Lecture 2: Sampling and Bias

Chapter 1.3

Goals for Today

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

Recall: What is statistics?

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

Point 2 is just as, if not, more important than point 3.

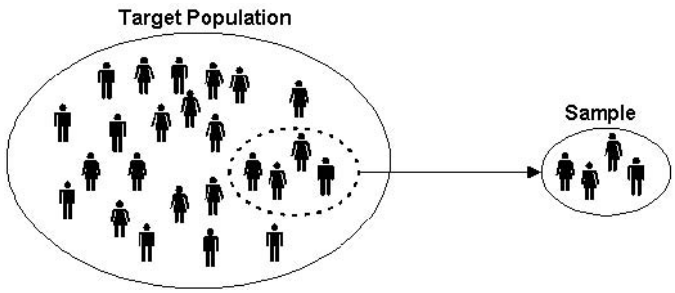
Populations and Samples

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

1. What proportion of Vermonters smoke?
2. What are the sexual behaviors of males and female Americans in 1948?

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.



Important:

Populations and Samples

So say we take we collect a sample of 100 Vermonters and poll their smoking habits in two ways:

- We stand outside a tobacco store and randomly select 100 people walking nearby. This is a non-representative sample AKA a biased sample.
- We have a list of all citizens of the state and randomly select 100 people from this list. This sample is representative and hence the poll's results can generalize to all of VT.

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.

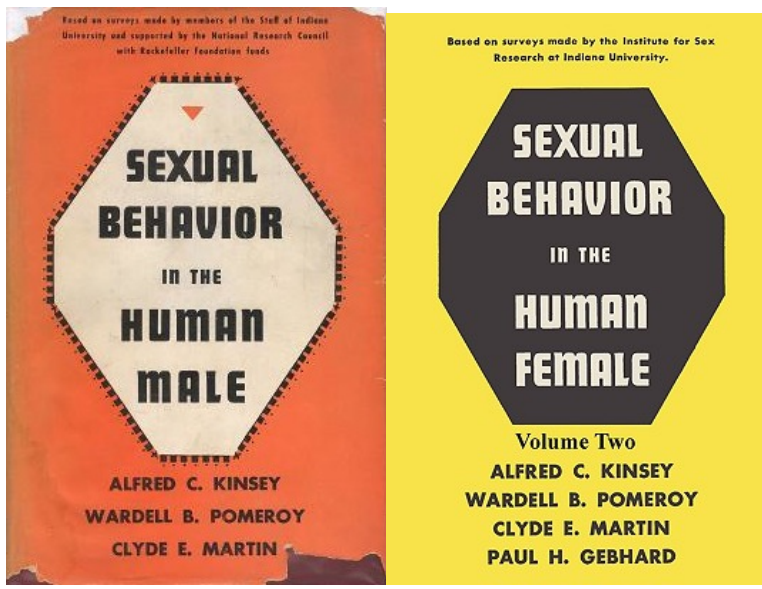
Statistics in Society: Kinsey's Questions/Research Problem

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).



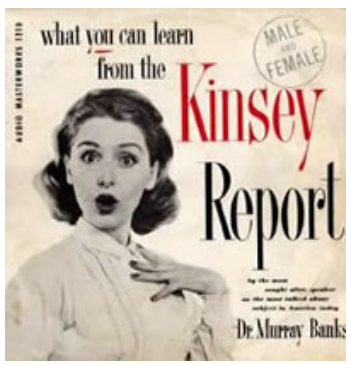
Statistics in Society: Conclusions of Kinsey Reports

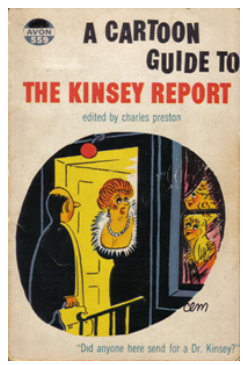
Kinsey claimed, among other things

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.

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: If you want your results to generalize beyond just your sample to your study population, your sampling scheme has to as representative as feasible.

Another Example: Facebook

In the news

- NPR's story on younger users (link)
- Facebook Envy (link)

Let's consider a hypothetical scenario where we compare 15 life occurrences between you and your friend rated between 0 (lowest) and 5 (highest).

Another Example: Facebook

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg
you																
friend																

Another Example: Facebook

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg
you	3	1	2	3	3	3	3	1	1	3	4	2	1	4	3	2.5
friend																

Another Example: Facebook

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg
you	3	1	2	3	3	3	3	1	1	3	4	2	1	4	3	2.5
friend	1	5	2	0	4	2	4	1	0	1	4	4	1	5	3	2.5

Another Example: Facebook

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg
you	3	1	2	3	3	3	3	1	1	3	4	2	1	4	3	2.5
friend's FB		5			4		4				4	4		5		4.3
friend	1	5	2	0	4	2	4	1	0	1	4	4	1	5	3	2.5

The selective "Facebook image curation" your friend performed is a form of selection bias!

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?