Sampling Theory and Methods

Spring 2008

C. L. Williams

Chapter 6 Stratified Sampling

Outline

1. Stratified Sampling

Chapter 6-Stratified Random Sampling

In the previous chapter we introduced the concept of stratification and discussed the reasons why stratification is used as a strategy in designing sample surveys. We also introduced notation commonly used by statisticians in discussing population characteristics and estimation procedures appropriate for stratified sampling. In this chapter we will discuss one type of stratified sampling in considerable detail, namely stratified random sampling.

Illustrative Example.

Consider the population of 14 families shown in Table 5.6. Suppose we decide to take a sample of **two** families from stratum 1, **two** families from stratum 2, and **four** families from stratum 3. Then we have $n_1 = 2$, $n_2 = 2$, $n_3 = 4$, $N_1, = 3$, $N_2 = 5$, $N_3 = 6$, and $\mathrm{N} = 14$. Suppose we select elements $X_{1,2}$ and $X_{1,3}$ from stratum 1 (i.e., $x_{1,1} = 3$ and $x_{1,2} = 4$),$X_{2,2}$ and $X_{2,5}$ from stratum 2 (i.e., $x_{2,1} = 6$ and $x_{2,2} = 8$), and we select $X_{3,1}$, $X_{3,2}$, $X_{3,5}$, and $X_{3,6}$ from stratum 3 (i.e., $x_{3,1} = 2$, $x_{3,2} = 3$, $x_{3,3} = 2$, and $x_{3,4} = 3$).

Estimates of the total number of individuals in each block

$$\begin{aligned} t_1 &= \frac{3 \times (3+4)}{2} = 10.5 \\ t_2 &= \frac{5 \times (6+8)}{2} = 35 \\ t_3 &= \frac{6 \times (2+3+2+3)}{4} = 15 \\ \text{so that} \quad t_{str} &= t_1 + t_2 + t_3 \\ &= 60.5 \end{aligned}$$

Estimates of the mean number of individuals in each block

$$
\begin{aligned}
\overline{x}_1 &= \frac{(3+4)}{2} = 3.5 \\
\overline{x}_2 &= \frac{(6+8)}{2} = 7 \\
\overline{x}_3 &= \frac{(2+3+2+3)}{4} = 2.5 \\
\text{so that} \quad \overline{x}_{str} &= \frac{3\times 3.5}{14} + \frac{5\times 7}{14} + \frac{6\times 2.5}{14} \\
&= 4.32
\end{aligned}
$$

Estimates of the variances for family size in each block

$$\begin{aligned}
s_{1,x}^2 &= \frac{(3-3.5)^2+(4-3.5)^2}{1} = 0.5\\
s_{2,x}^2 &= \frac{(6-7)^2+(8-7)^2}{1} = 2\\
\text{and}\quad s_{3,x}^2 &= \frac{(2-2.5)^2+(3-2.5)^2(2-2.5)^2+(3-2.5)^2}{3} = 0.33
\end{aligned}$$

For proportions, if we let

$$y_{h,i} = \begin{cases} 1 & \text{if family size is 4 or more} \\ 0 & \text{if family size is less than 4} \end{cases}$$

$$\begin{aligned} p_{1,y} &= \frac{(0+1)}{2} = 0.5 \\ p_{2,y} &= \frac{(1+1)}{2} = 1.0 \\ p_{3,y} &= \frac{(0+0+0+0)}{4} = 0.0 \end{aligned}$$

Since a stratified random sample consists of L simple random samples, which are drawn separately and independently within each stratum, and since the estimated population mean, total, or proportion is a linear combination of the estimated individual stratum means, totals, or proportions obtained from the sample, it follows that the mean of the sampling distribution of any of these estimated values is equal to the corresponding linear combination of population parameters. In other words, population totals, means, and proportions, when estimated as indicated in relations (5.6), (5.7), and (5.8), are, under stratified random sampling, **unbiased estimates** of the corresponding population means, totals, and proportions.

$$E\left(t_{str}\right) = \sum_{h=1}^{L} E\left(t_h\right) = \sum_{h=1}^{L} X_{h+} = X(\text{population total})$$

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

$$E\left(\overline{x}_{str}\right) = \frac{\sum_{h=1}^{L}(N_h)E(\overline{x}_h)}{N} = \frac{\sum_{h=1}^{L}(N_h)\overline{X}_h}{N} = \overline{X}(\text{population mean})$$

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

$$\begin{aligned} E\left(p_{y,str}\right) &= \frac{\sum_{h=1}^{L} N_h P_{hy}}{N} = \frac{\sum_{h=1}^{L} Y_{h+}}{N} = P_y(\text{population proportion}) \\ SE(p_{y,str}) &= \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{P_{hy}(1-P_{hy})}{n_h}\left(\frac{N_h-n_h}{N_h-1}\right)} \end{aligned}$$

Illustrative Example.

In the example used earlier, strata are three city blocks (Table 5.6), the elementary units are families, and the variable is family size. We took a stratified random sample of **two** families from stratum 1, **two** from stratum 2, and **four** from stratum 3. Thus, we have $n_1 = 2,\ n_2 = 1,\ n_3 = 4,\ N_1 = 3,\ N_2 = 5,\ N_3 = 6.$

$$\begin{aligned}
\sigma^2_{1,x} &= \frac{(4-3.67)^2 + (3-3.67)^2 + (4-3.67)^2}{3} = 0.222 \\
\sigma^2_{2,x} &= \frac{2 \times (4-5.8)^2 + (6-5.8)^2 + (7-5.8)^2 + (8-5.8)^2}{5} = 2.56 \\
\sigma^2_{3,x} &= \frac{(2-2.33)^2 + (3-2.33)^2 + \cdots + (3-2.33)^2}{6} = 0.222
\end{aligned}$$

$P_1 = 0.67,\ P_2 = 1.00$, and $P_3 = 0$, where P, is the proportion of families in the i^{th} stratum with four or more persons.

