# 11 Sample Surveys II: Stratification

## 11.1 Introduction

*Stratified Random Sampling* is characterized by the following:
• the population is divided into a number of parts called *strata*, say into $K$ strata;
• a simple random sample is drawn **independently** from each part;
• unbiased estimates of the *population mean* are obtained using an appropriately weighted sum of the sample means from each *stratum*.

Stratification is employed for a number of reasons:
i) differences between the *population strata means* do not make a contribution to the sampling error of the estimate $\overline{y}_{st}$. Sampling error arises solely from variations among the sampling units that are in the same stratum. Thus, if we form strata in an appropriate way, then we can expect a gain in precision over simple random sampling.

ii) we can choose the size of the sample taken from each stratum - sampling fractions need not be the same in each stratum. This gives scope to do an efficient job allocating resources to the sampling within strata.

## 11.2 Notation

Let $Y_{ij}$ be the value of the variable $Y$ for the $j$-th member of the $i$-th stratum in the population.

Let $y_{ij}$ be the value of the variable $y$ for the $j$-th member taken from the $i$-th stratum from within the sample.

Also, let $N_i$ be the total number of sampling units in the $i$-th stratum, in which case, the total population size is given by $N = \sum_{i=1}^{K} N_i$.

The following table summarizes some of the other quantities of interest:

| | **Population** | **Sample** |
|---|---|---|
| Size | $N$ | $n$ |
| Strata sizes | $N_i$ | $n_i$ |
| Strata weights | $W_i = N_i/N$ | $w_i = n_i/n$ |
| Total | $T = \sum_{i=1}^{K} \sum_{j=1}^{N_i} Y_{ij}$ | $t = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}$ |
| Mean | $\overline{Y} = T/N$ | $\overline{y} = t/n$ |
| Strata Totals | $T_i = \sum_{j=1}^{N_i} Y_{ij}$ | $t_i = \sum_{j=1}^{n_i} y_{ij}$ |
| Strata Means | $\overline{Y}_i = T_i/N_i$ | $\overline{y}_i = t_i/n_i$ |
| Variances | $S^2 = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y})^2}{N-1}$ | $s^2 = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2}{n-1}$ |
| Strata Variances | $S_i^2 = \frac{\sum_{j=1}^{N_i} (Y_{ij} - \overline{Y}_i)^2}{N_i - 1}$ | $s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2}{n_i - 1}$ |

Also,
define $f_i = n_i/N_i$ to be the *sampling fraction for the $i$-th stratum*, and,
$f = n/N$ to be the *overall sampling fraction.*

## 11.3 Design Efficiency of Stratified Sampling

Define

$$\overline{y}_{st} = \sum_{i=1}^{K} W_i \overline{y}_i$$

to be the *stratified* (simple) *sample mean.*
By exploiting results derived under *simple random sampling*, and noting the independence of the samples drawn from different strata, we deduce that

$$E[\overline{y}_{st}] = \sum_{i=1}^{K} W_i E[\overline{y}_i] = \sum_{i=1}^{K} W_i \overline{Y}_i = \frac{1}{N} \sum_{i=1}^{K} N_i \overline{Y}_i = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} Y_{ij} = \overline{Y}. \tag{1}$$

and

$$\text{var}(\overline{y}_{st}) = \sum_{i=1}^{K} W_i^2 \text{var}(\overline{y}_i) = \sum_{i=1}^{K} W_i^2 \frac{S_i^2}{n_i}(1 - f_i) \tag{2}$$

by recalling that, under *simple random sampling* (SRS),

$$\text{var}(\overline{y}) = \frac{S^2}{n}(1 - f).$$

Compare this with the *overall sample average*:

$$\overline{y}' = \sum_{i=1}^{K} w_i \overline{y}_i = \frac{1}{n} \sum_{i=1}^{K} n_i \overline{y}_i$$

In general, this will differ from $\overline{y}_{st}$ unless $n_i/N_i = n/N$, i.e. $n_i \propto N_i$ for $i = 1, \ldots, K$: the sample sizes are proportional to the stratum sizes, which is *proportional allocation* (see later); such an allocation also ensures that $\overline{y}'$ is unbiased for $\overline{Y}$.

