Class 10: Sampling and Surveys (Text: Section 3.2)

Populations and Samples

If we talk to everyone in a population, we have taken a **census**. But this is often impractical, so we take a **sample** instead. We calculate a **statistic** from the sample (for example, the sample mean) and use it tell us something about a **population parameter** (for example, the population mean).

A population parameter is a fixed number, but generally unknown. The sample statistic is known, but can vary. If you take a new sample, you get a new value of the statistic. Because of the variation in the sample statistic, results about the population parameter should be reported with a **margin of error.** (eg, $\pm 3\%$)

The following notation is in common use:

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$

Ex**:** Election polls: Pollsters get information about how a small group of voters plans to vote (the sample), and infer how all voters (the population) will vote.

Statistics only give us good information about a population if the sample is *representative* of the population.

Polls: How Good Can They Be?[1]

Accuracy Record in Presidential Elections

Gallup Poll Accuracy Record

Year	Candidates	Final Gallup Survey %	Election Result %	Gallup Deviation %
2012	Obama	49	51	-2
	Romney	50	47	3
2008	Obama	55	53	2
	McCain	44	46	-2
2004	Bush	49	50.7	-1.7
	Kerry	49	48.3	0.7

1 http://www.gallup.com/poll/9442/Election-Polls-Accuracy-Record-Presidential-Elections.aspx

Polls: What Can Go Wrong?

Ex: TV Poll: Should UN Headquarters stay in New York?

Of 186,000 phone calls, 67% said "No"; in a nationwide random sample of 500 adults; 72% said "Yes". Why the difference? This is a **voluntary response poll**. Such polls are usually biased, as only those who feel strongly bother to respond.

Ex: 1936 *Literary Digest* Poll[2]

The magazine had successfully predicted winner in every presidential election since 1916. However in 1936:

- **Prediction:** Alfred Landon (R) 57%
  Franklin Roosevelt (D) 43%
- **Outcome:** Alfred Landon (R) 38%
  Franklin Roosevelt (D) 62%

How was the 1936 survey conducted?

- Mailed 10 million questionnaires to their subscribers, car owners, and addresses from telephone books, etc
- 2.4 million people responded

What went wrong? This was a **convenience sample**.
Selection Bias: the sample wasn't representative. In the Depression, only people who had money could afford magazines; cars and phones were less common and only owned by well-off people.
Non-response Bias: only certain people responded. Same problem as with a voluntary response sample: Tends to create bias.
At the same election, a young George Gallup predicted 56% for Roosevelt with a sample of 50,000 and also obtained the *Literary Digest*'s prediction with another sample of 3,000.

Ex: 1948 Presidential Election

Gallup, Roper, and Crossley predicted Thomas Dewey (with 50%) was the winner over Harry Truman (with 44%).

- **Outcome:** Truman won by 50% to 45%.

How was the 1948 survey conducted?

- **Quota Sampling:** Population divided into subgroups and interviewers interviewed a fixed quota in each of the subgroups. But interviewer was free to select who he interviewed, leading to bias.

2 http://whyfiles.org/009poll/fiasco.html. Obama photo-shopped imaged was after the Supreme Court ruling on ACA (health care) on June 28, 2012, when CNN and Fox reported the wrong outcome.

Sampling Methods: Good Design

Two steps to choosing a good sample.
- Identify the population we want to know about. This is the **sampling frame**.
- Choose respondents **randomly** in a **probability sample**.

How is this done?

Simple Random Sample (SRS)

In SRS, each individual in the population and each sample of size *n* has an equal chance of being selected.

Ex: Randomly draw names from a list of voters.

Would you ever want anything other than a SRS?
- Efficiency (or cost).
- Ability to focus on particular groups

In those cases, we use **stratified sampling** or **cluster sampling**

Stratified Random Sample

To get a stratified random sample
- Divide the population into groups of similar individuals, called **strata**.
- Choose a separate SRS in each stratum.
- Combine these SRSs to form the full sample.
- Strata can be associated with variable to be measured.

Advantages of Stratified Random Sample
- If individuals within each stratum are similar, then can get more precise estimates from smaller samples than with a SRS.
- Can examine and compare small groups with disproportionate stratified random sampling, or oversampling (e.g., rural households)

Ex. Several US states allow same-sex marriage. In November 2008 in California, Proposition 8 amended the state's constitution to limit marriage to be between a man and a woman. A September 18, 2008, poll[3] of 405 Californian likely voters reported that 37.8% supported Proposition 8. What is:

(a) Population: All California likely voters
Sample: The 405 voters who responded to the poll
Population parameter: True proportion supporting Prop 8
Sample statistic: The 37.8%

(b) Obama's presence on the presidential ticket was expected to draw in more African-American and younger voters. To ensure that the views of these two groups are represented in their polls, what sampling method should polling agencies have used? Stratified by race and age

Cluster Sampling

Population is broken down in groups, called clusters, and a sample of clusters is selected at random.
- Differs from stratified random sampling because:
  - Choose clusters at random to sample; sample all strata
  - Unlike in stratified sampling, we do *not* want the cluster variable to be correlated to the outcome. (Because we sample a random set of clusters, but we sample all strata.)
- Main goal of cluster sampling is to reduce travel or interview costs and the costs of drawing up lists.
  - For face-to-face interviews, need houses close together
  - To measure test scores, it is difficult to compile lists of all high school students in the country. Easier to get a list of school districts, then contact a random sample of districts.

