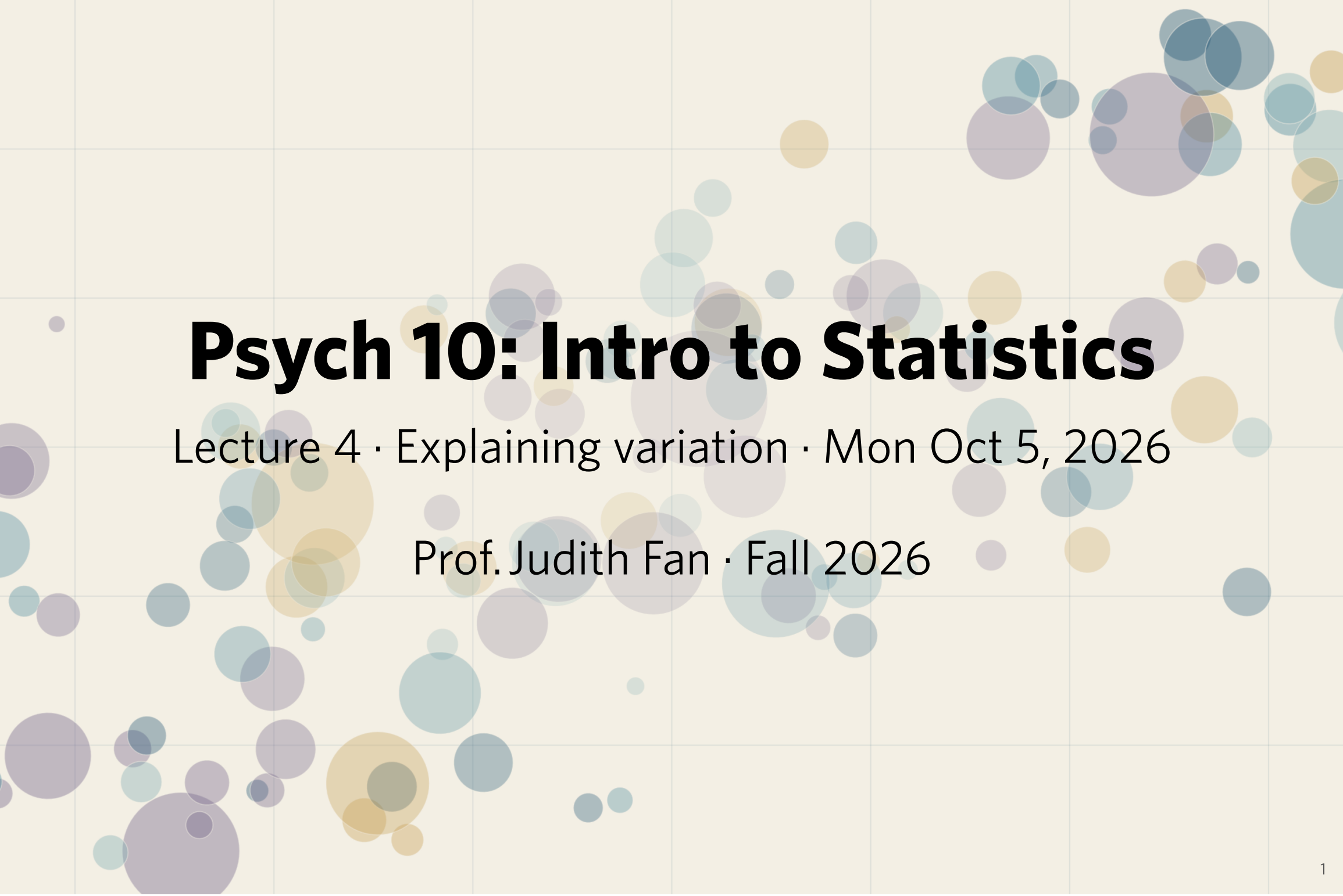




Psych 10: Intro to Statistics

Lecture 4 · Explaining variation · Mon Oct 5, 2026

Prof. Judith Fan · Fall 2026

Reminders & Announcements

- CourseKata **Chapter 4** and **Chapter 5** are both due this Friday. If you haven't started Chapter 4, start it today.
- Section this week: **Milestone 2**, due Sunday, October 11 on Gradescope.
- Friday: **Case #2, *The Vault of Fallen Labels***, in your tutorial session. Check your discussion section's Slack channel for your time slot. Bring a pen.

Today

Explaining variation

1. Using one variable to make a guess about another
2. Plotting one variable against another
3. Variation obscured by a bar plot showing only averages
4. "Explaining" in statistics is not the same thing as "causing": the tennis study

**1 Using one variable to make a
guess about another**

Guess one student's phone time yesterday

A student in this class is picked at random. **How many hours were they on their phone yesterday?**

Write a number in the margin of your worksheet.

Now you learn that they have a part-time job. **Does that make you want to pick a different number?**



What “explains variation” means

- If knowing one thing about a student (do they have a job?) helps you guess another thing about them (how many hours were they on their phone?), then the first variable **explains some of the variation** in the second.
- A better guess is the whole meaning of “explains variation”. It doesn’t promise a correct guess.
- Having a job explains some of the variation in phone hours if students with jobs and students without them tend to spend different amounts of time on their phones.

Written as a word equation

phone hours = job + other stuff

- Say it as: some of the variation in phone hours is explained by whether a student has a job, and the rest is other stuff.
- The **outcome** goes on the left and the **explanatory variable** on the right. “Other stuff” is every other source of differences among students.
- It’s a sentence written like an equation. The = sign means “is explained by”; it is not the = of arithmetic.

