Psych 10: Intro to Statistics

Review session · Wed Nov 11, 2026

Prof. Judith Fan · Fall 2026

Reminders & Announcements

- **Attendance slip:** please write your name, SUNet ID and section on the slip now. There is no worksheet today.
- **Much of today's class is office hours** with the teaching team, here in the lecture hall.
- The **answer key for Monday's practice exam** is on the Resources page of the course website, next to the formula sheet.
- The **exam** is next Wednesday, November 18, in class. It covers CourseKata Chapters 1 to 12, and you will have the formula sheet.
- **Friday, 1:30 to 2:50 pm**: TA office hours in your section's tutorial room, in place of the tutorial. Drop in with your exam questions; attendance credit is automatic.
- CourseKata **Chapter 12** is due Friday.
- Section this week: **Milestone 7**, due Sunday, November 15.

How to prepare for the exam

In order of priority:

1. **Monday's practice exam and its key**, on the Resources page: the closest thing to the real exam. Work it a second time without the key, then compare your reasoning with the key's; every mismatch is a place to focus your studying.
2. **The pre-practice questions and their key**, on the same page.
3. **Your lecture worksheets and the Friday tutorials**, starting with the ones that gave you the most trouble.
4. **The formula sheet**, which you will have during the exam: get to know it before exam day.
5. **CourseKata Chapters 1 to 12**, especially the sections the worksheets and tutorials kept returning to.

Where to get help: today, Friday's office hours, and a word with me after class.

Your question for today

For much of today's class, we'll break out into office hours with the teaching team.

On your attendance slip, please write one question you'd like to bring to office hours.

Be as specific as you can about the concept or the practice-exam item.

With your neighbor

Compare your question with your neighbors' questions.

Does anyone near you have a similar question?

Some topics you mentioned in advance

Here were some of the topics that were submitted through the form:

- **The sum of squares, degrees of freedom, MS, the MAD and the F ratio**: the meaning of each one, and why the table reports one F and two MS values but no MS Total. **PRE**, and how it differs from F.
- **Choosing**: mean or median; which graph; which statistic (SS, variance, SD, PRE, F); categorical or quantitative.
- **How the concepts and formulas build on one another**: the SD, z-scores and shape; PRE, the partition of SS and the data-generating process.

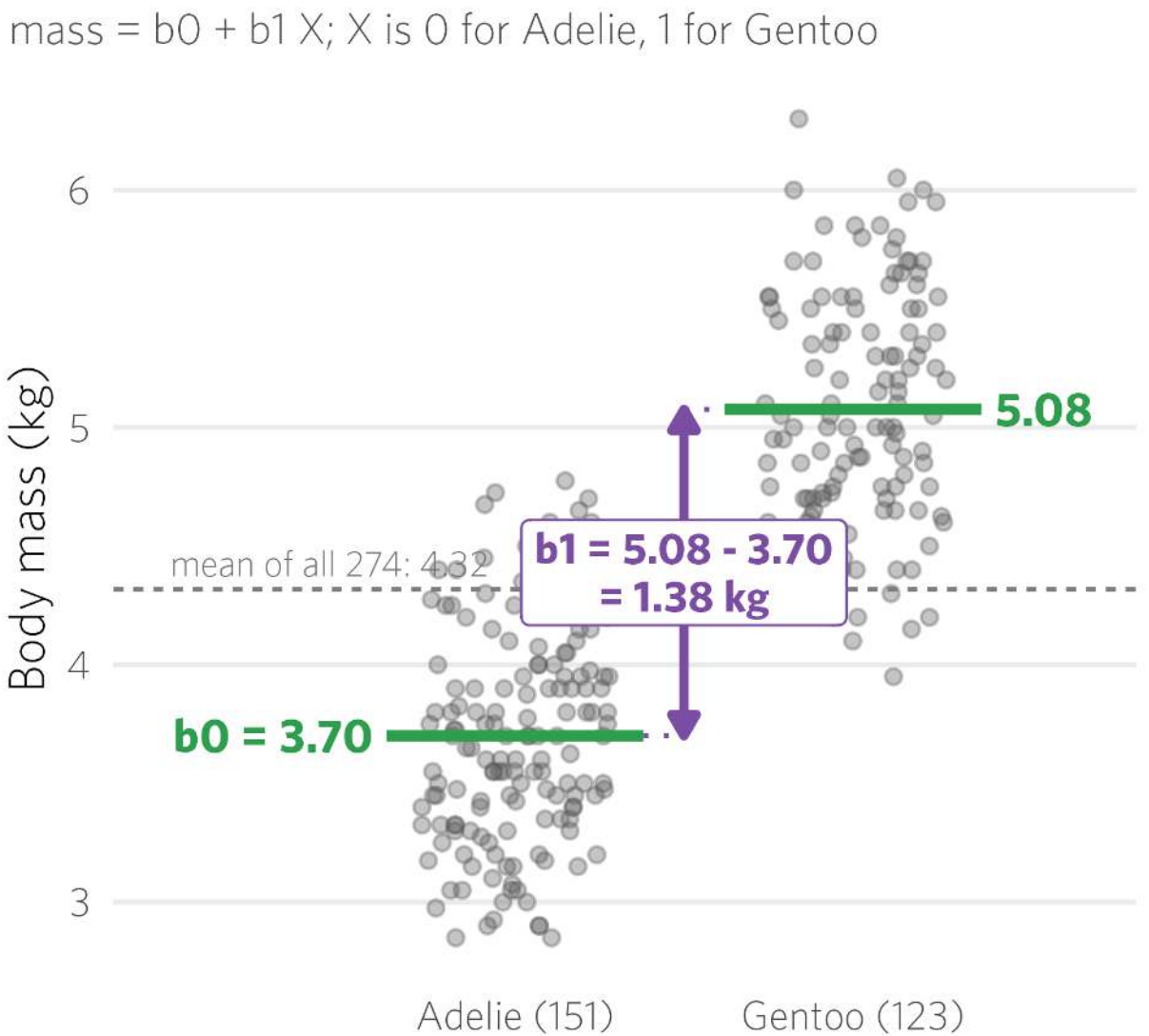
Today

Review session

1. The quantities in the ANOVA table, and the question each one answers
2. Choosing a summary, a graph, and a statistic
3. How the pieces fit together: shuffling and the p-value

Q. Office hours with the teaching team

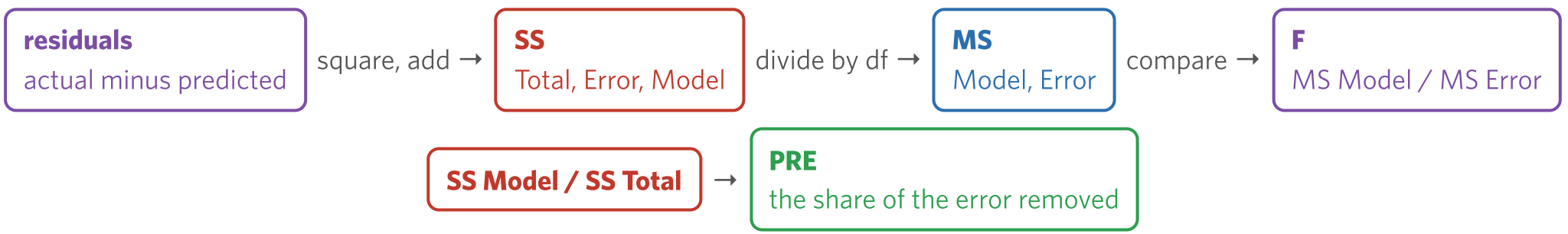
The species model on the 274 penguins



- Researchers at Palmer Station weighed 274 Adelie and Gentoo penguins.
- The species model predicts each penguin's mass with its species mean.
- **With your neighbor:** which quantity on this figure measures the distance between one penguin's mass and its predicted mass?

Every quantity in the table is built from the residuals

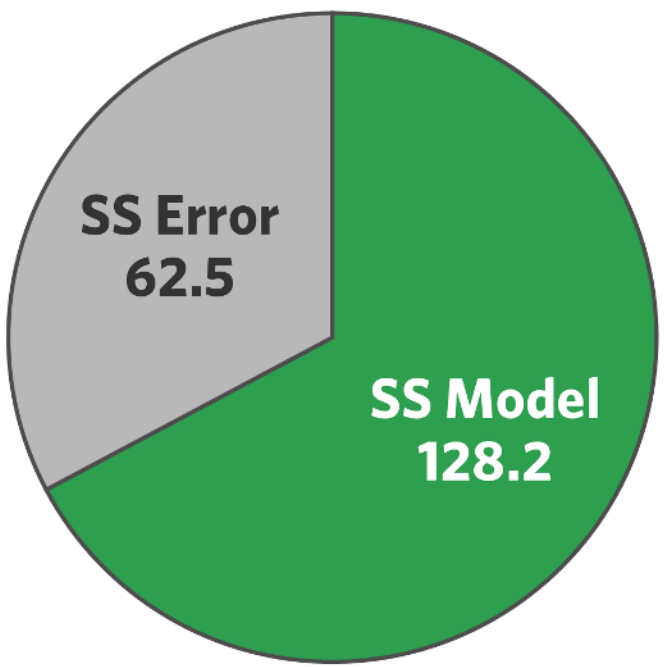
$Y_i = \hat{Y_i} + e_i$ each penguin's mass = the model's prediction for it + its residual



- Each box answers one question about the model.

How SS Total splits into SS Model and SS Error

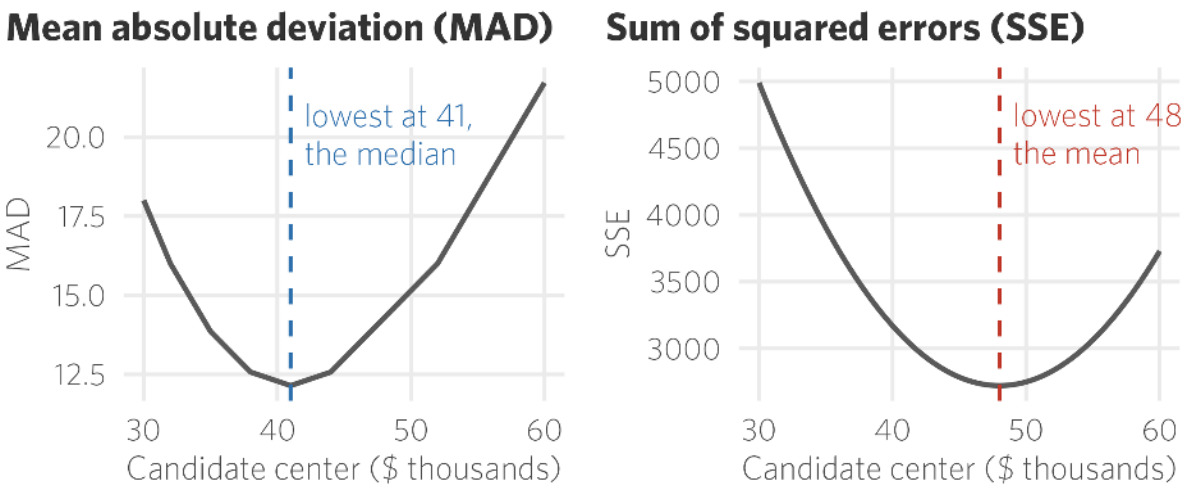
The species model's partition of SS Total
The whole circle is SS Total = 190.8, the empty model's error.
Green: the error the model removed. Grey: the error left.