$$
\begin{aligned}
SE(\overline{x}_{str}) &= \sqrt{\sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{hx}^2}{n_h} \left(\frac{N_h - n_h}{N_h - 1}\right)} \\
&= \left[\left(\frac{3}{14}\right)^2 \times \left(\frac{0.222}{2}\right) \times \left(\frac{3-2}{3-1}\right) + \right. \\
&\quad \left(\frac{5}{14}\right)^2 \times \left(\frac{2.56}{2}\right) \times \left(\frac{5-2}{5-1}\right) + \\
&\quad \left. \left(\frac{6}{14}\right)^2 \times \left(\frac{0.222}{4}\right) \times \left(\frac{6-4}{6-1}\right)\right]^{1/2} \\
&= 0.359
\end{aligned}
$$

$$
\begin{aligned}
SE(p_{y,str}) &= \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{P_{hy}(1-P_{hy})}{n_h}\left(\frac{N_h-n_h}{N_h-1}\right)} \\
&\quad \left[\left(\frac{3}{14}\right)^2 \times \left(\frac{(0.67)(0.33)}{2}\right) \times \left(\frac{3-2}{3-1}\right) + \right. \\
&\quad \left(\frac{5}{14}\right)^2 \times \left(\frac{(1)(0)}{2}\right) \times \left(\frac{5-2}{5-1}\right) + \\
&\quad \left.\left(\frac{6}{14}\right)^2 \times \left(\frac{(0)(1)}{4}\right) \times \left(\frac{6-4}{6-1}\right)\right]^{1/2} \\
&= 0.0504
\end{aligned}
$$

$$\begin{aligned}
\sigma^2_{hx} &= \frac{(N_h - 1)s^2_{hx}}{N_h} \\
s^2_{hx} &= \frac{\displaystyle\sum_{i=1}^{n_h} (x_{h,i} - \overline{x}_h)^2}{n_h - 1}
\end{aligned}$$

Illustrative Example

Recall

$$\begin{array}{rclrclrclrcl}
\overline{x}_1 &=& 3.5 & t_1 &=& 10.5 & s^2_{1,x} &=& 0.5 & \widehat{\sigma}^2_{1,x} &=& 0.33 \\
\overline{x}_2 &=& 7 & t_2 &=& 35 & s^2_{2,x} &=& 2 & \widehat{\sigma}^2_{2,x} &=& 1.6 \\
\overline{x}_3 &=& 2.5 & t_3 &=& 15 & s^2_{3,x} &=& 0.33 & \widehat{\sigma}^2_{3,x} &=& 0.275
\end{array}$$

$$t_{str} = 10.5 + 35 + 15 = 60.5$$

$$\begin{aligned}
\widehat{SE}(t_{str}) &= \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{\sigma^2_{hx}}{n_h}\left(\frac{N_h-n_h}{N_h-1}\right)} \\
& \left[(3)^2 \times \left(\frac{.5}{2}\right) \times \left(\frac{3-2}{3}\right) + \right. \\
& (5)^2 \times \left(\frac{2}{2}\right) \times \left(\frac{5-2}{5}\right) + \\
& \left. (6)^2 \times \left(\frac{0.33}{4}\right) \times \left(\frac{6-4}{6}\right)\right]^{1/2} \\
&= 4.09
\end{aligned}$$

so that

$$\begin{aligned}
t_{str} - 1.96 \times SE(t_{str}) \leq & \quad X \quad \leq t_{str} + 1.96 \times SE(t_{str}) \\
60.5 - 1.96 \times 4.09 \leq & \quad X \quad \leq 60.5 + 1.96 \times 4.09 \\
52.48 \leq & \quad X \quad \leq 68.52
\end{aligned}$$

In Chapter 3, we showed that under simple random sampling, estimated means, totals, and proportions for subgroups are unbiased estimates of the corresponding population means, totals, and proportions for the subgroups. This is not necessarily true in stratified random sampling, as is shown in the next example.

Illustrative Example.

Let us consider the data given in Table 6.1. If we let $\overline{X}_I$ denote the average price among the five **independent,I** pharmacies in the combined two communities, we see that $\overline{X}_I = 11.60$. Suppose we take a stratified random sample of six pharmacies from stratum I and three pharmacies from stratum 2 for purposes of estimating $\overline{X}_I$. Suppose also that we do not know before the sampling whether a given pharmacy is an independent or an affiliate of a chain. Our estimate $\overline{x}_{I,str}$ of $\overline{X}_I$ is given by

$$\overline{x}_{I,str} = \frac{\sum_{h=1}^{2} N_h \overline{x}_{I,str}}{N}$$

where $\overline{x}_{I,str}$, is the estimated mean for the independent pharmacies obtained from the sample taken in stratum h. There are seven possible samples of six pharmacies that can be taken in stratum 1, and there are four possible samples of three pharmacies that can be taken in stratum 2. These samples and the estimated mean for each sample are listed in Table 6.2.

Table: 6.1 Retail Prices of 20 Capsules of a Tranquilizer in All Pharmacies in Two Communities (Strata)

Community	Pharmacy	Type*	*Price of Drug ($)*
1	1	C	10.00
	2	I	9.00
	3	I	12.00
	4	I	11.00
	5	C	9.00
	6	C	9.50
	7	C	9.90
2	1	I	13.50
	2	I	12.50
	3	C	12.00
	4	C	11.00

*I = independent; C = chain.

Table: 6.2 Possible Samples for the Stratified Random Sample

Stratum 1		*Stratum 2*	
Pharmacies in Sample	$\overline{x}_{I,1}$ (\$)	Pharmacies in Sample	$\overline{x}_{I,2}$ (\$)
1,2,3,4,5,6	10.67	1,2,3	13.00
1,2,3,4,5,7	10.67	1,2,4	13.00
1,2,3,4,6,7	10.67	1,3,4	13.50
1,2,3,5,6,7	10.50	2,3,4	12.50
1,2,4,5,6,7	10.00		
1,3,4,5,6,7	11.50		
2,3,4,5,6,7	10.67		

There are $\begin{pmatrix} 7 \\ 6 \end{pmatrix} \times \begin{pmatrix} 4 \\ 3 \end{pmatrix} = 28$ possible values of $\overline{x}_{I,str} = \frac{(7\overline{x}_{I1}+4\overline{x}_{I1})}{11}$. The sampling distribution of $\overline{x}_{I,str}$ is shown in Table 6.3. The mean E($\overline{x}_{I,str}$) of the distribution of $\overline{x}_{I,str}$ over the 28 samples is equal to \$11.52, which is not equal to \$11.60, the value of $\overline{X}_I$, the mean price over the five independent pharmacies in the two communities.