Which variable is the outcome? Some examples!

1. Students who sleep less spend more time on their phones.
2. Seniors are late to class more often than frosh.
3. Students who spend more time on their phones get less sleep.



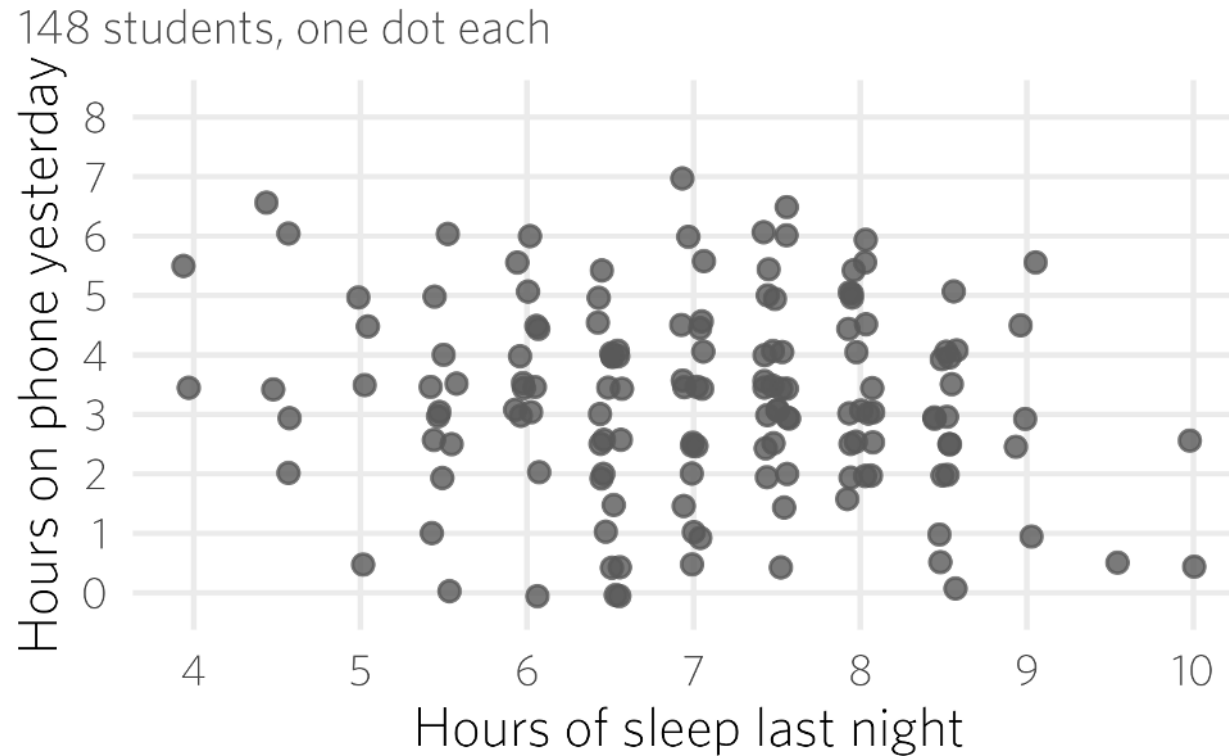
2 Plotting one variable against another

Different kinds of variables call for different kinds of plots

	explanatory variable is quantitative	explanatory variable is categorical
outcome is quantitative	a scatter plot	one dot per case, a box plot per group, or a histogram per group
outcome is categorical	(rare in this course)	the proportion in each group, or a table of counts

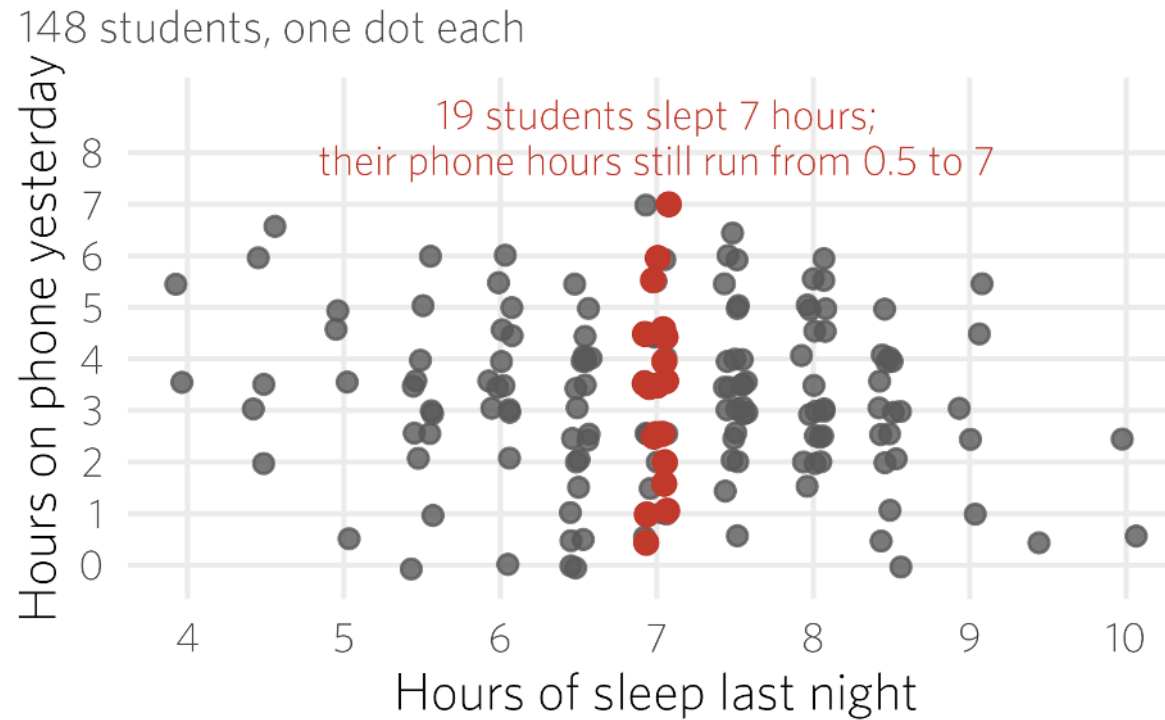
The next few slides show one example of each, all from the same survey of 148 students.

