Lecture 3: Data Collection and Sampling Strategies

Sources of Data

Three common sources of data we'll discuss:

1.

2.

3.

Example: Does the health of a male cricket impact its ability to successfully find a mate?

Observational Studies vs. Experiments

Experiments have one **major** advantage over observational studies:

Observational studies cannot be used to establish causation due to...

Example: Ice cream sales

Example: "Miracle drugs" and weight loss

Example: A childcare study enrolled 1364 infants in 1991 and followed them through age 6. Researchers found the more time children spent in childcare from birth to $4\frac{1}{2}$, the more adults tended to rate them as assertive, disobedient, and aggressive.

Type of data collection?

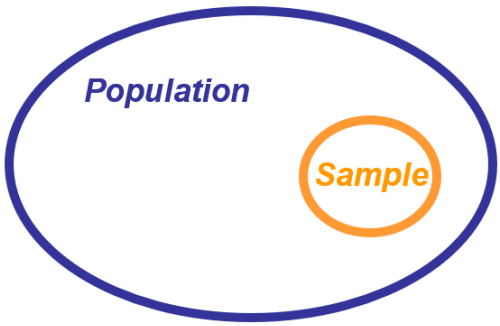
Explanatory and response variables?

Possible lurking/confounding variables?

An _____ was probably impossible here but, hypothetically, how might it have proceeded?

Observational Studies: Terminology

1. **Population**:



2. **Sample**:

Example: We want to know the distribution of student loan amounts for UNC undergraduates.

- What would a census look like?

- How about a sample survey?

Census vs. sample survey: pros and cons

Cooking metaphor:

For your inference to be valid,

Example: Battery manufacturer

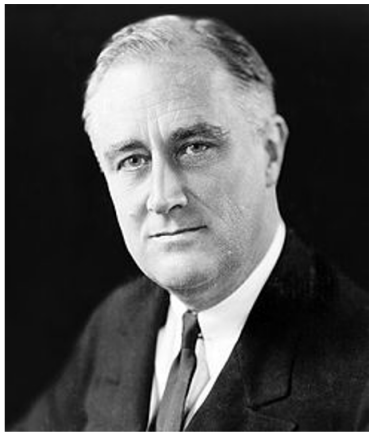
Population of interest?

How to choose the 24 batteries for inspection?

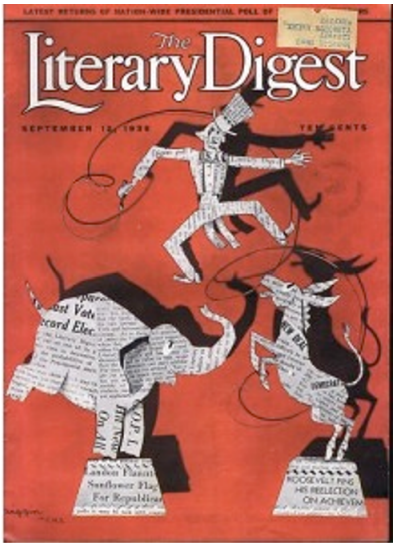
Sampling Strategies for Observational Studies

Major pitfall: **sampling bias**

In 1936, Alf Landon was the Republican nominee opposing the re-election of Franklin Roosevelt.



- The Literary Digest magazine polled about 10 million Americans, and got responses from about 2.4 million.
- Poll showed that Landon would likely be the overwhelming winner and Roosevelt would get only 43% of vote.
- Election result: Roosevelt won, with 62% of the vote.
- The magazine was completely discredited because of the poll, and was soon discontinued.