Table: Sampling Distribution of $\overline{x}_{I,str}$

$\overline{x}_{I,str}$		$\overline{x}_{I,str}$	
($)	f	($)	f
11.510	8	12.046	2
11.692	4	11.590	1
11.330	4	11.228	1
11.410	2	11.272	1
11.090	2	12.228	1
10.910	1	11.864	1
Total			28

Once we decide to use stratified sampling, and once we specify the strata and the total number, n, of sample elements, the next important decision we must make is that of allocation or specification of how many elements are to be taken from each stratum under the constraint that a total of n elements is to be taken over all strata. As we will see in this section, the standard errors of the estimated population parameters may be reduced considerably if careful thought is given to allocation.

In equal allocation, the same number of elements are sampled from each stratum. In other words, for each stratum, h, the sample size is given by

$$n_h = \frac{n}{L}$$

Equal allocation would be the allocation of choice if the primary objective of the sample survey is to test hypotheses about differences among the strata with respect to levels of variables of interest, under the assumption that within stratum variances were equal. If this assumption could not be made, then the allocation of choice for testing such hypotheses would be given by

$$n_h = \frac{\sigma_{hx}}{\sum_{h=1}^{L} \sigma_{hx}} \times n$$

Self-Weighting Samples

In proportional allocation, the sampling fraction n_h/N_h, is specified to be the same for each stratum, which implies also that the overall sampling fraction n/N is the fraction taken from each stratum. In other words, the number of elements n_1, taken from each stratum is given by

$$n_h = N_h \times \frac{n}{N}$$

$$\overline{x}_{str} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{h,i}}{n}$$

Illustrative Example.

Number of general hospitals located in four strata, where a stratum is composed of one or more geographical regions within Illinois. We wish to take a stratified random sample of 51 hospitals from among the 255 hospitals, and we wish to use proportional allocation. Then letting $N = 255$ and $n = 51$, we have from relation (6.9),

$$\begin{aligned}
n_1 &= (44)\left(\frac{51}{255}\right) = 8.8 \approx 9 \\
n_2 &= (116)\left(\frac{51}{255}\right) = 23.2 \approx 23 \\
n_3 &= (48)\left(\frac{51}{255}\right) = 9.6 \approx 10 \\
n_4 &= (47)\left(\frac{51}{255}\right) = 9.4 \approx 9
\end{aligned}$$

Table 6.4

Table: General Hospitals in Illinois by Geographical Stratum,1971

Stratum	No. of General Hospitals
1	44
2	116
3	48
4	47
Total	255

So take 9 elements (hospitals) from stratum 1; 23 from stratum 2; 10 from stratum 3; and 9 from stratum 4.

The sampling fractions within each stratum are

$$\begin{aligned}
\frac{n_1}{N_1} &= \frac{9}{44} = 0.2045 \\
\frac{n_2}{N_2} &= \frac{23}{116} = 0.1983 \\
\frac{n_3}{N_3} &= \frac{10}{48} = 0.2083 \\
\frac{n_4}{N_4} &= \frac{9}{47} = 0.1915
\end{aligned}$$

and as it should be $\displaystyle\sum_{h=1}^{4} n_h = n$.

The slight differences in sampling fractions among the strata are due to the fact that the required allocation given by relation (6.9) does not necessarily yield integer values. Thus, the n_h's, taken are those specified by Equation (6.9), but rounded up or down to the nearest integer. These minor differences among sampling fractions are generally ignored in constructing the estimates, and the sample is generally treated as if it were exactly a self-weighting sample.

In proportional allocation, the variance, $Var(\overline{x}_{str})$, of an estimated mean, $\overline{x}_{str}$ obtained from relation (6.10) with n_h set equal to $N_h(n/N)$ becomes

$$Var(\overline{x}_{str}) = \left(\frac{N-n}{N^2}\right)\sum_{h=1}^{L}\left(\frac{N_h^2}{N_h-1}\right)\left(\frac{\sigma_{hx}^2}{n}\right)$$

If all the N_h are reasonably large, the expression reduces to the approximation given by

$$Var(\overline{x}_{str}) \approx \left(\frac{N-n}{N}\right)\left(\frac{\sigma^2_{wx}}{n}\right)$$

where

$$\sigma^2_{wx} = \frac{\sum_{h=1}^{L} N_h \sigma^2_{hx}}{N}$$

The *sample estimate* of $Var(\overline{x}_{str})$ is given by

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

Note that relation (6.12) has a form that is very similar to the formula for the variance of an estimated mean under simple random sampling. The formula for the standard error was given in Box 3.1, and the square of this quantity, the variance, is given as

$$Var(\overline{x}) = \left(\frac{N-n}{N-1}\right)\left(\frac{\sigma_x^2}{n}\right)$$

$$Var(\overline{x}_{str}) \approx \left(\frac{N-n}{N}\right)\left(\frac{\sigma_{wx}^2}{n}\right)(6.12)$$

where

$$\sigma_{wx}^2 = \frac{\sum_{h=1}^{L} N_h \sigma_{hx}^2}{N}$$

The difference between the two formulas is that for proportional allocation in stratified random sampling, the population variance σ_x^2 is replaced by σ_{wx}^2, which is a weighted average of the individual variances σ_{hx}^2 of the distribution of among elements within each stratum. The weights in σ_{wx}^2 are proportional to N_h, the number of elements in each stratum.

Comparison of relations (6.12) and (6.14) indicates that stratified random sampling with proportional allocation will yield an estimated mean having lower variance than that obtained from simple random sampling whenever σ_{wx}^2 is less than σ_x^2. But note that, as in analysis of variance methodology, the population variance σ_x^2 may be partitioned into the two components σ_{bx}^2 and σ_{wx}^2

$$\sigma_x^2 = \sigma_{bx}^2 + \sigma_{wx}^2$$

where

$$\sigma_{bx}^2 = \frac{\sum_{h=1}^{L} N_h \left(\overline{X}_h - \overline{X}\right)^2}{N}$$

and σ_{wx}^2 is as given in 6.13.

