Road Safety Data, Collection, Transfer and Analysis

Deliverable 6.2B
Sampling techniques and naturalistic driving study designs

Please refer to this report as follows:
Commandeur, Jacques J.F. (2012) Sampling techniques and naturalistic driving study designs, Deliverable 6.2B of the EC FP7 project DaCoTA.

Grant agreement No TREN/FP7/TR/233659/"DaCoTA"
Theme: Sustainable Surface Transport: Collaborative project
Project Coordinator:
Professor Pete Thomas, Vehicle Safety Research Centre, ESRI
Loughborough University, Ashby Road, Loughborough, LE11 3TU, UK

Project Start date: 01/01/2010 **Duration 36 months**

Organisation name of lead contractor for this Deliverable:
SWOV Institute for Road Safety Research, The Netherlands

Report Author(s):
Jacques J.F. Commandeur, SWOV

Due date of Deliverable **31-12-2012** **Submission date:** **07/11/2012**

Project co-funded by the European Commission within the Seventh Framework Programme

Dissemination Level

PU **Public**

Contents

Chapter 1

Introduction

In this document we provide an overview of sampling and estimation methods that can be used to obtain population values of risk exposure data and safety performance indicators based on naturalistic driving study designs. More specifically, we discuss how to determine the optimal sample size required for the estimation of such population values based on a probabilistic sample of that same population, and with a predefined level of precision. Examples of population values of interest are the mean or the total of the number of kilometers traveled by all drivers of a car in a country, and the percentage of these same drivers wearing a seat belt. We restrict ourselves to probabilistic sampling techniques because non-probabilistic sampling techniques like convenience and snowball sampling do not lend themselves to the evaluation of the statistical properties of parameter estimates from a sample, and are therefore unfit for the estimation of sample size.

In Chapters 2 and 3, where simple random sampling and stratified random sampling are introduced, respectively, we assume that a complete sampling frame is available for all units in the population of interest. In Chapter 4 we discuss how to sample from a population with unequal probabilities, and why this can be useful. In Chapter 5 we extend the discussion to the situation where a complete sampling frame is not available, and present multi-stage sampling. In Chapter 6 we present two alternative methods of estimation of population parameters from a sample: the ratio and the regression estimator. In Chapter 7 we consider the possibilities and implications of repeated sampling of the same population. In all these chapters we are only concerned with the quantification of *sampling error*, and its consequences for the estimation of sample size. However, there are other types of potential errors as well, and these will be discussed in Chapter 8. Finally, considering all these aspects of sampling techniques, in Chapter 9 we provide a list of recommendations for the study design of the collection of data based on naturalistic driving observations.

As will become clear in the following chapters, a key concept in deciding about the size of a sample is the concept of *precision*. Precision quantifies how closely the *sample estimate* of a population parameter (such as a mean or a total or a percentage) corresponds to the *actual value* of the population parameter. Keeping everything else fixed, the following rule applies: given a certain sampling strategy, the higher the precision we impose on our estimate, the larger the sample should be. If we only tolerate an absolute 1% error in our estimate of a population characteristic like the percentage of car drivers in a country wearing a seat belt, for example, a larger sample will be required than if we can settle for an absolute 5% error in our estimate. Supposing that the true percentage of seat belt wearing is 80%, the latter

more liberal precision implies that we can expect the estimated percentage to be within the range of $80 \pm 5\%$ (i.e., somewhere between 75% and 85%), while the former more conservative precision yields an estimated percentage with a range of 79% to 81%. With a precision of only 5% it will therefore not be possible to detect effects of road safety measures on seat belt wearing that are smaller than 10%, while the more conservative precision of 1% allows us to detect much smaller effects, if any.

The results presented in this document are based on the classic textbook of Cochran (1977), on Moors and Muilwijk (1975) and Hays (1970), and on several internet sources.

Chapter 2

Simple random sampling

2.1 Introduction

Simple random sampling is a method of selecting n units out of the N units in the population such that every one of the possible distinct samples has an equal chance of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N, and a series of random numbers between 1 and N is then drawn by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units bearing these n numbers is the sample.

Since a number that has been drawn is removed from the population for all subsequent draws, this method is called random sampling *without replacement*. It is also possible to use random sampling *with replacement*, in which case at any draw, all N members of the population are given an equal chance of being drawn, no matter how often they already have been drawn.

Let the population consist of N units denoted by $y_1, y_2, \ldots, y_N$. Let the sample consist of $n < N$ units denoted by $y_1, y_2, \ldots, y_n$. For totals and means we have the following definitions.

Table 2.1: Definitions of totals and means.

	Population	Sample
Total:	$Y = \sum_{i=1}^{N} y_i = y_1 + y_2 + \cdots + y_N$	$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n$
Mean:	$\bar{Y} = \frac{Y}{N} = \frac{y_1+y_2+\cdots+y_N}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$	$\bar{y} = \frac{y_1+y_2+\cdots+y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$

The interest in sampling centers most frequently on four characteristics of the population:

1. the mean $\bar{Y}$, e.g., the average number of motor vehicle kilometers driven;
2. the total Y, e.g., the total number of motor vehicle kilometers driven;
3. the ratio of two totals or means $R = Y/X = \bar{Y}/\bar{X}$, e.g., the total number of fatalities divided by the total number of motor vehicle kilometres driven;
4. the proportion of units that belong to some defined class, e.g., the proportion of drivers wearing a seat belt.

Table 2.2: Population characteristic and their estimators.

Population characteristic	Estimator
Population mean $\mu = \bar{Y}$	$\hat{\bar{Y}} = \bar{y} =$ sample mean
Population total $N\mu = Y$	$\hat{Y} = N\bar{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$
Population ratio R	$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$

The symbol $\hat{}$ denotes an estimate of a population characteristic made from a sample.

In the formula for $\hat{Y}$ in Table 2.2, the factor $\frac{N}{n}$ by which the sample total is multiplied is also called the *expansion* or *raising* of *inflation* factor. Its inverse $\frac{n}{N}$, the ratio of the size of the sample to that of the population, is called the *sampling fraction* and is denoted by the letter f.

For the next couple of chapters two essential assumptions are being made:

- the sampling distribution of the statistic G is normally distributed,
- the only error in the estimate of the population parameter θ is due to random sampling.

In Chapter 8 we will also consider other sources of error than the sampling error.

2.2 Properties of the mean of a sample

Generally, suppose one is interested in estimating the value of population parameter θ, and one is considering the use of some sample statistic G as an estimate of the value of θ. Then an estimate of the parameter θ made from the sample statistic G is said to be *unbiased* if

$$E(G) = \theta. \tag{2.1}$$

That is, the sample quantity G is unbiased as an estimator of θ if the expectation of G is θ; in other words, G averaged over all possible random samples of size n is exactly equal to the true population value θ. If samples are drawn without replacement, there are $M = \frac{N!}{n!(N-n)!}$ possible random samples. If samples of size n are drawn with replacement, the total number of possible random samples is $M = N^n$.

For example, consider the mean of a simple random sample as an estimator of the mean of the population. In this case

$$G = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n},$$

and

$$\theta = \mu = \bar{Y}.$$

The question now is, is

$$E(\bar{y}) = \mu = \bar{Y}? \tag{2.2}$$

The answer is yes, because

$$E(\bar{y}) = E(\frac{y_1 + y_2 + \cdots + y_n}{n}). \tag{2.3}$$

But since

$$E(aX) = aE(X) \tag{2.4}$$

for some constant a and random variable X, and

$$E(X + Y + Z) = E(X) + E(Y) + E(Z) \tag{2.5}$$

for some random variables X, Y, and Z, it follows from (2.3) that

$$E(\bar{y}) = \frac{E(y_1) + E(y_2) + \cdots + E(y_n)}{n}.$$

But any $E(y)$ is μ by definition, for observations taken at random from the same population, and therefore

$$E(\bar{y}) = \frac{nE(y)}{n} = \mu = \bar{Y}.$$

The mean $\bar{y}$ of a random sample is an unbiased estimator of $\bar{Y}$, the population mean. Analogously we find that

$$\hat{Y} = N\bar{y} \tag{2.6}$$

is an unbiased estimator of the population total Y.

2.3 Properties of the variance of a sample

We define the sample variance as

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2}{n-1} - \frac{n}{n-1}\bar{y}^2. \tag{2.7}$$

Letting

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} \tag{2.8}$$

denote the population variance, the question again is, does s^2 satisfy (2.1), i.e., is the sample variance (2.7) an unbiased estimator of the population variance (2.8)? Or expressed mathematically:

$$E(s^2) = S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}?$$

In order to investigate the relation between $E(s^2)$ and S^2, we start by noting that it follows from (2.7) that

$$E(s^2) = E\left(\frac{\sum_{i=1}^n y_i^2}{n-1} - \frac{n}{n-1}\bar{y}^2\right) = E\left(\frac{\sum_{i=1}^n y_i^2}{n-1}\right) - E\left(\frac{n}{n-1}\bar{y}^2\right). \tag{2.9}$$

Applying (2.4) and (2.5) to the first term on the right we see that

$$E\left(\frac{\sum_{i=1}^n y_i^2}{n-1}\right) = \frac{\sum_{i=1}^n E(y_i^2)}{n-1}.$$

But since the population variance is defined as

$$S^2 = E[y_i - E(y_i)]^2 = E(y_i^2) - [E(y_i)]^2 = E(y_i^2) - \bar{Y}^2 = E(y_i^2) - \mu^2,$$

we also have that

$$E(y_i^2) = S^2 + \bar{Y}^2 = S^2 + \mu^2 \tag{2.10}$$

for any observation i. This means that

$$\frac{\sum_{i=1}^n E(y_i^2)}{n-1} = \frac{\sum_{i=1}^n (S^2 + \mu^2)}{n-1} = \frac{n}{n-1}(S^2 + \mu^2). \tag{2.11}$$

Generally, for any sample statistic G, we may consider its *sampling distribution*. This is a theoretical probability distribution that shows the functional relation between the possible values of some summary characteristic G of n cases drawn at random and the probability (density) associated with each value over all possible samples of size n from a particular population. In general, the sampling distribution of values for a sample statistic will not be the same as the population distribution, unless the sample sizes considered satisfy $n = 1$. Sampling distributions differ from population distributions in that the random variable is always the value of some statistic based on a sample of n cases.

Since a sample statistic is a random variable, the mean and variance of any sampling distribution are defined in the usual way. That is, let G be any sample statistic, then its expectation or mean is

$$E(G) = \mu_G,$$

and its variance is

$$\sigma_G^2 = E(G - \mu_G)^2 = E(G^2) - [E(G)]^2 = E(G^2) - \mu_G^2.$$

So, if the statistic G is the mean $\bar{y}$ of a random sample, for example, then

$$E(\bar{y}) = \mu_{\bar{y}} = \mu,$$

meaning that the mean of the sampling distribution of means is the same as the population mean, and

$$\sigma_{\bar{y}}^2 = E(\bar{y}^2) - [E(\bar{y})]^2 = E(\bar{y}^2) - \mu_{\bar{y}}^2 = E(\bar{y}^2) - \mu^2, \tag{2.12}$$

and therefore

$$E(\bar{y}^2) = \sigma_{\bar{y}}^2 + \mu^2. \tag{2.13}$$

Combining (2.9), (2.11), and (2.13) we obtain

$$E(s^2) = \frac{n}{n-1}(S^2 + \mu^2) - \frac{n}{n-1}(\sigma_{\bar{y}}^2 + \mu^2) = \frac{n}{n-1}(S^2 - \sigma_{\bar{y}}^2), \tag{2.14}$$

which shows that the expectation of the sample variance (2.7) is equal to the difference between the population variance S^2 and the variance of the sampling distribution of means $\sigma_{\bar{y}}^2$, up to a factor $\frac{n}{n-1}$.

Further working out (2.12), and assuming that the sample mean is based on n *independent* observations, and letting any pair of these observations be denoted by i and j, with scores y_i and y_j, then the square of the sample mean is

$$\bar{y}^2 = \frac{(y_1 + y_2 + \cdots + y_n)^2}{n^2} = \frac{y_1^2 + y_2^2 + \cdots + y_n^2 + 2\sum_{i<j} y_i y_j}{n^2}, \tag{2.15}$$

the sum of the squared scores, plus twice the sum of the cross-products of all pairs of scores, all divided by n^2. For a pair of *independent* observations i and j, we have that

$$E(y_i y_j) = E(y_i)E(y_j) = \mu^2. \tag{2.16}$$

Combining (2.10) and (2.16) we therefore find that

$$\begin{aligned} E(\bar{y}^2) =& \frac{E(y_1^2) + E(y_2^2) + \cdots + E(y_n^2) + 2\sum_{i<j} E(y_i y_j)}{n^2} \\ =& \frac{n(S^2 + \mu^2) + n(n-1)\mu^2}{n^2} \\ =& \frac{nS^2 + n\mu^2 + n^2\mu^2 - n\mu^2}{n^2} \\ =& \frac{nS^2 + n^2\mu^2}{n^2} = \frac{S^2}{n} + \mu^2. \end{aligned} \tag{2.17}$$

Substitution of (2.17) in (2.12) finally yields the *variance of the mean*

$$\sigma_{\bar{y}}^2 = E(\bar{y}^2) - \mu^2 = \frac{S^2}{n}. \tag{2.18}$$

The variance of the sampling distribution of means for independent samples of size n is always equal to the population variance divided by the sample size, S^2/n. The *standard error of the mean* then equals

$$\sigma_{\bar{y}} = \sqrt{\sigma_{\bar{y}}^2} = \frac{S}{\sqrt{n}}. \tag{2.19}$$

It follows from (2.18) that the variance of the sampling distribution of means is exactly equal to the population variance when $n = 1$ only. It also follows from (2.18) that the variance of the sampling distribution of means gets smaller and smaller as the sample size

n gets larger and larger, as one would expect on intuitive grounds. Stated differently, the larger the sample size, the more probable it is that the sample mean comes arbitrarily close to the population mean, a fact that is also known as the law of large numbers. Of course, if the sample is large enough to embrace the entire population, then there is no difference whatsoever between the sample mean $\bar{y}$ and the population mean $\bar{Y} = \mu$.

Substitution of (2.18) in (2.14) yields

$$E(s^2) = \frac{n}{n-1}(S^2 - \sigma^2_{\bar{y}}) = \frac{n}{n-1}(S^2 - \frac{S^2}{n}) = \frac{n}{n-1}S^2 - \frac{S^2}{n-1} = \frac{(n-1)S^2}{n-1} = S^2, \quad (2.20)$$

which proves that the sample variance (2.7) is an unbiased estimator of the population variance S^2. We therefore add a hat to definition (2.7) of the sample variance in order to indicate that it is an unbiased estimator of the population variance:

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}. \quad (2.21)$$

Given a sample of size n, the standard deviation of the population, S, is estimated simply by taking the square root of the unbiased estimator $\hat{s}^2$ in (2.21):

$$\text{estimated } S = \sqrt{\hat{s}^2} = \hat{s}, \quad (2.22)$$

while the standard error of the mean is also estimated by using the unbiased estimate $\hat{s}^2$ in (2.21) as follows

$$\text{estimated } \sigma_{\bar{y}} = \frac{\text{estimated } S}{\sqrt{n}} = \sqrt{\frac{\hat{s}^2}{n}} = \frac{\hat{s}}{\sqrt{n}}. \quad (2.23)$$

Example 2-1. Suppose we have a population consisting of $N = 5$ units with values 2, 4, 6, 8, and 10. For this population the total, the mean, and the variance are $Y = \sum_{i=1}^{N} y_i = 30$, $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{30}{5} = 6$, and $S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1} = 10$, respectively. From this population we can draw a total of $M = \frac{N!}{n!(N-n)!} = \frac{5!}{2!3!} = 10$ different simple random samples of $n = 2$ units without replacement. For each sample we calculate its mean $\bar{y}$, its total $N\bar{y}$, and its variance $\hat{s}^2$, see Table 2.3.

By averaging the values of $\hat{\bar{Y}}$, $\hat{Y}$, and $\hat{s}^2$ for this sample space in Table 2.3 we find their expectations to be equal to $E(\hat{\bar{Y}}) = \frac{1}{M}\sum_{j=1}^{M}\bar{y}_j = \frac{1}{10}\sum_{j=1}^{10}\bar{y}_j = \frac{60}{10} = 6 = \mu = \bar{Y}$, $E(\hat{Y}) = \frac{1}{M}\sum_{j=1}^{M}\hat{Y}_j = \frac{300}{10} = 30 = Y = N\mu$, and $E(\hat{s}^2) = \frac{1}{M}\sum_{j=1}^{M}\hat{s}^2_j = \frac{100}{10} = 10 = S^2$, confirming the unbiasedness of these estimates.

2.4 Sampling from finite populations

So far, we assumed that samples are drawn from very large populations. When samples are drawn from populations that are relatively small, the sample mean and the sample total are still unbiased estimators of the population mean and population total, respectively. However, when samples are drawn from populations that are relatively small, and – more specifically

Table 2.3: All possible simple random samples of $n = 2$ units without replacement from a population of $N = 5$ units with values 2, 4, 6, 8, and 10.

Sample		$\hat{\bar{Y}} = \bar{y}$	$\hat{Y}$	$\hat{s}^2$	$(\bar{y} - \bar{Y})^2$
2	4	3	15	2	9
2	6	4	20	8	4
2	8	5	25	18	1
2	10	6	30	32	0
4	6	5	25	2	1
4	8	6	30	8	0
4	10	7	35	18	1
6	8	7	35	2	1
6	10	8	40	8	4
8	10	9	45	2	9
Total		60	300	100	30

– when the sampling fraction is relatively large, i.e., when $\frac{n}{N} > 0.1$, then this must be accounted for in the formulas for the variances and standard errors.

Specifically, for a population of N cases in all, from which samples of size n are drawn, the sampling variance of the mean is

$$\mathrm{V}(\bar{y}) = \sigma^2_{\bar{y}} = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1-f). \tag{2.24}$$

The ratio $\frac{N-n}{N}$ in (2.24) is called the *finite population correction* (fpc), see Cochran (1977). The sampling variance of the mean tends to be somewhat smaller for a fixed value of n when sampling is from a finite population than when it is from an infinite population. Note that here the size of $\sigma^2_{\bar{y}}$ depends both on N, the total number in the population, and n, the sample size. It follows from (2.24) that the standard error of the mean is

$$\sigma_{\bar{y}} = \frac{S}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} = \frac{S}{\sqrt{n}}\sqrt{1-f}. \tag{2.25}$$

For finite populations the sampling variance of the total $\hat{Y} = N\bar{y}$ is

$$\mathrm{V}(\hat{Y}) = \sigma^2_{\hat{Y}} = \frac{N^2S^2}{n}\frac{(N-n)}{N} = \frac{N^2S^2}{n}(1-f), \tag{2.26}$$

while its standard error is

$$\sigma_{\hat{Y}} = \frac{NS}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} = \frac{NS}{\sqrt{n}}\sqrt{1-f}. \tag{2.27}$$

Cochran (1977, p.26) further shows that the unbiased sample estimator of the variance of the mean $\bar{y}$ is

$$\mathrm{v}(\bar{y}) = \text{estimated}\,\sigma^2_{\bar{y}} = s^2_{\bar{y}} = \frac{\hat{s}^2}{n}\left(\frac{N-n}{N}\right) = \frac{\hat{s}^2}{n}(1-f), \tag{2.28}$$

and that the unbiased sample estimator of the variance of the total $\hat{Y} = N\bar{y}$ is

$$\mathrm{v}(\hat{Y}) = \text{estimated}\,\sigma^2_{\hat{Y}} = s^2_{\hat{Y}} = \frac{N^2\hat{s}^2}{n}\left(\frac{N-n}{N}\right) = \frac{N^2\hat{s}^2}{n}(1-f), \tag{2.29}$$

while the corresponding standard errors are

$$\text{estimated}\,\sigma_{\bar{y}} = s_{\bar{y}} = \frac{\hat{s}}{\sqrt{n}}\sqrt{1-f}, \tag{2.30}$$

and

$$\text{estimated}\,\sigma_{\hat{Y}} = s_{\hat{Y}} = \frac{N\hat{s}}{\sqrt{n}}\sqrt{1-f}, \tag{2.31}$$

respectively. In all these formulas $\hat{s}$ is defined as in (2.21). As Cochran (1977, p.27) mentions the latter two estimates are slightly biased, but for most applications the bias is unimportant.

In general, standard errors of the estimated population mean and total are used for three important purposes:

- to compare the precision of simple random sampling with that obtained with other methods of sampling, see the examples in Chapters 3, 5, and 6;
- to estimate the size of the sample needed in a survey that is being planned, see, e.g., Sections 2.8, 2.10, 3.6, 3.9, 5.2.3, 5.2.4, 6.1.3, and 6.2.3;
- to estimate the precision actually attained in a survey that has been completed.

Their calculation requires S^2, the population variance. In practice this will not be known, but it can be estimated from the sample data.

Example 2-2. Suppose we have the same population as in Example 2-1, consisting of $N = 5$ units with values 2, 4, 6, 8, and 10, and with mean and variance equal to $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{30}{5} = 6$, and $S^2 = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N-1} = 10$, respectively. When drawing all possible $M = \frac{N!}{n!(N-n)!} = \frac{5!}{2!3!} = 10$ simple random samples of size $n = 2$ *without replacement*, in Table 2.3 we see that

$$\sigma^2_{\bar{y}} = \frac{1}{M}\sum_{j=1}^{M}(\bar{y}_j - \bar{Y})^2 = \frac{1}{10}\sum_{j=1}^{10}(\bar{y}_j - 6)^2 = \frac{30}{10} = 3,$$

while it follows from (2.24) that – in this example –

$$\mathrm{V}(\bar{y}) = \sigma^2_{\bar{y}} = \frac{S^2}{n}\frac{(N-n)}{N} = \left(\frac{10}{2}\right)\left(\frac{3}{5}\right) = 3.$$

This illustrates that (2.24) indeed is an unbiased estimator of the variance of the sampling distribution of the mean.

2.5 Sampling with replacement

So far, we have discussed results for single random sampling without replacement of the units in the population. For large populations the fact that samples are taken with or without replacement can safely be ignored. However, when the population under study is relatively small, the process of sampling with replacement has a real effect on the sampling distribution.

Even in this situation the sample mean is still an unbiased estimator of the population mean, regardless of the size of the population sampled. However, when samples are drawn with replacement, for finite populations the variance of the mean is

$$\mathrm{V}(\bar{y}) = \sigma^2_{\bar{y}} = \frac{N-1}{N}\frac{S^2}{n}. \tag{2.32}$$

Example 2-3. Suppose we again have the same population as in Example 2-1 consisting of $N = 5$ units with values 2, 4, 6, 8, and 10. As mentioned before, for this population the mean and the variance are $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{30}{5} = 6$, and $S^2 = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N-1} = 10$, respectively. From this population we can draw a total of $M = N^n = 5^2 = 25$ different simple random samples of $n = 2$ units *with replacement*. For each sample we calculate its mean $\bar{y}$, see Table 2.4.

From Table 2.4 we see that

$$E(\bar{y}) = \frac{1}{M}\sum_{j=1}^{M}\bar{y}_j = \frac{1}{25}\sum_{j=1}^{25}\bar{y}_j = \frac{150}{25} = 6,$$

which illustrates that the mean of the sample is still an unbiased estimator of the mean of the population when drawing simple random samples with replacement. Moreover, it follows from Table 2.4 that

$$\sigma^2_{\bar{y}} = \frac{1}{M}\sum_{j=1}^{M}(\bar{y}_j - \bar{Y})^2 = \frac{1}{25}\sum_{j=1}^{25}(\bar{y}_j - 6)^2 = \frac{100}{25} = 4,$$

and from (2.32) that – in this example –

$$\mathrm{V}(\bar{y}) = \sigma^2_{\bar{y}} = \frac{S^2}{n}\frac{(N-1)}{N} = \left(\frac{10}{2}\right)\left(\frac{4}{5}\right) = 4,$$

confirming that (2.32) indeed yields an unbiased estimate of the variance of the sampling distribution of the mean when drawing simple random samples with replacement.

2.6 Unbiased estimators of proportions

When dealing with *dichotomous* (qualitative) variables with only two categories with A units in the population belonging to the first category and $N - A$ in the population units belonging to the second category, we have a proportion of $P = A/N$ in the first category and a proportion of $Q = 1 - P$ in the second category. Letting a denote the number of units in

Table 2.4: All possible simple random samples of $n = 2$ units *with replacement* from a population of $N = 5$ units with values 2, 4, 6, 8, and 10.

Sample		$\hat{\bar{Y}} = \bar{y}$	$(\bar{y} - \bar{Y})^2$
2	2	2	16
2	4	3	9
2	6	4	4
2	8	5	1
2	10	6	0
4	2	3	9
4	4	4	4
4	6	5	1
4	8	6	0
4	10	7	1
6	2	4	4
6	4	5	1
6	6	6	0
6	8	7	1
6	10	8	4
8	2	5	1
8	4	6	0
8	6	7	1
8	8	8	4
8	10	9	9
10	2	6	0
10	4	7	1
10	6	8	4
10	8	9	9
10	10	10	16
Total		150	100

the first category in the sample, then it can be shown that $p = a/n$ is an unbiased estimator of P (i.e., $E(p) = P$), and

$$\hat{s}^2 = \frac{n}{n-1} pq \tag{2.33}$$

(with $q = 1 - p$) is an unbiased estimator of

$$S^2 = \frac{N}{N-1} PQ, \tag{2.34}$$

the variance of the proportion in the population, see Cochran (1977, p.51). For dichotomous variables we further have that the variance of the distribution of proportions in samples of size n is

$$\mathrm{V}(p) = \sigma_P^2 = \frac{N}{N-1} \frac{PQ}{n} \left(1 - \frac{n}{N}\right) = \left(\frac{N-n}{N-1}\right) \frac{PQ}{n}. \tag{2.35}$$

If p and P are the sample and population *percentages*, respectively, falling into class C, (2.35) continues to hold for the variance of p. The square root of the latter variance is the *standard error of a proportion*, and is denoted by σ_P. For populations that are not too small (say $N > 20$), the ratio $N/(N-1)$ in (2.35) may be dropped yielding

$$\mathrm{V}(p) = \sigma_P^2 = \frac{PQ}{n}\left(1 - \frac{n}{N}\right). \tag{2.36}$$

Moreover, if the sampling fraction is also not too large (say $n < 0.1N$), then the latter formula can be further simplified to the well-known expression:

$$\mathrm{V}(p) = \sigma_P^2 = \frac{PQ}{n}, \tag{2.37}$$

see Moors and Muilwijk (1975, p.19). An unbiased estimator of the latter variance from the sample is obtained with:

$$\mathrm{v}(p) = \text{estimated}\,\sigma_P^2 = s_{\bar{y}}^2 = \frac{\hat{s}^2}{n}\frac{N-n}{N} = s_p^2 = \frac{pq}{n-1}\left(1 - \frac{n}{N}\right), \tag{2.38}$$

compare with (2.28). It follows that if N is very large relative to n, so that the finite population correction is negligible, an unbiased estimate of the variance of p is

$$\mathrm{v}(p) = s_p^2 = \frac{pq}{n-1},$$

see Cochran (1977, p.51-52).

Table 2.5: Values of PQ and $\sqrt{PQ}$. P is the population percentage in class C.

P	0	10	20	30	40	50	60	70	80	90	100
PQ	0	900	1600	2100	2400	2500	2400	2100	1600	900	0
$\sqrt{PQ}$	0	30	40	46	49	50	49	46	40	30	0

Equation (2.35) shows how the variance of the estimated proportion or percentage changes with P, for fixed n and N. If the finite population correction is ignored, we have (2.37). The function PQ and its square root are shown in Table 2.5. These functions may be regarded as the variance and standard deviation, respectively, for a sample of size $n = 1$. From Table 2.5 it is clear that the variances and their square roots have their greatest values when $P = Q = 50$, i.e., when the population is equally divided between the two classes. *This means that – for dichotomous variables – the choice for $P = 50$ always results in an upper limit for the required sample size.* Moreover, the standard error of p changes relatively little when P lies anywhere between 30 and 70%.

The variance of $\hat{A} = Np$, the estimated total number of units in class C, is

$$\mathrm{V}(\hat{A}) = \sigma_{Np}^2 = \frac{N-n}{N-1}\frac{N^2 PQ}{n}, \tag{2.39}$$

and an unbiased estimator of the variance of $\hat{A} = Np$, the estimated total number of units in class C in the population from the sample is obtained with:

$$\mathrm{v}(\hat{A}) = \text{estimated}\,\sigma_{Np}^2 = s_{Np}^2 = \frac{N(N-n)}{n-1}pq. \tag{2.40}$$

2.7 Confidence intervals

For large enough samples it can be assumed that the estimates of the mean $\bar{y}$ and the total $\hat{Y}$ are normally distributed about the population values $\mu = \bar{Y}$ and $N\mu = Y$ respectively, and confidence intervals for these estimates can be constructed as follows. For the mean $\bar{Y}$ we have

$$\bar{y} - ts_{\bar{y}} \leq \bar{Y} \leq \bar{y} + ts_{\bar{y}}, \tag{2.41}$$

where $s_{\bar{y}}$ is defined in (2.30) and t is the value of the normal deviate corresponding to the desired confidence probability if $n \geq 50$, and t is the value in the Student t table with $(n-1)$ degrees of freedom if $n < 50$. For small samples with very skew distributions, however, special methods are needed.

For large enough samples, the value of the normal deviate corresponding to the 95% confidence probability is $t = 1.96$, and (2.41) equals

$$\bar{y} - 1.96s_{\bar{y}} \leq \bar{Y} \leq \bar{y} + 1.96s_{\bar{y}}. \tag{2.42}$$

Formula (2.42) expresses that the inequality holds with a *confidence* of 95%. On average, for 95 out of 100 samples the confidence interval will contain the actual population mean. The lower and upper limits calculated for a specific sample are called the 95% *confidence limits* of the population mean, and the maximum distance of $1.96s_{\bar{y}}$ between the sample and the population mean is called the 95% *confidence interval*. The ratio $\frac{1.96s_{\bar{y}}}{\bar{y}}$ is known as the *relative* confidence interval; it is 1.96 times the variation coefficient of $\bar{y}$.

For a qualitative variable, the confidence interval is

$$p - ts_p \leq P \leq p + ts_p, \tag{2.43}$$

where s_p is the square root of (2.38).

For the total Y the confidence interval is calculated as

$$N\bar{y} - ts_{\hat{Y}} \leq Y \leq N\bar{y} + ts_{\hat{Y}}, \tag{2.44}$$

where $s_{\hat{Y}}$ is defined in (2.31). See Cochran (1977, p.27) and Hays (1970, Chapter 10).

Example 2-4. Signatures to a petition were collected on $N = 676$ sheets. Each sheet had enough space for 42 signatures, but on many sheets a smaller number of signatures had been collected. The number of signatures per sheet were counted on a random sample of $n = 50$ sheets (about 7% of the sample), with the results shown in Table 2.6. Estimate the total number of signatures to the petition and the 80% confidence limits.

