Lecture 2: Data Collection + Sampling

Chapter 1.3

Goals for Today

- Understand important considerations about data collection, in particular sampling.
- Food for thought about the next lecture: establishing causality

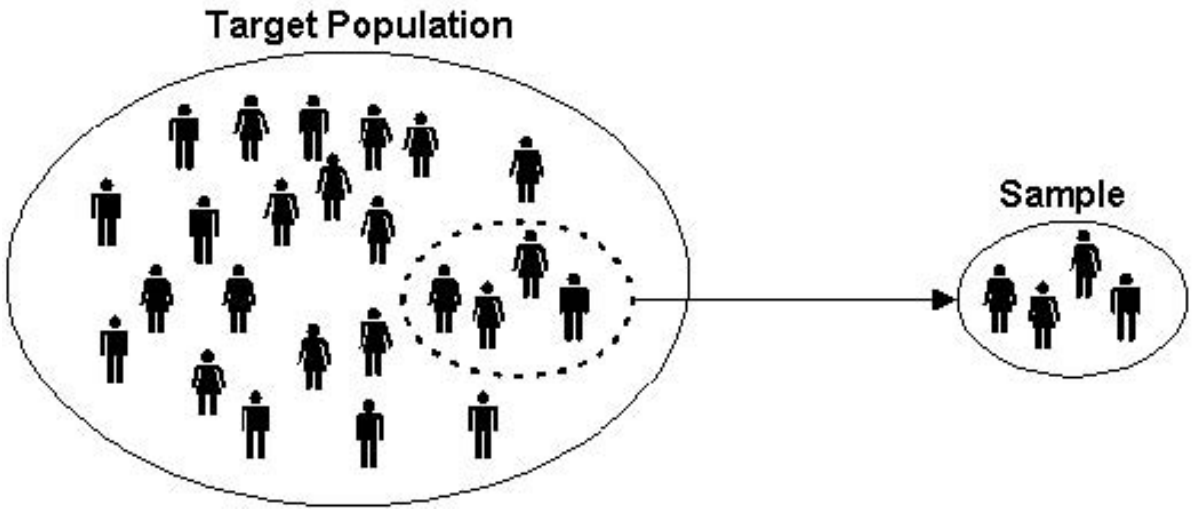
Populations and Samples

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

1. What proportion of Oregonians smoke?
2. What proportion of televisions in Afghanistan were tuned into President Obama's state of the union address yesterday?
3. What is the average mercury content in swordfish in the Atlantic Ocean?

Populations and Samples

However, it is often really unrealistic/inconvenient/expensive to collect data for every case in the population. Therefore, we take a sample of cases.



If the sample is representative of the desired population then our results will generalize. This is called generalizability.

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, the way for your sample to be representative is to randomly sample from the population. But this is easier said than done (more later).

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 about sexual matters.

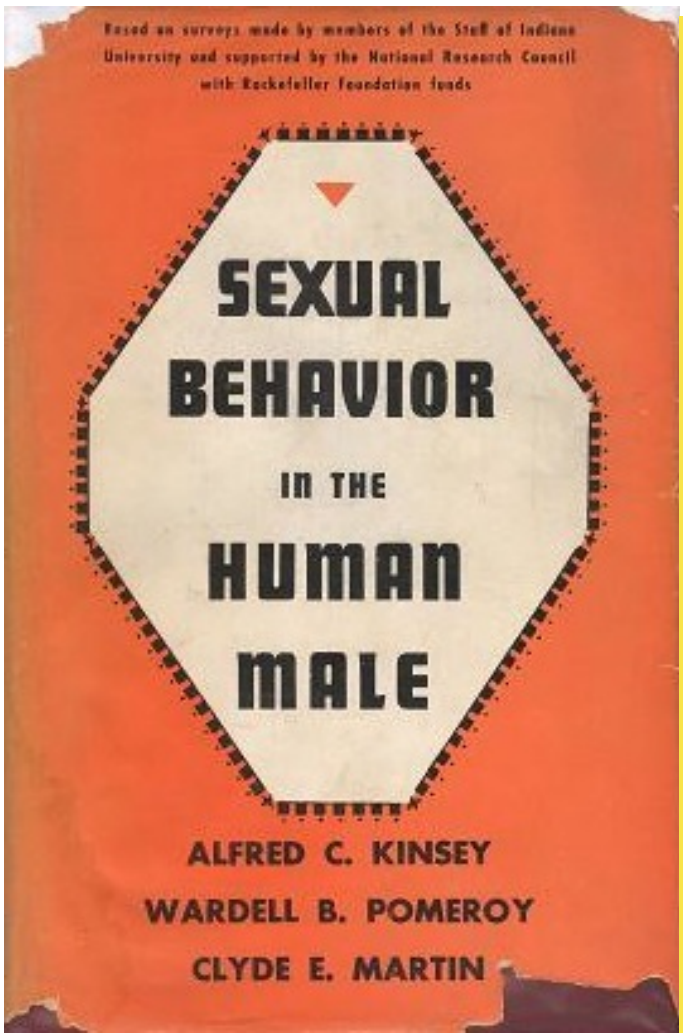
Statistics in Society: Kinsey's Questions/Research Problem

