Stratified Sampling

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Reading Assignment:
Scheaffer, Mendenhall, Ott, & Gerow
Chapter 5
Goals for this Lecture

• Define stratified sampling
• Discuss reasons for stratified sampling
• Derive estimators for the population
  – Mean
  – Total
  – Percentage
• Sample size calculation and allocation
  – Sampling proportional to strata size
• Discuss design effects
Stratified Sampling

- **Stratified sampling** divides the sampling frame up into strata from which separate probability samples are drawn.

- **Examples**
  - GWI survey, needed to obtain information from members of each military service.
  - For external validity, WMD survey had to sample large urban areas.
  - In a particular country, survey in support of IW requires information on both Muslim and Christian communities.
Reasons for Stratified Sampling

• More precision
  – Assuming strata are relatively homogeneous, can reduce the variance in the sample statistic(s)
  – So, get better information for same sample size

• Estimates (of a particular precision) needed for subgroups
  – May be necessary to meet survey’s objectives
  – May have to oversample to get sufficient observations from small strata

• Cost
  – May be able to reduce cost per observation if population stratified into relevant groupings
An Example

### Frame population of 20 persons sorted alphabetically, with SRS sample realization of size $n = 4$.

<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>Haberman, 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>

One SRS of Size 4

### One Stratified Random Sample of Total Size 4

<table>
<thead>
<tr>
<th>Record</th>
<th>Name</th>
<th>Group</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>Haberman, 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 4.6 Frame population of 20 persons sorted by group, with stratified element sample of size $n = 4$ from each stratum.

Strata

- When selecting a stratified random sample, must clearly specify the strata
  - Non-overlapping categories into which each sampling unit must be classified
  - Sampling units can only be in one strata
  - Strata based on information about whole population

- Can have more than one type of classification
  - E.g., officer/enlisted and male/female
  - Then each sampling unit (person in this case) must be classified into one of four stratum
Some Notation

- Additional notation for stratified sampling
  - $L$ is the number of strata
  - $N_i$ is the number of sampling units in stratum $i$
  - $n_i$ is the sample size in stratum $i$
  - $N$ is the total number of sampling units in the population: $N = N_1 + N_2 + \ldots + N_L$

- In this lecture, we use SRS within each strata
  - That does not necessarily mean each strata has the same selection probabilities
  - And often more complex sampling done within the strata
Mean Estimation

- Within a strata, probability a unit is sampled is \( n_i / N_i \)

- Via the SRS lecture, we know that an unbiased estimate of the mean of strata \( i \) is
  \[
  \frac{1}{N_i} \sum_{j=1}^{n_i} \frac{1}{\pi_i} y_{ij} = \frac{1}{N_i} \sum_{j=1}^{n_i} \frac{1}{n_i / N_i} y_{ij} = \frac{1}{N_i} \sum_{j=1}^{n_i} N_i y_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_i
  \]

- So the estimated total for strata \( i \) is \( \hat{\tau}_i = N_i \bar{y}_i \)

- And the population mean estimate is just the sum of all the estimated totals divided by the population size
  \[
  \bar{y}_{st} = \frac{\hat{\tau}}{N} = \frac{1}{N} \sum_{i=1}^{L} \hat{\tau}_i = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_i
  \]
Mean Estimation Summary

• Estimator for the mean: \( \overline{y}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \overline{y}_i \)

• Variance of \( \overline{y}_{st} \):

\[
\widehat{\text{Var}}(\overline{y}_{st}) = \text{Var}\left( \frac{1}{N} \sum_{i=1}^{L} N_i \overline{y}_i \right) \\
= \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \text{Var}(\overline{y}_i) \\
= \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \left( \frac{s_i^2}{n_i} \right)
\]

• Bound on the error of estimation: \( 2\sqrt{\widehat{\text{Var}}(\overline{y}_{st})} \)
Example: TV Viewing Time (1)

- Survey to estimate average TV viewing time
- Population consists of two urban areas and rural residents
- Reasons to stratify:
  - Similarity of viewing habits by towns and rural regions
  - Cost: perhaps cheaper to survey each region separately (?)
- Sample allocated proportional to strata size