Answer. Since the average number of signatures in this sample of 50 sheets is equal to

$$\bar{y} = \frac{\sum_{i=1}^{19} f_i y_i}{\sum_{i=1}^{19} f_i} = \frac{(23)(42) + (4)(41) + \cdots + (1)(3)}{23 + 4 + \cdots + 1} = \frac{1471}{50} = 29.42$$

signatures per sheet, the estimated total number of signatures in the population of 676 sheets equals

$$\hat{y} = N\bar{y} = (676)(29.42) = 19887.92.$$

Table 2.6: Results for a sample of 50 petition sheets: y_i is the number of signatures, and f_i the frequency of the number of sheets in the sample with y_i signatures.

y_i	42	41	36	32	29	27	23	19	16	15
f_i	23	4	1	1	1	2	1	1	2	2
y_i	14	11	10	9	7	6	5	4	3	
f_i	1	1	1	1	1	3	2	1	1	

An unbiased estimate of the variance of the population is obtained from (2.21), i.e., from

$$\hat{s}^2 = \frac{\sum_{i=1}^{19}(f_i(y_i - \bar{y}))^2}{n-1} = \frac{\sum_{i=1}^{19}(f_i(y_i - \bar{y}))^2}{49} = \frac{11220.18}{49} = 228.9832653,$$

with standard deviation

$$\hat{s} = \sqrt{\hat{s}^2} = \sqrt{228.9832653} = 15.13219301,$$

see (2.22). According to (2.31) the standard error of the total then equals

$$s_{\hat{Y}} = \frac{N\hat{s}}{\sqrt{n}}\sqrt{1-f} = \frac{(676)(15.13219301)}{\sqrt{50}}\sqrt{1-\frac{50}{676}} = 1392.122281.$$

The 80% confidence limits of the total are therefore

$$N\bar{y} - 1.28s_{\hat{Y}} \leq Y \leq N\bar{y} + 1.28s_{\hat{Y}},$$

and thus

$$19887.92 - (1.28)(1392.122281) \leq Y \leq 19887.92 + (1.28)(1392.122281),$$

meaning that

$$18106.00348 \leq Y \leq 21669.83652$$

with a confidence of 80%, i.e., in four out of five samples.

2.8 Estimation of sample size

As sample size increases, the distribution of the means of simple random samples from the same population approaches the normal distribution more and more, *irrespective whether the variable of interest in the population is normally distributed or not.* This is – loosely stated – the famous central limit theorem in statistics. In that case the mean $\bar{y}$ in the sample will therefore be located between $\mu - 1.96\sigma_{\bar{y}}$ and $\mu + 1.96\sigma_{\bar{y}}$ with a 95% probability, i.e.,

$$\mu - 1.96\sigma_{\bar{y}} \leq \bar{y} \leq \mu + 1.96\sigma_{\bar{y}}, \tag{2.45}$$

with a probability of 95%. It follows from (2.45) that

$$\bar{y} - 1.96\sigma_{\bar{y}} \leq \mu \leq \bar{y} + 1.96\sigma_{\bar{y}} \tag{2.46}$$

with the same probability. Substraction of $\bar{y}$ from (2.46) gives

$$-1.96\sigma_{\bar{y}} \leq \mu - \bar{y} \leq +1.96\sigma_{\bar{y}} \tag{2.47}$$

and therefore

$$|\mu - \bar{y}| \leq 1.96\sigma_{\bar{y}} \tag{2.48}$$

with a 95% probability.

Substituting (2.25) in (2.48) we see that – for a continuous variable –

$$|\mu - \bar{y}| \leq 1.96\sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)} \tag{2.49}$$

with a 95% probability. Assuming that we want the absolute error in our sample estimate $\bar{y}$ of the population mean μ to be no larger than $d = |\mu - \bar{y}|$, say, with a 95% probability, then from the latter equation we can obtain a formula for the required minimal sample size n:

$$d = t\sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}, \tag{2.50}$$

Note that we have replaced the value 1.96 corresponding to a probability of 95% with t in (2.50) in order to be as general as possible. If we want a 95% probability then $t = 1.96$; for a 90% probability $t = 1.64$, for a 99% probability $t = 2.58$, et cetera. Solving (2.50) for n we find that the minimal sample size should be

$$n = \frac{t^2 S^2}{\frac{t^2 S^2}{N} + d^2}. \tag{2.51}$$

For very large N this formula simplifies into

$$n_0 = \frac{t^2 S^2}{d^2}, \tag{2.52}$$

and we only need to estimate S. In this case it should be checked whether $n_0 < 0.1N$. If not we should apply the finite population correction and use

$$n = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{2.53}$$

which is identical to (2.51), as it is not very difficult to verify.

In order to establish the value of S in the population sometimes a small pilot study is done. Or reasonable estimates can be found from previous studies in the same research field, from studies in similar research fields, or based on theoretical grounds. Even with a rough estimation we can thus obtain a useful indication about the required sample size. Note that the formulas in this section are based on the normal distribution. It should therefore be

checked that the estimated n is large enough. If not, a larger sample should be taken then was calculated.

So far formulas have been provided for the estimation of sample size based on the *absolute error* $d = |\mu - \bar{y}|$ in the estimation of μ. If a *relative error* of the estimate of μ is used then

$$d\bar{Y} = t\sigma_{\bar{y}}, \tag{2.54}$$

or, upon substitution of (2.24),

$$d\bar{Y} = t\sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}. \tag{2.55}$$

In this case, d is a number satisfying $0 < d < 1$. Solving (2.55) for n, we obtain

$$n = \frac{t^2S^2}{\frac{t^2S^2}{N} + d^2\bar{Y}^2}.$$

The latter formula can also be written as

$$n = \frac{\left(\frac{tS}{d\bar{Y}}\right)^2}{1 + \frac{1}{N}\left(\frac{tS}{d\bar{Y}}\right)^2}, \tag{2.56}$$

see Cochran (1977, p.77). For very large N this formula simplifies into

$$n_0 = \frac{t^2S^2}{d^2\bar{Y}^2}. \tag{2.57}$$

The required sample size can also be expressed in terms of the *coefficient of variation* $c = \frac{S}{\bar{Y}}$. Substitution of $S = c\bar{Y}$ in (2.56) yields

$$n = \frac{t^2c^2}{\frac{t^2c^2}{N} + d^2}. \tag{2.58}$$

Example 2-5. For a bed of silver maple seedlings 1 ft wide and 430 ft long, it was found by complete enumeration that $\mu = \bar{Y} = 19$ and $S^2 = 85.6$, these being the true population values. The sampling unit was 1 ft of the length of the bed, so that $N = 430$. With simple random sampling, how many units must be taken to estimate μ within 10% of accuracy, apart from a chance of 1 in 20?

Answer. Using (2.57) as a first approximation, and since $d\bar{Y} = (0.1)(19) = 1.9$ and $t = 1.96$, we obtain

$$n_0 = \frac{(1.96^2)(85.6)}{1.9^2} = 91.09.$$

However, since $n_0/N = 91.09/430 = 0.21$ is not negligible, we use (2.53) yielding

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{91.09}{1.21} = 75.28.$$

Since $100(75.28/430) = 17.5$, almost 18% of the bed has to be counted in order to attain the precision desired.

For the estimation of sample size based on the absolute error $d = |N\mu - N\bar{y}|$ of the *total* $N\mu = Y$ in the population we have that

$$d = t\sigma_{\hat{Y}}, \tag{2.59}$$

or, upon substitution of (2.26),

$$d = t\sqrt{\frac{N^2S^2}{n}\left(\frac{N-n}{N}\right)}. \tag{2.60}$$

Solving (2.60) for n, we obtain

$$n = \frac{t^2S^2}{\frac{t^2S^2}{N} + \frac{d^2}{N^2}}. \tag{2.61}$$

If the finite population correction is negligible, i.e., if $n < 0.1N$, then the latter formula simplifies into

$$n_0 = \frac{t^2N^2S^2}{d^2}. \tag{2.62}$$

If an estimation of the sample size is required based on the relative error $dN\bar{Y}$ of the total $N\mu = Y$, where $0 < d < 1$, we need to solve

$$dN\bar{Y} = t\sqrt{\frac{N^2S^2}{n}\left(\frac{N-n}{N}\right)}, \tag{2.63}$$

yielding for n

$$n = \frac{t^2S^2}{\frac{t^2S^2}{N} + d^2\bar{Y}^2}. \tag{2.64}$$

For large N the latter formula simplifies into

$$n_0 = \frac{t^2S^2}{d^2\bar{Y}^2}. \tag{2.65}$$

Example 2-6. In a certain year we have a population of $N = 7$ million car drivers in a country. On average these road users drive $\bar{Y} = 15,000$ kilometers that year, with a standard deviation of $S = 5,000$ kilometers. The total amount of kilometers driven by car drivers that year is therefore $N\bar{Y} = 105$ billion kilometers. With simple random sampling, how many drivers must be selected to estimate the latter total $N\bar{Y}$ within 5% of accuracy, apart from a chance of 1 in 20?

Answer. Applying (2.65) as a first approximation with $d = 0.05$ and $t = 1.96$ we find that

$$n_0 = \frac{(1.96^2)(5,000^2)}{(0.05^2)(15,000^2)} = 170.7.$$

Since the sampling fraction $\frac{n_0}{N} = \frac{171}{7000000}$ is negligible the required sample size is $n = 171$. However, if we require a precision of 1% instead of 5%, the minimal sample size should be

$$n_0 = \frac{(1.96^2)(5,000^2)}{(0.01^2)(15,000^2)} = 4268.$$

The sampling fraction $\frac{n_0}{N} = \frac{4268}{7000000} = 0.0007$ is still negligible in this case, and the required sample size for this amount of precision is $n = 4268$.

For a *qualitative* variable we have for an absolute error that

$$|p - P| \leq t\sigma_P \tag{2.66}$$

which – upon substitution of (2.35) in (2.66), and letting $d = |p - P|$ again denote the value of this absolute error – yields

$$d = t\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}. \tag{2.67}$$

Solving (2.67) for n we find that

$$n = \frac{\frac{t^2PQ}{d^2}}{1 + \frac{1}{N}\left(\frac{t^2PQ}{d^2} - 1\right)}. \tag{2.68}$$

For practical use, an advance estimate p of P is substituted in this formula. If N is large, a first approximation is

$$n_0 = \frac{t^2pq}{d^2}. \tag{2.69}$$

In practice we therefore first calculate n_0. If the sampling fraction n_0/N is negligible, n_0 is a satisfactory approximation of n in (2.68). If not, comparison of (2.68) and (2.69) shows that n is obtained as

$$n = \frac{n_0}{1 + (n_0 - 1)/N} \doteq \frac{n_0}{1 + (n_0/N)}, \tag{2.70}$$

where $\doteq$ stands for "is approximately equal to".

It may be noted that it is easier to obtain an estimate of the required sample size for a qualitative variable than for a quantitative variable. For a quantitative variable we need to

have an estimate of the variance S^2 of the variable in the population if we use an absolute error, see (2.51) and (2.61), and estimates of both of the mean $\bar{Y}$ and the variance S^2 of the variable in the population if we use a relative error, see (2.56) and (2.64). For a qualitative variable, on the other hand, we only need an estimate of the proportion or percentage P in the population, see (2.68). Even if we do not know this proportion or percentage, it is still possible to obtain an estimate of the required sample size by using a value of $P = 0.5$ in (2.68), since this yields an upper bound for the required sample size (see also Section 2.6).

Example 2-7. An anthropologist is preparing to study the inhabitants of some island. Among other things, he wishes to estimate the percentage of inhabitants belonging to blood group O. The anthropologist wants this percentage to be correct within 5% in the sense that, if the sample shows 43% to have blood group O, the percentage for the whole island is sure to lie within 38% and 48%. He also is willing to take a 1 in 20 chance of getting an unlucky sample. The total population of the island is $N = 3200$. How large should the sample be?

Answer. In technical terms, the proportion p from the sample is to lie in the range $P \pm 5$, except for a 1 in 20 chance. Thus, $d = |p - P| = 0.05$ in (2.68), and $t = 1.96$. Moreover, assuming the worst case scenario where $P = 0.5$ (see Section 2.6), it follows from (2.69) that – as a first approximation – the sample size should be

$$n_0 = \frac{1.96^2(0.5)(0.5)}{0.05^2} = \frac{0.9604}{0.0025} = 384.2.$$

Since $n_0/N = 384/3200 = 0.12$, which is larger than 0.1, the finite population correction (fpc) is needed. Correcting for the fpc by application of (2.70) results in an estimated sample size of

$$n = \frac{384}{1 + (384 - 1)/3200} = 343.$$

Example 2-8. For the same population as in Example 2-6 we wish to estimate the percentage of car drivers wearing their seat belt. We want this percentage to be correct within 5% in the sense that, if the sample shows 80% to wear a seat belt, the percentage for the whole car driver population is sure to lie within 75% and 85%. We also are willing to take a 1 in 20 chance of getting an unlucky sample. How large should *this* sample be?

Answer. The proportion p from the sample is to lie in the range $P \pm 0.05$, except for a 1 in 20 chance. Thus, $d = |p - P| = 0.05$ in (2.68), and $t = 1.96$. Moreover, assuming the worst case scenario where $P = 0.5$ (see Section 2.6), it follows from (2.69) that – as a first approximation – the sample size should be

$$n_0 = \frac{1.96^2(0.5)(0.5)}{0.05^2} = \frac{0.49}{0.0025} = 384.2.$$

Since $n_0/N = 384/5,000,000 \doteq 0$, the finite population correction (fpc) is not needed. Assuming that $P = 0.8$ in the population we obtain

$$n_0 = \frac{1.96^2(0.8)(0.2)}{0.05^2} = \frac{2.4586}{0.0025} = 246.$$

Again assuming $P = 0.8$, but now requiring a precision of 1% instead of 5% we have

$$n_0 = \frac{1.96^2(0.8)(0.2)}{0.01^2} = \frac{0.614656}{0.0001} = 6,147,$$

where the fpc is still not needed.

Sometimes, particularly when estimating the total number NP of units in class C, we wish to control the *relative* error instead of the absolute error in Np; for example, we may wish to estimate NP with an error not exceeding 10%. Then

$$dP = t\sigma_P, \tag{2.71}$$

or, upon substitution of (2.35),

$$dP = t\sqrt{\frac{PQ}{n}\left(\frac{N-n}{N-1}\right)}. \tag{2.72}$$

In this case, d is again a number satisfying $0 < d < 1$. For this specification, replace d by dP in formulas (2.68) and (2.69). From (2.69) we then get

$$n_0 = \frac{t^2pq}{d^2p^2} = \frac{t^2}{d^2}\frac{q}{p}, \tag{2.73}$$

while formula (2.70) is unchanged.

We end by noting that the value of d in formulas (2.50), (2.55), (2.60), (2.67), (2.71) is known as the *sampling error*. For given population variance S^2, population mean $\bar{Y}$, and/or population proportion P, whichever is appropriate, and population size N, sample size n, and confidence limit t, these formulas can therefore also be used to calculate the sampling error of a particular sampling design.

Example 2-9. From a population of 10,000 car drivers with a percentage of 80 wearing the seat belt, a simple random sample of 400 car drivers is drawn. Assuming a 1 in 20 chance of getting an unlucky sample, what is the sampling error of this sample?

Answer. Since $N = 10,000$, $P = 0.8$, $n = 400$, and $t = 1.96$ in this situation, we apply (2.67) which gives

$$d = t\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}} = 1.96\sqrt{\left(\frac{10,000-400}{10,000-1}\right)\frac{(0.8)(0.2)}{400}} = 0.048.$$

The absolute sampling error in this sampling design is therefore 4.8%.

2.9 Sample size with more than one item

In many surveys information has to be collected on more than one item. One method of determining sample size is to select those items that are considered the most vital, and then

to estimate the minimal sample size needed for each of these vital items separately. When the single item estimations of n have been completed, it is time to make decisions based on the results. If the estimations are all reasonably close, and the largest n is within budget, then this sample size is used. When the n's are quite different, however, selecting the largest n may result in an overall precision that is much higher than originally intended, or in a very expensive survey. In this case, the desired precision may be lowered for some items thus allowing for the use of a smaller sample size. Sometimes the estimated n's vary so wildly that items simply have to be dropped from the survey altogether in view of the resources available and the precision required for the purpose of the survey.

Example 2-10. Again consider Examples 2-6 and 2-8, where it was found – using simple random sampling – that the estimation of the total distance traveled by car drivers with a precision of 1% required a minimal sample size of $n = 4,268$ car drivers, while the estimation of the total percentage of seat belt wearing by the same population with a precision of 1% required a minimal sample size of $n = 6,147$ car drivers. If we want to obtain estimates for both these items with a precision of 1%, we therefore should select a simple random sample of $n = 6,147$ car drivers.

2.10 Sample size when estimates are needed for subpopulations

In Section 2.6 we discussed how to obtain an unbiased estimate of the number of units in the population having a certain property. It often happens that the category is so important that we are not only interested in its size but also in the mean or total of the variable of interest in this category. Such categories are also known as *subpopulations* or *domains of study*.

Consider a population of N units, of which A units belong to a certain subpopulation. We will denote parameters corresponding to this subpopulation with the subscript s, and Y_s therefore refers to the total of variable y in the subpopulation, and $\bar{Y}_s$ to its mean. Then

$$Y_s = \sum_{i=1}^{A} y_i \text{ and } \bar{Y}_s = \frac{1}{A}\sum_{i=1}^{A} y_i, \tag{2.74}$$

are the *subtotal* and the *submean* of the population, respectively, and – letting a denote the number of elements in a random sample of this subpopulation –

$$\bar{y}_s = \frac{1}{a}\sum_{i=1}^{a} y_i \tag{2.75}$$

is an unbiased estimator of the submean of the population. The variance of y in the subpopulation is

$$S_s^2 = \frac{1}{A-1}\sum_{i=1}^{A}(y_i - \bar{Y}_s)^2, \tag{2.76}$$

and an unbiased sample estimator of this variance is

$$\hat{s}_s^2 = \frac{1}{a-1}\sum_{i=1}^{a}(y_i - \bar{y}_s)^2, \tag{2.77}$$

compare with (2.21).

Formulas (2.24) and (2.28) for the variance of the mean and the sample estimator of the variance of the mean still apply to subpopulations, in which case they can be written as

$$\mathrm{V}(\bar{y}_s) = \sigma_{\bar{y}_s}^2 = \frac{S_s^2}{a}(1 - \frac{a}{A}), \tag{2.78}$$

and

$$\mathrm{v}(\bar{y}_s) = \text{estimated}\,\sigma_{\bar{y}_s}^2 = s_{\bar{y}_s}^2 = \frac{\hat{s}_s^2}{a}(1 - \frac{a}{A}), \tag{2.79}$$

respectively. Analogous to (2.41), confidence limits for the sample estimate of the submean (2.75) can be constructed with

$$\bar{y}_s - t s_{\bar{y}_s} \leq \bar{Y}_s \leq \bar{y}_s + t s_{\bar{y}_s}, \tag{2.80}$$

where $s_{\bar{y}_s}$ is the square root of (2.79) and t is again the value of the normal deviate corresponding to the desired confidence probability if $n \geq 50$, and the value in the Student t table with $(n-1)$ degrees of freedom if $n < 50$. For the finite population correction in (2.78) and (2.79), $1 - \frac{n}{N}$ can be used instead of $1 - \frac{a}{A}$ when A is unknown. For unknown A, therefore, formulas (2.78) and (2.79) turn into

$$\mathrm{V}(\bar{y}_s) = \sigma_{\bar{y}_s}^2 = \frac{S_s^2}{a}(1 - \frac{n}{N}), \tag{2.81}$$

and

$$\mathrm{v}(\bar{y}_s) = \text{estimated}\,\sigma_{\bar{y}_s}^2 = s_{\bar{y}_s}^2 = \frac{\hat{s}_s^2}{a}(1 - \frac{n}{N}), \tag{2.82}$$

respectively.

As far as the estimation of the subtotal is concerned, no problems arise as long as the size A of the subpopulation is known. An unbiased estimator of the total is then

$$A\bar{y}_s = \frac{A}{a}\sum_{i=1}^{a} y_i, \tag{2.83}$$

and all the results in Sections 2.4 and 2.7 apply again. So formulas (2.26) and (2.29) for the variance of the total and the sample estimator of the variance of the total also apply to subpopulations, in which case they can be written as

$$\mathrm{V}(\hat{Y}_s) = \sigma_{\hat{Y}_s}^2 = \frac{A^2 S_s^2}{a}(1 - \frac{a}{A}), \tag{2.84}$$

and

$$\mathrm{v}(\hat{Y}_s) = \text{estimated}\,\sigma^2_{\hat{Y}_s} = s^2_{\hat{Y}_s} = \frac{A^2\hat{s}^2_s}{a}(1-\frac{a}{A}), \tag{2.85}$$

respectively.

However, when the size A of the subpopulation is *unknown*, special measures need to be taken because it is then not possible to use estimator (2.83). As will be illustrated in Example 2-11, not knowing the value of A introduces extra uncertainty, resulting in a loss of precision in the estimation of the subtotal of the population. The estimation of Y_s with unknown A is handled as follows. Let

$$y_i^{'} = \begin{cases} y_i, & \text{if the unit is in the subpopulation,} \\ 0, & \text{if the unit is not in the subpopulation.} \end{cases} \tag{2.86}$$

The population total of this new variable is

$$Y^{'} = \sum_{i=1}^{N} y_i^{'} = \sum_{i=1}^{A} y_i^{'} + \sum_{i=A+1}^{N} y_i^{'} = \sum_{i=1}^{A} y_i^{'} + 0 = Y_s, \tag{2.87}$$

which means that the population subtotal of y_i is equal to the population total of $y_i^{'}$. An unbiased estimator of $Y^{'} = Y_s$ is therefore obtained by multiplying the sample mean of $y_i^{'}$ with N:

$$N\bar{y}^{'} = N\frac{1}{n}\sum_{i=1}^{n} y_i^{'} = \hat{Y}_s. \tag{2.88}$$

The sampling variance of the latter total is

$$\mathrm{V}(\hat{Y}_s) = \sigma^2_{\hat{Y}_s} = \frac{N^2S^{'2}}{n}(1-\frac{n}{N}) \tag{2.89}$$

with

$$S^{'2} = \frac{1}{N-1}(\sum_{i=1}^{N} y_i^{'2} - \frac{Y_s^2}{N}). \tag{2.90}$$

A sample estimate of the variance of $\hat{Y}_s$ is

$$\mathrm{v}(\hat{Y}_s) = \text{estimated}\,\sigma^2_{\hat{Y}_s} = \frac{N^2s^{'2}}{n}(1-\frac{n}{N}), \tag{2.91}$$

with

$$s^{'2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i^{'} - \bar{y}^{'})^2. \tag{2.92}$$

The methods of this section also apply to surveys in which the frame used contains units that do not belong to the population as it has been defined.

Example 2-11. The number of children living with their parents is determined in a simple random sample of $n = 200$ families from a large population consisting of 2.6 million families, meaning that the finite population correction may be ignored. The frequency distribution of the number of children in the sample is shown in Table 2.7. First estimate the average number of children per family in the total population, including its 95% confidence interval. Next estimate the average number of children in families with at least four children, and its standard error. Finally estimate the *total* number of children and its standard error in families with at least four children, in the following two situations: 1) the total number of families with four or more children in the population is known, and 2) this total number is unknown.

Table 2.7: Number of children in a simple random sample of 200 families.

Number of children y_j	Frequency f_j	$f_j y_j$	$f_j y_j^2$
0	62	0	0
1	51	51	51
2	39	78	156
3	22	66	198
4	12	48	192
5	7	35	175
6	4	24	144
8	2	16	128
10	1	10	100
Total	200	328	1144

Answers. Let $k = 9$ denote the total number of categories in the frequency distribution of Table 2.7. Then an unbiased estimate of the average number of children per family in the population is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{j=1}^{k} f_j y_j}{\sum_{j=1}^{k} f_j} = \frac{328}{200} = 1.64.$$

An unbiased estimate of the variance in the population is

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{j=1}^{k} f_j y_j^2 - \frac{(\sum_{j=1}^{k} f_j y_j)^2}{\sum_{j=1}^{k} f_j}}{(\sum_{j=1}^{k} f_j) - 1} = \frac{1144 - \frac{(328)^2}{200}}{199} = 3.0456,$$

and the estimated standard error of the mean therefore equals

$$s_{\bar{y}} = \sqrt{\frac{\hat{s}^2}{n}} = \sqrt{\frac{3.0456}{200}} = 0.1234.$$

The 95% upper and lower confidence limits of the mean are $\bar{y} - 1.96 s_{\bar{y}} = 1.64 - (1.96)(0.1234) = 1.398$ and $\bar{y} + 1.96 s_{\bar{y}} = 1.64 + (1.96)(0.1234) = 1.882$, respectively.

Letting the index j only refer to the categories 4 through 10 in Table 2.7, an unbiased estimate of the submean of families with at least four children is

$$\bar{y}_s = \frac{1}{a}\sum_{i=1}^{a} y_i = \frac{\sum_{j=1}^{k} y_j}{\sum_{j=1}^{k} f_j} = \frac{133}{26} = 5.12.$$

An unbiased sample estimate of the variance of the latter submean equals

$$\hat{s}_s^2 = \frac{1}{a-1}\sum_{i=1}^{a}(y_i - \bar{y}_s)^2 = \frac{\sum f_j y_j^2 - \frac{(\sum f_j y_j)^2}{\sum f_j}}{(\sum f_j) - 1} = \frac{739 - \frac{(133)^2}{26}}{26-1} = 2.3462,$$

see (2.77), meaning that the sample estimate of the standard error of this submean is

$$s_{\bar{y}_s} = \sqrt{\frac{\hat{s}_s^2}{a}} = \sqrt{\frac{2.3462}{26}} = 0.3004.$$

We next show how to estimate the *subtotal* number of children in families with four or more children – and its standard error –, in the situation that the total number A of families with four or more children in the population is known, based on the sample data in Table 2.7.

When the size A of the subpopulation of families with four or more children is known, sample estimates of the subtotal and its standard error are obtained with (2.83) and (2.85), yielding $(A)(5.12)$ and $(A)(0.3004)$ in the present situation, respectively. The coefficient of variation of this estimate of the subtotal is therefore $\frac{As_{\bar{y}_s}}{A\bar{y}_s} = \frac{0.3004}{5.12} = 0.0587$.

When A is unknown, an estimate of the subtotal is calculated with (2.88) yielding

$$\hat{Y}_s = N\bar{y}' = N\frac{1}{n}\sum_{i=1}^{n} y_i' = (2.6)(\frac{1}{200})(133) = 1.729\,\text{million}.$$

The standard error of the latter subtotal is found with (2.91), which requires the calculation of (2.92):

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y' - \bar{y}')^2 = \frac{\sum_{j=1}^{k} f_j y_j'^2 - \frac{(\sum_{j=1}^{k} f_j y_j')^2}{\sum_{j=1}^{k} f_j}}{(\sum_{j=1}^{k} f_j) - 1} = \frac{739 - \frac{133^2}{200}}{199} = 3.269.$$

Note that the index j in the latter formula now again runs over *all nine categories* of the variable in Table 2.7. But the values of y_j corresponding to the categories 0 through 3 (not being part of the subpopulation of families of four children or more) are all set equal to zero in this case. The standard error of the subtotal therefore equals

$$s_{\hat{Y}_s} = \sqrt{\frac{N^2 s'^2}{n}} = \frac{N s'}{\sqrt{n}} = \frac{2.6\sqrt{3.269}}{\sqrt{200}} = 0.333.$$

When A is unknown, the coefficient of variation of the estimate equals $\frac{s_{\hat{Y}_s}}{\hat{Y}_s} = \frac{0.333}{3.269} = 0.19$. This is a more than three-fold increase compared with the coefficient of variation of 0.0587 when the size A of the population is known, which illustrates the increased uncertainty that is introduced in the estimation of subtotals when the size of the subpopulation of interest is not known.

We end this section by discussing how to estimate sample size n in simple random sampling when precision requirements are not only imposed on parameter estimates of the population of interest, but also on parameter estimates of a subpopulation. We again distinguish two situations: one where A – the size of the subpopulation – is known, and the other where A is unknown.

For an absolute error of the submean no larger than $d = |\bar{Y}_s - \bar{y}_s|$, the minimum sample size is determined by

$$d = t\sigma_{\hat{Y}_s}, \tag{2.93}$$

where t is the value of the normal deviate corresponding to the desired confidence probability, as before. When A is known, formula (2.78) for $\sigma_{\hat{Y}_s}$ can be substituted in (2.93), yielding

$$d = t\sqrt{\frac{S_s^2}{a}(1 - \frac{a}{A})}. \tag{2.94}$$

Solving (2.94) with respect to a gives

$$a = \frac{t^2 S_s^2}{d^2 + \frac{1}{A}t^2 S_s^2}. \tag{2.95}$$

The minimal sample size n is then obtained from

$$n = \frac{aN}{A}. \tag{2.96}$$

When A is unknown, formula (2.81) is substituted in (2.93) which results in

$$d = t\sqrt{\frac{S_s^2}{a}(1 - \frac{n}{N})}. \tag{2.97}$$

In this case we have one equation with two unknowns: a and n. This is solved by making a guess, say A_g, about the size of the subpopulation, and then substituting the expected value of $a = \frac{A_g n}{N}$ in (2.96). Solving the result with respect to n yields

$$n = \frac{t^2 S_s^2 N}{A_g d^2 + t^2 S_s^2}. \tag{2.98}$$

When in doubt, a value for A_g should be chosen that is as small as reasonably possible, since the sample size n will then be on the safe side.

Similar derivations apply for the derivation of minimal sample size in the estimation of subtotals in simple random sampling.

Chapter 3

Stratified random sampling

3.1 Introduction

Simple random sampling as discussed in the previous chapter is not necessarily the most efficient sampling strategy. One possible way to improve precision in the parameter estimates of the population – and thus to reduce the required sample size for their estimation – is to divide the population of size N into a number of mutually exclusive sub-populations of sizes $N_1, N_2, \ldots, N_L$ such that

$$N = N_1 + N_2 + \cdots + N_L,$$

and then to apply simple random sampling to each of these L sub-populations separately. These mutually exclusive sub-populations (e.g., males and females) are called *strata*. The sample sizes in the strata are denoted by $n_1, n_2, \ldots, n_L$, respectively. This sampling procedure is called *stratified random sampling*.

Generally, with stratified random sampling considerable gains in precision of the estimates or considerable reduction in costs can be obtained when the population variance of the variable of interest is (much) smaller within each stratum than in the total population. Examples of such situations will be provided below. Since most guidelines can be given for this situation, the largest part of this chapter is devoted to this case.

However, other important reasons for using stratification can also be:

- The population of interest is registered in more than one sampling frame. If these frames are located in different places one is almost forced to draw separate samples from the corresponding parts of the population.
- The method of observation or of sampling or of estimation must be performed differently in different parts of the population.
- If the survey requires precise estimates for certain subdivisions of the total population, it is advisable to treat these subdivisions as separate strata.

In Sections 3.2 through 3.9 we will first assume that the number and types of strata have already been decided upon and constructed. The problem of how to construct the strata and of how many strata there should be is taken up in Section 3.10.

3.2 Properties of the parameter estimates

Before embarking on the theory of stratified random sampling we first establish the notation shown in Table 3.1.

Table 3.1: Notation for stratum h; h denotes the stratum and i the unit within the stratum.

N_h	total number of units
n_h	number of units in sample
y_{hi}	value for the ith unit
$W_h = \frac{N_h}{N}$	stratum weight
$f_h = \frac{n_h}{N_h}$	sampling fraction in stratum
$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{ih}}{N_h}$	true mean
$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{ih}}{n_h}$	sample mean
$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{ih} - \bar{Y}_h)^2}{N_h - 1}$	true variance

Let L denote the number of strata, and let $\bar{y}_h$ be an unbiased estimator of $\bar{Y}_h$ in every stratum. Then,

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \bar{y}_h, \tag{3.1}$$

is an unbiased estimator of the population mean $\bar{Y}$, see Cochran (1977, p.91) for the proof. If $\frac{n_h}{n} = \frac{N_h}{N}$, i.e. if $\frac{n_h}{N_h} = \frac{h}{N}$ meaning that $f_h = f$ in every stratum, then the sampling fraction is the same in all strata. This special type of stratification is known as stratification with *proportional allocation* of the n_h.

