

Psych 10: Intro to Statistics

Lecture 9 · Comparing group models · Wed Oct 21, 2026

Prof. Judith Fan · Fall 2026

Reminders & Announcements

- CourseKata **Chapters 7 and 8** are due Friday.
- **Milestone 4** is 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.

Today

Comparing group models



Claude Code demo
with TA Anna Kelley



Reading the model's
two numbers



A model with three
means



The cost of a
parameter

The hour opens with a live Claude Code demonstration led by TA Anna Kelley, about 20 minutes; the worksheet then starts with the part Monday's class did not reach.

Pro tips: AI agentic coding

- A live demonstration with **Anna Kelley**.
- The question: do students who sleep more score higher on quizzes?
- The data: a small synthetic file, `sleep_quiz_demo.csv`, with each student's hours of sleep and quiz score.
- Claude Code runs from the Terminal in RStudio. It reads your project files, writes and runs R code, and reports back.

What to take from the demo

- **Give context.** Tell the AI what you're working on and what you want.
- **Stay in control.** Know the files it reads and the actions you approve.
- **Verify.** Read the code and the results, not only the AI's summary.
- **Stay engaged.** Work in small steps, ask questions, and use the AI to build your understanding.
- The AI can do much of the work of research, but not take responsibility for it. You are the researcher.