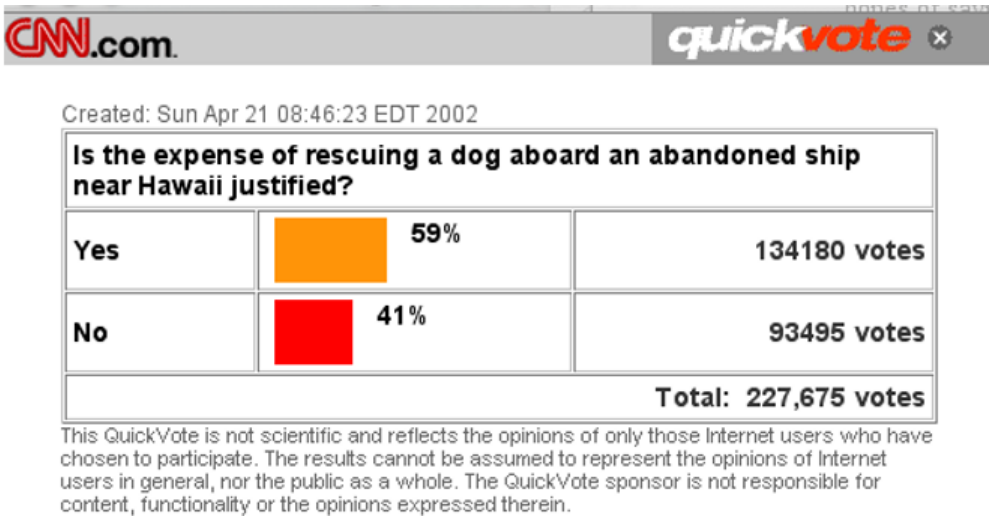
What went wrong?

Other possible sources of sampling bias:

- **Non-response**

- **Voluntary response**

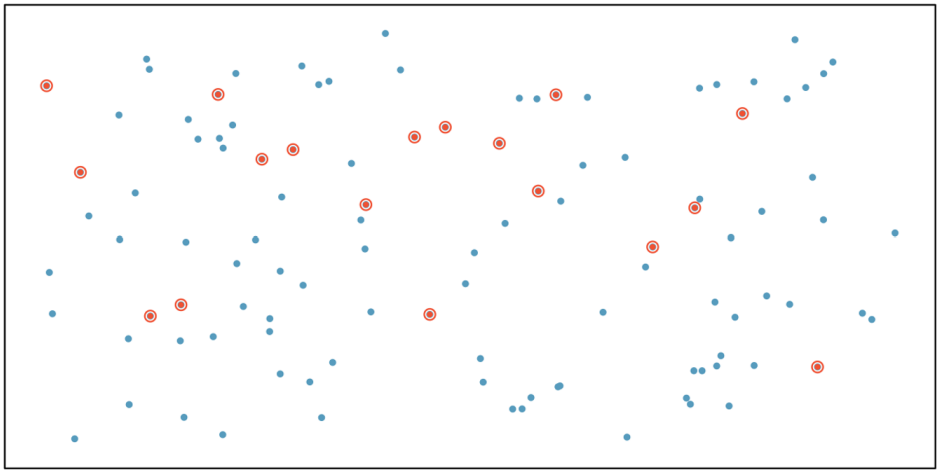
- **Convenience**



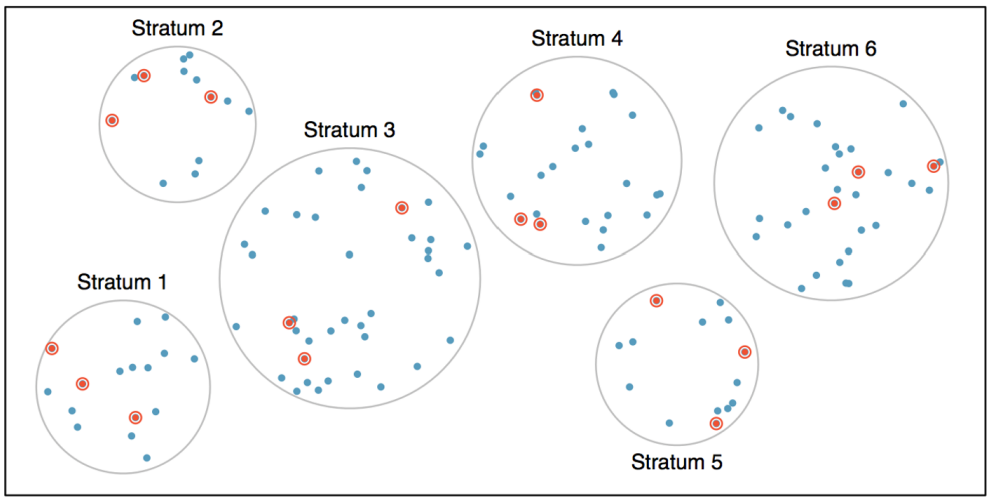
Arguably the most fundamental characteristic of good sampling techniques that seek to avoid bias is...

