Sample Design and Sampling Error

Professor Ron Fricker
Naval Postgraduate School
Monterey, California
Goals for this Lecture

• Introduction to sampling
  – Convenience vs. random sampling
  – Types of random sampling: simple random, stratified, and cluster sampling

• Define “margin of error”

• Basic power calculations under SRS
  – Continuous vs. binary response

• Finite population correction
Sampling for Statistical Inference

Unobserved population statistic

Sample statistic

Inference

Sample

Sampling
Good Statistical Inference is (Almost) All About Good About Sampling

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

• Using a random sample is protection against (unknowingly) selecting a biased sample

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

- **Convenience sample**: Individuals in the population decide to join the sample
  - 900 number and other call-in polls
  - Website surveys (often)
  - Shopper and visitor surveys

- **Random sample**: Individuals or units are chosen randomly from the population
  - Whether or not part of the sample is not individual’s choice/decision
Types of Random Sampling

• Simple random sample (SRS): any two samples of the same size are equally likely to be selected

• Some other possible random sampling methods:
  – Stratified sampling
    • Divide population into nonoverlapping, homogeneous groups and then draw a SRS from each group
  – Cluster sampling
    • Data naturally occurs in clusters
    • Use SRS to select clusters
Basic statistical methods assume:
- Population is of infinite size (or so large as to be essentially infinite)
- Sample size is a small fraction of the population
- Sample is drawn from the population via SRS

Under these conditions, can do the usual calculations:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{and} \quad Var(\bar{y}) = \frac{s^2}{n} = \frac{1}{n} \times \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$
Finite Population Correction

• Previous calculations assumed an infinite population size
  – Fine when fraction being surveyed less than 5%
  – Then $f = n / N \approx 0$

• If surveying more than 5 percent of the population, must adjust the estimated sample variance using a finite population correction

$$\text{Var}(\bar{y}) = (1 - f) s^2 / n = \left( \frac{N - n}{N} \right) \frac{s^2}{n}$$
Standard Error Estimates

- For various sample sizes, standard errors for an infinite-sized population and one with N=300
  - Binary question
  - Conservative $p=0.5$ assumption

<table>
<thead>
<tr>
<th>n</th>
<th>fpc</th>
<th>E (w/out fpc)</th>
<th>E (w/ fpc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.96</td>
<td>0.196</td>
<td>0.188</td>
</tr>
<tr>
<td>50</td>
<td>0.91</td>
<td>0.139</td>
<td>0.127</td>
</tr>
<tr>
<td>75</td>
<td>0.87</td>
<td>0.113</td>
<td>0.098</td>
</tr>
<tr>
<td>100</td>
<td>0.82</td>
<td>0.098</td>
<td>0.080</td>
</tr>
<tr>
<td>125</td>
<td>0.76</td>
<td>0.088</td>
<td>0.067</td>
</tr>
<tr>
<td>150</td>
<td>0.71</td>
<td>0.080</td>
<td>0.057</td>
</tr>
<tr>
<td>175</td>
<td>0.65</td>
<td>0.074</td>
<td>0.048</td>
</tr>
<tr>
<td>200</td>
<td>0.58</td>
<td>0.069</td>
<td>0.040</td>
</tr>
<tr>
<td>225</td>
<td>0.50</td>
<td>0.065</td>
<td>0.033</td>
</tr>
<tr>
<td>250</td>
<td>0.41</td>
<td>0.062</td>
<td>0.025</td>
</tr>
<tr>
<td>275</td>
<td>0.29</td>
<td>0.059</td>
<td>0.017</td>
</tr>
<tr>
<td>300</td>
<td>0.00</td>
<td>0.057</td>
<td>0.000</td>
</tr>
</tbody>
</table>
What Does “Margin of Error” Mean?

