

## Topic 2: Data Collection and Sampling Strategies

### Sources of Data

Three common sources of data we'll discuss:

1. **Anecdotes** Individual stories;  
single or few cases  
providing evidence → **BAD (as scientific inquiry)**

**'Supermothers' and  
grandfather lift 1 ton Renault  
Clio off trapped schoolboy**

MIKE HAS LOST 30LBS OF FAT

MIKE

RATING: ★★★★☆

AGE GROUP: 46 - 60

GENDER: Male

GOAL:

• Fat Loss



## Do Vaccines Cause Autism?

2. **Observational studies** Passive data collection;  
gather data without  
attempting to influence  
responses → **GOOD (as scientific inquiry)**
3. **Experiments** Active data collection;  
deliberately influence  
responses with some  
treatment

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

Observational study:

Catch some male crickets  
Measure their health (somehow)  
Release the crickets  
Track to see which ones successfully mate

Experiment:

Catch some male crickets  
Infect half the crickets with an intestinal parasite  
Release the crickets  
Track to see which ones successfully mate

### Observational Studies vs. Experiments

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

Experiments are the **ONLY** reliable method of...

### ESTABLISHING CAUSATION

**Causation:** cause-and-effect, e.g.  
"change in X **causes** change in Y"

Thus, experiments are considered the gold standard of research methods. But observational studies can still be extremely useful!

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

...their lack of ability to control for  
**lurking/confounding variables**

**Confounding variable:** a factor other than  
your explanatory variable that might affect  
what you observe in your response variable

Example: Ice cream sales

An observational study of New York beach towns found a strong positive relationship between ice cream sales and deaths by drowning.

Can we conclude that ice cream causes drowning?

NO - confounding variables like temperature, higher crowd volumes, the fact that this is a beach town instead of (e.g.) a state fair...

Observational studies require high awareness of confounding variables!

Example: "Miracle drugs" and weight loss

An experiment took 100 people interested in losing weight, gave 50 of them a new weight loss drug, and did not give the drug to the other 50. Greater weight loss was observed in the treatment group.

Can we conclude that the drug is effective?

NO - confounding variables like diet, exercise regimen, lifestyle, metabolism, underlying health conditions, genetics, motivation...

Experiments still have to worry about confounding variables!

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?

Observational study

Explanatory and response variables?

Explanatory: Amount of time in childcare

Response: behavior ratings by adults

Possible lurking/confounding variables?

Conditions of parents - how many parents in the home, socioeconomic status

Home environment - pets, siblings, etc

Conditions at childcare - class size, behavior of others, resources

Who is doing the rating?

...the list goes on!

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

Hypothetically...

- control for home environment by restricting to kids in the same household or similar households (by socioeconomic status, # of parents, siblings/pets...)
- dictate how much time each child spends in childcare
  - This one is ethically murky, depending on implementation

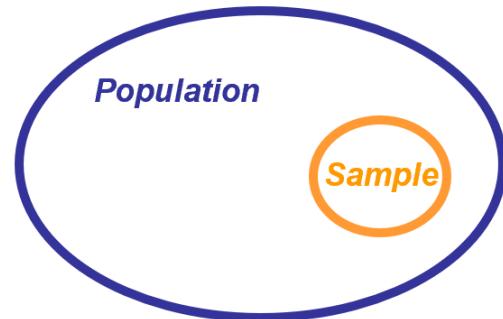
**WATCH OUT** when thinking about making adjustments like this! There's a reason we do observational studies instead of all-experiments-all-the-time, and it's usually related to ethics, logistics, or both.

## Observational Studies: Terminology

### 1. Population:

**Population:** the entire group you want to study / say something about

Ex: UNC undergraduates  
 Americans  
 crickets in North Carolina  
 swordfish in the Atlantic Ocean  
 patients with Type I diabetes  
 bottles of Coca-Cola produced at the Charlotte bottling plant



### 2. Sample:

**Sample:** a subset of the population

It is often not feasible - due to time, funding, or other logistic constraints - to acquire data on an entire population.

We use samples to make "educated guesses" about a population.

(These are VERY GOOD "educated guesses".)

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

- What would a census look like?

**Census:** a collection of data from the full population

Here, a census would require aggregating the student loan amounts for every single undergraduate student at UNC.

- How about a sample survey?

**Survey:** a collection of data from a sample

A survey might look like collecting the student loan amounts of 1000 students.

For the survey to be useful, you should make sure the sample is evenly balanced across graduating classes and fields of study...

