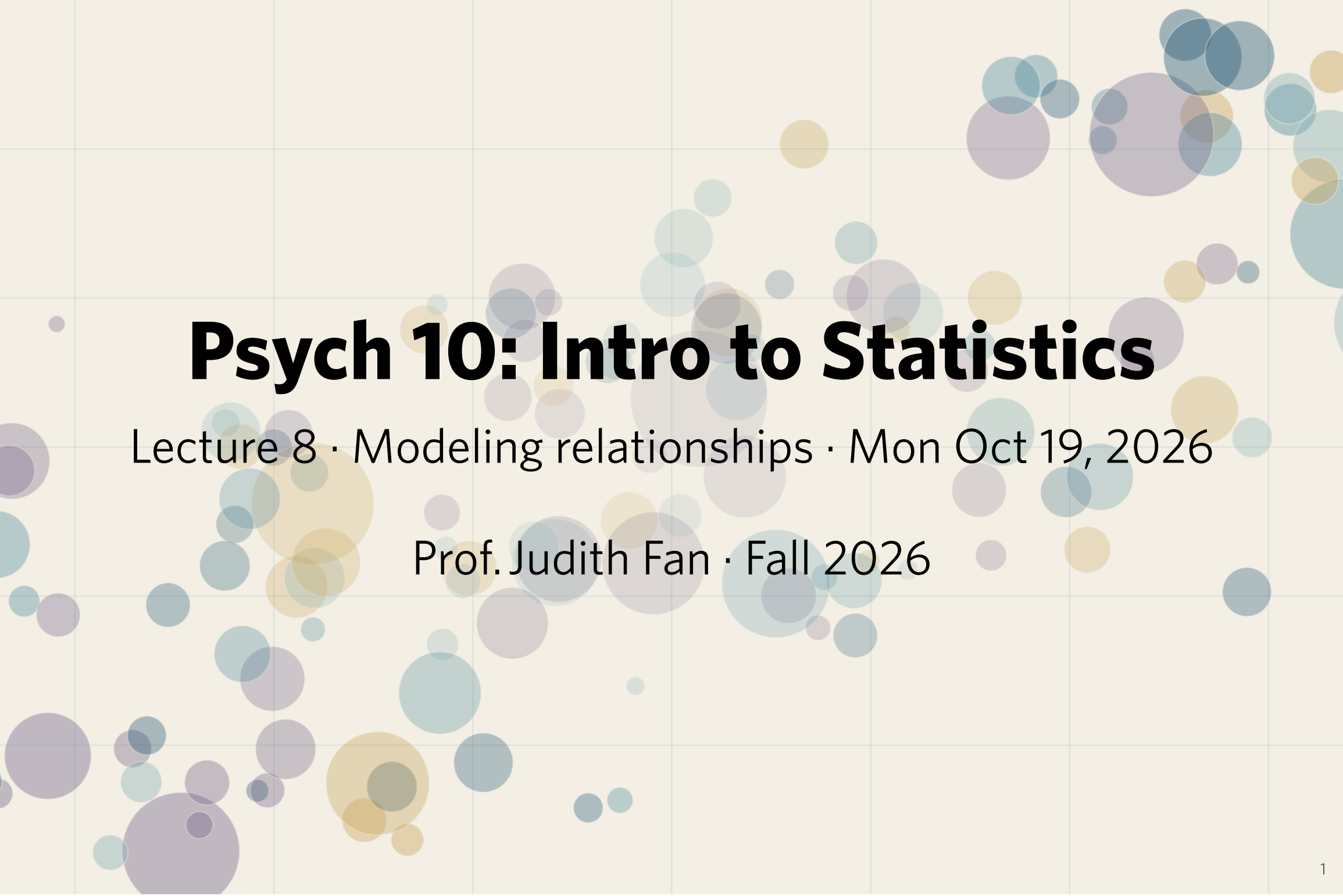




Psych 10: Intro to Statistics

Lecture 8 · Modeling relationships · Mon Oct 19, 2026

Prof. Judith Fan · Fall 2026

Reminders & Announcements

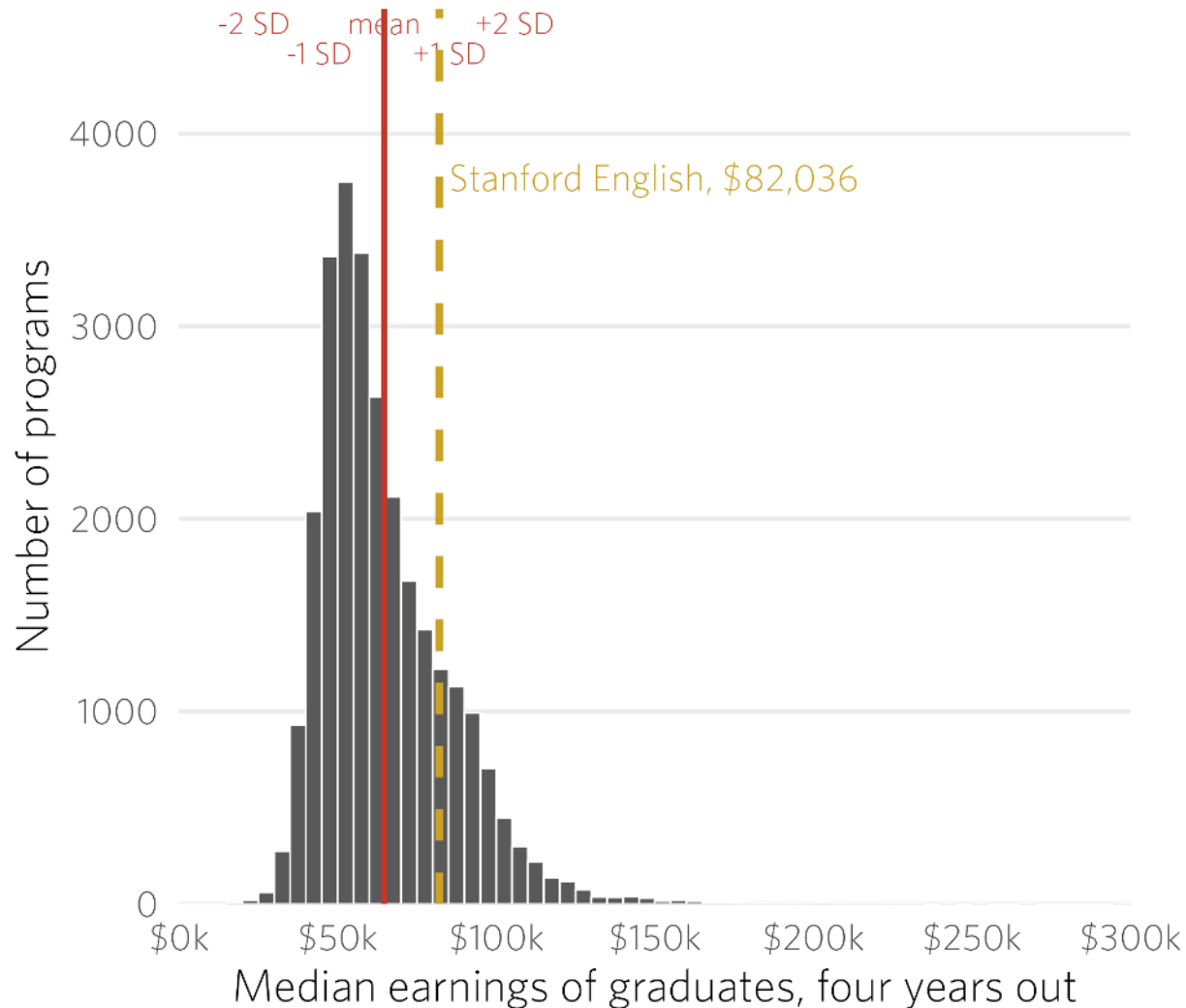
- CourseKata **Chapters 7 and 8** are due this Friday.
- Section this week: **Milestone 4**, *Deciding on a research question*, due Sunday, October 25 on Gradescope.
- Friday: **Case #4, The Reaction Room**, in your tutorial session. Check your discussion section's Slack channel for your time slot. Bring a pen.
- The **practice exam** is three weeks from today, Monday, November 9, in class. The exam is Wednesday, November 18.
- **Office hours:** today (Monday) 3–4pm in 420-300 as usual. No office hours this Friday, October 23.

