Survey Research and Sampling Theory

- Survey research: relies primarily on probability sampling
- Uses findings from a smaller sample to generalize to a larger target population
- Focus on *sampling theory*

Sampling Theory

- **Descriptive**: describing attributes of particular sample
- **Inferential**: generalizing beyond sample to larger population
What is sampling?

- Process of selecting subset of observations from among many possible observations to draw conclusions about a larger population
- **Representativeness**: each individual in larger population has an equal chance of being chosen for sample

Benefits of sampling

1) Sample is representative

2) Ability to calculate accuracy of sample (standard error)
9 Definitions

1) Population: target population
2) Study population: population used for study
3) Sampling unit: units used for selection
4) Sampling frame: list of sampling units used
5) Observation unit: person from which data are collected
6) Binomial variable: variable with two responses

9 definitions

7) Statistic vs. parameter: summary description of variable in a sample vs. summary description in target population
8) Sampling error (s.e.): determines accuracy of sample
9) Confidence levels: way of expressing degrees of confidence
Statistic vs. parameter

<table>
<thead>
<tr>
<th></th>
<th>Sample Statistic</th>
<th>Population Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{X}$</td>
<td>$\mu$ (mu)</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$s$</td>
<td>$\sigma$ (sigma)</td>
</tr>
</tbody>
</table>

Standard error

$$s.e. = \sqrt{\frac{p \times q}{n}}$$

$p = \text{proportion saying yes}$
$q = \text{proportion saying no (1-p)}$
$n = \text{sample size}$
Standard error

s.e. = $\sqrt{\frac{.70 \times .30}{1000}}$
Standard error

\[ \text{s.e.} = \sqrt{\frac{.21}{1000}} \]

\[ \text{s.e.} = .01449 \]

\[ \text{s.e.} = 1.449, \text{ or } 1.4\% \ (\text{.1 decimal place}) \]
### Properties of standard errors: different values of p’s and q’s

<table>
<thead>
<tr>
<th>p</th>
<th>q</th>
<th>n</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>.20</td>
<td>.80</td>
<td>100</td>
<td>.04</td>
</tr>
<tr>
<td>.40</td>
<td>.60</td>
<td>100</td>
<td>.049</td>
</tr>
<tr>
<td>.50</td>
<td>.50</td>
<td>100</td>
<td>.05</td>
</tr>
<tr>
<td>.60</td>
<td>.40</td>
<td>100</td>
<td>.049</td>
</tr>
<tr>
<td>.80</td>
<td>.20</td>
<td>100</td>
<td>.04</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>100</td>
<td>0.0</td>
</tr>
</tbody>
</table>

### Properties of standard errors: different sample sizes

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>q</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.5</td>
<td>.5</td>
<td>.07</td>
</tr>
<tr>
<td>100</td>
<td>.5</td>
<td>.5</td>
<td>.05</td>
</tr>
<tr>
<td>150</td>
<td>.5</td>
<td>.5</td>
<td>.04</td>
</tr>
<tr>
<td>200</td>
<td>.5</td>
<td>.5</td>
<td>.035</td>
</tr>
<tr>
<td>250</td>
<td>.5</td>
<td>.5</td>
<td>.032</td>
</tr>
<tr>
<td>1000</td>
<td>.5</td>
<td>.5</td>
<td>.016</td>
</tr>
</tbody>
</table>
Examples:
Roper Public Opinion Survey

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>q</th>
<th>n</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN (10/13-15/06)</td>
<td>.36</td>
<td>.64</td>
<td>1012</td>
<td>1.5%</td>
</tr>
<tr>
<td>ABC (10/8/01)</td>
<td>.92</td>
<td>.08</td>
<td>1009</td>
<td>.85%</td>
</tr>
<tr>
<td>Gallup (8/16/01)</td>
<td>.57</td>
<td>.43</td>
<td>1013</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

p=approve of Pres. Bush, q=do not approve/no opinion

Probability sampling designs

1) **Simple random sample (SRS):**
   --decide on sampling frame
   --give number to everyone on list
   --use table of random numbers to choose people
2) **Systematic sample with a random start:**

-- simpler than SRS

-- choose every $k^{th}$ element, where $k = \#$ in population / $\#$ in sample (sampling interval)
Example:
 systematic sampling

✓ Target population = 1000
✓ Sample population = 100
✓ k = 1000/100
✓ k = 10

Example:
 systematic sampling

✓ k = 10
✓ Start randomly with number between 1 and k
✓ Randomly select 6, then: 16, 26, 36 . . . through 996
✓ Watch out for periodicity!
Probability sampling designs

3) **Stratified sample**

--ensures different groups are adequately represented in sample
--increases accuracy in estimating population parameter
--reduces s.e.
Example: stratified sample

<table>
<thead>
<tr>
<th></th>
<th>population %</th>
<th>sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglo</td>
<td>70%</td>
<td>70</td>
</tr>
<tr>
<td>Black</td>
<td>20%</td>
<td>20</td>
</tr>
<tr>
<td>Hispanics</td>
<td>10%</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

Probability sampling designs

4) *Multistage cluster sample*:
   --used when there is no list of names
   --create a sampling frame via set of “stages”
Multistage cluster sampling: 3 stages

1) **Stage 1**: define area using map (divided into blocks)
   --Choose blocks via SRS or systematic sampling
   --Blocks = PSU (primary sampling unit)

2) **Stage 2**: list and number all dwelling on selected blocks only
   --SRS or systematic sampling of dwellings
   --Dwellings = secondary sampling unit
Multistage cluster sampling: 3 stages

3) **Stage 3:**
--interview HH or randomly selected member of dwelling

Multistage cluster sampling: Implications

✓ Sampling error increased:
  --Each sampling unit increases sampling error
✓ General guideline: maximize number of clusters and minimize number of elements within cluster
✓ Why?