Census vs. sample survey: pros and cons

**CENSUS:** Pro - accuracy (comprehensive data!)  
Con - difficulty (may not be logically feasible)

**SURVEY:** Pro - quicker/easier (needs less resources)  
Con - less accurate (lots less data...)

Imagine cooking a pot of soup. To test if there is enough salt in your soup, you taste the broth.

Cooking metaphor:

You **infer** the state of the full pot of broth (population) according to this taste (sample) in order to decide whether to add more salt.

**Inference:** a conclusion (or "educated guess") about the state / behavior / condition of an entire population based on observations of a sample.

For your inference to be valid,

...your pot needs to be well-stirred, so that the taste test is **representative** of the rest of the pot.

Example: Coca-Cola bottling plant

We want to check bottle quality! We have access to a full day's production for sample collection, and need 24 bottles.

Population of interest?

All bottles of Coca-Cola produced by the plant

(Note: by restricting to sampling from a particular day, we may already not be fully representing the population. But this is a necessary logistic compromise.)

How to choose 24 bottles for inspection?

Some options:

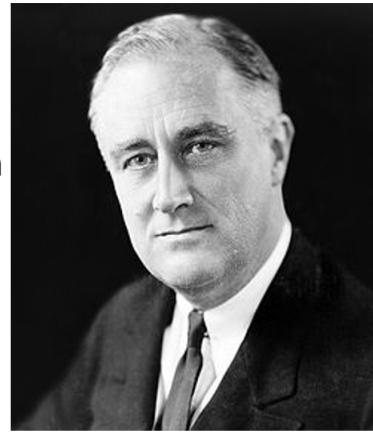
- Grab 1 bottle per hour, over a day
- Take every nth bottle, where n is equal to the total number produced in a day divided by 24
- Take a day's production, toss them all in a bin, mix them around, and select 24 at random (I advise that you do not attempt this method literally)

## Sampling Strategies for Observational Studies

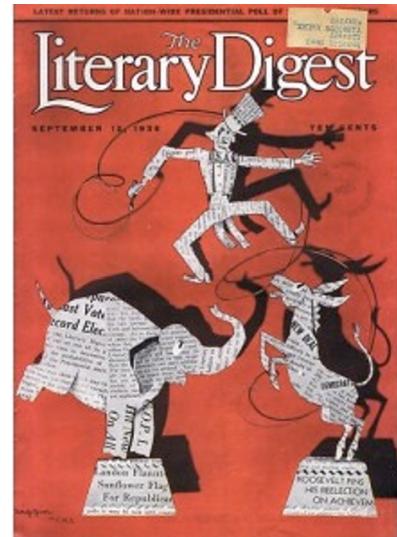
Major pitfall: **sampling bias**



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?

### SAMPLING BIAS!!!

Sampled from:

- Readers of Literary Digest
- Registered automobile owners
- People with working phone numbers



Resulting sample:

- was predominately wealthy
- tended towards mostly well-educated
- Mostly omitted working-class Americans!

Other possible sources of sampling bias:

- Non-response

Majority of sampled individuals do not respond to survey



Even if your PLANNED sample was representative, the RESULTING sample may not be!

- Voluntary response

Mostly hear from people with strong opinions (and free time)



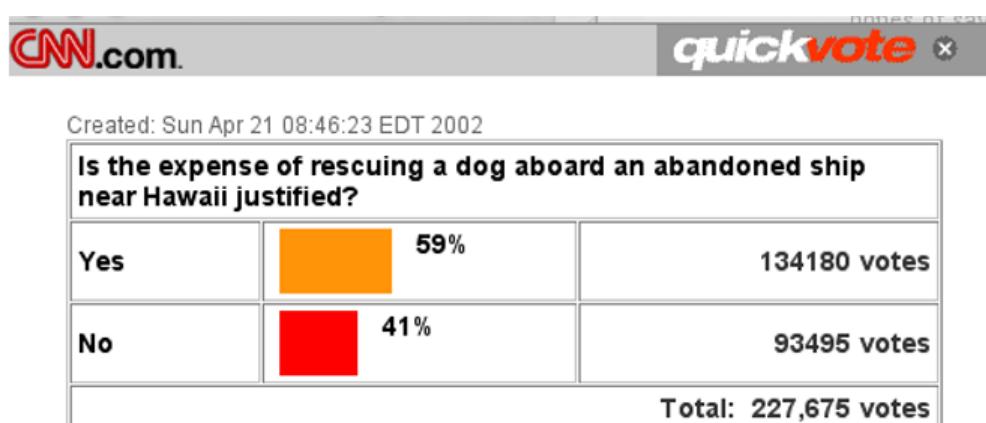
Even if your PLANNED sample was representative, the RESULTING sample may not be!