### Table 5.1

<table>
<thead>
<tr>
<th>Town A</th>
<th>Town B</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>27</td>
<td>8</td>
</tr>
<tr>
<td>43</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>36</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>39</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>28</td>
<td>49</td>
<td>30</td>
</tr>
<tr>
<td>28</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>29</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example: TV Viewing Time (2)

![Box plots of television-viewing time]

**Figure 5.1**
Box plots of television-viewing time

**Table 5.2**
Summary of the data from Table 5.1

<table>
<thead>
<tr>
<th>Stratum</th>
<th>( N )</th>
<th>( n )</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Town A</td>
<td>155</td>
<td>20</td>
<td>33.90</td>
<td>34.50</td>
<td>5.95</td>
</tr>
<tr>
<td>Town B</td>
<td>62</td>
<td>8</td>
<td>25.12</td>
<td>26.00</td>
<td>15.25</td>
</tr>
<tr>
<td>Rural</td>
<td>93</td>
<td>12</td>
<td>19.00</td>
<td>17.50</td>
<td>9.36</td>
</tr>
</tbody>
</table>

Example: TV Viewing Time (3)

A numerical descriptive summary of the data is shown in Table 5.2.

(a) From Table 5.2 and Eq. (5.1),

\[
\bar{y}_{st} = \frac{1}{N} \left[ N_1 \bar{y}_1 + N_2 \bar{y}_2 + \cdots + N_L \bar{y}_L \right]
\]

\[
= \frac{1}{310} \left[ (155)(33.900) + (62)(25.125) + (93)(19.00) \right]
\]

\[
= 27.7
\]

is the best estimate of the average number of hours per week that all households in the county spend watching television. Also,

\[
\hat{V}(\bar{y}_{st}) = \frac{1}{N^2} \sum N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \left( \frac{s_i^2}{n_i} \right)
\]

\[
= \frac{1}{(310)^2} \left[ \frac{(155)^2(0.871)(5.95)^2}{20} + \frac{(62)^2(0.871)(15.25)^2}{8} + \frac{(93)^2(0.871)(9.36)^2}{12} \right]
\]

\[
= 1.97
\]

The estimate of the population mean with an approximate 2-SD bound on the error of estimation is given by

\[
\bar{y}_{st} \pm 2 \sqrt{\hat{V}(\bar{y}_{st})} \quad \text{or} \quad 27.675 \pm 2 \sqrt{1.97} \quad \text{or} \quad 27.7 \pm 2.8
\]

Example: TV Viewing Time (4)

- Correct analysis:
  \[ \bar{y}_{st} = 27.7, \quad \hat{\sigma}_{\bar{y}_{st}} = 1.4, \quad 95\% \text{ CI} = (24.9, \ 30.5) \]

- What if you just treated this as a SRS?
  - Incorrect analysis with fpc:
    \[ \bar{y}_{SRS} = 27.7, \quad \hat{\sigma}_{\bar{y}} = 10.6, \quad 95\% \text{ CI} = (6.5, \ 48.9) \]
  - Incorrect analysis without fpc:
    \[ \bar{y}_{SRS}' = 27.7, \quad \hat{\sigma}_{\bar{y}} = 11.3, \quad 95\% \text{ CI} = (5.1, \ 50.3) \]
The design effect (aka deff) compares how a complex sampling design, in this case stratified sampling, compares to SRS.

\[
d^2 = \frac{\text{Var}(\bar{Y}_{st})}{\text{Var}(\bar{Y}_{SRS})} = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left( 1 - \frac{n_i}{N_i} \right) s_i^2 / n_i \frac{\left( 1 - \frac{n}{N} \right) s^2 / n}{\left( 1 - \frac{n}{N} \right) s^2 / n}
\]

Design effect can be greater or less than 1

- But with reasonably homogeneous strata, almost always get decrease in variance
- For this example,

\[
d^2 = \frac{\text{Var}(\bar{Y}_{st})}{\text{Var}(\bar{Y}_{SRS})} = \frac{1.97}{112.36} = 0.02
\]
Total Estimation Summary

