Lecture 1: Laying the Foundations + Terminology

Chapters 1.1-1.2

Goals for Today

- Go over the syllabus
- Show some fun examples
- Discuss how to evaluate the efficacy of a treatment
- Describe the different kinds of variables we'll consider

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

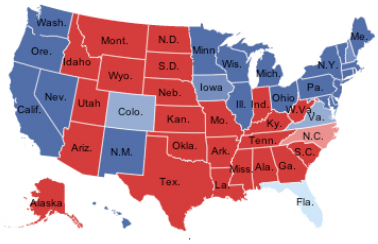
1. Identify the scientific question or problem
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion and communicate it

Statistics concerns itself with points 2 through 4.

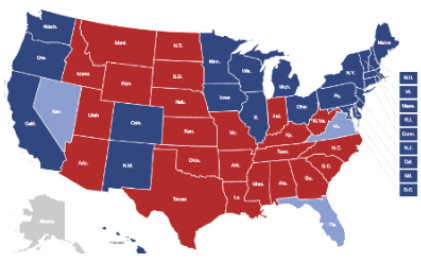
Your Majors

Biology	Economics
11	5
History	Environmental Studies
4	3
Mathematics	Psychology
3	3
Biochem and Molecular Biology	Chemistry
2	2
International Policy Studies	Linguistics
2	2
Undecided	Anthropology
2	1
Economics/Mathematics	Environmental Studies-Hist
1	1
Environmental Studies-Pol Sci	Physics
1	1
Sociology	
1	

Example: 2012 Election - Nate Silver's Predictions vs Actual Results



Nate Silver's Map



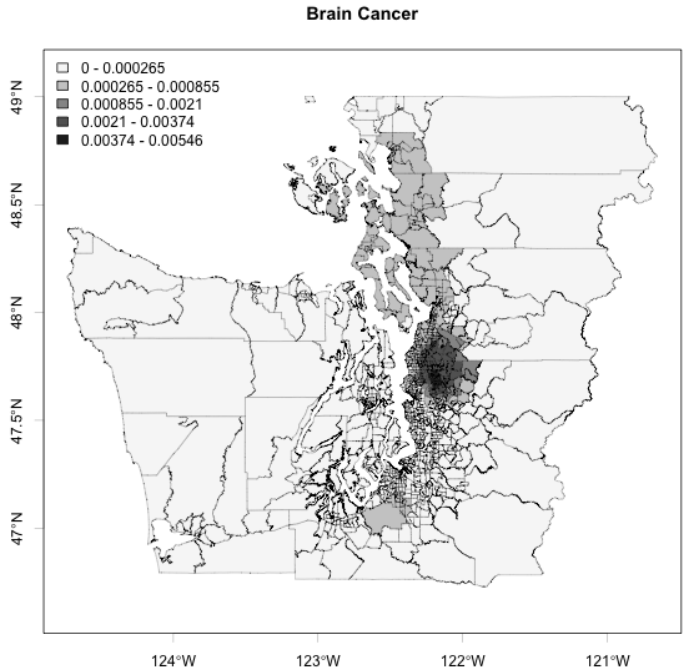
The Actual Map

Example: Brain & Breast Cancer in Western Washington

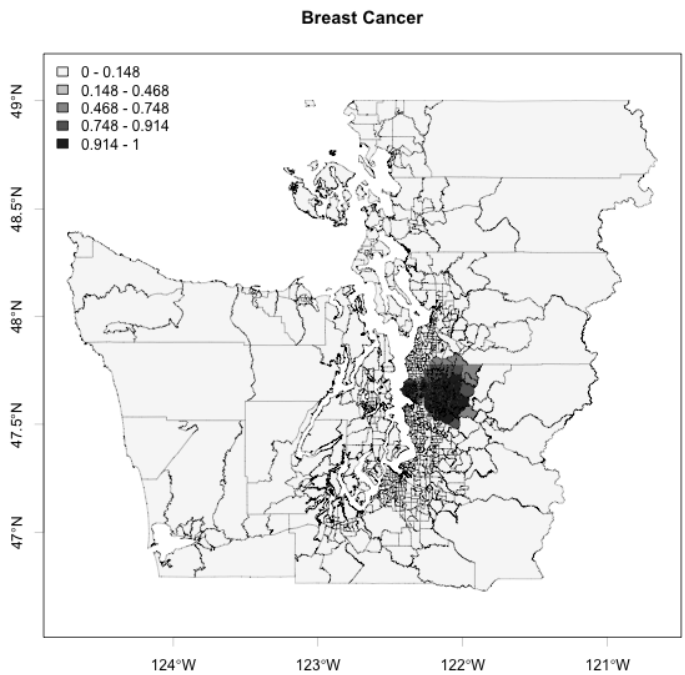
My PhD dissertation involved detecting cancer "clusters": areas of residual spatial variation of disease risk.

We modeled the (Bayesian) probability of cluster membership for each of the $n = 887$ census tracts in Western Washington in 2000, using cancer data from 1995–2005, controlling for age, race, and gender.

