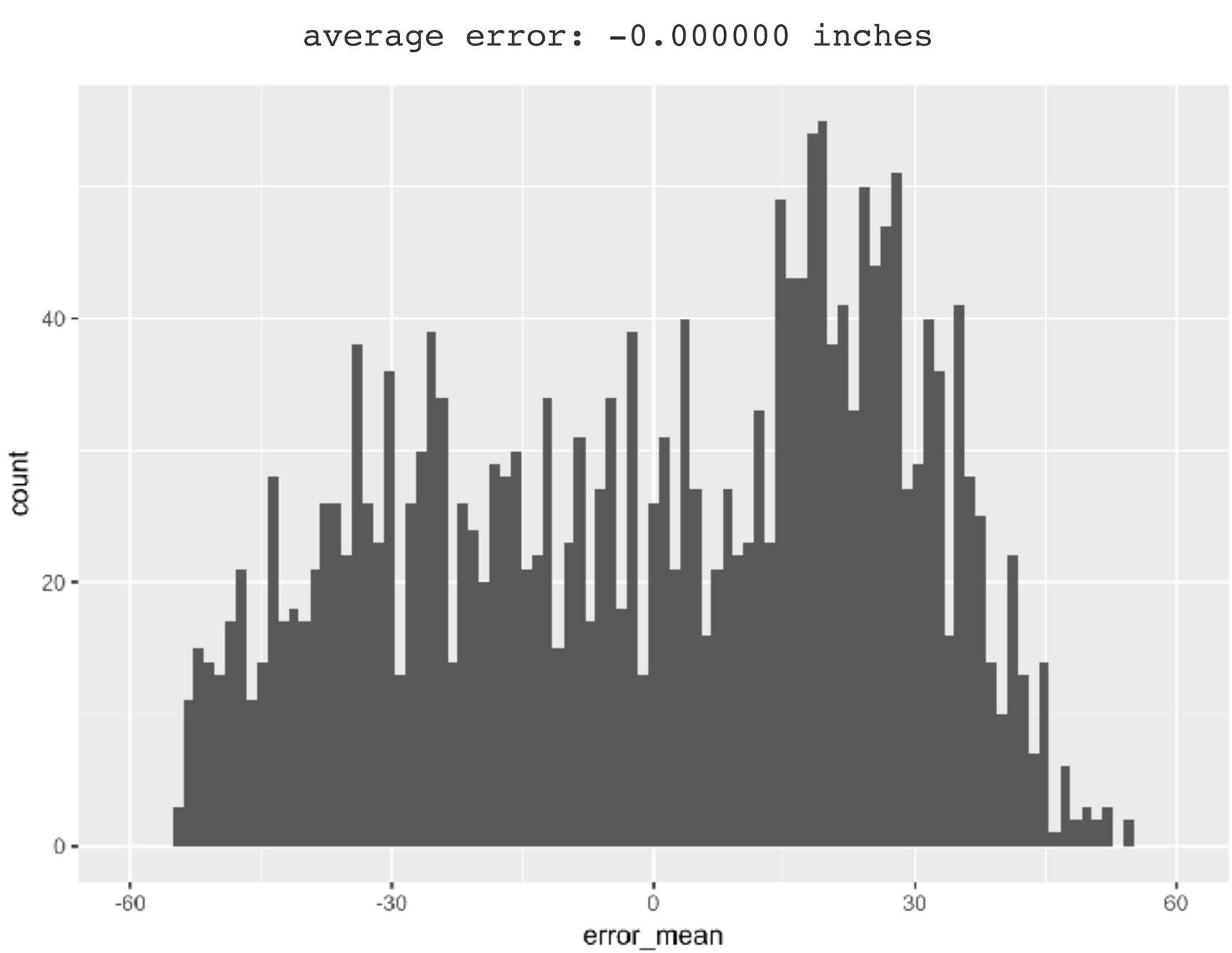
# Sum of errors from the mean is zero

```
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

```r
error_mean <- NHANES_child$Height - mean(NHANES_child$Height)

ggplot(NULL,aes(error_mean)) +
  geom_histogram(bins=100) + xlim(-60,60)
```



# Building an even better model

- The average error for mean is zero
  - But there are still errors, sometimes positive and sometimes negative
- The “best” estimate is one that minimizes errors overall (both positive and negative)
- We can quantify the total error by squaring the errors and adding them up

$$sum\,of\,squared\,errors = \sum_{i=1}^{n} (x_i - \hat{x})^2$$

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

We take the mean of the squared errors by dividing SSE by the number of values, and then take the square root:

$$mean\,squared\,error = \frac{SSE}{N}$$

```
print(paste('average squared error:',mean(error_mean**2)))
```

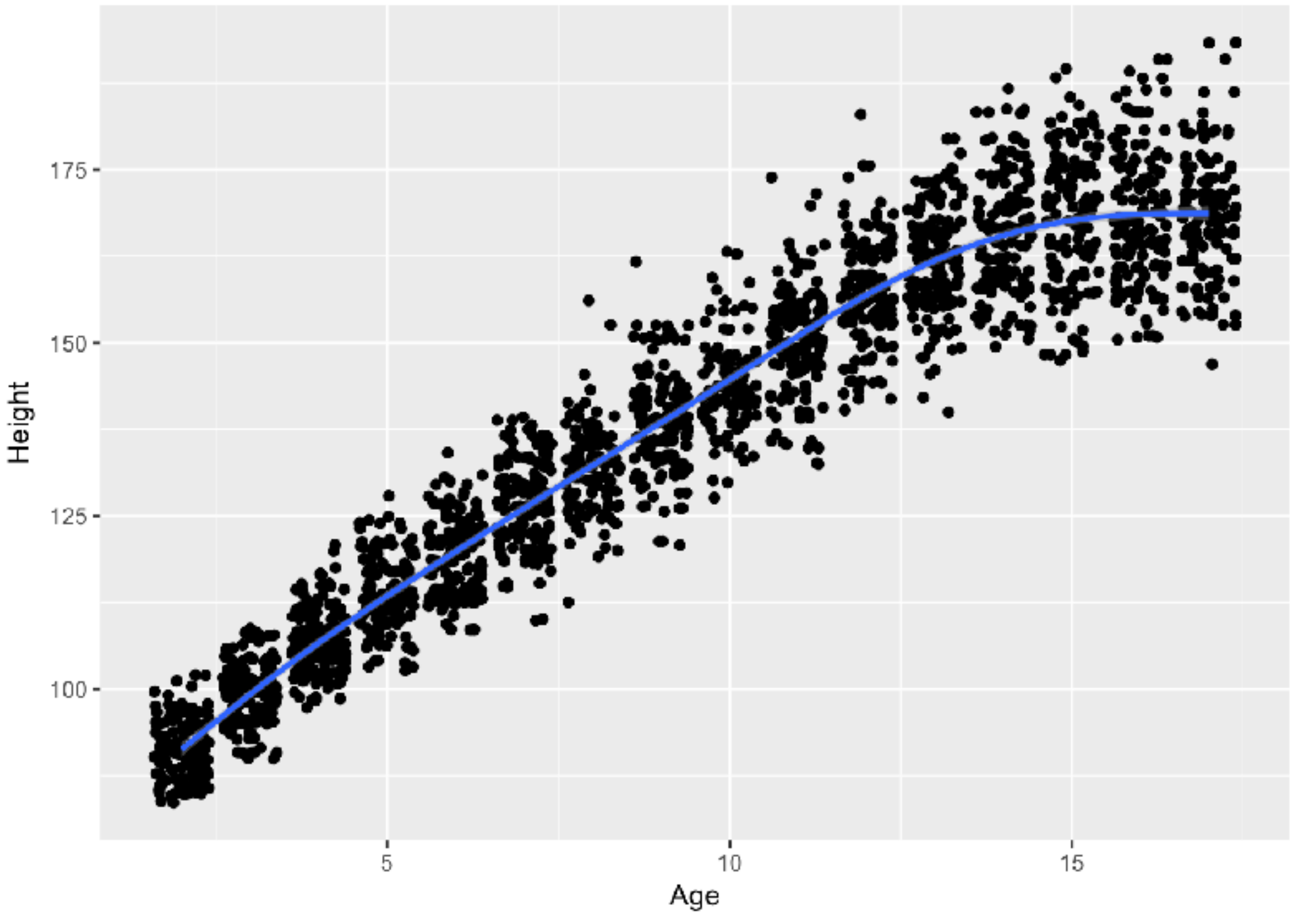
mean squared error: 720.05

This tells us that while on average we make no error, for any individual we could actually make quite a big error (~27 $inches^2$ on average).

Could we make the model any better? What else do we know about these individuals that might help us better estimate their height?

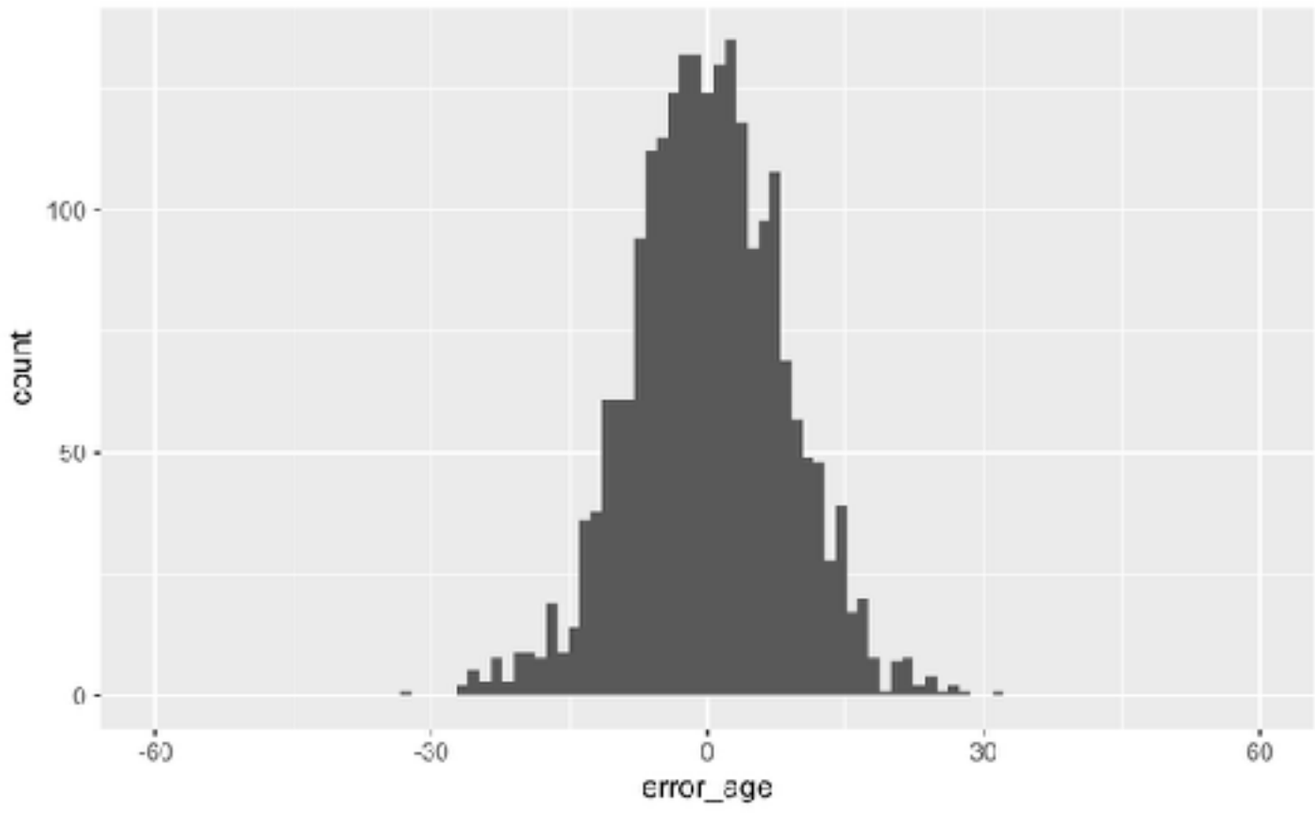
What about their age? Let's plot height versus age and see how they are related.