PRE = SS Model / SS Total = 128.2 / 190.8 = 0.67

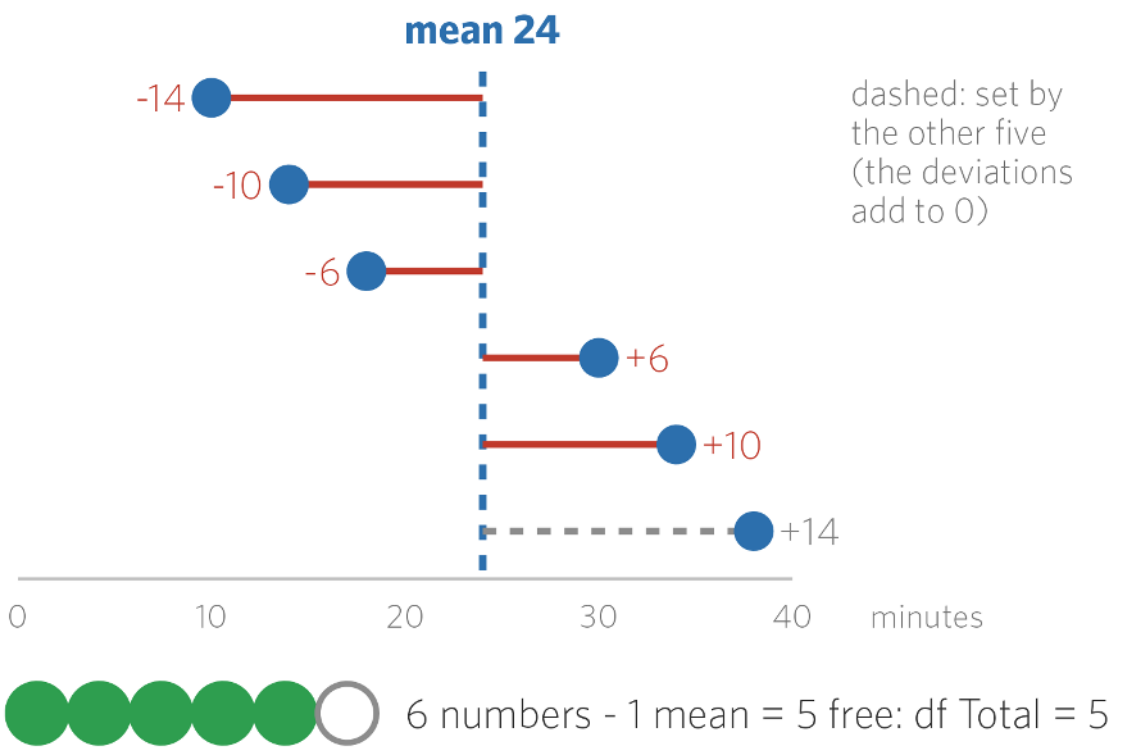
- **SS Total** is the empty model's error: every penguin's squared distance from the one mean, the whole of the error.
- **SS Error** is the species model's error: the squared residuals from the species means, the remaining error.
- **SS Model** = SS Total - SS Error: the error removed by the model.
- SS Total = SS Model + SS Error, for every model.

Why the table is built from squares rather than from the MAD



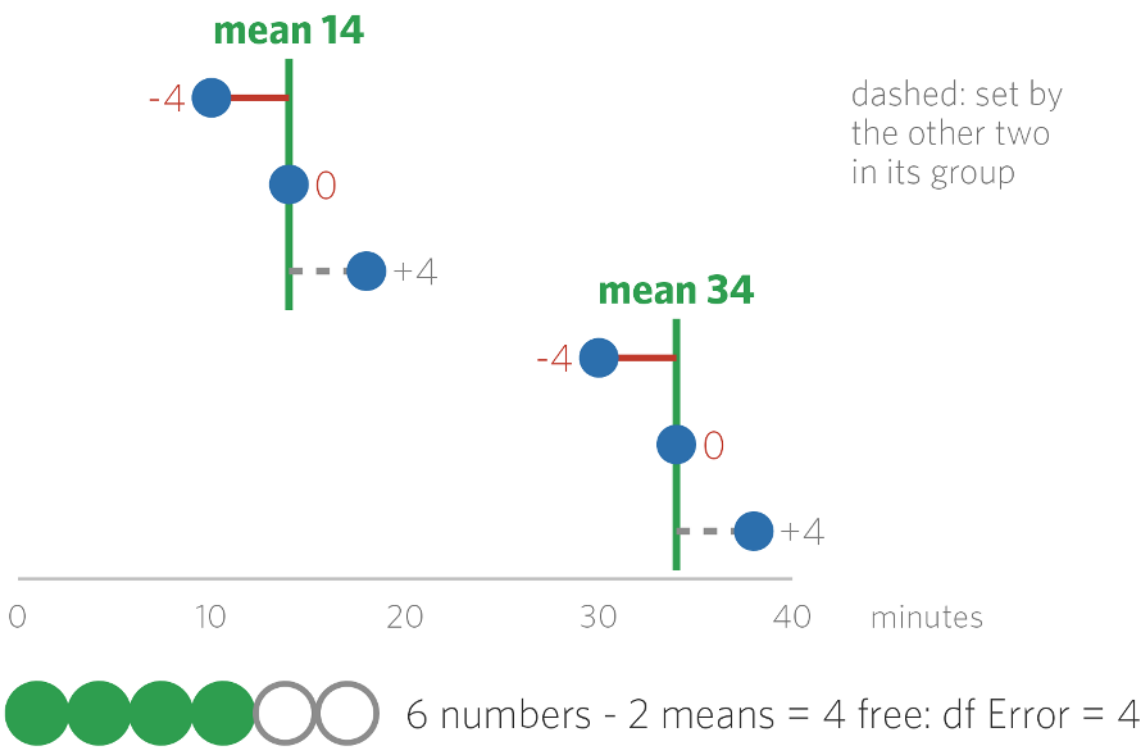
- Both curves measure a one-number model's total error. SAD adds up the distances from the prediction; SS adds up their squares; the MAD is SAD divided by n.
- The median gives the smallest SAD; the mean gives the smallest SS.
- Squared errors split into parts: SS Total = SS Model + SS Error holds for squares and not for distances, so the table, PRE and F all rest on squares.

The numbers left free after the mean is fixed: df Total



- Six commute times are six free numbers: 6 degrees of freedom.
- Fix their mean at 24. The first five deviations from the mean can take any values.
- The sixth is set by the other five, because the deviations add to zero: **df Total = n - 1 = 5**.

After two group means are fixed: df Error and df Model



- Fix the two group means, 14 and 34; two deviations in each group are free.

- The third in each group is set by the other two: **df Error = n - parameters = 4**. The second mean is one parameter more than the empty model's: **df Model = 1**.

- For the 274 penguins and two species means: df Total 273, df Model 1, df Error 272.

Error per degree of freedom: MS

The species model

	SS	df	MS	F	PRE
Model	128.2	1	128.2	558	0.67
Error	62.5	272	0.230		
Total	190.8	273	0.699		

- **MS Error** = SS Error / df Error: the variance of the residuals. Its square root, 0.48 kg, is the typical residual.
- **MS Model** = SS Model / df Model: the error removed per parameter spent.
- Dividing by df makes a total from 272 residuals comparable with a total from one parameter.

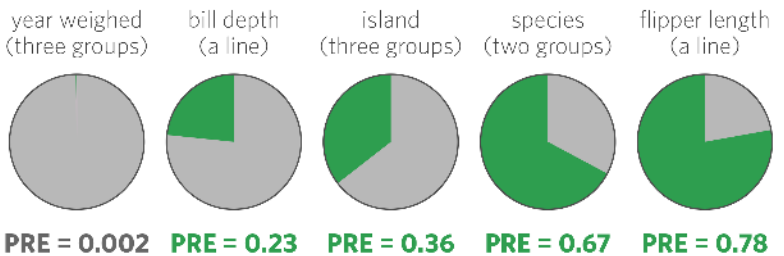
Why the table has one F and one PRE

		one number per row			*one number per table*
	SS	**df**	**MS**	**F**	**PRE**
Model	128.2	1	128.2	558	0.67
Error	62.5	272	0.230		
Total	190.8	273	0.699		

- **SS, df and MS describe one row each**: a sum of squares, its degrees of freedom, and their ratio. MS Total, 0.699 kg^2, is the variance of the 274 masses; its square root is the SD, 0.84 kg.
- **F and PRE each compare two rows**, so each is one number for the whole table: F = MS Model / MS Error, and PRE = SS Model / SS Total. Both sit in the Model row: both describe the model's improvement.
- The df column, read in reverse, gives the number of cases, df Total + 1, and the number of parameters beyond the mean, df Model.

PRE and F answer two different questions

Five models of body mass on the same 274 penguins. Green: SS Model. Grey: SS Error.



$$F = \frac{PRE \,/\, \mathrm{df}_{\mathrm{Model}}}{(1 - PRE) \,/\, \mathrm{df}_{\mathrm{Error}}}$$

- **PRE** is the share of SS Total a model removes: 0 if the species means equaled the overall mean, 1 if every penguin sat at its species mean.
- **F** takes the degrees of freedom into account: error removed per parameter, against error left per degree of freedom, so the same PRE gives a larger F from more penguins.

What the table says about the penguins

	SS	df	MS	F	PRE
Model	128.2	1	128.2	558	0.67
Error	62.5	272	0.230		
Total	190.8	273	0.699		

- **PRE = 0.67** says that species accounts for 67% of the variation in mass among these 274 penguins: a description of this sample.
- **SS Error fell** from 190.8 to 62.5, but it falls for any grouping, even labels dealt from a hat. The evidence for a difference between the species comes from **F = 558** and its p-value, not from the drop alone.
- **A species mean**, 5.08 kg for the Gentoos, is the model's prediction, not any one penguin's mass.

Today

Review session

1. The quantities in the ANOVA table, and the question each one answers
2. Choosing a summary, a graph, and a statistic
3. How the pieces fit together: shuffling and the p-value

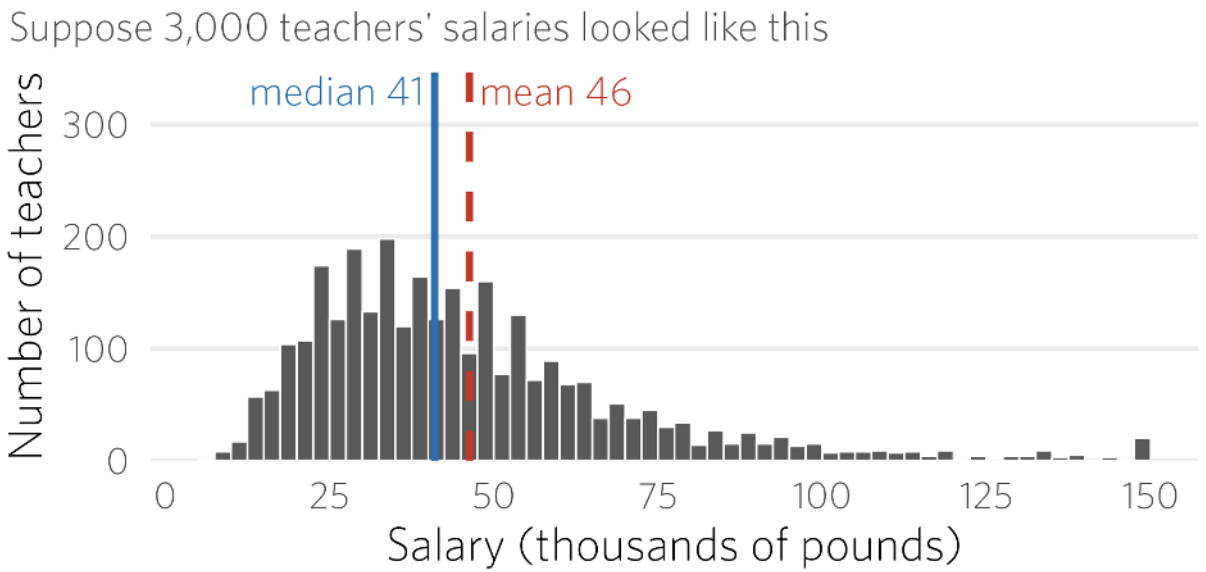
Q. Office hours with the teaching team

Is the variable categorical or quantitative?