Today

Comparing group models



Claude Code demo
with TA Anna Kelley



Reading the model's
two numbers



A model with three
means

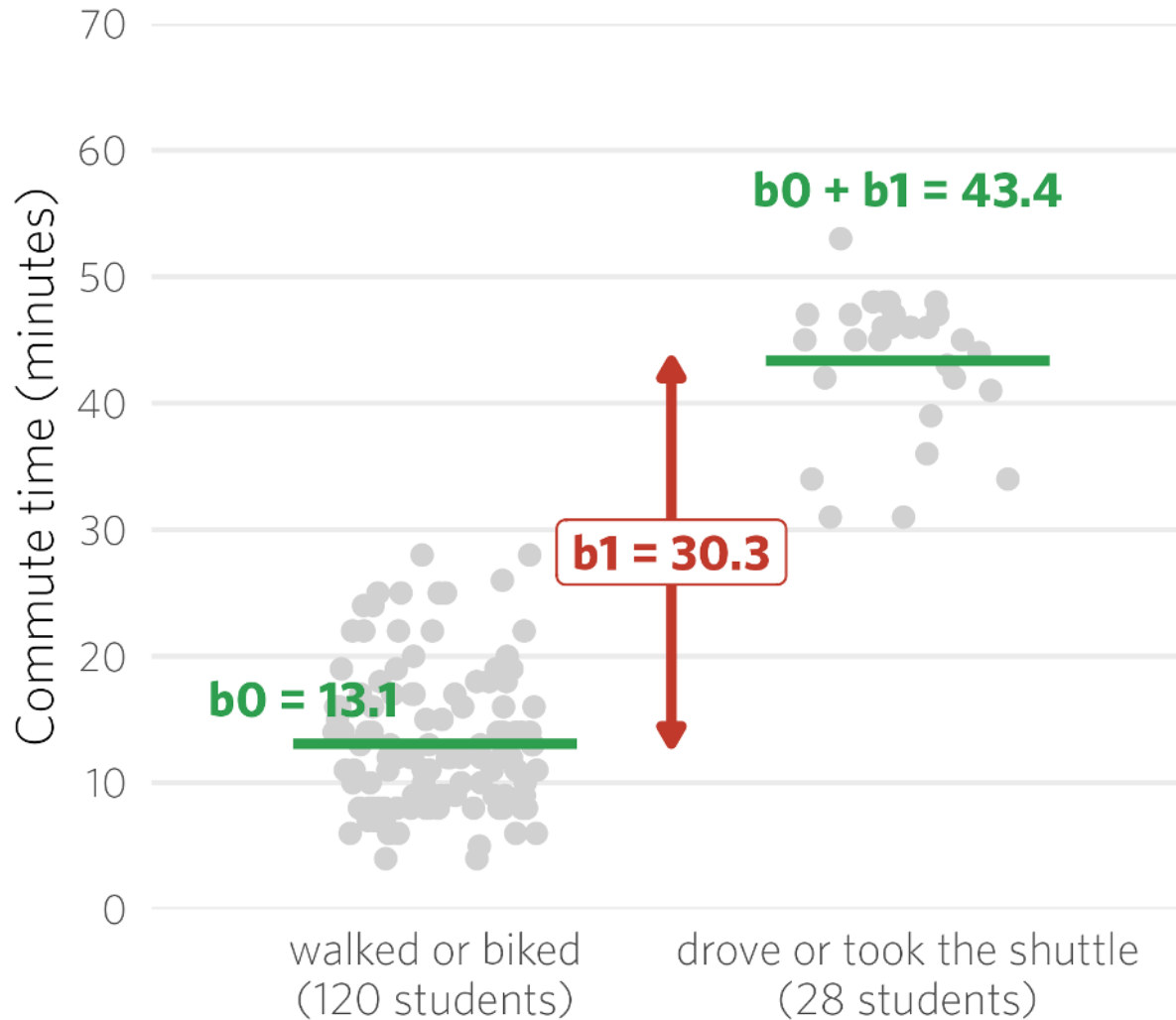


The cost of a
parameter