```
ggplot(NHANES_child,aes(x=Age,y=Height)) +
  geom_point(position='jitter') +
  geom_smooth()
```



```r
# find the best fitting model to predict height given age
model_age <- lm(Height ~ Age, data = NHANES_child)

# the predict() function uses the fitted model to predict values
for each person
predicted_age <- predict(model_age)

error_age <- NHANES_child$Height - predicted_age
sprintf('average squared error: %f inches',mean(error_age**2))
```

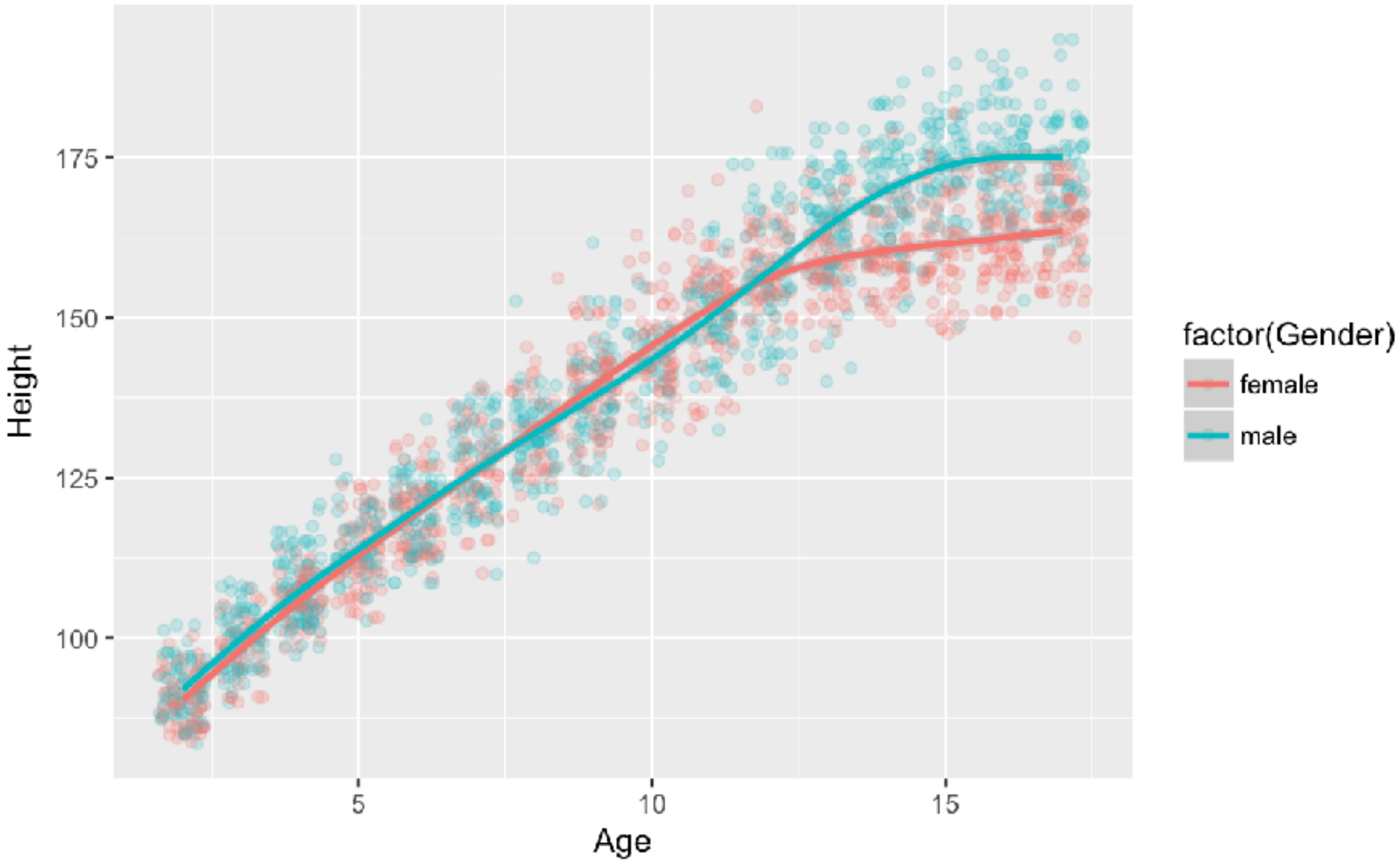
mean squared error: 69.61 inches



# What else do we know?

- What other variables might be related to height?

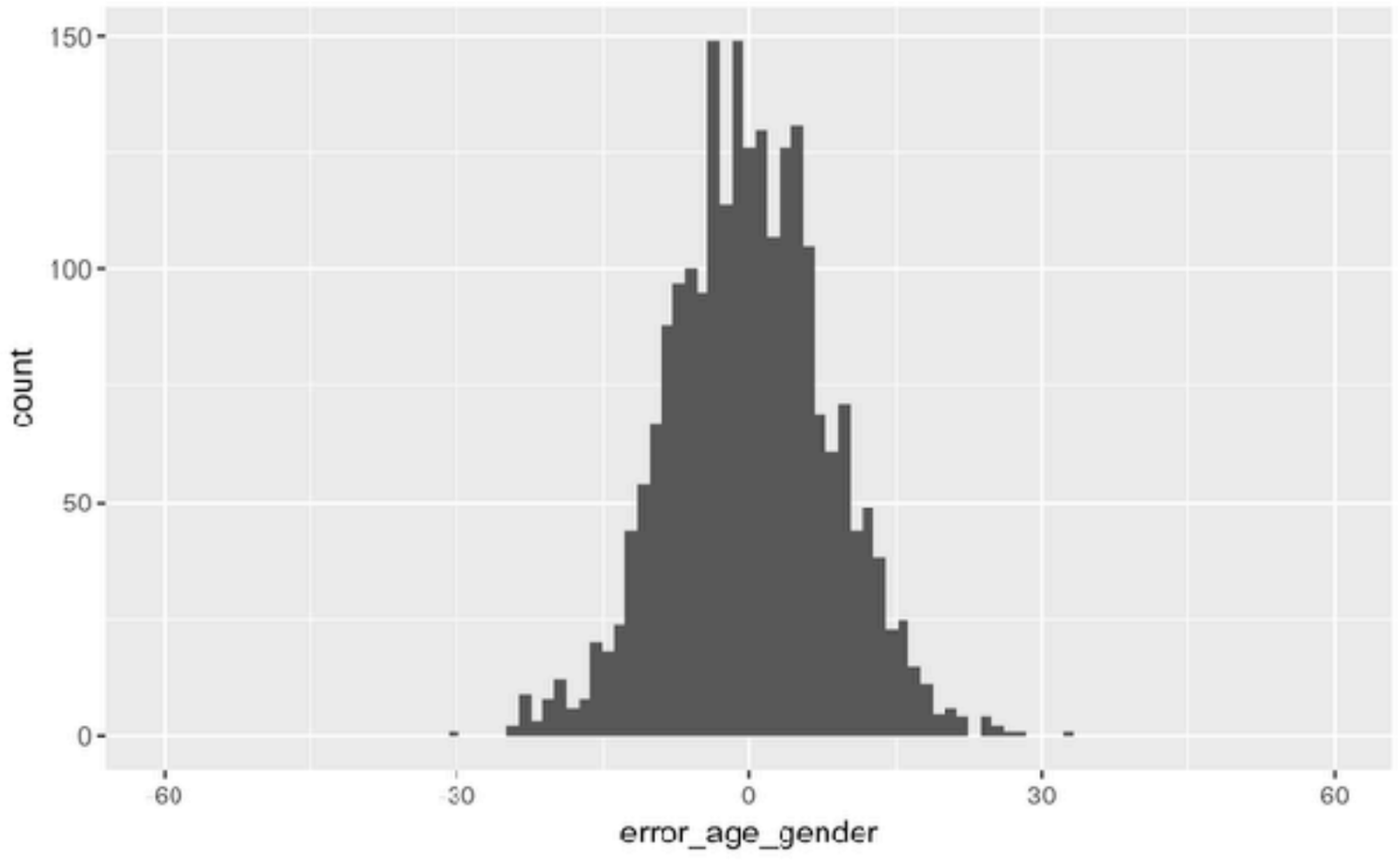
```
ggplot(NHANES_child,aes(x=Age,y=Height)) +
  geom_point(aes(colour = factor(Gender)),position = "jitter",alpha=0.2) +
  geom_smooth(aes(group=factor(Gender),colour = factor(Gender)))
```



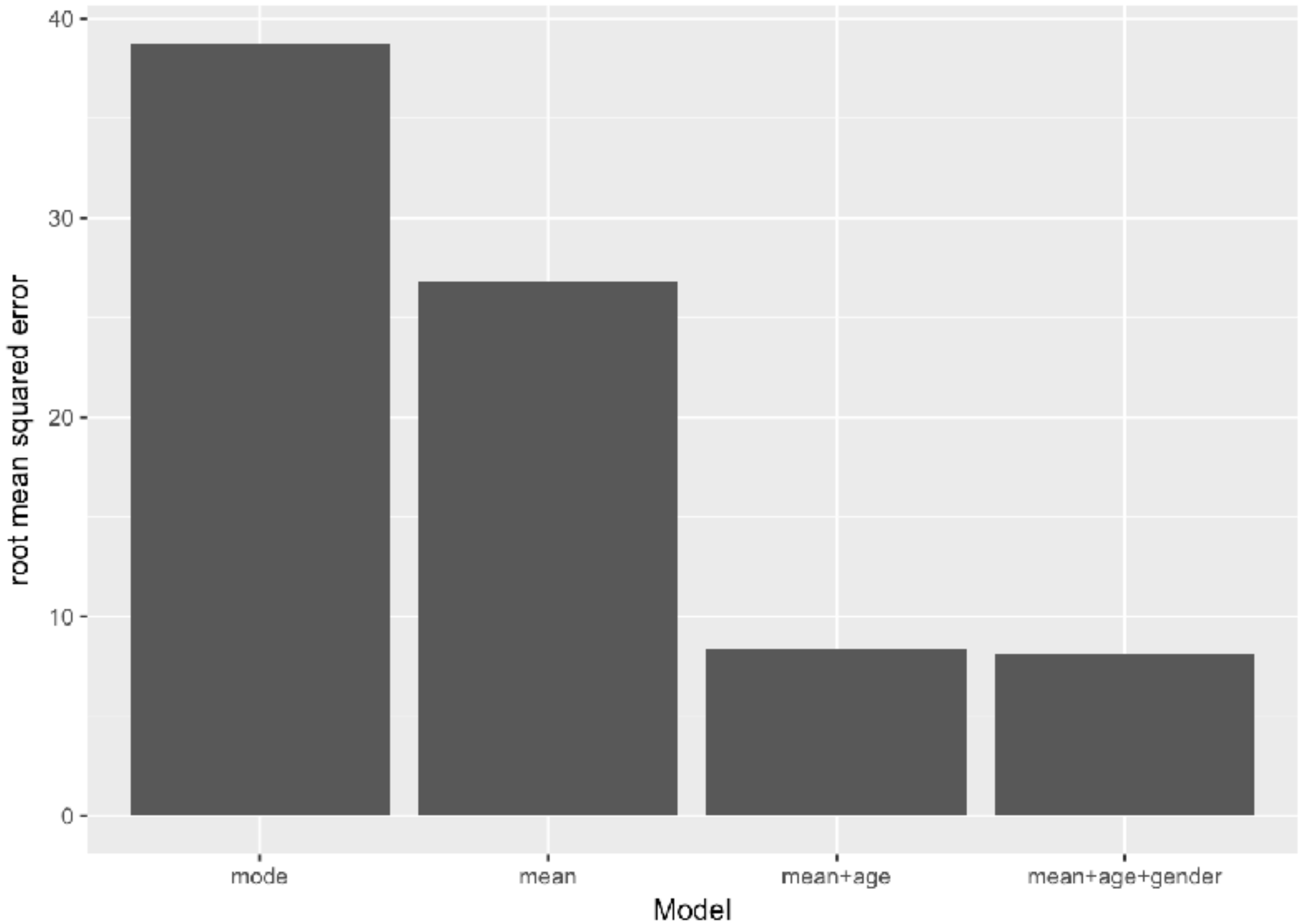
```
model_age_gender <- lm(Height ~ Age + Gender,
                       data=NHANES_child)
predicted_age_gender <- predict(model_age_gender)
error_age_gender <- NHANES_child$Height - predicted_age_gender
```

mean squared error: 66.42 inches

model: height = 84.33 + 5.47*Age + 3.57*Gender



```
error_df <- data.frame(error=c(mean(error**2),mean(error_mean**2),
                               mean(error_age**2),mean(error_age_gender**2)))
row.names(error_df) <- c('mode','mean','age','age+gender')
error_df$RMSE <- sqrt(error_df$error)
ggplot(error_df,aes(x=row.names(error_df),y=RMSE)) +
  geom_col() +ylab('root mean squared error') + xlab('Model') +
  scale_x_discrete(limits = c('mode','mean','age','age+gender'))
```