Variable of a penguin	Kind	Why
species: Adelie or Gentoo	categorical	the values are labels with no order
body mass in kg; flipper length in mm	quantitative	a larger value means more of the quantity, so a mean of the values has a meaning
the number on its leg band	categorical	a label written as a number: band 60 is not more of anything than band 20
the year it was weighed: 2007, 2008, 2009	either	as a quantity when the question is about a trend across years (a line); as three groups when the question is whether the years differ

- The test: does a larger value mean a larger amount of the quantity, so that a mean of the values has a meaning? A number can be a label.
- Some variables can be treated either way, and the question being asked settles the treatment.

When to report the mean and when to report the median



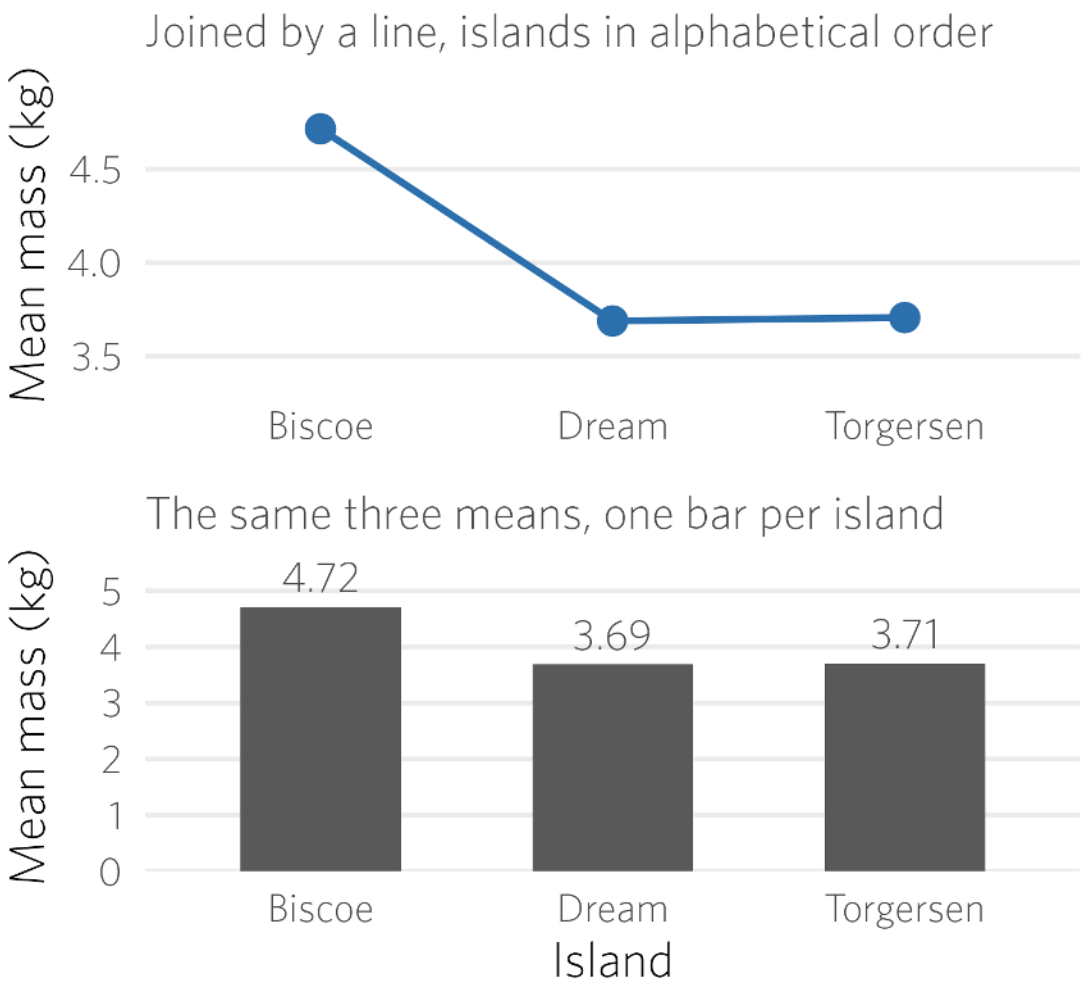
- The **mean** is the balance point and the total divided by n. Report it when the total matters, and use it in every model of this course, because it has the smallest SS.
- The **median** is the middle case. Report it for a typical case when the distribution has a long tail, as here, where the tail of high salaries pulls the mean (46) above the median (41).
- When the two differ, report both; the gap between them describes the skew.

The graph for each kind of variable

The variables	The graph	What it shows
one categorical	a bar chart of counts or proportions	how many cases in each category
one quantitative	a histogram, or a box plot	the shape, center and spread; the five-number summary and outliers
a quantitative outcome by a categorical variable	box plots side by side, or every case with the group means marked	the difference between the groups and the overlap within each group
two quantitative variables	a scatterplot, with the line when a line is the model	the direction, strength and form of the relationship
a quantity over time	points joined by a line, only because time has an order	the trend

- A chart of two bars, one per group mean, omits the spread within each group.
- A line joining categories that have no order suggests a trend that is not in the data.

A slope on a label and a line through labels



- A regression of mass on island number gives b_1 = -0.58 kg per island with Biscoe 1, Dream 2, Torgersen 3, and -0.04 with Torgersen 1, Biscoe 2, Dream 3: the same penguins, a different number. A label has no slope; its model is the island model, with three means.
- The top chart's line suggests that mass falls from Biscoe to Torgersen, but the islands have no order. The bars below suggest no order.

The question behind each statistic

Statistic	What it measures	Units	Grows with n?	Use it to
SS	a model's total squared error	squared units	yes	compare two models fit to the same data
variance	the spread of the values: squared error per degree of freedom	squared units	no	measure spread; the SD, its square root, carries the same information in the outcome's units
SD	the typical residual	the outcome's units	no	describe the spread of one variable, or of a model's residuals; compute z-scores
MAD	the typical distance from the mean	the outcome's units	no	describe spread when the median is the center
PRE	the share of the error a model removed	none, 0 to 1	no	say how much of the variation a model accounts for, in this sample
F	error removed per parameter, over error left per degree of freedom	none	yes	judge whether the improvement exceeds the improvement from chance

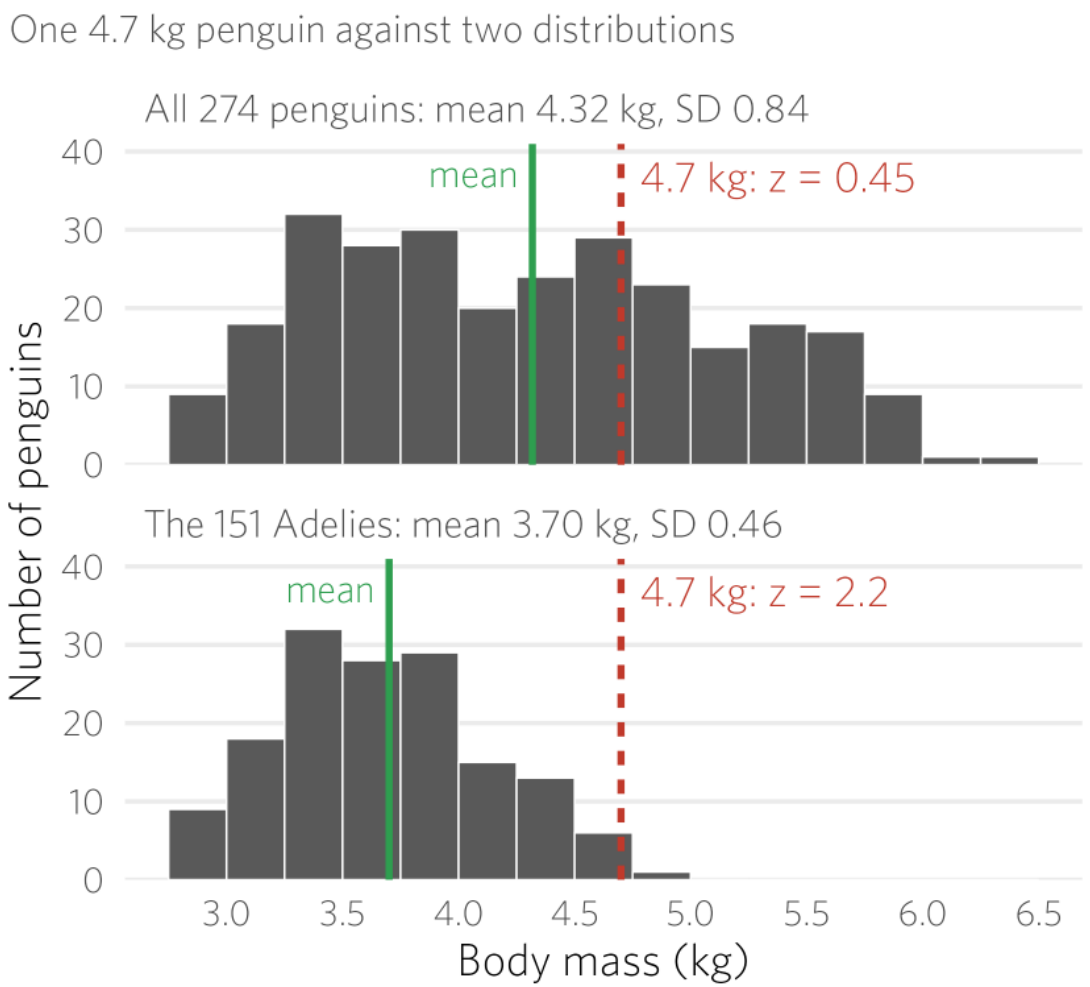
Today

Review session

1. The quantities in the ANOVA table, and the question each one answers
2. Choosing a summary, a graph, and a statistic
3. How the pieces fit together: shuffling and the p-value

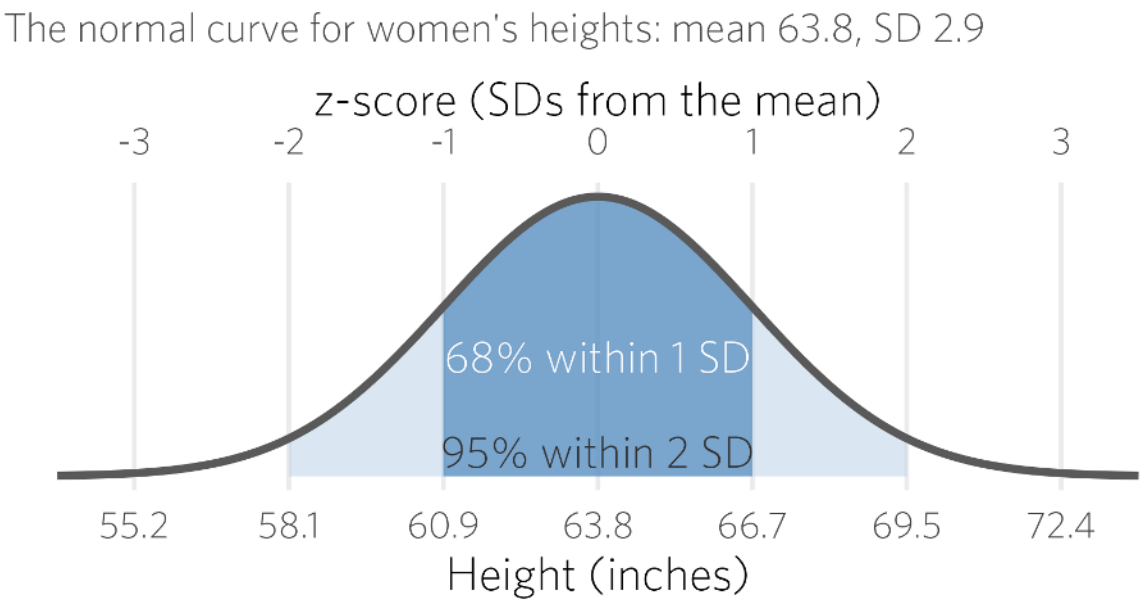
Q. Office hours with the teaching team

How unusual is a 4.7 kg penguin?