3 Field Research Corporation, September 18, 2008

Sampling and Non-Sampling Errors

Types of problem:
- Sampling not random
- Sampling from a group that is not representative of the population. (That is, wrong sampling frame.)
- Question wording leads to unreliable answers.

Non-Random Sampling

Selection Bias: the methodology for selecting people in the sample is not representative of the population.

Ex: 1936 Literary Digest Poll; 1948 Dewey/Truman

Avoid **under-coverage** of some groups:

Ex: homeless, prison population, those in dormitories, or without phones.

Non-response Bias: The people who participate may be different than those who choose not to participate.

Ex: 1936 Literary Digest Poll.

Ex. Phone surveys (people who work evenings excluded)

Ex: Polls conducted only in English.

Since **cell phones** have unlisted numbers, phone surveys may become unreliable again. See Pew Research.[4]

Wrong Sampling Frame

Ex: Medical texts were original written with instructions for treating a 70 kg man. (70 kg = 154 lb) As a result women and children often got inappropriate dosage and treatment.

Ex: Current dosages are being revised because the increase in obesity means that heavy patients may be getting doses that are too small.

Reporting Bias/Response Bias: Wording of questions: An important issue in polling!

People may not report how they would actually behave or what they actually believe.

Ex: If asked how much they would *theoretically* pay to protect the Arctic National Wildlife Refuge, people often report more than what they would *actually* pay.)

Ex: How would you find out what fraction of a population used deodorant? Asking "Did you use deodorant today?" is not likely to produce a reliable answer. Better to give a list of items and ask respondents to check those they had used that day..

Ex: In 1998, Scots were asked if they would vote in favor of "*independence for Scotland*" and if they supported "*an independent Scotland separate from the UK*"[5] Which question got 51% and which got 34%? First: 51%, second 34%

4 "How Serious Is Polling's Cell-Only Problem? The Landline-less Are Different and Their Numbers Are Growing Fast." http://pewresearch.org/pubs/515/polling-cell-only-problem June 20, 2007 updated Sept 23, 2008 http://pewresearch.org/pubs/964/cell-phones-and-the-2008-vote-an-update

5 "All Set for Independence?" *The Economist*, September 12, 1998.

Two Lancet Studies of Civilian Deaths in Iraq: Study Design

Lancet I (Nov '04):
33 clusters of 30 households in each sample, estimating change in death rate before and after Mar '03. Compared wartime and prewar death rates, obtained by memory of family members. Asked to see death certificates.

Lancet II (Oct '06):
47 clusters of 40 households each. Estimating change in death rate before and after the start of the war.
Data obtained from 1849 households, containing 12,801 people.
Death certificates for 92%.

- What difficulties do you foresee with this method? Do you expect to the results are too low? Too high? Impossible to tell?
- Huge variation depending on where clusters are. Lancet I drew Falluja and decided to omit it, considering it an outlier. (2/3 of the violent deaths reported were there.)
- Baseline depended on memories of people interviewed may not be reliable: If prewar deaths more easily forgotten, especially infant deaths, the estimates are too high.
- Families which were completely wiped out would not be available to be interviewed. Makes estimate too low.
- Death of combatants might have been hidden; makes estimates too low.
- Too small a sample size—makes result more variable and less reliable.
- *Slate* claims[6] that only households living on main commercial streets were surveyed; these may not be same as other households. (*Slate* says more likely to be targeted.)

Results:
Lancet I in 2004: 98,000 excess deaths;
95% Confidence interval (8,000, 194,000)
Lancet II in 2006: 654,965 excess deaths; 601,027 violent;
95% Confidence interval (392,979, 942,636)

Ex: What are the confidence intervals telling us?

Roughly: Range of possible values we might expect to obtain from other samples of the same type.

Recent Study

An October 2013 study surveyed 2000 Iraqi households widely spread out in Iraq.[7] The new study gives
95% Confidence interval (48,000, 751,000) excess deaths
Why such a difference from previous students?

- Families that suffered particularly badly may have been more inclined to escape, making estimates too low.

6 http://www.slate.com/blogs/the_world_/2013/10/18/new_study_estimates_half_a_million_casualties_from_iraq_war_but_how_reliable.html
7 "Mortality in Iraq Associated with the 2003–2011 War and Occupation: Findings from a National Cluster Sample Survey by the University Collaborative Iraq Mortality Study" by A. Hagopian, et al, in PLOS Online, October 2013.