Monday's class ended before Part 3 of its worksheet; it is today's Part 1, on the heights of 7,182 NHANES adults.

b_0 and b_1 on the figure

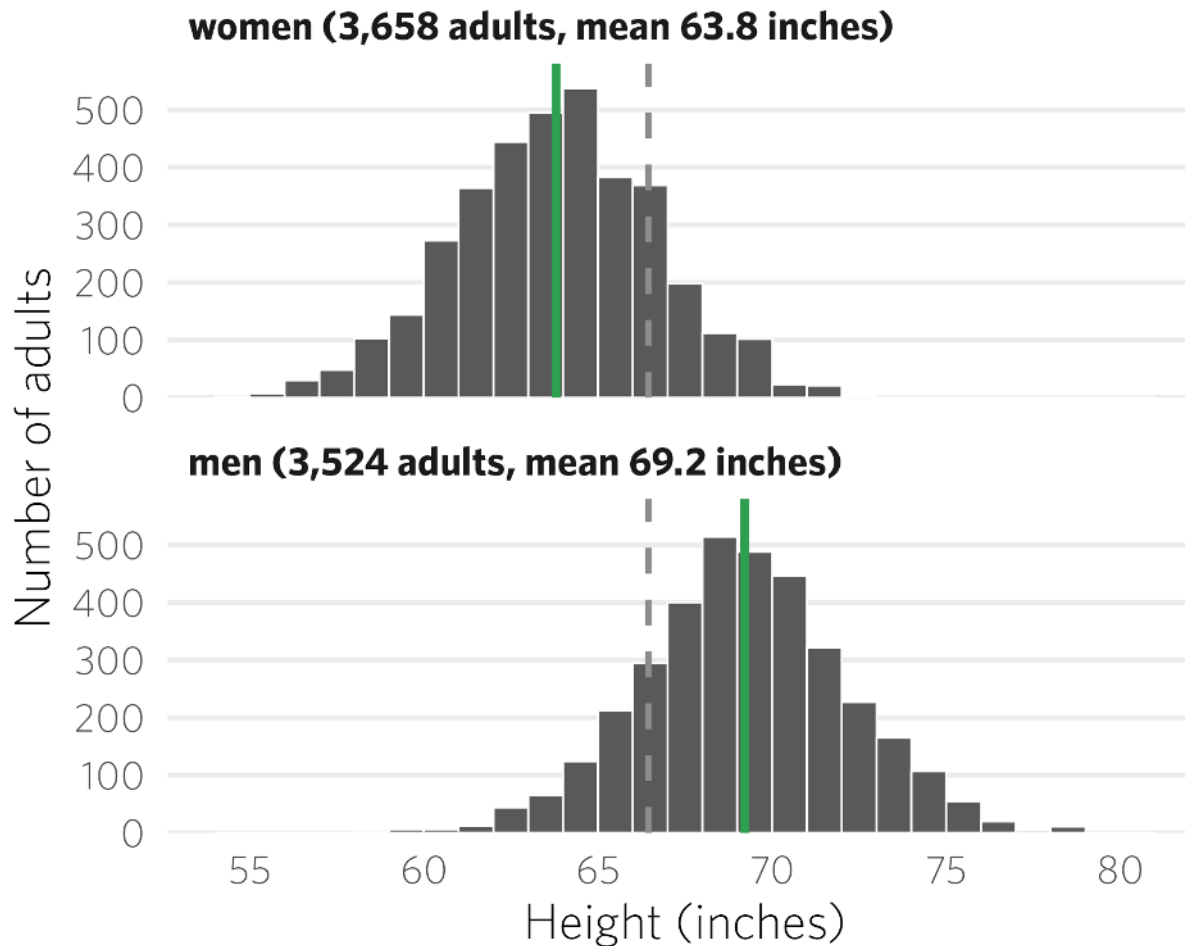
The two numbers R reports and the two group means