Both quantitative: a scatter plot



One dot per student. The outcome goes on the vertical axis.

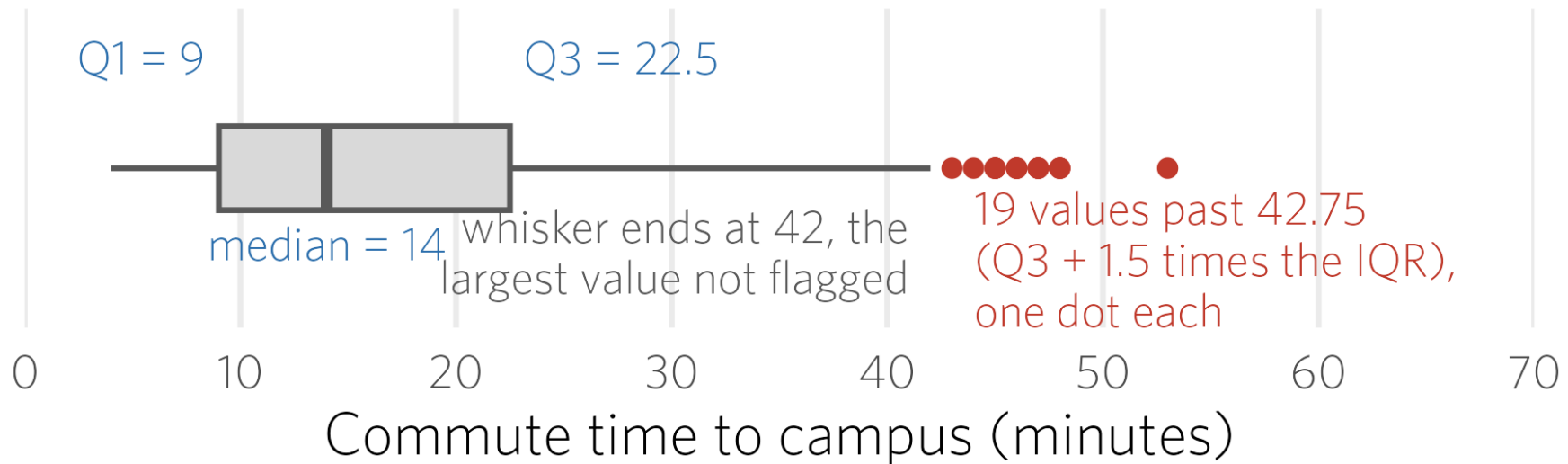
Same sleep, very different phone time



Take the students who slept exactly 7 hours. Their phone time still runs from about 0 to 7 hours. Sleep can't account for that vertical spread; it's all other stuff.

How to read a box plot

Box plot of the 148 commute times



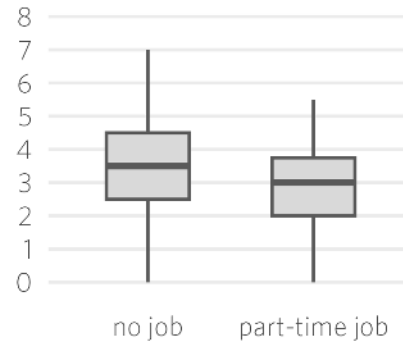
- The box is the middle half of the students; the line is the median.
- The whiskers reach the rest; a dot is a value far from the box.

