Lecture 3: Observational Studies + Randomized Experiments + Confounding + Simpsons's Paradox

Chapter 1.4

Goals for Today

- We illustrate the difference between
  - an observational study
  - a randomized experiment, where the treatment is assigned at random.
- Introduce the notion of confounding AKA lurking variables
- Discuss Simpson's Paradox (not in textbook).

Going Back to Previous Example

Going back to the study on

 $\Longrightarrow$ 

- The explanatory variable was: sleeping with your shoes on
- The response variable was: waking up with a headache
- The doctor hypothesized a causal relationship

Confounding Variable AKA Lurking Variable

This is an example of confounding. A confounding variable affects both the explanatory and response variable. So if:

Controlling for Potential Confounding

One way to control for (i.e. take into account) confounding is to do an exhaustive search for all such variables. This is not always practical.

Another way is via an experiment where we randomly assign individuals to a treatment or a control group in a randomized experiment.

Back to Shoes and Headaches

So imagine we recruit 10,000 people for our study and randomly assign 5000 people to each of:

- Treatment: sleep with shoes on
- Control: sleep with shoes off

In this table

Group	n	# with headache
Treatment	5000	n_1
Control	5000	n_2
Total	10,000	$n_1 + n_2$

n_1 and n_2 won't be very different.

Observational Studies vs Randomized Experiments

The key word from the study design above was randomly assign.

- Observational studies: a study where researchers have no control over who receives the treatment
- Randomized experiments: a study where researchers not only have control over who receives the treatment, but also make the assignments at random.

Observational Studies vs Randomized Experiments

Conclusion: The study introduced at the end of the last lecture is an observational study, so we cannot conclude that wearing shoes when you sleep causes you wake up with a headache.

Mantra: Correlation is not causation Just because two variables appear to be associated/correlated, does not mean that one is causing the other.

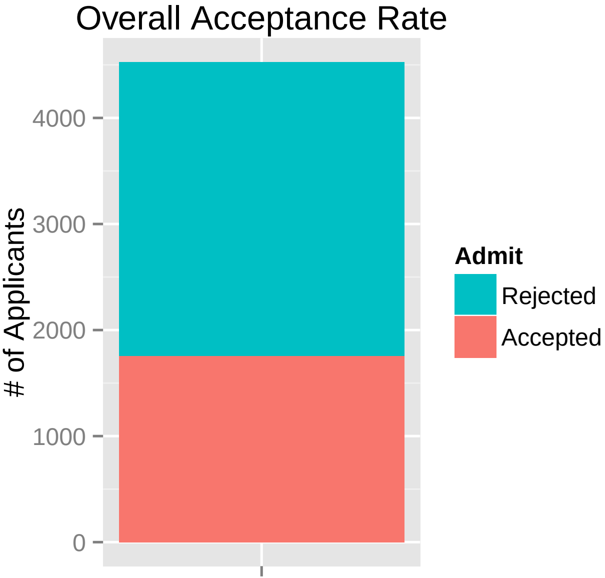
- Spurious correlations: http://www.tylervigen.com/
- Saturday Morning Breakfast Cereal: http://www.smbc-comics.com/?id=3129

Well-Known Example of Confounding

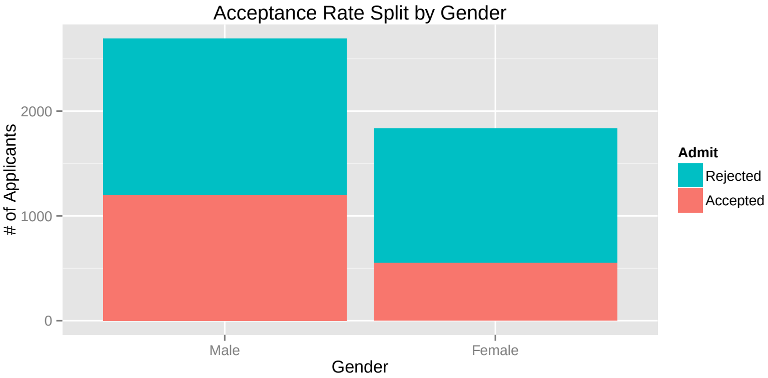
A famous example of an unaccounted for confounding variable having serious repercussions was when the UC Berkeley was sued in 1973 for bias against women who had applied for admission to graduate schools.