- Monday's model: the 148 commute times by how students got to campus.
- $b_0 = 13.1$ 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.

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 1

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 answer two ways: as the fraction $SS \text{ Model} / SS \text{ Total}$ with the two numbers filled in, and as a decimal.

Worksheet - Part 1, 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 not quite right. Say which two, and write a corrected version of each sentence.
- 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 1a-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 1c · 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 1d - 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 1e - 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.

Today

Comparing group models

A light gray circle containing the text "AI" in white.

Claude Code demo
with TA Anna Kelley

A light gray circle containing the number "1" in white.

Reading the model's
two numbers

A yellow circle with a black border containing the number "2" in white.

A model with three
means

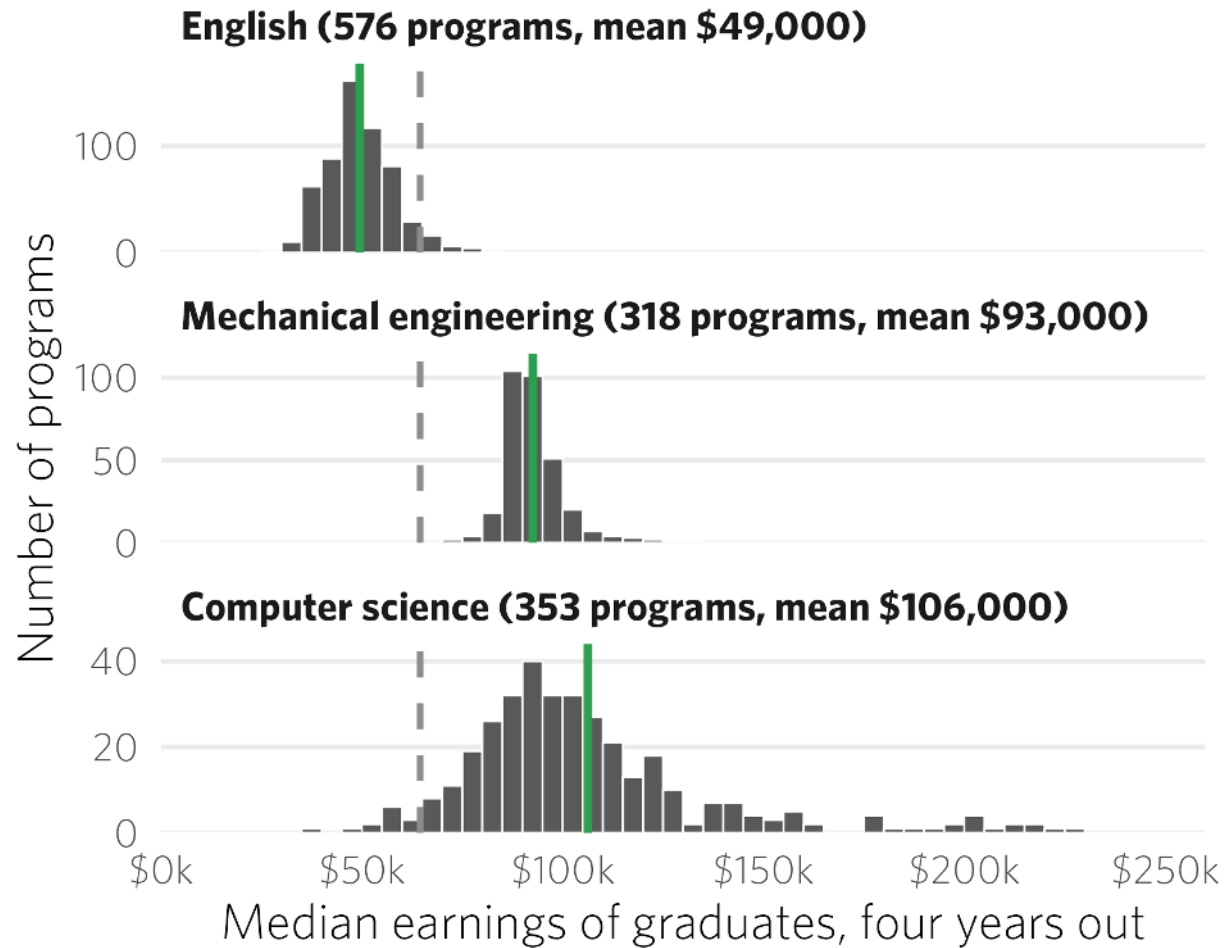
A yellow circle containing the number "3" in white.