Thus the ratio of the variance of the estimated mean $\overline{x}$ under simple random sampling to that of $\overline{x}_{str}$, the estimated mean under stratified random sampling with proportional allocation is given by,

$$\frac{Var(\overline{x})}{Var(\overline{x}_{str})} = \frac{\sigma_{bx}^2 + \sigma_{wx}^2}{\sigma_{wx}^2} = 1 + \frac{\sigma_{bx}^2}{\sigma_{wx}^2}$$

This ratio is always greater than or equal to unity, and the extent to which it differs from unity depends on the size of the ratio $\frac{\sigma_{bx}^2}{\sigma_{wx}^2}$. When this ratio is large, the estimated mean under stratified random sampling with proportional allocation will have a smaller variance than the corresponding estimate under simple random sampling. The component σ_{bx}^2 represents the variance *among the stratum means*, whereas the component and σ_{wx}^2 represents the variance among the elements *within the same stratum*. If the stratum means $\overline{X}_h$ are of the same order of magnitude, then little or nothing is gained by using stratified random sampling rather than simple random sampling.

On the other hand, if the stratum means are very different, it is likely that considerable reduction in the variance of an estimated mean can be obtained by use of stratified random sampling rather than simple random sampling. This makes sense intuitively because the purpose of stratification is to group the elements, in advance of the sampling, into strata on the basis of their similarity with respect to the values of a variable or a set of variables. If the elements within each stratum have very similar values of the variable being measured, then it would be difficult to obtain a "bad" sample, since each stratum is represented in the sample. A reliable estimate could then be obtained by sampling a small number of elements within each stratum. On the other hand, if the stratum means are very similar, then there is no point to stratification, and the extra effort required to take a stratified sample would not result in an improved estimate.

Illustrative Example.

Let us suppose that we wish to estimate the average number of hospital admissions for major trauma conditions per county among 82 counties in Illinois having general hospitals. A sample of counties will be taken, and the admission records of all hospitals in the sample counties will be reviewed for major trauma admissions. If it is reasonable to assume that there may be a strong correlation between the number of hospital beds among general hospitals within a county and the total number of admissions for major trauma conditions, then it would make sense to stratify by number of hospital beds. So this is the sampling plan that is chosen.

In Table 6.5, counties in Illinois are grouped into two strata on the basis of number of hospital beds. Stratum 1 consists of those counties having 1-399 beds, and stratum 2 consists of those having 400 beds or more.

From Table 6.5 we calculate the following:Illustrative Example

$$\begin{array}{rclrclrcl}
\overline{X}_1 & = & 123.91 & \widehat{\sigma}^2_{1,x} & = & 6,131,3 & N_1 & = & 65 \\
\overline{X}_2 & = & 871.59 & \widehat{\sigma}^2_{2,x} & = & 77,287.92 & N_2 & = & 17 \\
\overline{X} & = & 278.92 & \widehat{\sigma}^2_{x} & = & 112,751.93 & N & = & 82
\end{array}$$

where $\mathfrak{X}$ = the number of beds.

From relations (6.15) and (6.13) we have

$$\begin{aligned}
\sigma_{bx}^2 &= \frac{65 \times (123.91 - 278.92)^2 + 17 \times (871.59 - 278.92)^2}{82} \\
&= 91,868.39 \\
\sigma_{wx}^2 &= \frac{65 \times (6,131.63) + 17 \times (77,287.92)}{82} \\
&= 20,883.54
\end{aligned}$$

$$\frac{Var(\overline{x})}{Var(\overline{x}_{str})} = 1 + \frac{91,868.39}{20,883.54} = 5.4.$$

Therefore, we conclude that in terms of reduction of the variance of an estimated mean, stratification is likely to be of great benefit in this situation if, in fact, admissions for multiple trauma and number of beds are related, since the variance under stratification is less than 20% of the variance under simple random sampling.

Often proportional allocation is not the type of allocation that would result in an estimated total, mean, or proportion having the lowest variance among all Possible ways of allocating a total sample of n elements among the L strata. It can be shown that the allocation of n sample units into each stratum that will yield an estimated total, mean, or proportion for a variable $\mathfrak{X}$ having minimum variance is given by

$$n_h = \left(\frac{N_h \sigma_{hx}}{\sum_{h=1}^{L} N_h \sigma_{hx}} \right) (n)$$

Illustrative Example

Using the Table 6.5 and assume a close relationship between number of beds and number of admissions for major trauma. From relation (6.17), the following allocation of *25* sample elements will produce the estimated mean having the lowest variance:

$$\begin{aligned} n_1 &= \left(\frac{65\sqrt{6131.63}}{65\sqrt{6131.63}+17\sqrt{77,287.92}}\right)\times(25) \\ &= 12.96 \approx 13 \\ n_2 &= \left(\frac{17\sqrt{77,287.92}}{65\sqrt{6131.63}+17\sqrt{77,287.92}}\right)\times(25) \\ &= 12.04 \approx 12 \end{aligned}$$

Thus, the optimal allocation of the 25 sample elements is 13 elements from stratum 1 and 12 elements from stratum 2. Proportional allocation would have specified that 20 elements be taken from stratum 1 and 5 elements from stratum 2.

The standard error of an estimated mean under stratified random sampling with optimal allocation is given by relation (6.2), which is valid in general for any type of allocation under stratified random sampling. For the data in Table 6.5, using number of beds as the characteristic of interest, we have for optimal allocation from Equation (6.1),

$$\begin{aligned}
\widehat{SE}(\overline{x}_{str}) &= \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{\sigma_{hx}^2}{n_h}\left(\frac{N_h-n_h}{N_h-1}\right)} \\
&= \left[\left(\frac{65}{82}\right)^2 \times \left(\frac{6,131.63}{13}\right) \times \left(\frac{65-13}{65-1}\right) + \right. \\
&\quad \left.\left(\frac{17}{82}\right)^2 \times \left(\frac{77,287.92}{12}\right) \times \left(\frac{17-12}{17-1}\right)\right]^{1/2} \\
&= 18.09
\end{aligned}$$

For proportional allocation we have, taking the square root of the expression in Equation (6.11),

$$
\begin{aligned}
\sqrt{Var(\overline{x}_{str})} &= \sqrt{\left(\frac{N-n}{N^2}\right)\sum_{h=1}^{L}\left(\frac{N_h^2}{N_h-1}\right)\left(\frac{\sigma_{hx}^2}{n}\right)} \\
\sqrt{Var(\overline{x}_{str})} &= \Bigg[\left(\frac{82-65}{82^2}\right)\Bigg\{\left(\frac{65^2}{64}\right)\times\left(\frac{6,131.63}{25}\right)+ \\
&\qquad \left(\frac{17^2}{16}\right)\times\left(\frac{77,287.92}{25}\right)\Bigg\}\Bigg]^{1/2} \\
&= 24.71
\end{aligned}
$$

Thus we see that for these data the estimated mean under optimal allocation has a standard error considerably lower than that under proportional allocation.