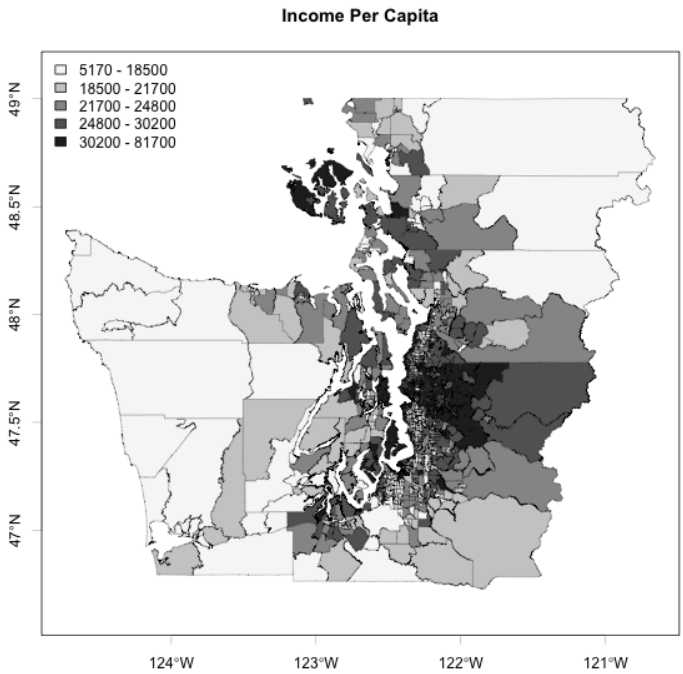
Brain Cancer Controlling for Age, Race, & Gender



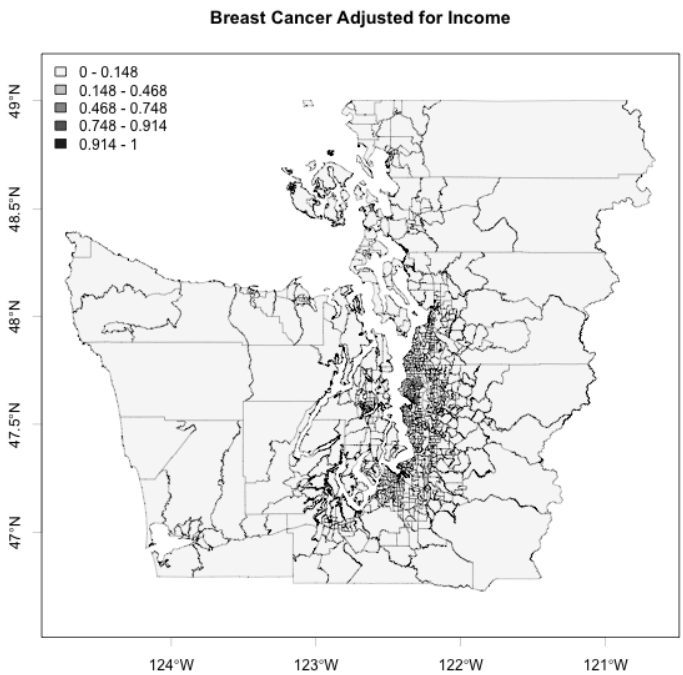
Breast Cancer Controlling for Age, Race, & Gender



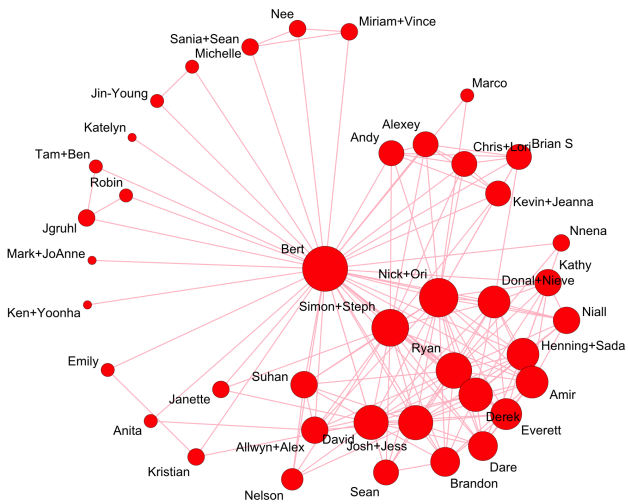
Income per Capita Quintiles



Breast Cancer Adjusted for Income as Well



Example: Social Network Display of a Recent Party I Had



Say we want answer the following questions:

- Does a new kind of cognitive therapy alter levels of depression in patients?
- Or you question the effectiveness of antioxidants in preventing cancer.
- Will reassuring potential new users to a gambling website that we won't spam them increase the sign-up rate?

