Sampling Theory and Methods

Spring 2008

C. L. Williams

Chapter 5 Stratified Sampling

Outline

Stratification
and Stratified Random Sampling

Simple random sampling and systematic sampling, the two types of sampling discussed up to this point, each involve taking a sample from the population as a whole; neither requires identification of sub-domains or subgroups before the sample is taken. Sometimes, however, the sampling frame can be partitioned into groups or strata, and the sampling can be performed separately within each stratum. The resulting sampling design is called *stratified sampling*. If simple random sampling is used to select the sample within each of the strata, the sample design is called *stratified random sampling*.

WHAT IS A STRATIFIED RANDOM SAMPLE?

A stratified random sample, as indicated above, is a sampling plan in which a population is divided into L **mutually exclusive and exhaustive** strata, and a simple random sample of n_h elements is taken within each stratum h. The sampling is performed independently within each stratum. In essence, we can think of a stratified random sampling scheme as being L separate simple random samples.

Operationally, a stratified random sample is taken in the same way as a simple random sample, but the sampling is done separately and independently within each stratum. If we let $N_1, N_2, N_3, \ldots, N_L$ represent the number of sampling units within each stratum, and $n_1, n_2, n_3, \ldots, n_L$ represent the number of randomly selected sampling units within each stratum, then the total number of possible stratified random samples is equal to

$$\left(\begin{array}{c} N_1 \\ n_1 \end{array} \right) \times \left(\begin{array}{c} N_1 \\ n_1 \end{array} \right) \cdots \times \left(\begin{array}{c} N_L \\ n_L \end{array} \right)$$

which is less than or equal to $\left(\begin{array}{c} N \\ n \end{array} \right)$, the total number of possible simple random samples.

WHY STRATIFIED SAMPLING?

Stratified sampling is used in certain types of surveys because it combines the conceptual simplicity of simple random sampling with potentially significant gains in reliability. It is a convenient technique to use whenever we wish to obtain separate estimates for population parameters for each sub-domain within an overall population and, in addition, wish to ensure that our sample is representative of the population.

Examples: Hospital beds

We wish to estimate the total number of beds in the hospitals of a certain state. We know that the majority of the hospitals are *small* or *middle sized* and that there are only a few very *large* hospitals. We also know that these very large hospitals account for a substantial portion of the total number of beds. Now suppose we decide to select a simple random sample of the hospitals in the state, determine the number of beds in each one so selected and, using the methods of Chapter 3, estimate the total number of beds among all hospitals in the entire state. The problem with this procedure is that there is a good chance that our sample may contain either too many or too few of the very large hospitals.

As a result, the sample may not adequately represent the population. Our solution to this problem is to stratify the sampling units (hospitals) prior to sampling, into three groups on the basis of size (i.e., small, middle, large), and then select, using simple random sampling techniques, certain numbers of hospitals from each of the three groups. An estimate of the total number of beds can then be obtained from the combined results of the three strata. This is the essence of stratified sampling.

Notation for stratification

Population quantities

Note a slight change in notation, but it's easier follow

- $X_{h,j}$-value of the jth unit in stratum h.
- $T_h = \sum_{j=1}^{N_h} X_{h,j}$-population total in stratum h.
- $T = \sum_{h=1}^{L} T_h$-population total.
- $\overline{X}_h = \frac{\sum_{j=1}^{N_h} X_{h,j}}{N_h}$-population mean in stratum h.
- $\overline{X} = \frac{T}{N} = \frac{\sum_{h=1}^{L}\sum_{j=1}^{N_h} X_{h,j}}{N}$ -overall population mean.