# What statistical models do you think have the biggest impact on your life?

**Top**

# What makes a model “good”?

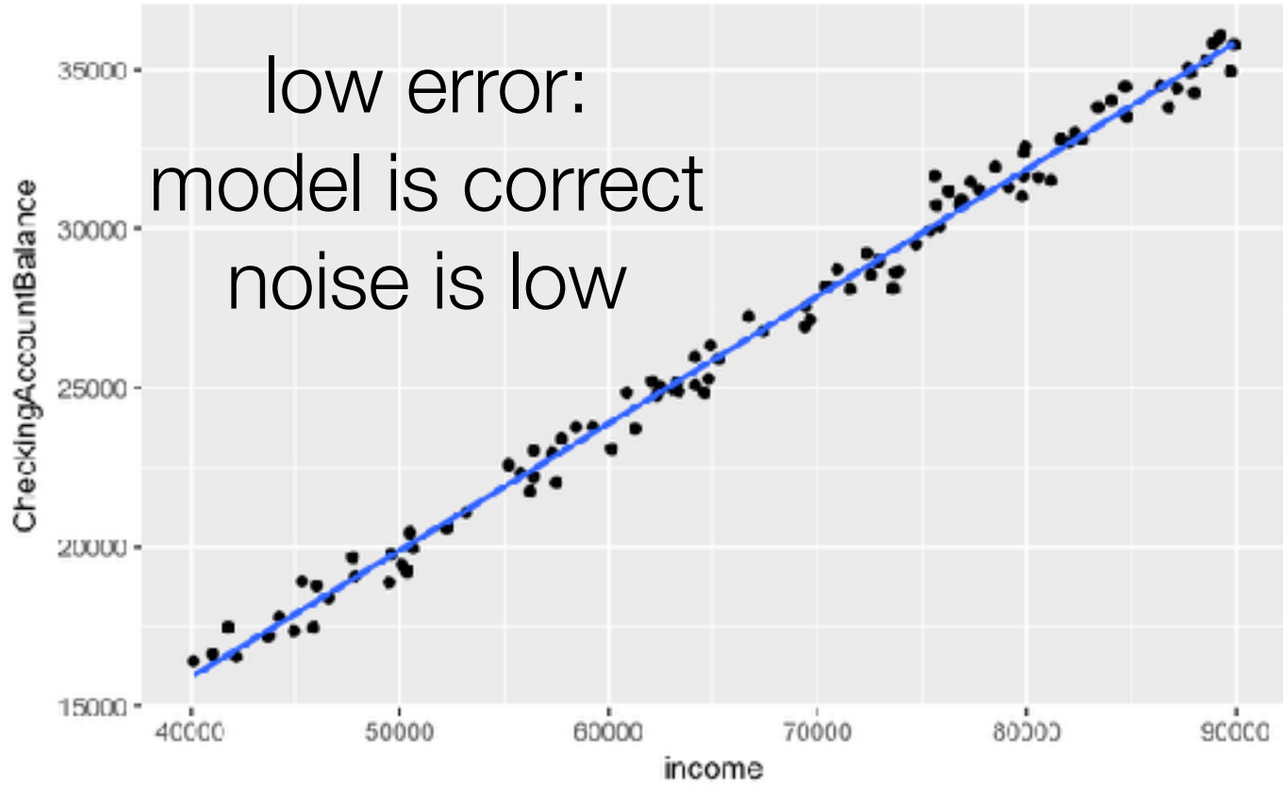
- Describes our dataset well
  - the error for the fitted data is low
- Generalizes to other data
  - the error for a new dataset is low
- These two are often in conflict!

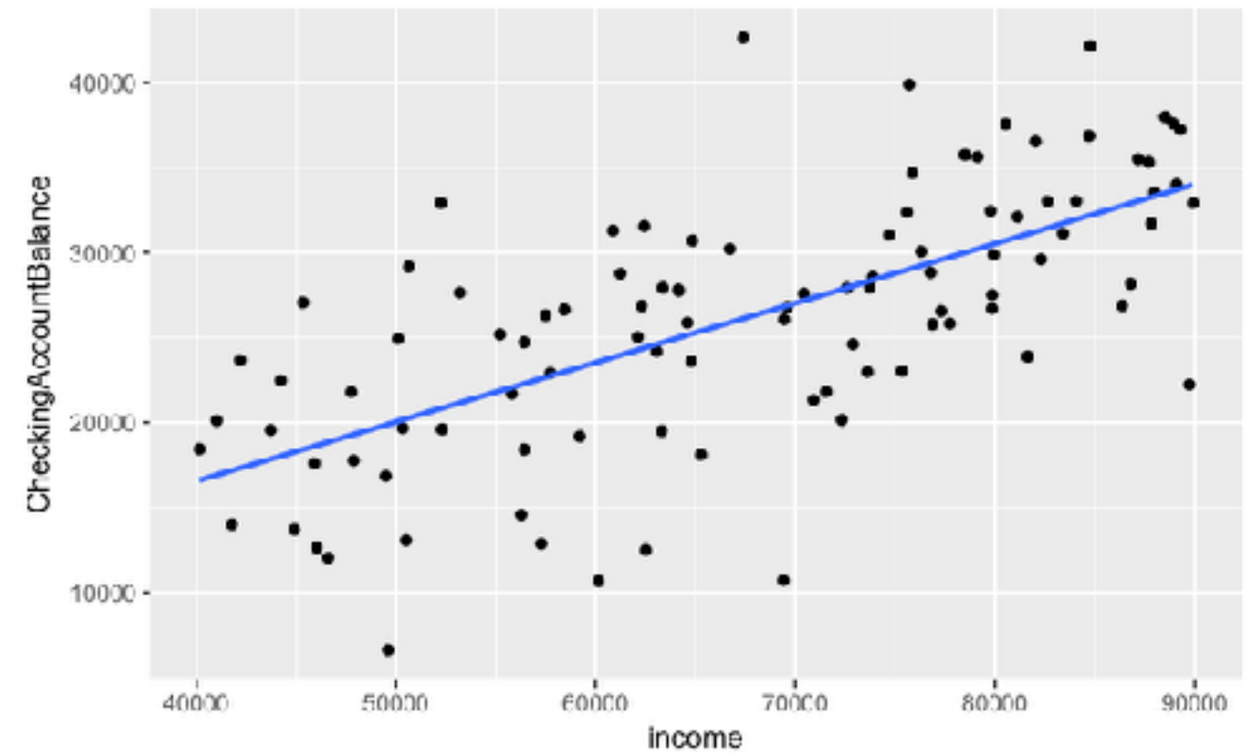
# Sources of error:

- Remember the basic model:
  - outcome = model + error
- Error can come from two sources:
  - The model is incorrect
  - The measurements have random error (“noise”)

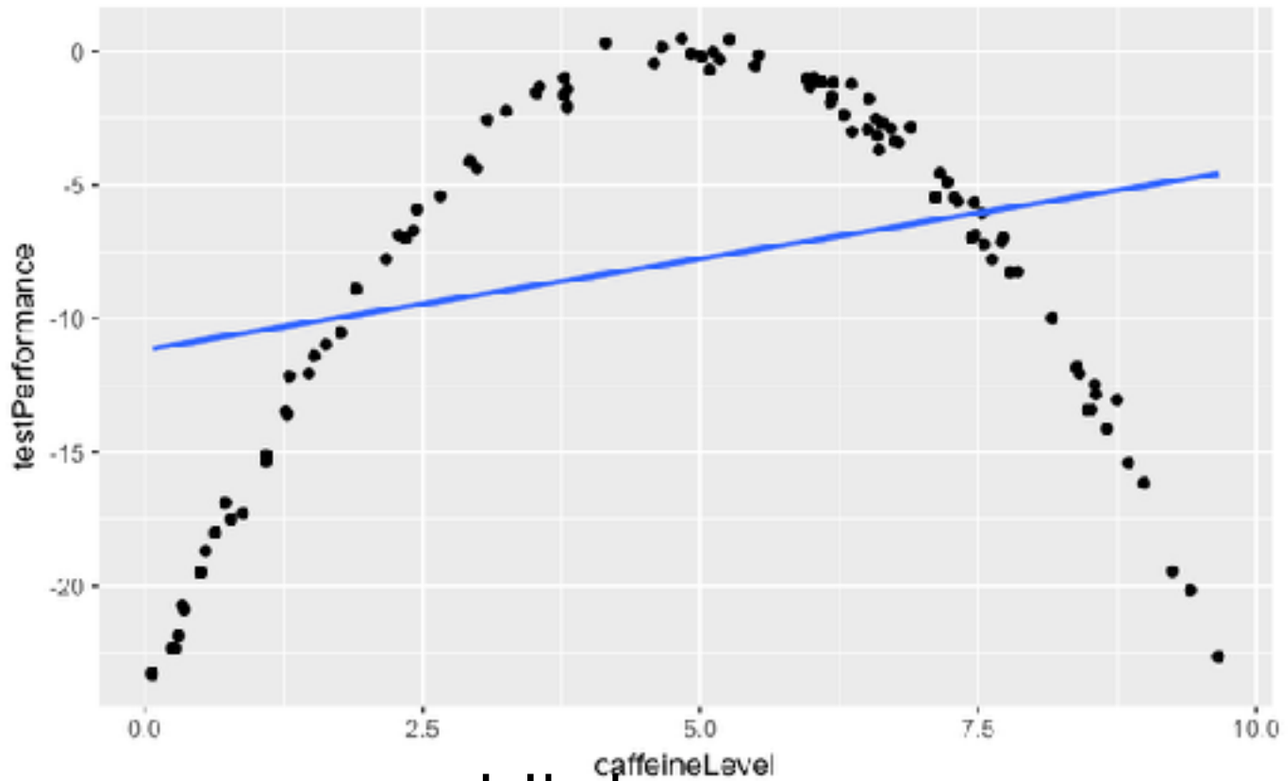
Error can come from two sources:

- incorrect model
- noisy data




High error:
model is correct
noise is high