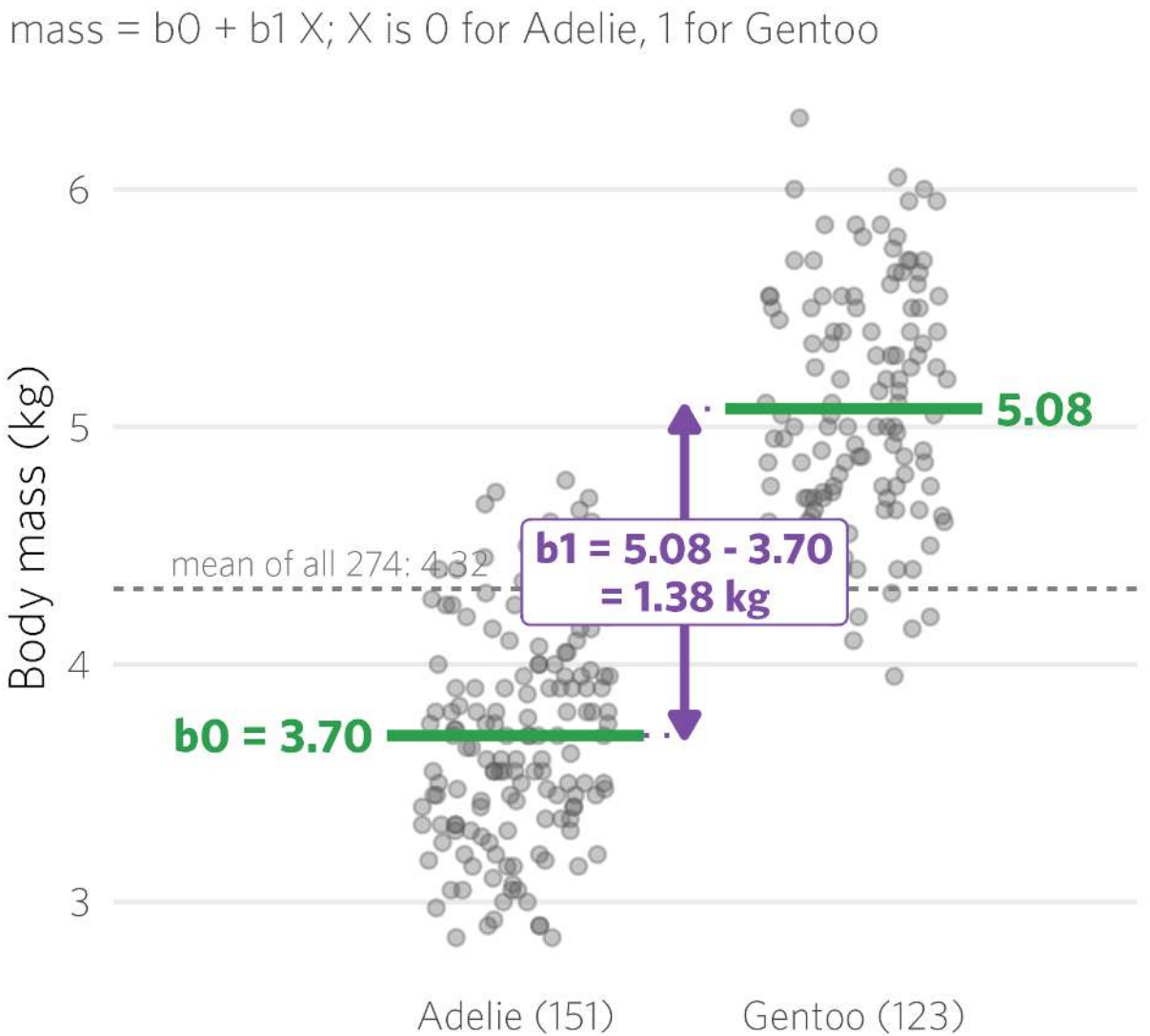
- Among all 274, the **SD** of 0.84 kg is the typical distance from the mean, and the **z-score** (4.7 - 4.32) / 0.84 = 0.45 makes this a fairly ordinary mass.
- It is an Adelie. Among the 151 Adelies, mean 3.70 and SD 0.46, its z-score is 2.2, an unusually large mass for an Adelie.
- A z-score answers "compared with what?": choose the group the question is about.

The 68% and 95% of the normal distribution



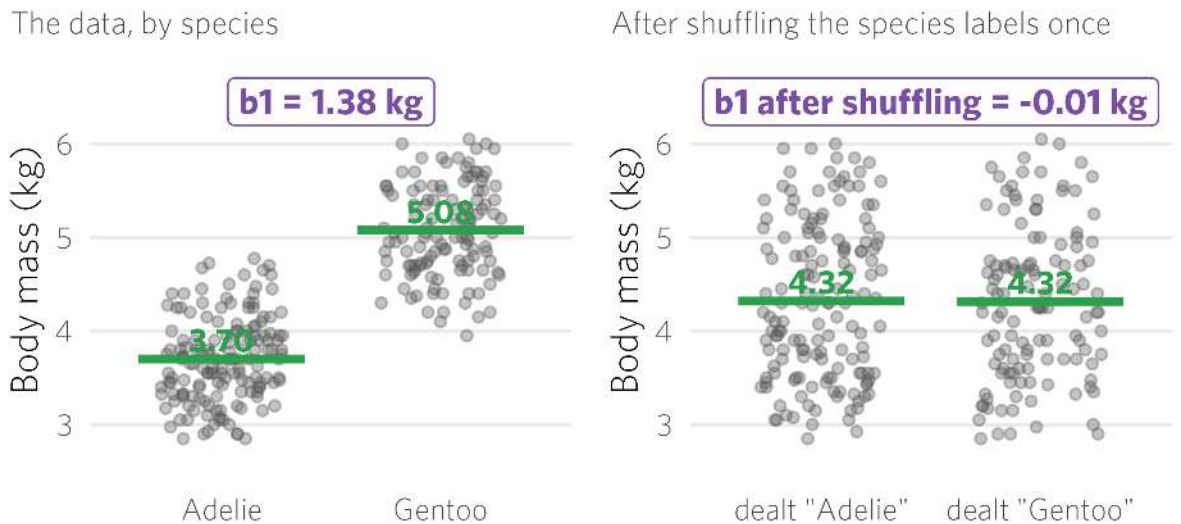
- z-scoring moves the center to 0 and the SD to 1 and leaves the shape unchanged: two humps stay two humps.
- For one shape, the normal, a z-score converts into a share of the cases: about 68% within 1 SD, about 95% within 2.
- Check the histogram first: the rule does not apply to the two-humped masses. It applies well to the means of many random samples, for almost any shape of data.

Could chance produce a 1.38 kg difference?



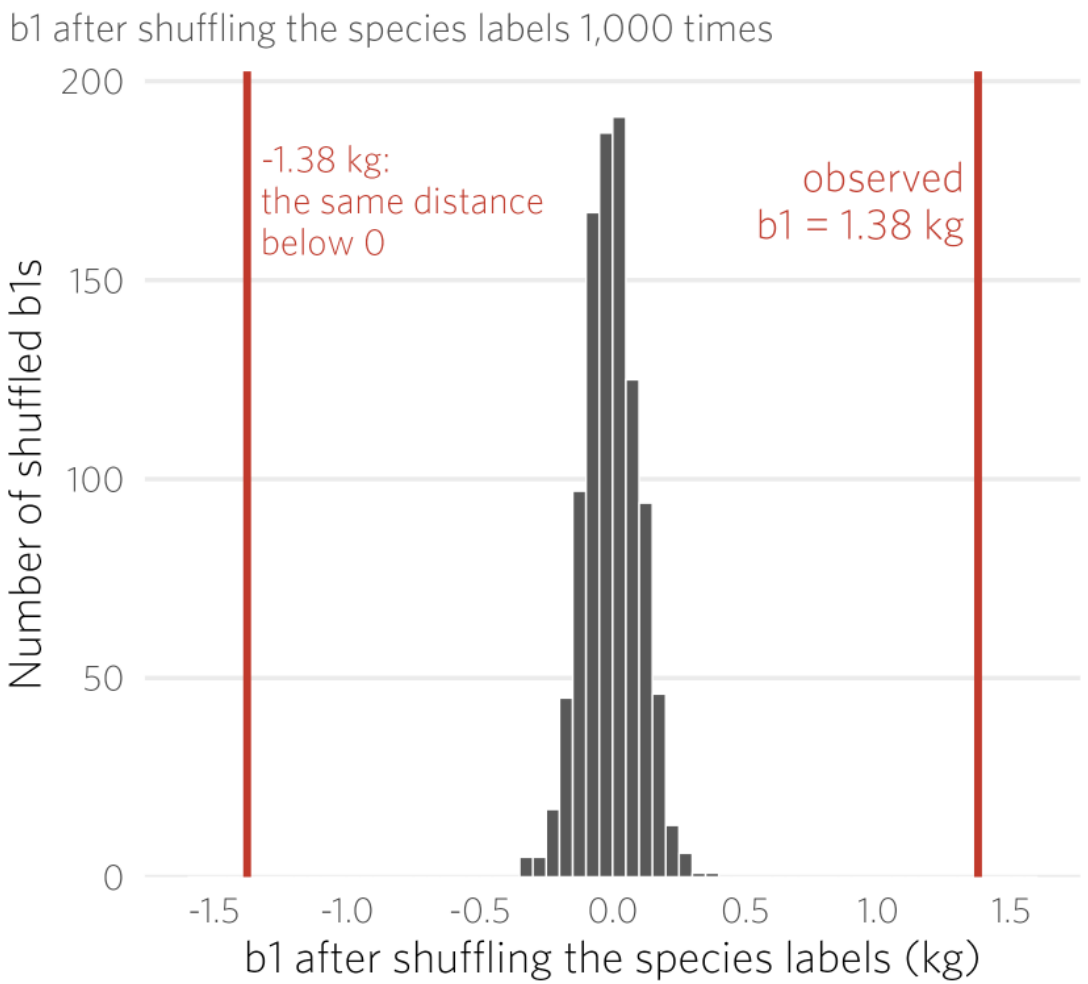
- The species model's b_1 = 1.38 kg was computed from these 274 penguins. Another 274 would give a different b_1.
- The question for inference: could a process in which species makes no difference to mass, $\beta_1 = 0$, produce a b_1 this large by chance?
- To answer it we need to know what b_1 values chance alone would produce. No such penguins exist to weigh, so we approximate them from these data, by shuffling.