Your project so far

- **Milestone 1.** A claim in the news that reports a number, and its source dataset.
- **Milestone 2.** The dataset's cases and variables, and a first plot.
- **Milestone 3.** Your group chose one claim and one dataset.
- **Milestone 4, this week.** A research question about that claim and dataset. An example follows, with a dataset you have seen.

One number summarizes one variable

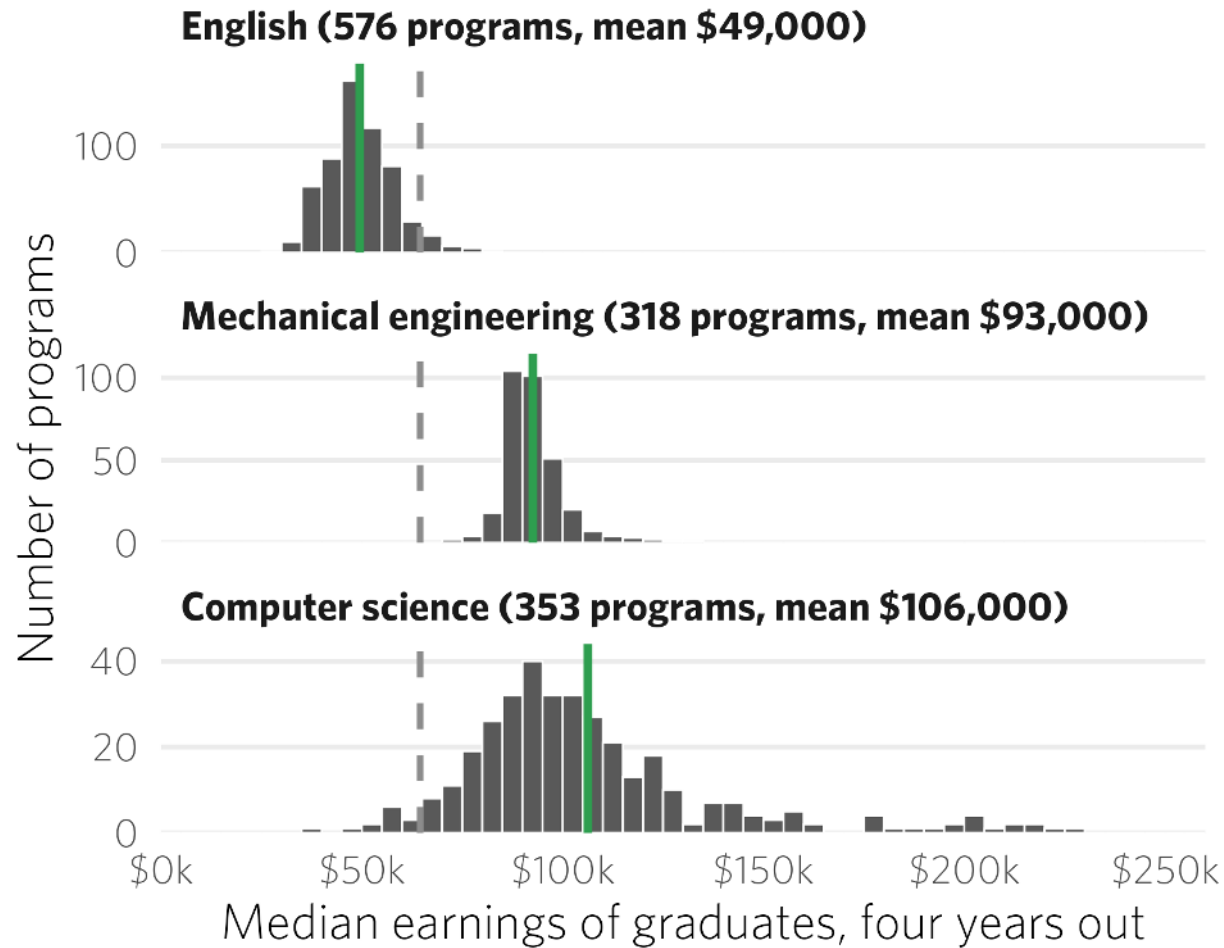
27,247 bachelor's programs: mean \$64,700, SD \$20,600



- The Times' figure: \$64,500 for the typical program, four years out.
- It is the mean of one variable, median earnings, across 27,247 programs.
- It is a model: one prediction for every program, and each deviation from it is error.
- The error is large: programs range from about \$20,000 to over \$200,000.

What explains the variation?

Dashed: the Times' \$64,500 for all programs.
Green: each major's mean.



- The \$64,500 doesn't tell you which programs lie above it and which below.
- But a second variable in the same dataset might: which major.
- A research question you could ask: **How much of the variation in program earnings does the major account for?**
- In words: *earnings = major + other stuff.*

Three processes behind the gap

- **The major itself.** The fields that hire each major pay different wages. In the data: the same gap at every school.
- **Which schools offer each major.** Engineering is taught mainly at large universities, whose graduates earn more in every major. In the data: a smaller gap within one school.
- **Who chooses each major.** Students planning careers in teaching or writing more often choose English, and those careers pay less whatever the major. In the data: a smaller gap within one line of work.
- The second and third are confounds: other variables that could, by themselves, produce the association.
- When coming up with three processes, try to include at least one confound.

What Milestone 4 asks for

- Commit to one claim and one dataset. Name the **outcome variable**, the one the claim summarizes, and an **explanatory variable** from the same dataset, each quantitative or categorical.
- Write the word equation, *outcome = explanatory variable + other stuff*, and the direction and approximate size you expect, in the outcome's units.
- Write three processes that could produce the relationship, and for each, what you would see in the data. Try to include at least one confound.
- Look for other variables that could distinguish the processes, and pose a second question built on one such variable.