Optimal Allocation and Economics

Suppose now that the cost of sampling an elementary unit is not the same for each stratum. Then the *total cost* C, of taking a sample of n_1 elements from stratum 1, n_2 elements from stratum 2, and so forth, is given by

$$C = \sum_{h=1}^{L} n_h C_h$$

where C_h is the cost of sampling an elementary unit in stratum h.

For a given sample size n, the allocation that will yield an estimate having the lowest variance per unit cost is given by

$$n_h = \left(\frac{N_h \sigma_{hx} / \sqrt{C_h}}{\sum_{h=1}^{L} N_h \sigma_{hx} / \sqrt{C_h}} \right) (n)$$

Similarly, if the total cost of taking the sample is fixed at C, the allocation that will yield the estimated mean having the lowest standard error at fixed cost C is given by

$$n_h = \left(\frac{N_h \sigma_{hx} / \sqrt{C_h}}{\sum_{h=1}^{L} N_h \sigma_{hx} \sqrt{C_h}} \right) \times (C)$$

Illustrative Example.

Let us suppose that a corporation has *260,000* accident reports available over a period of time and that a sample survey is being contemplated for purposes of estimating the average number of days of work lost per accident. Of the *260,000* accident reports, *150,000* are coded and *110,000* are uncoded. The coded forms could be processed on the computer directly, whereas the uncoded forms must first be coded before processing. Approximately $*10,000* is available for selecting the sample and coding and processing the data. With this in mind, it is desired to find the best way of allocating the sample elements among coded and uncoded forms. In the terminology of stratified sampling we have

$$\begin{aligned} N_1 &= \text{150,000 coded forms (stratum 1)} \\ N_2 &= \text{110,000 uncoded forms (stratum 2)} \\ C &= \$10,000 \end{aligned}$$

Let us suppose that the cost of sampling and processing sample forms is equal to $C_1 = \$0.32$ for a coded form and $C_2 = \$0.98$ for an uncoded form; that is,

$$\begin{aligned} n_1 &= \frac{\frac{150{,}000 \times \left(\frac{\sigma_{2x}}{2}\right)}{\sqrt{0.32}}}{150,000 \times \left(\frac{\sigma_{2x}}{2}\right) \times \sqrt{0.32} + 110,000 \times \sigma_{2x} \times \sqrt{0.98}} \times 10,000 \\ &\approx 8,762 \end{aligned}$$

$$\begin{aligned} n_2 &= \frac{\frac{110{,}000 \times \left(\frac{\sigma_{2x}}{2}\right)}{\sqrt{0.98}}}{150,000 \times \left(\frac{\sigma_{2x}}{2}\right) \times \sqrt{0.32} + 110,000 \times \sigma_{2x} \times \sqrt{0.98}} \times 10,000 \\ &\approx 7,343. \end{aligned}$$

Thus we would take a sample of 8762 coded reports and 7343 uncoded reports. We can verify that the total cost of the sampling is equal to \$10,000 by substituting the values for C_1, C_2, n_1, and n_2 into the relation for $C = 8762 \times 0.32 + 7343 \times 0.98 = \$10{,}000$.

We note that in order to obtain the optimal allocation, it is not necessary to know the actual values of the σ_{hx}. If we can express each σ_{hx} in terms of one of them (e.g.,σ_{rx} as was done in the example discussed above, then σ_{rx} appears as a common factor in both the numerator and denominator and therefore can be canceled.

One problem often encountered in optimal allocation, either with or without costs being taken into consideration, is that the optimal sample size h may be greater than N_h, the total number of elements in the stratum. When this occurs, we set h equal to N_h for each stratum having optimal allocation greater than N_h. Then we reallocate the remaining sample to other strata as specified by the algorithm of obtaining optimal allocation.

For example, let us consider the summary data from three strata:

Stratum	$\mathbf{N_h}$	σ_{hx}
1	100	50
2	110	10
3	120	5

If we wish to allocate a total sample of 140 elements to each stratum by using optimal allocation, we have, by relation (6.17), $n_1 = 104$, $n_2 = 23$ and $n_3 = 13$.

We would then take $n_1 = N_1 = 100$ and allocate the four remaining elements to strata 2 and 3 according to relation (6.17) as follows:

$$\begin{aligned}
n_2 &= \left[\frac{110 \times 10}{110 \times 10 + 120 \times 5}\right](4) \\
&= 2.6 \approx 3 \\
n_3 &= \left[\frac{1200 \times 5}{110 \times 10 + 120 \times 5}\right](4) \\
&= 1.4 \approx 1
\end{aligned}$$

Thus the final optimum allocation is $n_1 = 100$, $n_2 = 26$ and $n_3 = 14$.

In the planning of a sample survey for which stratified random sampling sampling is indicated, it is often a good strategy to calculate the optimal allocation for the most important variables in the survey. If the allocation differs among the variables, some compromise allocation might be considered (such as the mean of the optimal h over all variables of importance). Also, proportional allocation should be given some consideration. If the standard errors anticipated under proportional allocation are not much higher than those anticipated under optimal allocation, then the simplicity and convenience of proportional allocation may offset the small reduction in standard error under optimal allocation, and proportional allocation may be the best choice.

Illustrative Example: Case Study.

This example from a recent study of elderly twins illustrates the use of optimal allocation in stratified random sampling. The objective of this study was to test a method for identifying elderly twins (65 years and older) from lists of living Medicare beneficiaries. Studies on monozygotic and dizygotic twins are extremely useful in providing insight into the relative contribution of genetic and non-genetic influences on health and disease, and twins so identified would be placed in a registry for possible participation in future medical investigations.