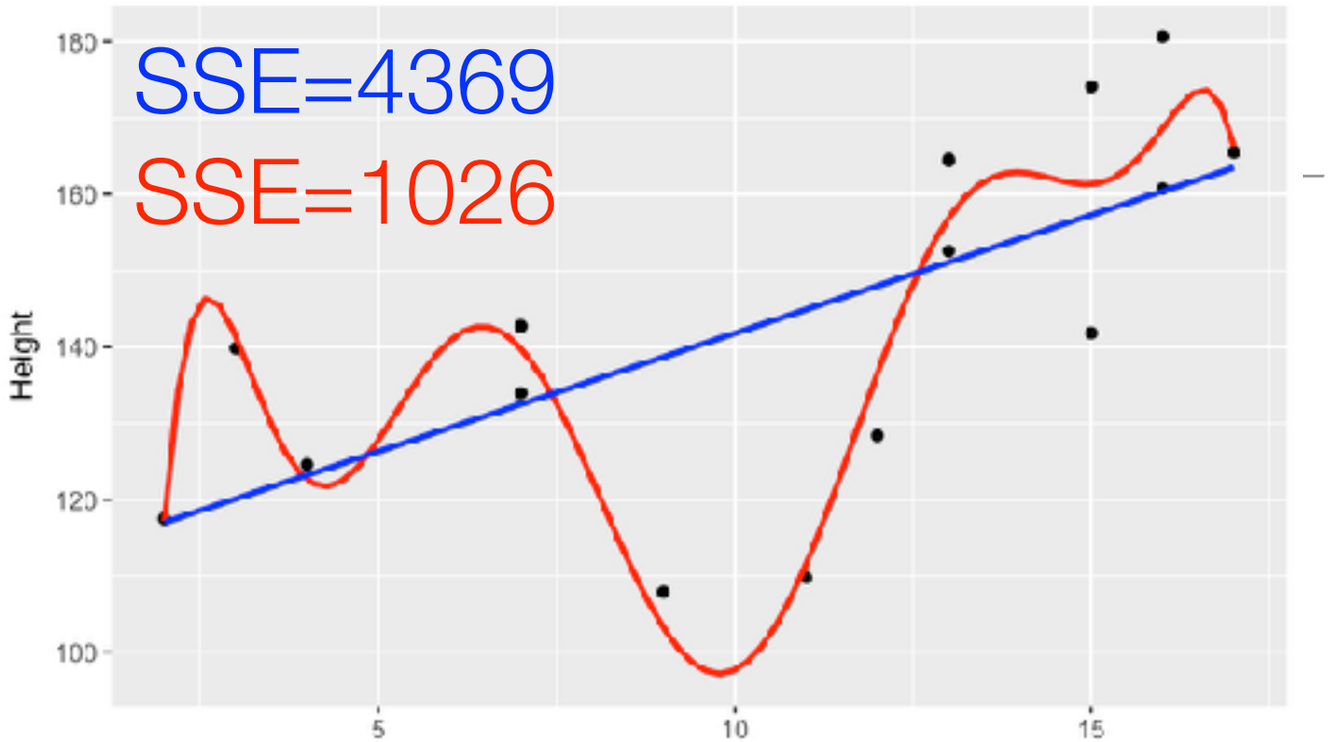



High error:
model is wrong
noise is low

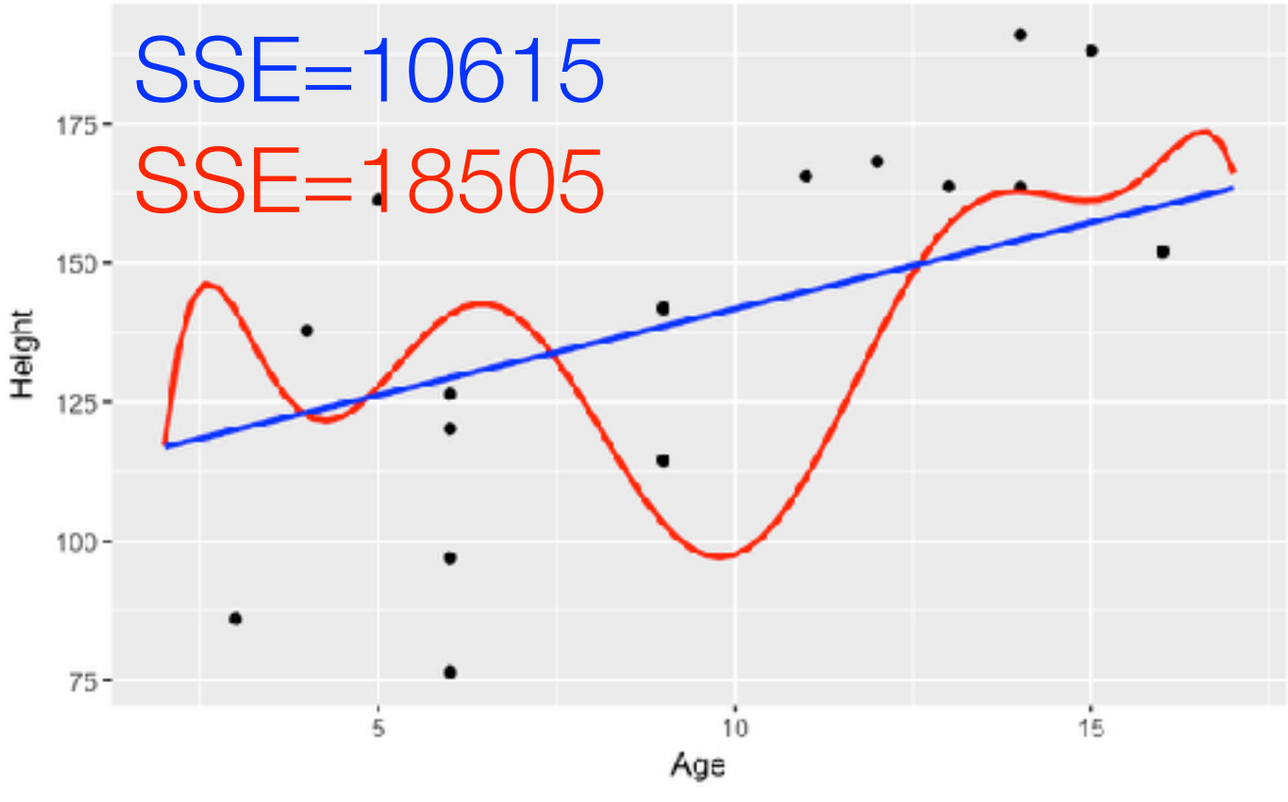
# Overfitting

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

Original sample

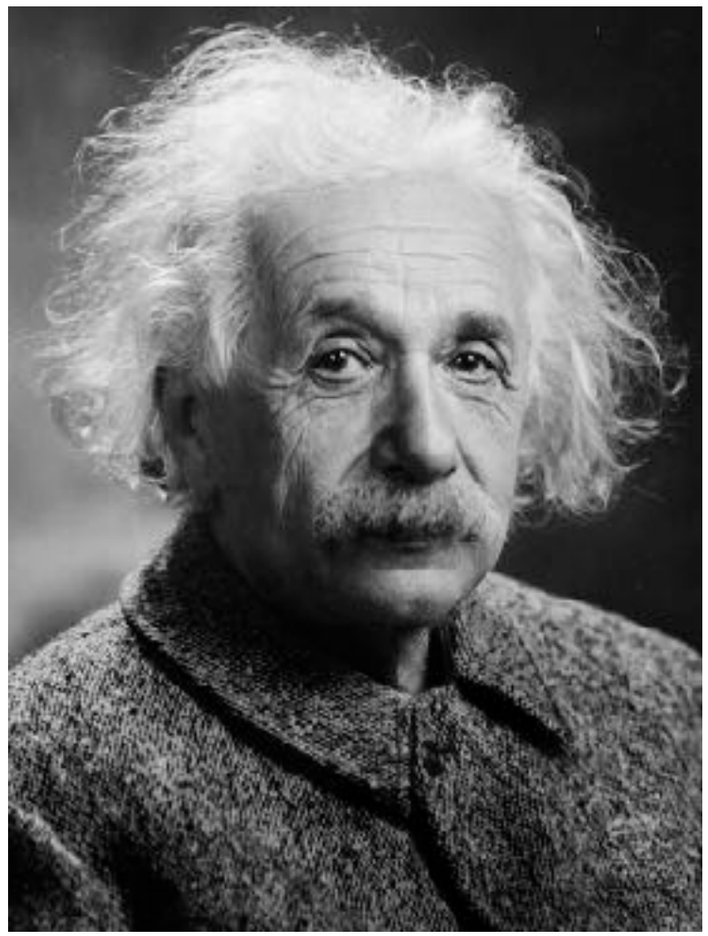



New sample



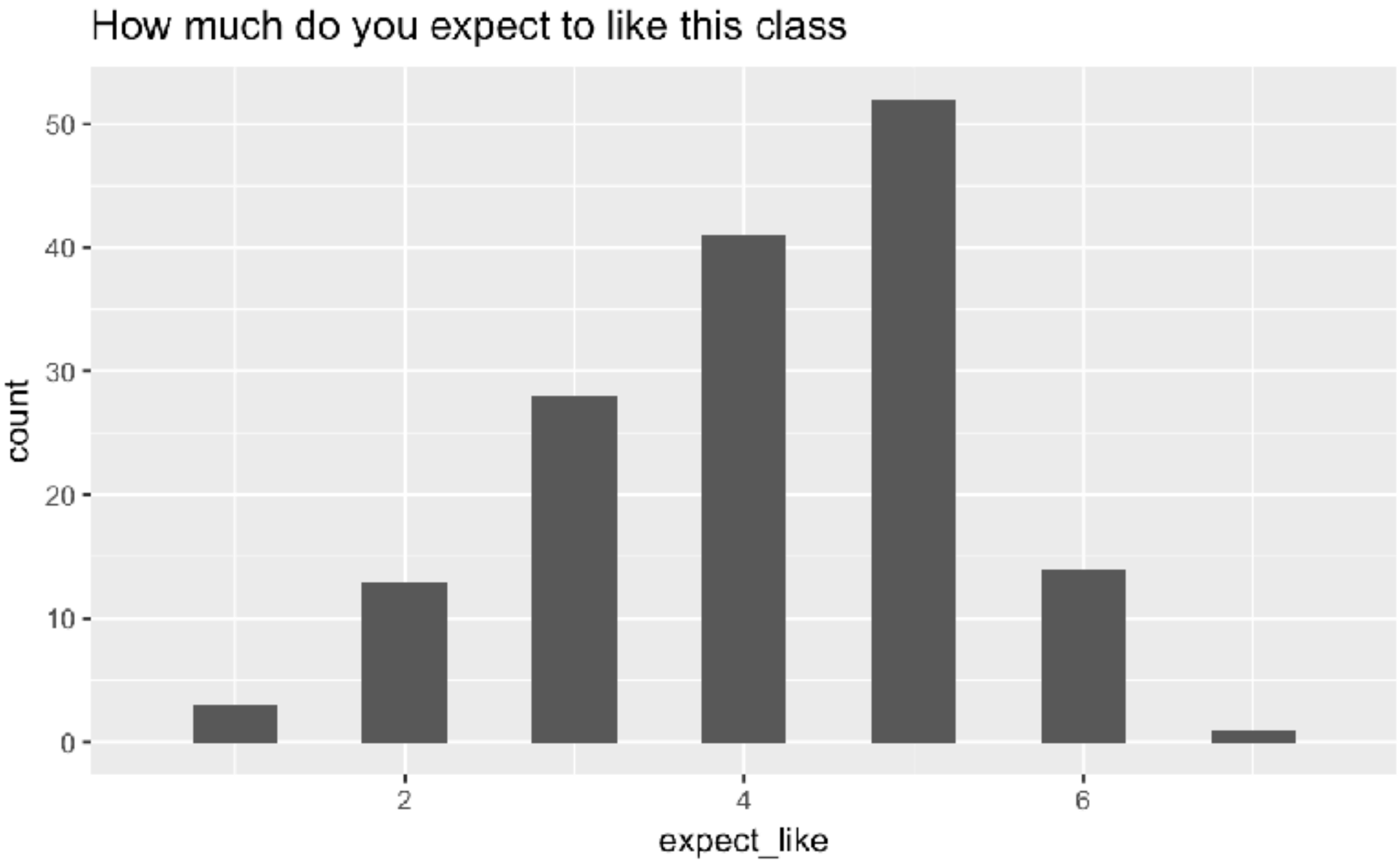
# The principle of parsimony

- "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, 1933
- Paraphrased as "everything should be as simple as it can be, but not simpler"



# The simplest model: Central tendency

- What is the most typical value?



# Mean (aka average)

sample mean

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

population mean

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

same formula, different symbols