What one shuffle does to the data



- Deal the 151 Adelie and 123 Gentoo labels out at random, and keep every mass as recorded. After the deal, species carries no information about mass.
- The difference between the two new group means, here -0.01 kg, is a b_1 produced by chance alone.
- One deal gives one such b_1. To judge the observed 1.38 kg we need the full range of b_1 values from chance alone, so we repeat the deal many times.

A thousand shuffles of the species labels

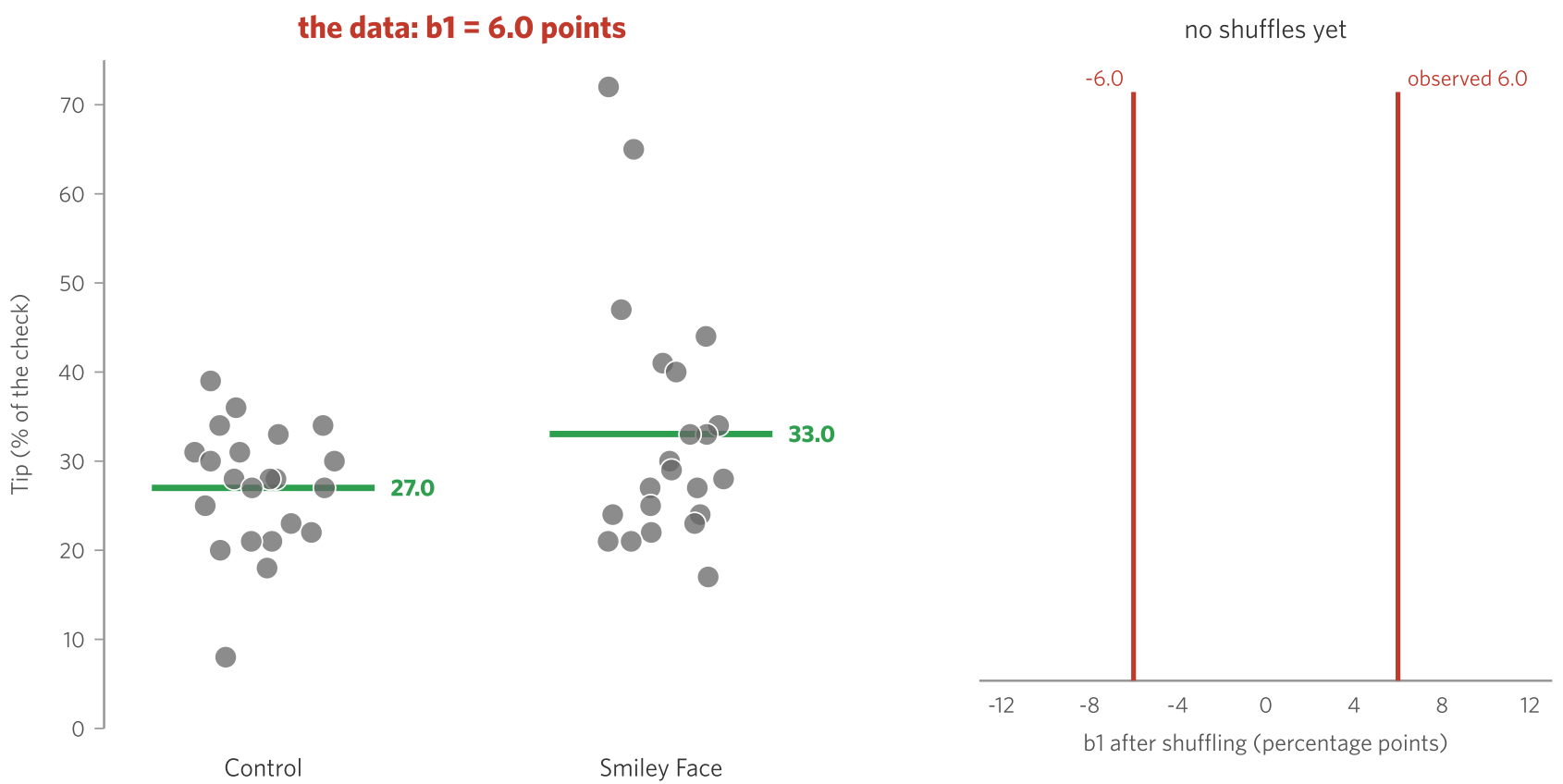


- 1,000 deals give 1,000 b_1 values from chance alone, all between about -0.4 and 0.4 kg.
- The observed 1.38 kg lies well outside them: in 1,000 tries, chance alone did not once produce a difference this large.
- The sample's b_1 estimates the process's β_1; the shuffles approximate the b_1 values a process with $\beta_1 = 0$ would produce. PRE = 0.67 describes these 274 penguins; the shuffles are about the process behind them.

Shuffling the 44 tips, one deal at a time

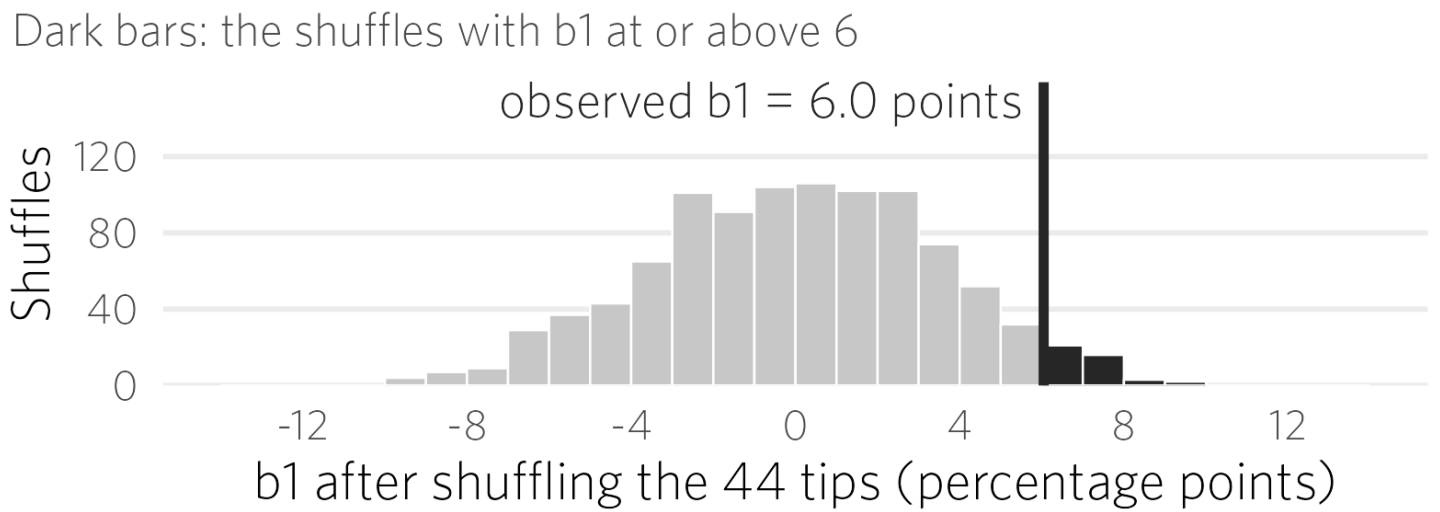
A second experiment, with a less clear answer: 22 of 44 tables, chosen by coin flip, got a smiley face on the check and tipped 6.0 points more. Could chance alone produce 6.0 points?

Reset to the data | Shuffle once | Shuffle 10 more | Shuffle to 1,000



Each dot is one table's tip. A shuffle deals the 44 labels out again at random; every tip stays at its height. Each shuffle adds its b_1 to the histogram.

A thousand shuffles of the 44 tips

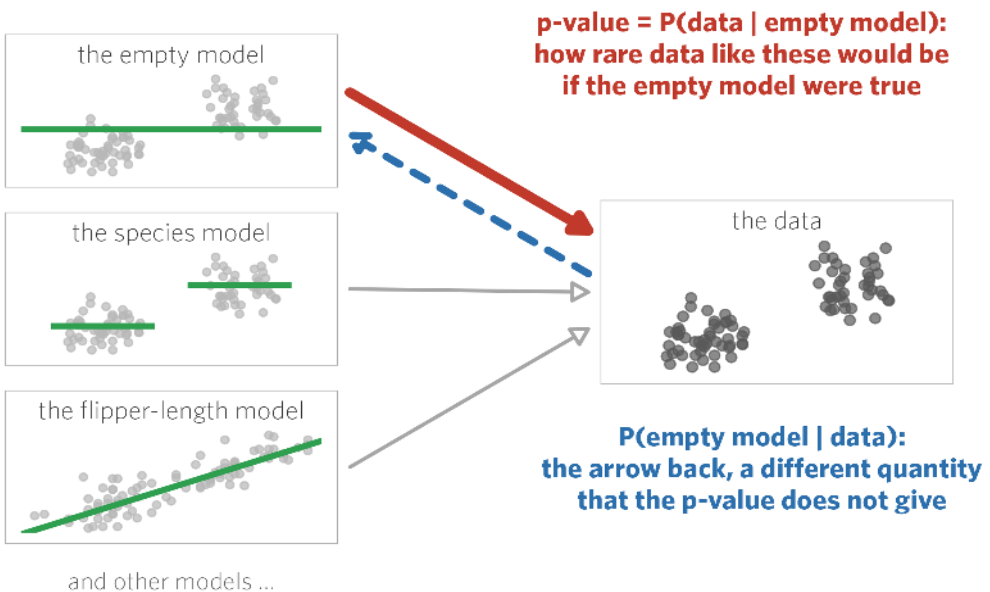


- 1,000 shuffles of the same 44 tips, with the observed b_1 = 6.0 marked. This time the observed value sits inside the distribution, near its upper edge.
- About 4 in 100 shuffles reached 6.0 or more (the dark bars), and about 9 in 100 fell at least 6.0 from 0 in either direction. So chance alone would produce a difference this far from 0 about 9 times in 100: p = .09.

The p-value and the 5% line

- The **p-value**: if the empty model were true, the probability of a b_1 at least as far from 0 as the observed one, in either direction. For the smiley face the shuffles put it at about .09. For the penguins' species difference it is below .001, since no shuffle came near 1.38.
- By a long-standing convention, a result that chance would produce in fewer than 5 of 100 tries is treated as unusual enough to deserve a closer look: the line is at .05. The line is a convenient cutoff from the 1920s, not a fixed rule.
- Below the line the empty model is **rejected**, as for the species difference; at or above it the empty model is **kept** for now, as for the smiley face: chance remains a plausible account of the data, which is not the same as evidence of no effect.

The p-value is about the data, not the empty model



The p-value is $P(\text{data} \mid \text{empty model})$, the red arrow: how rare data like these would be if the empty model were true.