If the samples are drawn independently in different strata, then

$$\mathrm{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h), \tag{3.2}$$

where $V(\bar{y}_h)$ is the variance of $\bar{y}_h$ over repeated samples from stratum h.

Moreover, for stratified random sampling, the variance of the estimate $\bar{y}_{st}$ is

$$\mathrm{V}(\bar{y}_{st}) = \sigma^2_{\bar{y}_{st}} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h). \tag{3.3}$$

There are a number of special cases of (3.3). If the sampling fractions n_h/N_h are negligible in all strata, then

$$\mathrm{V}(\bar{y}_{st}) = \sigma^2_{\bar{y}_{st}} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}. \tag{3.4}$$

With proportional allocation we may substitute

$$n_h = \frac{nN_h}{N}$$

in (3.3) yielding

$$\mathrm{V}(\bar{y}_{st}) = \sigma^2_{\bar{y}_{st}} = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2. \tag{3.5}$$

Finally, if sampling is proportional and the variances in all strata have the same value, S_w^2, (3.5) further simplifies into

$$\mathrm{V}(\bar{y}_{st}) = \sigma^2_{\bar{y}_{st}} = \frac{S_w^2}{n}(1-f). \tag{3.6}$$

If $\hat{Y}_{st} = N\bar{y}_{st}$ is the estimate of the population total Y, then

$$\mathrm{V}(\hat{Y}_{st}) = \sigma^2_{\hat{Y}_{st}} = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} \tag{3.7}$$

is the variance of the sampling distribution of the total $\hat{Y}_{st}$ of the stratified population.

Example 3-1. Given is a hypothetical population consisting of $N = 7$ units with values 3, 6, 6, 8 ,15, 18, and 28. For this population the total, the mean, and the variance are $Y = \sum_{i=1}^{N} y_i = 84$, $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{84}{7} = 12$, and $S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1} = \frac{470}{6} = 78\frac{1}{3}$, respectively. Moreover, if we draw a simple random sample of $n = 5$ units from this population, then it follows from (2.26) that the sampling variance of the total $\hat{Y} = N\bar{y}$ equals

$$\mathrm{V}(\hat{Y}) = \sigma^2_{\hat{Y}} = \frac{N^2S^2}{n}(1-f) = \frac{(7^2)(78\frac{1}{3})}{5}(1-\frac{5}{7}) = 219\frac{1}{3}.$$

Now suppose that we divide this population into the two strata shown in Table 3.2.

Table 3.2: Population of $N = 7$ units divided into two strata of $N_1 = 4$ and $N_2 = 3$ units.

Stratum	Values				Total
1	3	6	6	15	30
2	8	18	28		54

We first of all note that the population totals in the two strata are $Y_1 = 30$ and $Y_2 = 54$, respectively, and that the corresponding population means are $\bar{Y}_1 = \frac{30}{4} = 7.5$ and $\bar{Y}_2 = \frac{54}{3} = 18$, respectively, so that the grand mean in the population is $\sum_{h=1}^{L} W_h\bar{Y}_h = (\frac{4}{7})(7.5) + (\frac{3}{7})(18) = 12 = \bar{Y}$. The variances of the units within these two strata of the population are $S_1^2 = \frac{\sum_{i=1}^{N_1}(y_{i1} - \bar{Y}_1)^2}{N_1 - 1} = \frac{81}{3} = 27$ and $S_2^2 = \frac{\sum_{i=1}^{N_2}(y_{i2} - \bar{Y}_2)^2}{N_2 - 1} = \frac{200}{2} = 100$, respectively.

We now draw a simple random sample of $n_1 = 3$ units from stratum 1, and a simple random sample of $n_2 = 2$ units from stratum 2. Since there are $M = \Pi_{h=1}^{L} \frac{N_h!}{n_h!(N_h-n_h)!}$ possible samples in stratified random sampling without replacement, we have a total of $M = (\frac{N_1!}{n_1!(N_1-n_1)!})(\frac{N_2!}{n_2!(N_2-n_2)!}) = (\frac{4!}{3!1!})(\frac{3!}{2!1!}) = (4)(3) = 12$ possible random samples in the present situation. All these 12 possible random samples are given in Table 3.3.

Table 3.3: All possible simple random samples of size $n_1 = 3$ and $n_2 = 2$ without replacement from the two population strata in Table 3.2, and their statistics.

Sample number	Stratum 1			Stratum 2		$\bar{y}_1$	$\bar{y}_2$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}$
1	3	6	6	8	18	5	13	20	39	59
2	3	6	6	8	28	5	18	20	54	74
3	3	6	6	18	18	5	23	20	69	89
4	3	6	15	8	28	8	13	32	39	71
5	3	6	15	8	28	8	18	32	54	86
6	3	6	15	18	28	8	23	32	69	101
7	3	6	15	8	18	8	13	32	39	71
8	3	6	15	8	28	8	18	32	54	86
9	3	6	15	18	28	8	23	32	69	101
10	6	6	15	8	18	9	13	36	39	75
11	6	6	15	8	28	9	18	36	54	90
12	6	6	15	18	28	9	23	36	69	105
Total						90	216	360	648	1008
Expectation= Total/12						7.5	18	30	54	84

A comparison of the expectations in Table 3.3 with the population parameters shows that the estimators are indeed unbiased, since $E(\bar{y}_h) = \bar{Y}_h$ for $h = 1, 2$, $E(\hat{Y}_h) = Y_h$ for $h = 1, 2$, $E(\hat{Y}) = Y$, and also $E(\hat{\bar{Y}}) = \bar{Y}$ because $\frac{84}{7} = 12$.

Moreover, from (3.7) we obtain the variance of the sampling distribution of the stratified population total

$$\begin{aligned} \mathrm{V}(\hat{Y}_{st}) =& \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = N_1(N_1 - n_1)\frac{S_1^2}{n_1} + N_2(N_2 - n_2)\frac{S_2^2}{n_2} \\ =& 4(4-3)\frac{27}{3} + 3(3-2)\frac{100}{2} = (4)(9) + (3)(50) = 186. \end{aligned} \tag{3.8}$$

For a simple random sample of $n_1 = 2$ and $n_2 = 3$ from stratum 1 and 2 in Table 3.2, on the other hand, it follows from (3.7) that

$$\begin{aligned} \mathrm{V}(\hat{Y}_{st}) =& N_1(N_1 - n_1)\frac{S_1^2}{n_1} + N_2(N_2 - n_2)\frac{S_2^2}{n_2} \\ =& 4(4-2)\frac{27}{2} + 3(3-3)\frac{100}{3} = (8)(\frac{27}{2}) = (4)(27) = 108. \end{aligned} \tag{3.9}$$

A summary of the results for this example are given in Table 3.4. As the latter table clearly indicates, for a sample of $n = 5$ units drawn from a population of $N = 7$ units, the

Table 3.4: Sampling variances for population of $N = 7$ units using two stratified random sampling schemes.

Type of sampling	Sample size		Sampling variance of the total
simple random sample	$n = 5$		$219\frac{1}{3}$
stratified random sample	$n_1 = 3$	$n_2 = 2$	186
stratified random sample	$n_1 = 2$	$n_2 = 3$	108

precision of the estimated population total is the worst for a simple random sample, better for a stratified random sample with $n_1 = 3$ and $n_2 = 2$, and best for a stratified random sample with $n_1 = 2$ and $n_2 = 3$.

Example 3-2. We have the 1930 number of inhabitants, in thousands, of $N = 64$ large cities in the United States, see Table 3.5. The cities are arranged in two strata, the first containing the 16 largest cities, and the second the remaining 48 cities.

The total number of inhabitants in all 64 cities is to be estimated from a sample of size $n = 24$. Calculate the standard errors of the estimated total for 1) a simple random sample, 2) a stratified random sample with proportional allocation, and 3) a stratified random sample with 12 units drawn from each stratum.

Table 3.5: Population sizes of $N = 64$ large cities (in 1000's) in the United States in 1930.

Stratum			
1	2		
900	364	209	113
822	317	183	115
781	328	163	123
805	302	253	154
670	288	232	140
1238	291	260	119
573	253	201	130
634	291	147	127
578	308	292	100
487	272	164	107
442	284	143	114
451	255	169	111
459	270	139	163
464	214	170	116
400	195	150	122
366	260	143	134

Answer. The total and variance of the complete population is $Y = 19,568$ and $S^2 = 52,448$. The three estimates are denoted by $\hat{Y}$, $\hat{Y}_{st(prop)}$, and $\hat{Y}_{st(equal)}$.

1. Using (2.26) we find that – for simple random sampling – the sampling variance of the

total $\hat{Y}$ equals

$$\mathrm{V}(\hat{Y}) = \sigma^2_{\hat{Y}} = \frac{N^2 S^2}{n}(1-f) = \frac{(64^2)(52,448)}{24}\left(\frac{64-24}{24}\right) = 5,594,453,$$

with a standard error of $\sigma(\hat{Y}) = \sqrt{5,594,453} = 2365$.

2. The population variances in the two strata are $S_1^2 = 53,843$ and $S_2^2 = 5581$, respectively. Note that the population variance of the 16 largest cities in the first stratum is almost ten times as large as the population variance of the 48 cities in the second stratum. In proportional allocation we have $n_1 = 6$ and $n_2 = 18$, since $\frac{n_1}{N_1} = \frac{n_2}{N_2}$ in this case. Applying (3.7) to this situation we obtain

$$\begin{aligned}\mathrm{V}(\hat{Y}_{st(prop)}) = \sigma^2_{\hat{Y}_{st(prop)}} &= N_1(N_1-n_1)\frac{S_1^2}{n_1} + N_2(N_2-n_2)\frac{S_2^2}{n_2} \\ &= 16(16-6)\frac{53,843}{6} + 48(48-18)\frac{5581}{18} = 1,882,293.334,\end{aligned}$$

from which it follows that the standard error $\sigma(\hat{Y}_{st(prop)})$ equals $\sqrt{1,882,293.334} = 1372$ for proportional stratified random sampling.

3. Applying (3.7) to an equal allocation with $n_1 = n_2 = 12$ we find that

$$\begin{aligned}\mathrm{V}(\hat{Y}_{st(prop)}) = \sigma^2_{\hat{Y}_{st(prop)}} &= N_1(N_1-n_1)\frac{S_1^2}{n_1} + N_2(N_2-n_2)\frac{S_2^2}{n_2} \\ &= 16(16-12)\frac{53,843}{12} + 48(48-12)\frac{5581}{12} = 1,090,826.667,\end{aligned}$$

meaning that the standard error $\sigma(\hat{Y}_{st(equals)})$ is $\sqrt{1,090,826.667} = 1044$ for stratified random sampling using equal allocation.

We end example 3-2 by noting that equal sample sizes in the two strata are more precise than proportional allocation in this case, and that both stratified random sampling schemes are greatly superior to simple random sampling because the units in the two strata are much more homogeneous than the units in the complete population.

3.3 The estimated variances and confidence limits

It is clear from (2.21) in Section 2.3 that an unbiased estimator of the population variance in stratum h is

$$\hat{s}_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2, \tag{3.10}$$

if a simple random sample is taken within each stratum. This means that with stratified random sampling

$$\mathrm{v}(\bar{y}_{st}) = \text{estimated}\,\sigma^2_{\bar{y}_{st}} = \hat{s}^2_{\bar{y}_{st}} = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{\hat{s}_h^2}{n_h}, \tag{3.11}$$

is an unbiased estimator of the variance of $\bar{y}_{st}$, see Cochran (1977, p.95) for a proof.

If $\bar{y}_{st}$ is normally distributed, it follows from (3.11) that the confidence limits for the *population mean* are equal to

$$\bar{y}_{st} \pm t\hat{s}_{\bar{y}_{st}}, \tag{3.12}$$

while those of the *population total* are

$$N\bar{y}_{st} \pm tN\hat{s}_{\bar{y}_{st}}, \tag{3.13}$$

where t can be read from tables of the normal distribution.

3.4 Optimum allocation to strata

As we saw in the examples discussed in Section 3.2, for a sample of given size n it can make quite a difference for the precision of the population parameter estimates how the units in the total sample are allocated to the L strata of the population. In this section we present strategies for the optimization of the allocation of sample units to the L strata, temporarily assuming that the sample size is given and known. The more usual and general situation where the total sample size is unknown and also has to be estimated will be taken up in Section 3.6.

For an optimal allocation with given sample size, also called *Neyman allocation*, the partitioning of n over the L strata is chosen such that the variance of $\mathrm{V}(\bar{y})$ in (3.3) is minimized. It can be proved that this is achieved by taking

$$n_h = n\frac{N_h S_h}{\sum N_h S_h}, \tag{3.14}$$

for $h = 1, \ldots, L$. The latter formula implies that the sample size n_h in a stratum should be larger when the population size or the standard deviation in that stratum is larger. In other words, from large and/or heterogenous strata more units should be sampled. The value of the variance for optimum allocation with fixed n is obtained by substituting (3.14) into the general formula (3.3) for $\mathrm{V}(\bar{y}_{st})$ yielding the minimal variance equal to:

$$\mathrm{V}_{min}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}, \tag{3.15}$$

where the second term on the right represents the finite population correction.

Example 3-3. For the hypothetical data in Example 3-1, the application of (3.14) for a given sample size of $n = 5$ to the two strata yields

$$n_1 = n\frac{N_1 S_1}{N_1 S_1 + N_2 S_2} = 5\frac{4\sqrt{27}}{4\sqrt{27} + 3\sqrt{100}} = 2.05$$

for stratum 1 and

$$n_2 = n\frac{N_2 S_2}{N_1 S_1 + N_2 S_2} = 5\frac{3\sqrt{100}}{4\sqrt{27} + 3\sqrt{100}} = 2.95$$

for stratum 2. Rounding these numbers the optimal allocation is $n_1 = 2$ and $n_2 = 3$ which indeed gives the smallest variation, see also Table 3.4.

Example 3-4. For the data in Example 3-2, the application of (3.14) for a given sample size of $n = 24$ to the two strata results in

$$n_1 = n\frac{N_1S_1}{N_1S_1 + N_2S_2} = 24\frac{16\sqrt{53,843}}{16\sqrt{53,843} + 48\sqrt{5581}} = 12.21$$

for stratum 1 and

$$n_2 = n\frac{N_2S_2}{N_1S_1 + N_2S_2} = 24\frac{48\sqrt{5581}}{16\sqrt{53,843} + 48\sqrt{5581}} = 11.79$$

for stratum 2, showing – after rounding – that an equal allocation of $n_1 = n_2 = 12$ is indeed the optimal solution.

Example 3-5. In order to estimate the mean turnover of companies in inland shipping a sample of 300 companies is drawn. The population is stratified by the number of ships owned by each company. The numbers of companies in these strata are known, and estimates of the standard deviations of the turnover are also available, see Table 3.6. Find the optimal allocation of the sample of 300 companies to the three strata, and compare the precision of the latter Neyman allocation with a proportional allocation.

Table 3.6: Turnover of inland shipping companies in three strata.

Stratum h	Number of ships per company	Number of companies N_h	Standard deviation S_h of the turnover (in 1000)	N_hS_h	Optimal number n_h
1	1	850	6	5,100	113
2	2-4	280	15	4,200	93
3	≥ 5	60	70	4,200	93
Total		1,190	-	13,500	299

Answer. In order to find the optimal allocation of given $n = 300$ for the data in Table 3.6, we use (3.14) and obtain $n_1 = 300\frac{5,100}{13,500} = 113$, $n_2 = 300\frac{4,200}{13,500} = 93$, and $n_3 = 300\frac{4,200}{13,500} = 93$, as shown in the last column of Table 3.6. We see that there is a problem with the optimal n_3 because the optimal allocation would require a larger sample than the total number of companies in this stratum of the population. This situation is known as the problem of *over-sampling*. The solution is to allocate all $N_3 = 60$ companies to n_3, and then to reapply proportional allocation to the remaining 240 companies in the sample, and 1130 companies in the population. Since $\sum_{h=1}^{2} N_hS_h = 9,300$ in this case, we obtain sample sizes of $n_1 = 240\frac{5,100}{9,300} = 132$ and $n_1 = 240\frac{4,200}{9,300} = 108$ companies for the first two strata. The

variance of the estimated mean turnover in this sample can be calculated with (3.3), yielding

$$\mathrm{V}(\bar{y}_{st}) = \left(\frac{850}{1190}\right)^2 \left(\frac{6^2}{132}\right)\left(1-\frac{132}{850}\right) + \left(\frac{280}{1190}\right)^2 \left(\frac{15^2}{108}\right)\left(1-\frac{108}{280}\right) + \left(\frac{60}{1190}\right)^2 \left(\frac{70^2}{60}\right)\left(1-\frac{60}{60}\right) = 0.1884.$$

Note that stratum 3 does not contribute to this variance because the complete stratum of the population is sampled.

If we use proportional allocation, on the other hand, the n_h should satisfy $\frac{n_h}{n} = \frac{N_h}{N}$ for $h = 1, \ldots, L$, from which it follows that $n_h = n\frac{N_h}{N}$ for $h = 1, \ldots, L$. For the data in Table 3.6 this yields $n_1 = 300\frac{850}{1190} = 214$, $n_2 = 300\frac{280}{1190} = 71$, and $n_3 = 300\frac{60}{1190} = 15$. The variance of the estimated mean turnover in this sample can be calculated with (3.5), yielding

$$\begin{aligned}\mathrm{V}(\bar{y}_{st}) =& \frac{1-\frac{n}{N}}{n}\sum_{h=1}^{L} W_h S_h^2 = \frac{1-\frac{n}{N}}{nN}\sum_{h=1}^{L} N_h S_h^2 \\ =& \frac{1-\frac{300}{1190}}{(300)(1190)}\left((850)(6^2)+(280)(15^2)+(60)(70^2)\right) = 0.8120.\end{aligned}$$

Comparing the variance for proportional allocation with that for Neyman allocation, we see that the variance for proportional allocation is more than four times larger than that for Neyman allocation. This is caused by the very large differences between the standard deviations in the three strata.

So far we assumed that the total sample size is given or fixed. Sometimes this is not the best way to proceed. Generally, sampling methods are designed to obtain results that are *as accurate as possible for costs that are as small as possible*. This cost aspect can be taken into account into the calculations by not assuming a given sample size but a given budget. If the variable costs are different from stratum to stratum, then the simplest cost function that can be considered is the linear cost function

$$C = \sum_{h=1}^{L} n_h c_h, \tag{3.16}$$

where c_h is the variable cost of one unit in stratum h. In this situation the variance (3.3) of the estimated mean $\bar{y}_{st}$ is a minimum for a specified total cost C, and the cost is a minimum for a specified variance $\mathrm{V}(\bar{y}_{st})$ when $n_h \propto W_h S_h/\sqrt{c_h}$ (where $\propto$ means "proportional to").

Proof. Let a_h and b_h be two sets of arbitrary numbers ($h = 1, \ldots, L$). Then the Cauchy-Schwartz inequality states that

$$(\sum a_h^2)(\sum b_h^2) \geq (\sum a_h b_h)^2, \tag{3.17}$$

with equality if and only if $a_h = \lambda b_h$ for all $h = 1, \ldots, L$ where λ is some positive number. The product on the left of (3.17) is therefore globally minimized by choosing a_h proportional

to b_h for all $h = 1, \ldots, L$. We now use the Cauchy-Schwartz inequality to minimize the product $\mathrm{V}(\bar{y}_{st})C$ with respect to n_h, and first note that

$$\mathrm{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{N_h}. \tag{3.18}$$

Since the last term $\sum_{h=1}^{L} W_h^2 \frac{S_h^2}{N_h}$ in (3.18) is a constant with respect to n_h, the minimization of $\mathrm{V}(\bar{y}_{st})C$ is equivalent to the minimization of $\left(\sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}\right) C$. Now define $a_h = W_h \frac{S_h}{\sqrt{n_h}}$ and $b_h = \sqrt{n_h c_h}$, then it follows from (3.17) that

$$\left(\sum W_h^2 \frac{S_h^2}{n_h}\right)\left(\sum n_h c_h\right) \geq \left(\sum W_h S_h \sqrt{c_h}\right)^2, \tag{3.19}$$

and the global minimum $(\sum W_h S_h \sqrt{c_h})^2$ is obtained by choosing $\sqrt{n_h c_h} = \lambda W_h \frac{S_h}{\sqrt{n_h}}$, which can also be written as $n_h c_h = \lambda^2 W_h^2 \frac{S_h^2}{n_h}$ and thus as $n_h^2 c_h = \lambda^2 W_h^2 S_h^2$ from which it follows that

$$n_h = \lambda \frac{W_h S_h}{\sqrt{c_h}}, \tag{3.20}$$

for $h = 1, \ldots, L$, and for some proportionality constant λ. This ends the proof.

It follows from (3.20) that

$$n = \sum_{h=1}^{L} n_h = \lambda \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}},$$

and λ can therefore be written as

$$\lambda = \frac{n}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}. \tag{3.21}$$

Substitution of (3.21) in (3.20) yields

$$n_h = \frac{n W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}, \tag{3.22}$$

and therefore

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}.$$

The fact that n_h should be proportional to $W_h S_h / \sqrt{c_h}$ implies that for a given stratum a larger random sample should be taken if

1. the stratum is larger;
2. the variability within the stratum is larger;
3. sampling is cheaper in the stratum.

If the total cost C of the survey is fixed then the optimal total sample size is found by substitution of (3.22) in (3.16) yielding

$$n = \frac{C \sum_{h=1}^{L} N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}. \tag{3.23}$$

On the other hand, if the variance $\mathrm{V}(\bar{y}_{st})$ is fixed, then the optimal total sample size is found by substitution of (3.22) in (3.3) yielding, after quite some algebra,

$$n = \frac{(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}) \sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}{\mathrm{V}(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}, \tag{3.24}$$

with $W_h = N_h/N$.

Example 3-6. In Example 3-5 no account was taken for possible differences in observation costs between the three strata. For companies owning only one ship, information has to be obtained from the captain of the ship who will often be traveling, implicating high costs for the data collection process. On the other hand, companies with more ships have an office on shore where the required information can be obtained, implicating lower costs. Assume that the costs of an interview for a one ship company are 30 euro's while those for a company with more than one ship are 15 euro's. Also assume that the total budget for variable costs in the survey is 6480,- euro's. Find the optimal allocation with fixed budget.

Table 3.7: Turnover of inland shipping companies in three strata, including variable costs.

Stratum h	N_h	S_h (in 1000 euro)	c_h	$N_h S_h \sqrt{c_h}$	$n_h c_h$	n_h
1	850	6	30	27,933.85	3000	100
2	280	15	15	16,266.53	1740	116
3	60	70	15	16,266.53	1740	116
Total	1,190	-	-	60,466.91	6480	332

Answer. The optimal allocation with fixed budget is found using the numbers given in Table 3.7. Given the budget of 6480 euro's, the optimal sample size is obtained from (3.23) yielding

$$n = \frac{C \sum_{h=1}^{L} N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}} = \frac{(6480)(3100)}{60,455.91} = 332.215.$$

The stratum sample sizes are then obtained from (3.22)

$$n_h = \frac{n W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}} = \frac{n N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} = \frac{332 N_h S_h / \sqrt{c_h}}{3100},$$

giving $n_1 = 99.72$, $n_2 = 116.14$, and $n_3 = 116.14$, or rounded $n_1 = 100$, $n_2 = 116$, and $n_3 = 116$, see the last column in Table 3.7.

Again we have over-sampling in the third stratum. So we allocate all 60 companies in the third stratum to n_1 which costs $(60)(15) = 900$ euro's, and optimally reallocate the remaining $6480 - 900 = 5580$ euro's to the first two strata. This is achieved by recomputing (3.23) with $C = 5580$ and summing over the first two strata only:

$$n = \frac{C \sum_{h=1}^{2} N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{2} N_h S_h \sqrt{c_h}} = \frac{(5580)(2015.56)}{44,200.38} = 254.45,$$

or $n = 254$ rounded. The sample sizes are again obtained from (3.22)

$$n_h = \frac{n N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} = \frac{254 N_h S_h / \sqrt{c_h}}{2015.56},$$

giving $n_1 = 117.34$ and $n_2 = 136.66$, or $n_1 = 117$ and $n_2 = 137$ rounded. The variance of the estimated mean turnover in this stratified sample of $n = 117 + 137 + 60 = 314$ companies is

$$\mathrm{V}(\bar{y}_{st}) = \left(\frac{850}{1190}\right)^2 \left(\frac{6^2}{117}\right) \left(1 - \frac{117}{850}\right) + \left(\frac{280}{1190}\right)^2 \left(\frac{15^2}{137}\right) \left(1 - \frac{137}{280}\right) + \left(\frac{60}{1190}\right)^2 \left(\frac{70^2}{60}\right) \left(1 - \frac{60}{60}\right) = 0.1818,$$

see (3.3). Note that stratum 3 again does not contribute to this variance because the complete stratum of the population is sampled.

It may be noted that if $c_h = c$, i.e., if the cost per unit is the same in all strata, then the total cost equals $C = cn$, and optimum allocation for fixed cost reduces to optimum allocation for fixed sample size, see (3.14). It is also very interesting to note that (3.24) then can be written as

$$n = \frac{(\sum_{h=1}^{L} W_h S_h \sqrt{c}) \sum_{h=1}^{L} W_h S_h / \sqrt{c}}{\mathrm{V}(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{\mathrm{V}(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}, \tag{3.25}$$

which is formula (3.33) for the minimal total sample size required under optimum allocation, as we will see in Section 3.6!

3.5 Precision gains of stratified versus simple random sampling

In this section a comparison is made between simple random sampling and stratified random sampling with proportional and optimum allocation. If we let V_{ran}, V_{prop}, and V_{opt} denote the corresponding variances of the estimated population means, and if terms $1/N_h$ can be ignored, it can be proven that

$$V_{opt} \leq V_{prop} \leq V_{ran}, \tag{3.26}$$

where the optimum allocation is for fixed n, that is, with n_h proportional to $N_h S_h$, see Cochran (1977, pp.99). This means that the precision of optimum allocation for fixed n is always higher than or equal to proportional stratification, and that the precision of proportional stratification is always higher than or equal to simple random sampling.

Ideally the value of y itself should be used for stratification, the quantity to be measured, because this would guarantee no overlap between strata, and the variance within strata would be much smaller than the overall-variance, especially when there are many strata. This is illustrated by Example 3-2 in Section 3.2 where the population consisted of the size of 64 cities in 1930, stratified by size. Although there were only two strata, proportional stratification reduced the standard error from 2365 to 1372, while the optimum stratification under Neyman allocation with $n_1 = n_2 = 12$ produced a further reduction to 1044.

In practice this is not possible of course, but it implies that large gains in precision can be obtained if the following conditions are satisfied:

1. the population is composed of institutions varying widely in size;
2. the principal variables to be measured are closely related to the sizes of the institutions;
3. a good measure of size is available for setting up the strata.

All this implies that we should preferably stratify on an auxiliary variable *that is highly correlated with* y. Such a variable is also known as a *stratification variable*.

The larger the differences between the strata means or totals are, the more proportional allocation will outperform simple random sampling in terms of precision. It is therefore important to construct the strata in such a way that the differences between the means or totals are as large as possible. This is achieved by collecting units with small y values in one stratum, and units with large y values in another stratum. The effect will be that the differences within strata will be relatively small, thus splitting a heterogeneous population into homogeneous subgroups.

Whereas proportional allocation only takes differences between strata means or totals into account, in optimal allocation the differences in standard errors and costs between strata are also taken into consideration. This implies that – compared to proportional allocation – Neyman allocation yields population parameter estimates that are more precise when the differences between the S_h are larger. Optimal allocation for given budget, on the other hand, is more precise than Neyman allocation when the differences of the variable costs between strata are larger.

3.6 Estimation of sample size with continuous data

Formulas for the determination of sample size under an estimated optimum allocation were given in Section 3.4. In this section formulas are presented for any allocation. It is assumed that the estimate has a specified variance V. If, instead, the margin of error d has been specified (see Section 2.8), we use $V = \left(\frac{d}{t}\right)^2$, where t is again the normal deviate corresponding to the allowable probability that the error will exceed the desired margin d.

We start with the situation where we want to estimate the *population mean* $\bar{Y}$. Let s_h denote the estimate of S_h and let $n_h = w_h n$ where the w_h have been chosen. Then it follows

from formula (3.3) for the variance of the sampling distribution of the mean $\bar{y}_{st}$ that

$$\begin{aligned}\mathrm{V}(\bar{y}_{st}) =& \sigma^2_{\bar{y}_{st}} = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h}(1 - \frac{n_h}{N_h}) = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h} - \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h}\frac{n_h}{N_h} \\ =& \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{w_h n} - \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{N_h} = \frac{1}{n}\sum_{h=1}^{L} W_h^2 \frac{s_h^2}{w_h} - \sum_{h=1}^{L} \frac{N_h^2}{N^2}\frac{s_h^2}{N_h} \\ =& \frac{1}{n}\sum_{h=1}^{L} W_h^2 \frac{s_h^2}{w_h} - \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N} s_h^2 = \frac{1}{n}\sum_{h=1}^{L} W_h^2 \frac{s_h^2}{w_h} - \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2.\end{aligned}$$

Thus, the anticipated sampling variance $\mathrm{V}(\bar{y}_{st})$ is

$$\mathrm{V}(\bar{y}_{st}) = \frac{1}{n}\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{w_h} - \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2, \tag{3.27}$$

with $W_h = N_h/N$ and $w_h = n_h/n$.

Proceeding just as in Section 2.8, if we let $d = |\mu - \bar{y}_{st}|$ be the absolute sampling error d that we are willing to tolerate with some predefined probability, then this means that $d \leq t\sigma_{\bar{y}_{st}}$, from which the minimal sample size can be determined by solving

$$d = t\sqrt{\frac{1}{n}\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{w_h} - \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2} \tag{3.28}$$

with respect to n. This yields the following general formula for the minimal sample size n

$$n = \frac{\sum \frac{W_h^2 s_h^2}{w_h}}{V + \frac{1}{N}\sum W_h s_h^2}, \tag{3.29}$$

where $V = \left(\frac{d}{t}\right)^2$. If the finite population correction in (3.29) is ignored, we have as a first approximation

$$n_0 = \frac{1}{V}\sum \frac{W_h^2 s_h^2}{w_h}. \tag{3.30}$$

If n_0/N is not negligible, the sample size may be calculated as

$$n = \frac{n_0}{1 + \frac{1}{NV}\sum W_h s_h^2}. \tag{3.31}$$

When optimum allocation for fixed n is required where w_h should be proportional to $W_h s_h$, we know that

$$n_h = n\frac{N_h S_h}{\sum N_h S_h},$$

see (3.14), from which it follows that

$$w_h = \frac{n_h}{n} = \frac{N_h S_h}{\sum N_h S_h} = \frac{W_h S_h}{\sum W_h S_h}. \tag{3.32}$$

Substitution of (3.32) in (3.29) yields the sample size formula under optimum allocation:

$$n = \frac{(\sum W_h s_h)^2}{V + \frac{1}{N}\sum W_h s_h^2}. \tag{3.33}$$

Substitution of $w_h = W_h = N_h/N$ for proportional allocation in (3.29) shows that

$$n_0 = \frac{\sum W_h s_h^2}{V} \tag{3.34}$$

for negligible fpc and

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \tag{3.35}$$

if n_0/N is not negligible.