Explaining variation with a second variable

- **Weeks 1 to 4, Chapters 1 to 6:** one variable at a time. Its distribution, its mean as a model, and the error around the mean.
- **Week 5, Chapters 7 and 8:** a second variable accounts for part of that variation. The model is now one mean per group.
- **Milestone 4:** the same question about your own dataset. How much of the variation in the variable the claim summarizes does a second variable account for?

Today

Modeling relationships



A mean for each group



How much error did the
groups remove?

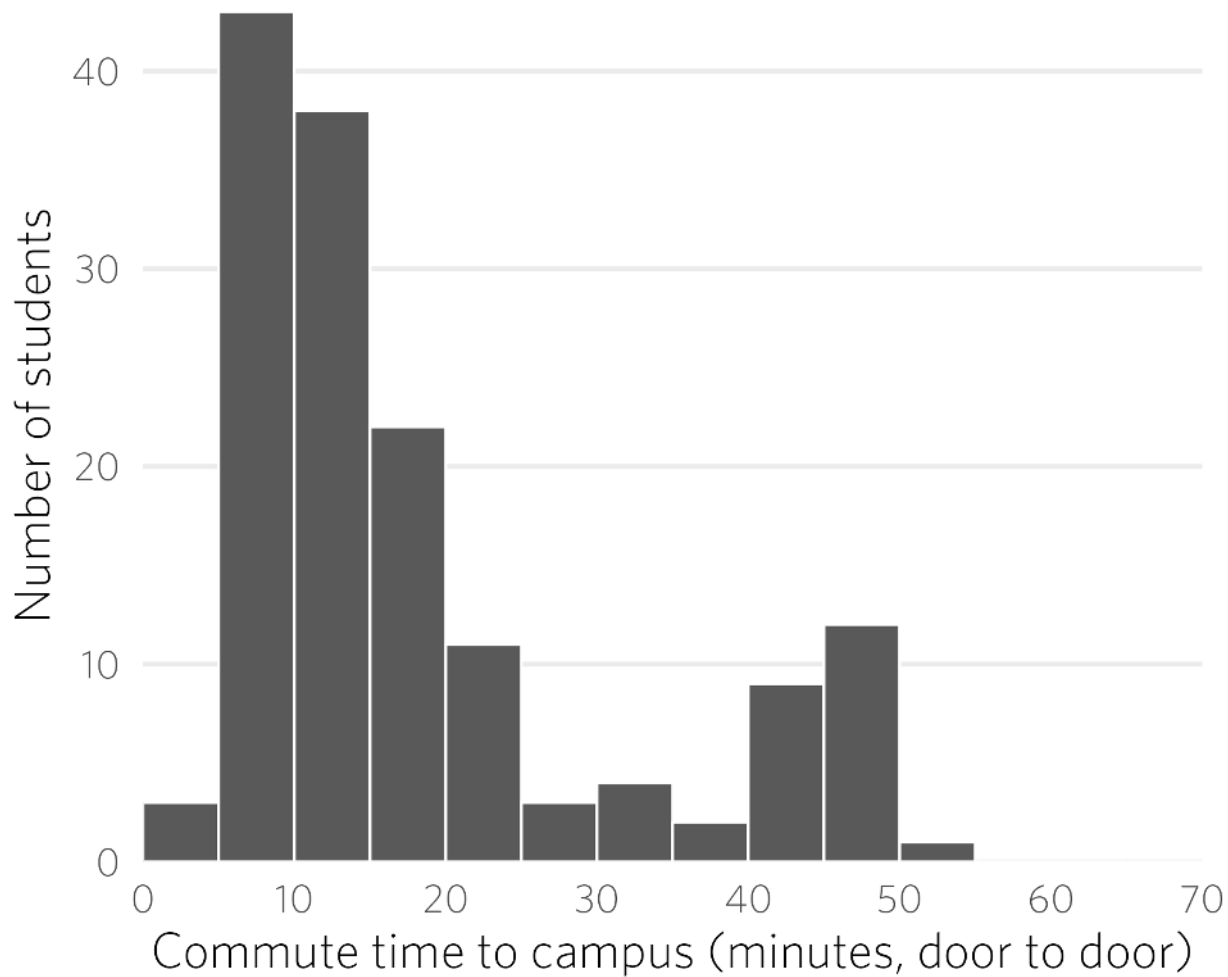


Reading the model's two
numbers

The running example: the 148 commute times from Week 2, split by how each student got to campus.

The 148 commute times from Week 2

Simulated class survey, $n = 148$



- In Week 2 the worksheet asked what process produced the second peak near 45 minutes.
- The survey also recorded how each student got to campus.

One model for everyone

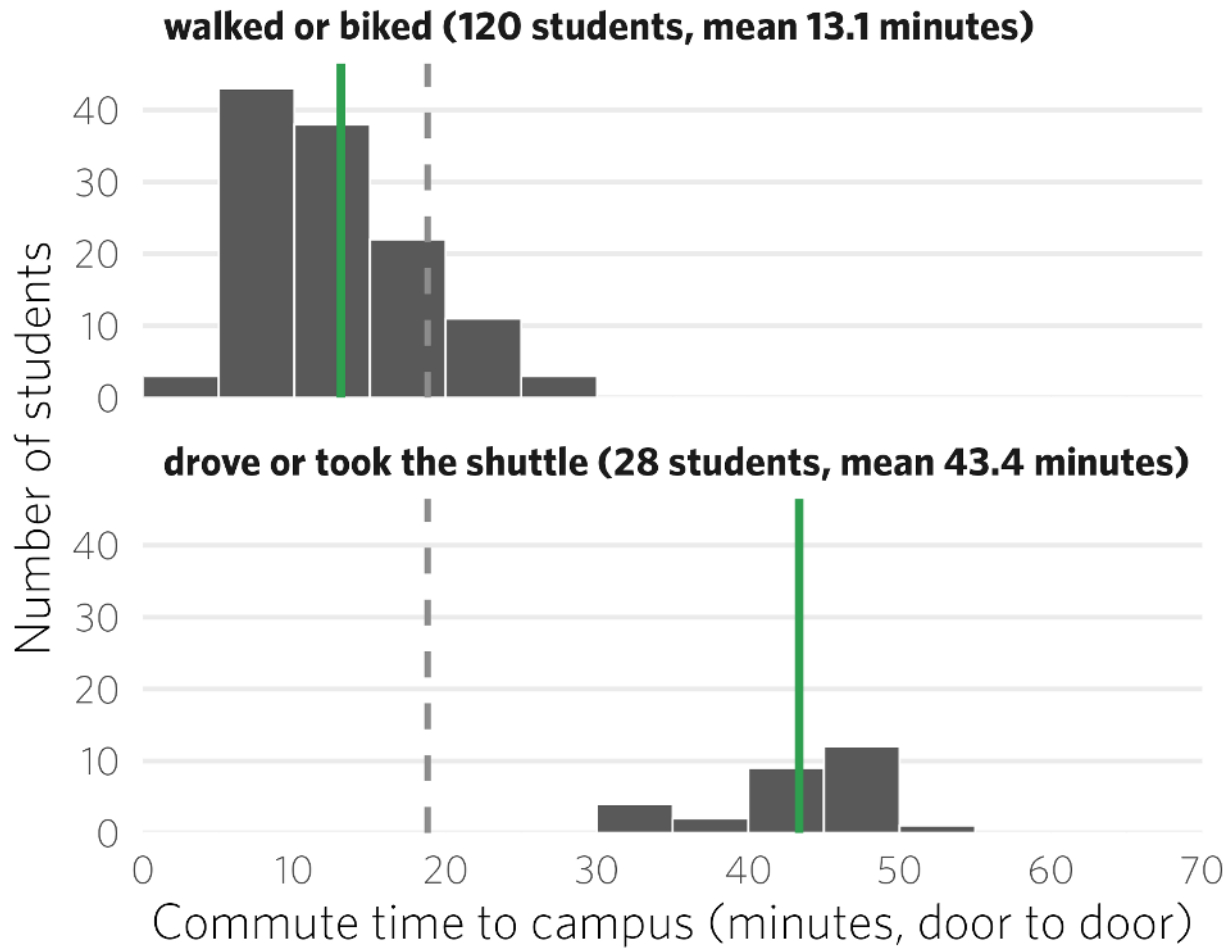
One model for everyone: the mean, 18.8 minutes