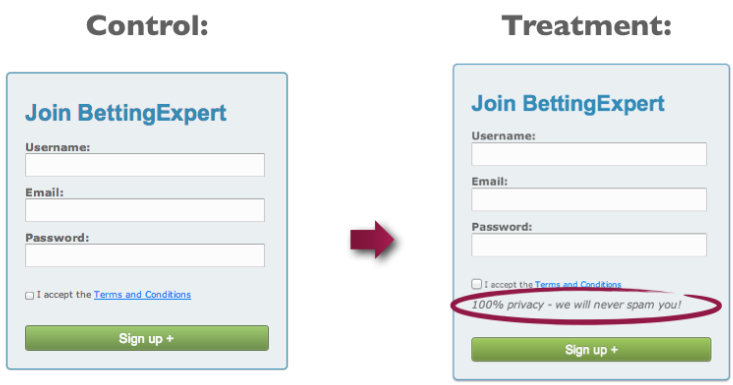
Evaluating the efficacy of a 'treatment'

In all the above cases, you are questioning the efficacy of a treatment/intervention. One way to evaluate the efficacy is via an experiment where you define

- A control group: the "business as usual" baseline group
- A treatment group: the group that receives/is subject to the treatment/intervention

and make comparisons.

Website Experiments



Example of a treatment vs control

Two other examples in the media of late

- Facebook's tinkering with user's emotions (link)
- OkCupid's admission that they experiment on human beings (link)

Variables

A variable is a description of any characteristic whose value may change from one unit in the population to the next:

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis
- column represents the variables corresponding to a particular observation

It is almost always the case that

- n is the number of observations
- p is the number of variables

Data Summaries

Consider the variable "federal spending per capita" in each of the 3,143 counties in the US. One can hardly digest this:

```
   [1]   6.068095   6.139862   8.752158   7.122016   5.130910   9.973062   9.311835  15.439218
   [9]   8.613707   7.104621   6.324061  10.640378   9.781442   8.982702   6.840035  20.330684
  [17]   9.687698  11.080738   7.839761   9.461856   9.650295   7.760627  25.774791  13.948106
  ...
[3121]   7.520731  10.246400   3.106800  17.679572   4.824044   7.247212   8.484211   8.794626
[3129]   9.829593   8.100945  17.090715   4.855849   6.621378  22.587359  10.813260  11.422522
[3137]   9.580265   4.368986   5.062138   6.236968   4.549105   8.713817   6.694784
```

Data Summaries

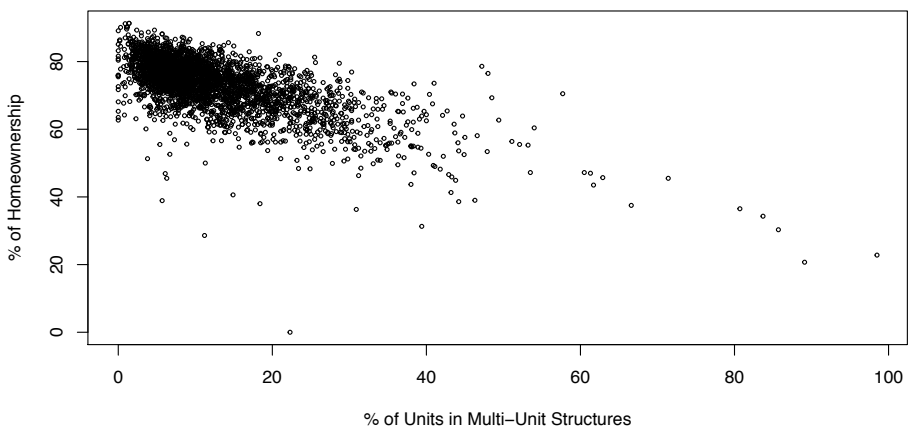
We can't interpret all the data at once; we need to boil it down via summary statistics, single numbers summarizing a large amount of data.

