Sampling and Survey

April 22, 2009
1. **Introduction**
   - Sampling in life

2. **Basic Concepts**
   - Survey Design
   - Sampling
   - How to select the sample
   - Probability Sampling methods
   - Sources of errors in sampling

3. **Sampling methods**
   - Simple Random Sample
   - Stratified Random Sample
   - Systematic Sampling
   - Cluster sampling
Sampling-Induction

- Sampling food and drink
- Sampling a new product to be introduced or already in the market.
- Your example of preference.
When observed instances are identical or almost so, only a few are needed to be observed.

Repeated observations are for confirmation.

When items indeed vary, drawing conclusion based on one or few samples may be risky.

Repeated observations are then required to make strong inferences.
Sampling is applied to

It lets finding out characteristics and events in:

- Human’s
- Plant populations
- Physical objects
- Animals’s
- Etc.
Meeting the objectives

In order to meet easily the survey’s objectives, the following are to be considered

- **Subject matter issues.**
  - Population to be surveyed
  - Statistics to be obtained
  - Data to be collected
  - Time periods
  - Accuracy
  - Analysis
  - Reports’ design and date of delivery.

- **Operational issues.**
  - Methods for obtaining the data
  - Survey data processing
  - Monitoring operational performance
  - Timetables for survey operations
• Administrative issues.
  • Project approvals
  • Finance
  • Recruiting, Training, Supervision, remuneration, transportation, Office space, Equipment, supplies, communications, relation with public, informants, staff, etc.

• Sampling issues
  • Sample frame
  • Sample size and its allocation
  • Domains of study
  • Methods for estimating the survey results and their sample and nonsample errors.
**Element** is an object in which a measurement is taken.

**Population** is a collection of elements about which we wish to make an inference.

**Sampling units** are non-overlapping collections of elements from the population that cover the entire pop.

**Frame** is a list of sampling units.

**Sample** is a collection of sampling units drawn from a frame or frames.
The goal is estimate population parameters ($\theta$) from the sample such as

- Mean
- Total
- Proportions
- Etc.

**To determine the sampling method and the sample size**

- Every item contains a certain amount of information.
- The quantity of information a sample has depends on the number of items in it and the variability.
- $\hat{\theta}$ estimates $\theta$ in the population, then

$$\text{error of estimation} = |\theta - \hat{\theta}| < B \text{ (the bound)}$$

- and, the certainty

$$P[|\theta - \hat{\theta}| < B] = (1 - \alpha)$$
The common value for $B = 2\sigma_{\hat{\theta}}$

Hence, the method of sampling is chosen, based on the following: Choose the sampling method with the highest certainty and the lowest error of estimation, together with the minimum cost.
1. **Simple random sample.** Any sample of size $n$ has the same chance of being selected. Property. It will contain as much information from the population as any other sample survey design if that population were homogeneous.

2. **Stratified random sample.** Nonetheless, suppose that pop consists of items clearly identified as belonging to any of a number of groups (i.e. people belonging to either group of low income or high income), estimation from that pop will be remarkably more accurate if the sample comes from all of the groups. This is achieved by knowing the presence of auxiliary variables, that define the groups or strata.
3 Cluster Sampling. Usually used in urban areas. This method consists of simple random selection of groups. Such as city blocks, families, or apartment buildings. Once a given group is selected, all the items in the group are sampled.

4 Systematic Sample. Sometimes a list of items is available. This method consists on randomly selecting one item close to the beginning of the list and proceed the selection of the sample every tenth or fifteenth item thereafter.
Errors of non-observation.

- **Sampling.** The data observed in a sample will not precisely mirror the data in the pop regardless how careful the measurements are done. The most common error is the Sample Bias.
- **Coverage.** The sampling frame does not match up perfectly with the target pop.
- **Non-response.** The inability to contact the sampled item, the inability of an interviewed to come up with the answer, or just because the interviewed refuses to answer.
Error of observation.