The cost of a
parameter

The example: the three majors from the Times' guide, as a model with three means.

Three majors from the Times' guide

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



- Monday's question: how much of the variation in program earnings does the major account for?
- Three majors, three means: English \$49,000, mechanical engineering \$93,000, computer science \$106,000.

The model as R reports it

```
1 lm(earnings ~ major, data = programs)
```

Coefficients:

(Intercept)	majorMechanical Engineering	majorComputer Science
49,453	43,139	56,803

- Three numbers, for a model with three means.
- The label `English` is absent.
- **Which of the three means is 49,453? Where are the other two?**

Three means from three numbers

$$b_0 = 49,453$$

$$b_0 + b_1 = 49,453 + 43,139 = 92,592$$

$$b_0 + b_2 = 49,453 + 56,803 = 106,256$$

$$\text{PRE} = 0.65$$

- The intercept is the mean of the reference group, English, whose label is absent.
- $b_1 = 43,139$ is added to b_0 to reach the mechanical engineering mean.
- $b_2 = 56,803$ is added to b_0 too, never to b_1 .
- Proportional reduction in error (PRE) 0.65: the major accounts for 65% of the variation in earnings among these 1,247 programs.

The model in symbols

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

program's major	X_{1i} (mechanical engineering?)	X_{2i} (computer science?)	prediction
English	0	0	b_0
mechanical engineering	1	0	$b_0 + b_1$
computer science	0	1	$b_0 + b_2$

- Three groups need two 0/1 variables, not three. A program that is neither mechanical engineering nor computer science is English.
- Each added group adds one variable and one parameter.

Worksheet - Part 2

A group model of program earnings by major was fitted to 1,247 programs from the Times' college guide: English, mechanical engineering, and computer science. R reports (in dollars, rounded to the nearest thousand):

(Intercept)	49,000
majorMechanical Engineering	43,000
majorComputer Science	57,000

- a.** What is the mean earnings of the computer science programs, in dollars?
- b.** A reader computes the computer science mean as $49,000 + 43,000 + 57,000 = 149,000$. One of the three numbers belongs to a different major and should not be in the sum. Write that number and the name of its major.
- c.** R used English as the reference group. If computer science were the reference group instead, what three numbers would the output show? Write all three, with their signs.