Using the `summary()` command in `R`:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   6.964   8.669   9.991  10.860 204.600       4
```


Relationships between variables

We can best display the relationship between two variables using a scatterplot AKA bivariate plot:



Relationships between variables

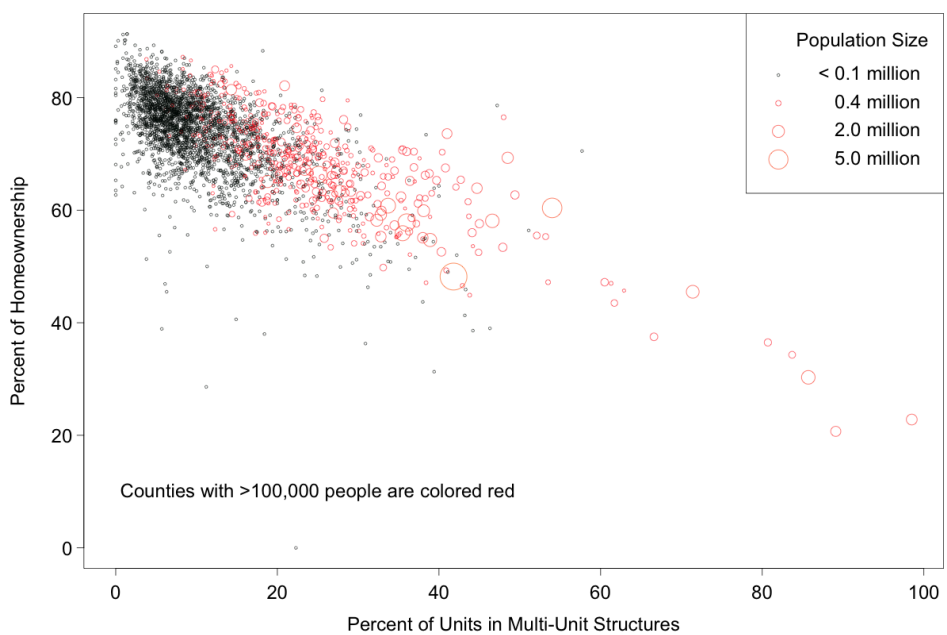
Almost always we are interested in the relationship between two or more variables.

A pair of variables are either related in some way (associated) or not (independent). No pair of variables are both associated and independent.

We can have either a negative association (as the value of one variable increases, the other decreases) or a positive association.

Relationships between variables

We can consider a third variable in the previous plot.



Lecture 2: Sampling and Bias

Chapter 1.3

Goals for Today

- Understand important considerations about data collection, in particular sampling.
- Food for thought about the next lecture: explanatory/response variables and causality.

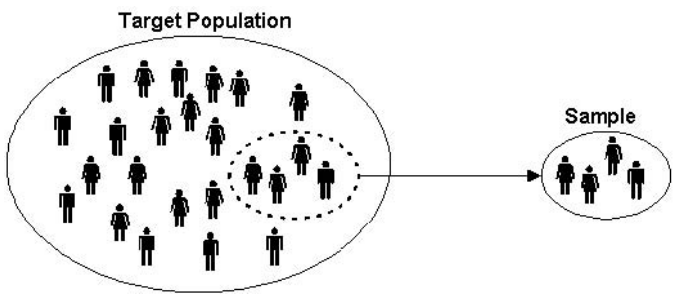
Populations and Samples

We want to make statements about some aspect of a study/target population.

1. What proportion of Oregonians smoke?
2. What are the sexual behaviors of males and female Americans in 1948?
3. What proportion of the Reed community believes they have personally experienced offensive, hostile, or intimidating conduct on campus?

Populations and Samples

It is often not feasible to collect data for every case in the population. If so, we take a sample of cases.



If the sample is representative of the desired population then our results will be generalizabile.