• **Margin of error** is just half the width of a 95 percent confidence interval

• Common survey terminology
  – Convention is a 3% margin of error
  – Means a 95% CI is the survey result +/- 3%

• To achieve a desired margin of error, must have the right sample size ($n$)
  – Power calculations are done by statisticians to figure out the required sample size to achieve a particular margin of error
To figure out how much data you need:

- Determine \( w \), the desired width of the confidence interval. Remember,

\[
w = 2z_{\alpha/2} \frac{s}{\sqrt{n}}
\]

- Get an estimate of \( s \) from somewhere, a pilot study, for instance

- Choose the confidence level: \( 100(1-\alpha) \)

- Required sample size:

\[
n = \frac{4z_{\alpha/2}^2 s^2}{w^2}
\]
SRS: How Much Data for $p$?

- To figure out how much data you need:
  - Choose $E$, the margin of error
  - Then, $E = z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p}) / n}$
  - Algebra gives required sample size:
    $$ n = \frac{z_{\alpha/2}^2 \hat{p}(1 - \hat{p})}{E^2} $$

- Can simplify further:
  - Estimate $p$ using worst case: 1/2
  - For 95% CI, approximate with $z_{\alpha/2} = 2$
  - Then, $n = 1 / E^2$
Power Calculations

• Some of the steps in determining required sample size:
  – Make conservative assumptions about sample variability (e.g., \( s \))
  – Use question with largest \( n \) (e.g., binary question)
  – Multiply by number of strata / groups (i.e., smallest unit of analysis)
  – Inflate to account for nonrespondents, missing data, etc.

• Can get tricky for non-SRS sampling methods
  – Don’t hesitate to consult a statistician
• **Stratified (random) sampling** divides the sampling frame up into strata from which separate random samples are drawn.

• Two main reasons, one practical and one statistical:
  - To ensure sufficient observations are drawn from small strata (i.e., to **oversample**)
    • Often necessary to meet survey’s objectives
  - To reduce the variance in the sample statistic(s)
    • Assuming strata are relatively homogeneous
An Example

<table>
<thead>
<tr>
<th>Record</th>
<th>Name</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bradburn, N.</td>
<td>High</td>
</tr>
<tr>
<td>2</td>
<td>Cochran, W.</td>
<td>Highest</td>
</tr>
<tr>
<td>3</td>
<td>Deming, W.</td>
<td>High</td>
</tr>
<tr>
<td>4</td>
<td>Fuller, W.</td>
<td>Medium</td>
</tr>
<tr>
<td>5</td>
<td>Habermann, H.</td>
<td>Medium</td>
</tr>
<tr>
<td>6</td>
<td>Hansen, M.</td>
<td>Low</td>
</tr>
<tr>
<td>7</td>
<td>Hunt, J.</td>
<td>Highest</td>
</tr>
<tr>
<td>8</td>
<td>Hyde, H.</td>
<td>High</td>
</tr>
<tr>
<td>9</td>
<td>Kalton, G.</td>
<td>Medium</td>
</tr>
<tr>
<td>10</td>
<td>Kish, L.</td>
<td>Low</td>
</tr>
<tr>
<td>11</td>
<td>Madow, W.</td>
<td>Highest</td>
</tr>
<tr>
<td>12</td>
<td>Mandela, N.</td>
<td>Highest</td>
</tr>
<tr>
<td>13</td>
<td>Norwood, J.</td>
<td>Medium</td>
</tr>
<tr>
<td>14</td>
<td>Rubin, D.</td>
<td>Low</td>
</tr>
<tr>
<td>15</td>
<td>Sheatsley, P.</td>
<td>Low</td>
</tr>
<tr>
<td>16</td>
<td>Steinberg, J.</td>
<td>Low</td>
</tr>
<tr>
<td>17</td>
<td>Sudman, S.</td>
<td>High</td>
</tr>
<tr>
<td>18</td>
<td>Wallman, K.</td>
<td>High</td>
</tr>
<tr>
<td>19</td>
<td>Wolfe, T.</td>
<td>Highest</td>
</tr>
<tr>
<td>20</td>
<td>Wooldley, T.</td>
<td>Medium</td>
</tr>
</tbody>
</table>