Corresponding quantities for the sample

Using SRS estimates within each stratum

Example

Stratified random sampling to estimate the size of the Nelchina herd of Alaskan caribou in February 1962. In January and early February, several sampling techniques were field-tested. The field tests told the investigators that several of the proposed sampling units, such as equal-flying-time sampling units, were difficult to implement in practice and that an *equal-area sampling unit of 4 square miles* (mi^2) would work well for the survey. The biologists used preliminary estimates of caribou densities to divide the area of interest into *six* strata; each stratum was then divided into a grid of 4-mi^2 *sampling units*. Stratum A, for example, contained $N_1 = 400$ sampling units; $n_1 = 98$ *sampling units* of these were randomly selected to be in the survey. The following data were reported:

Stratum	N_h	n_h	$\overline{x}_h$	s_h^2
A	400	98	24.1	5,575
B	30	10	25.6	4,064
C	61	37	267.6	347,556
D	18	6	179.0	22,798
E	70	39	293.7	123,578
F	120	21	33.2	9,795

Stratum	N_h	n_h	$\overline{x}_h$	s_h^2	$\hat{t} = N_h \overline{x}_h$	$\left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}$
A	400	98	24.1	5,575	9,640	6,872,040.82
B	30	10	25.6	4,064	768	243,840.00
C	61	37	267.6	347,556	16,324	13,751,945.51
D	18	6	179.0	22,798	3,222	820,728.00
E	70	39	293.7	123,578	20,559	6,876,006.67
F	120	21	33.2	9,795	3,984	5,541,171.43
total		211			54,497	34,105,732.43
sqrt(total)						5,840.01

With the data in this form, using a spreadsheet to do the calculations necessary for stratified sampling is easy. The spreadsheet shown in the Table simplifies the calculations that the estimated total number of caribou is 54,497 with standard error 5840. An approximate 95% CI for the total number of caribou is

$$54,497 \quad \pm \quad 1.96(5840) \quad = \quad [43,051, 65,943].$$

Illustrative Example.

A university group is interested in determining the necessity for child care for it's employees. The university group decides to conduct a sample survey to estimate the average amount of child care per week (in hours) that employees with small children would require. Suppose the university was surrounded by three towns (Town A, Town B and Town C) from which it's employees live. Time travel to these three towns differ, and each may currently offer different types of day care facilities.

Town A has a predominantly blue collar populous with many small rural households with small children. Town B is a more upscale community with more middle class households and fewer children at home, and Town C is mainly a farming community. A stratified random sample with three strata appears to be a more appropriate sample survey design because of the administrative conveniences and the similarities in child care needs, we would expect small variability within each stratum.

Stratum 1 *Town A*				*Stratum 2* *Town B*				*Stratum 3* *Town C*			
35	28	26	41	27	4	49	10	8	15	21	7
43	29	32	37	15	41	25	30	14	30	20	11
36	25	29	31					12	32	34	24
39	38	40	45								
28	27	35	34								

Stratum 1	*Stratum 2*	*Stratum 3*
$n_1 = 20$	$n_2 = 8$	$n_3 = 12$
$\overline{y}_1 = 33.9$	$\overline{y}_2 = 25.125$	$\overline{y}_3 = 19$
$s_1^2 = 35.358$	$s_2^2 = 232.411$	$s_3^2 = 87.636$
$N_1 = 155$	$N_2 = 62$	$N_3 = 93$

Illustrative Example.

Levy and Lemeshow

Suppose that a road having a length of 24 miles traverses areas that can be classified as urban and rural and that the road is divided into eight segments, each having a length equal to 3 miles. A sample of three segments is taken, and on each segment sampled, special equipment is installed for purposes of counting the number of total motor vehicle miles traveled by cars and trucks on the segment during a particular year. In addition, a record of all accidents occurring on each sample segment is kept. The number of truck miles and the number of accidents in which a truck was involved during a certain period are given in Table 5.1 for each of the eight segments in the population. Suppose that we take a simple random sample of three segments for purposes of estimating the total number of truck miles traveled on the road. There are 56 possible samples of three segments from the population of eight segments. The sampling distribution of the number of truck miles traveled on the road is given in Table 5.2.