- **Enumerators.** Have a direct and dramatic effect on the way a interviewed responds to a question. Reading without appropriate emphasis or intonation.
- **Respondents.** Inability to answer correctly.
- **Instrument.** The inability to clarify the measurement units a question refers to, inches or centimeter. Clarifying terms: employment or unemployment rates.
- **Method of data collection.** Either direct interview, by phone calls, or by mail, now by websites.
Simple Random Sample
**Definition.** *Simple Random Sample* is a sampling procedure such that a sample size $n$ is drawn from a pop size $N$, in a way that any sample of the same size has the same chance of being selected.

**How to draw a simple random sample.** Using *R – Software* load and install *Sampling* package. $srswor(n, N)$ function selects the sample for you.

**Estimation of a population mean, total, and proportion**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
<th>Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$</td>
<td>$\hat{V}(\bar{y}) = \frac{s^2}{n} (1 - \frac{n}{N})$</td>
<td>$2\sqrt{\hat{V}(\bar{y})}$</td>
</tr>
<tr>
<td>Total</td>
<td>$\hat{\tau} = N\bar{y} = \frac{N \sum_{i=1}^{n} y_i}{n}$</td>
<td>$\hat{V}(N\bar{y}) = N^2 \frac{s^2}{n} (1 - \frac{n}{N})$</td>
<td>$2\sqrt{\hat{V}(N\bar{y})}$</td>
</tr>
<tr>
<td>Proportion</td>
<td>$\hat{p} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$</td>
<td>$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1} (1 - \frac{n}{N})$</td>
<td>$2\sqrt{\hat{V}(\hat{p})}$</td>
</tr>
</tbody>
</table>
The following code does the sample selection and compute the CI for the mean of the Height for you. The data set used is the one you had accessed before in a file posted in the LISA — website, for this course.
Assuming $n = 10$ and $N = 57$.

\begin{verbatim}
n=10; N=57; sam=srswo(n,N);
srsexample=getdata(classssur,sam);
meanHeight=mean(classssur$Height,na.rm=TRUE)
varHeight=var(classssur$Height,na.rm=TRUE)
meanHeight.sample=mean(srsexample$Height,na.rm=TRUE)
varHeight.sample=var(srsexample$Height,na.rm=TRUE)
var.simpless.Height=(varHeight.sample/n)*(1-n/N)
B=2*sqrt(var.simpless.Height)
Estimation=data.frame(meanHeight.sample, var.simpless.Height,meanHeight.sample-B, meanHeight, meanHeight.sample+B)
colna=c("Height sample Mean","Height sample Var", "LCL","RealMean","UCL")
colnames(Estimation)=colna
\end{verbatim}
This is the output for the code above. (it is a 95% CI),

<table>
<thead>
<tr>
<th>$\bar{y}_{\text{Height}}$</th>
<th>$s_{\text{Height}}^2$</th>
<th>LCL</th>
<th>$\mu_{\text{Height}}$</th>
<th>UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>66.20</td>
<td>0.98</td>
<td>64.22</td>
<td>67.02</td>
<td>68.18</td>
</tr>
</tbody>
</table>
CI for estimating the population mean of the Height

The number of times the CI covers the real value is 96 out of 100.
Sample size to estimate $\mu$ with a bound on the error of estimation $B$ is given by

$$n = \frac{N\sigma^2}{(N - 1)\frac{B^2}{4} + \sigma^2}$$

Sample size to estimate $\tau$ with a bound on error $B$,

$$n = \frac{N\sigma^2}{(N - 1)\frac{B^2}{4N^2} + \sigma^2}$$

Sample size to estimate $p$ (proportion) with a bound on error $B$,

$$n = \frac{Npq}{(N - 1)\frac{B^2}{4} + pq}$$
**Example for the mean.** Assume that our data set is the frame of our population. We want to estimate the mean of the Height. From a previous study, we know that the standard deviation of the population’s Height is about 4.5. We also required an error of estimation $B=2$ inches. What is the needed sample size?

