Elements of the Sampling Problem

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Reading Assignment:
Scheaffer, Mendenhall, Ott, & Gerow, Chapter 2.1-2.3
Goals for this Lecture

• Some terminology

• Introduction to types of sampling
  – Convenience vs. random sampling
  – Types of random sampling
    • Simple random sampling (SRS)
    • Stratified sampling
    • Cluster sampling
    • Systematic sampling

• Examples of when and how things can go wrong
Terminology Review (1)

- An **element** is an object upon which a measurement is taken
  - In surveys of individuals, an element is a respondent
  - But it may not be (e.g., organizational-level surveys)
- A **population** is a collection of elements about which we wish to make an inference
  - Usually, all elements of the population are not measurable
- **Sampling units** are nonoverlapping collections of elements that cover the entire population
Terminology Review (2)

• A **frame** is a list of sampling units
  – An **explicit sample frame** is an actual list in which every sampling unit / element is enumerated
    • Example: List of NPS student e-mail addresses
  – An **implicit sample frame** is a methodology that allows for the selection of a sample such that the (non-zero) probability of selecting each element can be calculated
    • Example: Random digit dialing

• A **sample** is a collection of sampling units drawn from a single frame or from multiple frames
Types of Samples

- **Convenience sample**: Researcher does not control how observations get into the sample or does not randomly select sample.

- **Random sample**: Researcher controls how observations get selected into the sample using a randomization mechanism.

- **Latter more correctly referred to as a probability sample**:
  - Probability sampling required to do inference
  - A necessary but not sufficient condition
Good Sampling is Critical for Good Statistical Inference

• If we are to use a sample to infer something about a population, we need to:
  – Have some assurance that the sample is representative of the population (i.e., minimize the chance of bias)
  – Be able to quantify how far off our sample statistic could be from the population statistic (sampling error)

• Using a probability sample is protection against (unknowingly) selecting a biased sample
  – A high response rate is protection against nonresponse bias

• Classical statistics is all about quantifying uncertainty (i.e., sampling error) and using that information to determine statistical significance
The Inference Problem

We specify, know, and control this process.

So we can do the correct math here.

Population Parameter: Population Average

Sample Statistic: Sample Average
Error of Estimation

• Want to estimate some population quantity $\theta$
  – Could be an average or a total, for example
• It’s estimated with a statistic from the sample, $\hat{\theta}$
• Want the error of estimation to small, say less than some value $B$:
  \[
  \text{Error of estimation} = |\theta - \hat{\theta}| < B
  \]
• Generally cannot guarantee a bound on the error, so it’s typically stated in terms of a probability:
  \[\Pr \left( |\theta - \hat{\theta}| < B \right) = 1 - \alpha\]
Types of Random Sampling (1)

• **Simple random sample (SRS)**: any two samples of the same size are equally likely to be selected
  - One way to do, list all possible sample combinations of size $n$ (without replacement) out of the population, then randomly choose one

• All of the methods you’ve studied in previous statistics classes assumed SRS
  - More complicated types of random sampling require more complicated types of analysis
Types of Random Sampling (2)

• **Stratified sampling**: divide population into non-overlapping, homogeneous groups and then draw a SRS from each group.

• Most useful when population is more homogeneous within strata than across strata.

• E.g., want to survey service members about retention issues.
  – Likely should stratify by officer/enlisted, by branch of service.
  – Depending on objectives, could stratify further by rank and designator/NEC/MOS, etc.
Types of Random Sampling (3)

- **Cluster sampling**: often used if hard (or impossible) to enumerate population or when population naturally occurs in clusters
  - Data that occurs in clusters likely to be correlated within the cluster structure – analysis must account for correlation
  - Often used in surveying to minimize or mitigate fielding costs

- In cluster sampling, cluster is treated as the sampling unit
Types of Random Sampling (4)

- **Systematic sampling**: given a list of the \( N \) population members, sample every \( k^{th} \) unit on the list
  - Basis for how random searches are done of cars coming onto a base
  - Often useful for things like sampling visitors to a web site
- In addition to being convenient in some situations, it can sometimes give more precise estimates than SRS
Why Sample?

• E.g., why conduct an:
  – N1 survey of the reenlistment intentions of a sample of sailors
  – Nielsen survey of consumer television behaviors and preferences
  – Survey of a random sample of mess hall customers

• Rather than:
  – Conducting a census of every sailor (present and future) in the Navy
  – Evaluate how all television viewers watch TV
  – Survey of all mess hall customers for every meal

✓ Collecting data for whole populations is usually either too expensive or impossible
✓ Also, sampling error of a well-executed sample often smaller than the systematic error (bias) of a poorly-executed census
An Example of What Can Go Wrong: Using Poor Sampling Frame(s)

- Literary Digest presidential election poll: FDR vs Alfred Landon
  - Sampling frame taken from telephone numbers and club membership lists
    - Only one house in 4 had a phone
    - Rich tend to join clubs more than poor
    - Sampling frame coverage bias
  - Sent questionnaires to 10 million people
    - Only 2.4 million people responded
    - Large potential for nonresponse bias
Results of the Poll

- LD Prediction: FDR 43%, Landon 57%
- Election Result: FDR 62%, Landon 38%
- Largest error ever made by a major poll
  - Literary Digest had correctly forecast every election since 1916
  - Literary Digest went bankrupt soon after 1936
- The Gallup organization (a start-up in 1936)
  - Guessed the election more or less correctly using much less data than Literary Digest
Another Example of What Can Go Wrong: Non-Probability Sampling

• Even a survey chosen on purpose to be representative can suffer from bias
  – 1948 Gallup Poll
  – Quota sampling: Survey crews given quotas for the types of people to interview
    • Based on observable characteristics
    • Interview 6 men and 6 women, etc.
  – Specified so that the resulting sample will match the population characteristics
What's Wrong with Quota Sampling?