Part 2a-c · The three means and the reference group

a. $49,000 + 57,000 = 106,000$

b. $49,000 + 43,000 = 92,000$

c. (Intercept) 106,000
majorEnglish -57,000
majorMechanical Engineering -14,000

- **a.** The computer science number is added to the intercept, the English mean.
- **b.** The 43,000, which belongs to mechanical engineering. Each number is added to the intercept, never to the other number.
- **c.** With computer science as the reference group, the other two numbers are negative. The three means are unchanged.

Today

Comparing group models



Claude Code demo
with TA Anna Kelley



Reading the model's
two numbers



A model with three
means

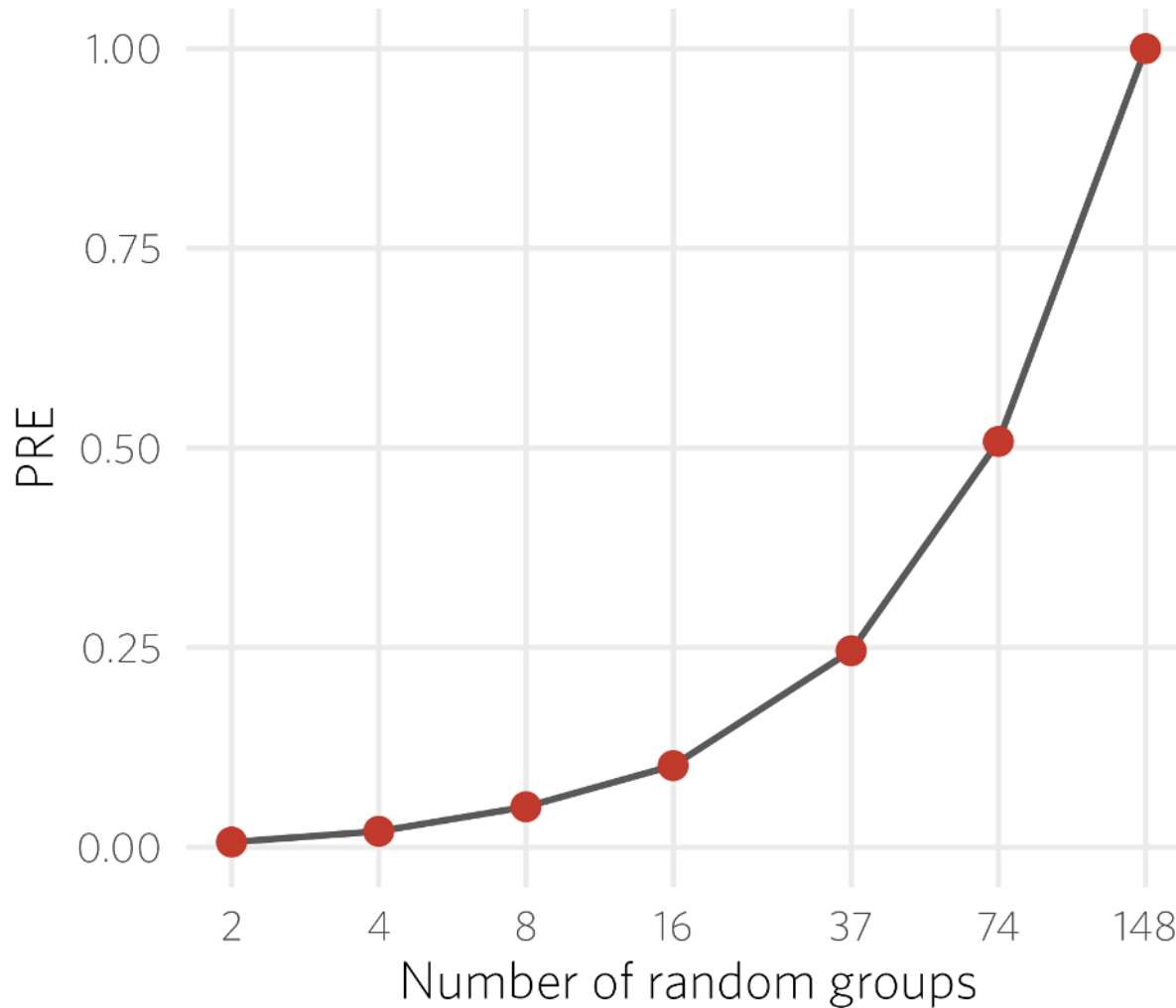


The cost of a
parameter

The example: the 148 commute times dealt into random groups, then Monday's eight students.

The error removed by random groups