*Soln.* We have to compute the sample size. Remember that $N=57$. Thus

$$n = \frac{57 \times 4.5^2}{(57 - 1) \frac{2^2}{4} + 4.5^2} = 15.14 \approx 16$$
**Example on proportion.** Now we want to estimate the proportion of women in this class. We do not know anything about women 😞!!!. We also required an error of estimation $B=0.05$. What is the sample size needed get our estimation.

**Soln.**

$$n = \frac{57(0.5)(0.5)}{(57 - 1) \frac{0.05^2}{4} + (0.5)(0.5)} = 50$$

This large value of the sample size is due to (1) We do not know much about the proportion, and $p=0.5$ is the value that maximizes the variance in the pop, and (2) because the population size is really small.
Stratified Random Sample

Outline
Introduction
Basic Concepts
Sampling methods

Simple Random Sample
Stratified Random Sample
Systematic Sampling
Cluster sampling

Sampling and Survey
• **Definition.** It is obtained by separating the population elements into homogenous groups, called *strata*.

• **How to draw a stratified.** Using *R — software*, the *strata* function in the *Sampling* package, does the selection for you.

• **Estimation of a population mean, total, and proportion**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
<th>Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{y}<em>{st} = \frac{1}{N} \sum</em>{i=1}^{L} N_i \bar{y}_i$</td>
<td>$\hat{V}(\bar{y}<em>{st}) = \frac{1}{N^2} \sum</em>{i=1}^{L} N_i^2 (1 - \frac{n_i}{N_i}) (\frac{s_i^2}{n_i})$</td>
<td>$2\sqrt{\hat{V}(\bar{y}_{st})}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N \bar{y}<em>{st} = \sum</em>{i=1}^{L} N_i \bar{y}_i$</td>
<td>$\hat{V}(N \bar{y}<em>{st}) = \sum</em>{i=1}^{L} N_i^2 (1 - \frac{n_i}{N_i}) (\frac{s_i^2}{n_i})$</td>
<td>$2\sqrt{\hat{V}(N \bar{y}_{st})}$</td>
</tr>
<tr>
<td>Proportion</td>
<td>$\hat{p}<em>{st} = \frac{1}{N} \sum</em>{i=1}^{L} N_i \hat{p}_i$</td>
<td>$\hat{V}(\hat{p}<em>{st}) = \frac{1}{N^2} \sum</em>{i=1}^{L} N_i^2 (1 - \frac{n_i}{N_i}) (\frac{\hat{p}_i \hat{q}_i}{n_i - 1})$</td>
<td>$2\sqrt{\hat{V}(\hat{p}_{st})}$</td>
</tr>
</tbody>
</table>
The following example is again based on our data set. Now we define Gender as the variable to stratify. This variable has two strata. We also want to estimate the mean of the Height. Using R, we get the following output.
### Table: Stratified random sample R-output