Record | Name         | Group |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Cochran, W.</td>
<td>Highest</td>
</tr>
<tr>
<td>7</td>
<td>Hunt, J.</td>
<td>Highest</td>
</tr>
<tr>
<td>11</td>
<td>Madow, W.</td>
<td>Highest</td>
</tr>
<tr>
<td>12</td>
<td>Mandela, N.</td>
<td>Highest</td>
</tr>
<tr>
<td>19</td>
<td>Wolfe, T.</td>
<td>Highest</td>
</tr>
<tr>
<td>1</td>
<td>Bradburn, N.</td>
<td>High</td>
</tr>
<tr>
<td>3</td>
<td>Deming, W.</td>
<td>High</td>
</tr>
<tr>
<td>8</td>
<td>Hyde, H.</td>
<td>High</td>
</tr>
<tr>
<td>17</td>
<td>Sudman, S.</td>
<td>High</td>
</tr>
<tr>
<td>18</td>
<td>Wallman, K.</td>
<td>High</td>
</tr>
<tr>
<td>4</td>
<td>Fuller, W.</td>
<td>Medium</td>
</tr>
<tr>
<td>5</td>
<td>Habermann, H.</td>
<td>Medium</td>
</tr>
<tr>
<td>9</td>
<td>Kalton, G.</td>
<td>Medium</td>
</tr>
<tr>
<td>13</td>
<td>Norwood, J.</td>
<td>Medium</td>
</tr>
<tr>
<td>20</td>
<td>Wooldley, T.</td>
<td>Medium</td>
</tr>
<tr>
<td>6</td>
<td>Hansen, M.</td>
<td>Low</td>
</tr>
<tr>
<td>10</td>
<td>Kish, L.</td>
<td>Low</td>
</tr>
<tr>
<td>14</td>
<td>Rubin, D.</td>
<td>Low</td>
</tr>
<tr>
<td>15</td>
<td>Sheatsley, P.</td>
<td>Low</td>
</tr>
<tr>
<td>16</td>
<td>Steinberg, J.</td>
<td>Low</td>
</tr>
</tbody>
</table>

Figure 4.5 Frame population of 20 persons sorted alphabetically, with SRS sample realization of size \( n = 4 \).

Figure 4.6 Frame population of 20 persons sorted by group, with stratified element sample of size \( n_s = 1 \) from each stratum.
Proportionate Allocation to Strata

- Sample size within each strata is proportional to strata size in population
- If N is population size and n is total sample size, then $n_h / n = N_h / N$ where
  - $N_h$ is the population size of stratum $h$
  - $n_h$ is the sample size for stratum $h$
- Estimate the population mean as

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \left( \frac{N_h}{N} \right) \bar{y}_h$$
Proportionate Allocation to Strata

• Assuming SRS in each strata, the estimate of the variance of the sample mean is

\[
v \left( \bar{y}_{st} \right) = \sum_{h=1}^{H} W_h^2 \left( 1 - f_h \right) s_h^2 / n_h
\]

\[
= \sum_{h=1}^{H} \left( \frac{N_h}{N} \right)^2 \left( \frac{N_h - n_h}{N_h} \right) \left( \frac{1}{n_h - 1} \right) \left( \frac{1}{n_h} \right) \sum_{i=1}^{n_h} \left( y_{hi} - \bar{y}_h \right)^2
\]

\[
= \sum_{h=1}^{H} \left( \frac{N_h \left( N_h - n_h \right)}{N n_h \left( n_h - 1 \right)} \right) \sum_{i=1}^{n_h} \left( y_{hi} - \bar{y}_h \right)^2
\]
Example

Table 4.2. Proportionate Stratified Random Sample Results from a School Population Divided into Three Urbanicity Strata