As a double check on these formulas it is interesting to note that when there is only one stratum in the population, i.e., when $L = 1$, we have that $N_h = N$, $W_h = N_h/N = 1$, and $w_h = n_h/n = 1$, meaning that (3.29) then simplifies to

$$\begin{aligned} n =& \frac{\sum \frac{W_h^2 s_h^2}{w_h}}{V + \frac{1}{N}\sum W_h s_h^2} = \frac{s^2}{V + \frac{1}{N}s^2} \\ =& \frac{s^2}{\frac{d^2}{t^2} + \frac{1}{N}s^2} = \frac{t^2 s^2}{d^2 + \frac{1}{N}t^2 s^2}, \end{aligned}$$

which is equal to the sample size formula (2.51) for simple random sampling, as expected.

When working with a *relative error* in estimating the mean of a population, d should be replaced by $d\bar{Y}$ with $0 < d < 1$. In this case V is therefore defined as $V = \left(\frac{d\bar{Y}}{t}\right)^2$ in formulas (3.29), (3.30), (3.31), (3.33), and (3.34).

When we want to estimate the *population total* $Y = N\bar{Y}$, then it follows from formula (3.7) for the variance of the sampling distribution of the total that

$$\begin{aligned} \mathrm{V}(\hat{Y}_{st}) =& \sigma^2_{\hat{Y}_{st}} = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} = \sum_{h=1}^{L} N_h^2 \frac{s_h^2}{n_h} - \sum_{h=1}^{L} N_h n_h \frac{s_h^2}{n_h} \\ =& \sum_{h=1}^{L} \frac{N_h^2 s_h^2}{w_h n} - \sum_{h=1}^{L} N_h s_h^2 = \frac{1}{n}\sum_{h=1}^{L} \frac{N_h^2 s_h^2}{w_h} - \sum_{h=1}^{L} N_h s_h^2. \end{aligned}$$

Letting $d = |Y - \hat{Y}_{st}| = |N\mu - N\bar{y}_{st}|$, we have that $d \leq t\sigma_{\hat{Y}_{st}}$ from which it follows that

$$d = t\sqrt{\frac{1}{n}\sum_{h=1}^{L} \frac{N_h^2 s_h^2}{w_h} - \sum_{h=1}^{L} N_h s_h^2}. \tag{3.36}$$

Solving (3.36) with respect to n, we obtain the following general formula for sample size:

$$n = \frac{\sum \frac{N_h^2 s_h^2}{w_h}}{V + \sum N_h s_h^2}, \tag{3.37}$$

with $V = \left(\frac{d}{t}\right)^2$. For Neyman allocation the formula is obtained by substitution of (3.32) in (3.37) yielding

$$n_0 = \frac{(\sum N_h s_h)^2}{V}, \tag{3.38}$$

if the fpc is negligible, and

$$n = \frac{n_0}{1 + \frac{1}{V}\sum N_h s_h^2}, \tag{3.39}$$

if it is not. For proportional allocation we substitute $w_h = N_h/N$ in (3.37) and obtain

$$n_0 = \frac{N}{V}\sum N_h s_h^2, \tag{3.40}$$

if the fpc is negligible, and

$$n = \frac{n_0}{1 + \frac{n_0}{N}}. \tag{3.41}$$

if it is not.

As a double check on these formulas it is again interesting to note that when there is only one stratum in the population, i.e., when $L = 1$, we have that $N_h = N$, $W_h = N_h/N = 1$, and $w_h = n_h/n = 1$, meaning that (3.37) then simplifies to

$$\begin{aligned} n =& \frac{\sum \frac{N_h^2 s_h^2}{w_h}}{V + \sum N_h s_h^2} = \frac{N^2 s^2}{V + N s^2}, \\ =& \frac{N^2 s^2}{\frac{d^2}{t^2} + N s^2} = \frac{t^2 s^2}{\frac{d^2}{N^2} + \frac{t^2 s^2}{N}}, \end{aligned}$$

which is equal to the sample size formula (2.61) for simple random sampling, as it should be.

If we finally use a *relative* sampling error dY with $0 < d < 1$ in order to obtain an estimate of the population total Y, then V in (3.37), (3.38), (3.39), and (3.40) is defined as

$$V = \left(\frac{dY}{t}\right)^2 = \left(\frac{dN\bar{Y}}{t}\right)^2. \tag{3.42}$$

Example 3-7. In order to estimate enrollments for the 1946-1947 academic year in a population of 196 colleges in the United States, these colleges were arranged in six strata. The first five strata were constructed by size of institution, while the sixth stratum contained colleges for women only. Estimates s_h of the S_h were computed from the results for the 1943-1944 academic year. An optimum stratification based on these s_h was used. The objective was a coefficient of variation of 5% in the estimated total enrollment. In 1943 the total enrollment for this group of 196 colleges was $56,472$. Table 3.8 shows the values of N_h, s_h, and $N_h s_h$ which were known before determining n. Find the sample size n, and the optimal allocation of the estimated sample size to the six strata.

Table 3.8: Data for estimating sample size.

Stratum	N_h	s_h	$N_h s_h$	n_h
1	13	325	4,225	9
2	18	190	3,420	7
3	26	189	4,914	10
4	42	82	3,444	7
5	73	86	6,278	13
6	24	190	4,560	10
Totals	196		26,841	56

Answer. Since the coefficient of variation (cv) of a total is $\frac{S}{N\bar{Y}} = \frac{S}{56{,}472}$, and we require a cv of 5%, the desired standard error S is $S = (0.05)(56,472) = 2823.6$ from which it follows that the required variance is $V = S^2 = 2823.6^2 = 7972716.96$. The estimated sample size under optimal allocation is then found with (3.39) yielding

$$n = \frac{(\sum N_h s_h)^2}{V + \sum N_h s_h^2} = \frac{26,841^2}{7972716.96 + 4,640,387} = 57.12.$$

Neyman allocation to the six strata then follows from (3.14)

$$n_h = n\frac{N_h s_h}{\sum N_h s_h} = 57.12\frac{N_h s_h}{26,841},$$

giving – after rounding – the numbers shown in the last column of Table 3.8.

Example 3-8. Continuing Example 2-6 in Section 2.8, suppose that we stratify the same population into two strata of $N_1 = 3$ and $N_2 = 4$ million car drivers, and that the mean number of kilometers driven in these two strata are $\bar{Y}_1 = 7000$ and $\bar{Y}_2 = 21000$, respectively. Note that the mean number of kilometers driven in the total population is still $\frac{N_1\bar{Y}_1 + N_2\bar{Y}_2}{N} = 15000 = \bar{Y}$, as before. Determine the required sample size for estimation of the population total under proportional and optimum allocation 1) when the standard deviations in the two strata are $S_1 = S_2 = 5000$, 2) when the standard deviations in the two strata are $S_1 = S_2 = 3000$, and 3) when the standard deviations in the two strata are $S_1 = 5000$ and $S_2 = 1000$. Use a precision of 1% in estimating the population total with a probability of 1 out of 20 of being unlucky.

Answer. With this precision and probability, $d = 0.01$ and $t = 1.96$ meaning that $V = \left(\frac{dN\bar{Y}}{t}\right)^2 = \left(\frac{(0.01)(7,000,000)(15,000)}{1.96}\right)^2$, see (3.42). Applying formula (3.40) to the situation that $S_1 = S_2 = 5000$ for proportional allocation we obtain

$$n_0 = \frac{N}{V}\sum N_h s_h^2 = \frac{7,000,000}{V}((3,000,000)(5,000^2) + (4,000,000)(5,000^2)) = 4,268,$$

with the assignment to the individual strata following from $n_h = n_0\frac{N_h}{N}$ yielding $n_1 = 1,829$

and $n_2 = 2,439$. Applying formula (3.38) for optimum allocation we again find that

$$n_0 = \frac{(\sum N_h s_h)^2}{V} = \frac{((3,000,000)(5,000) + (4,000,000)(5,000))^2}{V} = 4,268,$$

with the assignment to the individual strata following from (3.14), and yielding $n_1 = 1,829$ and $n_2 = 2,439$. Note that these results are exactly the same as for simple random sampling, see Example 2-6.

When $S_1 = S_2 = 3,000$, we have for proportional allocation that

$$n_0 = \frac{N}{V}\sum N_h s_h^2 = \frac{7,000,000}{V}((3,000,000)(3,000^2) + (4,000,000)(3,000^2)) = 1,537,$$

with the assignment to the individual strata following from $n_h = n_0\frac{N_h}{N}$ yielding $n_1 = 659$ and $n_2 = 878$, while for optimum allocation, we also obtain

$$n_0 = \frac{(\sum N_h s_h)^2}{V} = \frac{((3,000,000)(3,000) + (4,000,000)(3,000))^2}{V} = 1,537,$$

with the assignment to the individual strata following from (3.14), yielding $n_1 = 659$ and $n_2 = 878$, just as with proportional allocation. This shows that stratified random sampling is more precise than simple random sampling, as long as the standard deviations in the population strata are – on the whole – smaller than the standard deviation in the total population. It also shows that proportional and optimum allocation have the same precision as long as the standard deviations in the strata are all equal.

When $S_1 = 5,000$ and $S_2 = 1,000$, on the other hand, we have for proportional allocation that

$$n_0 = \frac{N}{V}\sum N_h s_h^2 = \frac{7,000,000}{V}((3,000,000)(5,000^2) + (4,000,000)(1,000^2)) = 1,927,$$

with the assignment to the individual strata following from $n_h = n_0\frac{N_h}{N}$ yielding $n_1 = 826$ and $n_2 = 1101$, and for optimum allocation

$$n_0 = \frac{(\sum N_h s_h)^2}{V} = \frac{((3,000,000)(5,000) + (4,000,000)(1,000))^2}{V} = 1,258,$$

with the assignment to the individual strata following from (3.14), yielding $n_1 = 993$ and $n_2 = 265$.

Summarizing, this example confirms that

- For given precision, stratified random sampling with proportional and/or optimum allocation always requires a smaller sample size than simple random sampling, unless the standard deviations in the population strata are all equal to the standard deviation of the total population;

- For given precision, proportional and optimum allocation require the same sample size when the standard deviations in the population strata are the same. When the standard deviations in the population strata are different, however, the required sample size for optimum allocation is always smaller than that for proportional allocation.

3.7 Stratified sampling for proportions

If we want to estimate the proportion of units in the population that fall into some defined class C, the ideal stratification is obtained by subdividing the population into two strata: the first stratum containing all units that belong to class C, and the second stratum containing the rest of the population. If this can not be achieved, we try to construct strata such that the proportion in class C varies as much as possible from stratum to stratum.

Letting

$$P_h = \frac{A_h}{N_h} \text{ and } p_h = \frac{a_h}{n_h}$$

denote the proportions of units in class C in the hth stratum and in the sample from that stratum, respectively, the unbiased estimate of the proportion in the whole population appropriate to stratified random sampling is

$$p_{st} = \sum \frac{N_h p_h}{N}, \tag{3.43}$$

and the variance of p_{st} is

$$\mathrm{V}(p_{st}) = \sigma^2_{p_{st}} = \frac{1}{N^2} \sum \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}, \tag{3.44}$$

see Cochran (1977, p.108) for a proof.

In almost all applications the terms $1/N_h$ will be negligible, even if the finite population is not negligible, in which case (3.44) simplifies into

$$\mathrm{V}(p_{st}) = \sigma^2_{p_{st}} = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{P_h Q_h}{n_h} = \sum \frac{W_h^2 P_h Q_h}{n_h} (1 - f_h). \tag{3.45}$$

When the fpc can also be ignored we obtain

$$\mathrm{V}(p_{st}) = \sigma^2_{p_{st}} = \sum \frac{W_h^2 P_h Q_h}{n_h}. \tag{3.46}$$

With proportional allocation where $n_h = n\frac{N_h}{N}$ for $h = 1, \ldots, L$, the variance of the sampling distribution of p_{st} is

$$\mathrm{V}(p_{st}) = \sigma^2_{p_{st}} = \frac{N - n}{N} \frac{1}{nN} \sum \frac{N_h^2 P_h Q_h}{N_h - 1} \doteq \frac{1 - f}{n} \sum W_h P_h Q_h. \tag{3.47}$$

The sample estimate of the variance is obtained by substituting $p_h q_h/(n_h - 1)$ for the unknown $P_h Q_h / n_h$ in any of the formulas above.

For Neyman allocation the variance $\mathrm{V}(p_{st})$ is minimized for *fixed sample size* by choosing

$$n_h \propto N_h \sqrt{N_h/(N_h - 1)} \sqrt{P_h Q_h} \doteq N_h \sqrt{P_h Q_h}$$

from which it follows that

$$n_h \doteq n \frac{N_h \sqrt{P_h Q_h}}{\sum N_h \sqrt{P_h Q_h}}. \tag{3.48}$$

Substitution of (3.48) in (3.45) yields – after some algebra – the minimum variance of the sampling distribution of proportions under optimum allocation with given sample size:

$$\mathrm{V}_{min}(p_{st}) = \frac{(\sum W_h\sqrt{P_hQ_h})^2}{n} - \frac{\sum W_hP_hQ_h}{N}, \tag{3.49}$$

where the second term in (3.49) is the finite population correction.

For Neyman allocation the variance $\mathrm{V}(p_{st})$ is minimized for *fixed cost* $C = \sum c_h n_h$ by taking

$$n_h \doteq n\frac{N_h\sqrt{P_hQ_h/c_h}}{\sum N_h\sqrt{P_hQ_h/c_h}}. \tag{3.50}$$

The value of n is found as in Section 3.4.

3.8 Gains in precision in stratified sampling for proportions

When the variable costs per unit are the same in all strata, the following two rules usually apply

1. the gain in precision from stratified random sampling over simple random sampling is small or modest unless the P_h vary greatly from stratum to stratum;

2. optimum allocation for fixed n hardly improves over proportional allocation if all P_h lie between 0.1 and 0.9.

Table 3.9: Relative precision of simple versus stratified random sampling for proportions.

P_h			simple $100n\mathrm{V}(p)/(1-f)$ $= 100PQ$	stratified $100n\mathrm{V}(p_{st})/(1-f)$ $= 100(\frac{1}{3})\sum P_hQ_h$	relative precision $100PQ/((\frac{1}{3})\sum P_hQ_h)$
0.4	0.5	0.6	2500	2433	103%
0.3	0.5	0.7	2500	2233	112%
0.2	0.5	0.8	2500	1900	132%
0.1	0.5	0.9	2500	1433	174%

As an illustration of the first rule consider the data in Table 3.9 where stratified random with proportional allocation is compared with simple random sampling for three strata of equal size, meaning that $W_h = \frac{1}{3}$. Four cases are included, ranging from very similar proportions (0.4, 0.5, and 0.6) in the three strata, to very different proportions (0.1, 0.5, 0.9). In all cases the proportion in the total population equals $P = 0.5$. The variances of simple random sampling (where $\mathrm{V}(p)$ is calculated from (2.36)) and of stratified random sampling (where $\mathrm{V}(p_{st})$ is calculated according to (3.47)) corresponding to these four cases are given in the second and the third column of Table 3.9, respectively. The last column, which contains

Table 3.10: Relative precision of proportional versus optimal allocation for proportions.

P_0	0.4 or 0.6	0.3 or 0.7	0.2 or 0.8	0.1 or 0.9	0.05 or 0.95
RP(%)	100.0	99.8	98.8	94.1	86.6

the relative precision of stratified over simple random sampling, indicates that the gain in precision is only large in the last two cases.

As far as the second rule is concerned, ignoring the fpc it follows from (3.49) and (3.47) that the minimum variances for optimum allocation and proportional allocation for fixed n are equal to

$$\mathrm{V}_{opt}(p_{st}) = \frac{(\sum W_h\sqrt{P_hQ_h})^2}{n} \text{ and } \mathrm{V}_{prop}(p_{st}) = \frac{\sum W_hP_hQ_h}{n}$$

respectively. The relative precision of optimal over proportional allocation is therefore

$$\frac{\mathrm{V}_{opt}(p_{st})}{\mathrm{V}_{prop}(p_{st})} = \frac{(\sum W_h\sqrt{P_hQ_h})^2}{\sum W_hP_hQ_h}. \tag{3.51}$$

Assuming that we have two strata of equal size (where $W_1 = W_2$) and fixing P_1 on 0.5, for varying values of P_2 the relative precisions obtained from (3.51) have been collected in Table 3.10. It is clear from this table that the gain in precision for optimum allocation is very limited, even for proportions as small as 0.1 (or as large as 0.9) in stratum 2. In most cases the simplicity of proportional stratification more than compensates for this minor loss in precision, at least when the variable costs are identical in all strata. It follows from Table 3.10, however, that considerable gains can be achieved with optimum allocation if the proportions are very small (or very large) *and* there are large differential costs in different strata.

3.9 Estimation of sample size for proportions

Formulas for the estimation of sample size for proportions can be derived from the more general formulas in Section 3.6. Letting V denote the desired variance in the estimate of the proportion P for the whole population, it follows from (3.47) that the formulas for proportional allocation are

$$n_0 = \frac{\sum W_hp_hq_h}{V}\,,\; n = \frac{n_0}{1+\frac{n_0}{N}}, \tag{3.52}$$

while it follows from (3.49) that the formulas for optimum allocation are

$$n_0 = \frac{(\sum W_h\sqrt{p_hq_h})^2}{V}\,,\; n = \frac{n_0}{1+\frac{1}{NV}\sum W_hp_hq_h}, \tag{3.53}$$

where n_0 is the first approximation ignoring the fpc, and n is the corrected value if the fpc is not negligible. In the latter formulas the factors $N_h/(N_h-1)$ are assumed to be (almost)

equal to one. The same formulas apply if p_h, q_h, V, et cetera, are expressed in percentages. If we use an absolute error $d = |p - P|$, then $V = \left(\frac{d}{t}\right)^2$; if a relative error dP is applied where $0 < d < 1$, then $V = \left(\frac{dP}{t}\right)^2$. Finally, if the *total number* NP in the population in class C needs to be estimated, then all variances should be multiplied by N^2.

Example 3-9. Continuing Examples 2-6 and 2-8 in Section 2.8, for the same population we again wish to estimate the percentage of car drivers wearing a seat belt, but now with the same strata as in Example 3-8. So we have $N_1 = 3$ and $N_2 = 4$ million car drivers meaning that $W_1 = \frac{3}{7}$ and $W_2 = \frac{4}{7}$. Let the proportion of car drivers wearing a seat belt in these two strata be $P_1 = 0.7$ and $P_2 = 0.875$. Note that the proportion in the total population then still is the same as in Example 2-8, since $p_{st} = W_1P_1 + W_2P_2 = (\frac{3}{7})(0.7) + (\frac{4}{7})(0.875) = 0.8$ according to (3.43). How large should the sample be when we want the proportion to lie in the range $P \pm 0.05$ except for a 1 in 20 chance (use both proportional and optimum allocation)? And what if we want a precision of $P \pm 0.01$ except for a 1 in 20 chance?

Answer. In this situation $t = 1.96$. For a precision of 5% we have that $d = 0.05$ and applying (3.52) we need a sample of

$$n_0 = \frac{\sum W_h p_h q_h}{V} = \frac{(\frac{3}{7})(0.7)(0.3) + (\frac{4}{7})(0.875)(0.125)}{(\frac{0.05}{1.96})^2} = 234$$

car drivers when using proportional allocation (with $n_1 = W_1 n_0 = 100$ and $n_2 = W_2 n_0 = 134$), and applying (3.53) we need a sample of

$$n_0 = \frac{(\sum W_h \sqrt{p_h q_h})^2}{V} = \frac{(\frac{3}{7})\sqrt{(0.7)(0.3)} + (\frac{4}{7})\sqrt{(0.875)(0.125)}}{(\frac{0.05}{1.96})^2} = 228$$

car drivers when using optimum allocation with $n_1 = 116$ and $n_2 = 112$ from (3.48).

For a precision of 1%, on the other hand, $d = 0.01$ and we obtain

$$n_0 = \frac{(\frac{3}{7})(0.7)(0.3) + (\frac{4}{7})(0.875)(0.125)}{(\frac{0.01}{1.96})^2} = 5,858$$

car drivers from (3.52) when using proportional allocation (with $n_1 = W_1 n_0 = 2,511$ and $n_2 = W_2 n_0 = 3,348$), and

$$n_0 = \frac{(\frac{3}{7})\sqrt{(0.7)(0.3)} + (\frac{4}{7})\sqrt{(0.875)(0.125)}}{(\frac{0.01}{1.96})^2} = 5,705$$

car drivers from (3.53) when using optimum allocation with $n_1 = 2,908$ and $n_2 = 2,798$ from (3.48).

In this example we see that the improvement of optimum allocation over proportional allocation is indeed quite small, as already discussed in Section 3.8. But we also note a considerable improvement in precision of stratified random sampling over simple random sampling for proportions, especially for the 1% situation (compare with Example 2-8).

3.10 Choice and construction of strata

So far, we have assumed that the type and the number of strata L in stratified random sampling are given and known. In practice, however, it is the responsibility of the researcher to construct these strata. In this section we discuss the following issues:

- What information is most useful for the construction of strata?
- What boundaries for the strata should be used?
- How many strata should be used?

The answers to these questions depend upon the purpose of the use of stratified sampling. If stratification was chosen for practical reasons, such as a decentralised sampling frame or the need for different observation methods in different parts of the population (see Section 3.1), then there is hardly any choice. Often, however, stratification is used in order to improve the precision of the estimates. In this case, it makes a difference whether we are primarily interested in the total population or in parts of that population. The requirement that the estimates must be as precise as possible for both the total population *and* its parts is usually impossible to fulfill. The optimal stratifications and allocations for these two conditions will usually be quite different, and a compromise will have to be made. When subpopulations are involved Moors and Muilwijk (1975) advise to choose these subpopulations as strata because sample data referring to subpopulations that contain only part of a stratum are relatively imprecise, as discussed in Section 2.10 and Section 3.11.

For the situation where the primary purpose is to make comparisons between different strata, we refer to Cochran (1977, Section 5A.13). Here we restrict the discussion to the situation where we require estimates to be as precise as possible for the population as a whole.

As we already mentioned in Section 3.5, concerning the choice of the stratification variable(s), we should preferably stratify on an auxiliary variable that is highly correlated with y. When the stratification variable is continuous, and when we know its frequency distribution in the population, and this frequency distribution has more classes than the number of strata, then the *cumulative f-rule* of Dalenius and Hodges (1957) is a practical method for deciding which classes should be merged. This works as follows. Compute the square root of the frequency f of each class interval of the stratification variable, and choose those limits for the strata for which the total frequency per stratum is as equal as possible. If the interval sizes are different the class frequencies must first be multiplied by the corresponding interval size before their square root is taken. By choosing equal sample sizes for the chosen number of strata the allocation is approximately optimal. See Example 3-10 for an illustration of the cumulative f-rule taken from Moors and Muilwijk (1975).

Example 3-10. We want to estimate the production in a branch of business by stratified random sampling. From a previous year we have the frequency distribution of the value of production shown in Table 3.11. Use the cumulative f-rule to derive the optimal limits of three strata from this table.

Answer. Since the interval sizes of the classes in the frequency distribution are different, we first multiply the class frequencies f with their interval sizes b, before taking their square

Table 3.11: Frequency distribution of a stratification variable, and the cumulative f-rule.

Class interval (in thousand euros)	Number of companies f	Interval size b	$\sqrt{bf}$	Cumulated $\sqrt{bf}$	Strata limits
< 100	3000	100	547	547	
$100 < 250$	2000	150	547	1094	< 250
$250 < 500$	1000	250	500	1594	
$500 < 1000$	400	500	447	2041	$250 < 1000$
$1000 < 2500$	100	1500	387	2428	
$2500 < 10000$	20	7500	387	2815	≥ 2500
≥ 10000	10	20000	447	3262	
Total	6530		3262		

root. The limits of the strata are then chosen in such a way that each of the three strata contains approximately one third of the total of $\sqrt{bf}$. The exact limits should have been $\frac{3262}{3} = 1087$ and $\frac{(2)(3262)}{3} = 2175$; the limits in the last column of Table 3.11 are the closest approximation to these numbers. The result is that the total sample size is distributed evenly over the three strata. Note that the value of $b = 20000$ for the highest class interval is based on the fact that the largest company had a production value of 30000 euro.

So far we have assumed that the sizes of the strata in the population are known. Should this not be the case then the method of double sampling for stratification discussed in Section 7.3.3 can be considered.

As concerns the *number* of strata, each increase in the number of strata in principle has the effect of decreasing the variance. In practice, however, it is found that the gain in precision for more than six to ten strata is usually very small. The exact limit depends on the strength of association between the stratification and the research variable: the higher their correlation, the more strata can be chosen. However, since more strata also involve higher costs the number of strata is seldom larger than ten.

This applies especially when only one stratification variable is used and when estimates are only required for the total population. When there are several stratification variables it is generally better to stratify with respect to different stratification variables simultaneously, using not too many classes for each variable, for example three or four.

If separate estimates are required for parts of the population it is often a good idea to use a fairly large number of strata. As stated earlier, these strata should preferably be chosen in such a way that the subpopulations of interest can be obtained by joining these strata together.

A practical application of stratified random sampling for the estimation of the total number of kilometers driven by passenger cars is provided in Molnar, Moritz, Smeets, Buelens, and Dohmen (2009). They base their estimation on data from a Dutch commercial database called "Nationale Autopas " (which is Dutch for "National Car Pass") containing odometer readings collected at the moment a car visits a service station for a periodic roadworthiness check (called APK in Dutch) or a car service. Besides a fixed sum for handling-costs, the

owners of the database also charge a small amount for each odometer reading. The available budget per year is limited. Therefore, the sample design is developed to achieve a maximum precision of the traffic estimates for given budget.

In another database of the Dutch Road Traffic authorities, all motor vehicles in the Netherlands are registered. The latter database also contains several technical features of each car. This information was used to set up an optimal allocation scheme for stratified random sampling. While sampling Molnar, Moritz, Smeets, Buelens, and Dohmen (2009) stratified by the following variables:

- year of construction of the car,
- fuel type, three types: 1) Petrol, 2) Diesel, and 3) Other, and
- ownership, two categories: 1) private and 2) business.

The use of these stratification variables improves the precision of the sample estimate of the total number of kilometers driven by passenger cars because newer cars tend to be used more intensively than older cars, cars running on diesel and gas oil tend to cover more kilometers than cars running on petrol, and cars used for business also tend to be used more intensively than privately owned cars. For estimates of the variances in each of these strata, Molnar, Moritz, Smeets, Buelens, and Dohmen (2009) used older samples from the NAP database.

3.11 Subpopulations

When we have subpopulations that cut through the strata, we need to distinguish between three situations:

- The size of the subpopulation is known in each stratum;
- We only know the total size of the subpopulation;
- We do not know anything about the size of the subpopulation.

We first consider the most simple situation where the size of the subpopulation is known in each stratum $h = 1, \ldots, L$. Let A_h denote this size, and a_h the size of the subpopulation in the sample drawn from stratum h. Further let $\bar{Y}_{sh}$ and S^2_{sh} denote the mean and variance of the subpopulation in stratum h. Then

$$\bar{y}_{sh} = \frac{1}{a_h} \sum_{i=1}^{a_h} y_{hi} \tag{3.54}$$

and

$$s^2_{sh} = \frac{1}{a_h - 1} \sum_{i=1}^{a_h} (y_{hi} - \bar{y}_{sh})^2 \tag{3.55}$$

are unbiased estimators of these subpopulation parameters in stratum h. For the total of the complete subpopulation we have the estimators

$$\hat{Y}_s = \sum_{h=1}^{L} A_h \bar{y}_{sh} = \sum_{h=1}^{L} \frac{A_h}{a_h} \sum_{i=1}^{a_h} y_{hi} \tag{3.56}$$

and

$$\mathrm{v}(\hat{Y}_s) = \sum_{h=1}^{L} \frac{A_h^2 s_{sh}^2}{a_h} \left(1 - \frac{a_h}{A_h}\right). \tag{3.57}$$

Since the total number of elements in the subpopulation is $A = \sum_{h=1}^{L} A_h$, dividing (3.56) and (3.57) by A and A^2 respectively yields unbiased estimators of the submean and its variance:

$$\hat{\bar{Y}}_s = \frac{\hat{Y}_s}{A} = \frac{1}{A} \sum_{h=1}^{L} A_h \bar{y}_{sh} \tag{3.58}$$

and

$$\mathrm{v}(\hat{\bar{Y}}_s) = \frac{v(\hat{Y}_s)}{A^2} = \frac{1}{A^2} \sum_{h=1}^{L} \frac{A_h^2 s_{sh}^2}{a_h} \left(1 - \frac{a_h}{A_h}\right). \tag{3.59}$$

Next, we consider the situation where the total size A of the subpopulation is known, but not the sizes of A_h per stratum. Formulas (3.56), (3.57), (3.58), and (3.59) are then no longer applicable. Just as in Section 2.10 we therefore need to introduce the variable

$$y_i^{'} = \begin{cases} y_i, & \text{if the unit is in the subpopulation,} \\ 0, & \text{if the unit is not in the subpopulation.} \end{cases}$$

An unbiased estimator of the mean of this variable in stratum h is

$$\bar{y}_h^{'} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}^{'} = \frac{1}{n_h} \sum_{i=1}^{a_h} y_{hi} = \frac{a_h}{n_h} \bar{y}_{sh},$$

as a result of (3.54). Replacement of the ratio $\frac{A_h}{a_h}$ in (3.56) with $\frac{N_h}{n_h}$ yields the following estimator of the subtotal in stratum h:

$$N_h \bar{y}_h^{'} = \frac{N_h}{n_h} \sum_{i=1}^{a_h} y_{hi}, \tag{3.60}$$

see also (2.88). The estimator of the grand subtotal in the population is then

$$\hat{Y}_s = \sum_{h=1}^{L} N_h \bar{y}_h^{'} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{a_h} y_{hi}. \tag{3.61}$$

Analogous to (2.92) the estimated variance in each stratum h within the subpopulation is

$$s_h^{'2} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi}^{'} - \bar{y}_h^{'})^2, \tag{3.62}$$

and the variance estimator of the grand subtotal $\hat{Y}_s$ in the population is

$$\mathrm{v}(\hat{Y}_s) = \text{estimated } \sigma_{\hat{Y}_s}^2 = \sum_{h=1}^{L} \frac{N_h^2 s_h^{'2}}{n_h} \left(1 - \frac{n_h}{N_h}\right). \tag{3.63}$$

For estimates of the submean in the population we divide (3.61) by A and (3.63) by A^2.