Since approximately 1 in every 100 deliveries results in a multiple birth, any attempt at screening unselected lists of individuals for identification of twins would be prohibitively expensive. The following characteristics of twins, how- ever, might be used to obtain modified lists that have a higher prevalence of twins:

- Both members of a twin pair (with very rare exceptions) are born in the same place and have the same date of birth.
- They are also of the same race.
- Both members of a male-male twin pair will have the same last name and a different first name.
- Members of a twin pair are very likely to have Social Security numbers (SS#) that are very close to each other.

From living male beneficiaries, pairs were constructed consisting of individuals having the same date of birth, the same state of birth, the same last name, and a different first name. The pairs so obtained were then assigned a number representing the difference in their SS#. This difference (called sequence difference) was obtained by a complex algorithm [47]; the size of this sequence difference being proportional to the length of time separating the issuance dates of the SS card to each member of the pair.

From living female beneficiaries, female-female pairs were constructed from records that had the same date and state of birth and the same first seven digits of the SS#. (Surnames could not be used as was done with the males because of name change upon marriage.) Male-female pairs were constructed by the same algorithm used to construct female-female pairs. The data set constructed as described above consisted of 255,848 paired records categorized into six classes according to race (white/African-American) and sex (M-M, M-F, F-F). Each of these six groups was further subdivided into three classes based on the size of the sequence difference in SS# (first quartile/second and third quartiles/fourth quartile). Those pairs having sequence differences in the first quartile represent SS#'s issued relatively close in time, and so on.

Table 6.6 indicates the number of pairs that were obtained in each of the 18 “strata” defined above. A pilot survey of approximately 1000 pairs was to be conducted, having as its objective the estimation in each of the six race-sex groups of the proportion of pairs in this database that are truly twins. This was considered important as a test of whether or not this methodology would produce a database that has a high prevalence of twins. A sample of pairs was to be taken, and each individual sampled was to be queried on his/her twin status. The design of the sample for this pilot survey was to be that of stratified random sampling with optimal allocation applied separately to each of the six race-sex groups.

With this in mind, the formula for allocation of sample into the three SS# sequence difference groups is given by Equation (6.17), where $\sigma_{hx} = \sqrt{p_{hX}(1-p_{hX})}$, and p_{hX} is the proportion of twins in stratum h ($h = 1, 2, 3$). The proportions, p_{hX} are not known, and it is necessary to make some "educated" guesses concerning their values. The overall prevalence rate of twins in the master file of Medicare beneficiaries is likely to be about 1, the rate in the population as a whole. We would then expect that the algorithms used to construct the set of pairs would yield a very much higher prevalence of twins-let us assume a 10-fold higher prevalence-which would be a rate of 10. We further assume that the prevalence of twins in the first sequence difference quartile is four times that in the middle two quartiles and eight times that in the fourth quartile.

$$\frac{\sum_{h=1}^{3} N_h p_{hx}}{\sum_{h=1}^{3} N_h}$$

where N_h is the total number of pairs in stratum h.
We can set $p_{1x} = 8p_{3x}$ and $p_{2x} = 2p_{3x}$ so that

$$\frac{8 \times p_{3x} \times 39,872 + 2 \times p_{3x} \times 79,727 + p_{3x} \times 39,872}{39,872 + 79,727 + 39,872} = 0.10$$

or

$$p_{3x} = 0.0308$$

It then follows $p_{2x} = 0.0615$ and $p_{1x} = 0.2461$.

$$\begin{aligned}
\sigma_{3x} &= [(0.0308)(1-0.0308)]^{0.5} = 0.1728 \\
\sigma_{2x} &= [(0.0615)(1-0.0615)]^{0.5} = 0.2402 \\
\sigma_{1x} &= [(0.2461)(1-0.2461)]^{0.5} = 0.4307
\end{aligned}$$

$\text{n} = \frac{1000}{6} = 167.$

$$\begin{aligned} n_1 &= \frac{39,872 \times 0.4307}{39,872 \times 0.4307 + 79,727 \times 0.2402 + 39,872 \times 0.1728} \times 167 \\ &\approx 66 \\ n_2 &= \frac{79,727 \times 0.2402}{39,872 \times 0.4307 + 79,727 \times 0.2402 + 39,872 \times 0.1728} \times 167 \\ &\approx 74 \\ n_3 &= \frac{39,872 \times 0.1728}{39,872 \times 0.4307 + 79,727 \times 0.2402 + 39,872 \times 0.1728} \times 167 \\ &\approx 27 \end{aligned}$$

STRATIFICATION AFTER SAMPLING

A sample design in which the sampling plan is that of simple random sampling but the estimation procedure is that appropriate for stratified random sampling can sometimes produce estimates having standard errors that are not much higher than those obtained by stratified random sampling. The advantage of this design is that it eliminates the inconvenience, or impossibility, of grouping the elements into strata in advance of the sampling. It is known as *stratification after sampling or poststratification*.

For example, it may be of interest to estimate the proportion of premature births in a given hospital during the past year. It is known from past experience that the prematurity rate among blacks is higher than the corresponding rate for whites. However, to stratify the entire set of hospital records by racial group would be impractical, since racial group is recorded in the records and all records would have to be inspected to do such stratification prior to the sampling. However, if the total number of blacks and the total number of whites who have entered the hospital during the year for deliveries is known (as it may well be by the hospital administration), a simple random sample may be stratified after the sampling to improve the precision of the estimate.

Let $\overline{x}_{pstr}$ and Var($\overline{x}_{pstr}$) represent the *poststratification sample mean* sample mean and variance of its sampling distribution respectively. Then

$$\overline{x}_{pstr} = \sum_{h=1}^{L} \left(\frac{N_h}{N} \right) \overline{x}_h$$

and

$$Var\left(\overline{x}_{pstr}\right) = \left(\frac{N-n}{nN} \right) \sum_{h=1}^{L} \frac{N_h}{N} S_{hx}^2 + \left(\frac{1}{n^2} \right) \sum_{h=1}^{L} \frac{N_h}{N} S_{hx}^2 \left(\frac{N-n_h}{N} \right)$$

where

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

The first term in Var($\overline{x}_{pstr}$) is approximately the variance of an estimated mean under stratified sampling with proportional allocation. The second term increases the variance and reflects the fact that the n1, in the resulting sample are random variables. The second term will generally be small when the sample size n is large. Although S^2_{hx} is not known, it can be estimated by s^2_{hx}. [relation (5.9)] and the *sample estimate* of Var($\overline{x}_{pstr}$) is given by

$$Var\left(\overline{x}_{pstr}\right) = \left(\frac{N-n}{nN}\right)\sum_{h=1}^{L}\frac{N_h}{N}s^2_{hx} + \left(\frac{1}{n^2}\right)\sum_{h=1}^{L}s^2_{hx}\left(\frac{N-n_h}{N}\right)$$

Expressions similar to relations (6.20) and (6.21) can be derived for the variance of estimated poststratified totals and proportions, as well as the estimated variances from the sample information. Now let us look at an example of how poststratification can be useful in reducing sampling error.