# The mean as a balancing point

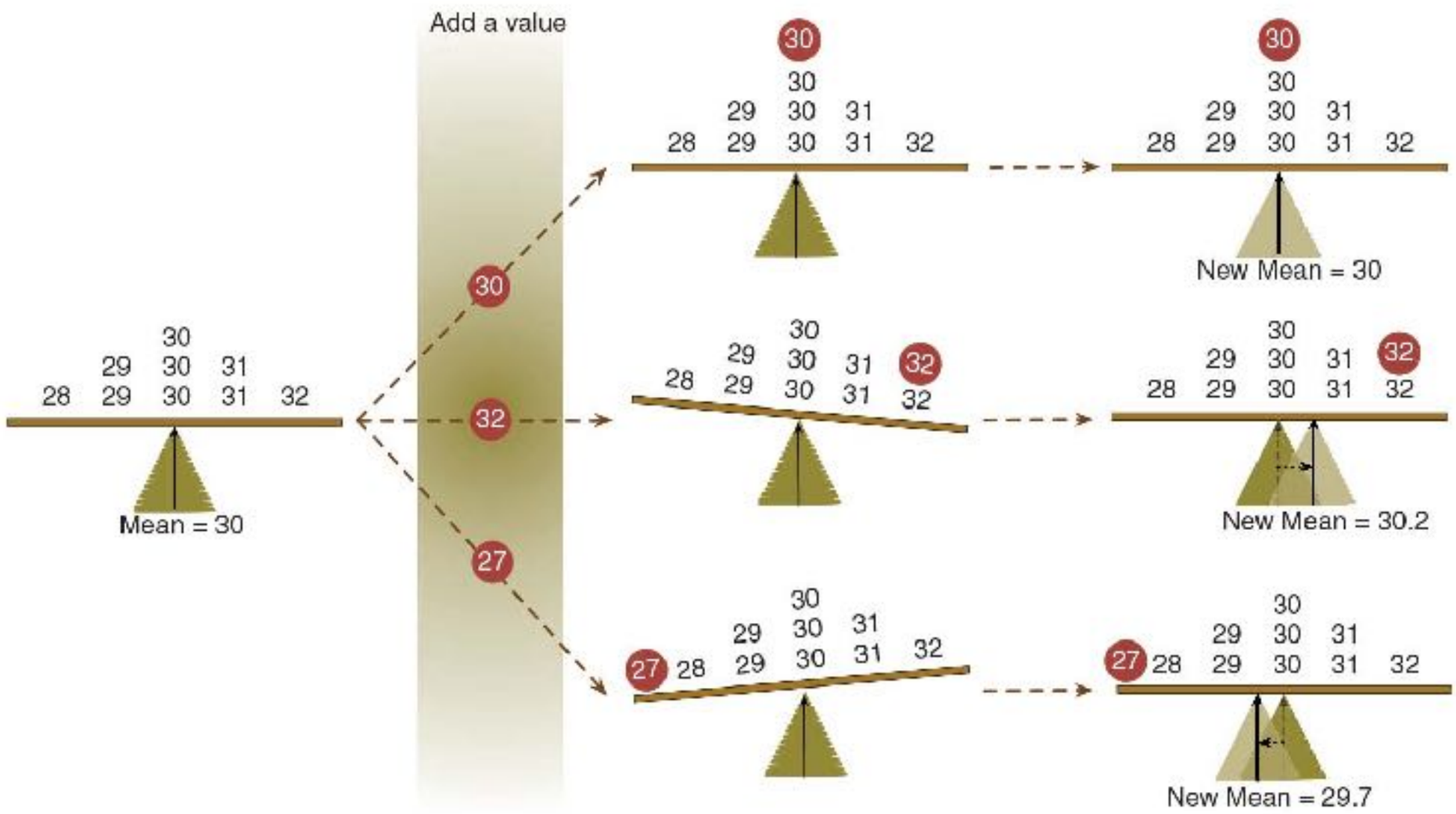



*Figure 4.7* *The mean is the balancing point of a distribution*

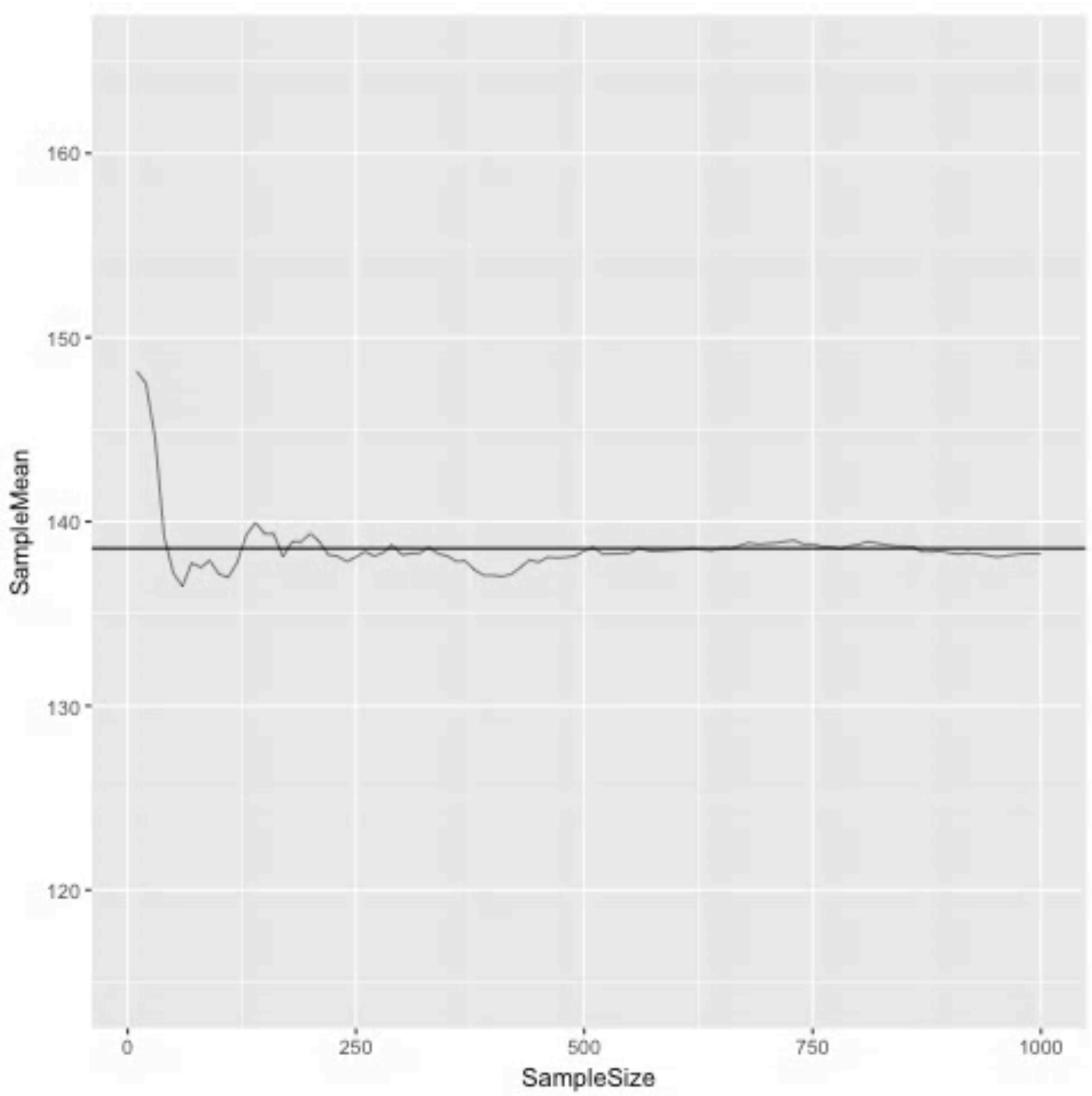
# The mean is the “best” estimate

- The mean is the value that minimizes the sum of squared errors
  - This is the statistical definition of being the “best” estimate
  - We proved this earlier
    - But we can also demonstrate it using R, which you will do in your next problem set…

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

# Estimating the mean accurately can require lots of data

- Data: Height of children from NHANES (2,223 children)
- Mean height: 138.5 in
- What happens if we take smaller samples from this group?