<table>
<thead>
<tr>
<th>TABLE 2.2</th>
<th>Comparison of Gallup and Roper samples with U.S. population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentages</td>
</tr>
<tr>
<td>Sex</td>
<td>49.1*</td>
</tr>
<tr>
<td>Male</td>
<td>50.9</td>
</tr>
<tr>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>34.3</td>
</tr>
<tr>
<td>21–34</td>
<td>30.9</td>
</tr>
<tr>
<td>35–49</td>
<td>34.8</td>
</tr>
<tr>
<td>50 and over</td>
<td></td>
</tr>
<tr>
<td>Education (last school attended)</td>
<td>43.5</td>
</tr>
<tr>
<td>Grade school or less</td>
<td>43.4</td>
</tr>
<tr>
<td>High school</td>
<td>13.0</td>
</tr>
<tr>
<td>College</td>
<td>96.1</td>
</tr>
<tr>
<td>Color</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>3.9</td>
</tr>
<tr>
<td>Negro</td>
<td>85.8</td>
</tr>
<tr>
<td>Veteran status</td>
<td>14.2</td>
</tr>
<tr>
<td>Male veterans</td>
<td>82.5</td>
</tr>
<tr>
<td>All others</td>
<td></td>
</tr>
<tr>
<td>Labor union membership of males</td>
<td>17.5</td>
</tr>
<tr>
<td>Member</td>
<td>82.5</td>
</tr>
<tr>
<td>Non Member</td>
<td></td>
</tr>
<tr>
<td>Number of respondents</td>
<td>2972</td>
</tr>
</tbody>
</table>

*Population aged 21 years and over as of November 1948.

Results

**TABLE 2.1**
The election polls of 1948

<table>
<thead>
<tr>
<th></th>
<th>Dewey</th>
<th>Truman</th>
<th>Thurmond</th>
<th>Wallace</th>
<th>Total*</th>
</tr>
</thead>
<tbody>
<tr>
<td>National vote</td>
<td>45.1</td>
<td>49.5</td>
<td>2.4</td>
<td>2.4</td>
<td>99.4</td>
</tr>
<tr>
<td>Crossley</td>
<td>49.9</td>
<td>44.8</td>
<td>1.6</td>
<td>3.3</td>
<td>99.6</td>
</tr>
<tr>
<td>Gallup</td>
<td>49.5</td>
<td>44.5</td>
<td>2.0</td>
<td>4.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Roper</td>
<td>52.2</td>
<td>37.1</td>
<td>5.2</td>
<td>4.3</td>
<td>98.8</td>
</tr>
</tbody>
</table>

*Exclusive of percentages for minor candidates. Gallup percentages calculated on total vote for four principal candidates.

Who Would You Interview?

- Republican
- Democrat

✓ Survey interviewers showed republican bias
### Table 2.3
Gallup poll accuracy

<table>
<thead>
<tr>
<th>Year</th>
<th>Gallup final survey</th>
<th>Election result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1936</td>
<td>55.7% Roosevelt (D)</td>
<td>62.5% Roosevelt</td>
</tr>
<tr>
<td>1940</td>
<td>52.0% Roosevelt (D)</td>
<td>55.0% Roosevelt</td>
</tr>
<tr>
<td>1944</td>
<td>51.5% Roosevelt (D)</td>
<td>52.3% Roosevelt</td>
</tr>
<tr>
<td>1948</td>
<td>44.5% Truman (D)</td>
<td>49.9% Truman</td>
</tr>
<tr>
<td>1952</td>
<td>51.0% Eisenhower (R)</td>
<td>55.4% Eisenhower</td>
</tr>
<tr>
<td>1956</td>
<td>59.5% Eisenhower (R)</td>
<td>57.8% Eisenhower</td>
</tr>
<tr>
<td>1960</td>
<td>51.0% Kennedy (D)</td>
<td>50.1% Kennedy</td>
</tr>
<tr>
<td>1964</td>
<td>64.0% Johnson (D)</td>
<td>61.3% Johnson</td>
</tr>
<tr>
<td>1968</td>
<td>43.0% Nixon (R)</td>
<td>43.5% Nixon</td>
</tr>
<tr>
<td>1972</td>
<td>62.0% Nixon (R)</td>
<td>61.8% Nixon</td>
</tr>
<tr>
<td>1976</td>
<td>48.0% Carter (D)</td>
<td>50.0% Carter</td>
</tr>
<tr>
<td>1980</td>
<td>47.0% Reagan (R)</td>
<td>50.8% Reagan</td>
</tr>
<tr>
<td>1984</td>
<td>59.0% Reagan (R)</td>
<td>59.2% Reagan</td>
</tr>
</tbody>
</table>

The figure shown is the winner's percentage of the Democratic Republican vote, except in the elections of 1948, 1968, and 1976.

SMO&G Chapter 2.4 & 2.5

• Rest of chapter is Sheaffer, Mendenhall, Ott & Gerrow’s summary of survey design
  – Sources of errors
  – Reducing errors
  – Designing questionnaires

• We covered this material in Dillman and Fowler
  – Read if you want to get another perspective (that essentially reinforces all we’ve done so far)
What We Have Covered

• Introduced some terminology
• Discussed various types of sampling
  – Convenience vs. random sampling
  – Types of random sampling
    • Simple random sampling (SRS)
    • Stratified sampling
    • Cluster sampling
    • Systematic sampling
• Examined some examples of when and how things can go wrong