- The mean of all 148 times is 18.8 minutes.
- **One student is picked at random. What commute time do you predict for that student?**
- **Now you learn that the student took the shuttle. What commute time do you predict for a shuttle rider?**

A mean for each group

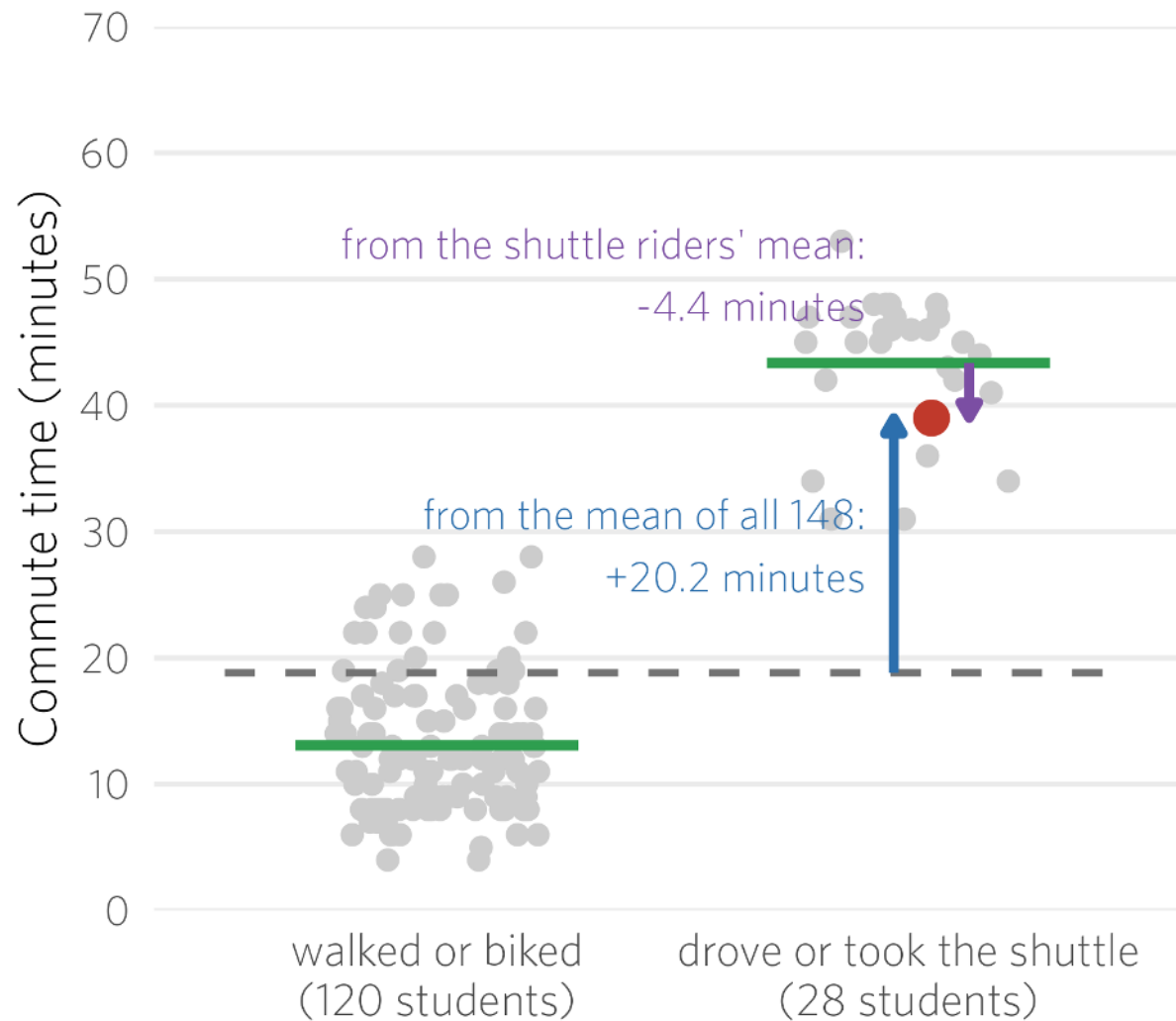
Dashed: the mean of all 148 students, 18.8 minutes.
Green: each group's mean.



- Walkers and bikers average 13.1 minutes; drivers and shuttle riders, 43.4.
- Predicting each student from their group's mean replaces one mean with two.
- In words: *commute time* = *how the student got to campus + other stuff*. $DATA = MODEL + ERROR$, with two means in the model.

A shuttle rider's residual under each model

One shuttle rider's residual under each model



- This student took 39 minutes.
- Empty model: $39 - 18.8 = +20.2$ minutes.
- Group model: $39 - 43.4 = -4.4$ minutes.
- A residual is defined relative to a model, so each student has one per model.

Worksheet - Part 1

In the Week 2 survey, 148 students reported their commute time. The mean of all 148 is 19 minutes. The 120 students who walked or biked have a mean of 13 minutes; the 28 who drove or took the shuttle have a mean of 43 minutes. (All three means are rounded to whole minutes.)

- a.** One student walked and took 22 minutes. The empty model predicts 19 minutes for every student. Compute this student's residual under the empty model, in minutes, with its sign.
- b.** The group model predicts each student's time with their own group's mean. Compute the same student's residual under the group model, in minutes, with its sign.
- c.** For this one student, which model has the smaller residual? Write *empty model* or *group model*.
- d.** Part c compared the two models for one student. To compare them for all 148 students, you would compute one quantity for each model. Name that quantity.