Finally, we consider the often encountered situation where even A, the total size of the subpopulation, is unknown. In this case (3.61) and (3.63) can still be used for the estimation of the subtotal and its variance because the value of A is not required in this formulas. However, to obtain estimates of the submean and its variance, we need to estimate the value of A. Since the observed sample fraction $\frac{a_h}{n_h}$ can be considered an estimate of the unknown population fraction $\frac{A_h}{N_h}$ in each stratum (and thus $\frac{A_h}{N_h} \doteq \frac{a_h}{n_h}$), an estimate of A_h in each stratum is obtained with $\hat{A}_h = \frac{N_h a_h}{n_h}$, and summing over all strata yields the estimate

$$\hat{A} = \sum_{h=1}^{L} \frac{N_h a_h}{n_h} \tag{3.64}$$

for the size of the subpopulation. This value can then be used to estimate the submean with

$$\hat{\bar{Y}}_s = \frac{1}{\hat{A}} \sum_{h=1}^{L} N_h \bar{y}_h'. \tag{3.65}$$

Unfortunately, we can not simply divide (3.63) by $\hat{A}^2$ in order to obtain an estimate of the variance of the submean, because submean (3.65) is now a ratio of two estimators. As will be discussed in Section 6.1 this situation requires a different variance formula. Letting

$$s_{rh}^2 = s_{sh}^2 + \frac{a_h}{n_h - 1} \left(1 - \frac{a_h}{n_h}\right) (\bar{y}_{sh} - \hat{\bar{Y}}_s)^2, \tag{3.66}$$

where the subscript r stands for ratio, the variance of the submean is

$$\mathrm{v}(\hat{\bar{Y}}_s) = \text{estimated}\, \sigma^2_{\hat{\bar{Y}}_s} = \frac{1}{\hat{A}^2} \sum_{h=1}^{L} \frac{N_h^2 s_{rh}^2}{n_h} \left(1 - \frac{n_h}{N_h}\right). \tag{3.67}$$

We end by noting – and this will not come as a surprise – that not knowing the values of the A_h's and of A in the population comes with a price: a decrease in precision of the estimated population parameters.

Chapter 4

Sampling with unequal probabilities

In Chapter 2 we have considered simple random sampling where the N units in the population all have the same probability $\frac{1}{N}$ of ending up in the sample. In Chapter 3 we encountered situations where the probabilities of being drawn are different for different groups of elements (i.e., strata) in the population. In this chapter we consider the possibility of drawing elements with probabilities that are different for each element of the population. With this procedure, precision of population parameter estimates can sometimes be improved even more than in stratified random sampling.

4.1 The mean and total, and their variances

Let z_i denote the probability of element i $(i = 1, \ldots, N)$ being drawn from the population, with $\sum_{i=1}^{N} z_i = 1$. When a sample of n elements is drawn with replacement from this population,

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{z_i}, \tag{4.1}$$

is an unbiased estimator of the total of the population. It is interesting to note that, when $z_i = \frac{1}{N}$ for all $i = 1, \ldots, N$, (4.1) yields

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{z_i} = \frac{1}{n} \sum_{i=1}^{n} N y_i = N \frac{1}{n} \sum_{i=1}^{n} y_i = N\bar{y},$$

as in simple random sampling, see (2.6). The variance of this estimator is

$$\mathrm{V}(\hat{Y}) = \sigma^2_{\hat{Y}} = \frac{1}{n} \sum_{i=1}^{N} z_i (\frac{y_i}{z_i} - Y)^2. \tag{4.2}$$

A unbiased sample estimator of this variance is

$$\mathrm{v}(\hat{Y}) = \text{estimated}\, \sigma^2_{\hat{Y}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\frac{y_i}{z_i} - \hat{Y})^2. \tag{4.3}$$

For the mean and its variance we obtain the unbiased estimators

$$\hat{\hat{Y}} = \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i}{z_i} \tag{4.4}$$

from (4.1), and

$$\mathrm{v}(\hat{\hat{Y}}) = \text{estimated}\,\sigma^2_{\hat{\hat{Y}}} = \frac{1}{n(n-1)N^2}\sum_{i=1}^{n}(\frac{y_i}{z_i} - \hat{Y})^2 \tag{4.5}$$

from (4.3). As is illustrated in Example 4-1, compared with random sampling with equal probabilities, the precision of the estimates is improved the most when the unequal probabilities are chosen to be proportional to the values of the variable of interest. However, when the probabilities are chosen inversely proportional to the values of the variable of interest, then the precision of the estimates is even worse than in sampling with equal probabilities.

Example 4-1. Consider an artificial population of three elements, with probabilities of being drawn and values on the variable of interest shown in Table 4.1. What are the variances of the total for a random sample of $n = 1$ 1) with the probabilities given in Table 4.1, 2) with equal probabilities, and 3) with probabilities equal to $z_1 = 0.5$, $z_1 = 0.3$, and $z_1 = 0.2$?

Table 4.1: Population of $N = 3$ elements with probabilities z_i of being drawn.

i	y_i	probability z_i
1	3	0.2
2	9	0.3
3	18	0.5
Total	30	1.0

Answer. The variance of the total for the data in Table 4.1 is

$$\mathrm{V}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{N} z_i(\frac{y_i}{z_i} - Y)^2 = 0.2(\frac{3}{0.2} - 30)^2 + 0.3(\frac{9}{0.3} - 30)^2 + 0.5(\frac{18}{0.5} - 30)^2$$
$$= 45 + 0 + 18 = 63.$$

When using equal probabilities we find that

$$\mathrm{V}(\hat{Y}) = \frac{1}{3}((3)(3) - 30)^2 + \frac{1}{3}((9)(3) - 30)^2 + \frac{1}{3}((18)(3) - 30)^2 = 147 + 3 + 192 = 342.$$

With probabilities equal to $z_1 = 0.5$, $z_1 = 0.3$, and $z_1 = 0.2$, on the other hand, we obtain

$$\mathrm{V}(\hat{Y}) = 0.5(\frac{3}{0.5} - 30)^2 + 0.3(\frac{9}{0.3} - 30)^2 + 0.2(\frac{18}{0.2} - 30)^2 = 288 + 0 + 720 = 1,008.$$

Note that the variance of the total is the smallest for the probabilities in Table 4.1, because these probabilities are approximately proportional to the values of y_i in the population. Choosing probabilities that are approximately inversely proportional to the values of y_i in the population, on the other hand, results in a variance of the total that is even larger than when equal probabilities are used.

4.2 Sampling with unequal probabilities in practice

Suppose we have an artificial population of three elements only: y_1, y_2, and y_3. As in Table 4.1, we want to draw y_1 with a probability of $z_1 = 0.2$, y_2 with a probability of $z_2 = 0.3$, and y_3 with a probability of $z_3 = 0.5$. Note that these probabilities add up to 1, as they should. We now want to draw a random sample of $n = 1$ element from this population. How should we proceed?

We first of all note that it is always possible to write each of the N probabilities z_i $(i = 1, \ldots, N)$ as a ratio f_i/F_N such that f_i is an integer and $F_N = \sum_{i=1}^{N} f_i$. In this example, we can write $z_1 = \frac{2}{10}$, $z_2 = \frac{3}{10}$, and $z_3 = \frac{5}{10}$ with $F_N = 2 + 3 + 5 = 10$. Now draw a random number between 1 and $F_N = 10$ inclusive. If this random number equals 1 or 2, then the first element y_1 is added to the sample, if the random number equals 3, 4, or 5, then the sample is y_2, and if it is 6, 7, 8, 9, or 10, then the sample is y_3.

Table 4.2: Population of $N = 9$ municipalities and number of inhabitants (rounded).

Municipality	Number of inhabitants	Cumulative number of inhabitants
1	22,600	22,600
2	32,200	54,800
3	87,800	142,600
4	20,100	162,700
5	22,500	185,200
6	41,000	226,200
7	27,300	253,500
8	24,300	277,800
9	22,000	299,800

Next, suppose we want to draw a random sample of four out of a population of nine municipalities with probabilities proportional to the number of inhabitants of each municipality, see Table 4.2. Then the number of inhabitants are first cumulated, and four random numbers are drawn from the range $1, \ldots, 2998$, the total number of inhabitants in hundreds in the population. Should the numbers 669, 2503, 2349, and 2952 happen to be drawn, for example, then the sample consists of municipalities 3, 7, again 7, and 9 respectively.

Chapter 5

Two-stage sampling

So far we have assumed that a complete sampling frame is available, containing a list of all the N elements of the population of interest, from which a simple or stratified random sample of size n can easily be drawn. When such a list is not available, two-stage sampling is an alternative option. In two-stage sampling, first a sample of n *groups of units*, often called *primary units*, is drawn from the population. Then, in the second step, a sample of m second-stage units or subunits, also called *secondary units*, is drawn from each of the n chosen primary units. In this case it is assumed that a sampling frame for the primary units is available, but not necessarily for all the second-stage units. Should the secondary units be again made up of tertiary units, then there is a third stage, in which case we are dealing with a three-stage sampling design. Other general terms for this type of sampling design are *subsampling* and *multi-stage sampling*.

As an example, consider a survey amongst people of 12 years and older of a country. In the first stage a sample of n primary units is drawn from all the N municipalities of that country. In the second stage m persons of 12 years and older are drawn from each of the n municipal registers. The advantage of this sampling design is that a list of all N municipalities of country is easily obtained, while the list of all people of 12 years and older of a country is not available. The second stage then only requires sampling from n municipal registers.

When the primary units all contain the same number of subunits, this is called subsampling or two-stage sampling with units of equal size. This situation is discussed in Section 5.1. When the primary units all contain a different number of subunits, we have subsampling or two-stage sampling with units of unequal size. This situation is taken up in Section 5.2.

To complicate matters further, at different stages different types of sampling designs can be used, e.g., we may use simple random sampling at all stages, or stratified random sampling at all stages, or simple random sampling at the first stage and stratified random sampling at the second stage, et cetera. Here we only discuss simple random sampling in a two-stage situation. For sample size estimation when stratified random sampling is used at each stage, we refer to Section 10.9 of Cochran (1977).

When *all* of the secondary or subunits of the sampled primary units are observed then the resulting sampling design is called *cluster sampling*, a sampling technique that we do not discuss in this document. For the reader interested in the details of cluster sampling we refer to Cochran (1977, Chapters 9 and 9A).

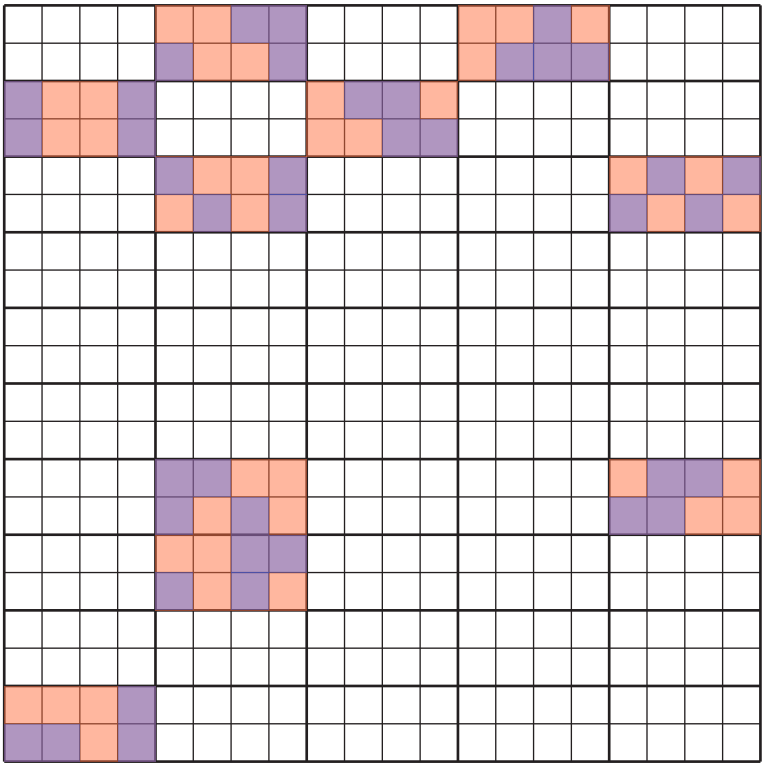

Figure 5.1: Two-stage sampling example.

5.1 Units of equal size

Let the total number of primary units in the population be denoted by N, from which n units are sampled, and let the total number of secondary units in each primary unit of the population be denoted by M from which m secondary units are sampled. The total number of elements in the population is then NM, while the total sample size then consists of nm elements. In Figure 5.1 this is illustrated for the situation where the total population consists of 400 elements (i.e., cells). The population has further been divided into $N = 50$ primary units, each consisting of $M = 8$ secondary units. From a total of $N = 50$ primary units, $n = 10$ have been sampled in the first stage (as indicated by the red rectangles in Figure 5.1), and from each of these latter 10 chosen primary units, $m = 4$ secondary units have been sampled in the second stage (as indicated by the blue squares in Figure 5.1). This results in a sample of $nm = (10)(4) = 40$ secondary units or elements from a population with a total of 400 elements.

We further use the following notation.

y_{ij} =value obtained for the jth subunit in the ith primary unit

$\bar{y}_i = \sum_{j=1}^{m} \frac{y_{ij}}{m}$ = sample mean in the ith primary unit

$\bar{\bar{y}} = \sum_{i=1}^{n} \frac{\bar{y}_i}{n}$ = overall sample mean

$S_1^2 = \frac{\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2}{N-1}$ = population variance among primary unit means

$S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2}{N(M-1)}$ = population variance among subunits within primary units

If the n units and the m subunits from each chosen unit are selected by simple random sampling, then $\bar{\bar{y}}$ is an unbiased estimator of $\mu = \bar{\bar{Y}}$ with variance

$$\mathrm{V}(\bar{\bar{y}}) = \sigma_{\bar{\bar{y}}}^2 = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn}. \tag{5.1}$$

Moreover, an unbiased estimator of $\mathrm{V}(\bar{\bar{y}})$ from the sample is

$$\mathrm{v}(\bar{\bar{y}}) = \hat{s}_{\bar{\bar{y}}}^2 = \left(\frac{1-f_1}{n}\right)s_1^2 + \left(\frac{f_1(1-f_2)}{mn}\right)s_2^2, \tag{5.2}$$

where $f_1 = n/N$, $f_2 = m/M$, and

$$s_1^2 = \frac{\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2}{n-1} \text{ and } s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2}{n(m-1)}, \tag{5.3}$$

see Cochran (1977, pp.277-278) for proofs of (5.1) and (5.2).

It is interesting to note if $m = M$, i.e., if $f_2 = 1$, then formula (5.1) – and the corresponding sampling scheme – reduces to that for *cluster sampling* (not discussed in this document). On the other hand, if $n = N$ the formula is that for proportional stratified random sampling as discussed in Chapter 3, because the primary units may then be regarded as strata, all of which are sampled.

Example 5-1. Suppose we have a population of 24 elements with $N = 6$ primary units, each consisting of $M = 4$ secondary units, see Table 5.1. What is the variance of the distribution of the mean when a random sample of $n = 4$ primary units is drawn in the first stage from this population, and then a random sample of $m = 1$ secondary units from each of these $n = 4$ primary units in the second stage? And what is this variance for $n = 2$ and $m = 2$? And for $n = 1$ and $m = 4$?

Answer. The sums $Y_i = \sum_{j=1}^{M}$ $(i = 1, \ldots, N)$ and means $\bar{Y}_i$ $(j = 1, \ldots, N)$ of the $M = 4$ secondary units in each primary unit of the population are given in the sixth and seventh column of Table 5.1, respectively. The overall mean of the population is $\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \frac{1}{6}(30) = 5$, which can also be calculated as $\bar{\bar{Y}} = \frac{1}{NM}\sum_{i=1}^{N} Y_i = \frac{1}{(6)(4)}(120) = 5$.

Table 5.1: An artificial population of 24 elements with $N = 6$ primary units (in the rows), each consisting of $M = 4$ secondary or subunits (in the columns).

	1	2	3	4	Y_i	$\bar{Y}_i$
1	4	3	2	1	10	2.5
2	2	6	4	8	20	5.0
3	3	4	1	2	10	2.5
4	8	2	4	6	20	5.0
5	12	3	2	1	30	7.5
6	12	6	9	3	30	7.5
Total					120	30

Table 5.2: Variances for three different two-stage sampling designs from the population in Table 5.1.

sampling design		variance $\left(\frac{N-n}{N}\right)\frac{S_1^2}{n}$	variance $\left(\frac{M-m}{M}\right)\frac{S_2^2}{mn}$	
n	m	in first stage	in second stage	total variance
4	1	0.417	1.458	1.875
2	2	1.667	0.972	2.639
1	4	4.167	0	4.167

The population variance among primary unit means in Table 5.1 is $S_1^2 = 5$ and that among subunits within primary units is $S_2^2 = 7.778$. For a random sample of $n = 4$ primary units from this population, followed by a random sample of $m = 1$ secondary units from each of these $n = 4$ primary units, i.e., for the situation that $n = 4$ and $m = 1$, the two terms at the right of (5.1) are $\left(\frac{N-n}{N}\right)\frac{S_1^2}{n} = \left(\frac{6-4}{6}\right)\frac{5}{4} = 0.417$ and $\left(\frac{M-m}{M}\right)\frac{S_2^2}{mn} = \left(\frac{4-1}{4}\right)\frac{7.778}{(1)(4)} = 1.458$. The total variance of the mean is therefore $0.417 + 1.458 = 1.875$, see the second row of Table 5.2. Analogous calculations for $n = 2$ and $m = 2$, and for $n = 1$ and $m = 4$ yield the variances shown in the second and third rows of Table 5.2, respectively. We see that the variance decreases when m increases, but also that the variance in the first stage increases faster as n gets smaller. For given sample size nm, the total variance is therefore smallest (and the precision largest) when n is as large as possible.

In two-stage sampling it is almost always true that – for constant sample size nm – the variance is smallest when n is as large as possible, and m therefore as small as possible. However, when choosing for a certain two-stage sampling design it is equally important to consider the amount of money involved in collecting the data. As noted earlier, a multi-stage sampling design is often used because no sampling frame of the total population is available. But an increase in the number of sampled primary units also increases the costs of the survey, because a sampling frame of secondary units will have to be constructed or consulted for each of the sampled primary units.

The previous formulas presented so far apply to two-stage sampling of quantitative vari-

ables. For qualitative variables (i.e., percentages and proportions) the subunits are classified into two classes and we estimate the proportion that falls in the first class. The previous formulas can then be applied by defining $y_{ij} = 1$ if the corresponding subunit falls into the first class and as $y_{ij} = 0$ otherwise. Let $p_i = a_i/m$ be the proportion falling in the first class in the subsample from the ith unit. Then $\bar{p} = \frac{\sum p_i}{n}$ and an unbiased estimator of $\mathrm{V}(\bar{p})$ from the sample is

$$\mathrm{v}(\bar{p}) = \hat{s}^2_{\bar{p}} = \frac{1-f_1}{n(n-1)}\sum_{i=1}^{n}(p_i-\bar{p})^2 + \frac{f_1(1-f_2)}{n^2(m-1)}\sum_{i=1}^{n} p_i q_i. \tag{5.4}$$

Example 5-2. In a study of plant disease plants were grown in 160 small plots containing nine plants each. A random sample of 40 plots was taken and three random plants in each sampled plot were checked for the presence of disease. It was found that 22 plots had no diseased plants out of three, 11 had one, 4 had two, and 3 had three. Estimate the proportion of diseased plants and its standard error.

Answer. We have $N = 160$, $n = 40$, $M = 9$, and $m = 3$. The proportions p_i of diseased plants per sampled plot are 0 in 22 plots, $\frac{1}{3}$ in 11 plots, $\frac{2}{3}$ in 4 plots, and $\frac{3}{3} = 1$ in 3 plots. The mean proportion of these 40 plots is therefore $\bar{p} = \frac{1}{n}\sum p_i = \frac{1}{40}((22)(0) + (11)(\frac{1}{3}) + (4)(\frac{2}{3}) + (3)(1))) = \frac{1}{40}\frac{28}{3} = \frac{7}{30} = 0.233$. Then $\sum_{i=1}^{n}(p_i - \bar{p})^2 = 3.822$, while $\sum_{i=1}^{n} p_i q_i = 3.333$, so that (5.4) yields

$$\mathrm{v}(\bar{p}) = \hat{s}^2_{\bar{p}} = \frac{1-\frac{40}{160}}{40(40-1)}(3.822) + \frac{\frac{40}{160}(1-\frac{3}{9})}{40^2(3-1)}(3.333) = 0.00201,$$

meaning that the standard error of the proportion is $\sqrt{0.00201} = 0.045$.

In two-stage sampling with units of equal size, an estimation of sample size is obtained by considering a cost function of the type

$$C = c_1 n + c_2 nm, \tag{5.5}$$

where c_1 denotes the cost of a primary unit, and c_2 the cost involved in a secondary or subunit. The first component of cost, $c_1 n$, is therefore proportional to the number of primary units in the sample, and the second, $c_2 nm$, is proportional to the total number of second stage elements in the sample. The optimal number of second stage elements in the sample is

$$m_{opt} = \frac{S_2}{\sqrt{S_1^2 - S_2^2/M}}\sqrt{c_1/c_2}, \tag{5.6}$$

provided that $S_1^2 > S_2^2/M$ after which m_{opt} is rounded to the nearest integer. Should $m_{opt} > M$ or $S_1^2 \leq S_2^2/M$ then take $m = M$, i.e., use cluster sampling.

Proof. In order to find the optimal value for n and m we minimize the product of the variance of the mean $\mathrm{V}(\bar{\bar{y}})$ in (5.1) and the cost function C in (5.5) with respect to n and m. It is not very difficult to verify that the variance (5.1) can also be written as

$$\mathrm{V}(\bar{\bar{y}}) = \frac{1}{n}\left(S_1^2 - \frac{S_2^2}{M}\right) + \frac{1}{mn}S_2^2 - \frac{1}{N}S_1^2, \tag{5.7}$$

and the last term on the right is a constant with respect to n and m. This means that the minimization of the product $\mathrm{V}(\bar{\bar{y}})C$ is equivalent to the minimization of

$$\left[\frac{1}{n}\left(S_1^2 - \frac{S_2^2}{M}\right) + \frac{1}{mn}S_2^2\right][c_1 n + c_2 nm] = \left[\left(S_1^2 - \frac{S_2^2}{M}\right) + \frac{1}{m}S_2^2\right][c_1 + c_2 m], \tag{5.8}$$

since the terms n cancel each other out. Working out the product (5.8) shows that we have to minimize

$$\frac{c_1 S_2^2}{m} + mc_2\left(S_1^2 - \frac{S_2^2}{M}\right) + c_2 S_2^2 + c_1\left(S_1^2 - \frac{S_2^2}{M}\right) \tag{5.9}$$

with respect to m. The optimum is obtained by setting the first order derivative of (5.9) equal to zero:

$$-\frac{c_1 S_2^2}{m^2} + c_2\left(S_1^2 - \frac{S_2^2}{M}\right) = 0 \tag{5.10}$$

and then solving (5.10) with respect to m. This yields (5.6), which completes the proof.

The value of n_{opt} can next be found by solving either the cost equation (5.5) or the variance equation (5.1) with respect to n, depending on which criterium has been defined beforehand. If it is the total cost C in (5.5) that is fixed, for example, then (5.6) is substituted in (5.5) and solved for n yielding

$$n_{opt} = \frac{C}{c_1 + m_{opt}c_2}. \tag{5.11}$$

If it is the variance $\mathrm{V}(\bar{\bar{y}})$ that is fixed, on the other hand, then (5.6) is substituted in (5.1) and solved for n yielding

$$n_{opt} = \frac{S_1^2 + \left(\frac{M - m_{opt}}{M}\right)\frac{S_2^2}{m_{opt}}}{\mathrm{V}(\bar{\bar{y}}) + \frac{S_1^2}{N}}. \tag{5.12}$$

In practice, the estimation of m and n requires estimates of c_1/c_2 and of S_2/S_1.

From (5.6) the following rules can be derived. When c_1 is much larger than c_2, a large m should be chosen and therefore a small n. We could even take $m = M$, which means that cluster sampling is used if c_1 (the preparation cost) is large and c_2 (the observation cost) is small. When c_1 is small in comparison with c_2, a small m should be chosen and a large n; in this situation we could even go for $n = N$, a stratified random sample.

Similarly, a relatively large value of S_2 compared with S_1 results in a large m. In other words, when the primary units are not very different, we only need to observe a few of them. Should S_1 be large in comparison with S_2, on the other hand, then we should choose n large and m small, because the secondary units within the primary units are then relatively homogeneous, while the primary units are relatively heterogeneous.

5.2 Units of unequal size

5.2.1 Sampling with equal probabilities

A situation encountered very often in practice is that the primary units vary in size. Let M_i denote the number of secondary units in primary unit i ($i = 1, \ldots, N$) of the population.

Then $M_0 = \sum_{i=1}^{N} M_i$ is the number of elements in the total population, and the mean size $\bar{M}$ of the primary units equals

$$\bar{M} = \frac{M_0}{N} = \frac{1}{N}\sum_{i=1}^{N} M_i. \tag{5.13}$$

The total of primary unit i is

$$Y_i = \sum_{j=1}^{M_i} y_{ij}, \tag{5.14}$$

so that the grand total in the population equals

$$Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}. \tag{5.15}$$

The average total in the N primary units of the population is

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{Y}{N}, \tag{5.16}$$

and the mean of primary unit i equals

$$\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} = \frac{1}{M_i} Y_i. \tag{5.17}$$

The grand mean of the population is therefore

$$\bar{\bar{Y}} = \frac{Y}{M_0} = \frac{\bar{Y}}{\bar{M}}. \tag{5.18}$$

Example 5-3. Consider the population shown in Table 5.3. Calculate the totals for each primary unit, the means for each primary unit, the average total of the primary units, the grand total of the population, and the grand mean of this population.

Answer. The totals Y_i and means $\bar{Y}_i$ of the primary units in this population are obtained from (5.14) and (5.17), respectively, and can be found in the fourth and fifth column of Table 5.3. The average total of the four primary units is obtained from (5.16) yielding $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{4}(9 + 7 + 36 + 28) = 20 = \frac{1}{N}Y$. The grand total of the population is $Y = 9 + 7 + 36 + 28 = 80$, and the grand mean is $\bar{\bar{Y}} = \frac{Y}{M_0} = \frac{80}{16} = 5 = \frac{20}{4} = \frac{\bar{Y}}{\bar{M}}$ since $\bar{M} = \frac{M_0}{N} = \frac{16}{4} = 4$.

Suppose we now draw n random primary units from the population without replacement in the first stage, and m_i secondary units from each of these n primary units without replacement in the second stage. Then the sample mean in primary unit i is

$$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}. \tag{5.19}$$

Table 5.3: An artificial population of 16 elements with $N = 4$ primary units (in the rows), consisting of 2, 4, 6, and 4 secondary or subunits (in the columns), respectively.

primary unit	secondary units	M_i	Y_i	$\bar{Y}_i = Y_i/M_i$	S_{2i}^2	$z_i = \frac{M_i}{M_0}$
1	4 5	2	9	4.50	0.5000	$\frac{2}{16} = 0.125$
2	1 1 3 2	4	7	1.75	0.9167	$\frac{4}{16} = 0.250$
3	6 7 6 5 5 7	6	36	6.00	0.8000	$\frac{6}{16} = 0.375$
4	8 6 8 6	4	28	7.00	1.3333	$\frac{4}{16} = 0.250$
Total		$M_0 = 16$	$Y = 80$			1.000

This means that

$$\hat{y}_i = M_i \bar{y}_i. \tag{5.20}$$

is an unbiased estimator of the total Y_i of this primary unit and – since this primary unit has been selected randomly – also of the average total of the primary units. Averaging over the n primary units in the sample yields the following unbiased estimator of the average total (5.16) of the primary units in the population:

$$\hat{\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i. \tag{5.21}$$

An unbiased estimator of the grand mean of the population then equals

$$\hat{\bar{y}}_u = \frac{\hat{\bar{y}}}{\bar{M}} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \hat{y}_i, \tag{5.22}$$

while an unbiased estimator of the grand total of the population is

$$\hat{y}_u = N\hat{\bar{y}} = \frac{N}{n} \sum_{i=1}^{n} \hat{y}_i = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i. \tag{5.23}$$

Note that

$$M_0 \hat{\bar{y}}_u = M_0 \frac{1}{n\bar{M}} \sum_{i=1}^{n} \hat{y}_i = M_0 \frac{1}{n\frac{\bar{M}_0}{N}} \sum_{i=1}^{n} \hat{y}_i = \frac{N}{n} \sum_{i=1}^{n} \hat{y}_i = \hat{y}_u, \tag{5.24}$$

as it should be.

The variance of the unbiased estimate (5.23) of the total of the population is

$$\mathrm{V}(\hat{y}_u) = \sigma_{\hat{y}_u}^2 = \frac{N^2}{n}(1 - f_1) \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{(N-1)} + \frac{N}{n} \sum_{i=1}^{N} \frac{M_i^2 (1 - f_{2i}) S_{2i}^2}{m_i}, \tag{5.25}$$

where $f_1 = \frac{n}{N}$, $f_{2i} = \frac{m_i}{M_i}$, and

$$S_{2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2, \tag{5.26}$$

is the variance among subunits in the ith unit. In the latter two formulas Y_i, $\bar{Y}$, and $\bar{Y}_i$ are as defined in (5.14), (5.16), and (5.17), respectively.

An unbiased sample estimator of this variance is provided by

$$\mathrm{v}(\hat{y}_u) = \hat{s}^2_{\hat{y}_u} = \frac{N^2(1-f_1)}{n}\frac{\sum_{i=1}^n(\hat{y}_i - \hat{\bar{y}})^2}{(n-1)} + \frac{N}{n}\sum_{i=1}^n \frac{M_i^2(1-f_{2i})s^2_{2i}}{m_i}, \tag{5.27}$$

where

$$s^2_{2i} = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2, \tag{5.28}$$

and $\hat{y}_i$, $\hat{\bar{y}}$, and $\bar{y}_i$ are defined as in (5.20), (5.21), and (5.19), respectively, see Cochran (1977, p.303).

The variance of the unbiased estimate (5.22) of the mean of the population is

$$\mathrm{V}(\hat{\bar{y}}_u) = \sigma^2_{\hat{\bar{y}}_u} = \frac{\mathrm{V}(\hat{y}_u)}{M_0^2}, \tag{5.29}$$

and an unbiased sample estimator of this variance is provided by

$$\mathrm{v}(\hat{\bar{y}}_u) = \hat{s}^2_{\hat{\bar{y}}_u} = \frac{\mathrm{v}(\hat{y}_u)}{M_0^2}. \tag{5.30}$$

Example 5-4. Again consider the population shown in Table 5.3. What is the variance of the distribution of the mean when a random sample of $n = 2$ primary units is drawn in the first stage from this population, and then a random sample of $m_i = 2$ secondary units ($i = 1, \ldots, 4$) from the primary units in the second stage? And what is this variance for $n = 2$ and a proportional allocation of $m_i = \frac{1}{2}M_i$ for $i = 1, \ldots, 4$?

Answer. The variance of the sampling distribution of the total is obtained from (5.25). The values of the S^2_{2i} defined in (5.28) for this population are given in the second last column of Table 5.3. For $n = 2$ and $m_i = 2$ for $i = 1, \ldots, 4$, (5.25) yields

$$\mathrm{V}(\hat{y}_u) = \frac{4^2}{2}(1-\frac{2}{4})\frac{(9-20)^2 + (7-20)^2 + (36-20)^2 + (28-20)^2}{(4-1)}$$
$$+\frac{4}{2}\left(\frac{2^2(1-\frac{2}{2})0.5}{2} + \frac{4^2(1-\frac{2}{4})0.9167}{2} + \frac{6^2(1-\frac{2}{6})0.8}{2} + \frac{4^2(1-\frac{2}{4})1.3333}{2}\right) = 850.53,$$

from which it follows that the variance of the sampling distribution of the mean equals

$$\mathrm{V}(\hat{\bar{y}}_u) = \frac{\mathrm{V}(\hat{y}_u)}{M_0^2} = \frac{850.33}{16^2} = 3.3224,$$

according to (5.29).

Analogous calculations for $n = 2$ and $m_i = \frac{1}{2}M_i$ (resulting in $m_1 = 1$, $m_2 = 2$, $m_3 = 3$, and $m_4 = 2$) give $\mathrm{V}(\hat{y}_u) = 842.93$ and $\mathrm{V}(\hat{\bar{y}}_u) = 3.2927$, a slight increase in precision.