Populations and Samples

So say we take a representative sample of 1000 Oregonians and poll their smoking habits. We can then generalize the results to the entire population of Oregon.

One example of a non-representative sample is a biased sample.

How do we take a representative sample? In its simplest form, you need to randomly sample from the entire population. But this is easier said than done.

Comment on the Representativeness of These Samples:

1. The Royal Air Force wants to study how resistant their airplanes are to bullets. They study the bullet holes on all the airplanes on the tarmac after an air battle against the Luftwaffe (German Air Force).
2. I want to know the average income of Reed graduates in the last 10 years. So I get the records of 10 randomly chosen Reedies. They all answer and I take the average.
3. Imagine it's 1993 i.e. almost all households have landlines. You want to know the average number of people in each household in Portland. You randomly pick out 500 phone numbers from the phone book and conduct a phone survey.
4. You want to know the prevalence of illegal downloading of TV shows among Reed students. You get the emails of 100 randomly chosen Reedies and ask them "How many times did you download a pirated TV show last week?"

Statistics in Society: Alfred Kinsey

In the mid 20th century, biologist/sexologist Alfred Kinsey wanted to study human sexuality.

At the time sexuality was an extremely taboo subject, very little research had been conducted at that point and Kinsey was astonished at the public's general ignorance.

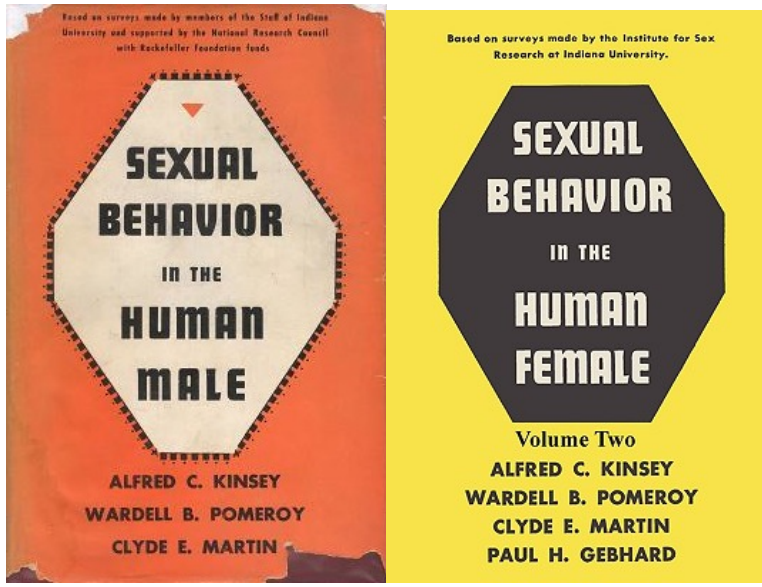
Statistics in Society: Kinsey's Questions/Research Problem