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>StudyHrs</th>
<th>SleepHrs</th>
<th>Job</th>
<th>Textpay</th>
<th>Gender</th>
<th>ID_unit</th>
<th>Stratum</th>
<th>nh</th>
<th>Nh</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>155</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>150</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>6.00</td>
<td>23.00</td>
</tr>
<tr>
<td>67</td>
<td>120</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>180</td>
<td>2</td>
<td>11</td>
<td>1</td>
<td>6.00</td>
<td>23.00</td>
</tr>
<tr>
<td>64</td>
<td>106</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td>200</td>
<td>2</td>
<td>21</td>
<td>1</td>
<td>6.00</td>
<td>23.00</td>
</tr>
<tr>
<td>67</td>
<td>na</td>
<td>15</td>
<td>7</td>
<td>2</td>
<td>198</td>
<td>2</td>
<td>33</td>
<td>1</td>
<td>6.00</td>
<td>23.00</td>
</tr>
<tr>
<td>na</td>
<td>an</td>
<td>14</td>
<td>4</td>
<td>2</td>
<td>260</td>
<td>2</td>
<td>45</td>
<td>1</td>
<td>6.00</td>
<td>23.00</td>
</tr>
<tr>
<td>68</td>
<td>122</td>
<td>13</td>
<td>6</td>
<td>1</td>
<td>250</td>
<td>2</td>
<td>55</td>
<td>1</td>
<td>6.00</td>
<td>23.00</td>
</tr>
<tr>
<td>72</td>
<td>160</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>111</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>6.00</td>
<td>34.00</td>
</tr>
<tr>
<td>70</td>
<td>160</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>245</td>
<td>1</td>
<td>17</td>
<td>2</td>
<td>6.00</td>
<td>34.00</td>
</tr>
<tr>
<td>67</td>
<td>147</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>220</td>
<td>1</td>
<td>31</td>
<td>2</td>
<td>6.00</td>
<td>34.00</td>
</tr>
<tr>
<td>75</td>
<td>160</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>200</td>
<td>1</td>
<td>40</td>
<td>2</td>
<td>6.00</td>
<td>34.00</td>
</tr>
<tr>
<td>67</td>
<td>153</td>
<td>15</td>
<td>9</td>
<td>2</td>
<td>90</td>
<td>1</td>
<td>44</td>
<td>2</td>
<td>6.00</td>
<td>34.00</td>
</tr>
<tr>
<td>58</td>
<td>155</td>
<td>16</td>
<td>6</td>
<td>1</td>
<td>200</td>
<td>1</td>
<td>57</td>
<td>2</td>
<td>6.00</td>
<td>34.00</td>
</tr>
</tbody>
</table>
In R, by the function *syvdesign* we analyzed the data and along with *syvmean* function, the following result is obtained,

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>$s_{Height}^2$</th>
<th>LCI</th>
<th>Real Mean</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>67.92</td>
<td>0.79</td>
<td>66.15</td>
<td>67.02</td>
<td>69.70</td>
</tr>
</tbody>
</table>

It is clear that this method gives a more precise mean estimation. It is because comparing the variances of both methods, this one above is smaller. Then a narrower CI.
The sample size is then computed

- The sample size required to estimate $\mu$ or $\tau$ with a bound $B$ on the error of estimation is

$$n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / w_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2},$$

where $D = \frac{B^2}{4}$ for $\mu$ or $D = \frac{B^2}{4N^2}$ for $\tau$, and $w_i$ is the fraction of observation allocated to stratum $i$. Thus,

$$w_i = n_i / n$$
The sample size required to estimate a proportion with a bound B on the error of estimation is

\[ n = \frac{1}{\sum_{i=1}^{L} N_i^2 p_i q_i / w_i} \left( N^2 \frac{B^2}{4} + \sum_{i=1}^{L} N_i p_i q_i \right) \]

Again,

\[ w_i = n_i / n \]
Example for estimating the mean. Going back to our data set. We want to estimate the mean under the strata scheme. We have two strata defined by Gender variable. We barely know that on each strata the standard deviation of the Height is approximately: 4 and 3. We know that \( N = 57 = N_{\text{Gender}=1}(= 23) + N_{\text{Gender}=2}(= 34) \).

We want a sample equally weighted across strata. Determine the sample size, with an error of estimation=2 inches. From our basic statistic courses, we know that the standard deviation is approximately one fourth of the range. Hence,

\[
D = 2^2 / 4 = 1
\]

\[
n = \frac{23^2(4)^2 / (1/2) + 34^2(3)^2 / (1/2)}{57^21 + (23(4)^2 + 34(3)^2)} = 9.6 \approx 10
\]

[Note: compare this sample size with the one obtained by Simple Random Sample]
Allocation

There are two most important ways to allocate the sample into strata:

- Considering the cost, which is the general allocation

$$n_i = n \left( \frac{N_i \sigma_i / \sqrt{c_i}}{\sum_{i=1}^{L} N_i \sigma_i / \sqrt{c_i}} \right)$$

- Neyman allocation, making $c_i = c, \forall \ i=1, \ldots, L$

$$n_i = n \left( \frac{N_i \sigma_i}{\sum_{i=1}^{L} N_i \sigma_i} \right)$$
Example. Continuation of the last example. Allocate the calculated sample size into those strata, assuming that the real standard deviation are those stated in the problem. Assuming also, equal costs.

\[ n_1 = 10 \left( \frac{23 \times 4}{23 \times 4 + 34 \times 3} \right) = 4.7 \approx 5 \]

\[ n_2 = 10 \left( \frac{34 \times 3}{23 \times 4 + 34 \times 3} \right) = 5.2 \approx 5 \]
Systematic Sample

N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1 26 51 76
2 27 52 77
3 28 53 78
4 29 54 79
5 30 55 80
6 31 56 81
7 32 57 82
8 33 58 83
9 34 59 84
10 35 60 85
11 36 61 86
12 37 62 87
13 38 63 88
14 39 64 89
15 40 65 90
16 41 66 91
17 42 67 92
18 43 68 93
19 44 69 94
20 45 70 95
21 46 71 96
22 47 72 97
23 48 73 98
24 49 74 99
25 50 75 100
• **Definition.** A sample is obtained by randomly selecting one element from the first $k \leq \frac{N}{n}$ elements and every $k^{th}$ element thereafter.

• **Pros and Cons.**
  - It is easier to perform in the field and therefore is less subject to selection errors.
  - Estimates can be more precise.
  - If the population is grouped, the sample will contain information among those groups.
  - The sample is spread uniformly over the population.
  - Without knowing the population, it could induce some of the sampling bias.
  - To get a more accurate estimation of the estimator’s variance multiple systematic samples are required.
**Estimation (From only one systematic sample).** Either a mean, a total, or a proportion estimation is required, the formulas are the same as those for Simple Random Sample.

**How to draw a systematic sample.** Using R, the function `UPsystematic` draws the sample for you. It is necessary to define a vector of probabilities. Since this method can be generalized to a unequal sampling probability. For this case, that probability is the same for every single item.
Exercise. Select a sample of size $n = 10$, systematically. Using R, we get the following table. It is only part of the table.

<table>
<thead>
<tr>
<th>ID_unit</th>
<th>Gender</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>71</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>2</td>
<td>66</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>2</td>
<td>61</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>2</td>
<td>64</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>2</td>
<td>63</td>
</tr>
<tr>
<td>36</td>
<td>36</td>
<td>2</td>
<td>65</td>
</tr>
<tr>
<td>42</td>
<td>42</td>
<td>1</td>
<td>73</td>
</tr>
<tr>
<td>47</td>
<td>47</td>
<td>1</td>
<td>75</td>
</tr>
<tr>
<td>53</td>
<td>53</td>
<td>2</td>
<td>68</td>
</tr>
</tbody>
</table>

It is clear that this induced sample bias. Do you notice it?
In the case of repeated systematic samples, we clearly can estimate the variance of \( \bar{y}_{sy} \) through ANOVA idea. In the following context. Assume the following table,

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Sample Number</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (initials)</td>
<td>( y_{11} ) ( y_{12} ) ( y_{1n_s} )</td>
<td>( \bar{y}_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( y_{21} ) ( y_{22} ) ( y_{2n_s} )</td>
<td>( \bar{y}_2 )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( c )</td>
<td>( y_{c1} ) ( y_{c2} ) ( y_{cn_s} )</td>
<td>( \bar{y}_k )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \bar{y} )</td>
</tr>
</tbody>
</table>

\[
\hat{V}(\hat{\mu}) = \frac{(N - n)}{N} \frac{\sum_{i=1}^{c} (\bar{y}_i - \bar{y})^2}{c(c-1)}
\]

Note:

\[
V(\bar{y}_{sy}) = \frac{\sigma^2}{n} (1 + (n - 1)\rho)
\]
How to draw multiple systematic samples?