- Estimator for the total:  \( \hat{\tau}_{st} = N\overline{y}_{st} = \sum_{i=1}^{L} N_i \overline{y}_i \)
- Variance of \( \hat{\tau}_{st} \):
  \[
  \widehat{\text{Var}}(\hat{\tau}_{st}) = \text{Var} \left( \sum_{i=1}^{L} N_i \overline{y}_i \right) = \sum_{i=1}^{L} N_i^2 \text{Var}(\overline{y}_i) = \sum_{i=1}^{L} N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \left( \frac{s_i^2}{n_i} \right)
  \]
- Bound on the error of estimation:  \( 2\sqrt{\text{Var}(\hat{\tau}_{st})} \)

2/1/13
Proportion Estimation Summary

• Proportion estimator: \( \hat{p}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \hat{p}_i \)

• Variance of \( \bar{y}_{st} \):

\[
\widehat{\text{Var}}(\hat{p}_{st}) = \text{Var}
\left( \frac{1}{N} \sum_{i=1}^{L} N_i \hat{p}_i \right)
\]

\[
= \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \text{Var}(\hat{p}_i)
\]

\[
= \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \left(\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}\right)
\]

• Bound on the error of estimation: \( 2\sqrt{\text{Var}(\hat{p}_{st})} \)
Sample Size Determination & Allocation

- Calculations more complicated as need to
  - Choose overall sample size $n$, and
  - Allocate sample to strata, $n_1 + n_2 + \cdots + n_L = n$

- “Best” allocation depends on
  - Purpose of the stratification
  - Number of sampling units in each stratum
  - Variability of sampling units within each stratum
  - Cost of surveying each sampling unit from each stratum
Sample Size Calculation for Estimating the Mean

• Begin as with SRS, setting $B = 2\sqrt{\text{Var}(\bar{y}_{st})}$
• Now, define $a_i$ as the proportion of the sample size to allocate to strata $i$, $n_i = n \times a_i$
• Then,

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 \sigma_i^2}{a_i}}{N^2 B^2 / 4 + \sum_{i=1}^{L} N_i^2 \sigma_i^2}$$
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 \( n_i / n = N_i / N \) where
  - \( N_i \) is the population size of stratum \( i \)
  - \( n_i \) is the sample size for stratum \( i \)
- Then
  \[
  n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / a_i}{N^2 B^2 / 4 + \sum_{i=1}^{L} N_i^2 \sigma_i^2} = \frac{\sum_{i=1}^{L} (N_i^2 \sigma_i^2) / (N_i / N)}{N^2 B^2 / 4 + \sum_{i=1}^{L} N_i^2 \sigma_i^2} = \frac{\sum_{i=1}^{L} N_i \sigma_i^2}{N B^2 / 4 + \frac{1}{N} \sum_{i=1}^{L} N_i \sigma_i^2}
  \]
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>$N_i$</th>
<th>$N_i/N$</th>
<th>$n_i$</th>
<th>$n_i/N_i$</th>
<th>$\bar{y}_i$</th>
<th>$s_i^2$</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 $\bar{y}_{st} = (0.4 \times 6) + (0.5 \times 5) + (0.1 \times 8) = 5.7$

Other Allocation Schemes

• Rather than allocating sample proportional to strata size, can
  – Allocate according to variability of the strata
    • Idea is to allocate more of the sample to strata that are more variable
    • Done right, can provide most precise population estimates
  – Allocate according to the cost of collection per strata
    • Idea is to allocate more of the sample to strata that cost less
Poststratification

• Poststratification used when strata variable(s) not observed until after survey completed
  – That is, stratify on data collected in the survey

• Useful when sample demographics turn out not to match the population demographics
  – As with stratification, still need to know the distribution of the strata variable in the population
  – But don’t need to know the values for each element in the sampling frame

• See text for more detail
What We Have Covered

• Defined stratified sampling
• Discussed reasons for stratified sampling
• Derived estimators for the population
  – Mean
  – Total
  – Percentage
• Sample size calculation and allocation
  – Sampling proportional to strata size
• Discussed design effects