  - start with a sample of size 10 and then increase by 10 up to 1000



# 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 |

mean income: $61,600

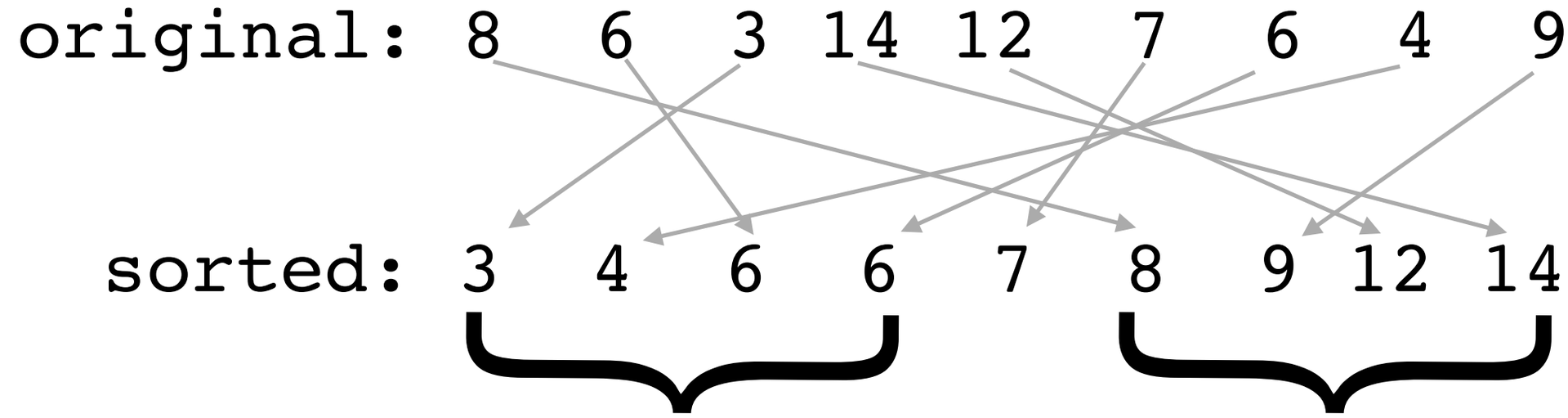
mean income: $10,848,400

# Breakouts!

- Come up with an example of a statistic that is relevant to public policy and that might be contaminated by outliers
- What effect could this have on policy decisions?
- How might you address the problem?

# Median

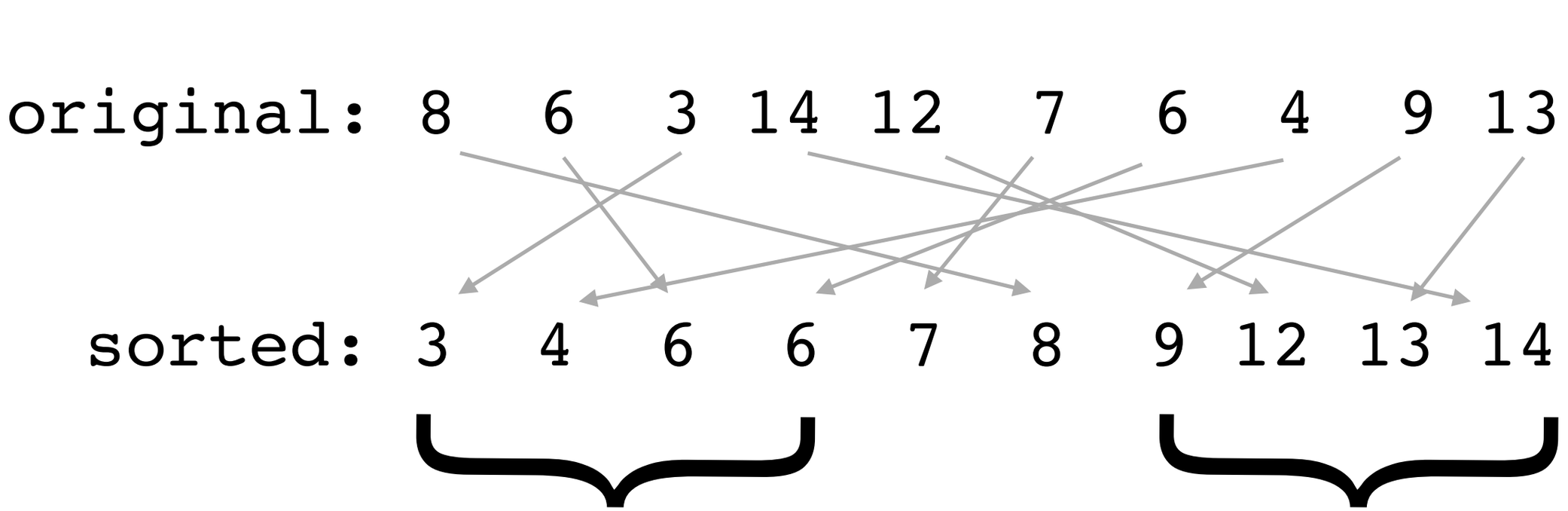
- When the scores are ordered from smallest to largest, the median is the middle score



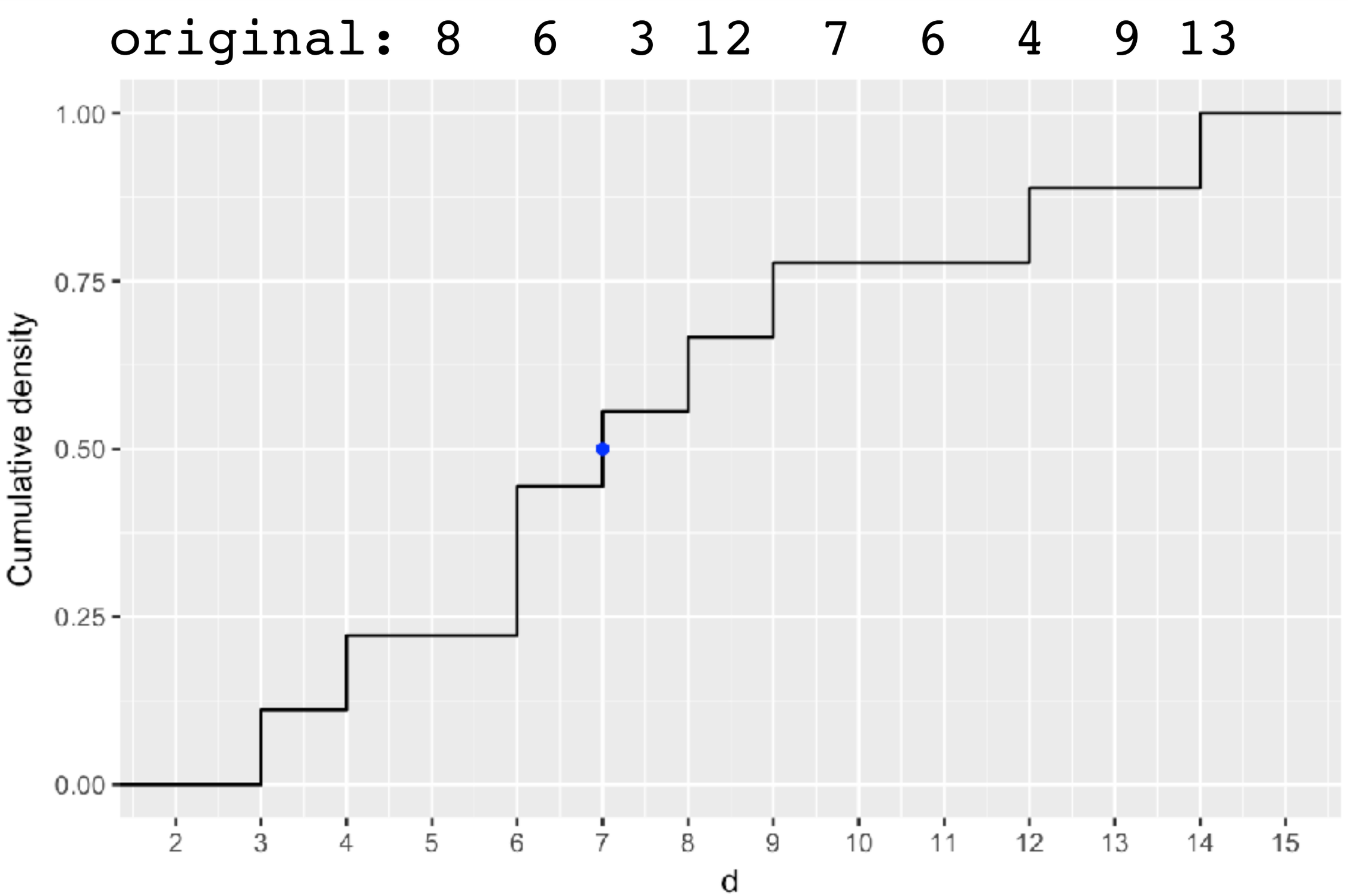

median = 7

# 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



# Median as the 50th percentile



# The median minimizes absolute error

- The mean minimizes the sum of squared errors

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

- The median minimizes the sum of absolute errors

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

Why do you think that matters?