Illustrative Example.

A veterinarian is interested in studying the annual veterinary costs of his clientele (who own either dogs or cats). From a separate record system, he knows that he sees 850 dogs and 450 cats regularly in his practice (these are numbers of animals, not numbers of visits). He knows that the information on type of animal (i.e., dog or cat) is contained in the medical records, but that it would take too much time to sort the records into strata defined by animal type. So he decides to select a simple random sample and then poststratify. He regards the poststratification process as necessary since he knows that, on average, it costs more to keep dogs healthy than to keep cats healthy.

He samples 50 records, recording the total amount of money spent (including medication) by the owners of the animals he saw over the past two years. The sampling results are given in Table 6.8. Now suppose that this sample of 50 animals is to be used to estimate the average annual expense of owning a dog or a cat. Then we have the following calculations (refer to Boxes 2.2 and 3.1):

$$\begin{aligned}\overline{x} &= \frac{\sum_{i=1}^{50} x_i}{50} \\ &= \frac{45.14 + 50.13 + \cdots + 39.26}{50} \\ &= 39.73\end{aligned}$$

$$\begin{aligned}s_x^2 &= \frac{\sum_{i=1}^{50} (x_i - \overline{x})^2}{(50-1)} \\ &= \frac{(45.14 - 39.73)^2 + (50.13 - 39.73)^2 + \cdots + (39.26 - 39.73)^2}{49} \\ &= 256.68\end{aligned}$$

$$\widehat{SE}\left(\overline{x}\right) = \left[\left(\frac{1300-50}{1300}\right)\left(\frac{256.68}{50}\right)\right]^{1/2}$$
$$\sqrt{4.936} = 2.2222$$

so that

$$\begin{aligned}
\overline{x} - 1.96 \times SE(\overline{x}) \leq \quad & \overline{X} \quad \leq \overline{x} + 1.96 \times SE(\overline{x}) \\
39.73 - 1.96 \times 2.222 \leq \quad & \overline{X} \quad \leq 39.73 + 1.96 \times 2.222 \\
35.38 \leq \quad & \overline{X} \quad \leq 44.08
\end{aligned}$$

Now we can use the known stratum totals in a poststratification process to obtain a more precise estimate of $\overline{X}$. The veterinarian knows that the number of dogs in his files is $N_1 = 850$ and the total number of cats is $N_2 = 450$. Stratifying the *50* animals in the sample given in Table 6.8 into dogs and cats yields $n_1 = 32$ dogs and $n_2 = 18$ cats in the sample. Then we have (refer to Box 2.2):

$$\begin{aligned}\overline{x}_1 &= \frac{45.14 + 50.13 + \cdots + 39.26}{32} \\ &= 49.86\end{aligned}$$

$$\begin{aligned}\overline{x}_2 &= \frac{27.15 + 23.39 + \cdots + 14.18}{28} \\ &= 21.71\end{aligned}$$

$$\begin{aligned}s_{x_1}^2 &= \frac{\displaystyle\sum_{i=1}^{50} (x_i - \overline{x})^2}{(50-1)} \\ &= \frac{(45.14 - 49.86)^2 + (50.13 - 49.86)^2 + \cdots + (39.26 - 49.86)^2}{31} \\ &= 70.22\end{aligned}$$

$$
\begin{aligned}
s_{x_2}^2 &= \frac{\sum_{i=1}^{50} (x_i - \overline{x})^2}{(50-1)} \\
&= \frac{27.15 - 21.71)^2 + (23.39 - 21.71)^2 + \cdots + (14.18 - 21.71)^2}{17} \\
&= 75.00
\end{aligned}
$$

$$
\begin{aligned}
\overline{x}_{pstr} &= \left(\frac{850}{1300}\right) \times 49.86 + \left(\frac{450}{1300}\right) \times 21.71 \;=\; 40.12 \\
\widehat{VAR}\left(\overline{x}_{pstr}\right) &= \left(\frac{1300-50}{50 \times 1300}\right) \left[\left(\frac{850}{1300}\right) \times 70.22 + \left(\frac{450}{1300}\right) \times 75\right] \\
&+ \left(\frac{1}{50^2}\right) \left[\left(\frac{1300-850}{1300}\right) \times 70.22 + \left(\frac{1300-450}{1300}\right) \times 7 \right. \\
&= 1.439 \\
\widehat{SE}\left(\overline{x}_{pstr}\right) &= \sqrt{1.439} \;=\; 1.20
\end{aligned}
$$

Hence the 95% confidence interval estimate of the population mean, $\overline{X}$, is given, by poststratification, as

$$
\begin{aligned}
\overline{x}_{pstr} - 1.96 \times SE(\overline{x}_{pstr}) \leq \; & \overline{X} \; \leq \overline{x}_{pstr} + 1.96 \times SE(\overline{x}_{pstr}) \\
40.12 - 1.96 \times 1.20 \leq \; & \overline{X} \; \leq 40.12 + 1.96 \times 1.20 \\
37.77 \leq \; & \overline{X} \; \leq 42.47
\end{aligned}
$$

HOW LARGE A SAMPLE IS NEEDED?

