Introduction: Surveys Are Everywhere

PSCI 3802 Chapter 1

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Polling Is Inescapable

When we hear "polling" we think of elections. But election polls are only the visible *tip of the iceberg.*

- The **unemployment rate** comes from two surveys (households + businesses), not a count of every job.
- *Any* demographic statistic that isn't sex, age, race, or home ownership (e.g. "38% of the US has a college degree") comes from a **survey**, not the Census.
- Even with good administrative records (COVID vaccine doses by race), the moment you need a *rate* ("what share of White Philadelphians got the vaccine?"), you are back to needing a survey for the denominator.

Why do elections feel different? Because we get to see the truth. *We argue for weeks about whether election polls were "accurate." Is the unemployment rate accurate? We have no idea. There is no "Unemployment Election" to benchmark it against. The error is always there; elections are the rare case where we get to see it.*

What Makes a Survey a Survey?

According to Groves:

> A survey is a **systematic** method for gathering information from a **sample** of entities for the purpose of constructing a **quantitative** description of the attributes of the larger **population** of which the entities are members.

- **Systematic**: a clearly defined, repeatable methodology.
- **Quantitative**: attributes measured systematically.

- **Sample → population**: the defining feature. Use a *part* to understand the *whole*.

A census (counting everyone) is not a survey. *Censuses are ancient (they show up in Numbers and in Luke). The genuinely new idea (only about a century old) is that we can* shortcut *the count by looking at a small random slice. It is new because the statistics needed to justify it are new.*

A Short History of Getting It Wrong

1936: the *Literary Digest* disaster. *Their method was "more data is better": buy every list you can (phone books, car registrations, club rosters), mail 10 million postcards, tabulate 2.4 million replies. They predicted Alf Landon would win with 55%. FDR won 61% and all but 8 electoral votes.* Two problems:

- *Sample-frame bias*: their lists (phone and car owners) skewed Republican.
- *Non-response bias*: Landon supporters mailed their postcards back at higher rates.

Gallup and quota sampling. *The "fix" addressed only the frame: set targets for population subgroups and fill them. Smaller, better-targeted data beat the Digest's 2.4 million. Gallup called the 1936, 1940, and 1944 elections, and the pollsters became stars.*

1948: Dewey Defeats Truman. *Quota sampling let interviewers choose* who *filled each quota, and they avoided the very poor and very rich ("the unshaven and sweaty"). Truman won. The lesson: even with the right quotas, human choice in* who *gets interviewed introduces bias.*

The resolution is random sampling. *Stop hand-picking respondents to hit targets. Let chance choose them.*

Playing God: Why Random Sampling Works

Build a population where we know *the truth: 50% male, 60% white, 25% rich, 55% college, and a vote that is split 50/50. (The traits are correlated, e.g. college and voting Republican are negatively related.)*

```
          male           white            rich
         0.501           0.599           0.251
       college vote.republican
         0.551           0.501
```

Sample every *member and of course you recover the population exactly. The remarkable claim is that a much* smaller *random sample recovers it* in expectation. *Watch the sample shrink:*

```
            male white  rich college vote.republican
n = 50000 0.502 0.601 0.253    0.55           0.505
n = 1000  0.524 0.601 0.263    0.56           0.505
n = 100   0.490 0.630 0.200    0.54           0.530
```

Smaller samples get noisier, but every one is approximately *right, with no systematic tilt.*

Population size is irrelevant. *A random sample of 1,000 works as well in China (1.4B) as in Canada (40M). Think of salting soup: stir well (that is the randomness), and one spoonful tells you the seasoning. The size of the pot does not matter.*

How Polls Go Wrong: Biased Sampling

Now break the randomness, the way the Digest did. Let college graduates be more likely to be interviewed (20% vs 15%):

```
          male           white            rich
         0.479           0.613           0.249
       college vote.republican
         0.622           0.489
```

We oversample the college-educated, so we overstate the college share by about seven points. But the traits are correlated, *so the same error drags* everything *off: the estimated Republican share falls too (here by a bit over a point), and it will always move in that direction, because college graduates lean Democratic.* A tilt on one variable quietly biases the ones you actually care about.

Takeaway: *the only way to understand a mass public is a survey, and a survey is only as good as its sampling. Random selection buys*

unbiased inference from a tiny sample; the moment selection depends on who people are, *bias creeps into every estimate. The population is the point.*