# The median is less sensitive to outliers

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

mean income: $61,600

median income: $64,000

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

mean income: $10,848,400

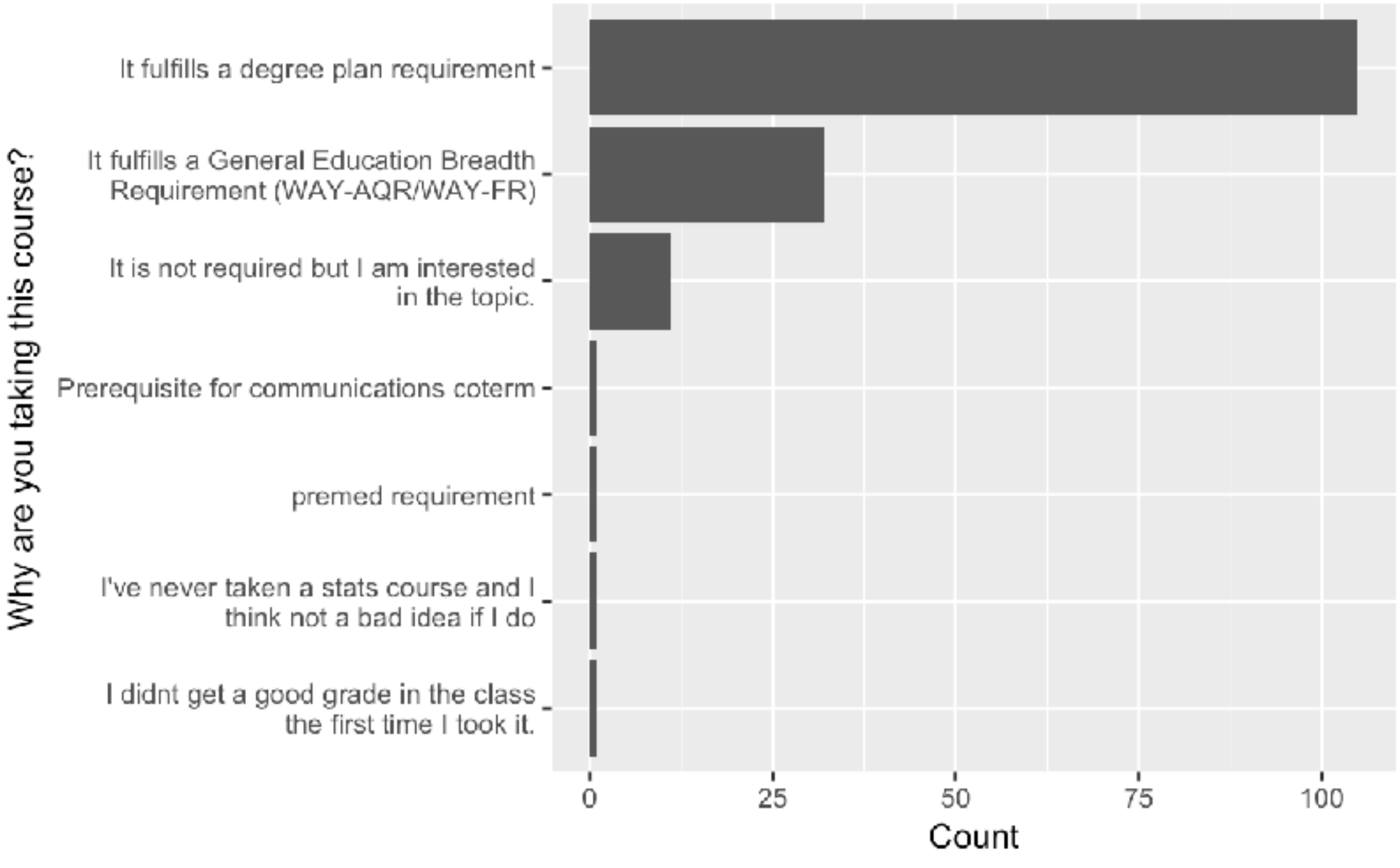
median income: $64,000

# 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
  - More on this later
- But the median is more robust
  - Less likely to be influenced by outliers
- Statistics is all about tradeoffs…

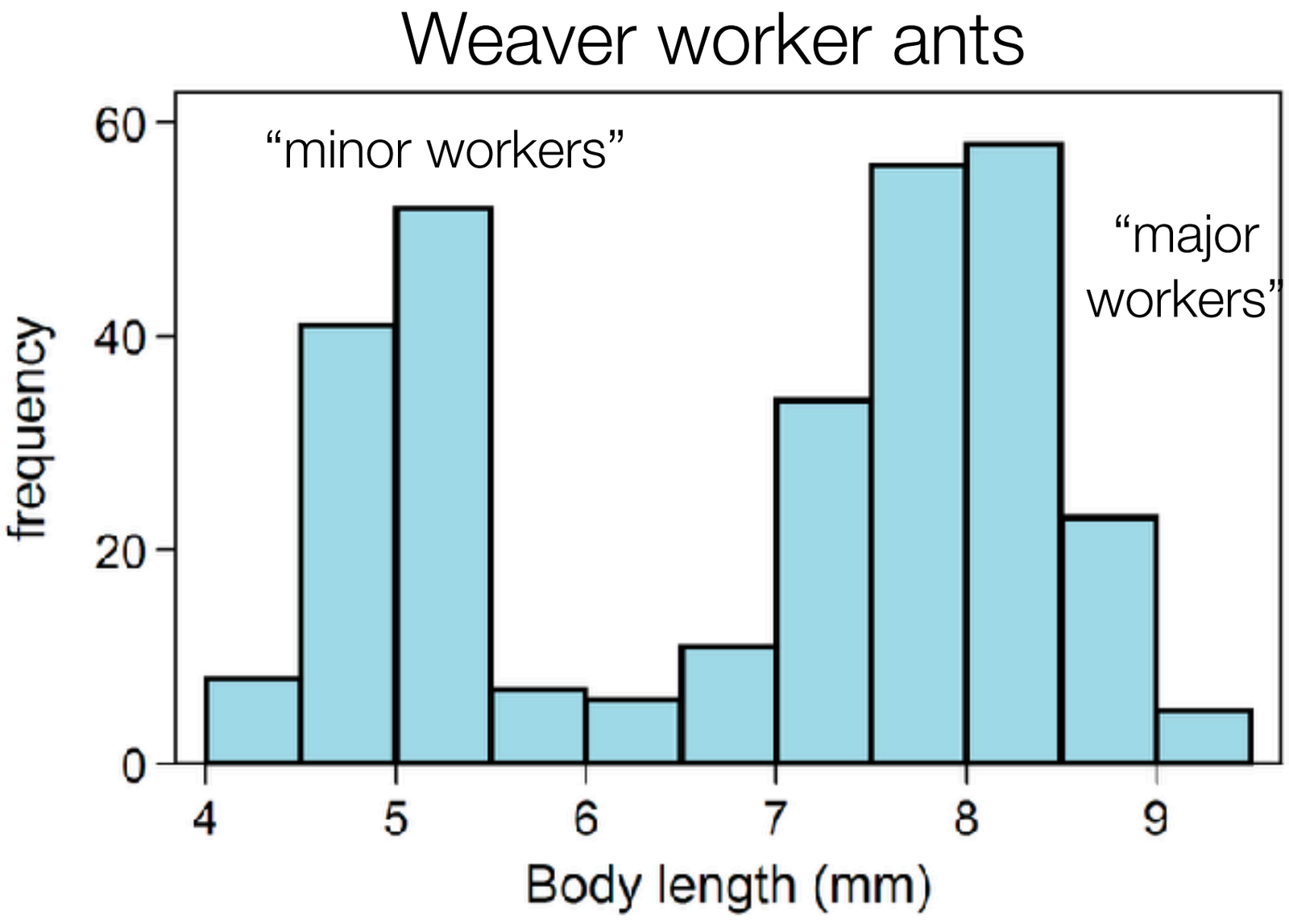
# Mode

- What is the most common value in the dataset?

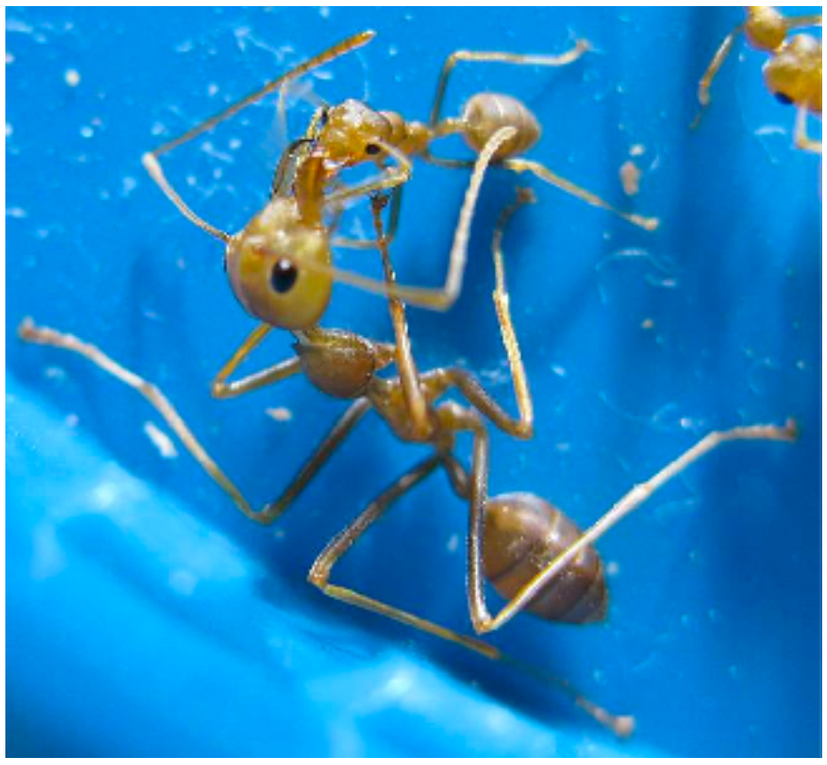


# Bimodal distributions

- There is not necessarily a single peak in the distribution




https://commons.wikimedia.org/wiki/File:BimodalAnts.png



Minor worker
grooming a major worker

https://termitesandants.blogspot.com/2010/04/oecophylla-smaragdina.html

# The fit of the sample mean: Variance and standard deviation

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

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

mean=6

$$SD = \sqrt{variance}$$

SSE: 20

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

SD=sqrt(5)=2.2

# Why we use N-1 when estimating the variance from a sample

- The variance of the population ($\sigma^2$) is defined as:
  - where $\mu$ is the population mean

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

- However, if we use this same equation with samples from the population, it is going to be biased on average - that is, we expect its value to be slightly different from the population value:

$$s^2 = \frac{n}{n-1}\sigma^2$$

- In order to get an unbiased estimate of the population variance from the sample data, we need to correct it:

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

# Degrees of freedom

- How many values are free to vary once the statistic is computed?