Let's consider the $n = 4526$ people who applied to the 6 largest departments.

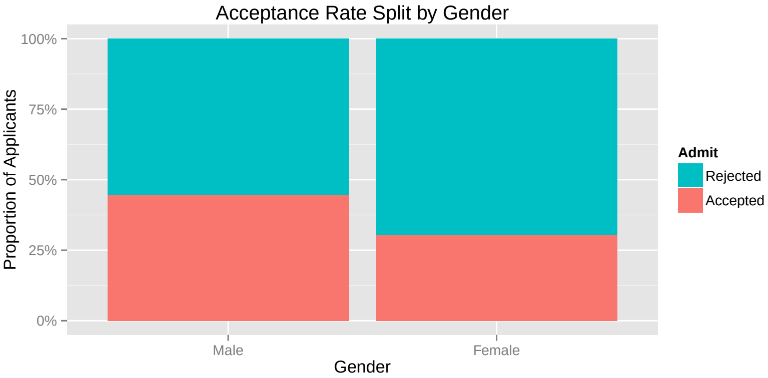
Of the $n = 4526$ applicants:



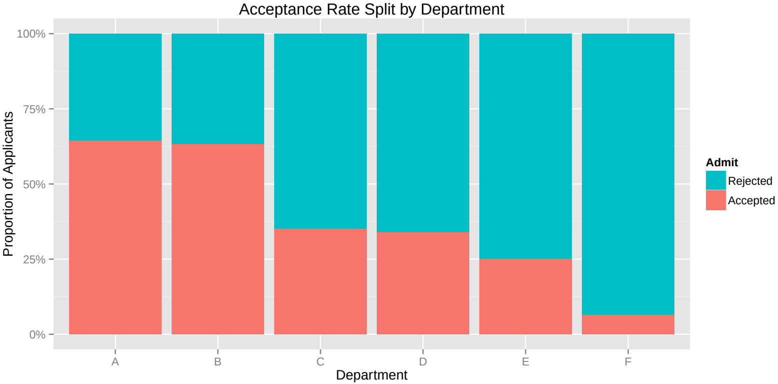
Split the counts by gender:



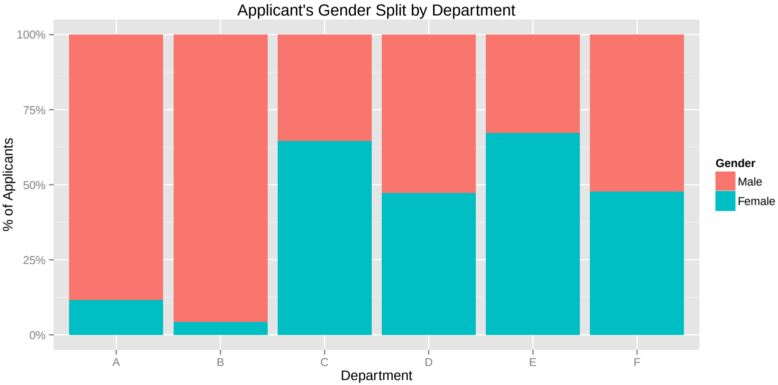
Look at proportions instead of counts:



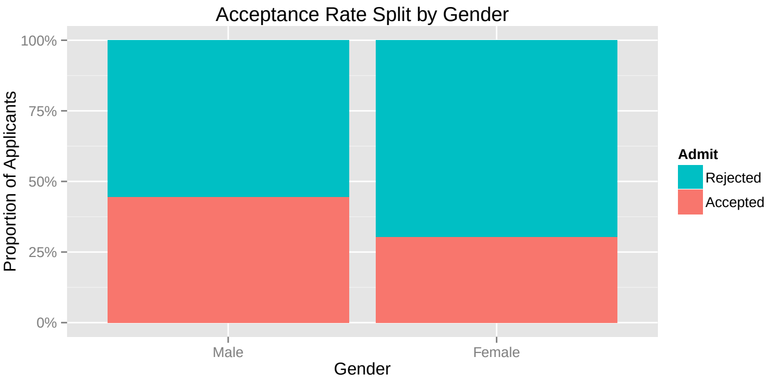
What was the “competitiveness” of departments?