Quantitative outcome, categorical explanatory variable: three plots

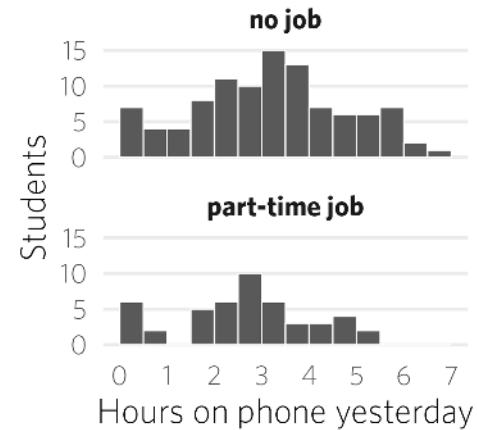
One dot per student



A box plot per group



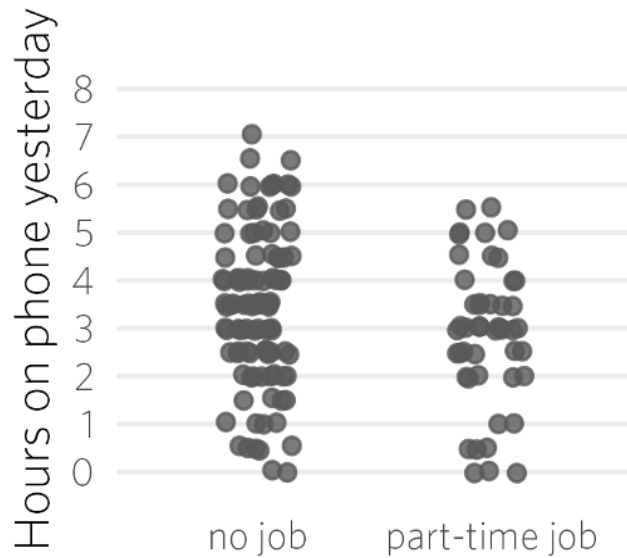
A histogram per group



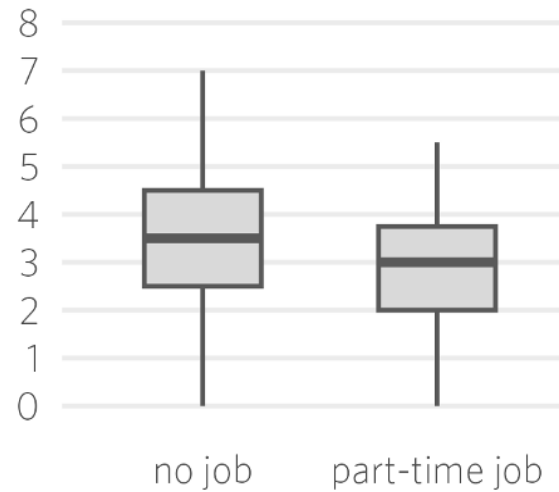
The same 148 students three ways: one dot per student (the dots are spread sideways so that each one is visible; a dot's position inside its column carries no information); a box plot per group; a histogram per group on the same axis.

With your neighbor

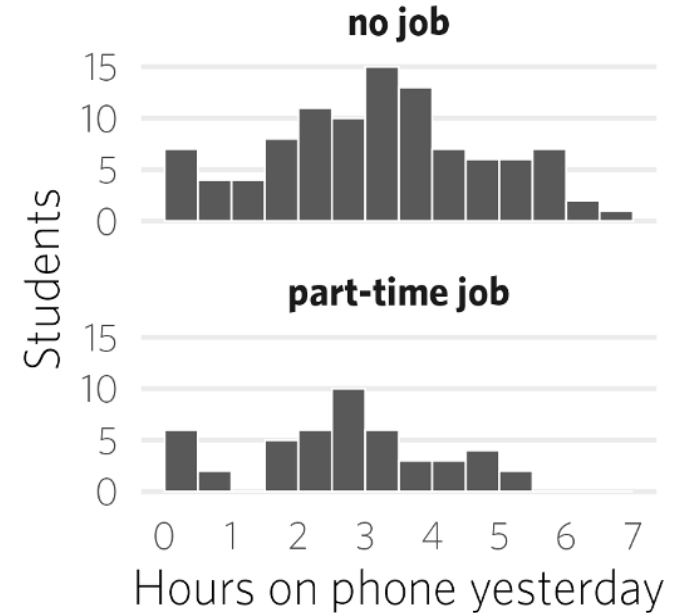
One dot per student



A box plot per group



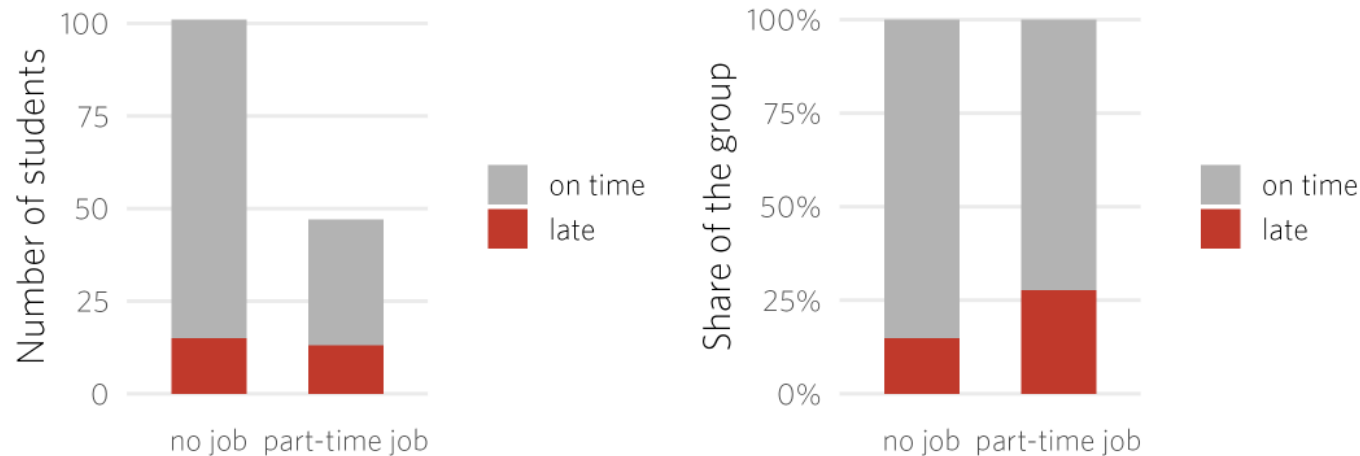
A histogram per group



Does having a job explain some of the variation in phone hours? Use any of the three plots.