Part 1a-c · The walker's residual under each model

a. $22 - 19 = +3$ minutes

b. $22 - 13 = +9$ minutes

c. $3 < 9$

- **a.** The empty model predicts 19 minutes for everyone; the residual is +3.
- **b.** The group model predicts 13 minutes for a walker; the residual is +9.
- **c.** The empty model, for this student. The model with less total error can still have the larger residual for one student.

Part 1d - Which model fits all 148 students better?

- For each model, square every student's residual and add up the 148 squares. That sum is the model's total error.
- The model with the smaller total fits better.
- The difference between the two totals is the error the groups removed.

Today

Modeling relationships



A mean for each group



How much error did the groups remove?



Reading the model's two numbers

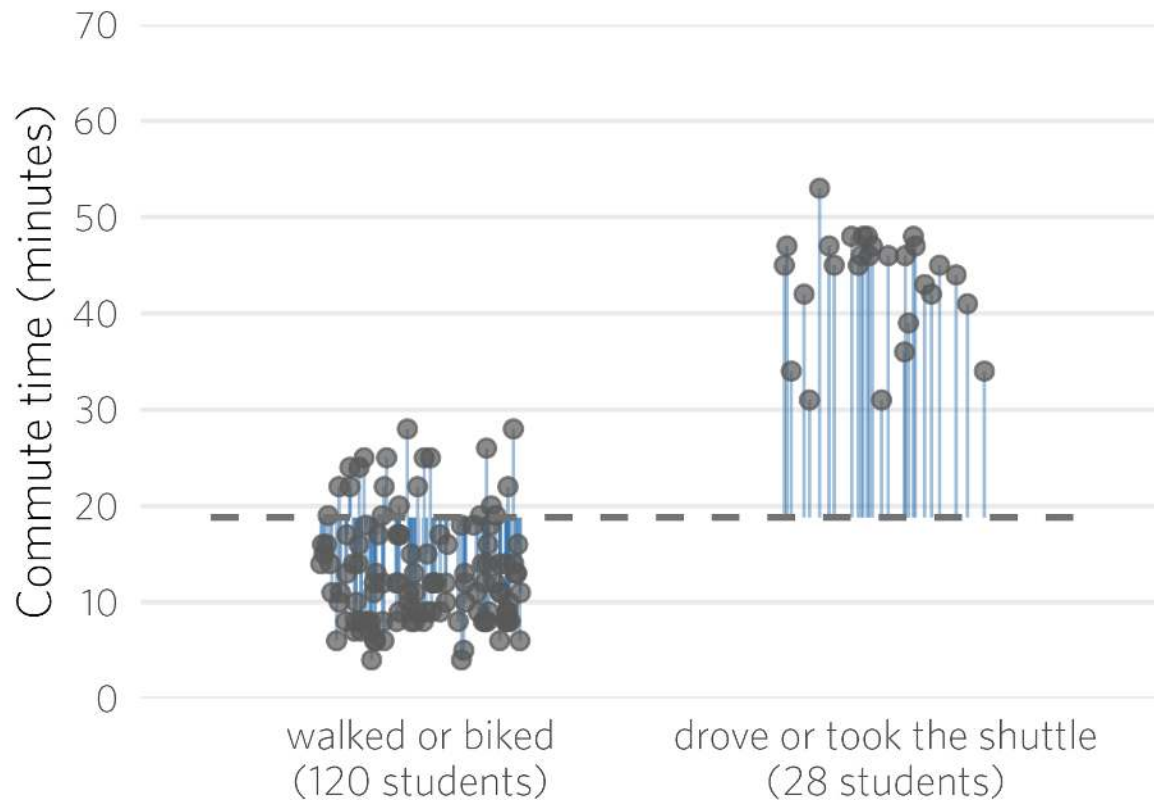
The sum of squared residuals under each model, and the difference between the two sums.

The empty model's total error: SS Total

$$PRE = \frac{SS_{\text{Total}} - SS_{\text{Error}}}{SS_{\text{Total}}}$$

SS_{Total}: each residual from the mean of all 148, squared and summed. The empty model's total error.

Residuals from the mean of all 148; SS Total = 25,335



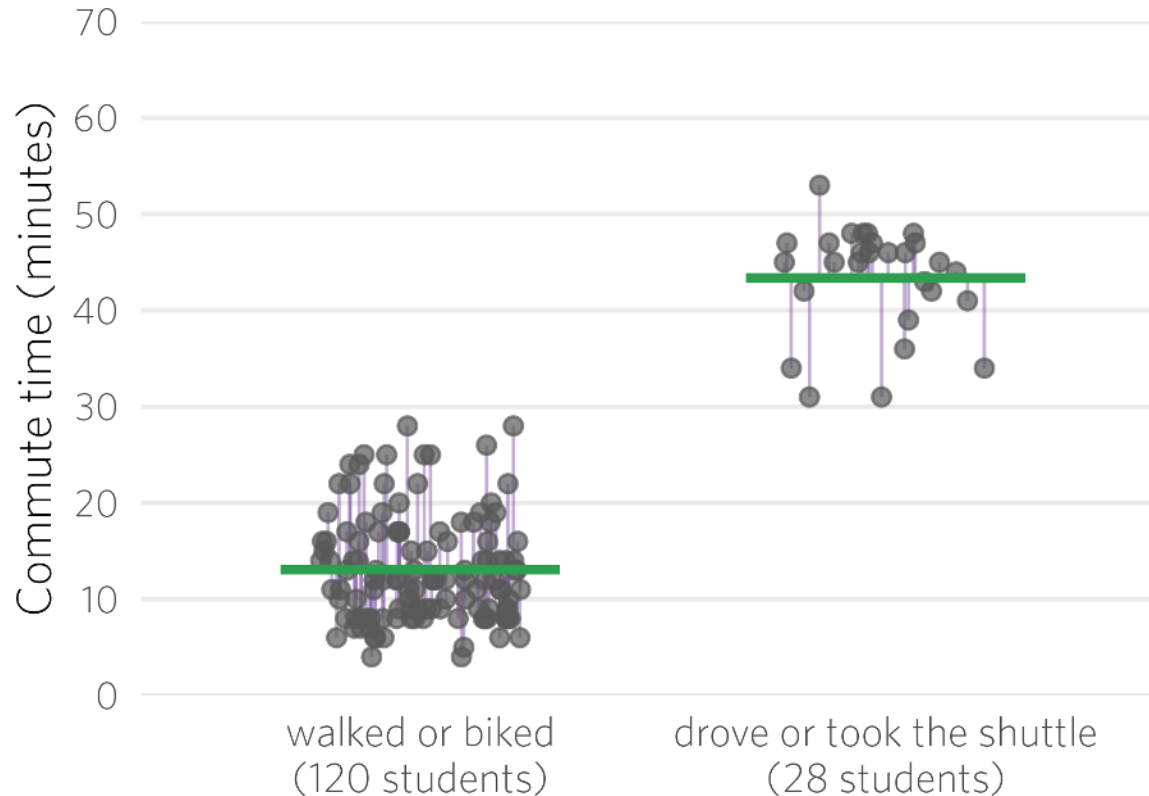
- SS Total = 25,335 square minutes.
- The sum of squares from Lecture 6, for the empty model.