Estimates of the total number of truck miles obtained in this way range from 15,026.67 to 60,938.67 with the mean of the sampling distribution of *population total* equal to 34,054, the total T, and the standard error of T (*total*) equal to 10,536.9. Now suppose that instead of taking a simple random sample of three segments from the population of eight segments, we first group the segments into two strata, one consisting of urban segments, the other consisting of rural segments, as shown in Table 5.3.

Table: 5.1 Truck Miles and Number of Accidents Involving Trucks by Type of Road Segment

Segment	Type	Number of Truck Miles ($\times$ 100)	Number of Accidents Involving Trucks
1	Urban	6327	8
2	Rural	2555	5
3	Urban	8691	9
4	Urban	7834	9
5	Rural	1586	5
6	Rural	2034	1
7	Rural	2015	9
8	Rural	3012	4

Table: 5.2 Sampling Distribution of *Sum total* for 56 Possible Samples of Three Segments

Segments in Sample	t	Segments in Sample	t	Segments in Sample	t	Segments in Sample	t
(1,2,3)	46,861.33	(1,4,8)	45,794.67	(2,4,7)	33,077.33	(3,5,8)	35,437.33
(1,2,4)	44,576.00	(1,5,6)	26,525.33	(2,4,8)	35,736.00	(3,6,7)	33,973.33
(1,2,5)	27,914.67	(1,5,7)	26,474.67	(2,5,6)	16,466.67	(3,6,8)	36,632.00
(1,2,6)	29,109.33	(1,5,8)	29,133.33	(2,5,7)	16,416.00	(3,7,8)	36,581.33
(1,2,7)	29,058.67	(1,6,7)	27,669.33	(2,5,8)	19,074.67	(4,5,6)	30,544.00
(1,2,8)	31,717.33	(1,6,8)	30,328.00	(2,6,7)	17,610.67	(4,5,7)	30,493.33
(1,3,4)	60,938.67	(1,7,8)	30,277.33	(2,6,8)	20,269.33	(4,5,8)	33,152.00
(1,3,5)	44,277.33	(2,3,4)	50,880.00	(2,7,8)	20,218.67	(4,6,7)	31,688.00
(1,3,6)	45,472.00	(2,3,5)	34,218.67	(3,4,5)	48,296.00	(4,6,8)	34,346.67
(1,3,7)	45,421.33	(2,3,6)	35,413.33	(3,4,6)	49,490.67	(4,7,8)	34,296.00
(1,3,8)	48,080.00	(2,3,7)	35,362.67	(3,4,7)	49,440.00	(5,6,7)	15,026.67
(1,4,5)	41,992.00	(2,3,8)	38,021.33	(3,4,8)	52,098.67	(5,6,8)	17,685.33
(1,4,6)	43,186.67	(2,4,5)	31,933.33	(3,5,6)	32,829.33	(5,7,8)	17,634.67
(1,4,7)	43,136.00	(2,4,6)	33,128.00	(3,5,7)	32,778.67	(6,7,8)	18,829.33

Table: 5.3 Two Strata for Data of Table 5.1

Stratum 1 (Urban Segments)		*Stratum 2 (Rural Segments)*	
Segment	Truck Miles $\times$ 1000	Segment	Truck Miles $\times$ 1000
1	6327	2	2555
3	8691	5	1586
4	7834	6	2034
		7	2015
		8	3012

We might now take a sample of one segment from stratum 1 and two segments from stratum 2 and estimate the total number of truck miles by the estimate t_{str} given by

$$t_{str} = t_{str_1} + t_{str_2}$$

where t_{str_1} = estimated number of truck miles in the three segments composing stratum 1, and
t_{str_2} = estimated number of truck miles in the five segments composing stratum 2

Table: 5.4 Sampling Distribution of T_{str} for 30 Possible Samples of Three Segments-Urban Segment 1