| x |
|---|
| 3 |
| 5 |
| 6 |
| 7 |
| 9 |

mean=6

| x |
|---|
| 3 |
| 5 |
| 6 |
| 7 |
| **?** |

$$\frac{3+5+6+7+x}{5} = 6$$

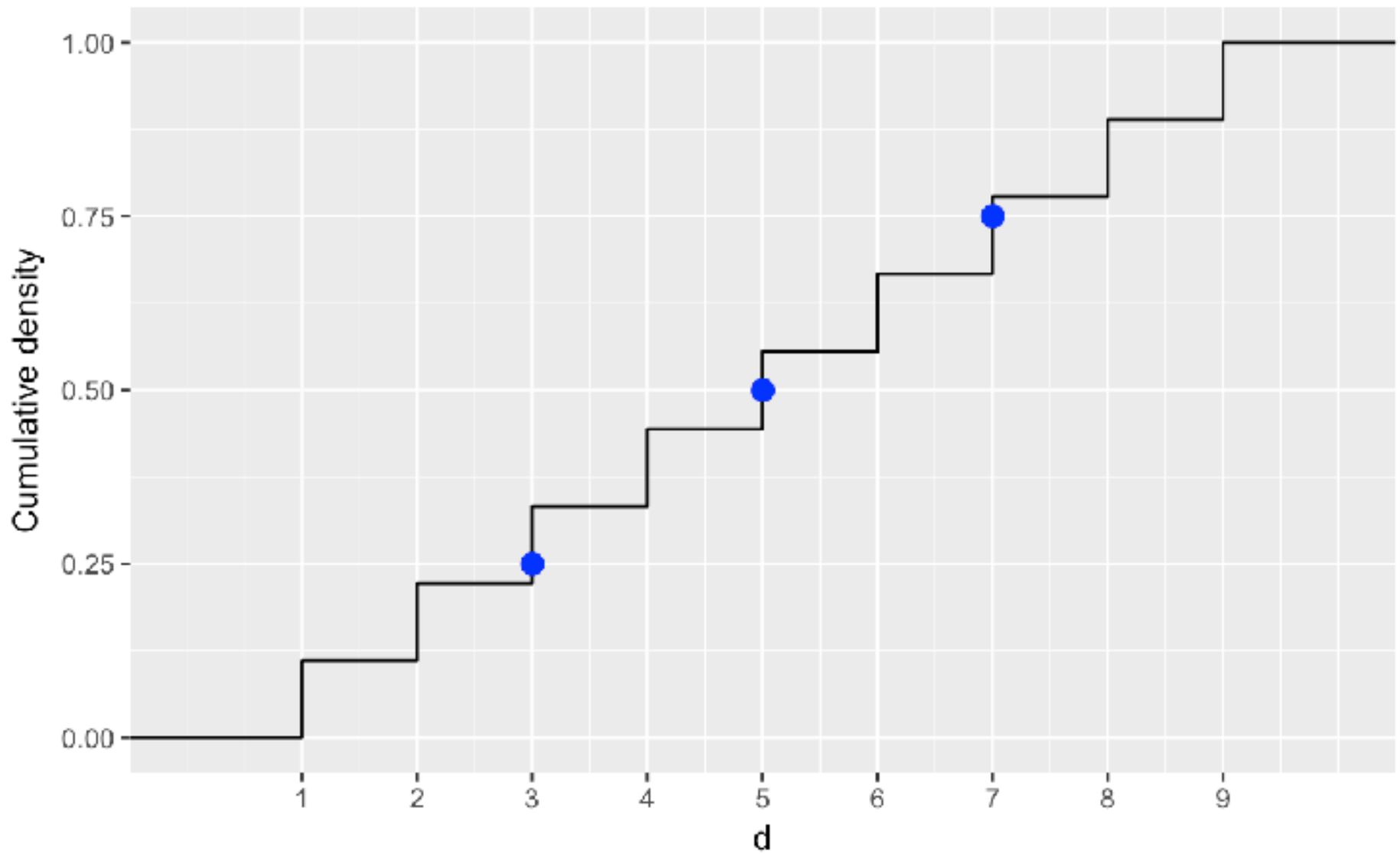
$$x = 6 * 5 - 21 = 9$$

Once the mean has been computed,
we only have n-1 *degrees of freedom*

# Robust measures of dispersion: interquartile range

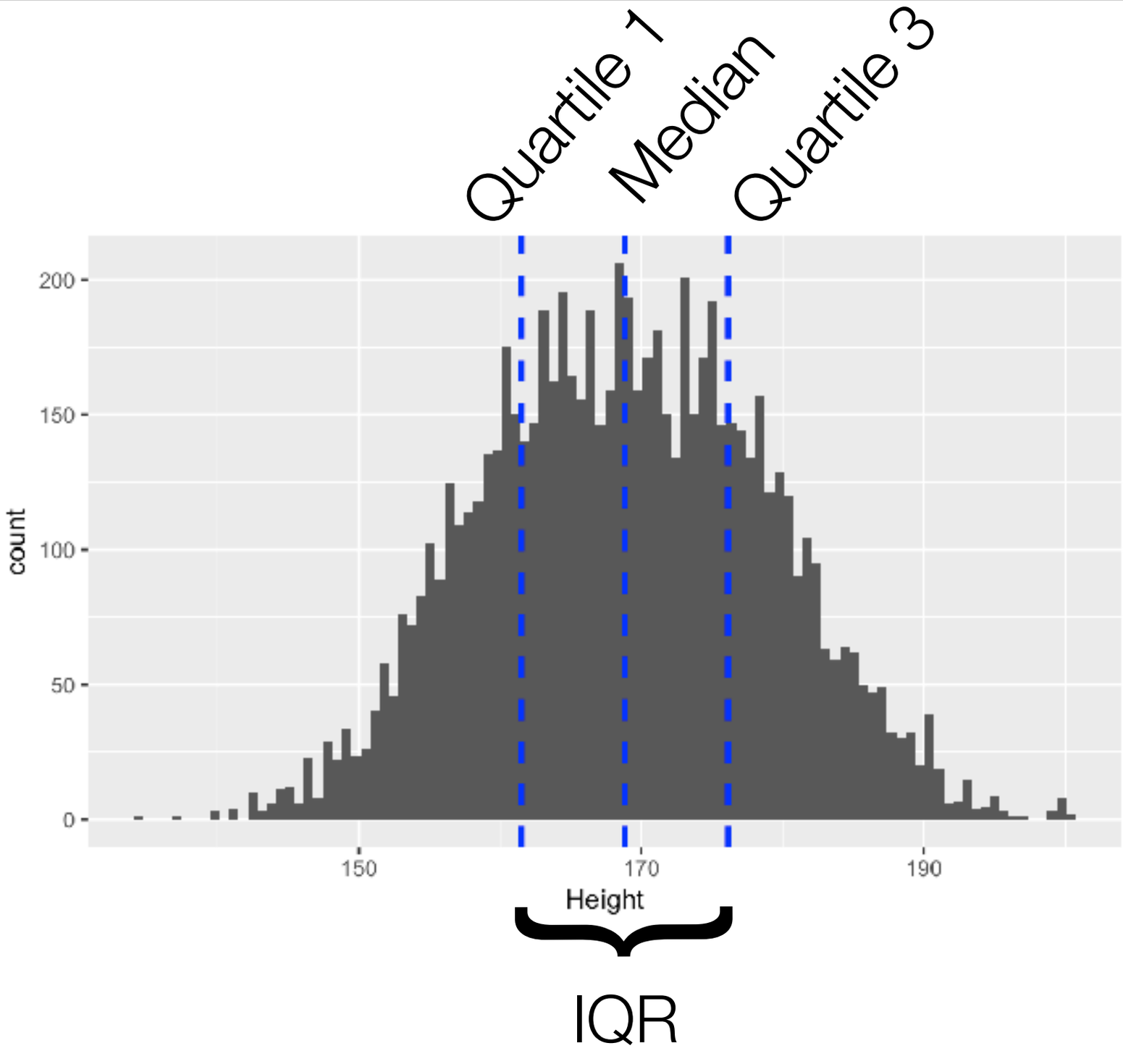
- Quartiles:
  - 25th, 50th, and 75th percentiles

d=seq(1,9)=c(1,2,3,4,5,6,7,8,9)

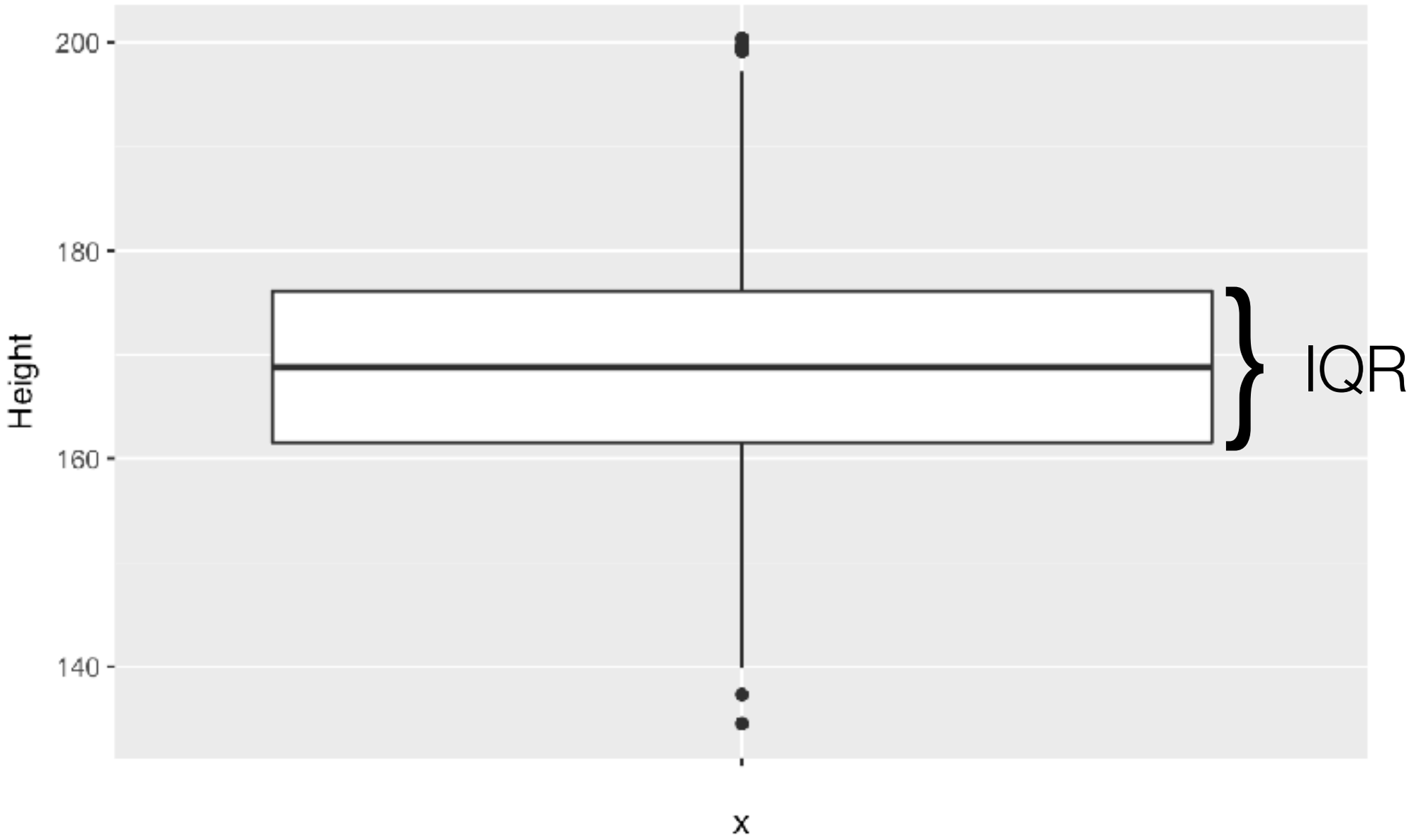


# Interquartile range on NHANES height

- IQR contains 50% of values
- vs.1 standard deviation, which contains ~34% of values
- If data are normally distributed:
  - IQR ~ SD*1.349



# Box plots and IQR



# Effect of outliers on estimates of dispersion

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

std deviation: $9,099

interquartile range: $8,000

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

std deviation: $24,122,479

interquartile range: $14,000

# Recap

- The basic statistical model: outcome = model + error
- A better fitting model is better, up to a point
- The simplest model is the central tendency of the data
- Measures of central tendency include the mean, median, and mode
- The fit of the central tendency is defined as the deviation