Both categorical: the share of each group

Counts: 15 late students without a job, 13 with one
Proportions: 15% of one group late, 28% of the other



Now the outcome is categorical too: on time or late, by job. There's no mean and no spread, only counts, and counts mislead when the groups differ in size (101 students have no job and 47 have one). Compare the share of each group: 15% late without a job, 28% with one.

The same numbers as a table

	no job	part-time job
on time	86	34
late	15	13
total	101	47

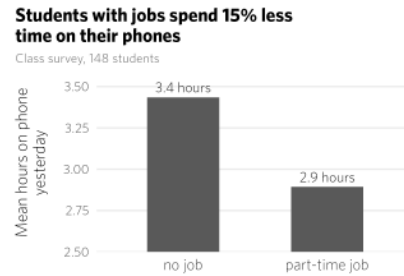
- 13 out of 47 is 28%: the share of students with a job who were late.
- 13 out of 28 is 46%: the share of late students who have a job.
- Both are true, and they answer different questions. Whenever you report a percentage, say what the whole is: 28% *of the students with a job*, or 46% *of the late students*.

3 Variation obscured by a bar plot showing only averages

A bar plot showing only averages

Newspapers and reports rarely print the dots, the box plots, or the histograms. They print a bar plot: one bar per group, as tall as the group's mean. Part 1 is one of those, made from our phone-time data.

Worksheet - Part 1



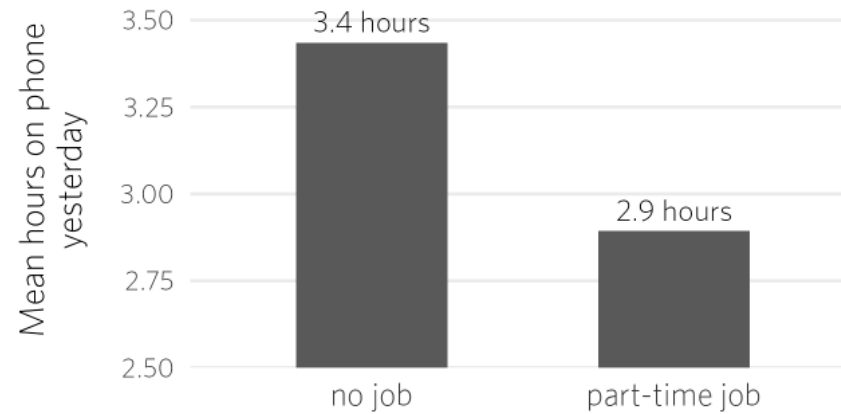
Suppose a campus newspaper printed this bar plot, with the headline “Students with jobs spend 15% less time on their phones.” The bars are the mean number of hours spent on the phone yesterday, for students without a job and for students with a part-time job.

- a.** The headline says 15% less. Check that with the two numbers printed on the bars: by what percentage is the job group’s mean below the no-job group’s mean?
- b.** Look at where the vertical axis starts. If it started at zero instead, would the two bars look nearly the same height, or very different?

Worksheet - Part 1, continued

Students with jobs spend 15% less time on their phones

Class survey, 148 students

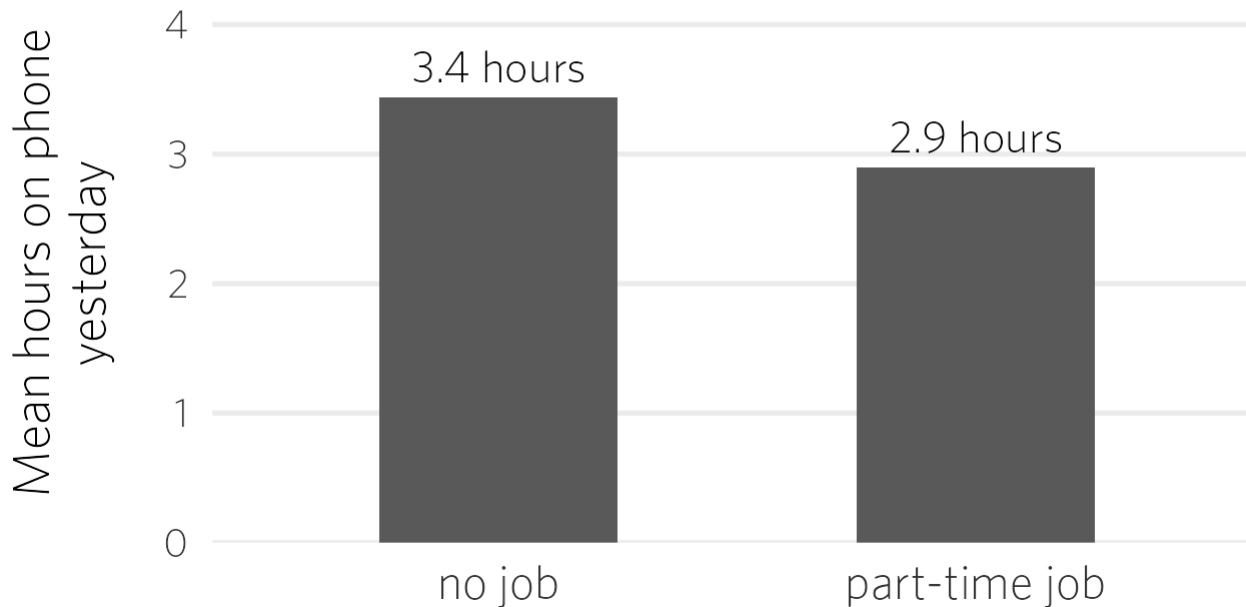


c. The two means differ by half an hour. From this bar plot alone, can you tell whether half an hour is a big difference compared with how much students *within the same group* differ from one another? What would a plot have to show before you could answer that question?