The group model's total error: SS Error

$$PRE = \frac{SS_{\text{Total}} - SS_{\text{Error}}}{SS_{\text{Total}}}$$

SS_{Error}: each residual from the student's own group mean, squared and summed. The group model's total error.

Residuals from each group's own mean; SS Error = 4,528



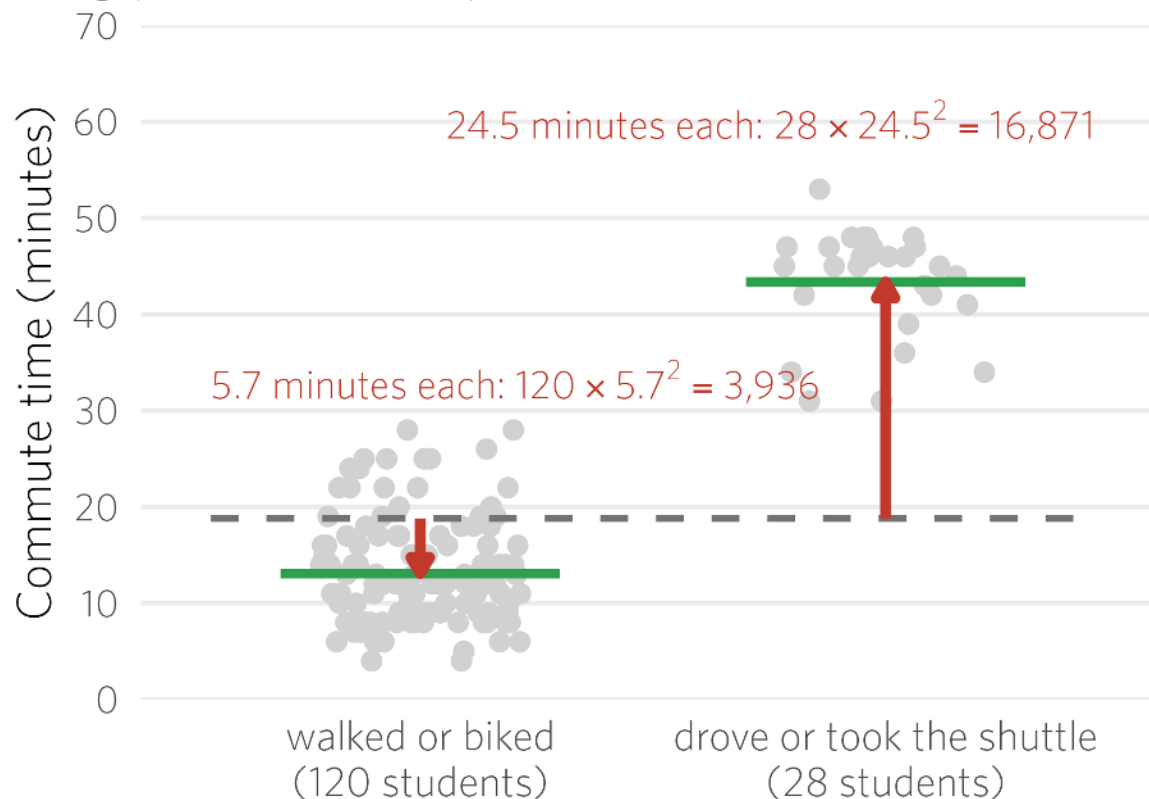
- SS Error = 4,528 square minutes.
- The same computation, with the group means as predictions.

The error the groups removed: SS Model

$$PRE = \frac{SS_{\text{Total}} - SS_{\text{Error}}}{SS_{\text{Total}}}$$

$SS_{\text{Total}} - SS_{\text{Error}}$, called **SS_{Model}** : the variation the group model accounts for.

The gap between the two predictions, for each student



- $SS_{\text{Model}} = 25,335 - 4,528 = 20,807$.
- Or student by student: each group mean's distance from 18.8, squared, times the group's size.
- With rounded means: $120 \times 5.7^2 + 28 \times 24.6^2 \approx 20,800$.

Proportional reduction in error (PRE)

$$PRE = \frac{SS_{\text{Total}} - SS_{\text{Error}}}{SS_{\text{Total}}} = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = \frac{20,807}{25,335} = 0.82$$

- SS Model is 20,807 square minutes. On its own, that number is hard to interpret.
- Divide it by SS Total: $20,807 / 25,335 = 0.82$.
- This is the **proportional reduction in error (PRE)**. The group model removes 82% of the empty model's error.
- In words: how students got to campus accounts for 82% of the variation in commute times.
- Other names for the same number: eta squared, R^2 .

Worksheet - Part 2

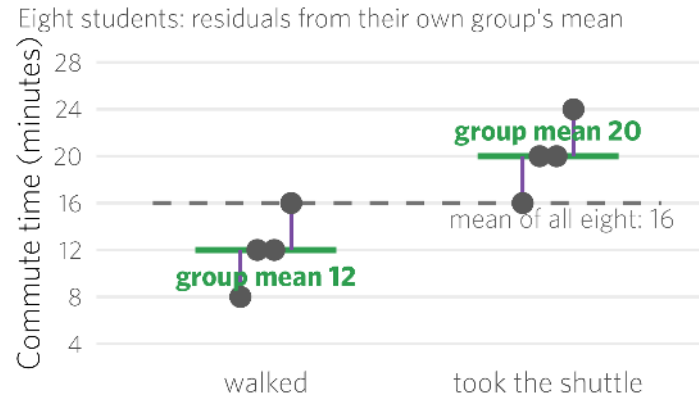
Eight students from one discussion section reported their commute times. The four who walked: **8, 12, 12, and 16** minutes. The four who took the shuttle: **16, 20, 20, and 24** minutes.

a. Compute three means: the walkers' mean, the shuttle riders' mean, and the mean of all eight students. Write all three, in minutes. Then fill in the rest of the table on your sheet; the first row is done.

Worksheet - Part 2, continued

- b.** Compute SS Total: add up the eight squared deviations from the mean of all eight students (the table's fourth column).
- c.** Compute SS Error: add up the eight squared deviations from each student's own group mean (the table's last column).
- d.** Compute $SS \text{ Model} = SS \text{ Total} - SS \text{ Error}$.
- e.** Compute the proportional reduction in error, $PRE = SS \text{ Model} / SS \text{ Total}$. Write it as a fraction and as a decimal.
- f.** Suppose the four shuttle riders had taken 36, 40, 40, and 44 minutes instead (each time 20 minutes longer). For each of SS Total, SS Error, and SS Model, write *larger*, *smaller*, or *unchanged*.