<table>
<thead>
<tr>
<th>Stratum</th>
<th>(N_h)</th>
<th>(W_h)</th>
<th>(n_h)</th>
<th>(f_h)</th>
<th>(\bar{y}_h)</th>
<th>(s_h^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central city schools</td>
<td>3200</td>
<td>0.4</td>
<td>192</td>
<td>0.06</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Other urban schools</td>
<td>4000</td>
<td>0.5</td>
<td>240</td>
<td>0.06</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Rural schools</td>
<td>800</td>
<td>0.1</td>
<td>48</td>
<td>0.06</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>8000</td>
<td>1.0</td>
<td>480</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Note \(\bar{y} = (6 + 5 + 8) / 3 = 6.3\) (simple average)
- However \(\bar{y}_{st} = (0.4 \times 6) + (0.5 \times 5) + (0.1 \times 8) = 5.7\)
• Design effect compares how variation from stratified sampling compares to SRS

\[ d^2 = \frac{v(\bar{y}_{st})}{v(\bar{y}_{SRS})} = \frac{\sum_{h=1}^{H} W_h^2 (1 - f_h) s_h^2 / n_h}{(1 - f) s^2 / n} \]

• Design effect can be greater or less than 1
• But with reasonably homogeneous strata, almost always get decrease in variance
  – Means smaller std errors and confidence intervals
Systematic Sampling

- **Systematic sampling**: can be a simple way to do stratified sampling (proportional to size)
  - Basic idea: take every $k^{th}$ element in list-based sampling frame, $k = N / n$
    - Sort frame by strata
    - If $k$ is fractional, round up or down
    - Select a random integer between 1 and $N$
    - Start at that element in the frame and take every $k^{th}$ element thereafter (if the end of the list is reached, go to the beginning and continue)

- Aka implicitly stratified sampling
### Systematic Selection

<table>
<thead>
<tr>
<th>Record</th>
<th>Name</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cochran, W.</td>
<td>Highest</td>
</tr>
<tr>
<td>2</td>
<td>Hunt, J.</td>
<td>Highest</td>
</tr>
<tr>
<td>3</td>
<td>Madow, W.</td>
<td>Highest</td>
</tr>
<tr>
<td>4</td>
<td>Mandela, N.</td>
<td>Highest</td>
</tr>
<tr>
<td>5</td>
<td>Wolfe, T.</td>
<td>Highest</td>
</tr>
<tr>
<td>6</td>
<td>Bradburn, N.</td>
<td>High</td>
</tr>
<tr>
<td>7</td>
<td>Deming, W.</td>
<td>High</td>
</tr>
<tr>
<td>8</td>
<td>Hyde, H.</td>
<td>High</td>
</tr>
<tr>
<td>9</td>
<td>Sudman, S.</td>
<td>High</td>
</tr>
<tr>
<td>10</td>
<td>Wallman, K.</td>
<td>High</td>
</tr>
<tr>
<td>11</td>
<td>Fuller, W.</td>
<td>Medium</td>
</tr>
<tr>
<td>12</td>
<td>Habermann, H.</td>
<td>Medium</td>
</tr>
<tr>
<td>13</td>
<td>Kalton, G.</td>
<td>Medium</td>
</tr>
<tr>
<td>14</td>
<td>Norwood, J.</td>
<td>Medium</td>
</tr>
<tr>
<td>15</td>
<td>Woolsey, T.</td>
<td>Medium</td>
</tr>
<tr>
<td>16</td>
<td>Hansen, M.</td>
<td>Low</td>
</tr>
<tr>
<td>17</td>
<td>Kish, L.</td>
<td>Low</td>
</tr>
<tr>
<td>18</td>
<td>Rubin, D.</td>
<td>Low</td>
</tr>
<tr>
<td>19</td>
<td>Shoutley, P.</td>
<td>Low</td>
</tr>
<tr>
<td>20</td>
<td>Steiner, J.</td>
<td>Low</td>
</tr>
</tbody>
</table>

**Systematic Selection, RS = 2**

- **→** Hunt, J.
- **→** Deming, W.
- **→** Habermann, H.
- **→** Kish, L.