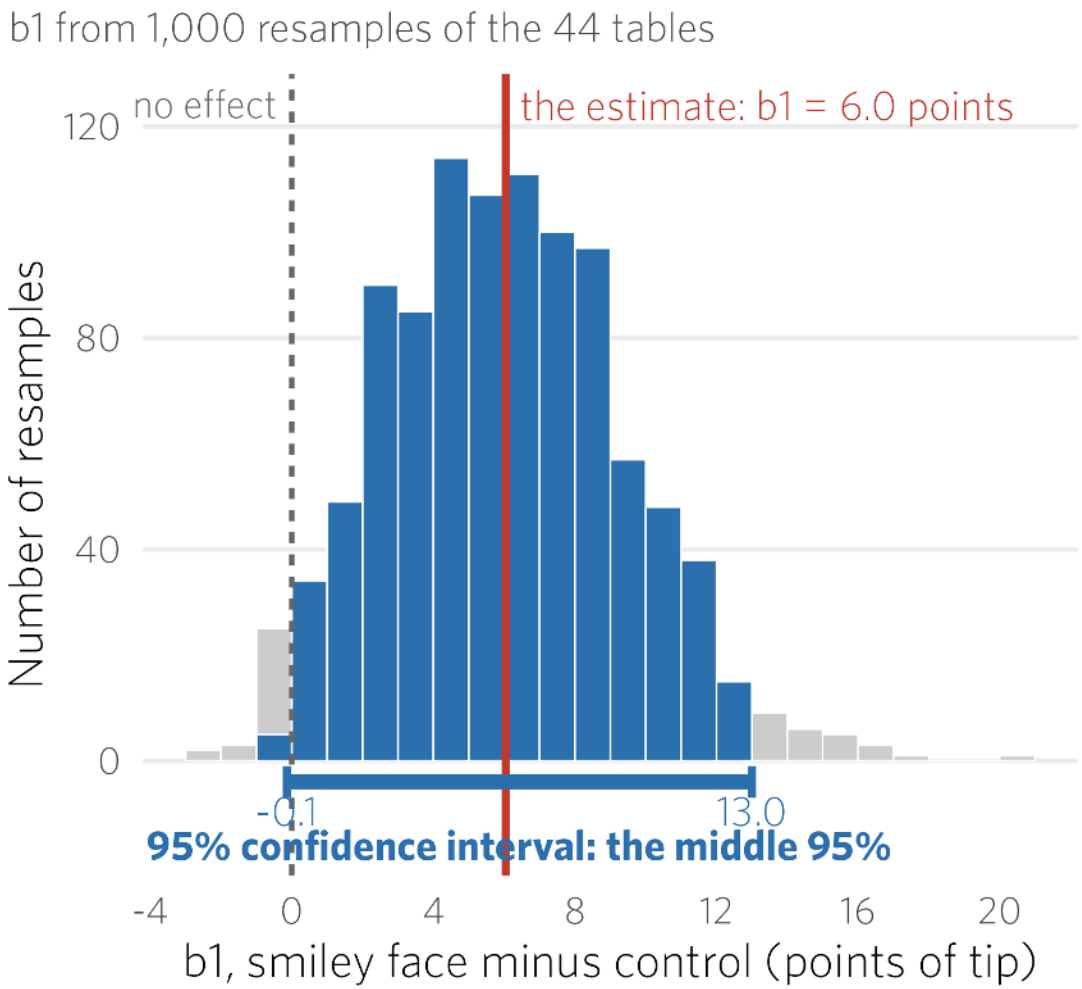
$P(\text{empty model} \mid \text{data})$, the blue arrow back, is a different quantity. A p-value of .09 does not mean a 9% chance that the smiley face made no difference.

Two misreadings of the smiley-face result

The result: 6.0 points more with the smiley face, and a difference at least that large in about 9 of every 100 shuffles.

- **An accurate summary.** Tips were higher on average with the smiley face, but a difference this large is fairly common under chance alone, so the question remains open.
- **Misreading: "The smiley face has no effect."** The data are consistent with no effect, and equally consistent with a true effect of several percentage points of tip; 44 tables are too few to separate the two possibilities.
- **Misreading: "The smiley face raises tips by 6.0 points."** The 6.0 points is the best estimate from these 44 tables, but another 44 tables would give a different estimate. A **confidence interval** puts a range around the estimate: here, the true effect is likely to lie somewhere between about -1 and 13 points. Any effect in that range, zero included, could plausibly have produced data like these.

The confidence interval



- A **confidence interval** is a range of values for the true effect that remain plausible in light of the data.
- Here, b_1 was computed 1,000 times, each time from a resample of the 44 tables. The estimates spread widely, because tips vary so much from table to table, and the interval covers the middle 95% of those estimates: about -1 to 13 points.
- The p-value considers one candidate value for the true effect, zero; the interval considers every candidate value.
- The 95% describes the method rather than this one interval: across many samples, about 95 in 100 intervals built this way would contain the true effect. Whether this particular interval is one of them is unknown.

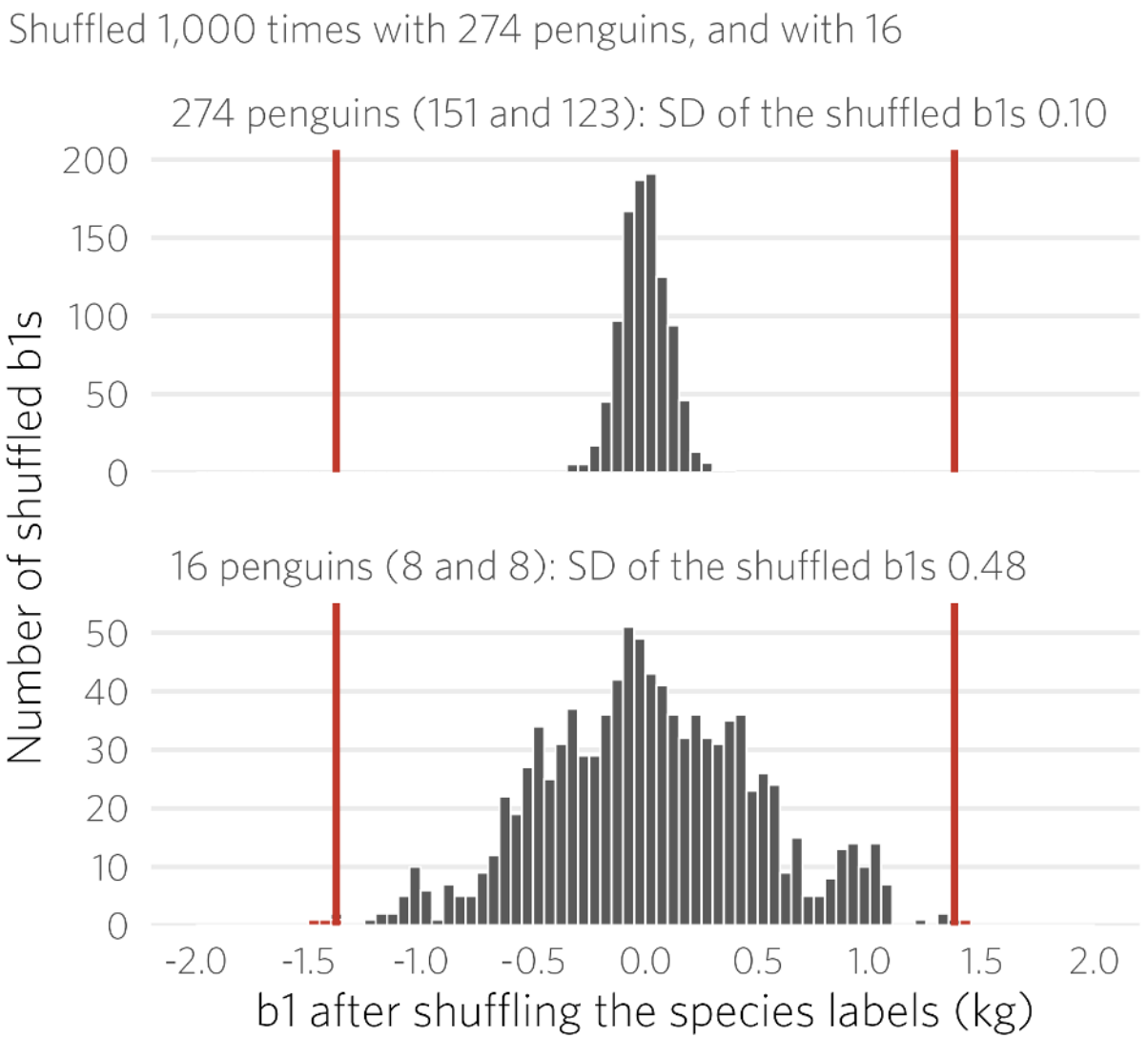
Reporting a confidence interval

The estimate first, then the interval, with its level:

	Estimate	95% confidence interval
Smiley face, 44 tables	b_1 = 6.0 points	-0.7 to 12.8 points
Species, 274 penguins	b_1 = 1.38 kg	1.26 to 1.49 kg

- In a sentence: "Tables with a smiley face tipped 6.0 percentage points more than control tables, 95% CI [-0.7, 12.8]."
- Reading the first interval: it includes 0, so an effect of zero remains plausible, and so does an effect as large as 12 points.
- Reading the second: it is narrow and far from 0, so a difference of zero is implausible, and the species difference is most likely somewhere between 1.26 and 1.49 kg.

The same b_1 from 16 penguins



- With 274 penguins, the shuffled b_1 values have an SD of 0.10 kg; with 16 penguins, 8 per species, 0.48 kg.
- A b_1 of 1.38 kg from 16 penguins would sit near the edge of what chance alone would produce: the same difference, but weaker evidence.
- The p-value depends on the number of cases as well as on b_1: it reflects the strength of the evidence for an effect, not the size of the effect.

Today

Review session

1. The quantities in the ANOVA table, and the question each one answers
2. Choosing a summary, a graph, and a statistic
3. How the pieces fit together: shuffling and the p-value



Office hours with the teaching team

Office hours with the teaching team

- Before you move, please pass your attendance slip to the aisle.
- For the rest of class, members of the teaching team are stationed around the lecture hall to answer your questions.
- Bring your question and your practice exam to any member of the teaching team.

Before Friday

- Finish CourseKata **Chapter 12**.
- Friday: TA office hours, 1:30 to 2:50 pm, in your section's tutorial room. Drop in with your exam questions.
- **Milestone 7** is due Sunday, November 15.
- The **exam** is Wednesday, November 18, in class.

Backup slides for the practice exam

How these slides work

- One slide per item from Monday's practice exam, in the order of the sheet: the question in grey, then the answer in bold and its reasons.
- The full key, with the rubric for each item, is on the Resources page of the course website.

Question 1 · Gym visits and heart rate

A gym recorded each member's visits in a month (2 to 14) and resting heart rate, in beats per minute. Members chose their own number of visits.

```
Coefficients:
(Intercept)       Visits
       80.0         -2.5

                SS  df       MS      F    PRE      p
Model     992.250   1  992.250 27.000 0.3600  .0000
Error    1764.000  48   36.750
Total    2756.250  49   56.250
```

- The model: predicted heart rate = 80 - 2.5 times the number of visits.

Question 1.1 · A bigger branch

Another branch fits the same model to four times as many members and also finds PRE = .36. Is its F larger than 27, smaller, or about the same?

- **Larger**, about four times as large.
- From the formula sheet: $F = \dfrac{PRE/\mathrm{df}_{\mathrm{Model}}}{(1 - PRE)/\mathrm{df}_{\mathrm{Error}}}$.
- This branch: $0.36/(0.64/48) = 27$. The bigger branch, with about 198 error df: $0.36/(0.64/198) \approx 111$.
- The same PRE is stronger evidence when it rests on four times as many members.

Question 1.2 · Reading the table

(i) How many members' data were used? (ii) Which coefficient does df Model = 1 belong to?

- **(i) 50 members.** df Total = n - 1 = 49, so n = 50. The one df taken off is spent on the empty model's mean.
- **(ii) The slope**, b_1 = -2.5. The empty model already estimates the mean; the regression model adds one parameter, the slope on visits.
- The intercept does not get this df: df Total has already taken the empty model's one parameter into account, and df Error, 48, is the remainder.

Question 1.3 · The AI assistant's summary

Rate each claim by how well the output justifies it: well, partly, or not at all.