The *design efficiency* is determined by comparing $\text{var}(\overline{y}_{st})$ with $\text{var}(\overline{y})$. Indeed, define

$$D_{eff} = \text{var}(\overline{y}_{st})/\text{var}(\overline{y}).$$

If the within-stratum sampling fractions are equal to each other, then the gain in efficiency depends on how small the $\{S_i^2\}$ are in comparison to $S^2$.
Indeed, since, in this case (see later),

$$\text{var}(\overline{y}_{st}) = \frac{1-f}{n} \sum_{i=1}^{K} \frac{N_i}{N} S_i^2$$

then

$$\text{var}(\overline{y}) - \text{var}(\overline{y}_{st}) = \frac{(1-f)}{n} \left\{ S^2 - \frac{1}{N} \sum_{i=1}^{K} N_i S_i^2 \right\}.$$

However,

$$S^2 = \frac{1}{N-1} \left\{ \sum_{i=1}^{K} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y})^2 \right\} = \frac{1}{N-1} \left\{ \sum_{i=1}^{K} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y}_i + \overline{Y}_i - \overline{Y})^2 \right\}$$
$$= \sum_{i=1}^{K} \left( \frac{N_i - 1}{N-1} \right) S_i^2 + \sum_{i=1}^{K} \frac{N_i}{N-1} (\overline{Y}_i - \overline{Y})^2.$$

If the $\{N_i\}$ are sufficiently large, then $\frac{N_i - 1}{N-1} \approx \frac{N_i}{N-1}$, and so

$$S^2 \approx \frac{1}{N} \left\{ \sum_{i=1}^{K} N_i S_i^2 + \sum_{i=1}^{K} N_i (\overline{Y}_i - \overline{Y})^2 \right\}.$$

Thus

$$\text{var}(\overline{y}) - \text{var}(\overline{y}_{st}) \approx \frac{(1-f)}{nN} \sum_{i=1}^{K} N_i (\overline{Y}_i - \overline{Y})^2 = \frac{(1-f)}{n} \sum_{i=1}^{K} W_i (\overline{Y}_i - \overline{Y})^2$$

which is strictly positive unless the $\{\overline{Y}_i\}$ are all equal to each other. Thus (under the proviso that the $\{N_i\}$ are sufficiently large), the *stratified sample mean* is usually more efficient than the *simple random mean*. A more precise analysis can be found in the next section.

## 11.4 Choice of Strata fractions

There are a number of possible strategies available for selecting the strata sampling fractions. We outline a few of these below, along with their relative efficiencies.

### 11.4.1 Equal Allocation

Set $n_i = n/K$, for $i = 1, \ldots, K$, i.e. equal 'case loads': this particular mechanism may be considered to be administratively convenient.

### 11.4.2 Proportional Allocation

Set $n_i$ proportional to $N_i$, i.e. $n_i \propto N_i$, which implies that $f_i = \frac{n_i}{N_i} = \frac{n}{N} = f$, and so $n_i = n \times W_i$, $i = 1, \ldots, K$.

Denote the corresponding estimate of $\overline{Y}$ by $\overline{y}_{st;p}$. Then

$$\begin{aligned}\text{var}(\overline{y}_{st,p}) &= \sum_{i=1}^{K} W_i^2 \frac{S_i^2}{n_i}(1 - f_i) \\ &= \frac{(1-f)}{n} \sum_{i=1}^{K} W_i S_i^2 = \frac{1-f}{n} S_w^2\end{aligned}$$

where $S_w^2$ is the *average within-strata variance.*

Then

$$\begin{aligned}D_{eff}(\overline{y}_{st;p}) &= \text{var}(\overline{y}_{st;p})/\text{var}(\overline{y}) \\ &= \frac{S_w^2}{n}(1-f)/\frac{S^2}{n}(1-f) = S_w^2/S^2.\end{aligned}$$

This shows that efficiency is achieved upon the basis of *homogeneity* within strata.

### 11.4.3 Optimum Allocation based on Cost

Problem: Minimize the variance of the estimate subject to a cost constraint.

Let $c_0$ be some fixed cost, and define $c_i$ to be the cost per sampled unit in stratum $i$. Then the total cost satisfies

$$c_0 + \sum_{i=1}^{K} n_i c_i \leq C \tag{3}$$

where $C$ is the available budget.

Thus we seek to minimize

$$\text{var}(\overline{y}_{st}) = \sum_{i=1}^{K} \frac{W_i^2 S_i^2 (1 - f_i)}{n_i}$$

subject to the constraint (3).