Part 2a-c · The two sums of squares



a. means 12, 20, and 16 minutes

$$\begin{aligned} SS_{\text{Total}} &= 64 + 16 + 16 + 0 + 0 + 16 + 16 + 64 \\ &= 192 \end{aligned}$$

$$\begin{aligned} SS_{\text{Error}} &= (16 + 0 + 0 + 16) + (16 + 0 + 0 + 16) \\ &= 32 + 32 = 64 \end{aligned}$$

- **a.** Walkers: $48 / 4 = 12$. Shuttle riders: $80 / 4 = 20$. All eight: $128 / 8 = 16$.
- **b.** Deviations from 16: -8, -4, -4, 0, 0, 4, 4, 8. Squared: 64, 16, 16, 0, 0, 16, 16, 64.
- **c.** Deviations from each group's own mean: -4, 0, 0, 4 in both groups. Squared: 16, 0, 0, 16 in both groups.

Part 2d-e · SS Model and PRE

$$\begin{aligned}SS_{\text{Model}} &= SS_{\text{Total}} - SS_{\text{Error}} \\ &= 192 - 64 = 128\end{aligned}$$

$$\text{check: } 8 \times 4^2 = 128$$

$$PRE = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = \frac{128}{192} = \frac{2}{3} \approx 0.67$$

- **d.** $SS_{\text{Model}} = 128$. This is the error the group means removed.
- Check: each student's group mean is 4 minutes from 16, and $8 \text{ students} \times 4^2 = 128$.
- **e.** $PRE = 2/3$. The proportional reduction in error: the group means removed two thirds of the empty model's error.
- In words: how students got to campus accounts for two thirds of the variation in these eight commute times.

Part 2f · Moving the shuttle riders

group means 12 and 40; all eight 26

$$SS_{\text{Error}} = 32 + 32 = 64$$

$$SS_{\text{Model}} = 8 \times 14^2 = 1,568$$

$$SS_{\text{Total}} = 1,568 + 64 = 1,632$$

$$PRE = 1,568/1,632 = 0.96$$

- The shuttle riders' mean becomes 40, and the mean of all eight becomes 26.
- SS Error: unchanged. The spread within each group is the same.
- SS Model and SS Total: larger. Each group mean is now 14 from the overall mean, not 4.
- PRE grows with the distance between group means, relative to the spread within groups.

The same partition for all 148 students

	SS	PRE
Model (how the student got to campus)	20,807	0.82
Error	4,528	
Total	25,335	

- R reports a model's fit in this layout.
- Three more columns, df, MS, and F, come on Wednesday.

Today

Modeling relationships

1

A mean for each group

2

How much error did the
groups remove?

3

Reading the model's two
numbers

R reports the two group means as a reference mean and a difference between means.

The model as R reports it

```
1 lm(minutes ~ travel, data = survey)
```

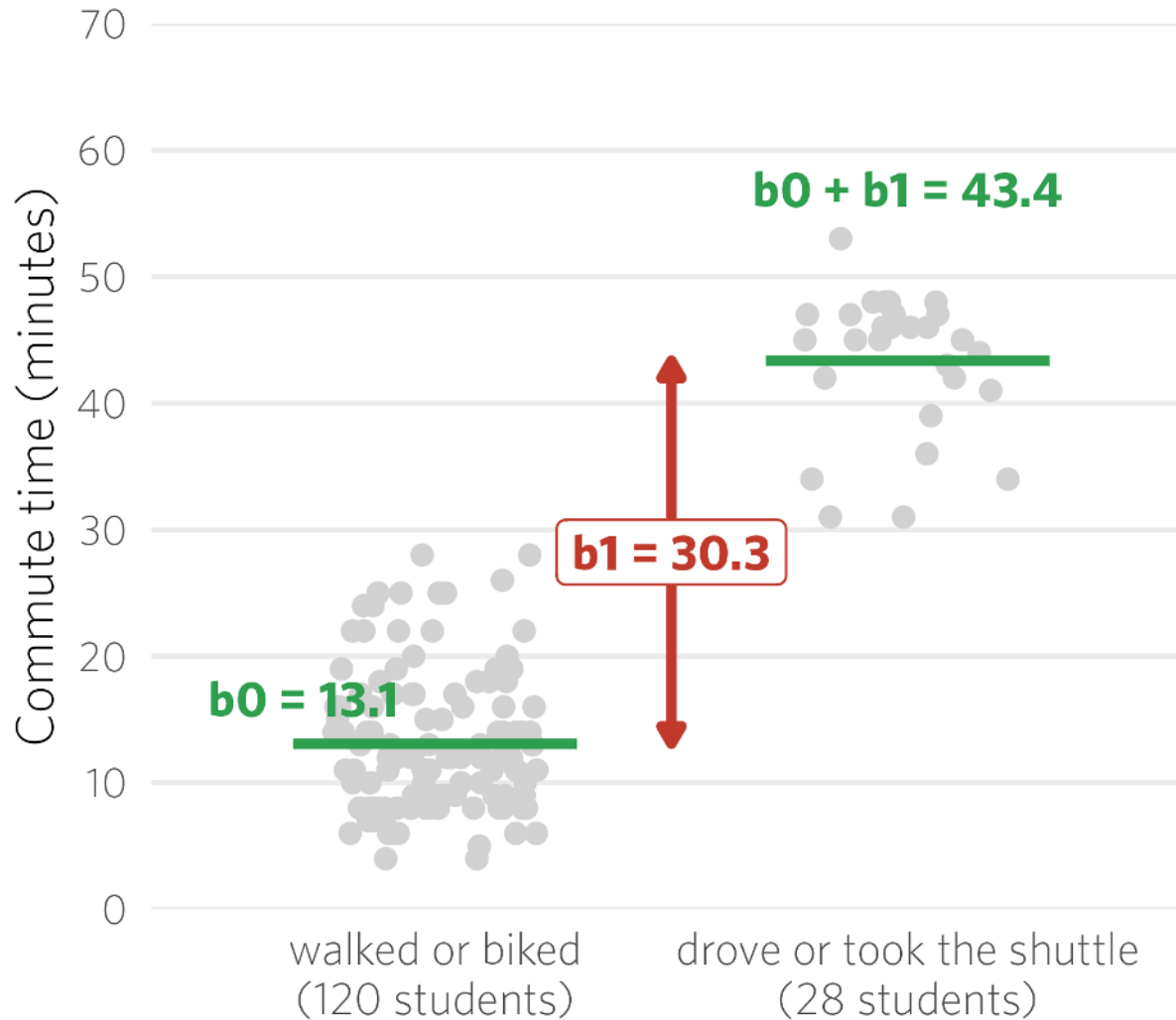
Coefficients:

(Intercept)	travelshuttle
13.08	30.27

- Two numbers, for a model with two means.
- Neither is 43.4, the shuttle riders' mean. The label `bike` is absent.
- **Which group's mean is 13.08? What is 30.27?**

b_0 and b_1 on the figure

The two numbers R reports and the two group means



- $b_0 = 13.1$ is the walkers' mean.
- $b_1 = 30.3$ is the difference between the means: $13.1 + 30.3 = 43.4$.
- R reports the second mean as a difference from the first.