Claim	Rating	Reason
[1] Each extra visit goes with about 2.5 bpm lower resting heart rate	**well**	the slope, read as an association
[2] A member who adds four visits will lower their heart rate by 10 bpm	**not at all**	members chose their visits; a difference between members is not the effect of a change in one member
[3] A member who never visits would have a heart rate of 80	**not at all**	every member visited at least twice; 0 visits is outside the data
[4] Visits explain 36% of the variation in heart rate	**well**	PRE = .36
[5] Because members chose, the results tell us nothing about visits and heart rate	**not at all**	self-selection limits causal claims, not the description of the association

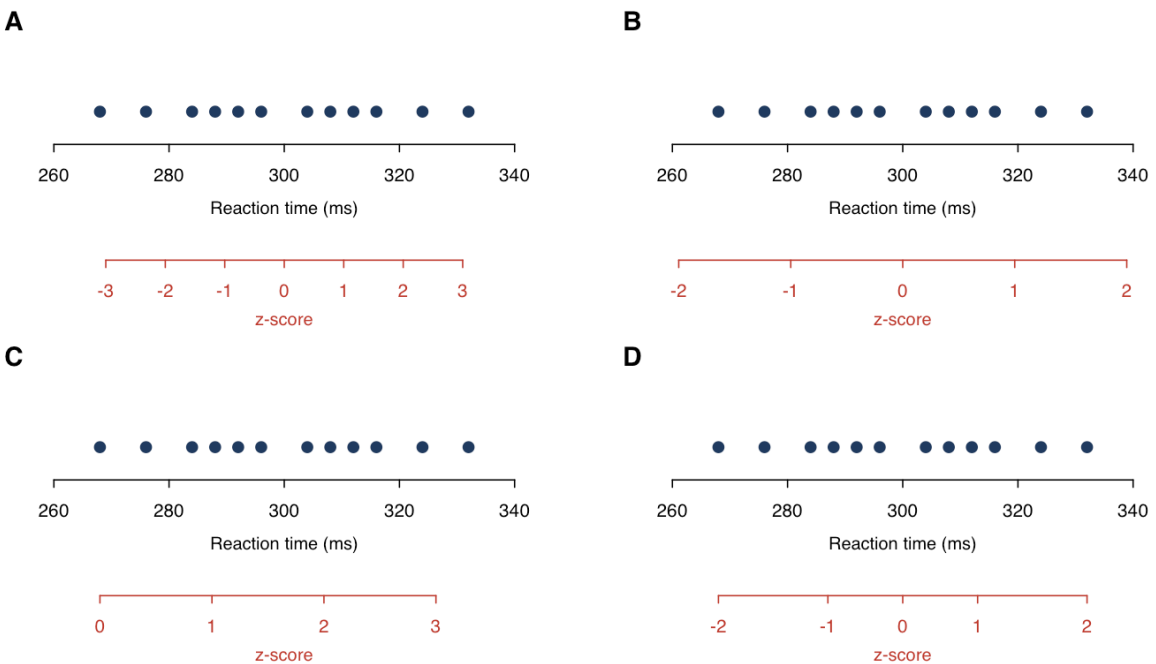
Question 2 · Noise and reaction time

12 volunteers were randomly assigned to three rooms, 4 per room. Reaction time (RT) in milliseconds.

```
favstats(~ RT, data = noise)
 min  Q1 median  Q3 max mean       sd  n
 268 287    300 313 332  300 19.44694 12
favstats(RT ~ Room, data = noise)
   Room min Q1 median  Q3 max mean       sd n
1  Loud 308 314    320 326 332  320 10.32796 4
2   Low 288 294    300 306 312  300 10.32796 4
3 Quiet 268 274    280 286 292  280 10.32796 4
```

- The empty model predicts 300 ms for every volunteer; its SS is about 4,160.

Question 2.1 (i) · Which axis shows z-scores?



- **B**: 0 sits at the mean, 300 ms, and one unit is one SD, about 19 ms.
- A takes its units from one room's SD, about 10 ms, so its units are too small.
- C starts at 0 at the fastest time, but times below the mean have negative z-scores.
- D labels the quartiles, as if z-scores were ranks.

Question 2.1 (ii) and (iii) · Two z-scores for 332 ms

About how many SDs above the mean is 332 ms, among all 12 and within the Loud room? Which comparison is the right one for a Loud-room volunteer?

- **Among all 12:** (332 - 300) / 19, about 1.6 SDs. **Within the Loud room:** (332 - 320) / 10, about 1.2 SDs. The larger value is the one for all 12.
- **The Loud-room comparison** is the right one: it judges the time against the other volunteers in the same room.
- The two values differ because they compare 332 with different distributions: the Loud room's mean is higher, so 332 is closer to it, and the room's SD is smaller.

Question 2.2 · One volunteer drops out

The 296-ms volunteer is removed from the 12. Is the SD of the remaining 11 larger, smaller, or about the same as 19.4 ms?

- **B, larger.**
- The SD comes from SS divided by n - 1. 296 is only 4 ms from the mean, so its squared error, 16, is a tiny part of the total of about 4,160.
- The total barely changes, but the divisor drops from 11 to 10: the variance rises from about 378 to about 414, and the SD from 19.4 to about 20.4 ms.
- Removing a time far from the mean, such as 332, would make the SD smaller.

Question 2.3 (i) · PRE at its two extremes

What would the Room model's PRE be if all three room means were exactly 300 ms? And if every time equaled its room's mean?

- **All room means at 300: PRE = 0.** The Room model makes the same predictions as the empty model and removes none of its error.
- **Every time at its room mean: PRE = 1.** The Room model leaves no error, so it removes all of the empty model's error.
- PRE is a proportion, from 0 to 1, so the second answer is 1, not 100.

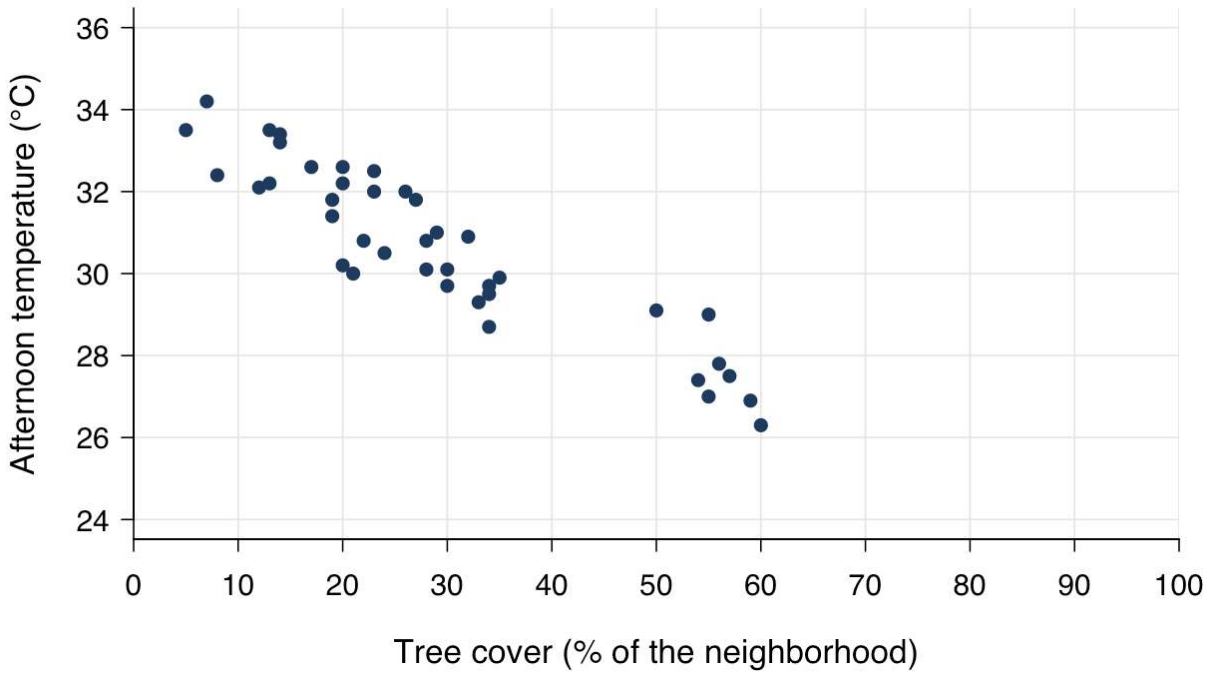
Question 2.3 (ii) · Three sentences about the Room model

- *[1] The room means have the smallest total squared error of any set of three predictions.* **Well.** Within each room, the mean is the prediction with the least squared error.
- *[2] The predictions are exactly right for everyone in the Low room.* **Not at all.** The model predicts 300 ms for all four, but their times run from 288 to 312.
- *[3] SS Error fell, so the rooms must differ in their effect.* **Not at all.** Adding groups always lowers SS Error, by chance alone if nothing else; F and the p-value judge the drop against the drops from shuffled labels.

Question 2.4 · Checking a classmate's work

- *(i) A classmate computes the empty model's SS as the SD squared, times 12.* **Too large.** The variance is SS / (n - 1), so SS is the variance times 11, not 12: about 4,160, not about 4,540.
- *(ii) "An SD of about 19 ms means each volunteer's time is about 19 ms from the mean."* **Partly.** 19 ms is the typical size of the errors, but the distances vary: 4 ms for the 296 and 304 volunteers, 32 ms for the 268 and 332 volunteers.

Question 3 · Trees and summer heat



- 40 neighborhoods: percent tree cover and mean summer afternoon temperature.
- A planner fits a straight line, not drawn here. Its value at 21% cover is about 31.5 °C.
- The one neighborhood measured at 21% cover was 30.0 °C.

Question 3.1 · Which prediction?

A new neighborhood in this city with 21% cover: which prediction is best supported by these data?

- **B, the line's value, about 31.5 °C.** The line draws on all 40 neighborhoods, a better guide than any single neighborhood.
- A, the 30.0 °C of the one measured neighborhood, rests on a single neighborhood and its chance variation.
- C, 23 °C at 95% cover, extends the line past the most shaded neighborhood measured, at about 60% cover.
- D, a desert city, applies the pattern to a city with a different climate from these 40 neighborhoods.

Question 3.2 · The hills

The leafiest neighborhoods are in the hills. A council member: “Planting trees cools neighborhoods.”

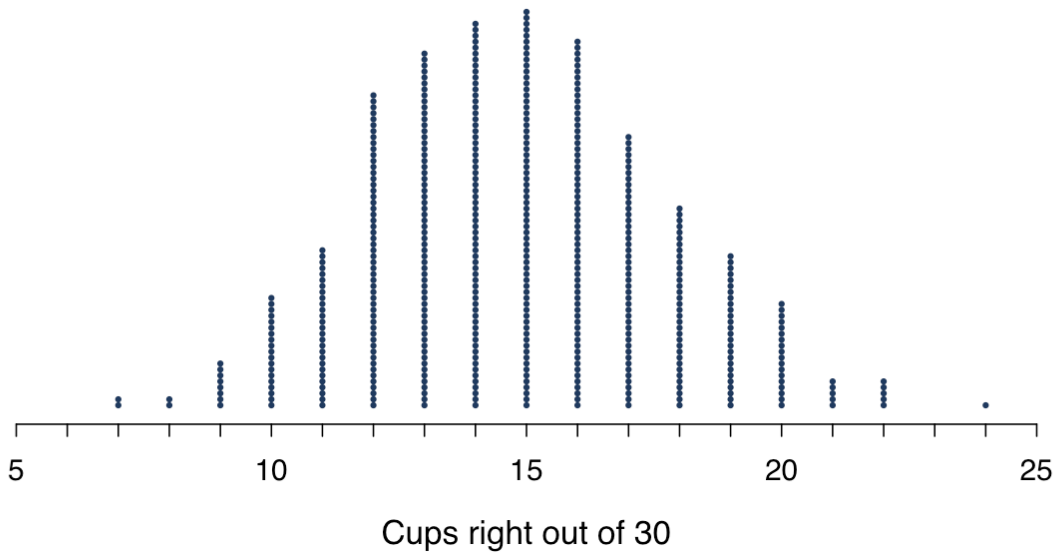
- *(i) If trees had no effect, how would elevation have to relate to temperature for the plot to show the same downward pattern?* **Higher neighborhoods would have to be cooler.** Then the leafy neighborhoods would be the cool ones even if the trees had no effect.
- *(ii) How well do these data support the claim?* **Partly, or not at all.** The pattern fits the claim, but elevation alone could produce it, so these data cannot separate trees from elevation. An experiment that assigns trees at random would settle the question.