To solve this problem, form the *Lagrangian*:

$$L(n_1, \ldots, n_K) = \text{var}(\overline{y}_{st}) + \lambda(c_0 + \sum_{i=1}^{K} n_i c_i - C).$$

Now

$$\frac{\partial L}{\partial n_j} = -\frac{W_j^2 S_j^2}{n_j^2} + \lambda c_j$$

which, upon equating to zero, yields

$$\lambda = \frac{W_j^2 S_j^2 / c_j}{n_j^2}.$$

This implies that

$$n_j \propto \frac{W_j S_j}{\sqrt{c_j}}.$$

Actually, if the costs across all the strata were equal, then the optimal allocation would be such that $n_j \propto W_j S_j$: we denote the resulting estimate by $\overline{y}_{st;o}$. It can be shown that

$$\text{var}(\overline{y}_{st;o}) = \frac{1}{n} \left( \sum_{i=1}^{K} W_i S_i \right)^2 - \frac{1}{N} \sum_{i=1}^{K} W_i S_i^2.$$

### 11.4.4 Disproportionate sampling

Disproportionate sampling is particularly worthwhile when the variable being measured has a highly skewed or asymmetrical distribution. It is often the case that such populations contain a few sampling units that have large values for this variable and many units with small values. Examples include: family income, house prices, cases tried by courts, sales turnover. We explore this allocation rule via the following example.

**Example 11.1 (An application of 'Disproportionate Sampling')**
Consider a population of 1019 tertiary colleges and universities in the U.S.A. Data on student enrollment for a certain academic year were as follows, stratified by size of institution.

| **Stratum** $i$ | **No. of students per institution** | **No. of institutions** $N_i$ | **Stratum Total** $T_i$ | **Stratum Mean** $\overline{Y}_i$ | **Std. Dev.** $S_i$ |
|---|---|---|---|---|---|
| 1 | <1000 | 661 | 292671 | 443 | 236 |
| 2 | 1000-3000 | 205 | 345302 | 1684 | 625 |
| 3 | 3000-10000 | 122 | 672728 | 5514 | 2008 |
| 4 | ≥10000 | 31 | 573693 | 18506 | 10023 |
| | | $N = 1019$ | $T = 1884394$ | | |

These data might be used to plan a sample designed to give a quick estimate of total registration in some future year.

Assume that the sampling costs per unit are the same in all strata and that we are aiming at a total sample size of $n = \sum_{i=1}^{4} n_i = 250$. Then, we could attempt to invoke the optimal allocation rule, i.e. that $n_j \propto N_j S_j$. Under this rule, $n_4$ would be

$$250 \times \frac{31 \times 10023}{(661 \times 236) + (205 \times 625) + (122 \times 2008) + (31 \times 10023)} \approx 92.$$

However, $N_4 = 31$: thus our 'disproportionate' allocation procedure says that we take a 100% sample from stratum 4 and then use the rule $n_i \propto N_i S_i$ to distribute the remaining sample over the other strata.

| **Stratum** $i$ | $n_i$ | **Sampling rate** $f_i$ |
|---|---|---|
| 1 | 65 | 10% |
| 2 | 53 | 26% |
| 3 | 101 | 83% |
| 4 | 31 | 100% |
| Total | 250 | |

## 11.5 Estimation of Totals

A number of unbiased estimators for $T$ can be formulated: these differ in their variability according to the sampling scheme from which they are generated.

Under *simple random sampling*, we use $\widehat{T} = N\overline{y}$, and so

$$E[\widehat{T}] = N \times E[\overline{y}] = N\overline{Y} = T$$
$$\text{var}(\widehat{T}) = N^2(1-f)S^2/n.$$

Under *stratified random sampling* with *proportional allocation*, we use

$$\widehat{T}_{st;p} = N\overline{y}_{st;p}.$$

Now

$$E[\widehat{T}_{st;p}] = E[N\overline{y}_{st;p}] = N \times E[\overline{y}_{st;p}] = N\overline{Y} = T$$

using (1), which is still applicable under *proportional allocation.*
Also

$$\text{var}(\widehat{T}_{st;p}) = \text{var}(N\overline{y}_{st;p}) = N^2 \times \text{var}(\overline{y}_{st;p}) = N^2 \times \frac{(1-f)}{n}\sum_{i=1}^{K} W_i S_i^2 = N^2 \frac{1-f}{n} S_w^2.$$

Finally, consider *stratified random sampling* with *optimal allocation* using equal costs for all sampling units. Then defining

$$\widehat{T}_{st;o} = N\overline{y}_{st;o}$$

it follows that

$$E[\widehat{T}_{st;o}] = E[N\overline{y}_{st;o}] = N \times E[\overline{y}_{st;o}] = N\overline{Y} = T$$

and

$$\text{var}(\widehat{T}_{st;o}) = N^2 \times \text{var}(\overline{y}_{st;o}) = N^2 \left\{ \frac{1}{n}\left(\sum_{i=1}^{K} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{K} W_i S_i^2 \right\}.$$