<table>
<thead>
<tr>
<th>One Sys</th>
<th>Multiple Sys</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=960</td>
<td>N=960</td>
</tr>
<tr>
<td>n=60</td>
<td>n=60</td>
</tr>
<tr>
<td>( k = \frac{960}{60} = 16 )</td>
<td>( 10k = 10 \times 16 = 160 = k' )</td>
</tr>
<tr>
<td>( c=1 ) w. ( n_s = n )</td>
<td>( c=6 ) w. ( n_s = 10 )</td>
</tr>
<tr>
<td>1 in ( k )</td>
<td>10 in ( k' )</td>
</tr>
</tbody>
</table>
Exercise. Based on the sample we got in the exercise above, using R, we got the following result. The sample mean of the Height and the standard error of this mean are given by, 

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>67.30</td>
<td>1.29</td>
</tr>
</tbody>
</table>

And

<table>
<thead>
<tr>
<th>LCI</th>
<th>Real mean</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>64.72</td>
<td>67.02</td>
<td>69.88</td>
</tr>
</tbody>
</table>

This CI is wide. It is because this sampling method induced a sample bias, it is also increasing the variability on the estimation.
**Textbook example.** The QC section uses systematic sampling to estimate the average amount of fill in 12-ounces cans. The data was obtained in a 1 in 15 sys. sample in one day. Estimate $\mu$ and place a bound on the error of estimation.

<table>
<thead>
<tr>
<th>Amount of fill (in ounces)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.00 11.91 11.87 12.05 11.75 11.85</td>
</tr>
<tr>
<td>11.97 11.98 12.01 11.87 11.93 11.98</td>
</tr>
<tr>
<td>12.01 12.03 11.98 11.91 11.95 11.87</td>
</tr>
<tr>
<td>12.03 11.98 11.87 11.93 11.97 12.05</td>
</tr>
<tr>
<td>12.01 12.00 11.90 11.94 11.93 12.02</td>
</tr>
<tr>
<td>11.80 11.83 11.88 11.89 12.05 12.04</td>
</tr>
</tbody>
</table>

This is clearly a one systematic sample.
Soln. To apply the analysis for a multiple systematic sampling, we assume that we have \( c = 6 \) clusters and \( n_s = 6 \).

- The sample mean is (which is the pop mean estimator):
  \[
  \bar{y} = \frac{12.0 + 11.97 + 12.01 + \ldots + 12.02 + 12.04}{36} = 11.95
  \]

- The sample variance, given that we can compute 6 different means which are:
  \[
  11.91 \mid 11.96 \mid 11.96 \mid 11.97 \mid 11.97 \mid 11.92,
  \]
  is
  \[
  \frac{\sum_{i=1}^{6} (\bar{y}_i - \bar{y})^2}{(6 - 1)} = 0.0007985.
  \]
Now, just using the formula for the variance of the multiple systematic sampling, and under the assumption that $N$ is unknown and large ($\frac{N-n}{N} = 1$), we get,

$$\hat{V}(\hat{\mu}) = \frac{0.0007985}{6} = 0.000133$$

- The bound is:
  $$2\sqrt{0.000133} = 0.023$$
- The 95% CI for the pop mean is:
  $$(11.92, 11.97)$$
**Definition.** A *cluster sample* is a probability sample in which each sampling unit is a "collection" of elements. Those collections could be,

- City blocks
- Land plots
- Forest regions
- etc.

It reduces the cost of obtaining observations by reducing the distances among elements.