"Good" sampling techniques:

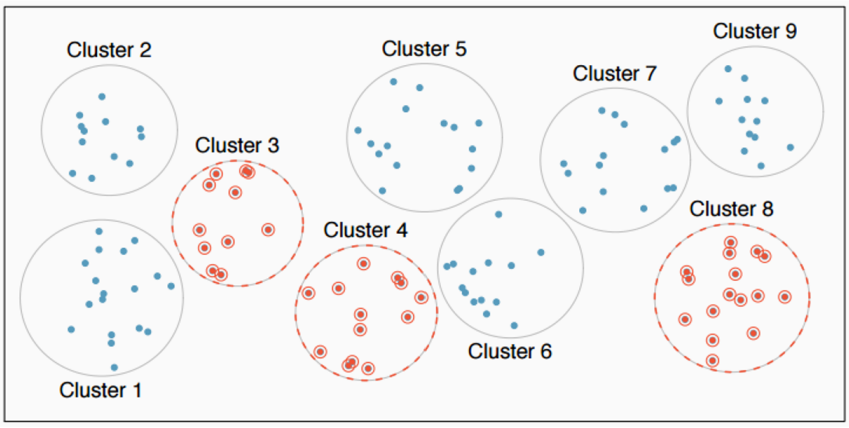
- **Simple Random Sample**



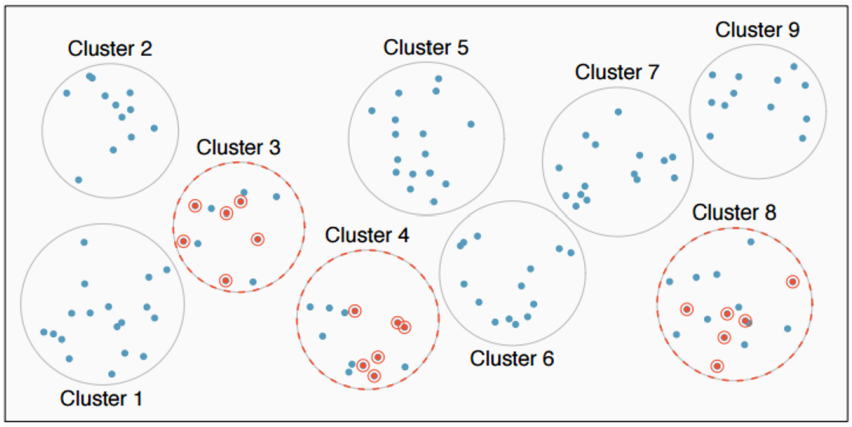
- **Stratified Sampling**



- **Cluster Sampling**



- **Multistage Sampling**



Other factors that can bias/influence results:

- Wording of questions

Survey of high-school students:

- "Which is easier for someone of your age to buy: cigarettes, beer, or marijuana?" (35%, 18%, 34%)
- "Which is easier for someone of your age to obtain: cigarettes, beer, or marijuana?" (39%, 27%, 19%)

Poverty Assistance:

- "Is US spending too much on assistance to the poor?" (13%)
- "Is US spending too much on welfare?" (44%)

- Framing of questions

Fewer people mention the economy in open-ended version

% answering that the issue matter most in deciding their vote for president in 2008

	Open-ended	Closed-ended
The economy	**35**	**58**
The war in Iraq	5	10
Health care	4	8
Terrorism	6	8
Energy policy	*	6
Other	43	8
Candidate mentions	9	–
Moral values/social issues	7	–
Taxes/distribution of income	7	–
Other issues	5	–
Other political mentions	3	–
Change	3	–
Other	9	–
Don't know	7	2
	100	100

Note: Open-ended figures reflect respondents' unprompted first response. Close-ended figures reflect respondents' first choice from five options read by the interviewer.
Source: Survey conducted November 2008.

PEW RESEARCH CENTER