## 11.6 Stratified Sampling for Attributes

Here it is assumed that $X_{ij} = 1$ if the $j$-th member of the $i$-th stratum belongs to a certain category, and $X_{ij} = 0$ otherwise. Thus the proportion of the population that belong to this category is $\overline{X} = \sum_{i=1}^{K} \sum_{j=1}^{N_i} X_{ij}/N$; let us denote this by $\pi$. Similarly, the proportion of the $i$-th stratum population that belong to this category is given by $\overline{X}_i = \sum_{j=1}^{N_i} X_{ij}/N_i$, which we denote by $\pi_i$.

Let $x_{ij}$ be equal to 1 if the $j$-th member of the $i$-th stratum within the sample possesses the attribute, and is equal to 0 otherwise.

Let $\overline{x}_i$ be the proportion of the $n_i$ members of the sample from the $i$-th stratum that belong to the category, i.e. $\sum_{j=1}^{n_i} x_{ij}/n_i$. The quantity $\pi$ can be estimated by

$$\widehat{\pi}_{st} = \sum_{i=1}^{K} W_i \overline{x}_i$$

This is unbiased since

$$\begin{aligned} E[\widehat{\pi}_{st}] &= \sum_{i=1}^{K} W_i E[\overline{x}_i] = \sum_{i=1}^{K} W_i \overline{X}_i \\ &= \frac{1}{N} \sum_{i=1}^{K} N_i \overline{X}_i = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} X_{ij} = \overline{X} = \pi. \end{aligned}$$

Also, using the fact[1] that $S_i^2 = \frac{N_i \pi_i (1-\pi_i)}{N_i - 1}$ , then

$$\mathrm{var}(\widehat{\pi}_{st}) = \sum_{i=1}^{K} W_i^2 \mathrm{var}(\overline{x}_i) = \sum_{i=1}^{K} W_i^2 \frac{\pi_i(1-\pi_i)}{n_i} \frac{1-f_i}{1-\frac{1}{N_i}} \approx \sum_{i=1}^{K} W_i^2 \frac{\pi_i(1-\pi_i)}{n_i}(1-f_i)$$

where the final expression holds if the $\{N_i\}$ are sufficiently large.

## 11.7 Choosing a Sample Size

We assume that sample size is determined in the context of a reasonable sampling scheme (optimality criteria for experiments or optimal sampling schemes for surveys). Criteria are usually of two types - statistical and/or financial.

### 11.7.1 Statistical Criteria

Classical textbook approach is to set up criteria independent of any financial constraint and determine the sample size necessary. Typical criteria are *confidence intervals* and *standard errors*.

[1] Under this regime, $S^2 = \frac{N\pi(1-\pi)}{N-1}$: see top of p.12 of lecture 10.

(a) <u>C.I. criterion</u>
Suppose that we have a finite population of size $N$ with standard deviation $S = \sqrt{S^2}$. We could estimate $\overline{Y}$ using $\overline{y}$ based on a sample of size $n$. Along with the assumption of normality, we have that

$$\overline{y} \sim N\left(\overline{Y}, \frac{S^2}{n}\left\{1 - \frac{n}{N}\right\}\right)$$

approximately, in which case the width of the 'approximate' $100(1-\alpha)\%$ C.I. is

$$2 \times z_{\alpha/2}\frac{S}{\sqrt{n}}\left(1 - \frac{n}{N}\right)^{1/2}.$$

For a 95% C.I., this is approximately equal to

$$4 \times \frac{S}{\sqrt{n}}\left(1 - \frac{n}{N}\right)^{1/2}$$

since $z_{\alpha/2} = 1.96$.

To ensure that the width of the interval is equal to $d$, say, we solve

$$d \approx 4 \times \frac{S}{\sqrt{n}}\left(1 - \frac{n}{N}\right)^{1/2}$$

which implies that

$$n \approx \frac{16S^2}{d^2 + \frac{16S^2}{N}} \approx \frac{16S^2}{d^2}$$

for large $N$.

(b) <u>S.E.Criterion</u>
Suppose that we require that the standard error of the estimate of the mean is no greater than $e$. Then

$$\frac{S}{\sqrt{n}}\left(1 - \frac{n}{N}\right)^{1/2} = S\left(\frac{1}{n} - \frac{1}{N}\right)^{1/2} = e$$

i.e.

$$\frac{S^2}{n} = e^2 + \frac{S^2}{N}$$

which implies that

$$n = \frac{S^2}{e^2 + \frac{S^2}{N}} \approx \frac{S^2}{e^2}$$

for large (stratum) sample sizes $N$.