Question 3.3 · The cool-pavement lottery

60 streets asked for a reflective coating; 30 are drawn at random for this summer, the other 30 for next year.

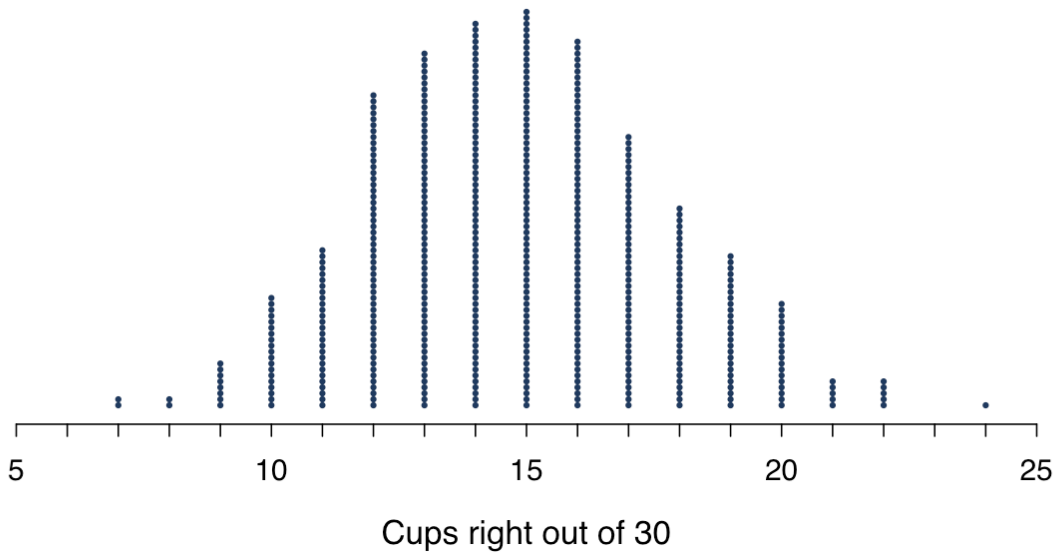
- *(i) Could comparing the two sets of streets show whether the coating cools streets?* **Yes.** The lottery spreads elevation, shade, traffic and every unmeasured factor about evenly across the two sets, so a clear temperature difference can be attributed to the coating.
- *(ii) "A lottery is unfair; you're experimenting on people." How well justified?* **Not at all.** Only 30 streets can be coated this summer either way; the order is set fairly by the lottery, every street is coated within a year, and the coating is harmless.

Question 4 · Telling fresh coffee from stale



- A barista tastes 30 cups, each fresh or stale by a coin flip, and is right on 19.
- The plot: 500 simulated guessers, each right with probability one half on each of 30 cups, one dot per guesser at the number right.

Question 4.1 · Nineteen of thirty



- *(i) How many of the 500 guessers got 19 or more right?* **About 55**, roughly 11 in every 100.
- *(ii) "So she can't really tell the difference."* **Not at all.** Guessing reaches 19 about 11 times in 100, so 19 of 30 is weak evidence of skill.
- Weak evidence of skill is not evidence of no skill: a modest skill could easily score 19 of 30 as well. Only a longer test could separate guessing from a modest skill.

Question 4.2 · A longer test

120 cups, 76 right, again about 63%. Compared with 19 of 30, how strong is this as evidence of skill?

- **C, stronger evidence**, because a guesser's share stays nearer one half on a longer test.
- On 30 cups, about 11 guessers in 100 reach 63%. On 120 cups, lucky and unlucky cups offset one another, so 63% by guessing is rare, about 2 in 1,000.
- A is wrong because the same share is not the same evidence: its rarity under guessing depends on the number of cups.
- B and D count mistakes or cups right without the length of the test, so the counts carry no information.

Question 4.3 · The tasting contest

40 baristas, 30 cups each; “gifted” means 20 or more right. A guesser reaches 20 about 6 times in 100.

- *(i) If all 40 were guessing, how many would be called gifted?* **About 2**: 40 times 6/100 = 2.4.
- *(ii) Three baristas score 20 or more. More than expected? What do the three scores show?* **Only slightly more than the 2 or so expected**, and 40 guessers often produce three. So the scores are weak evidence that any barista has a gift, and they cannot identify a gifted barista.
- Testing many baristas makes a few high scores likely by chance alone.

Question 4.4 · The retest

The three high scorers are retested on 30 new cups each.

- *(i) If they were guessing, what score would you expect on the retest?* **About 15**, most likely between 10 and 20. They were selected for high scores, so if luck produced those scores, the retest scores should be lower.
- *(ii) "Below 20 on the retest means no gift." A taster right on 65 of 100 cups scores below 20 about half the time. How well justified is the rule?* **Not at all.** A real but modest skill misses the cutoff about half the time, so the rule would dismiss about half of such tasters.
- A real skill does not produce a high score on every test; a decision needs several tests, or a test with many more cups.

Question 5.1 · Two surveys

A reading app with 40,000 users. Emailed survey: 4,000 replies, 85% like the app. A researcher's random sample of 200: 190 answered, 62% like the app.

- *(i) Is the 85% likely too high, too low, or about right?* **Too high.** The 4,000 chose to reply, and users who like the app are more likely to answer, so the replies over-represent satisfied users.
- *(ii) Which estimate would you trust more?* **The 62%.** The 200 were chosen at random and almost all answered, so they resemble all users, with some chance error but no bias in a known direction.
- 4,000 replies do not remove the bias: a larger sample with the same bias is a more precise estimate of the wrong quantity.

Question 5.2 · The district rankings

120 districts ranked by mean reading gain. The top 10 and the bottom 10 all have fewer than 30 users. A manager: "Small districts use the app best."

- *(i) How could the smallest districts end up at both ends even if size had no relation to the app's effect?* A mean of 20 students moves a lot when a few readers are unusually strong or weak; a mean of 2,000 students barely moves for the same readers. Small districts' means spread more widely in both directions, so the extremes, high and low, come from small districts.
- *(ii) How well justified is the manager's conclusion?* **Not at all.** Small districts are also at the bottom of the ranking, as chance variation predicts; the manager's explanation covers only the top.

Question 5.3 · The big trial

20,000 students randomly assigned to the app or not. The app group scored 0.4 points higher out of 100; none of 1,000 shuffles reached 0.4. Teachers call 5 points meaningful. The ad: "Proven to dramatically improve reading scores!"

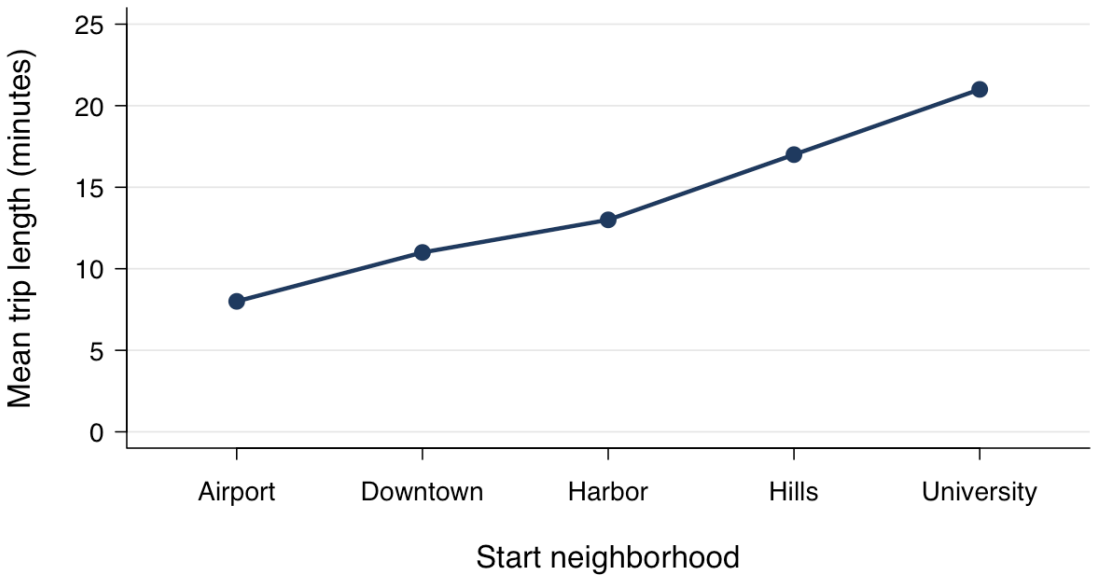
- *(i) What does the shuffle result show about whether the app has any effect?* Chance never produced 0.4 in 1,000 shuffles: strong evidence of some effect, and, because of the random assignment, the effect can be attributed to the app.
- *(ii) How well justified is the ad?* **Partly.** "Proven to improve": justified, the evidence is strong. "Dramatically": not justified, 0.4 points is under a tenth of the 5 points teachers consider meaningful.
- With 20,000 students, even a tiny effect stands out from chance. Strong evidence of an effect is not evidence of a large effect.

Question 6.1 · The analyst's two models

2,000 bike-share trips in May, with each trip's start station number (an ID from 1 to 80), start neighborhood, minutes, and temperature.

- *(a) Slope 0.05 on station number: "Trips from a station numbered one higher are 0.05 minutes longer."* **Not at all.** Station numbers are labels, not amounts: renumber the stations and the slope changes, with the same trips.
- *(b) Slope 0.3 on temperature: "Trips starting one degree warmer are 0.3 minutes longer."* **Well.** Both variables are amounts, and the sentence describes an association.
- *(ii) A better way to use station in a model?* A group model: predict each trip with the mean length of trips from its station, or its neighborhood; each station is its own group, with no order.

Question 6.2 · The neighborhood chart



- *(i) The horizontal axis: categorical or quantitative? Ordered?* Categorical, with no natural order; listed here alphabetically.
- *(ii) "Trips get longer from Airport to University."* **Not at all.** The line suggests a trend, but the "trend" depends only on the order of the neighborhoods on the axis.
- *(iii) A better chart?* Bars, or unconnected points, one per neighborhood: nothing joins the categories as if they lay on a scale.

Question 6.3 · Rain and riding

Rides were 5% lower on rain-forecast days. A difference of at least 5% either way came up in about 40 of every 100 shuffles. The most accurate summary for the operations team?

- **C.** Rides were a little lower, but a difference of this size is common under chance alone, so the data are not yet decisive.
- A turns "chance could explain the difference" into "the difference is chance"; such data are also consistent with a real effect.
- B treats the 5% as established, but the true effect could be zero, 5%, or larger.
- D misreads the 40 in 100: it is how often a difference this large arises with no forecast effect, not the probability of no effect.