Stratum 1	Stratum 2	t_{str_1} $(= 3\overline{x}_1)$	t_{str_2} $(= 5\overline{x}_2)$	t_{str} $= t_{str_1} + t_{str_2}$
1	(2,5)	18,981	10,352.50	29,333.50
1	(2,6)	18,981	11,472.50	30,453.50
1	(2,7)	18,981	11,425.00	30,406.00
1	(2,8)	18,981	13,917.50	32,898.50
1	(5,6)	18,981	9,050.00	28,031.00
1	(5,7)	18,981	9,002.50	27,983.50
1	(5,8)	18,981	11,495.00	30,476.00
1	(6,7)	18,981	10,122.50	29,103.50
1	(6,8)	18,981	12,615.00	31,596.00
1	(7,8)	18,981	12,567.50	31,548.50

Table: 5.4 Sampling Distribution of T_{str} for 30 Possible Samples of Three Segments-Urban Segment 2

Stratum 1	Stratum 2	t_{str_1} $(= 3\overline{x}_1)$	t_{str_2} $(= 5\overline{x}_2)$	t_{str} $= t_{str_1} + t_{str_2}$
3	(2,5)	26,073	10,352.50	36,425.50
3	(2,6)	26,073	11,472.50	37,545.50
3	(2,7)	26,073	11,425.00	37,498.00
3	(2,8)	26,073	13,917.50	39,990.50
3	(5,6)	26,073	9,050.00	35,123.00
3	(5,7)	26,073	9,002.50	35,075.50
3	(5,8)	26,073	11,495.00	37,568.00
3	(6,7)	26,073	10,122.50	36,195.50
3	(6,8)	26,073	12,615.00	38,688.00
3	(7,8)	26,073	12,567.50	38,640.50

Table: 5.4 Sampling Distribution of T_{str} for 30 Possible Samples of Three Segments-Urban Segment 3

Stratum 1	Stratum 2	t_{str_1} $(= 3\overline{x}_1)$	t_{str_2} $(= 5\overline{x}_2)$	t_{str} $= t_{str_1} + t_{str_2}$
4	(2,5)	23,502	10,352.50	33,854.50
4	(2,6)	23,502	11,472.50	34,974.50
4	(2,7)	23,502	11,425.00	34,927.00
4	(2,8)	23,502	13,917.50	37,419.50
4	(5,6)	23,502	9,050.00	32,552.00
4	(5,7)	23,502	9,002.50	32,504.50
4	(5,8)	23,502	11,495.00	34,997.00
4	(6,7)	23,502	10,122.50	33,624.50
4	(6,8)	23,502	12,615.00	36,117.00
4	(7,8)	23,502	12,567.50	36,069.50

Table: 5.5 Comparison of Results for Simple Random Sampling and Stratification

	Sampling Design	
	Simple Random Sampling	Stratification
Number of elements in sample	3	3*
Number of possible samples	56	30
Mean of distribution of estimated totals	34,054†	34,054†
Standard error of estimated total	10,536.9	3,297.6
Range of distributions of estimated totals	45,912	12,007

*One element from stratum 1, two from stratum 2; †This is also the population total.

Advantages of Stratification-Ad Nauseum

There are three major advantages of stratification over simple random sampling.

1. Given certain conditions, precision may be increased over simple random sampling (i.e., lower standard errors may result from the estimation procedure).
2. It is possible to obtain estimates for each of the strata that have specified precision.
3. It may be just as easy, for either political or administrative reasons, to collect information for a stratified sample as is possible for a simple random sample. If such is the case, there is little to lose by taking a stratified sample, since the resulting standard errors will rarely exceed those of simple random sampling.

