

Math 207: Statistics

Chapter 19: Sample Surveys

Population (parameters)

Sample (statistics)

Dr. Ralph Wojtowicz
Mathematics Department



SHENANDOAH
UNIVERSITY

1 Example

- Exercise 7 page 22

2 Political Polls

- Literary Digest Poll
- Dewey Defeats Truman
- Using Chance
- Chance Error and Bias

Exercise 7 page 22

(Hypothetical) In a clinical trial, data collection usually starts at “baseline,” when the subjects are recruited into the trial but before they are assigned to treatment or control. Data collection continues until the end of followup. Two clinical trials on prevention of heart attacks report baseline data on smoking, shown below. In one of these trials, the randomization did not work. Which one, and why?

		Number of persons	Percent who smoked
Trial (i)	Treatment	1,012	49.3%
	Control	997	69.0%
Trial (ii)	Treatment	997	59.3%
	Control	1,017	59.0%

- Randomization did not work in Trial (i).
- If the groups had been chosen at random, the percent of smokers would have been about the same in both treatment and control groups.
- In statistics, “randomization is your friend.”

The Literary Digest Poll: 1936

- Great Depression had resulted in great economic challenges prior to the 1936 presidential election.
- Franklin Delano Roosevelt (Democratic incumbent) was running against Alfred Landon (Republican governor of Kansas).
- The *Literary Digest* had correctly predicted the outcome of each presidential election since 1916.
- *Literary Digest* surveyed 2.4 million people (largest number of people every replying to a poll).
- *Digest's* prediction: overwhelming Landon victory!
- Election results: Roosevelt won 62% to 38%.
- Problems:
 - The *Digest* sampled from phone books, club membership lists and other sources that favored wealthy voters.
 - When a selection procedure is biased, taking a large sample doesn't help.

Dewey Defeats Truman: 1948



- 1948 presidential election: Harry Truman (Democratic incumbent who ascended to the presidency when FDR died in office) ran against Thomas Dewey (Republican district attorney from New York).
- Three major polls predicted Dewey would win: Crossley (for the Hearst newspapers), Gallup and Roper (for *Fortune* magazine)
- Quota sampling (e.g., select specified numbers of people to poll by gender, race, age, etc.) did not accurately reflect the population.

Using Chance

- Running a political poll is like sampling from a box of number tickets.
- Simple random sampling:
 - Tickets drawn at random without replacement.
 - All tickets have equal likelihood of being chosen
- Multistage cluster sampling:
 - Used to avoid challenges of randomly selected people being widely distributed across the country.
 - Divide country into regions, regions into cities, etc.
 - Randomly sample individuals from these specified areas.
 - Sample sizes are typically scaled to account for relative populations of areas.
- *The Islands*: website for simulating samples from a fictional population:
<https://islands.smp.uq.edu.au>
- Probability methods are highly effective! See Section 19.5.

Chance Error and Bias

- Running a political poll is like sampling from a box of number tickets.

253,785 0s 433,211 1s

- Bias: sample does not represent the actual distribution of 0s and 1s in the box.
- Chance error: $SE_{\%} = \frac{SD_{\text{box}}}{\sqrt{n}}$