Suppose that we wish to determine the number of elements needed to be 100 $\times$ (1 - α)% certain of obtaining from a stratified random sampling, an estimated mean, $\overline{x}_{str}$ that differs from the true mean $\overline{X}$ by no more than 100 $\times \epsilon$. This formulation is equivalent to that discussed earlier for simple random sampling and systematic sampling. The formula (valid for reasonably large N_h,) for the required n is as follows:

$$n = \frac{\left(\frac{Z^2_{\alpha/2}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h^2\sigma_{hx}^2}{\pi_h\overline{X}^2}\right)}{\epsilon^2 + \left(\frac{Z^2_{\alpha/2}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h^2\sigma_{hx}^2}{\overline{X}^2}\right)}$$

where

$$\pi_h = \frac{n_h}{n}$$

Relation (6.22) is valid for any type of allocation. It is also valid for the estimation of a population total. The analogous sample size formula for estimation of a population proportion P_y, from stratified random sampling is given by

$$n = \frac{\left(\frac{Z^2_{\alpha/2}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h^2 P_{hy} \times (1 - P_{hy})}{\pi_h P_y^2}\right)}{\epsilon^2 + \left(\frac{Z^2_{\alpha/2}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h^2 P_{hy} \times (1 - P_{hy})}{P_y^2}\right)}$$

We can see from relation (6.22) that its use requires more knowledge about parameters of the distribution than is likely to be available or than can be guessed with any degree of confidence. For this reason, relation (6.22) is unlikely to be of much help in actual practice. However, if one assumes proportional allocation, then relation (6.22) reduces to the form

$$n \;=\; \frac{N Z^2_{(\alpha/2)} \frac{\sigma^2_{wx}}{\overline{X}^2)}}{N\epsilon^2 + Z^2_{(\alpha/2)} \frac{\sigma^2_{wx}}{\overline{X}^2)}}$$

We can see from relation (6.22) that its use requires more knowledge about parameters of the distribution than is likely to be available or than can be guessed with any degree of confidence. For this reason, relation (6.22) is unlikely to be of much help in actual practice. However, if one assumes proportional allocation, then relation (6.22) reduces to the form

$$n \approx \frac{Z^2_{(\alpha/2)} \times \frac{N}{1+\gamma} \times V^2_x}{N\epsilon^2 + Z^2_{(\alpha/2)} \times \frac{1}{1+\gamma} \times V^2_x}$$

Illustrative Example.

Suppose that we are planning to take a sample of the members of a health maintenance organization (HMO) for purposes of estimating the average number of hospital episodes per person. The sample will be selected from membership lists grouped according to age (under 45 years; 45-64 years; 65 years and over). Let us suppose that the distributions of hospital episodes are available from national data (such as the National Health Interview Survey) and are as given in Table 6.9. Suppose further that the number of HMO members in each age group is as follows:

$$\begin{aligned} &\text{Age group 1:} & N_1 &= 600 \\ &\text{Age group 2:} & N_2 &= 500 \\ &\text{Age group 3:} & N_3 &= 400 \end{aligned}$$

Table 6.9

Table: Distribution of Hospital Episodes per Person per Year

	Age Group	Average Number of Hospital Episodes	Variance of Distribution of Hospital Episodes
1.	Under 45 years	0.164	0.245
2.	45-64 years	0.166	0.296
3.	65 years and over	0.236	0.436

$$\overline{X} = \frac{600 \times 0.164 + 500 \times 0.166 + 400.236}{1500} = 0.184$$

$$\begin{aligned}\sigma_{bx}^2 &= \frac{600 \times (0.164 - 0.184)^2 + 500 \times (0.166 - 0.184)^2 + 400 \times (0.236 - 0.}{1500} \\ &= 0.0009891\end{aligned}$$

$$\begin{aligned}\sigma_{wx}^2 &= \frac{600 \times 0.245 + 500 \times 0.296 + 400 \times 0.436}{1500} \\ &= 0.31293\end{aligned}$$

$$
\begin{aligned}
\sigma_x^2 &= 0.0009891 + 0.31293 = 0.31392 \\
V_x^2 &= \frac{0.31392}{0.184^2} = 9.27 \\
\gamma &= \frac{0.0009891}{0.31293} = 0.00316
\end{aligned}
$$

$$n = \frac{[(9 \times 1500)/(1_{+}0.00316)] \times 9.27}{[(9 \times 9.27)/(1 + 0.00316)] + 1500 \times (0.20)^2} = 872$$

$$\begin{aligned} n_1 &= 600 \times \frac{872}{1500} = 349 \\ n_2 &= 500 \times \frac{872}{1500} = 291 \\ n_3 &= 400 \times \frac{872}{1500} = 232 \end{aligned}$$

$$
\begin{aligned}
\pi_1 &= \frac{600\sqrt{0.245}}{600\sqrt{0.245}+500\sqrt{0.296}+400\sqrt{0.436}} = 0.356 \\
\pi_2 &= \frac{500\sqrt{0.296}}{600\sqrt{0.245}+500\sqrt{0.296}+400\sqrt{0.436}} = 0.327 \\
\pi_3 &= \frac{400\sqrt{0.436}}{600\sqrt{0.245}+500\sqrt{0.296}+400\sqrt{0.436}} = 0.317
\end{aligned}
$$

$$\begin{aligned} n &\approx \left[\frac{9}{(1500)^2}\right] \frac{\left[\frac{(600)^2(0.245)}{(0.356)(0.184)^2} + \frac{(500)^2(0.296)}{(0.327)(0.184)^2} + \frac{(400)^2(0.436)}{(0.317)(0.184)^2}\right]}{(0.2)^2 + \left[\frac{9}{(1500)^2}\right]\left[\frac{(600)(0.245)}{(0.184)^2} + \frac{(500)(0.296)}{(0.184)^2} + \frac{(400)(0.43}{(0.184)^2}\right.} \\ &= 860 \end{aligned}$$

$$\begin{aligned} n_1 &= n \times \pi_1 = 860 \times 0.356 = 306 \\ n_2 &= n \times \pi_2 = 860 \times 0.356 = 281 \\ n_3 &= n \times \pi_3 = 860 \times 0.356 = 273 \end{aligned}$$

$$n \approx \frac{\left(\frac{Z^2_{\alpha/2}}{N}\right)\left(\sum_{h=1}^{L}\frac{N_h P_{hy} \times (1 - P_{hy})}{\pi_h P_y^2}\right)}{\epsilon^2 + \left(\frac{Z^2_{\alpha/2}}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h P_{hy} \times (1 - P_{hy})}{P_y^2}\right)}$$