**How to draw a sample.** Take advantage of R, using *cluster* function in *Sampling* package. It is needed to have in the data set a clustering variable. Then a Simple Random Sample on the clusters is carried out.
Example. This is a new data set. For this data set select a sample of 15 states. The columns on this data set are (From US Bureau of the Census, 1995):

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 State</td>
<td>US states</td>
</tr>
<tr>
<td>2 ExpPP90 and 92</td>
<td>Expenditure (dls) per student</td>
</tr>
<tr>
<td>3 ExpPC90 and 92</td>
<td>Expenditure (dls) per capita</td>
</tr>
<tr>
<td>4 TeaSal90 and 92</td>
<td>Average teacher salary (thousand dls)</td>
</tr>
<tr>
<td>5 Comp1</td>
<td>% of residents over 25, w. complete High School, 1990</td>
</tr>
<tr>
<td>6 Dropout</td>
<td>% of people 16-19 of age dropped out HS, 1990</td>
</tr>
<tr>
<td>7 Region</td>
<td>1:NE, 2:Midwest, 3: S, 4:W</td>
</tr>
<tr>
<td>8 Pop</td>
<td>Population in millions, 1992</td>
</tr>
<tr>
<td>9 Enroll</td>
<td>Students enrolled, 1990 (thousands)</td>
</tr>
<tr>
<td>10 Teachers</td>
<td>Teachers, 1990 (thousands)</td>
</tr>
</tbody>
</table>
**Soln.** Using R, we got the following result.

<table>
<thead>
<tr>
<th>STATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Alaska</td>
</tr>
<tr>
<td>2 California</td>
</tr>
<tr>
<td>3 Colorado</td>
</tr>
<tr>
<td>4 Connecticut</td>
</tr>
<tr>
<td>5 DC</td>
</tr>
<tr>
<td>6 Florida</td>
</tr>
<tr>
<td>7 Indiana</td>
</tr>
<tr>
<td>8 Iowa</td>
</tr>
<tr>
<td>9 Louisiana</td>
</tr>
<tr>
<td>10 N_Dakota</td>
</tr>
<tr>
<td>11 New_Mexico</td>
</tr>
<tr>
<td>12 Ohio</td>
</tr>
<tr>
<td>13 Oklahoma</td>
</tr>
<tr>
<td>14 Pennsylvania</td>
</tr>
<tr>
<td>15 Texas</td>
</tr>
</tbody>
</table>
As it could be expected, estimations’ formulae become messier. Expressions for estimating the mean and total of a population, are

\[
\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}, \text{ and } \hat{\tau} = M\bar{y}
\]

It is needed to define some other terms,

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Number of clusters in the pop</td>
</tr>
<tr>
<td>n</td>
<td>Number of clusters selected</td>
</tr>
<tr>
<td>(m_i)</td>
<td>Number of elements in cluster (i)</td>
</tr>
<tr>
<td>(\bar{m})</td>
<td>Average cluster size for the sample</td>
</tr>
<tr>
<td>M</td>
<td>Number of elements in the pop</td>
</tr>
<tr>
<td>(\bar{M} = M/N)</td>
<td>Average cluster size in pop</td>
</tr>
<tr>
<td>(y_i)</td>
<td>Total of all observations in cluster (i)</td>
</tr>
</tbody>
</table>
The estimated variance of $\hat{\mu}$ and $\hat{\tau}$,

$$
\hat{V}(\bar{y}) = \frac{(N-n)}{Nn\bar{M}^2} s^2_r, \quad \text{and} \quad \hat{V}(\hat{\tau}) = M^2 \hat{V}(\bar{y}); \quad s^2_r = \frac{\sum_{i=1}^{n} (y_i - \bar{y}m_i)^2}{n-1}
$$

or for estimating $\hat{\tau}$ without $M$ dependence,

$$
N\bar{y}_t = \frac{N}{n} \sum_{i=1}^{n} y_i; \quad \hat{V}(\hat{\tau}) = N^2 \frac{(N-n)}{Nn} s^2_t; \quad s^2_t = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_t)^2}{n-1}
$$

$\bar{M}$ can be estimated by $\bar{m}$ if $M$ is unknown, for estimating $\hat{\mu}$.
**Exercise.** For the sample selected above, estimate the average teachers’ salary for the US, 1992. Place a bound of error of estimation.