- Convenience

Get data primarily from easy-to-reach survey participants



Sample can ONLY be representative of this easy-to-reach group; you miss out on all other data!



This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole. The QuickVote sponsor is not responsible for content, functionality or the opinions expressed therein.

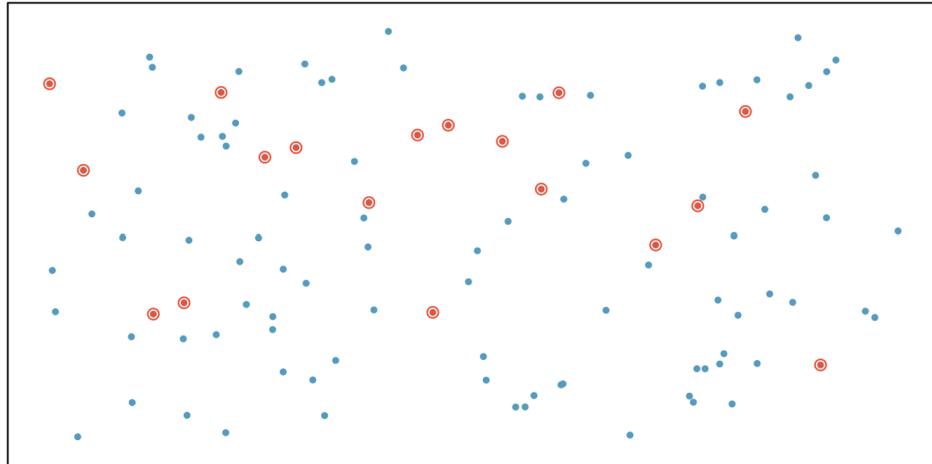
Arguably the most fundamental characteristic of good sampling techniques that seek to avoid bias is...

**RANDOM** selection of participants/observational units

“Good” sampling techniques:

- Simple Random Sample

Choose randomly from across whole population,  
e.g. "draw n names/participants out of a hat"

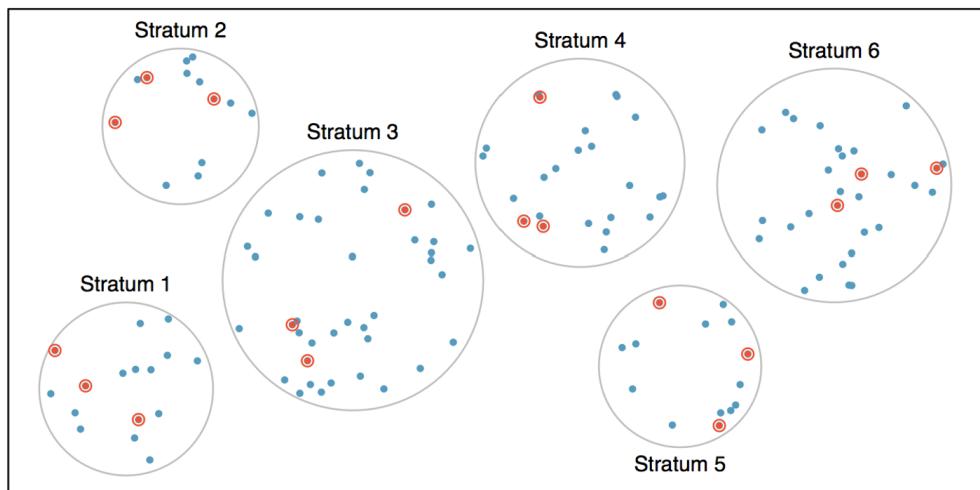


- Stratified Sampling      Divide population into strata, then take SRS from each

Stratified sampling is good when you have some characteristic which you believe will substantially influence responses/results.

Common ways to stratify:  
 - age (e.g. 18-39, 40-64, 65+...)  
 - work type (blue-collar, white-collar, service...)  
 - gender  
 - race/ethnicity

GOAL: ensure each stratum is in some sense "homogenous" with respect to your research question



We use stratified sampling to ensure we do not accidentally OMIT or UNDERREPRESENT key subgroups of the population.