5.2.2 Sampling with unequal probabilities

Let the primary units in two-stage sampling be selected with (unequal) probabilities z_i $(i = 1, \ldots, N)$ *with replacement*, where the z_i are positive numbers adding up to 1. Let the subsample of m_i subunits from the ith unit be randomly drawn *without replacement*. Note that z_i is now allowed to be different for each primary unit. Here we present the results of selecting the primary units with replacement because the formulas for the true and estimated variances of the estimates are relatively simple in this case, while those for sampling without replacement are much more complicated. For those interested in the formulas without replacement, we refer to Cochran (1977, Section 11.10). An unbiased estimator of the population total when sampling with replacement is

$$\hat{Y}_{ppz} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{z_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}, \tag{5.31}$$

from which it follows that

$$\hat{\bar{Y}}_{ppz} = \frac{\hat{Y}_{ppz}}{M_0}, \tag{5.32}$$

and its variance is

$$\mathrm{V}(\hat{Y}_{ppz}) = \sigma^2_{\hat{Y}_{ppz}} = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i} - Y\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i z_i}. \tag{5.33}$$

from which it follows that

$$\mathrm{V}(\hat{\bar{Y}}_{ppz}) = \sigma^2_{\hat{\bar{Y}}_{ppz}} = \frac{\mathrm{V}(\hat{Y}_{ppz})}{M_0^2}. \tag{5.34}$$

An unbiased sample estimator of the total $\mathrm{V}(\hat{Y}_{ppz})$ is

$$\mathrm{v}(\hat{Y}_{ppz}) = \hat{s}^2_{\hat{Y}_{ppz}} = \frac{\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{z_i} - \hat{Y}_{ppz}\right)^2}{n(n-1)}, \tag{5.35}$$

and an unbiased sample estimator of the mean $\mathrm{V}(\hat{\bar{Y}}_{ppz})$ is therefore

$$\mathrm{v}(\hat{\bar{Y}}_{ppz}) = \hat{s}^2_{\hat{\bar{Y}}_{ppz}} = \frac{\mathrm{v}(\hat{Y}_{ppz})}{M_0^2}. \tag{5.36}$$

see Cochran (1977, pp.306-307). The subscript *ppz* in these formulas is shorthand for "probabilities proportional to z_i".

When the probabilities z_i are chosen such that $z_i = \frac{M_i}{M_0}$ for $i = 1, \ldots, n$, this is called two-stage sampling with *probabilities proportional to size* of the primary units. In this case (5.31) and (5.32) can be written as

$$\hat{Y}_{pps} = \frac{M_0}{n}\sum_{i=1}^{n}\bar{y}_i, \tag{5.37}$$

and

$$\hat{\bar{Y}}_{pps} = \frac{\hat{Y}_{pps}}{M_0} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i, \tag{5.38}$$

respectively, while (5.35) and (5.36) can then be written as

$$\mathrm{v}(\hat{Y}_{pps}) = \hat{s}^2_{\hat{Y}_{pps}} = \frac{M_0^2 \sum_{i=1}^{n}(\bar{y}_i - \frac{1}{M_0}\hat{Y}_{pps})^2}{n(n-1)} = \frac{M_0^2 \sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_{pps})^2}{n(n-1)}, \tag{5.39}$$

and

$$\mathrm{v}(\hat{\bar{Y}}_{pps}) = \frac{\mathrm{v}(\hat{Y}_{pps})}{M_0^2} = \frac{\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_{pps})^2}{n(n-1)}, \tag{5.40}$$

respectively. The subscript *pps* in the latter formulas is shorthand for "probabilities proportional to size". Note that this type of sampling requires that the size M_i of each primary unit in the population is known.

Example 5-5. Again consider the population shown in Table 5.3. What is the variance of the distribution of the mean when a random sample of $n = 2$ primary units is drawn in the first stage from this population, and then a random sample of $m_i = 2$ secondary units $(i = 1, \ldots, 4)$ from the primary units in the second stage, and the primary units are selected with (unequal) probabilities proportional to size? And what is this variance for $n = 2$ and a proportional allocation of $m_i = \frac{1}{2}M_i$ for $i = 1, \ldots, 4$?

Answer. For probabilities proportional to size (i.e., with probabilities $z_1 = \frac{2}{16} = 0.125$, $z_2 = \frac{4}{16} = 0.25$, $z_3 = \frac{6}{16} = 0.375$, and $z_4 = \frac{4}{16} = 0.25$, see the last column in Table 5.3), the variance of the total for $n = 2$ and $m_i = 2$ for all i is found with (5.33):

$$\mathrm{V}(\hat{Y}_{ppz}) = \frac{1}{2}\left(0.125(\frac{9}{0.125} - 80)^2 + 0.25(\frac{7}{0.25} - 80)^2 + 0.375(\frac{36}{0.375} - 80)^2 + 0.25(\frac{28}{0.25} - 80)^2\right)$$
$$+\frac{1}{2}\left(\frac{(2^2)(1-\frac{2}{2})(0.5)}{(2)(0.125)} + \frac{(4^2)(1-\frac{2}{4})(0.9167)}{(2)(0.25)} + \frac{(6^2)(1-\frac{2}{6})(0.8)}{(2)(0.375)} + \frac{(4^2)(1-\frac{2}{4})(1.333)}{(2)(0.25)}\right) = 548.8.$$

The variance of the mean therefore equals

$$\mathrm{V}(\hat{\bar{Y}}_{ppz}) = \frac{\mathrm{V}(\hat{Y}_{ppz})}{M_0^2} = \frac{548.8}{16^2} = 2.1438.$$

Analogously, when a proportional allocation of $m_i = \frac{1}{2}M_i$ is used we obtain

$$\mathrm{V}(\hat{Y}_{ppz}) = 546.4,$$

and

$$\mathrm{V}(\hat{\bar{Y}}_{ppz}) = \frac{\mathrm{V}(\hat{Y}_{ppz})}{M_0^2} = \frac{546.4}{16^2} = 2.1344,$$

which is only a very slight improvement upon the previous result. Note, however, that the sampling with probabilities proportional to size shown in this example gives much more precise results than the sampling with equal probabilities discussed in Example 4-4.

5.2.3 Sample size estimation with equal probabilities

For equal probabilities, the formula of the variance of the sampling distribution of the total is

$$\mathrm{V}(\hat{y}_u) = \frac{N^2}{n}(1-f_1)\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{(N-1)} + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i},$$

with $f_{2i} = \frac{m_i}{M_i}$, see (5.25). When using a self-weighting design with $f_{2i} = \frac{m_i}{M_i} = f_2 = \frac{\bar{m}}{\bar{M}} = \frac{N\bar{m}}{M_0}$ for $i = 1, \ldots, N$, it follows that $m_i = \frac{N\bar{m}M_i}{M_0}$. Substitution of the latter in the second term of the variance formula yields

$$\begin{aligned}\frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_2)S_{2i}^2}{m_i} &= \frac{N}{n}\sum_{i=1}^{N}\frac{M_0M_i^2(1-f_2)S_{2i}^2}{N\bar{m}M_i} = \sum_{i=1}^{N}\frac{M_0M_i(1-f_2)S_{2i}^2}{n\bar{m}} \\ &= \sum_{i=1}^{N}\frac{M_0^2M_i(1-f_2)S_{2i}^2}{n\bar{m}M_0} = \frac{M_0^2(1-f_2)}{n\bar{m}}\sum_{i=1}^{N}\frac{M_i}{M_0}S_{2i}^2,\end{aligned}$$

meaning that (5.25) can then be written as

$$\mathrm{V}(\hat{y}_u) = \frac{N^2}{n}(1-f_1)\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{(N-1)} + \frac{M_0^2(1-f_2)}{n\bar{m}}\sum_{i=1}^{N}\frac{M_i}{M_0}S_{2i}^2. \tag{5.41}$$

Dividing (5.41) by $M_0^2 = (\bar{M}N)^2$ yields the variance of the *mean*:

$$\begin{aligned}\mathrm{V}(\hat{\bar{\bar{y}}}_u) =& \frac{\mathrm{V}(\hat{y}_u)}{M_0^2} = \frac{N^2}{n\bar{M}^2N^2}(1-f_1)\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{(N-1)} + \frac{(1-f_2)}{n\bar{m}}\sum_{i=1}^{N}\frac{M_i}{M_0}S_{2i}^2 \\ =& \frac{(1-f_1)}{n}\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{\bar{M}^2(N-1)} + \frac{(1-f_2)}{n\bar{m}}\sum_{i=1}^{N}\frac{M_i}{M_0}S_{2i}^2.\end{aligned} \tag{5.42}$$

Let $S_b^2 = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{\bar{M}^2(N-1)}$ and $S_2^2 = \sum_{i=1}^{N}\frac{M_i}{M_0}S_{2i}^2$, and substitute $f_2 = \frac{\bar{m}}{\bar{M}}$ in (5.42), and we obtain

$$\begin{aligned}\mathrm{V}(\hat{\bar{\bar{y}}}_u) =& \frac{(1-\frac{n}{N})}{n}S_b^2 + \frac{(1-\frac{\bar{m}}{\bar{M}})}{n\bar{m}}S_2^2 \\ =& \frac{1}{n}S_b^2 - \frac{1}{N}S_b^2 + \frac{1}{n\bar{m}}S_2^2 - \frac{1}{n\bar{M}}S_2^2 \\ =& \frac{1}{n}(S_b^2 - \frac{1}{\bar{M}}S_2^2) + \frac{1}{n\bar{m}}S_2^2 - \frac{1}{N}S_b^2.\end{aligned} \tag{5.43}$$

Let the cost function be defined as

$$\text{cost} = c_u n + c_2\sum_{i=1}^{n}m_i + c_l\sum_{i=1}^{n}M_i, \tag{5.44}$$

where

- c_u is the fixed cost per primary unit

- c_2 is the cost per subunit
- c_l the cost of listing per subunit in a selected unit

Formula (5.44) is not usable as such because the total cost depends on the particular set of units that is selected. Instead, we consider the average cost over n units, which equals

$$E(C) = c_u n + c_2 n\bar{m} + c_l n\bar{M} = (c_u + c_l\bar{M})n + c_2 n\bar{m} = c_1 n + c_2 n\bar{m}, \tag{5.45}$$

with $\bar{m} = f_2\bar{M}$.

Because $\frac{1}{N}S_b^2$ is a constant with respect to n and m, the minimization of the product $\mathrm{V}(\hat{\bar{\bar{y}}}_u)E(C)$ over n and m is equivalent with the minimization of

$$\left[\frac{1}{n}(S_b^2 - \frac{1}{\bar{M}}S_2^2) + \frac{1}{n\bar{m}}S_2^2\right][c_1 n + c_2 n\bar{m}] = \left[(S_b^2 - \frac{1}{\bar{M}}S_2^2) + \frac{1}{\bar{m}}S_2^2\right][c_1 + c_2\bar{m}] \tag{5.46}$$

over $\bar{m}$. Working out the product (5.46) yields

$$c_2\bar{m}(S_b^2 - \frac{1}{\bar{M}}S_2^2) + \frac{c_1 S_2^2}{\bar{m}} + c_2 S_2^2 + c_1(S_b^2 - \frac{1}{\bar{M}}S_2^2). \tag{5.47}$$

Setting the first derivative of (5.47) with respect to $\bar{m}$ equal to zero:

$$c_2(S_b^2 - \frac{1}{\bar{M}}S_2^2) - \frac{c_1 S_2^2}{\bar{m}^2} = 0,$$

and solving for $\bar{m}$ yields

$$\bar{m}_{opt} = \frac{S_2}{\sqrt{S_b^2 - \frac{1}{\bar{M}}S_2^2}}\sqrt{\frac{c_1}{c_2}}. \tag{5.48}$$

Once $\bar{m}_{opt}$ has been obtained from (5.48), for fixed cost the value of n_{opt} can be found by substituting $\bar{m}_{opt}$ in (5.45), while for fixed variance it can be found by substituting $\bar{m}_{opt}$ in (5.43).

5.2.4 Sample size estimation with unequal probabilities

Here, we only consider the situation where the unequal probabilities are chosen to be proportional to size, i.e. where $z_i = \frac{M_i}{M_0}$, and where the m_i are all equal (meaning that $m_i = m$ for all i). In that situation it can be proven that the variance of the mean (5.34) can be written as

$$\mathrm{V}(\hat{\bar{\bar{Y}}}_{pps}) = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M_0}(\bar{Y}_i - \hat{\bar{\bar{Y}}}_{pps})^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M_0}\frac{1 - \frac{m}{M_i}}{m}S_{2i}^2. \tag{5.49}$$

Letting $S_b^2 = \sum_{i=1}^{N}\frac{M_i}{M_0}(\bar{Y}_i - \hat{\bar{\bar{Y}}}_{pps})^2$, $S_2^2 = \sum_{i=1}^{N}\frac{M_i}{M_0}S_{2i}^2$, and $S_3^2 = \frac{1}{M_0}\sum_{i=1}^{N}S_{2i}^2$, (5.49) can be simplified to

$$\mathrm{V}(\hat{\bar{\bar{Y}}}_{pps}) = \frac{1}{n}(S_b^2 + \frac{1}{m}S_2^2 - S_3^2). \tag{5.50}$$

Since, for appropriate definitions of c_1 and c_2, the average cost of sampling n units is

$$E(C) = c_1 n + c_2 nm, \tag{5.51}$$

the optimal sample size is found by minimizing the product of (5.50) and (5.51):

$$\begin{aligned} \mathrm{V}(\hat{\bar{Y}}_{pps})E(C) =& (S_b^2 + \frac{1}{m}S_2^2 - S_3^2)(c_1 + c_2 m) \\ =& c_1(S_b^2 - S_3^2) + \frac{1}{m}c_1 S_2^2 + c_2 m(S_b^2 - S_3^2) + c_2 S_2^2. \end{aligned} \tag{5.52}$$

The global optimum is obtained by setting the first derivative of (5.52) with respect to m equal to zero:

$$-\frac{1}{m^2}c_1 S_2^2 + c_2(S_b^2 - S_3^2) = 0, \tag{5.53}$$

and then solving for m, yielding

$$m_{opt} = \sqrt{\frac{S_2^2}{S_b^2 - S_3^2}}\sqrt{\frac{c_1}{c_2}}. \tag{5.54}$$

Just as in the previous section, once m_{opt} has been determined with (5.54), for fixed cost the value of n_{opt} can be found by substituting m_{opt} in (5.51), while for fixed variance it can be obtained by substituting m_{opt} in (5.49).

Chapter 6

Alternative methods of estimation

In the previous chapters we have presented several sampling methods, and discussed the effect these methods have on the precision of the sample estimates of population parameters. Another important way to improve the precision of population estimates is to use other methods of estimation than just the sample mean, total or percentage. This is the topic of the present chapter. We discuss two alternative estimators: the ratio estimator and the regression estimator. As will become clear below, both estimators require information on auxiliary variables that are (highly) correlated with the variable of interest. Moreover, these alternative estimators can only be applied in the situation that both the variable of interest and the auxiliary variables are *quantitative* variables.

6.1 The ratio estimator

When we have a simple random sample of n observations y_i, the (unbiased) *direct estimator* of the population mean is given by

$$\hat{\bar{Y}}_d = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}, \tag{6.1}$$

see Chapter 2, where we use the subscript d in the present chapter to indicate that it is a direct estimator. Now suppose that we also know the sample and population mean of an auxiliary variable x, then we can use their ratio $\frac{\bar{X}}{\bar{x}}$ to change the direct estimator into the *ratio estimator*:

$$\hat{\bar{Y}}_R = \bar{y}\frac{\bar{X}}{\bar{x}}, \tag{6.2}$$

where the subscript R denotes the ratio estimator, which can also be written as

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}}\bar{X} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\bar{X}. \tag{6.3}$$

The total of the population is then estimated with

$$\hat{Y}_R = N\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}}X. \tag{6.4}$$

When y and x are positively correlated, the ratio estimator (6.3) is often more precise than the direct estimator (6.1). Intuitively this can be explained as follows. When y and x are positively correlated and $\bar{y}$ happens to be large, then $\bar{x}$ will also tend to be large. Multiplication of $\bar{y}$ with $\frac{\bar{X}}{\bar{x}}$ will therefore tend to result in a smaller value for $\bar{y}$. Conversely, when $\bar{y}$ happens to be small, then $\bar{x}$ will also tend to be smaller than $\bar{X}$, meaning that $\hat{\bar{Y}}_r$ will probably be larger than $\bar{y} = \hat{\bar{Y}}_d$. The result is that the fluctuation of $\hat{\bar{Y}}_r$, the ratio estimator of the population mean $\bar{Y}$, and therefore its variance, will be smaller than that of $\hat{\bar{Y}}_d$, the direct estimator of the population mean $\bar{Y}$.

In practice y and x are often the same variable measured at different time points or time periods. Auxiliary variable x is often also some sort of measure of the size of the units in the population. The part of (6.3) and (6.4) that is subject to random fluctuation is the ratio

$$\hat{R} = \frac{\bar{y}}{\bar{x}}, \tag{6.5}$$

which can be conceived of as the estimator of the population ratio

$$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}. \tag{6.6}$$

Defining

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}, \tag{6.7}$$

the ratio estimators of the mean and the total of the population can be written as

$$\hat{\bar{Y}}_R = \hat{R}\bar{X} \tag{6.8}$$

and

$$\hat{Y}_R = \hat{R}X, \tag{6.9}$$

respectively.

6.1.1 Properties of the ratio estimator and its variance

The values of $\bar{X}$ and X are fixed, and the statistical properties of (6.8) and (6.9) are therefore completely determined by the statistical properties of $\hat{R}$. The estimator $\hat{R}$ is an unbiased estimator of R if

$$E(\hat{R}) = R. \tag{6.10}$$

This is generally not the case, and $\hat{R}$ is therefore *biased*. However, for large enough samples $\bar{x}$ will be very similar to $\bar{X}$, in which case

$$E(\hat{R}) = E(\frac{\bar{y}}{\bar{x}}) \doteq E(\frac{\bar{y}}{\bar{X}}) = \frac{E(\bar{y})}{\bar{X}} = \frac{\bar{Y}}{\bar{X}} = R, \tag{6.11}$$

implying that the bias will be negligible in large samples.

Table 6.1: An artificial population of 3 units with values for two variables y and x.

i	1	2	3
y_i	12	8	16
x_i	4	6	8

Example 6-1. Determine all simple random samples without replacement of size two for the artificial population shown in Table 6.1. Calculate $E(\hat{R})$ from all these samples and compare with the actual population ratio R.

Answer. The total distinct number of simple random samples without replacement of size $n = 2$ from a population of $N = 3$ is $M = \frac{N!}{n!(N-n)!} = \frac{3!}{2!1!} = 3$. These three samples are shown in Table 6.2, including their means for y and for auxiliary variable x. The average ratio of the latter two means is $E(\hat{R}) = \frac{\sum_{j=1}^{M} \frac{\bar{y}_j}{\bar{x}_j}}{M} = \frac{6.047}{3} = 2.016$. The actual ratio equals $R = \frac{\bar{Y}}{\bar{X}} = \frac{12}{6} = 2$, showing that $\hat{R}$ is indeed biased.

Table 6.2: Calculation of expected and actual ratio based on all simple random samples of size two from artificial population in Table 6.1.

Units in the sample		$\bar{y}$	$\bar{x}$	$\hat{R} = \frac{\bar{y}}{\bar{x}}$
1	2	10	5	2
1	3	14	6	2.333
2	3	12	7	1.714
Total		36	18	6.047
Expectation		12	6	2.016

If variables y_i and x_i are measured on each unit of a simple random sample of size n, assumed large (so that $\bar{x} \doteq \bar{X}$), the mean squared error (MSE) and variance of $\hat{R} = \frac{\bar{y}}{\bar{x}}$ are each approximately

$$\mathrm{MSE}(\hat{R}) \doteq \mathrm{V}(\hat{R}) = \sigma^2_{\hat{R}} \doteq \frac{1 - \frac{n}{N}}{n\bar{X}^2} \left[\frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1} \right], \tag{6.12}$$

where $R = \frac{\bar{Y}}{\bar{X}}$ is the ratio of the population means, see Cochran (1977, pp.31-32) for a proof.

Similarly, assuming n large, the approximate variance of the ratio estimator of the population total is

$$\mathrm{V}(\hat{Y}_R) = \sigma^2_{\hat{Y}_R} \doteq \frac{N(N-n)}{n} \left[\frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1} \right], \tag{6.13}$$

while the approximate variance of the ratio estimator of the population mean is

$$\mathrm{V}(\hat{\bar{Y}}_R) = \sigma^2_{\hat{\bar{Y}}_R} \doteq \frac{1-\frac{n}{N}}{n}\left[\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}\right]. \tag{6.14}$$

A sample estimate of the term

$$S_R^2 = \frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \tag{6.15}$$

in (6.12), (6.13), and (6.14) is

$$s_R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1} \tag{6.16}$$

which has a bias of order $\frac{1}{n}$. This gives the following estimated variance of $\hat{R}$:

$$v(\hat{R}) = \text{estimated}\,\sigma^2_{\hat{R}} = \frac{1-\frac{n}{N}}{n\bar{X}^2}\left[\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1}\right], \tag{6.17}$$

where the sample estimate $\bar{x}$ is substituted in the denominator if $\bar{X}$ is not known. Similarly, the estimated variance of $\hat{Y}_R$ is

$$\mathrm{v}(\hat{Y}_R) = \text{estimated}\,\sigma^2_{\hat{Y}_R} = \frac{N(N-n)}{n}\left[\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1}\right], \tag{6.18}$$

while the estimated variance of $\hat{\bar{Y}}_R$ equals

$$\mathrm{v}(\hat{\bar{Y}}_R) = \text{estimated}\,\sigma^2_{\hat{\bar{Y}}_R} = \frac{1-\frac{n}{N}}{n}\left[\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1}\right]. \tag{6.19}$$

Example 6-2. Determine $\mathrm{V}(\hat{R})$ from Table 6.2. Compare this result with approximation (6.12) to calculate the variance of $\hat{R}$.

Answer. From Table 6.2 we obtain $\mathrm{V}(\hat{R}) = E(\hat{R} - E(\hat{R}))^2 = \frac{1}{3}[(2-2.016)^2 + (2.333 - 2.016)^2 + (1.714 - 2.016)^2] = 0.0640$. Formula (6.12), on the other hand, yields $\mathrm{V}(\hat{R}) = \frac{(1-\frac{2}{3})}{(2)(6^2)}\frac{32}{(3-1)} = 0.0741$, showing that (6.12) indeed only gives an approximation.

Example 6-3. For a shipment of apples we need to estimate the total weight of apple juice that can be extracted from these apples. We draw a random sample of 15 apples from the shipment, and determine both the weight of each apple in the sample, and the weight of the juice extracted (in pounds), see Table 6.3. We do not know the total number of apples in the shipment, but we do know that the total weight of the shipment is 2000 pounds. Based on this information, what is the total amount of apple juice we may expect to obtain from this shipment? And what is the estimated variance of this estimate?

Table 6.3: Apple juice weight and apple weight in a random sample of 15 apples

Apple juice weight (y_i)	Apple weight (x_i)	$(y_i - \hat{R}x_i)^2$
0.16	0.22	0.0002209
0.15	0.26	0.0004634
0.20	0.31	0.0000204
0.25	0.37	0.0000348
0.16	0.28	0.0006112
0.27	0.38	0.0003727
0.28	0.40	0.0002596
0.16	0.21	0.0004605
0.11	0.18	0.0000766
0.16	0.29	0.0009809
0.17	0.26	0.0000023
0.24	0.32	0.0008346
0.21	0.33	0.0000594
0.11	0.16	0.0000198
0.22	0.35	0.0001189
2.85	4.32	0.0045360
0.19	0.288	

Answer. In Figure 6.1 we have plotted the juice weight y_i obtained from each apple in the random sample against its weight x_i, together with the best fitting regression line. Visual inspection of this figure already suggests that the two variables are indeed strongly correlated; in fact, their correlation is 0.943.

The regression equation of the best fitting line for these two variables is $\hat{y} = -0.009 + 0.691x$, see Figure 6.2. The t-test for the regression coefficient `B` of the independent variable x shows that the linear relation between apple juice and apple weight is very significant ($p < 0.001$). Moreover, the t-test for `(Constant)` in Figure 6.2 indicates that the intercept does not significantly deviate from zero, meaning that the use of a ratio estimator is warranted in this case. If the intercept happens to significantly deviate from zero, the *regression estimator* should be used, as will be explained in Section 6.2.

We first of all note that it is not possible to apply the direct estimator $\hat{Y}_d = N\bar{y}$, see (2.6) in Section 2.2, to estimate the total weight of the apple juice because we do not know N, the total number of apples in the shipment. However, using (6.7) we find from Table 6.3 that

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{15} y_i}{\sum_{i=1}^{15} x_i} = \frac{0.19}{0.288} = \frac{2.85}{4.32} = 0.6597,$$

and, according to (6.9) the estimated total weight of apple juice in the shipment is

$$\hat{Y}_r = \hat{R}X = (0.6597)(2000) = 1319.44 \,\text{pounds}.$$

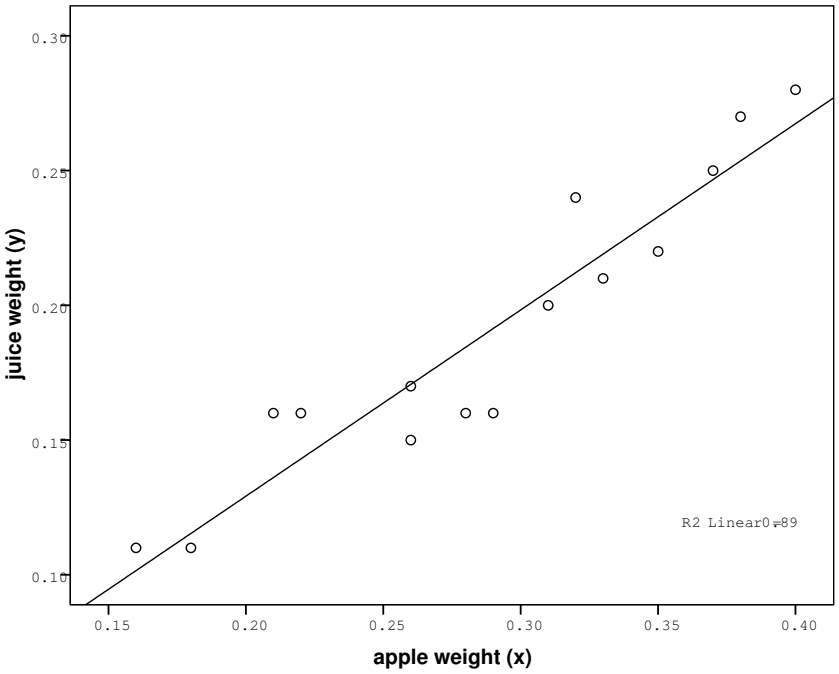

Figure 6.1: Scatter plot of juice weight y against apple weight x, including regression line

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.009	.020		-.451	.659
	apple weight	.691	.068	.943	10.233	.000

a. Dependent Variable: juice weight

Figure 6.2: Results of linear regression of juice weight y on apple weight x

It further follows from (6.18) that the estimated variance of this total equals

$$\mathrm{v}(\hat{Y}_R) = \frac{N(N-15)}{15}\left[\frac{0.0045360}{14}\right]. \tag{6.20}$$

But N, the total number of apples in the shipment, is unknown. Still, we can obtain an estimate of N using the following reasoning. The average weight of all the apples in the shipment is $\bar{X} = \frac{1}{N}X$, meaning that $N = \frac{X}{\bar{X}}$. We know that $X = 2000$. We do not know $\bar{X}$, the average weight in the population, but we may consider $\bar{x} = 0.288$, the average weight in the sample from Table 6.3, to be a reasonable approximation of $\bar{X}$. So $N = \frac{X}{\bar{X}} \doteq \frac{X}{\bar{x}} = \frac{2000}{0.288} = 6944.44$. Substituting this estimate for the total number of apples in (6.20) yields

$$\mathrm{v}(\hat{Y}_R) = \frac{6944.44(6944.44-15)}{15}\left[\frac{0.0045360}{14}\right] = (3208072.02)(0.000324) = 1039.42.$$

The point of this example is, of course, that the weight of the total shipment is easily determined while determination of the juice weight is much more time consuming.

Example 6-4. In this example taken from Cochran (1977, p.152) we have the number of inhabitants of a random sample of 49 cities in the United States drawn from a population of 169 large US cities, both for the year 1920 (x) and for the year 1930 (y), see Table 6.4. The true total X of the number of inhabitants in the population in 1920 is known to be $22,919$. Based on this information, estimate the total number of inhabitants in the population of 196 cities in 1930. Also calculate the variance of this ratio estimate, and compare the latter variance with that of the direct estimator for Y.

Answer. In Figure 6.3 we have plotted the number of inhabitants in 1930 from each city in the sample against the number of inhabitants in 1920, together with the best fitting regression line. Visual inspection of this scatter plot suggests that the two variables are indeed correlated; in fact, their correlation is 0.982.

The regression equation of the best fitting line for these two variables is $\hat{Y} = 8.384 + 1.158X$, see Figure 6.4. The t-test for the regression coefficient `B` is very significant ($p < 0.001$), meaning that there is indeed a strong linear relationship between the two variables. Also, the t-test for the `(Constant)` in Figure 6.4 indicates that the intercept is not significantly different from zero, at least not at the conventional 5% level. This means that a ratio estimate of the total is warranted in this case.