The same two means with the axis starting at zero

Students with jobs spend 15% less time on their phones

Class survey, 148 students



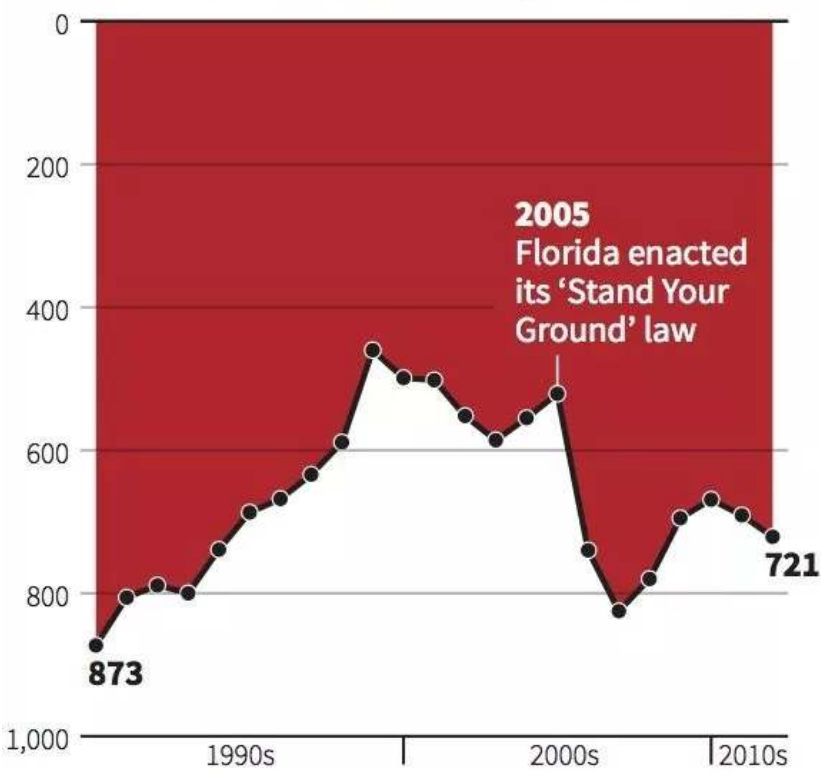
Where should the vertical axis start?

- **For a bar plot, usually at zero.** A bar's height stands for its number, so an axis that starts higher up turns a small difference into a large-looking one, as in the newspaper's chart. If you have a reason to start it somewhere else, make the starting point obvious to the reader.
- **For a line plot, often not at zero.** A line shows how something changes, and an axis that reaches all the way down to zero can flatten that change into nothing. Start the axis where it shows the data.
- Either way, the first thing to read on any plot is where the axis starts.

Gun deaths in Florida

Gun deaths in Florida

Number of murders committed using firearms



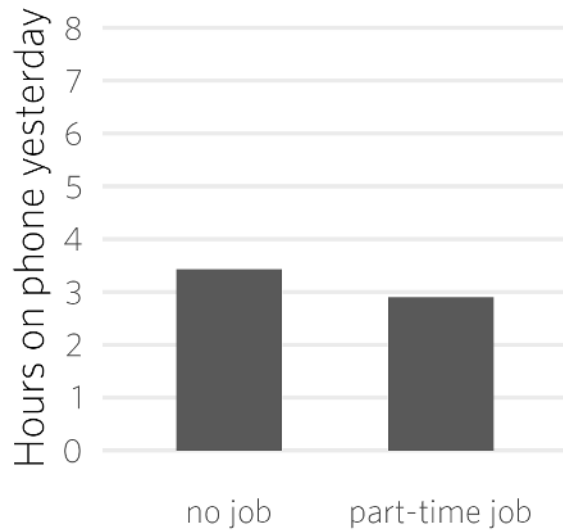
Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