Using R, we get the following 95% CI for estimating the population mean.

<table>
<thead>
<tr>
<th></th>
<th>(\bar{y})</th>
<th>(s_r)</th>
<th>LCI</th>
<th>Mean (pop)</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers’ mean</td>
<td>33.99</td>
<td>1.66</td>
<td>30.67</td>
<td>34.14</td>
<td>37.31</td>
</tr>
</tbody>
</table>

R estimated \(\bar{M}\) by \(\tilde{m}\).

Since we know the values for \(N\) and \(\bar{M}\), thus, doing the computation in Excel, the following results are obtained.

<table>
<thead>
<tr>
<th></th>
<th>(\bar{y})</th>
<th>(s_r)</th>
<th>LCI</th>
<th>Mean (pop)</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers’ mean salary</td>
<td>33.99</td>
<td>2.37</td>
<td>29.25</td>
<td>34.14</td>
<td>38.74</td>
</tr>
</tbody>
</table>
The sample size determination is done by the following formulas.

- To estimate $\mu$ with a bound $B$ on the error,

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2}$$

Where $D = (B\bar{M})^2/4$. $\sigma_r^2$ is estimated by $s_r^2$.

- To estimate $\tau$ with a bound $B$ on the error,

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2}$$

Where $D = B^2/(4N^2)$. $\sigma_r^2$ is estimated by $s_r^2$. 

The extension of these formulas to estimate proportions, is straightforward.
Example of the analysis in multinomial type data set
Should smoking be banned from the workplace?
A Time/Yankelovitch poll of 800 adult americans carried out on April 6-7, 1994 (Time april 18, 1994)

<table>
<thead>
<tr>
<th></th>
<th>Nonsmokers</th>
<th>Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banned</td>
<td>44%</td>
<td>8%</td>
</tr>
<tr>
<td>Special Areas</td>
<td>52%</td>
<td>80%</td>
</tr>
<tr>
<td>No restriction</td>
<td>3%</td>
<td>11%</td>
</tr>
<tr>
<td>Total</td>
<td>100%(600)</td>
<td>100%(200)</td>
</tr>
</tbody>
</table>

1. The true difference between the proportions choosing “banned” The prop choosing banned are independent of each other. Thus the 95% CI for the difference estimate is:

\[(.44 - .08) \pm 2 \sqrt{\frac{.44 \times .56}{600} + \frac{.08 \times .92}{200}}\]

\[0.36 \pm 0.06 = 30\%, 42\%\]
The true difference between the proportions of nonsmokers choosing "banned" and "especial areas". This is a multinomial type data, the estimate of the diff is,

\[
(.52 - .44) \pm 2 \sqrt{\frac{.44 \times .56}{600} + \frac{.52 \times .48}{600} + 2 \frac{.52 \times .44}{600}}
\]

\[
0.08 \pm 0.08 = 0.16\%
\]

You may ask why these formulas are used.
In a multinomial type data set, regardless the number of classes you have, for proportions we have the following expressions

$$\hat{p}_i = \sum_{j=1}^{n} \frac{y_{ij}}{n} \text{ for all } i = 1, \ldots, k(\text{classes})$$

$$V(\hat{p}_i) = \frac{p_i(1 - p_i)}{n} \text{ for all } i = 1, \ldots, k(\text{classes})$$

$$\text{Cov}(\hat{p}_i, \hat{p}_j) = -\frac{p_ip_j}{n} \text{ for all } (i \neq j) = 1, \ldots, k(\text{classes})$$

And

$$V(\hat{p}_i - \hat{p}_j) = V(\hat{p}_i) + V(\hat{p}_j) - 2\text{Cov}(\hat{p}_i, \hat{p}_j)$$