**Figure 4.7** Frame population of 20 persons sorted by group, with systematic selection, selection interval = 5, random start = 2.
What is Cluster Sampling?

• Units for survey occur in groups (clusters)

• **Cluster sampling**: use probability sampling to select clusters, survey all units in each cluster
  – Aka single stage cluster sampling

• There are more complicated types of cluster sampling such as **two-stage cluster sampling**
  – Select primary sampling units (PSUs) by probability sampling
  – Within each selected PSU, sample secondary sampling units (SSUs) via probability sampling
  – Survey all units in each selected SSU
Advantages and Disadvantages of Cluster Sampling

• Advantages:
  – For some populations, cannot construct list-based sampling frame
    • Can first sample by cluster (area, organizational, etc) then sample within
  – For some efforts, too expensive to conduct a SRS
    • E.g., drawing a SRS from the US population for an in-person interview

• Disadvantage: Cluster samples generally provide less precision than SRS or stratified samples
When To Use Cluster Sampling

• Use cluster sampling only when economically justified
  – I.e., when cost savings overcome (require) loss in precision

• Most likely to occur when
  – Constructing a complete list-based sampling frame is difficult, expensive, or impossible
  – The population is located in natural clusters (schools, city blocks, etc.)
Example #1: Small Village

Figure 4.4 A bird’s eye view of a population of 30 “●” and 30 “○” households clustered into six city blocks, from which two blocks are selected.
Example #2: NAEP

• Assume:
  – 40,000 4th grade classrooms in US
  – $B=25$ students per classroom

• Sampling procedure:
  – Select $a$ classrooms
  – Visit each classroom and collect data on all students
    • If $a = 8$, will have data on 200 students

• Note the differences from SRS
  – All groups of 200 students cannot be sampled
  – Students in each classroom more likely to be alike
Mean & Variance Computations

- Would calculate the mean test score as

\[ \bar{y} = \frac{\sum_{\alpha=1}^{a} \sum_{\beta=1}^{B} y_{\alpha\beta}}{aB} \]

- And the variance is

\[ v(\bar{y}) = \left( \frac{1 - f}{a} \right) s_a^2 \]

where

\[ s_a^2 = \left( \frac{1}{a - 1} \right) \sum_{\alpha=1}^{a} (\bar{y}_\alpha - \bar{y})^2 \]

- Key idea: Only the classrooms are random
The Design Effect

• Consider 8 classrooms with mean scores of 370, 370, 375, 375, 380, 380, 390, and 390
  – So, $\bar{y} = 378.75$
• Then $s_a^2 = 62.5$ and $v(\bar{y}) = 7.81$
• Suppose a SRS with $n=200$ gives $s^2 = 500$ so that $v_{SRS}(\bar{y}) = 2.50$
• Design effect ($d^2$) is the ratio of the variances:

$$d^2 = \left( \frac{v(\bar{y})}{v_{SRS}(\bar{y})} \right) = \frac{7.81}{2.50} = 3.13$$
• In example, the design effect says that clustering tripled the sampling variance.

• Means an increase in the standard error (and hence the confidence limits) of 77%.
  – Because $\sqrt{3.13} = 1.77$

• Says in this case we need almost twice the sample size as a SRS sample to get the same precision.
• I like to think about design effects in terms of effective sample size
  – What size SRS would give the same precision as the clustered sample?
• In previous example, we had $n = 200$ with $d^2 = 3.13$
  – The effective sample size is $n_{eff} = 200 / 3.13 = 64$
  – So we could have done a SRS of a sample of 64 and achieved the same precision
  – Would have meant going to $64/25 = 2.6$ times as many sites – perhaps unaffordable
What We Have Covered

• Introduction to sampling
  – Convenience vs. random sampling
  – Types of random sampling: simple random, stratified, and cluster sampling

• Defined “margin of error”

• Basic power calculations under SRS
  – Continuous vs. binary response

• Finite population correction