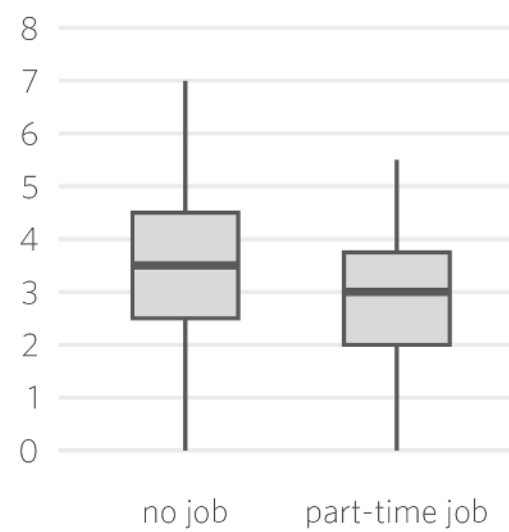
REUTERS

The variation the bar plot obscures

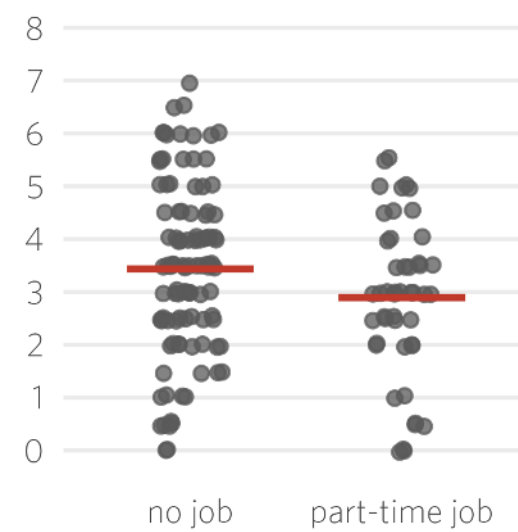
The two means



A box plot per group



Every student, with the means in red

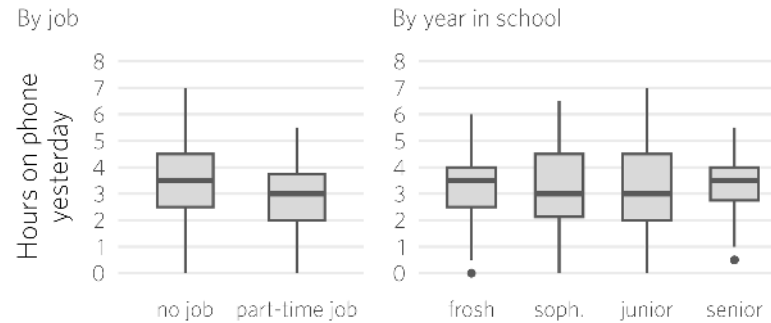


The two means differ by half an hour. Within each group, students range from 0 to 7 hours.

Between-group and within-group variation

- **Between-group variation:** the distance between the two groups' centers.
Here, half an hour.
- **Within-group variation:** the spread of the students inside each group.
Here, 0 to 7 hours in both groups.
- So having a job explains a little of the variation in phone hours. Most of it is other stuff.

Worksheet - Part 2



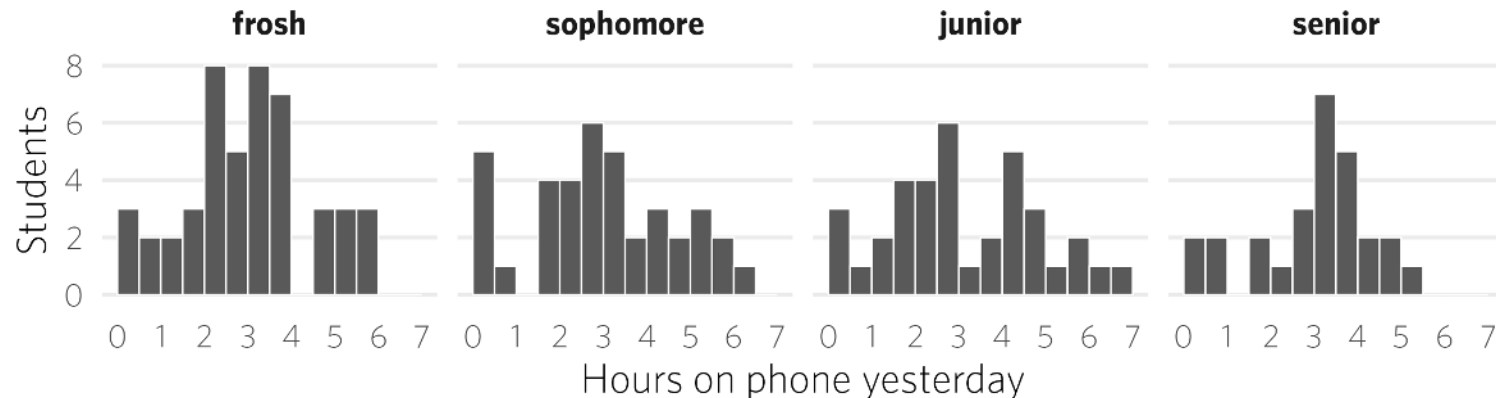
These plots show the same 148 students' phone hours broken down two ways: by job, and by year in school.

a. Which explanatory variable explains more of the variation in phone hours, job or year? Say which feature of the box plots you used: the distance between the medians, or the amount of overlap between the boxes.

b. Look at the students without a job. Is everyone in that group on their phone for about 3.5 hours a day? Say which part of the box plot is your evidence.

Which explains more, job or year?

A histogram per year, same axis



Neither one explains much of the variation. The four year medians are identical, the two job medians differ by half an hour, and the boxes overlap almost completely in both plots. Even the group with the highest median has students at 0 and 1 hours. For now, “a little” or “a lot” is as precise as we can get just by looking at the plots (we’ll learn how to put an actual number on it in a couple of weeks!).

4 “Explaining” in statistics is not the same thing as “causing”: the tennis study

The tennis study

So far, “explains” has just meant “helps you guess”. Here is a variable that helps a lot with the guess. Does that mean it *causes* the outcome?

In 2018, a group of researchers reported on about 8,600 adults in Copenhagen whom they had followed for 25 years. Compared with people who didn’t exercise at all, the tennis players lived about **9.7 years longer** on average! (Badminton: 6.2 years. Soccer: 4.7. Cycling: 3.7. Swimming: 3.4. Jogging: 3.2.)