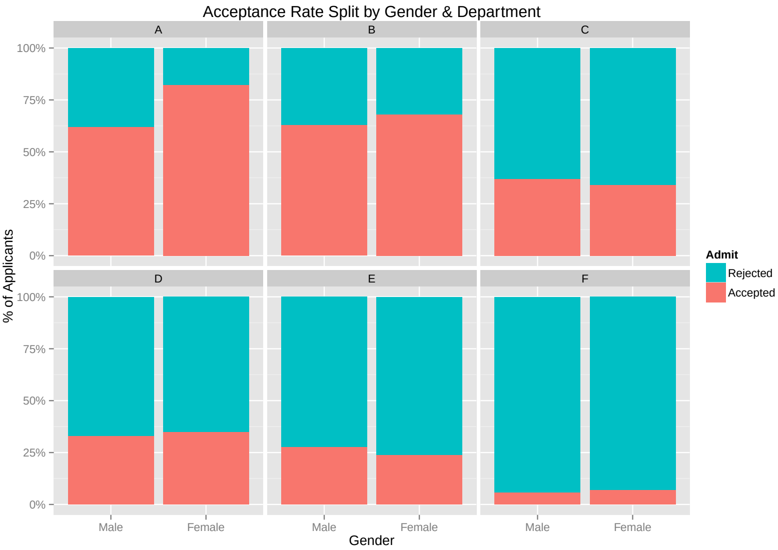
Where were the women applying?



So while in aggregate things looked like this:



You need to account for department!



Bickel et al.'s (1975) Explanation

There was the presence of a confounding variable: competitiveness of applying to the department, which is a function

- number of applicants
- number of available slots

So it wasn't that departments were discriminating against women, rather:

- women tended to apply to departments with high competition and hence lower admission rates, primarily the humanities.
- men tended to apply to departments with low competition and hence higher admission rates, primarily the sciences.

Bickel et al.'s (1975) Explanation

In fact, Bickel et al. found that "If the data are properly pooled...there is a small but statistically significant bias in favor of women."

This was the exact opposite claim of the lawsuit. This is known as Simpson's Paradox.

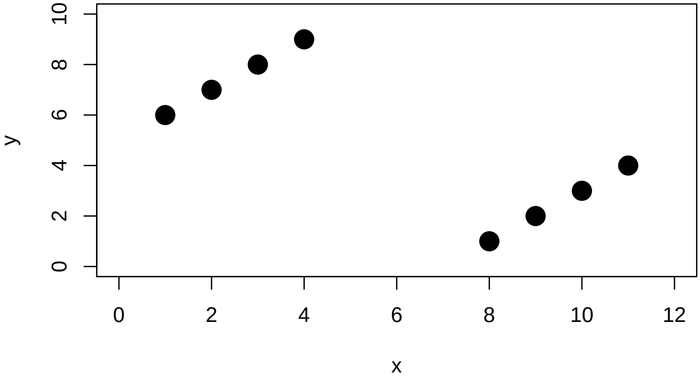
Simpson's Paradox

(From Wikipedia) Simpson's paradox occurs when a trend that appears in different groups of data disappears when these groups are combined, and the reverse trend appears for the aggregate data.

This is due to a confounding variable.