Proportions

To make inferences about proportions we have

$$\widehat{p}_{str} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{p}_h$$

and

$$\widehat{V}\left(\widehat{p}_{str}\right) = \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\widehat{p}_h \times (1 - \widehat{p}_h)}{n_h - 1}$$

Total number

Estimating the total number of population units have a this characteristic is similar:

$$\widehat{t}_{str} = \sum_{i=1}^{L} N_h \widehat{p}_h.$$

and

$$\widehat{V}\left(\widehat{t}_{str}\right) = N^2 \widehat{V}\left(\widehat{p}_{str}\right).$$

Example of Stratification when a proportion is estimated

The American Council of Learned Societies (ACLS) used a stratified random sample of selected ACLS societies in seven disciplines to study publication patterns and computer and library use among scholars who belong to one of the member organizations of the ACLS. The data are shown in following table. Ignoring the non-response for now and supposing there are no duplicate memberships, let's use the stratified sample to estimate the percentage and number of respondents of the major societies in those seven disciplines who are women.

Here, let N_h be the membership figures, and n_h be the valid returns.

Table: Data from ACLS Survey

Discipline	**Membership** N_h	**Number Mailed**	**Valid Returns,** n_h	**Female Members (%)**
Literature	9,100	915	636	38
Classics	1,950	633	451	27
Philosophy	5,500	658	481	18
History	10,850	855	611	19
Linguistics	2,100	667	493	36
Poli. Sci.	5,500	833	575	13
Sociology	9,000	824	588	26
Totals	44,000	5,385	3,835	

Then,

$$\begin{aligned}
\widehat{p}_{str} &= \sum_{h=1}^{7}\left(\frac{N_h}{N}\right)\widehat{p}_h = \left(\frac{9,100}{44,000}\right)(0.38) + \left(\frac{1950}{44,000}\right)(0.27) \\
&\quad + \left(\frac{5,500}{44,000}\right)(0.18) + \left(\frac{10,850}{44,000}\right)(0.19) \\
&\quad + \left(\frac{2,100}{44,000}\right)(0.36) + \left(\frac{5,500}{44,000}\right)(0.13) \\
&\quad + \left(\frac{9,000}{44,000}\right)(0.26) \\
&= 0.2465
\end{aligned}$$

and

$$\begin{aligned}
SE\left(\widehat{p}_{str}\right) &= \sqrt{\sum_{h=1}^{7}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\widehat{p}_h \times (1-\widehat{p}_h)}{n_h - 1}} \\
&= 0.0071.
\end{aligned}$$

And the estimated total number of female members in the societies is

$$\begin{aligned} \widehat{t}_{str} &= \sum_{i=1}^{7} N_h \widehat{p}_h \\ &= 44,000 \times 0.2465. \end{aligned}$$

and

$$\begin{aligned} SE\left(\widehat{t}_{str}\right) &= \sqrt{N^2 \widehat{V}\left(\widehat{p}_{str}\right)} \\ &= 44,000 \times 0.0071 \;=\; 312. \end{aligned}$$

Let us consider a population containing N elementary units that are grouped exclusively and exhaustively into L strata in such a way that stratum 1 contains N_1 elementary units, stratum 2 contains N_2 elementary units, . . . and stratum L contains N_L elementary units.

$$N = \sum_{h=1}^{L} N_h$$

Suppose we are considering a variable or characteristic $\mathfrak{X}$ in the population. This could be the total, the mean, the proportion, or any other population characteristic. Then $X_{h,j}$ would represent the value of the characteristic $\mathfrak{X}$ for the jth elementary unit within stratum h.

Illustrative Example.

Suppose we are interested in estimating the average daily pharmaceutical cost per patient at a hospital. We decide to stratify the hospital into services (medical, surgical, ob-gyn, all other services combined), and we define the elementary units as patients on any given day. Suppose that on a designated day, there are 250 patients in the hospital, of which 100 are medical, 75 surgical, 50 ob-gyn, and 25 other services. Then using the notation introduced above we have

$$N = 250 \quad N_1 = 100 \quad N_2 = 75 \quad N_3 = 50 \quad N_4 = 25$$

If $\mathfrak{X}$ designates the value of the variable 1 for the jth elementary unit within stratum h, then, for example, in stratum 2 we have