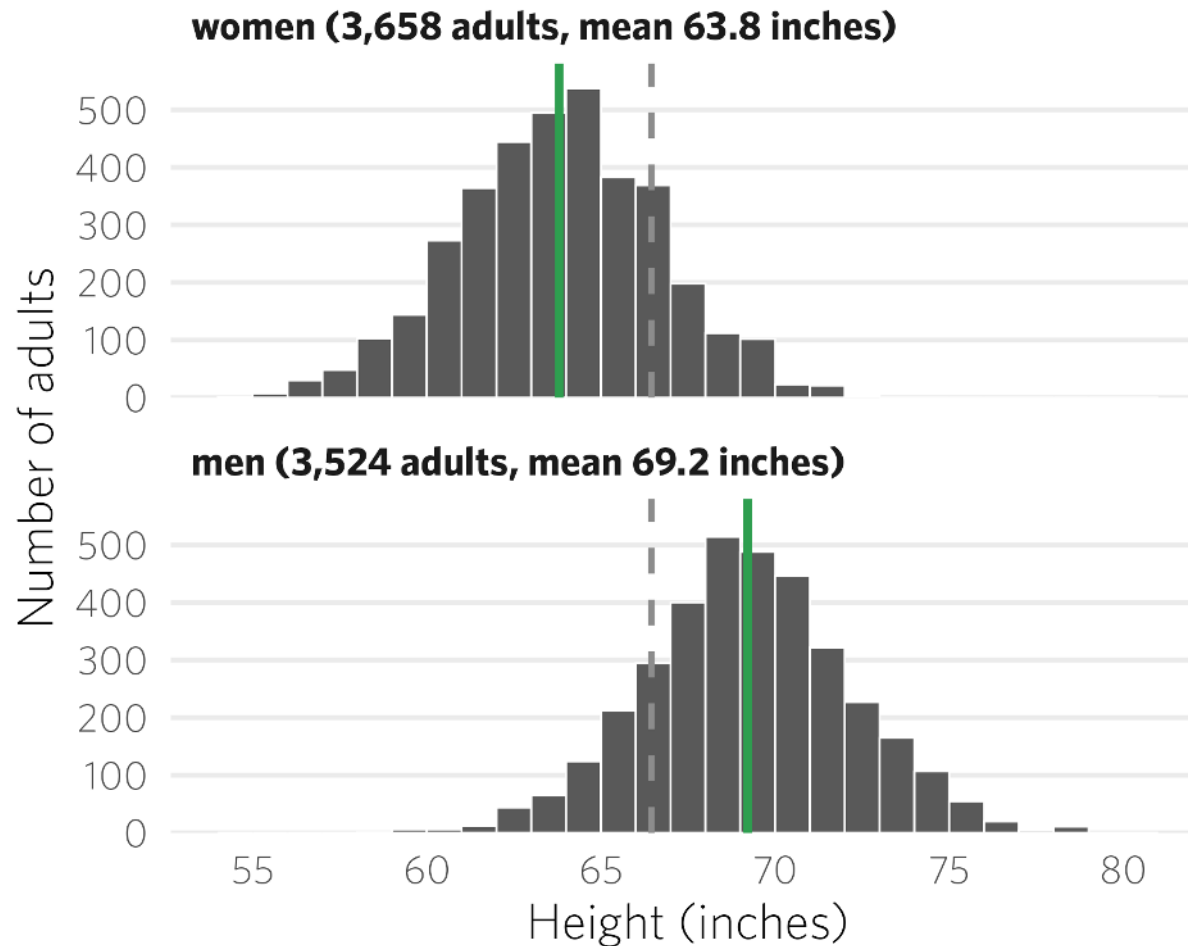
The model in symbols

$$Y_i = b_0 + b_1 X_i + e_i$$

- Y_i is student i 's commute time. e_i is the residual.
- b_0 is the walkers' mean. b_1 is the difference between the means.
- X_i is 0 for a walker and 1 for a shuttle rider. The 0 and 1 are labels, not quantities.
- A walker's prediction is b_0 . A shuttle rider's is $b_0 + b_1$.
- R's reference group is the alphabetically first label. With `walk` and `drive`, R would print `(Intercept) 43.4` and `travelwalk -30.3`. The predictions are the same.

Lecture 7's heights as a group model

7,182 adults aged 20 and over (NHANES, 2009-2012).
Dashed: the mean of all, 66.5 inches.



- The same model fits any quantitative outcome and two-group explanatory variable.
- Here: the heights of 7,182 NHANES adults, by sex.
- The R output and the sums of squares are on your worksheet.

Worksheet - Part 3

In the NHANES survey, 7,182 adults had their height measured. A group model of height by sex, fitted in R, gives these two numbers (in inches, rounded):

(Intercept)	64
Gendermale	5

and these sums of squares: **SS Model 52,000 · SS Error 61,000 · SS Total 113,000.**

- a.** What is the mean height of the men, in inches?
- b.** Write the model's prediction for a man in the form $b_0 + b_1X_i$, with numbers in place of b_0 , b_1 , and X_i .
- c.** What proportion of the variation in height does sex account for? Write the division and its result as a decimal.

Worksheet - Part 3, continued

- d.** An AI assistant summarized this output in three sentences: “The intercept of 64 is the mean height of the women. The coefficient of 5 for Gendermale is the mean height of the men. PRE = 0.46 means that the model predicts 46% of adults’ heights correctly.” Two of the three sentences are wrong. Say which two, and write a corrected version of each.
- e.** Sex accounts for about 46% of the variation in height. Name one other variable that could account for some of the other 54%.

Part 3a-b · The men's mean from the two numbers

a. $64 + 5 = 69$ inches

b. $64 + 5 \times 1 = 69$

$$64 + 5 \times 0 = 64$$

- **a.** $64 + 5 = 69$ inches: the women's mean plus the difference.
- **b.** $X_i = 1$ for a man, so $64 + 5$; $X_i = 0$ for a woman, so 64.

Part 3c · PRE for height

$$PRE = \frac{52,000}{113,000} \approx 0.46$$

- Sex accounts for about 46% of the variation in adult height.
- The histograms overlap: many women are taller than many men. That within-sex variation is the model's error.

Part 3d - The two errors in the summary

- Sentence 1 is correct: the intercept is the reference group's mean, the women's.
- Sentence 2 is wrong: 5 is the difference between the means. The men's mean is $64 + 5 = 69$ inches.
- Sentence 3 is wrong. $\text{PRE} = 0.46$ means the group model removes 46% of the total variation in height. It is not a share of adults; the model predicts no one's height exactly.

Part 3e · The other 54%

- Within each sex, heights have an SD of about 2.9 inches. That spread is the model's error.
- Variables associated with height within a sex: age (adults lose height after about 50), childhood nutrition, the parents' heights, country of birth.
- NHANES records age and country of birth.

Before Friday

- Finish CourseKata **Chapters 7 and 8**.
- Friday: **The Reaction Room**. Check your discussion section's Slack channel for your time slot.
- Milestone 4 is due Sunday.

Hand in your worksheet on the way out

On the last line, write down one thing you learned today.

Thanks!