PRE: the error removed grows with every group added



- Deal Monday's 148 students into groups at random, by drawing labels from a hat.
- PRE rises with every group added, however meaningless the labels.
- With 148 groups of one student each, $PRE = 1$, and the model predicts nothing for a 149th student.
- A model with one parameter per case is overfitted.

Degrees of freedom as a budget

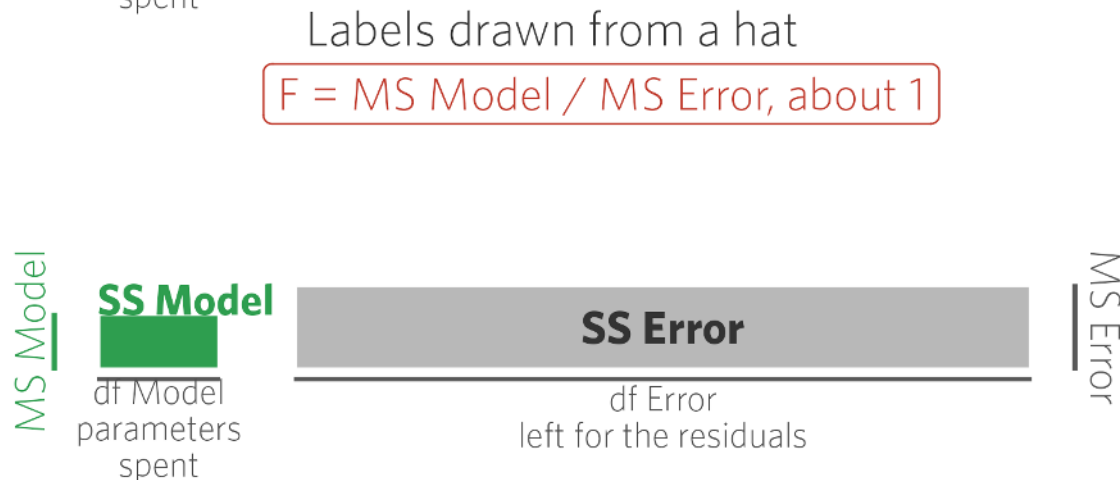
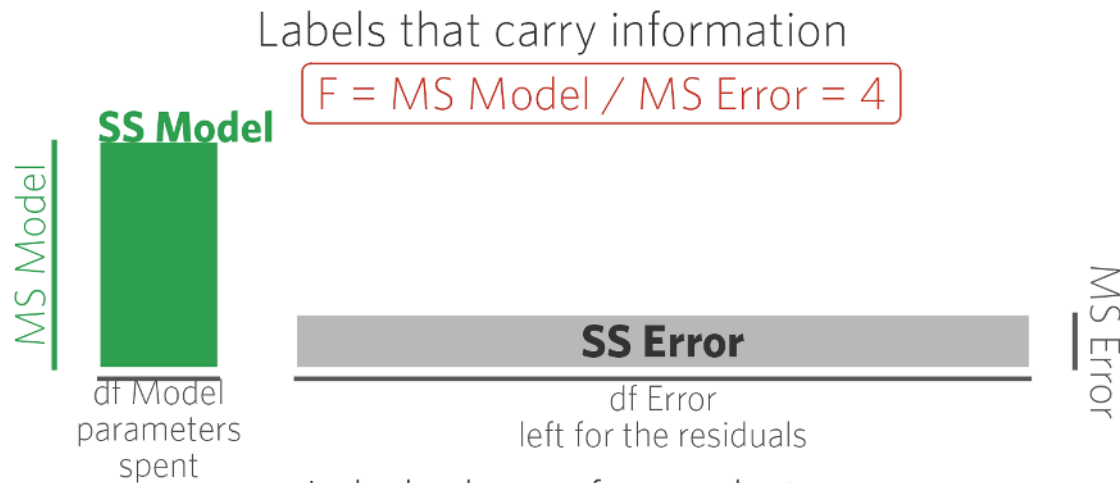
- The 148 commute times are 148 independent numbers: 148 degrees of freedom (df).
- The empty model spends one of them on the grand mean, leaving **df Total = 147**.
- **df Model** = the number of groups minus 1: one df per added group mean.
- **df Error** = the number of students minus the number of groups: the df left for the residuals.
- For the travel model: $\text{df Model} = 1$, $\text{df Error} = 146$, and $1 + 146 = 147$.