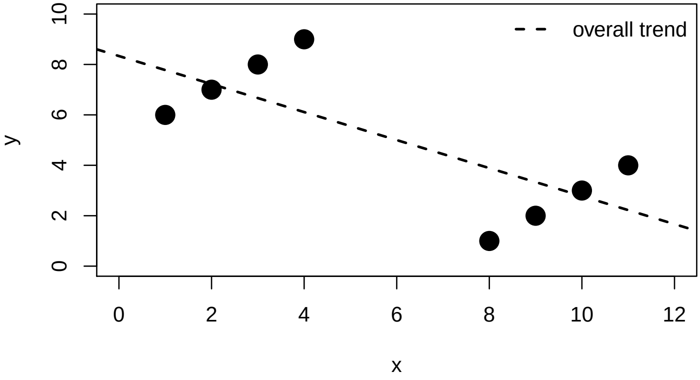
A Graphical Illustration of Simpson's Paradox

Say we have the following points:



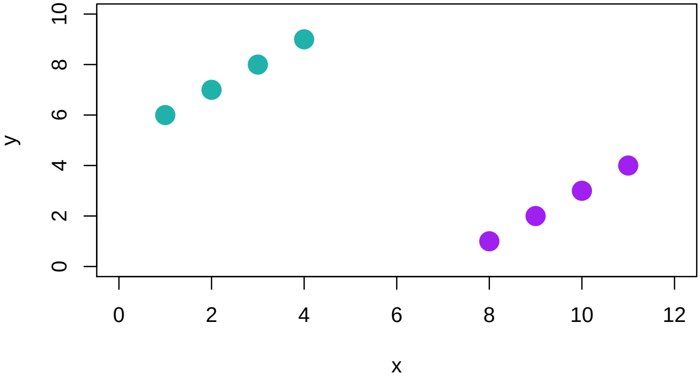
A Graphical Illustration of Simpson's Paradox

Overall, if we fit a single line, the explanatory variable x is negatively related with the outcome variable y:



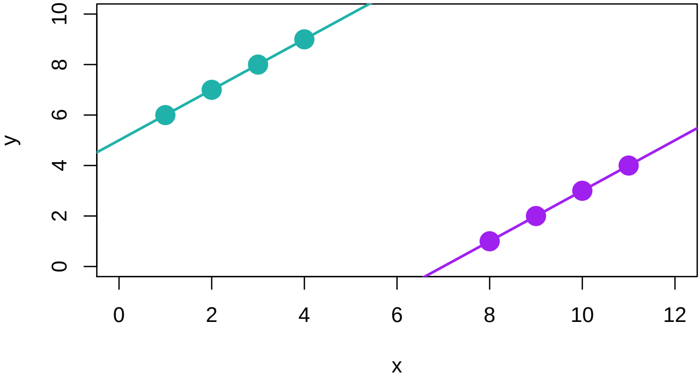
A Graphical Illustration of Simpson's Paradox

But say we consider a confounding variable, in this case color, and fit two separate lines for each group:



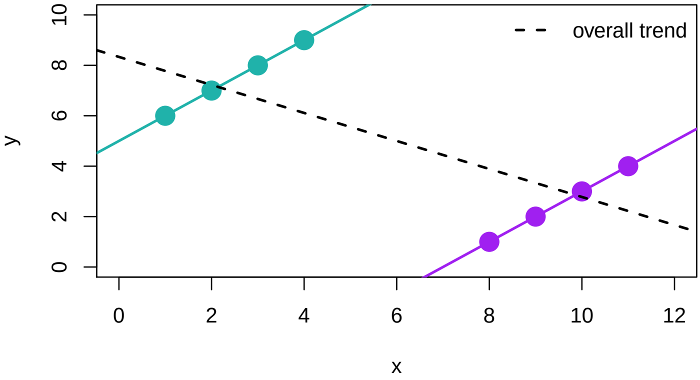
A Graphical Illustration of Simpson's Paradox

The subgroups now exhibit a positive relationship!



A Graphical Illustration of Simpson's Paradox

i.e. the trend in aggregate is the reverse of the trend in the subgroups (teal & purple).



Bickel et al.'s (1975) Conclusion

"The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system."

"Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects."

The original paper can be found here.

Next time

We will discuss

- Specific types of sampling beyond just simple random sampling, as this is not always feasible
- Experimental design: some key principles to keep in mind when evaluating the efficacy of treatments.