$X_{2,1}$= the value of variable $\mathfrak{X}$ for element 1 within stratum 2

$X_{2,2}$= the value of variable $\mathfrak{X}$ for element 2 within stratum 2

and so on, up to

$X_{2,75}$= the value of variable $\mathfrak{X}$ for element 75 within stratum 2

The *total* or aggregate amount of a variable $\mathfrak{X}$ *within stratum* h is defined by T_h as given by

$$T_h = \sum_{h=1}^{N_h} X_{h,j}$$

The *total for the whole population* is given by the sum of the stratum totals, or

$$\begin{aligned} T &= \sum_{h=1}^{L} \sum_{j=1}^{N_h} X_{h,j} \\ &= \sum_{h=1}^{L} T_h \end{aligned}$$

The *mean level of a characteristic* $\mathfrak{X}$ for a stratum h is denoted by $\overline{X}_h$, and is given by

$$\begin{aligned} \overline{X}_h &= \frac{\sum_{h=1}^{N_h} X_{h,j}}{N_h} \\ &= \frac{T_h}{N_h} \end{aligned}$$

The mean $\overline{X}$ of a variable $\mathfrak{X}$ for the entire population is given by

$$\begin{aligned} \overline{X} &= \frac{\sum_{h=1}^{L} N_h \overline{X}_h}{N} \\ &= \sum_{h=1}^{L} W_h \overline{X}_h \end{aligned}$$

where

$$W_h = \frac{N_h}{N}$$

The *variance* σ^2_{hx} of the distribution of a variable $\mathfrak{X}$ within a particular stratum h is defined as the average squared deviation about the stratum mean and is given by

$$\sigma^2_{hx} = \frac{\sum_{j=1}^{N_h} \left(X_{h,j} - \overline{X}_h\right)^2}{N_h}$$

are defined for each stratum in the same way as they were defined for a population that is not grouped into strata. The coefficient of variation for the distribution within a particular stratum is given by

$$V_{hx} = \frac{\sigma_{hx}}{\overline{X}_h}$$

Illustrative Example.

Consider a population of 14 families living on three city blocks. If we consider the families as elementary units, the blocks as strata, and family size as the characteristic $\mathfrak{X}$, we might have the situation shown in the following table

Table: 5.6 Strata for a Population of 14 Families

Block	Family	Family Size
1	1	4
	2	3
	3	4
2	1	4
	2	6
	3	4
	4	7
	5	8
3	1	2
	2	3
	3	2
	4	2
	5	2
	6	3

Estimates of Population Parameters

Stratum Specific

$$(\textbf{Total}):\ t_h = N_h \overline{x}_h$$

$$(\textbf{Mean}):\ \overline{x}_h = \frac{\sum_{j=1}^{n_h} x_{h,j}}{n_h}$$

$$(\textbf{Proportion}):\ p_{hy} = \frac{\sum_{j=1}^{n_h} y_{h,j}}{n_h}$$

Estimates of Population Parameters-Entire Population

$$(\textbf{Total}):\ t_{str} = \sum_{h=1}^{L} t_h$$

$$(\textbf{Mean}):\ \overline{x}_{str} = \frac{\sum_{h=1}^{L} N_h \overline{x}_h}{N}$$

$$(\textbf{Proportion}):\ p_{y,str} = \frac{\sum_{h=1}^{L} y_{h+}}{N}$$

Standard Error Estimates

$$\widehat{SE}\left(t_{str}\right) = \sqrt{\sum_{h=1}^{L} \frac{N_h^2 s_{hx}^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}$$

$$\widehat{SE}\left(\overline{x}_{str}\right) = \sqrt{\sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \frac{s_{hx}^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}$$

$$\widehat{SE}\, p_{y,str} = \sqrt{\sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \frac{p_{hy}(1 - p_{hy})}{n_h - 1}\left(\frac{N_h - n_h}{N_h}\right)}$$