What type of questions was Kinsey interested in? Using his 300 question survey, he hoped to address...

1. What percentage of Americans engaged in premarital and extramarital sex?
2. What were the homosexual tendencies of American males?
3. How common were oral sex and masturbation?
4. ...

Statistics in Society: Kinsey Reports

The results were published two books on human sexual behavior known as the "Kinsey Reports": Sexual Behavior in the Human Male (1948) and Female (1953).



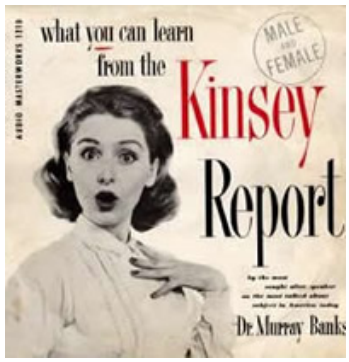
Statistics in Society: Conclusions of Kinsey Reports

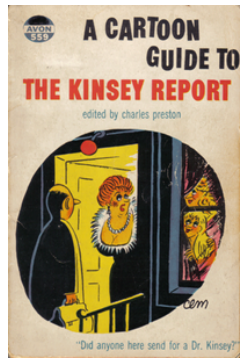
Kinsey claimed, among other things

1. 85% of white men had had premarital sex, 50% had had extra-marital sex
2. Kinsey wrote in 1948 that one in ten white men were more or less, exclusively homosexual for at least three years between the ages of 16 and 55.
3. Kinsey reported that oral sex was very common (70% of couples did it), masturbation was very common (almost 63%/92% of women/men did it)

Statistics in Society: Reaction to Kinsey Reports

Needless to say, people were taken quite aback.





There was also a huge conservative backlash against the reports.

Statistics in Society: Kinsey's Methods

What were his data collection methods? How did he sample his data? Focusing on the male report, my understanding is that

1. He did in fact base his conclusions on a very large sample size of 5300 males.
2. He sought out volunteers to answer his 300 question survey.
3. He recruited new people by asking previous respondents if they knew other people. This led to a large proportion of his sample to include prison populations and male prostitutes.