- Cluster Sampling

For convenience of sampling, break the population into smaller groups! Useful for very large populations

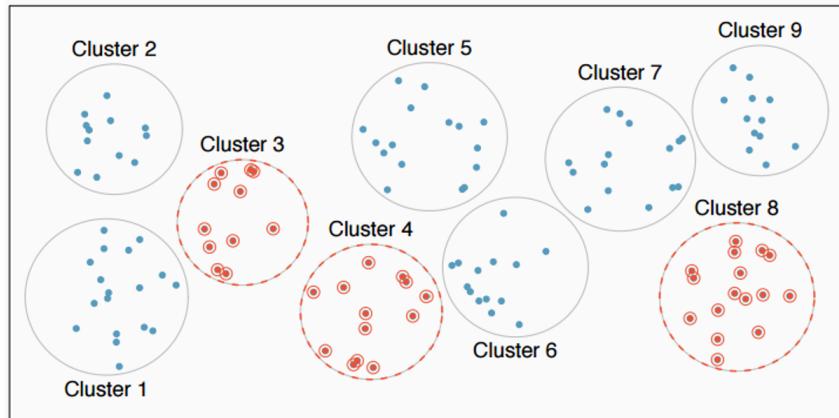
Method: use SRS to choose which clusters to sample from, then take WHOLE CLUSTER into sample

We want the clusters to be "heterogenous" (i.e. varied, diverse) in the same ways that the population is!

...but this also means this works best when the population looks a bit homogenous itself, OR when the variations won't influence your survey.

Some ways to make clusters:

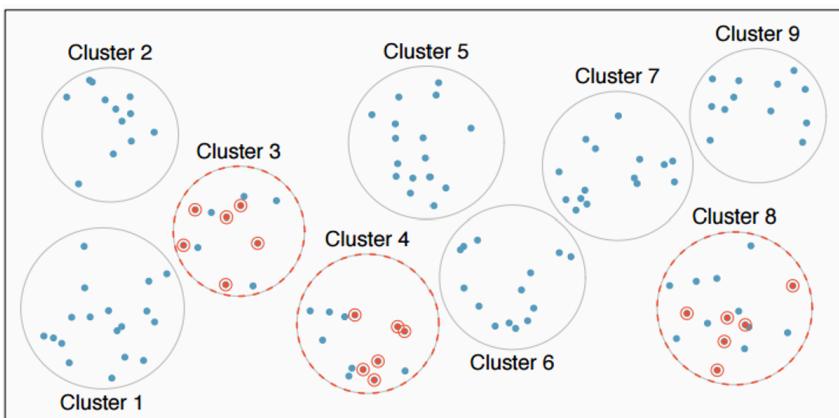
- cluster by region/area (zip code, section of ocean, village)
- cluster by depersonalized info (first 2 digits of Social Security Number)



- Multistage Sampling

When your population is WAY too big, pair cluster sampling with SRS within each cluster

Use this when your clusters are still too big to be fully sampled.



NOTE: you can also do other mix-and-match approaches! For example, you may want to pair stratified sampling with cluster sampling, to get a more convenient sampling approach that still allows you to account for differences between population subgroups.

HOWEVER, BE CAUTIOUS! You must always be able to justify your sampling methodology to others. Never make a sampling decision that doesn't have a firm reason behind it!

Other factors that can bias/influence results:

- Wording of questions

**CAUTION:** Word choice is EXTREMELY influential in surveys!

### Survey of high-school students:

"buy" vs "obtain" shifts opinions of high school students of how easy it is to gain possession of various substances

- "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%)

"assistance to the poor" vs "welfare" has major impact on opinions of American adults about the same programs

- Framing of questions

**CAUTION:** Question framing (such as the type of question you ask) is EXTREMELY influential in surveys!

### Fewer people mention the economy in open-ended version

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

Survey on important issues for determining voting decisions:

Economy was ranked much more important when it was LISTED as an option on a multiple-choice survey than when voters were given the SAME QUESTION in open-ended form!

Which ranking better represents the true opinion of American voters???

|                              | 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

**WARNING:** if your sample is biased and/or not representative, your sample statistic will NOT be close to the population parameter!

ONE LAST THING:

**Sample statistic:** a value, such as a proportion or percentage, computed from a sample, e.g. using survey results

All these numbers here are sample statistics.

The END GOAL of sampling is to compute sample statistics, to serve as estimates for population parameters.

**Population parameter:** the true number for that proportion/percentage/etc across the whole population