For the data in Table 6.4, we have that $\sum x_i = 5054$ and $\sum y_i = 6262$ from which it follows that $\hat{R} = \frac{6262}{5054} = 1.2390$ and $\hat{Y}_R = \hat{R}X = (1.2390)(22,919) = 28,397$. If we choose to neglect the information available from 1920 we find that $\hat{Y}_d = N\bar{y} = (196)(\frac{6262}{49}) = 25,048$ in 1930. The actual total for 1930 is 29,351. In terms of precision we find that $s_R^2 = \frac{\sum(y_i - \hat{R}x_i)^2}{n-1} = \frac{29782.88}{48} = 620.48$, meaning that estimated variance of this total is

$$\mathrm{v}(\hat{Y}_R) = \frac{196(196-49)}{49}(620.48) = 364,842.24,$$

see (6.18). Neglecting the information available from 1920, on the other hand, we would find

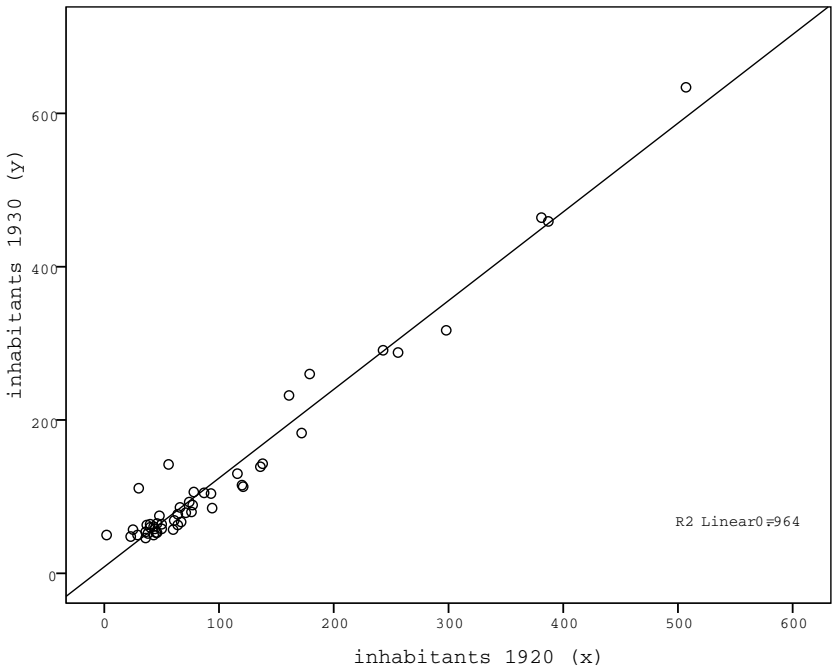

Figure 6.3: Scatter plot of number of inhabitants in 1930 against number of inhabitants in 1920, including regression line

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8.384	4.777		1.755	.086
	inhabitants 1920	1.158	.033	.982	35.383	.000

a. Dependent Variable: inhabitants 1930

Figure 6.4: Results of linear regression of number of inhabitants in 1930 on number of inhabitants in 1920

Table 6.4: Number of inhabitants (in 1000's) in 1920 (x_i) and 1930 (y_i) in a random sample of 49 US cities

x_i	y_i	x_i	y_i	x_i	y_i
76	80	2	50	243	291
138	143	507	634	87	105
67	67	179	260	30	111
29	50	121	113	71	79
381	464	50	64	256	288
23	48	44	58	43	61
37	63	77	89	25	57
120	115	64	63	94	85
61	69	64	77	43	50
387	459	56	142	298	317
93	104	40	60	36	46
172	183	40	64	161	232
78	106	38	52	74	93
66	86	136	139	45	53
60	57	116	130	36	54
46	65	46	53	50	58
				48	75

that

$$\mathrm{v}(\hat{Y}_d) = \frac{N^2 \hat{s_y}^2}{n}\left(\frac{N-n}{N}\right) = \frac{(196^2)(15,158.832)}{49}\left(\frac{196-49}{196}\right) = 8,913,393.22,$$

see (2.29). As far as this example is concerned, the ratio estimator therefore yields an estimate of the total number of inhabitants in the population of 196 US cities in 1930 that is more than 24 times more precise than the direct estimator.

6.1.2 The ratio estimator in stratified random sampling

In this section we discuss two ways in which a ratio estimate of the population total Y can be made in stratified random sampling, see Cochran (1977, p.164 cf). The first is to make a *separate* ratio estimate of the total of each stratum and then add the totals. The second way is to use a *combined* ratio estimate.

For the *separate* ratio estimator of the total we have the formulas

$$\hat{Y}_{Rs} = \sum_{h=1}^{L} \hat{R}_h X_h = \sum_{h=1}^{L} \frac{\bar{y}_h}{\bar{x}_h} X_h, \tag{6.21}$$

the subscript Rs standing for the *separate ratio estimator*, while its variance is estimated from the sample with

$$\mathrm{v}(\hat{Y}_{Rs}) = \text{estimated } \sigma^2_{\hat{Y}_{Rs}} = \sum_{h=1}^{L} \frac{N_h^2(1-\frac{n_h}{N_h})}{n_h}(s_{yh}^2 - 2\hat{R}_h s_{yxh} + \hat{R}_h^2 s_{xh}^2). \tag{6.22}$$

In these formulas, $\bar{y}_h$ and $\bar{x}_h$ are the sample means in stratum h, $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$, and X_h is the total of the auxiliary variable X in stratum h. Moreover, s_{yh} and s_{xh} are the sample standard deviations in stratum h and

$$s_{yxh} = \frac{\sum i = 1^{n_h}(y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{n_h - 1}, \tag{6.23}$$

i.e., the sample covariance between y and x in stratum h.

In order to calculate the separate ratio estimator we need to know the totals X_h in stratum h. If these stratum totals are unknown, but the population total X is known, then the *combined* ratio estimator is used:

$$\hat{Y}_{Rc} = \hat{R}X = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h} X, \tag{6.24}$$

the subscript Rc standing for the *combined ratio estimator*, where $\hat{R} = \frac{\hat{Y}}{\hat{X}}$, and $\hat{Y}$ and $\hat{X}$ have been estimated from the stratified sample using the direct estimator. Its estimated variance is

$$\mathrm{v}(\hat{Y}_{Rc}) = \text{estimated}\,\sigma^2_{\hat{Y}_{Rc}} = \sum_{h=1}^{L} \frac{N_h^2(1 - \frac{n_h}{N_h})}{n_h}(s_{yh}^2 - 2\hat{R}s_{yxh} + \hat{R}^2 s_{xh}^2). \tag{6.25}$$

If the sample sizes per stratum are small, it is better to use the combined ratio estimator than the separate ratio estimator, because the bias can become quite large for the latter estimator in this case. On the other hand, the separate ratio estimator is usually more precise, especially when there are large differences between the stratum ratios R_h.

6.1.3 Estimation of sample size for the ratio estimator

To find the sample size needed to estimate $\mu = \bar{Y}$ when the ratio estimator is used, let $d = |\mu - \bar{y}|$ be the margin of error we are willing to tolerate, let t be the value of the normal deviate corresponding to the desired confidence probability. Then the minimum sample size required follows from

$$d = t\sigma_{\hat{\bar{Y}}_R} = t\sqrt{\left(\frac{N-n}{N}\right)\frac{S_R^2}{n}} \tag{6.26}$$

with S_R^2 defined in (6.15). Solving (6.26) with respect to n we obtain

$$n = \frac{t^2 S_R^2}{d^2 + \frac{t^2 S_R^2}{N}}, \tag{6.27}$$

which reduces to

$$n_0 = \frac{t^2 S_R^2}{d^2} \tag{6.28}$$

when N is very large.

For the estimation of sample size based on the absolute error $d = |N\mu - N\bar{y}|$ of the *total* $N\mu = Y$ in the population we have that

$$d = t\sigma_{\hat{Y}_R}, \tag{6.29}$$

or, upon substitution of (6.13) in (6.29),

$$d = t\sqrt{\frac{N(N-n)S_R^2}{n}} \tag{6.30}$$

with S_R^2 defined in (6.15). Solving (6.30) for n, we find that

$$n = \frac{t^2 S_R^2}{\frac{t^2 S_R^2}{N} + \frac{d^2}{N^2}}, \tag{6.31}$$

or – if $\frac{n}{N} < 0.1$ –

$$n_0 = \frac{t^2 N^2 S_R^2}{d^2}. \tag{6.32}$$

Example 6-5. For the data in Example 6-4, what is – using a ratio estimator – the minimum sample size required in order to estimate the total number of inhabitants in the population of US cities with an absolute error margin of 1000, apart from a chance of 1 in 20? And what sample size would be needed using the direct estimator of this total?

Answer. Since $d = 1000$, $t = 1.96$, and $N = 196$ in this example, and letting $s_R^2 = 620.48$ be used as an estimate of S_R^2, it follows from (6.32) that

$$n_0 = \frac{t^2 N^2 S_R^2}{d^2} = \frac{(1.96^2)(196^2)(620.48)}{1000^2} = 91.57,$$

But since $\frac{n_0}{N} = \frac{92}{196} = 0.47 > 0.1$ we had better use (6.31) in order to correct for the finite population. This gives

$$n = \frac{t^2 S_R^2}{\frac{t^2 S_R^2}{N} + \frac{d^2}{N^2}} = \frac{(1.96^2)(620.48)}{\frac{(1.96^2)(620.48)}{196} + \frac{1000^2}{196^2}} = \frac{2383.64}{12.16 + 26.03} = 62.42.$$

The ratio estimator therefore only requires a sample size of 63 US cities. If we use the direct estimator, on the other hand, and acknowledge that $s_y^2 = 15,158.83$ (the variance of the sampled 49 US cities in 1930 in Table 6.4) may be used as an estimator of S_y^2, the actual variance in the population of 169 US cities in 1930, we find that

$$n = \frac{t^2 S_y^2}{\frac{t^2 S_y^2}{N} + \frac{d^2}{N^2}} = \frac{(1.96^2)(15,158.83)}{\frac{(1.96^2)(15,158.83)}{196} + \frac{1000^2}{196^2}} = 180.21,$$

according to (2.61) in Section 2.8. In this example, the direct estimator therefore requires a sample that covers almost 92% of the population, while the ratio estimator requires a sample of only 32% of the population, an almost threefold improvement.

6.2 The regression estimator

Like the ratio estimator, the linear regression estimator is used to increase precision by way of an auxiliary variable x_i that is correlated with y_i. Inspection of the relation between x_i and y_i may reveal that the two variables are approximately linearly related but that the regression line does not go through the origin. This suggests an estimator based on the linear regression of y_i on x_i rather than on the ratio of the two variables.

6.2.1 Properties of the regression estimator and its variance

Letting y_i and x_i be known for every unit in the sample, and let the population mean $\mu_x = \bar{X}$ of the x_i also be known. Then the linear regression estimator of $\bar{Y}$, the population mean of the y_i, is

$$\bar{y}_L = \bar{y} + b(\bar{X} - \bar{x}), \tag{6.33}$$

where the subscript L denotes *linear regression* and b is an estimate of the change in y when x is increased by one unit. For an estimate of the population total Y, we take

$$\hat{Y}_L = N\bar{y}_L. \tag{6.34}$$

Just like the ratio estimator, the regression estimator is also biased.

It is interesting to note some special cases of (6.33). If we let $b = 0$, (6.33) reduces to $\bar{y}_L = \bar{y}$, and we have the direct estimator of a simple random sample. If we choose $b = \frac{\bar{y}}{\bar{x}}$, we obtain from (6.33):

$$\bar{y}_L = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{\bar{Y}}_R,$$

the ratio estimator. If we finally set $b = 1$, and sample the whole population we find that

$$\bar{y}_L = \bar{y} + (\bar{X} - \bar{X}) = \bar{y} = \bar{Y}.$$

If b in (6.33) is the least squares estimate of β, the regression coefficient in the population, that is if

$$b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{6.35}$$

then in simple random samples of size n, with n large, the approximate variance of regression estimator (6.33) of the population mean is

$$\mathrm{V}(\hat{\bar{Y}}_L) = \sigma^2_{\hat{\bar{Y}}_L} \doteq \frac{1 - \frac{n}{N}}{n} S_y^2(1 - \rho^2), \tag{6.36}$$

where $\rho = \frac{S_{yx}}{S_y S_x}$, the population correlation between y and x, S_y and S_x are the standard deviation of y and x in the population, respectively, and S_{yx} is their population covariance. See Cochran (1977, p.194) for a proof. A sample estimate of this approximate variance, assuming large samples, is obtained with

$$\mathrm{v}(\hat{\bar{Y}}_L) = \text{estimated}\,\sigma^2_{\hat{\bar{Y}}_L} \doteq \frac{1 - \frac{n}{N}}{n(n-2)}\left(\sum(y_i - \bar{y})^2 - \frac{[\sum(y_i - \bar{y})(x_i - \bar{x})]^2}{\sum(x_i - \bar{x})^2}\right), \tag{6.37}$$

where we note that

$$SS_{\text{residual}} = \left(\sum (y_i - \bar{y})^2 - \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2} \right), \tag{6.38}$$

i.e., it is the residual sum of squares for the linear regression of y on x in the sample. The approximate variance of regression estimator (6.34) of the population total is

$$\mathrm{V}(\hat{Y}_L) = \sigma^2_{\hat{Y}_L} \doteq N(N-n)\frac{S_y^2}{n}(1-\rho^2), \tag{6.39}$$

while a sample estimate of this approximate variance, assuming large samples, is obtained with

$$\mathrm{v}(\hat{Y}_L) = \text{estimated}\, \sigma^2_{\hat{Y}_L} \doteq \frac{N(N-n)}{n(n-2)} \left(\sum (y_i - \bar{y})^2 - \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2} \right). \tag{6.40}$$

Example 6-6. A mathematics achievement test was given to 486 students before entering a certain university who then took a calculus course. At the end of the course a simple random sample of 10 students is taken and their score on a calculus test is obtained. It is known that the average score on the achievement test was 52 for all 486 students. The scores of the ten students on the achievement and calculus tests are given in Table 6.5. Based on these scores, what is the estimated average score on the calculus test in the total population of 486 students? And what is the estimated variance of this estimate? What are the answers to these two questions for the ratio estimator? And for the direct estimator?

Table 6.5: Calculus scores and achievement test scores in a random sample of 10 students

Student	Calculus score (y_i)	Achievement test score (x_i)
1	65	39
2	78	43
3	52	21
4	82	64
5	92	57
6	89	47
7	73	28
8	98	75
9	56	34
10	75	52
Total	760	460
Mean	76	46

Answer. In Figure 6.5 we have plotted the calculus test scores in the sample against the achievement test scores, together with the best fitting regression line. Visual inspection of this scatter plot suggests that the two variables are indeed correlated; in fact, their correlation

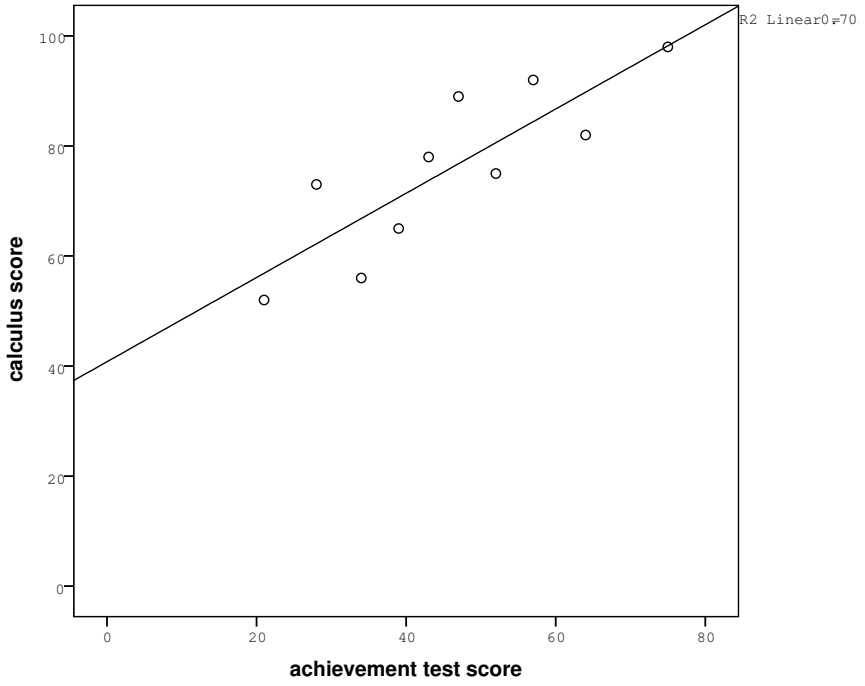

Figure 6.5: Scatter plot of calculus score against achievement test score, including regression line

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.784	8.507		4.794	.001
	achievement test score	.766	.175	.840	4.375	.002

a. Dependent Variable: calculus score

Figure 6.6: Results of linear regression of calculus score on achievement test score: parameter estimates

is 0.982. It is also clear from this figure, however, that the regression line does not go through the origin, meaning that the regression estimator should be used.

The regression equation of the best fitting line on the scores of the 10 students on these two variables is $\hat{Y} = 40.784 + 0.765562X$, see Figure 6.6. The t-test for the regression coefficient `B` is very significant ($p < 0.01$), meaning that there is indeed a strong linear relationship between the two variables. The t-test for the `(Constant)` in Figure 6.6 indicates that the intercept also very significantly deviates from zero ($p < 0.01$), confirming that the regression estimator of the average should be used in this case.

Since the sample means are $\bar{y} = 76$ and $\bar{x} = 46$, see Table 6.5, the regression estimate follows from (6.33) yielding

$$\bar{y}_L = \bar{y} + b(\bar{X} - \bar{x}) = 76 + (0.765562)(52 - 46) = 80.5934.$$

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1449.974	1	1449.974	19.141	.002[b]
	Residual	606.026	8	75.753		
	Total	2056.000	9			

a. Dependent Variable: calculus score

b. Predictors: (Constant), achievement test score

Figure 6.7: Results of linear regression of calculus score on achievement test score: ANOVA table

The sample estimate of the variance follows from (6.37):

$$\mathrm{v}(\hat{\bar{Y}}_L) \doteq \frac{1-\frac{10}{486}}{10(10-2)}\left(\sum(y_i-\bar{y})^2 - \frac{[\sum(y_i-\bar{y})(x_i-\bar{x})]^2}{\sum(x_i-\bar{x})^2}\right)$$
$$= (0.01224279835)\left(2056 - \frac{1894^2}{2474}\right) = (0.01224279835)(606.025869) = 7.42.$$

Using the ratio estimator (6.2) we obtain

$$\hat{\bar{Y}}_R = \bar{y}\frac{\bar{X}}{\bar{x}} = (76)(\frac{52}{46}) = 85.91,$$

while the estimated variance of $\hat{\bar{Y}}_R$ equals

$$\mathrm{v}(\hat{\bar{Y}}_R) = \frac{1-\frac{n}{N}}{n}\left[\frac{\sum_{i=1}^n (y_i-\hat{R}x_i)^2}{n-1}\right] = \left(\frac{1-\frac{10}{486}}{10}\right)(283.421)$$
$$= (0.09794238683)(283.421) = 27.76,$$

according to (6.19).

The (unbiased) direct estimator of the population mean is given by

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^n y_i = \bar{y} = 76,$$

see Chapter 2; according to (2.28) its variance is

$$\mathrm{v}(\hat{y}) = \frac{1-\frac{n}{N}}{n}\hat{s}^2 = \left(\frac{1-\frac{10}{486}}{10}\right)(228.44) = (0.09794238683)(228.44) = 22.37.$$

So we not only see that the regression estimator is the most precise, but also that the ratio estimator is even less precise than the direct estimator in this case.

6.2.2 The regression estimator in stratified random sampling

As with the ratio estimator, Cochran (1977, p.200 cf) also discusses two ways in which a regression estimate of the population total Y can be made in stratified random sampling.

The first way is again to make a *separate* regression estimate of the total of each stratum and then add the totals. The second way is to use a *combined* regression estimate.

For the *separate* regression estimator of the total we have the formulas

$$\hat{Y}_{Ls} = \sum_{h=1}^{L} N_h[\bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)], \tag{6.41}$$

the subscript Ls standing for the *separate linear regression estimator*, while its variance is estimated from the sample with

$$\mathrm{v}(\hat{Y}_{Ls}) = \text{estimated}\,\sigma^2_{\hat{Y}_{Ls}} = \sum_{h=1}^{L} \frac{N_h^2(1-\frac{n_h}{N_h})}{n_h}(s_{yh}^2 - 2b_h s_{yxh} + b_h^2 s_{xh}^2). \tag{6.42}$$

In these formulas, $\bar{y}_h$ and $\bar{x}_h$ are the sample means in stratum h, and $\bar{X}_h$ is the population mean of the auxiliary variable X in stratum h. Moreover, s_{yh} and s_{xh} are the sample standard deviations in stratum h, s_{yxh} is the sample covariance between y and x in stratum h (see (6.23) for the formula), and

$$b_h = \frac{s_{yxh}}{s_{xh}^2}, \tag{6.43}$$

i.e., the sample regression coefficient in the linear regression of y on x in stratum h.

This estimator is to be recommended when the regression coefficients β_h in the population are very different, and the sample sizes n_h are not too small. When the n_h are small there is the danger of bias, and it is better to use the *combined linear regression estimator*. In this case, the same regression coefficient b_c is used in each stratum:

$$\hat{Y}_{Ls} = \sum_{h=1}^{L} N_h[\bar{y}_h + b_c(\bar{X}_h - \bar{x}_h)], \tag{6.44}$$

with variance

$$\mathrm{v}(\hat{Y}_{Lc}) = \text{estimated}\,\sigma^2_{\hat{Y}_{Lc}} = \sum_{h=1}^{L} \frac{N_h^2(1-\frac{n_h}{N_h})}{n_h}(s_{yh}^2 - 2b_c s_{yxh} + b_c^2 s_{xh}^2). \tag{6.45}$$

The optimal value of b_c is the following weighted mean of the b_h:

$$b_c = \sum_{h=1}^{L} \frac{N_h^2(1-\frac{n_h}{N_h})s_{xh}^2/n_h}{\sum_{h=1}^{L} N_h^2(1-\frac{n_h}{N_h})s_{xh}^2/n_h} b_h, \tag{6.46}$$

which – upon substitution of (6.43) – can be written as

$$b_c = \sum_{h=1}^{L} \frac{N_h^2(1-\frac{n_h}{N_h})s_{yxh}/n_h}{\sum_{h=1}^{L} N_h^2(1-\frac{n_h}{N_h})s_{xh}^2/n_h}, \tag{6.47}$$

see Cochran (1977, Section 7.10).

6.2.3 Estimation of sample size for the regression estimator

To find the sample size needed to estimate $\mu = \bar{Y}$ when the regression estimator is used, let $d = |\mu - \bar{y}|$ be the margin of error we are willing to tolerate, let t be the value of the normal deviate corresponding to the desired confidence probability. Then the minimum sample size required follows from

$$d = t\sigma_{\hat{\bar{Y}}_L} = t\sqrt{\left(\frac{N-n}{N}\right)\frac{S_y^2}{n}(1-\rho^2)}, \tag{6.48}$$

see (6.36). Solving (6.48) with respect to n we obtain

$$n = \frac{t^2 S_y^2(1-\rho^2)}{d^2 + \frac{t^2 S_y^2(1-\rho^2)}{N}}, \tag{6.49}$$

which reduces to

$$n_0 = \frac{t^2 S_y^2(1-\rho^2)}{d^2} \tag{6.50}$$

when N is very large.

For the estimation of sample size based on the absolute error $d = |N\mu - N\bar{y}|$ of the *total* $N\mu = Y$ in the population we have that

$$d = t\sigma_{\hat{Y}_L}, \tag{6.51}$$

or, upon substitution of (6.39) in (6.51),

$$d = t\sqrt{N(N-n)\frac{S_y^2}{n}(1-\rho^2)}. \tag{6.52}$$

Solving (6.52) for n, we find that

$$n = \frac{t^2 S_y^2(1-\rho^2)}{\frac{d^2}{N^2} + \frac{t^2 S_y^2(1-\rho^2)}{N}}, \tag{6.53}$$

or – if $\frac{n}{N} < 0.1$ –

$$n_0 = \frac{t^2 N^2 S_y^2(1-\rho^2)}{d^2}. \tag{6.54}$$

Chapter 7

Systematic and repeated sampling

So far we have assumed that we need information on the population at only one point or during only one period in time. We now consider the situation where we need to observe a population two or more times, or during a prolonged period of time. One reason why we would like to draw two consecutive samples from the population is in order to obtain auxiliary information from the first sample, and then use this information to obtain a more precise estimate of the variable of interest in the second sample. This situation is called double sampling and is discussed in Section 7.3. A second reason why we could be interested in drawing two or more consecutive samples of the population is in order to be able to evaluate changes due to the effect of external forces acting on the population. The question then arises: is it best to repeatedly observe the same or different sampling units each time? This is the topic of Section 7.4.

First, however, we discuss yet another sampling method: systematic sampling.

7.1 Systematic sampling

In systematic sampling it is assumed that the N units in the population are numbered 1 to N in some order. To select a systematic sample of n units, a unit at random is selected from the first k units and every kth unit thereafter. If $k = 15$ for example, and if the first randomly selected unit is number 13, then the subsequent units are numbers 28, 43, 58, et cetera. The selection of the first unit determines the whole sample. In order for the sample to "cover" the whole population the value of the integer k should be chosen equal to, or approximately equal to $\frac{N}{n}$. This is called an *every kth* systematic sample.

Consider the artificial population shown in Table 7.1. If a systematic sample of 4 elements is to be selected from this population, then k is chosen to be $k = \frac{N}{n} = \frac{32}{4} = 8$ in order to cover the whole population. If the first randomly selected number happens to be 5, then the sample consists of element numbers 5, 13, 21, and 29, with values 11, 27, 49, and 65, respectively. This is row 5 in Table 7.1.

Since N is not in general an integral multiple of k, different systematic samples from the same finitie population may vary by one unit in size. With $N = 23$ and $k = 5$, for example, a total of five different systematic samples may be obtained. The three samples starting with unit numbers 1, 2, and 3 have $n = 5$, while those starting with unit numbers 4 and 5 have $n = 4$. This introduces a disturbance in the theory of systematic sampling. The disturbance is negligible if n is larger than 50 and will be ignored here. Even when n is small it is unlikely

Table 7.1: Population of 32 units numbered from 1 to 32

unit number	value y_i	unit number	value y_i	unit number	value y_i	unit number	value y_i	Total	Mean
1	1	9	13	17	37	25	53	104	26
2	3	10	17	18	39	26	61	120	30
3	3	11	19	19	43	27	63	128	32
4	5	12	21	20	41	28	57	124	31
5	11	13	27	21	49	29	65	152	38
6	11	14	25	22	47	30	65	148	37
7	11	15	31	23	51	31	71	164	41
8	15	16	33	24	57	32	75	180	45

to be large. Cochran (1977, p.206) also discusses an alternative systematic sampling method providing both a constant sample size and an unbiased sample mean.

A systematic sample can be conceived of as a kind of stratified sample with one observation in each stratum. However, whereas in stratified random sampling random numbers have to be drawn for each sample unit, in systematic sampling the first random number specifies for all strata which fixed unit is selected.

The first advantage of systematic sampling over random sampling is therefore that it only requires the drawing of one random number instead of n. And we only need to count simply k units further instead of using a count that changes all the time. If the units in the population are in a certain order, systematic sampling also guarantees that the sample is spread evenly over the population, whereas in random sampling some subpopulations may accidentally be underrepresented or even not represented at all.

If $N = nk$, the mean of a systematic sample, which we will denote by y_{sy}, is an unbiased estimate of $\bar{Y}$ for a randomly located systematic sample. Let y_{ij} denote the jth element of the ith systematic sample, $j = 1, \ldots, n$, $i = 1, \ldots, k$, and let $\bar{y}_i$ denote the mean of the ith sample. Then the variance of the mean of a systematic sample of size n is

$$\mathrm{V}(\bar{y}_{sy}) = \sigma^2_{\bar{y}_{sy}} = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}, \tag{7.1}$$

where

$$S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \tag{7.2}$$

is the variance among units that lie within the same systematic sample, see Cochran (1977, p.208) for a proof.

Example 7-0. Consider the artificial population in Table 7.1. What is the variance of the mean for a simple random sample of size $n = 4$ from this population? And the variance of the mean for a systematic sample of size $n = 4$?

Answer. The variance of the mean for a simple random sample of size $n = 4$ from this population is obtained from (2.18):

$$\mathrm{V}(\bar{y}) = \sigma^2_{\bar{y}} = \frac{S^2}{n},$$

with S^2 as defined in (2.8). Since the mean of the data in Table 7.1 is $\bar{Y} = \frac{1120}{32} = 35$, the variance of y in the population is

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}, = \frac{15,840}{31} = 510.9677,$$

from which it follows that the variance of the mean for a simple random sample of size $n = 4$ is

$$\mathrm{V}(\bar{y}) = \frac{510.9677}{4} = 127.74.$$

For the variance of the mean of a systematic sample from the population in Table 7.1 we first calculate (7.2):

$$S^2_{wsy} = \frac{1}{8(4-1)}(14,720) = 613.3333,$$

from which it follows that

$$\mathrm{V}(\bar{y}_{sy}) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S^2_{wsy} = \frac{32-1}{32}(510.9677) - \frac{8(4-1)}{32}(613.3333) = 35.$$

We conclude that – for these data at least – the systematic sample estimate is much more precise than the simple random sample estimate.

So how does systematic sampling generally compare with simple and stratified random sampling? We may distinguish four situations.

1. The manner in which the units of the population are ordered in the sampling frame is completely unrelated to the values of the variable of interest. This is often true when units are ordered alphabetically. In this situation a systematical sample will be as precise as a simple random sample, and the variance formulas of Chapter 2 can be used.

2. The numbering of the elements of the population in the sampling frame is based on some qualitative characteristic, e.g. the location where someone is living. The relative frequency of these characteristics in a systematic sample will then be very similar to their frequency in the population. We have basically applied proportional stratification, which is more precise than simple random sampling, see also Chapter 3.

3. The numbers assigned to the units of the population in the sampling frame globally increase or decrease with the values of a continuous auxiliary variable (e.g., some measure of size) that is highly correlated with the variable of interest. This has the effect that the means of the research variable are very different in each population layer of k elements. The implicit stratification then obtained with systematic sampling is much more precise than simple random sampling, see Example 7-0 where the numbering of the units is clearly correlated with the values of the y_i.

4. When there happens to be a periodic fluctuation in the population – meaning that population elements with rank numbers in the sampling frame that are a multiple of a certain value c have about the same value – then the variance of a systematic sample will underestimate the real variance in the population.

In systematic sampling there are some problems to obtain sample estimates of the variance of the mean. According to Moors and Muilwijk (1975) one often approximates this variance by using formula (2.28) for simple random sampling without replacement (in systematic sampling an element can never be selected more than once). In the just mentioned situation 1, no problems arise with this formula, while in situations 2 and 3 the variance will be estimated as being too large, meaning that we then obtain an estimate of precision that can only err on the safe side. Cochran (1977, p.225) provides dedicated formulas for calculating sample estimates of the variance of the mean in situations 2 and 3, but also warns that these variance estimates can be badly biased when applied to the wrong situation.

7.2 Stratified systematic sampling

Just as a simple random sample can be selected from each stratum in stratified random sampling, we can also draw a systematic sample from each stratum. In the latter case the starting points of the systematic samples in each stratum should be independently determined. This method will be more precise than stratified random sampling if systematic sampling within strata is more precise than simple random sampling with strata.

If $\bar{y}_{syh}$ is the mean of the systematic sample in stratum h, $h = 1, \ldots, L$, the estimate of the population mean $\bar{Y}$ and its variance are

$$\bar{y}_{stsy} = \sum_{h=1}^{L} W_h \bar{y}_{syh}, \tag{7.3}$$

and

$$\mathrm{V}(\bar{y}_{stsy}) = \sigma^2_{\bar{y}_{stsy}} = \sum_{h=1}^{L} W_h^2 \mathrm{V}(\bar{y}_{syh}). \tag{7.4}$$

The problem of finding a sample estimate of this variance is the same as discussed in the previous section.

7.3 Double sampling

Double sampling is a sampling method where initially a random sample is selected for the purpose of obtaining auxiliary information only, after which a second random sample is drawn from the first sample in order to observe the actual variable of interest in addition to the auxiliary variable. Double sampling, also known as two-phase sampling, is not only useful for obtaining auxiliary information for ratio and regression estimation (see Chapter 6), but also for finding information for stratified random sampling (see Chapter 3) and for handling non-response.

7.3.1 Double sampling for ratio estimation

Let y_i be the variable of interest, x_i be the auxiliary variable, n' be the number of units in the first sample, which includes the second sample, and let $n < n'$ be the number of units in the second sample. Variables x_i and y_i are only observed in the second sample. In the first sample only x_i is observed. It is assumed that the observation of the y_i is expensive whereas the observation of the x_i is (much) cheaper.