What could be some issues?

Response of the American Statistical Association

The American Statistical Association criticized the sampling procedure. In particular, John Tukey, one of the most eminent statisticians of the time, said

> *"A random selection of three people would have been better than a group of 300 chosen by Mr. Kinsey."*

Even though the Kinsey Report was groundbreaking and contributed much to the field of sexology by bringing many topics to the forefront, Kinsey's statements were not generalizable to the general public.

Reed Campus Climate Survey

During the 2012-2013 academic year Reed contracted Rankin & Associates Consulting to conduct the Campus Climate Survey to "examine the learning, living, and working environment at Reed College."

On page v and iii of the Executive Summary:
http://www.reed.edu/institutional_diversity/campus_climate.html

Examples of Different Types of Bias:

Moral of the Story

For you:

1. the consumer of statistics: Ask yourself what was the study design?
   - Who is the study population?
   - Who are the respondents and how were they selected?
2. the producer of statistics: think about how you will collect your data beforehand. If you want your results to generalize beyond just your sample to your study population, your sampling scheme has to as representative as feasible.

Explanatory and Response Variables

Example: A medical doctor pours over some his patients' medical records and observes:

$\Longrightarrow$

He then posits the following causal relationship:

- Explanatory variable: sleeping with shoes on
- Response variable: waking up with headaches

What's wrong with hypotheses?