### 11.7.2 Cost Criteria

(a) Costing the Survey or Study
The costs may be detailed under headings such as set-up and initial design, pilot survey/experiment, analysis of pilot and refinement of design, interview training, data collection, data entry, data screening, analysis, report presentation.

Some costs are fixed, others monotonic in the sample size (not necessarily linear). Overall cost generally of the form:

$$\text{Cost} = \text{Constant} + \sum_{i=1}^{K} g_i(\text{stratum sample size})$$

where the function $g_i(\cdot)$ is often of a form involving either steps or slopes. For example, the steps may correspond to the need for another interviewer.

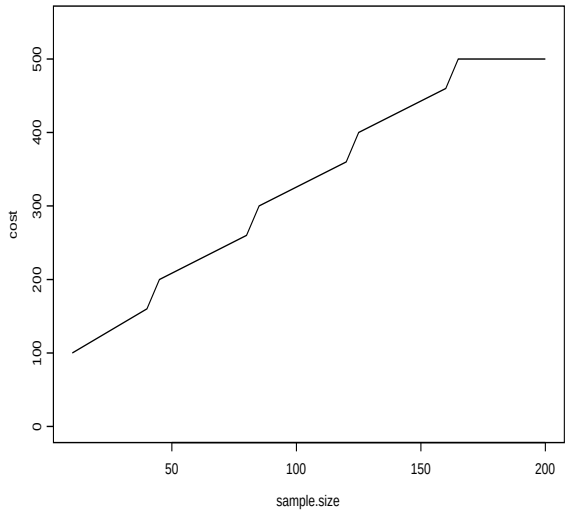


Figure 1: cost vs. sample size

(b) Evaluating a Return
This is more difficult as we require the 'value' of determining a particular *population parameter* to a desired degree of accuracy or the 'value' of detecting a particular *change in magnitude* of a particular process.

Might start with a statistical criterion curve, standard error say, plotted against sample size. This needs to be converted to a monetary value curve.

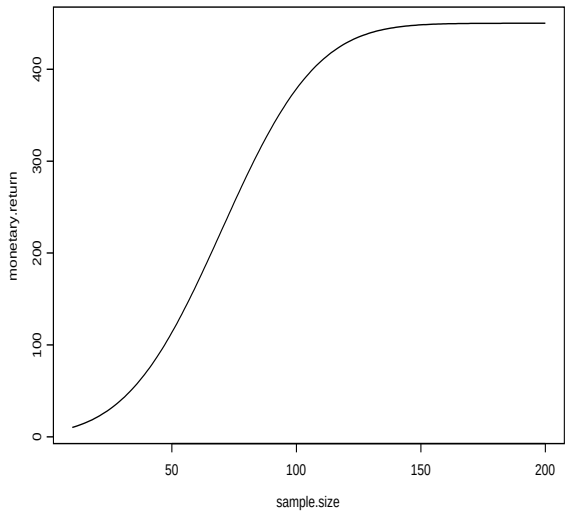


Figure 2: monetary return vs. sample size

(c) 'Cost-Benefit' Analysis

We can overlay cost and benefit curves. Some samples will be too small to give any benefit. Others are clearly too large. An appropriate trade-off of cost and benefit is required.

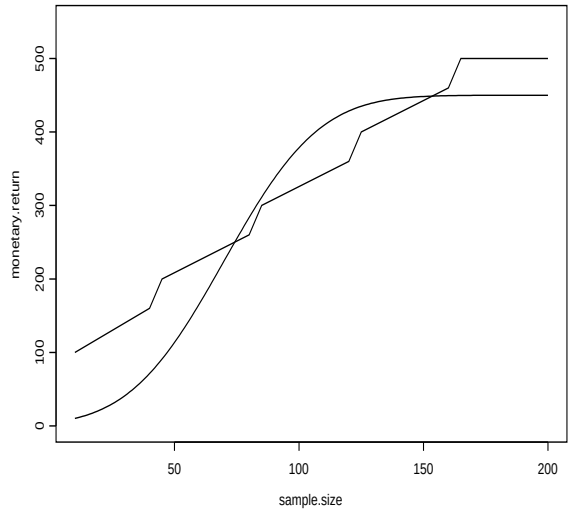


Figure 3: 'Cost-Benefit' analysis