Error per degree of freedom (MS)

- **MS** is a sum of squares divided by its degrees of freedom: the error per degree of freedom.
- **MS Error** = $4,528 / 146 = 31.0$ square minutes: the variance of the residuals.
- **MS Model** = $20,807 / 1 = 20,807$: the error removed per parameter spent.

What the F ratio compares

Width: degrees of freedom. Height: error per degree of freedom (MS).
Area: a sum of squares. Both pictures have the same total area,
SS Total: the same data, labelled two ways.

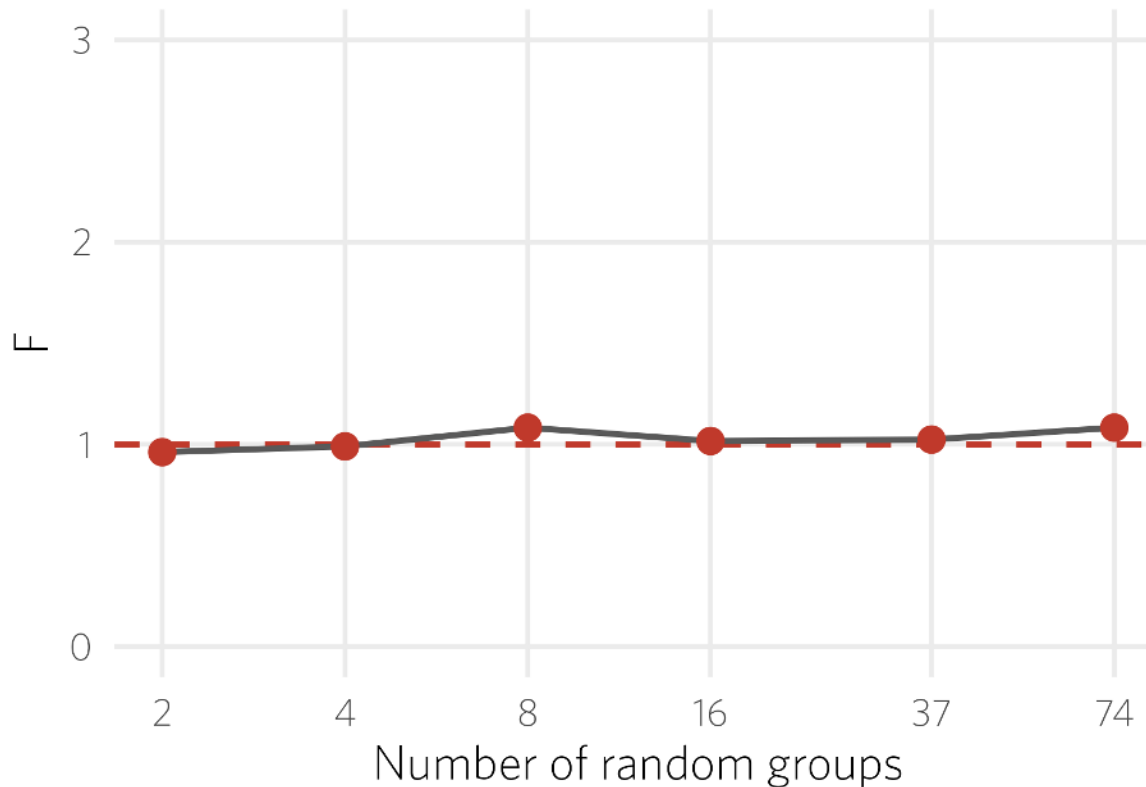


- Each block is a sum of squares: its width the degrees of freedom, its height the error per degree of freedom, MS.
- Green: the error removed per parameter spent. Grey: the error left per degree of freedom the residuals keep.
- F is the green height divided by the grey height.

The F ratio

$$F = \frac{MS_{\text{Model}}}{MS_{\text{Error}}} = \frac{\text{error removed per parameter spent}}{\text{error remaining per degree of freedom left}}$$

F stays near 1 whatever the number of random groups



- Random labels: a parameter removes about as much error as remains per degree of freedom, so F is near 1.
- A large F: the parameters removed far more than their share.

The table for how students got to campus

	SS	df	MS	F	PRE
Model (how the student got to campus)	20,807	1	20,807	671	0.82
Error	4,528	146	31.0		
Total	25,335	147			

- $F = 671$: the one parameter removed 671 times as much error as remains per degree of freedom.
- R's table has one more column, p , the subject of Week 7.

Worksheet - Part 3

Monday's eight students, four walkers and four shuttle riders: **SS Model = 128, SS Error = 64, SS Total = 192.** Fill in the table on your sheet as you go.

- a. Fill in the df column: df Model, df Error, and df Total. (The empty model spends one of the eight degrees of freedom; each additional group mean spends one more.)
- b. Fill in the MS column: MS Model and MS Error, each SS divided by its df. Round to one decimal.
- c. Compute $F = MS \text{ Model} / MS \text{ Error}$ and write it in the table, rounded to a whole number.

	SS	df	MS	F	PRE
Model (walked or took the shuttle)	128				
Error	64				
Total	192				

Worksheet - Part 3, continued

d. Now suppose each of the eight students is in a group of their own, so the model has eight means. Write that model's SS Error, PRE, and df Error, one number each.

e. A ninth student from the same section is about to report a commute time. Which model gives a usable prediction for that student: the two-group model or the eight-group model? Write your choice and one sentence of reason.

Part 3a-c · The table for the eight students

a. $df = 1, 6, 7$

b. $MS_{\text{Model}} = 128/1 = 128$

$$MS_{\text{Error}} = 64/6 \approx 10.7$$

c. $F = 128/10.7 \approx 12$

- **a.** $df_{\text{Model}} = 1$, one mean beyond the empty model's. $df_{\text{Error}} = 8 - 2 = 6$. $df_{\text{Total}} = 8 - 1 = 7$.
- **b.** $MS_{\text{Error}} = 10.7$ square minutes: the variance of the eight residuals.
- **c.** $F = 12$: the one parameter removed twelve times as much error as remains per degree of freedom.

Part 3d-e · One mean per student

d. $SS_{\text{Error}} = 0$

$$PRE = 192/192 = 1$$

$$df_{\text{Error}} = 8 - 8 = 0$$

e. $F = \frac{192/7}{0/0}$ undefined

- **d.** Every prediction equals the student's own time: no error left, $PRE = 1$, and no degrees of freedom left.
- **e.** The two-group model. The eight-group model has eight means, and nothing says which one a ninth student gets.
- With $df_{\text{Error}} = 0$ there is no estimate of the error around a prediction, and F cannot be computed.
- A perfect fit to these eight students is not a model of the ninth.

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!