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.

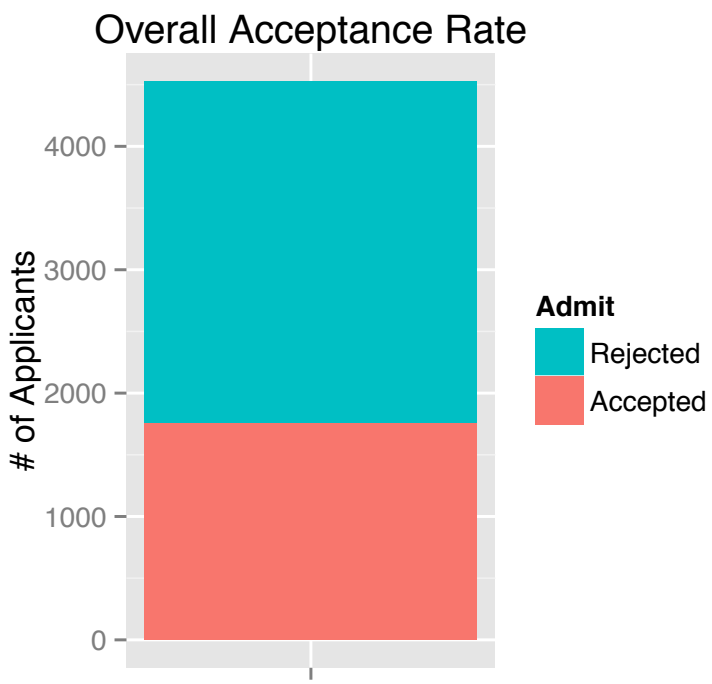
- Spurious correlations: http://www.tylervigen.com/
- Saturday Morning Breakfast Cereal: http://www.smbc-comics.com/?id=3129

Well-Known Example of Confounding

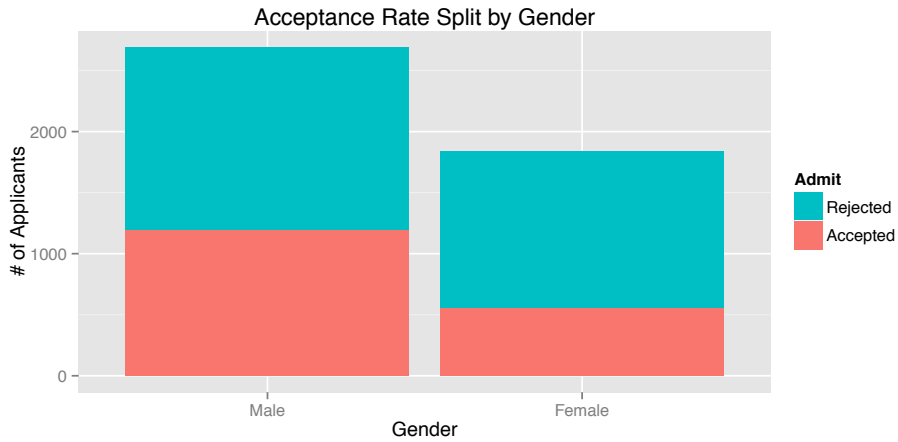
A famous example of an unaccounted for confounding variable having serious repercussions was when the UC Berkeley was sued in 1973 for bias against women who had applied for admission to graduate schools.

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

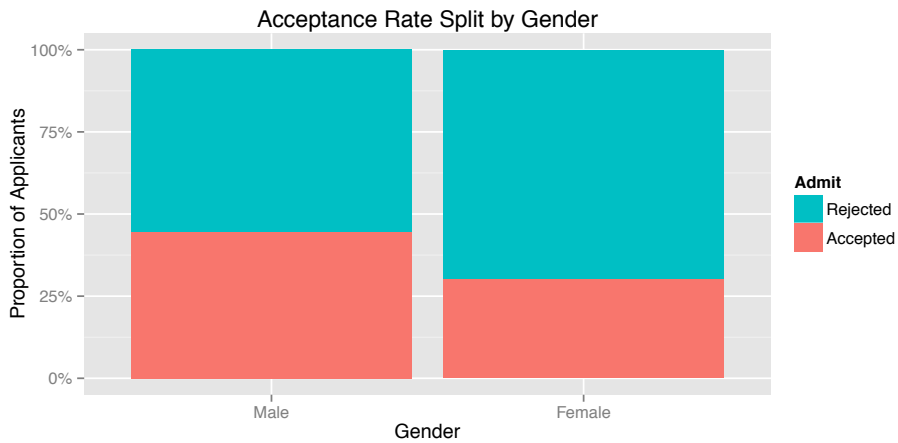
Of the $n = 4526$ applicants:



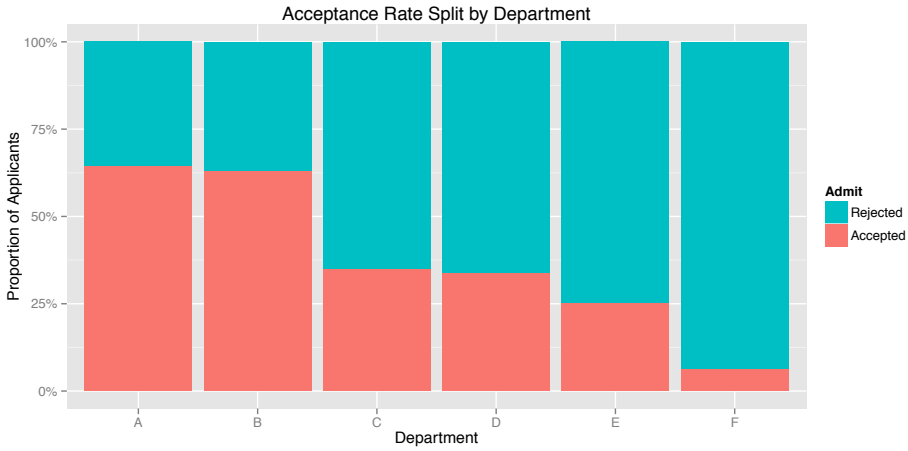
Split the counts by gender:



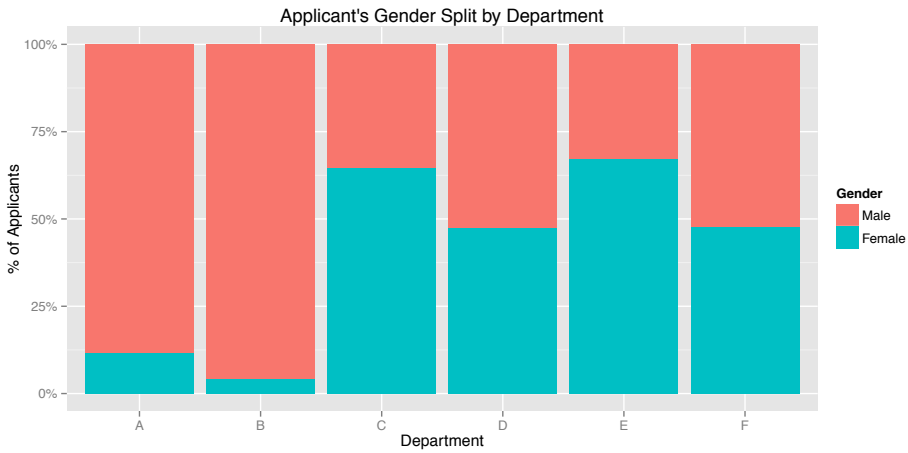
Look at proportions instead of counts:



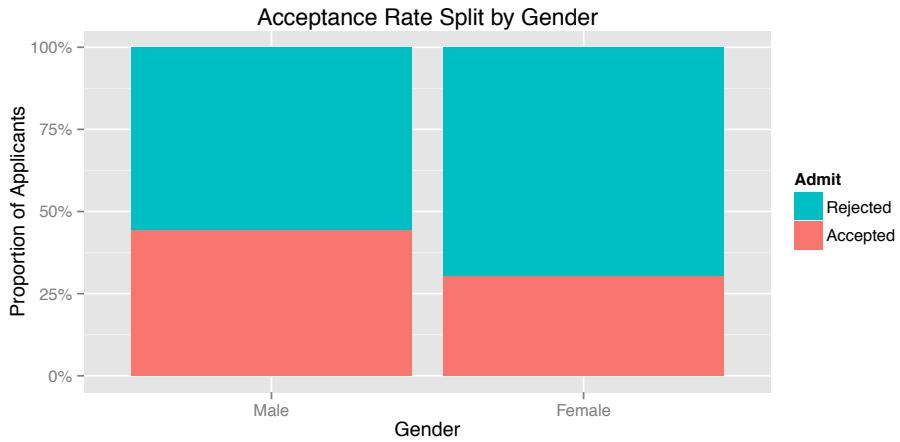
What was the "competitiveness" of departments?



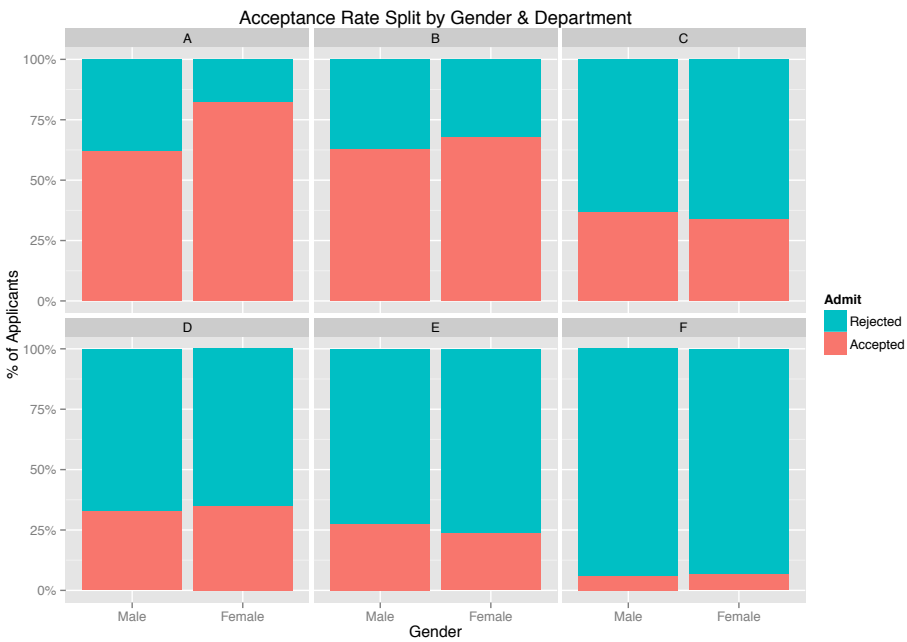
Where were the women applying?



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.

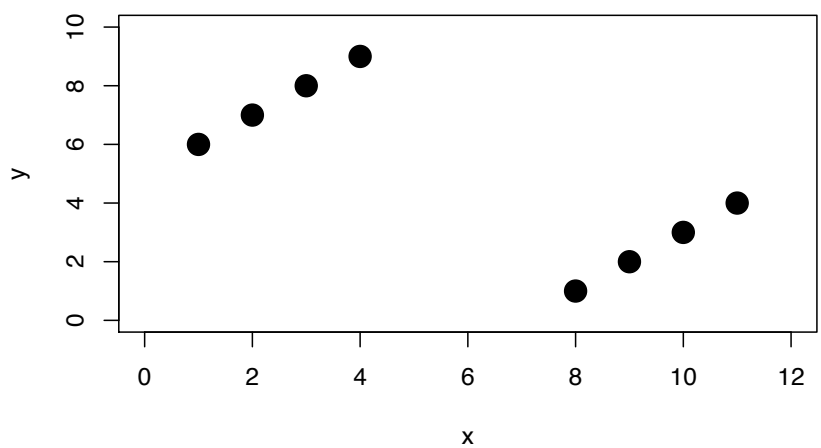
Simpson's Paradox

(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.

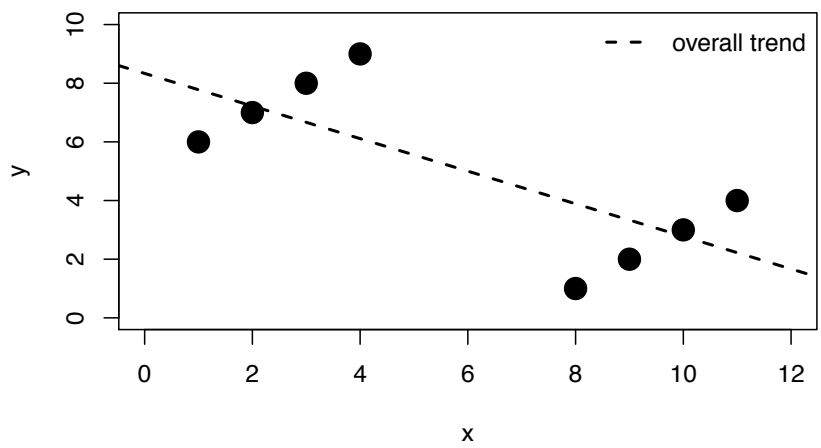
A Graphical Illustration of Simpson's Paradox

Say we have the following points:



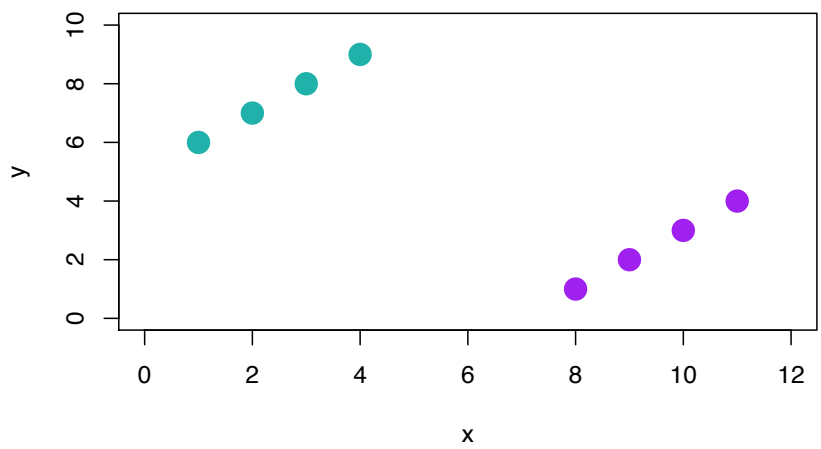
A Graphical Illustration of Simpson's Paradox

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



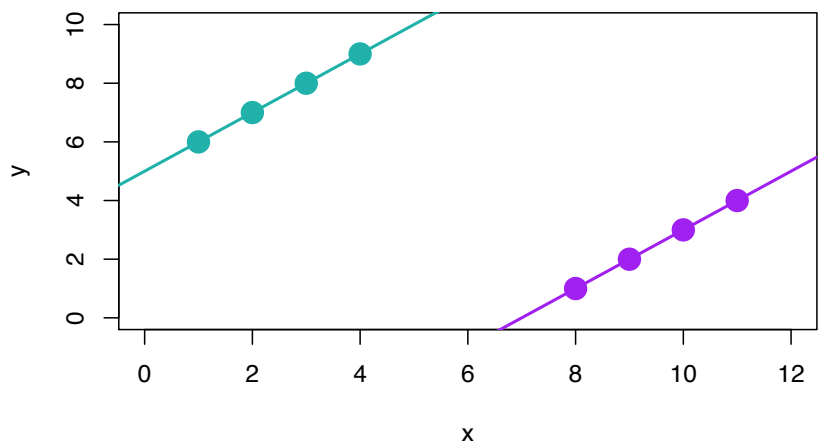
A Graphical Illustration of Simpson's Paradox

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



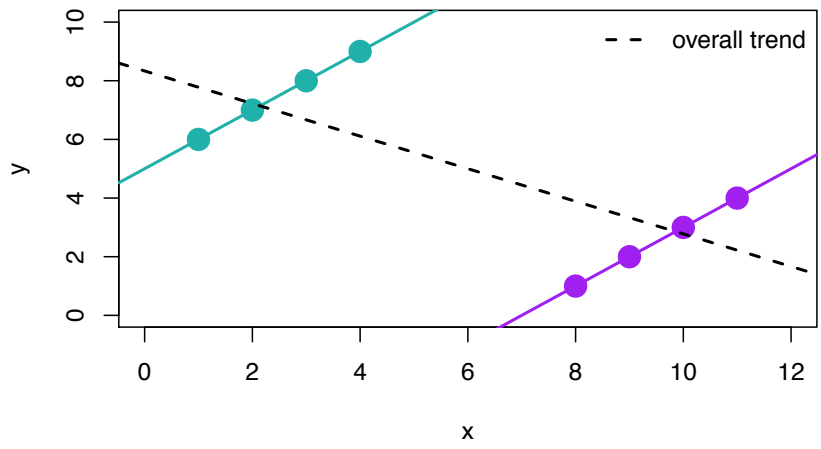
A Graphical Illustration of Simpson's Paradox

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.