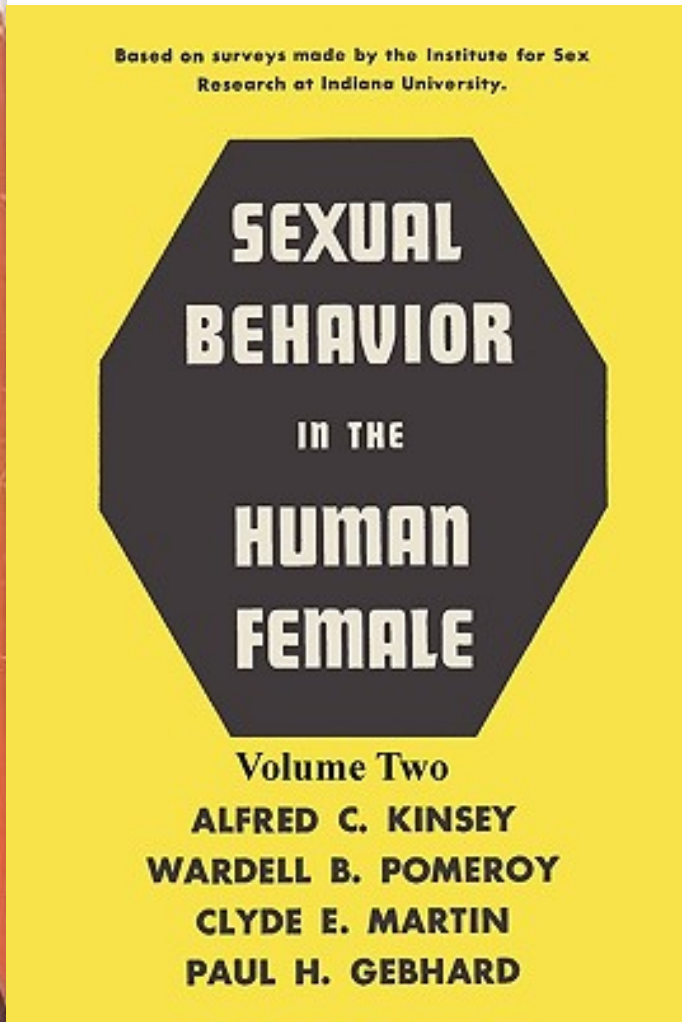
What type of questions was Kinsey interested in? Using his 300 question survey, he hoped to address...

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 Sexual Behavior in the Human Female (1953).





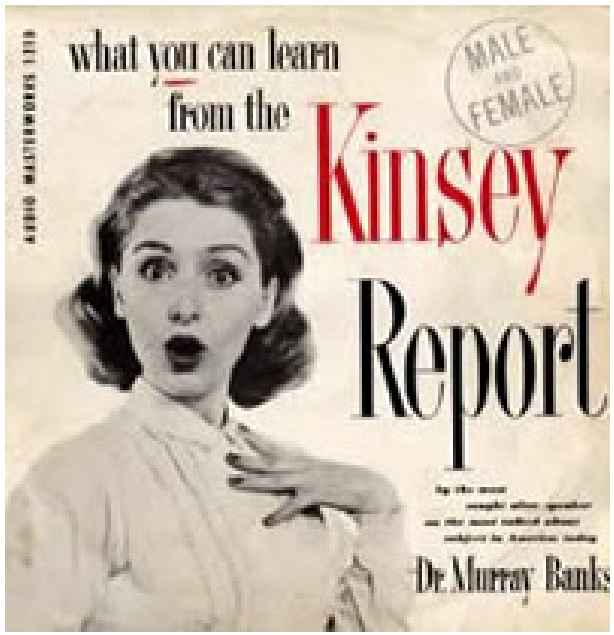
Statistics in Society: Conclusions of Kinsey Reports

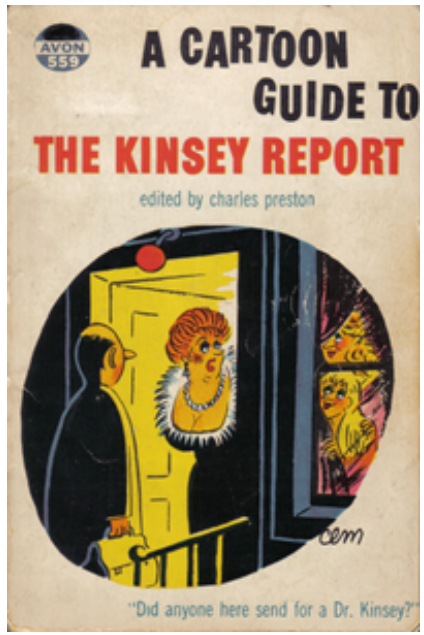
Kinsey claimed, among other things

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

BUT WAIT! 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 gathered convenience samples of people who knew each other. 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, for instance his "one in ten" statement, were not generalizable to the general public.

Moral of the Story

For you:

1. the consumer of statistics: before trusting any conclusions of studies making statements about a population based on a sample, you should ask: What was the study design? What was the sampling scheme?
2. the producer of statistics AKA the researcher: put a lot of thought into 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. Consult a statistician if you're not sure!

Different Types of Bias:

1. Volunteer bias: individuals who are more willing to participate have a higher chance of being sampled.
2. Convenience sample bias: individuals who are easily accessible are more likely to be included.
3. Survival bias: large segments of the population who "died" are not sampled.
4. Selection bias: some individuals are more likely to be selected for study than others.

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?