Schnohr et al., *Mayo Clinic Proceedings*, 2018

Worksheet - Part 3

A 2018 study followed about 8,600 adults in Copenhagen for 25 years. **People who played tennis lived about 9.7 years longer, on average, than people who did no exercise.**

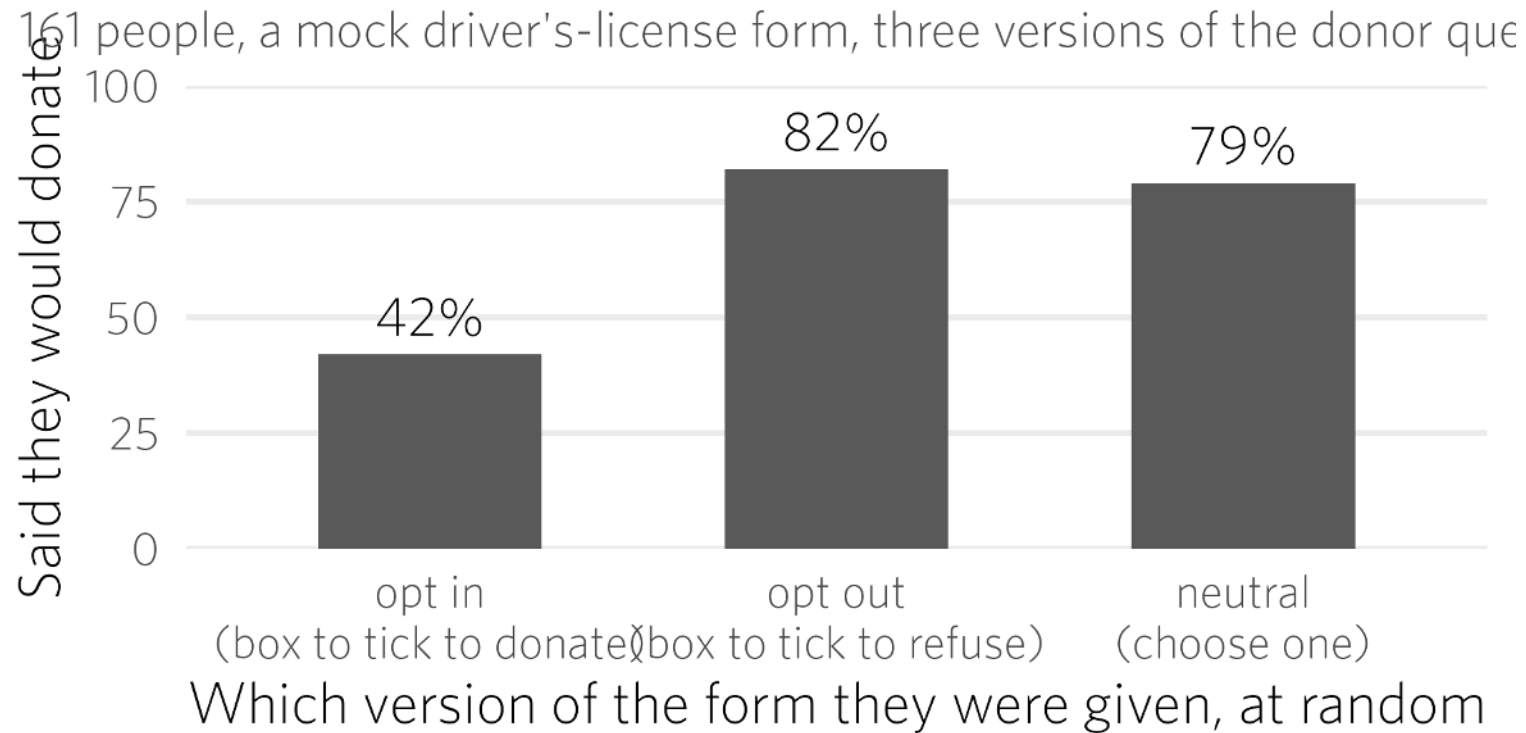
- a.** Write down **two different processes** that could have produced that difference: two different explanations for how it could have come about, not two versions of one explanation.
- b.** Ask a neighbor for one of theirs and write it down as your third. Then check each other's: does your neighbor's process seem plausible to you, and does yours to them?
- c.** Circle the one of the three you find most plausible.
- d.** Pick one of the others. What would you look up, or go out and measure, to tell whether it or the one you circled is what produced the difference?

Let's hear some

Four things to say about the tennis finding

- **Maybe the arrow points the other way.** People who are already sick don't take up tennis, and people who get sick give up the game.
- **Maybe something else explains both.** People who are wealthier, or healthier to begin with, are more likely to play tennis *and* more likely to live a long time.
- **Nobody was assigned to tennis.** The researchers just watched and measured, so in the end they were comparing the kind of people who choose tennis with the kind of people who do no exercise.
- **What would settle it?** Assign thousands of people at random to tennis or to no exercise, and follow them for 25 years. With random assignment (who gets which treatment is settled by a coin flip, not by the person or by the researcher), you can say "causes". With random sampling (who is in the study is settled the same way), you can say "in general". The Copenhagen study had neither kind of randomness.

An experiment on organ-donation forms



Here's what it looks like when you *do* assign at random: the three groups should be alike in every other way, so the form is the only thing that could have made the difference. And it made a big difference!

Takeaway

- In statistics, “explaining” is not the same thing as “causing”. Tennis explains some of the variation in how long these Danes lived, but whether it *caused* any of it is a separate question.
- If you want to say “causes”, you need a **experimental study** that assigns the treatment at random. Collecting more data, on its own, doesn’t license causal claims.

Before Friday

- Finish CourseKata **Chapters 4 and 5**.
- Friday: **The Vault of Fallen Labels**. Check your discussion section's Slack channel for your time slot.
- Milestone 2 is due Sunday.

Hand in your worksheet on the way out

On the last two lines, write down something new you learned today. Thanks!