If the first sample is used to obtain

$$\bar{x}' = \frac{1}{n'} \sum_{i=1}^{n'} x_i, \tag{7.5}$$

as an estimate of $\bar{X}$ in a ratio estimate of the population mean $\bar{Y}$, the ratio estimator of the population mean $\bar{Y}$ is

$$\bar{y}_R = \hat{R}\bar{x}' = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \bar{x}', \tag{7.6}$$

with approximate variance equal to

$$\mathrm{V}(\bar{y}_R) = \sigma^2_{\bar{y}_R} \doteq \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) (S_y^2 - 2RS_{yx} + R^2 S_x^2). \tag{7.7}$$

and estimated variance equal to

$$\mathrm{v}(\hat{Y}_R) = \text{estimated}\, \sigma^2_{\hat{Y}_R} = \frac{(N-n')}{N} \frac{s_y^2}{n'} + \frac{n'-n}{n'n(n-1)} \sum_{i=1}^{n} (y_i - \hat{R}x_i)^2. \tag{7.8}$$

The ratio estimator of the population total Y is

$$\hat{Y}_R = \hat{R}\hat{X} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \hat{X}, \tag{7.9}$$

and

$$\hat{X} = \frac{N}{n'} \sum_{i=1}^{n'} x_i. \tag{7.10}$$

The estimated variance of this estimator is

$$\mathrm{v}(\hat{Y}_R) = \text{estimated}\, \sigma^2_{\hat{Y}_R} = N(N-n') \frac{s_y^2}{n'} + N^2 \frac{n'-n}{n'n(n-1)} \sum_{i=1}^{n} (y_i - \hat{R}x_i)^2. \tag{7.11}$$

Example 7-1. A forest resource manager is interested in estimating the total number of dead trees in a 400 acre area of heavy infestation. He divides the area in 200 plots of equal size and uses arial photo counts to find the number of dead trees in a random sample of 18 plots. He then randomly samples 8 plots out of these 18 plots and conducts a ground count

Table 7.2: Counts of dead trees in a random sample of 8 plots

Plot	arial photo count first sample (x_i')	arial photo count second sample (x_i)	ground count (y_i)
1	5		
2	7	7	9
3	10	10	13
4	6		
5	7	7	10
6	9	9	11
7	3		
8	6		
9	8		
10	11		
11	5		
12	9	9	10
13	12		
14	13		
15	3	3	4
16	20	20	25
17	15	15	17
18	4		
Total	153	80	99

on these 8 plots. The resulting counts are given in Table 7.2. Based on these counts, what is the estimated total number of dead trees in the 400 acre area? And what is the estimated variance of this estimate?

Answer. In this example $N = 200$, $n' = 18$, and $n = 8$, and the estimated total number of dead trees in the 400 acre area is

$$\hat{Y}_R = \hat{R}\hat{X} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \frac{N}{n'} \sum_{i=1}^{n'} x_i = \left(\frac{99}{80}\right)\left(\frac{200}{18}\right)(153) = 2,103.75.$$

Since $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 39.4384$ and $\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2 = 6.1928$, the estimated variance of this estimate equals

$$\begin{aligned}
\mathrm{v}(\hat{Y}_R) =& N(N-n')\frac{s^2}{n'} + N^2 \frac{n'-n}{n'n(n-1)} \sum_{i=1}^{n}(y_i - \hat{R}x_i)^2 \\
=& 200(200-18)\left(\frac{39.4384}{18}\right) + 200^2 \left(\frac{18-8}{(18)(8)(8-1)}\right)(6.1928) \\
=& 79,753.20889 + 2,457.460317 = 82,210.66921.
\end{aligned}$$

7.3.2 Double sampling for regression estimation

In some applications of double sampling the auxiliary variable x_i is used to obtain a regression estimate of the population mean $\bar{Y}$. Just as in double sampling for ratio estimation, in the first (large) sample of size n', only x_i is measured. In the second step, a random subsample of size $n = \frac{n'}{k}$ is taken, and both x_i and y_i are measured.

Letting y_i and x_i be known for every unit in the sample, and let the population mean $\mu_x = \bar{X}$ of the x_i also be known. Then the linear regression estimator of $\bar{Y}$, the population mean of the y_i, is

$$\bar{y}_L = \bar{y} + b(\bar{x}' - \bar{x}), \tag{7.12}$$

where $\bar{x}'$ and $\bar{x}$ are the means of the x_i in the first and second samples, and b is the least squares regression coefficient of y_i on x_i, computed for the second sample. Just as in one-stage sampling, this regression estimator will be biased, see also Section 6.2.

Assuming random sampling and $\frac{1}{n'}$ and $\frac{1}{n}$ negligible withe respect to 1, the approximate variance of regression estimator (7.12) of the population mean is

$$\mathrm{V}(\hat{\bar{Y}}_L) = \sigma^2_{\hat{\bar{Y}}_L} \doteq \frac{S_y^2(1-\rho^2)}{n} + \frac{S_y^2\rho^2}{n'} - \frac{S_y^2}{N}, \tag{7.13}$$

where $\rho = \frac{S_{yx}}{S_y S_x}$, the population correlation between y and x, S_y and S_x are the standard deviation of y and x in the population, respectively, and S_{yx} is their population covariance. See Cochran (1977, p.339) for a proof.

If terms in $\frac{1}{n}$ are negligible, a sample estimate of $\mathrm{V}(\hat{\bar{Y}}_L)$ is

$$\mathrm{v}(\hat{\bar{Y}}_L) = \text{estimated}\,\sigma^2_{\hat{\bar{Y}}_L} \doteq \frac{s^2_{y.x}}{n} + \frac{s_y^2 - s^2_{y.x}}{n'} - \frac{s_y^2}{N}, \tag{7.14}$$

where

$$s^2_{y.x} = \frac{1}{n-2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2(\sum_{i=1}^{n}(x_i - \bar{x})^2)\right) \tag{7.15}$$

is an unbiased estimate of $S_y^2(1-\rho^2)$ and

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{7.16}$$

is an unbiased estimate of S_y^2.

If the second sample is small and terms in $\frac{1}{n}$ are not negligible relative to 1, an estimate of variance is

$$\mathrm{v}(\hat{\bar{Y}}_L) = \text{estimated}\,\sigma^2_{\hat{\bar{Y}}_L} = s^2_{y.x}\left(\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2)}\right) + \frac{s_y^2 - s^2_{y.x}}{n'} - \frac{s_y^2}{N}, \tag{7.17}$$

see also Cochran (1977, p.343).

7.3.3 Double sampling for stratification

Double sampling can also be used when stratified random sampling is the method of preference, but the sizes of the different strata in the population are unknown. A large simple random sample of size n' is drawn from the total population of N units, and the only observation made in this fist large sample is how many units n'_h $(h = 1, \ldots, L)$ are in each stratum. Unbiased estimates of the population stratum sizes $W_h = \frac{N_h}{N}$ are then obtained from $w_h = \frac{n'_h}{n'}$. Next, a second sample is selected by stratified random sampling from the first sample. These units are classified into strata with n_h units selected from the n'_h sample units in stratum h, and the variable of interest y_{hi} is obtained for each unit in this second sample. Letting $\bar{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$ denote the sample mean of stratum h in the second sample, an unbiased estimate of the population mean $\bar{Y} = \sum W_h \bar{Y}_h$ is obtained with

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h. \tag{7.18}$$

The objective of the first sample is therefore to estimate the strata weights W_h; that of the second sample is to estimate the strata means $\bar{Y}_h$. The variance of $\bar{y}_{st}$ is

$$\mathrm{V}(\bar{y}_{st}) = \sigma^2_{\bar{y}_{st}} = S^2 \left(\frac{1}{n'} - \frac{1}{N} \right) + \sum_{h=1}^{L} \frac{W_h S_h^2}{n'} \left(\frac{n'_h}{n_h} - 1 \right), \tag{7.19}$$

where S^2 is the population variance. An unbiased sample estimate for this variance is

$$\begin{aligned} \mathrm{v}(\bar{y}_{st}) = \text{estimated}\, \sigma^2_{\bar{y}_{st}} &= \left(\frac{N - n'}{N} \right) \left(\frac{1}{n' - 1} \right) \sum_{h=1}^{L} w_h (\bar{y}_h - \bar{y}_{st})^2 \\ &+ \frac{N-1}{N} \sum_{h=1}^{L} \left(\frac{n'_h - 1}{n' - 1} - \frac{n_h - 1}{N - 1} \right) \frac{w_h s_h^2}{n_h}, \end{aligned} \tag{7.20}$$

where s_h^2 is the stratum sample variance in the second sample.

Example 7-2. A shoe store wants to estimate the average number of pairs of shoes owned by the students living in a certain college town neighbourhood. They reason that a stratified sample based on gender could be a good approach but do not know the distribution of the gender in that neighbourhood. They also do not know the gender of the respondent until after contacting them. They therefore use double sampling by first contacting 160 randomly selected students in that neighbourhood and ask them about their gender. They find that this first sample consists of 64 males and 96 females. They next randomly sample 8 males and 12 females from the 64 males and 96 females in this first sample, and ask them to count the number of pairs of shoes that they have at home and report back to them. The data for this second sample are given in Table 7.3. Based on these observations, what is the estimated average number of pairs of shoes these students own? And what is the estimated variance of this average?

Answer. In this example the size of the first sample is $n' = 160$, with $n'_1 = 64$ males and $n'_2 = 96$ females. The unknown proportion $W_1 = \frac{N_1}{N}$ of males in the population is

Table 7.3: Number of pairs of shoes owned by 8 male and 12 female students

	Male	Female
	5	17
	6	19
	9	13
	5	16
	9	8
	7	11
	5	15
	8	19
		12
		13
		33
		20
Total	54	196

therefore estimated to be $w_1 = \frac{n'_1}{n'} = \frac{64}{160} = 0.4$, and the unknown proportion $W_2 = \frac{N_2}{N}$ of females in the population is estimated to be $w_2 = \frac{n'_2}{n'} = \frac{96}{160} = 0.6$. In the second sample, the average number of pairs of shoes for males is found to be $\bar{y}_1 = \frac{54}{8} = 6.75$ with a variance of $s_1^2 = 3.073$, and that for females is found to be $\bar{y}_2 = \frac{196}{12} = 16.33$ with a variance of $s_2^2 = 40.5769$, see Table 7.3. The estimated average number of pairs of shoes owned by the total student population in that neighbourhood is therefore estimated to be

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h = (0.4)(6.75) + (0.6)(16.33) = 12.498.$$

We do not know N, the total number of students in that neighbourhood. However, if we assume it to be quite large, (7.20) can be simplified to

$$\mathrm{v}(\bar{y}_{st}) = \left(\frac{1}{n'-1}\right) \sum_{h=1}^{L} w_h (\bar{y}_h - \bar{y}_{st})^2 + \sum_{h=1}^{L} \left(\frac{n'_h - 1}{n'-1}\right) \frac{w_h s_h^2}{n_h},$$

yielding

$$\mathrm{v}(\bar{y}_{st}) = \left(\frac{1}{160-1}\right) \left(0.4(6.75-12.498)^2 + 0.6(16.33-12.498)^2\right)$$
$$+ \left(\frac{64-1}{160-1}\right) \frac{(0.4)(3.073)}{8} + \left(\frac{96-1}{160-1}\right) \frac{(0.6)(40.5769)}{12} = 0.13853 + 1.2731 = 1.4116.$$

7.3.4 Double sampling for non-response

An ingenious application of double sampling is the adjustment for non-response in a survey. In this case the first sample of size n' is a simple random sample from a population of N

units. These units are then stratified in two strata: n_1' contains those sample units that respond, and $n_2' = n' - n_1'$ consists of the sample units that do not respond. In the second phase another random sample of size $n_2 = \frac{n_2'}{k} < n_2'$ is drawn from the n_2' non-respondents in the first round and a massive effort is made to obtain information on the variable of interest for all these n_2 units. We have then a double sampling setting where $n_1' = n_1$, and n_2 is the size of the second sample.

Example 7-3. In a university of 1000 students, a questionnaire is mailed to a simple random sample of 106 students asking them about the amount of time they spend studying each week. Out of these 106 students 46 respond. From the 60 non-respondents, a simple random sample of 20 is selected and intensive efforts are made by telephone and personal visit to obtain responses. This yields the data shown in Table 7.4. Based on these observations, what is the estimated average number of hours these students study each week? And what is the estimated variance of this average?

Table 7.4: Double sampling for average number of hours spent studying per week

	Students responding to questionnaire (first round)	Students contacted and responding to telephone and visit (second round)
Sample mean	20.5	10.9
Sample standard deviation	6.2	5.1
Sample size	46	20

Answer. In this example the first sample of $n' = 106$ students can be stratified into the following two strata: the respondents with a size of $n_1' = 46$ and the non-respondents with a size of $n_2' = 60$. The unknown proportion $W_1 = \frac{N_1}{N}$ of respondents in the population is therefore estimated to be $w_1 = \frac{n_1'}{n'} = \frac{46}{106} = 0.434$, and the unknown proportion $W_2 = \frac{N_2}{N}$ of non-respondents in the population is estimated to be $w_2 = \frac{n_2'}{n'} = \frac{60}{106} = 0.566$. In the next step a second random sample of $n_2 = 20$ is drawn from the $n_2' = 60$ students in the non-respondent stratum of the first sample. The average of the first stratum in the first sample is $y_1 = 20.5$ hours with a standard deviation of $s_1 = 6.2$; the average of the second stratum in the second sample is $y_2 = 10.9$ hours with a standard deviation of $s_2 = 5.1$, see Table 7.4. The estimated average number of hours spent studying in the total population of $N = 1000$ students is therefore estimated to be

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h = (0.434)(20.5) + (0.566)(10.9) = 15.0664.$$

According to (7.20), and because $n_1' = n_1 = 46$ in this case, the estimated variance of this

mean is

$$\begin{aligned}\mathrm{v}(\bar{y}_{st}) &= \left(\frac{N-n'}{N}\right)\left(\frac{1}{n'-1}\right)\sum_{h=1}^{L} w_h(\bar{y}_h-\bar{y}_{st})^2 + \frac{N-1}{N}\sum_{h=1}^{L}\left(\frac{n'_h-1}{n'-1}-\frac{n_h-1}{N-1}\right)\frac{w_h s_h^2}{n_h}\\ &= \left(\frac{1000-106}{1000}\right)\left(\frac{1}{106-1}\right)\left(0.434(20.5-15.0664)^2+0.566(10.9-15.0664)^2\right)\\ &+\frac{1000-1}{1000}\left((\frac{46-1}{106-1}-\frac{46-1}{1000-1})\frac{(0.434)(6.2)^2}{46}+(\frac{60-1}{106-1}-\frac{20-1}{1000-1})\frac{(0.566)(5.1)^2}{20}\right)\\ &= 0.1928+0.5382=0.731.\end{aligned}$$

Correcting for non-response with the double sampling procedure yields a substantially smaller estimate of 15 for the average number of hours spent studying by these students than the estimate of 20.5 hours found for the first phase respondents only.

With double sampling for non-response it is also possible to estimate the required sample size of n' for given budget. Let the cost of taking the sample be

$$c_0 n' + c_1 n'_1 + \frac{c_2 n'_2}{k}, \tag{7.21}$$

where the c's are the costs per unit: c_0 is the cost of making the first attempt, c_1 is the cost of processing the results from the first attempt,and c_2 is the cost of getting and processing the data in the second stratum. Letting W_1 and W_2 denote the population proportions in the two strata of respondents and non-respondents then the expected cost is

$$C = c_0 n' + c_1 W_1 n' + \frac{c_2 W_2 n'}{k}, \tag{7.22}$$

since $W_h n' = \frac{N_h}{N} n' = \frac{n'_h}{n'} n' = n'_h$, $w_h = \frac{n'_h}{n'}$ being an unbiased estimate of $W_h = \frac{N_h}{N}$. It follows from (7.18) that an unbiased estimate of $\bar{Y}$ is obtained with

$$\bar{y}' = w_1\bar{y}_1 + w_2\bar{y}_2 = \frac{n'_1\bar{y}_1 + n'_2\bar{y}_2}{n'}, \tag{7.23}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the means of the samples of sizes $n_1 = n'_1$ and $n_2 = \frac{n'_2}{k}$. From (7.20) we obtain

$$\begin{aligned}\mathrm{V}(\bar{y}') =& S^2\left(\frac{1}{n'}-\frac{1}{N}\right)+\frac{W_1 S_1^2}{n'}\left(\frac{n'_1}{n_1}-1\right)+\frac{W_2 S_2^2}{n'}\left(\frac{n'_2}{n_2}-1\right)\\ =& S^2\left(\frac{1}{n'}-\frac{1}{N}\right)+\frac{W_2 S_2^2}{n'}\left(\frac{n'_2}{n_2}-1\right),\end{aligned} \tag{7.24}$$

since $n'_1 = n_1$. Substitution of $n_2 = \frac{n'_2}{k}$ in (7.24) yields

$$\mathrm{V}(\bar{y}') = S^2\left(\frac{1}{n'}-\frac{1}{N}\right)+\frac{W_2 S_2^2(k-1)}{n'}. \tag{7.25}$$

The quantities n' and k are then chosen to minimize the product $C(V+\frac{S^2}{N})$. From (7.25) and (7.22) we have

$$V + \frac{S^2}{N} = \frac{(S^2 - W_2 S_2^2)}{n'} + \frac{kW_2 S_2^2}{n'}. \tag{7.26}$$

and

$$C = (c_0 + c_1 W_1)n' + \frac{c_2 W_2 n'}{k}. \tag{7.27}$$

Optimizing $C(V + \frac{S^2}{N})$ with respect to k, it is not very difficult to verify that the optimum is found for

$$k_{opt} = \sqrt{\frac{c_2(S^2 - W_2 S_2^2)}{S_2^2(c_0 + c_1 W_1)}}. \tag{7.28}$$

The initial sample size n' can be chosen either to minimize C for given V, or V for given C by solving n' from (7.26) or (7.27). If V is specified, it follows from (7.26) that

$$n'_{opt} = \frac{N(S^2 + (k-1)W_2 S_2^2)}{NV + S^2}, \tag{7.29}$$

where V is the variance (7.25) of the estimated population mean. For given total budget C, on the other hand, it follows from (7.27) that

$$n'_{opt} = \frac{kC}{k(c_0 + C_1 W_1) + c_2 W_2}. \tag{7.30}$$

The solution requires a knowledge of W_2. This can often be estimated from previous experience. Moreover, in addition to S^2, the variance in the population whose value must be estimated in advance in any problem of sample size, we also need an estimate of S_2^2, the variance in the non-response stratum. The latter value may be harder to predict, and it will probably not be the same as S^2. In surveys made by mail of economic enterprise, for example, the respondents tend to be larger operators, with larger between-unit variances than the non-respondents.

If W_2 is not well-known, Cochran (1977) suggests to obtain an approximation of n'_{opt} from (7.28) and (7.29) by substituting a range of assumed values of W_2 between 0 and a safe upper limit, and then to use the maximum n'_{opt} in this series as the initial sample size n'. When the replies to the mail survey have been received, the value of n'_2 is known. In order to obtain a value for k with this method, the variance $\mathrm{V}_c(\bar{y}')$ conditional on the known values of n_2 and n' should be used. This variance is

$$\mathrm{V}_c(\bar{y}') = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \frac{n'_2 S_2^2(k-1)}{n'^2}. \tag{7.31}$$

Equation (7.31) is solved to find the k that gives the desired conditional variance. The cost for this method is usually only slightly higher than the optimum cost for known W_2.

With stratified sampling, the optimum values of the n'_h and the k_h in the individual strata are rather complex. According to Cochran (1977), a good approximation is to estimate first,

by the methods in Section 3.4 and Section 3.6, the sample sizes n_h that would be required in the strata if there were no non-response. Then, from (7.29), if $W_2 = 0$, we have

$$n_0' = \frac{NS^2}{NV + S^2}, \tag{7.32}$$

and (7.29) can therefore be written as

$$n_{opt}' = n_0 \left(1 + \frac{(k-1)W_2 S_2^2}{S^2}\right). \tag{7.33}$$

Applying the latter equation separately to each stratum gives an approximation to the optimal n_h'. The values of k_h are found by applying (7.28) in each stratum.

These techniques can also be used with the ratio or the regression estimator. With the ratio estimator, the quantities S^2 and S_2^2 are replaced by S_d^2 and S_{2d}^2, where $d_i = y_i - Rx_i$. With the regression estimator, S^2 becomes $S^2(1-\rho^2)$ and S_2^2 becomes $S_2^2(1-\rho^2)$.

Example 7-4. In a university of $1,000$ students, we want to find the average amount of money the students spend on living. We expect the non-response to be about 40%. It is also expected that the respondents have a larger variance than the non-respondents. The overall variance in the population is estimated to be $S^2 = 120$ while the variance for the non-respondents is estimated to be $S_2^2 = 80$. The initial costs of sampling each respondent is 0, the cost of a standard response is 1, and the cost of a call-back is 4. What is the number of students to sample in the first round, and how many students should be subjected to a call-back, if we require the variance of the estimator to be equal to 5?

Answer. We have $c_0 = 0$, $c_1 = 1$ and $c_2 = 4$. Moreover, $W_1 = 0.6$ and $W_2 = 0.4$. The optimal value for k, the fraction to sample from the non-respondents in the first sample, can be found with (7.28):

$$k_{opt} = \sqrt{\frac{c_2(S^2 - w_2 S_2^2)}{S_2^2(c_0 + c_1 w_1)}} = \sqrt{\frac{(4)(120 - (0.4)(80))}{(80)(1)(0.6)}} = \sqrt{\frac{352}{48}} = 2.71.$$

Since $V = 5$ the optimal size of the first sample is obtained using (7.29):

$$n_{opt}' = \frac{N(S^2 + (k-1)w_2 S_2^2)}{NV + S^2} = \frac{1000(120 + (2.71-1)(0.4)(80))}{(1000)(5) + 120} = \frac{174,720}{5,120} = 34.125.$$

Rounding up we have $n' = 35$, from which it follows that $n_2' = w_2 n' = (0.4)(35) = 14$. So for a precision of $V = 5$ we require a random sample of 35 students in the first phase, and we need to call back $n_2 = \frac{n_2'}{k} = \frac{14}{2.71} = 5.16$, i.e., 6 students in the second phase.

Example 7-5. In a survey the first sample is taken by mail and the response is expected to be about 50%. The precision desired is that which would be given by a simple random sample of size $1,000$ if there were no non-response. The cost of mailing a questionnaire is 10 cents, and the cost of processing the completed questionnaire is 40 cents. To carry out a

personal interview costs 4.10 euro's. How many questionnaires should be sent out and what percentage of the non-respondents should be interviewed?

Answer. We have $c_0 = 0.10$, $c_1 = 0.40$, and $c_2 = 4.50$. The precision desired is $V = \frac{S^2}{1000}$. Moreover, $W_1 = W_2 = 0.5$. If the variances S^2 and S_2^2 are assumed to be equal, and N is assumed to be large it follows from (7.28) that

$$k_{opt} = \sqrt{\frac{c_2(S^2 - W_2 S_2^2)}{S_2^2(c_0 + c_1 W_1)}} = \sqrt{\frac{c_2(1 - W_2)}{c_0 + c_1 W_1}} = \sqrt{\frac{(4.5)(0.5)}{0.10 + (0.40)(0.5)}} = 2.739,$$

and from (7.29) that

$$\begin{aligned} n'_{opt} =& \frac{N(S^2 + (k-1)W_2 S_2^2)}{NV + S^2} = \frac{(S^2 + (k-1)W_2 S_2^2)}{V + \frac{S^2}{N}} \doteq \frac{(S^2 + (k-1)W_2 S_2^2)}{V} \\ =& \frac{(S^2 + (k-1)W_2 S_2^2)}{\frac{S^2}{1{,}000}} = 1{,}000(1 + (k-1)W_2) = 1{,}000(1 + (2.739 - 1)(0.5)) = 1{,}869.5. \end{aligned}$$

This means that $1,870$ questionnaires should be mailed, and of the 935 expected not to be returned a random sample of $\frac{935}{2.739} = 341$ should be interviewed. The total cost of the survey is $C = (c_0 + c_1 W_1)n' + \frac{c_2 W_2 n'}{k} = (0.10 + (0.40)(0.5))(1,870) + \frac{(4.50)(0.5)(1,870)}{2.739} = 561 + 1,536 = 2,097$ euro's.

7.4 Continuous sampling

So far we have assumed that we need information on the population at only one point or during only one period in time. We now consider the situation where we need to investigate a population several times or during a prolonged period of time. The question then arises: is it best to repeatedly observe the same or different sampling units each time?

As discussed by Cochran (1977, Section 12.10), given the data from a consecutive series of samples, the answer to this question depends on the kind of quantity for which we may wish estimates:

- The *change* in $\bar{Y}$ from one occasion to the next?
- The *average value* of $\bar{Y}$ over *all occasions*?
- The *average value* of $\bar{Y}$ for the *most recent occasion*?

We first consider the case of two consecutive surveys. Let the population variables of interest on the two occasions be x and y, respectively, and their correlation be ρ. Further assume that both samples are random and of equal size, and that the finite population correction is negligible. Suppose that the randomly selected part W of the first sample is replaced by other randomly selected elements on the second occasion, while the other part $1 - W$ (the *permanent part*) remains the same. We distinguish three situations:

- The purpose of the data collection is the estimation of *changes* over time in the means $\bar{X}$ and $\bar{Y}$, or totals X and Y. This is most appropriate if we want to study the effects of forces that are known to have affected the population of interest. In this case we

want the estimated *difference* $\hat{\bar{X}} - \hat{\bar{Y}}$ to be as accurate as possible. Both the mean of the permanent part $\bar{x}'$ and the mean $\bar{x}''$ of the rest is then an unbiased estimator of $\bar{X}$. An unbiased estimator of $\bar{X}$ is obtained with the following weighted average:

$$\hat{\bar{X}} = W\bar{x}' + (1-W)\bar{x}'' \tag{7.34}$$

for the first survey, and

$$\hat{\bar{Y}} = W\bar{y}' + (1-W)\bar{y}'' \tag{7.35}$$

for the second survey, where $\bar{x}'$ and $\bar{y}'$ are obtained from the same sample fraction, and $\bar{x}''$ and $\bar{y}''$ from two different sample fractions. Because the different sample fractions are independent we have that

$$\mathrm{V}(\hat{\bar{Y}} - \hat{\bar{X}}) = W^2\mathrm{V}(\bar{y}' - \bar{x}') + (1-W)^2\mathrm{V}(\bar{y}'' - \bar{x}''). \tag{7.36}$$

For the two variances on the right side of (7.36) we find that

$$\mathrm{V}(\bar{y}' - \bar{x}') = \frac{1}{Wn}(S_x^2 - 2\rho S_x S_y + S_y^2), . \tag{7.37}$$

the sample size of this part being Wn, and

$$\mathrm{V}(\bar{y}'' - \bar{x}'') = \mathrm{V}(\bar{y}'') + \mathrm{V}(\bar{x}'') = \frac{1}{(1-W)n}(S_x^2 + S_y^2), \tag{7.38}$$

$\bar{y}''$ and $\bar{x}''$ being calculated from two independent samples of size $(1-W)n$. Substitution of (7.37) and (7.38) in (7.36) yields

$$\mathrm{V}(\hat{\bar{Y}} - \hat{\bar{X}}) = \frac{1}{n}(S_x^2 - 2W\rho S_x S_y + S_y^2). \tag{7.39}$$

When ρ is positive, which is the usual situation, it follows from (7.39) that the precision of the difference $\hat{\bar{X}} - \hat{\bar{Y}}$ is maximised by choosing W as large as possible. Since this largest value is $W = 1$, we see that precision in measuring change is largest when the random sample on the two occasions is *left unchanged*. In practice this is not always feasible, because part of the population ceases to exist (e.g., emigration) while another part of the population is renewed (e.g., immigration).

- The purpose of the data collection is to estimate the *sum* or *average* on the two occasions as accurately as possible. Again starting from (7.34) we now find that

$$\mathrm{V}(\hat{\bar{Y}} + \hat{\bar{X}}) = \frac{1}{n}(S_x^2 + 2W\rho S_x S_y + S_y^2). \tag{7.40}$$

Again assuming ρ to be positive, it follows from (7.40) that the precision of the sum $\hat{\bar{X}} + \hat{\bar{Y}}$ is maximised by choosing W as small as possible. Since this smallest value is $W = 0$, we see that precision in measuring the average or total of the two occasions is largest when the random sample of the first occasion is *completely replaced* on the second occasion.

- The purpose of the data collection is to estimate the value of $\hat{Y}$ or $\hat{\bar{Y}}$ as accurately as possible *on the second occasion*. In this case we can replace part, usually at least one half, of the sample and apply two-phase or double sampling (see Section 7.3) with a regression estimator.

When the population is observed on more than two occasions, for the already mentioned practical reasons one often uses *gradual rotation* of the sample. In the second year, for example, one quarter of the previous sample is replaced, in the third year yet another quarter of the original sample is replaced, et cetera, meaning that the whole sample will have been rotated after the fourth year. In this way, no one will be part of the sample for more than four years. On this topic Cochran (1977, Chapter 12) further concludes that "field costs are likely to be lower if the same units are retained for a number of occasions. If estimates of the change in the population total or mean are of interest, this factor also points toward matching more than half the units from one occasion to the next. It is convenient to keep the weights and the proportion matched constant, instead of changing them every occasion. ... increase in the proportion matched from $\frac{1}{2}$ to $\frac{3}{4}$ produces substantial gains in efficiency for the estimates of change at the expense of smaller losses in efficiency for the current estimates. The results suggest that retention of $\frac{1}{2}$, $\frac{3}{4}$, or $\frac{4}{5}$ from one occasion to the next may be a good practical policy if current estimates and estimates of change are both important."

Chapter 8

Other sources of error

In the previous chapters we only discussed sampling error as a potential source of deviation between population values and their sample estimates. Sampling errors arise due to the fact that we only investigate part of the population, and they disappear when the whole population is observed. In this chapter we discuss other sources of error not considered so far.

8.1 Problems with the sampling frame

In this section we follow the recommendations on (the solution of) problems with the sampling frame as discussed in Moors and Muilwijk (1975). The sampling frame is the administrative counterpart of the population of interest. It usually consists of one or more lists or databases containing all elements of the population, but this is not necessary as long as a systematic description is provided of the way in which the elements can be found. Military maps, for example, can be used as a frame for all parcels in a region, and these need not be put on a list. In the case of two-stage sampling (see Chapter 5), a frame for all secondary units is not required: it is enough to have a frame of all primary units together with frames of the elements within each of the *randomly drawn* primary units.

There are three problems that can arise with sampling frames: missing elements, elements that are included more than once, and elements that do not belong to the population of interest. In these cases three general approaches may be considered:

- We can try and correct the complete sampling frame. This approach, however, is both costly and time-consuming.
- We can redefine the population according to the possibilities offered by the frame.
- We can ignore the problems and consider the frame to be ideal. If the effect on the results is small, this approach can hardly be argued against.

Several of these problems arise because the frame always slightly lags behind the actual situation. Even when changes in the population are reported quickly, delays occur in the administrative processing of the mutations.

8.2 Missing elements

Incompleteness of the frame is a frequent problem, that is also often difficult to fix. We can have random missing elements, as a result of errors or mistakes, and a systematical lack of information for certain subpopulations. An example of the latter situations concerns new elements of the population that have not yet been processed. There are two conceivable methods to solve for incompleteness of the frame. The first method is to use two different frames of the same population, if available. By comparing the two frames, random missing elements may be tracked down, and often also systematic missing elements. With this first method we may again distinguish two cases:

- We use frame 1, with the missing elements in 1 supplemented from frame 2. This is only feasible when frame 2 is not too large.
- Frames 1 and 2 are used together. This, however, results in new problems because some elements will be registered in both frames, see also Section 8.4.

In the nineteen seventies in the Netherlands passenger cars were in principle registered twice, for example: once in the database of the "Dienst motorrijtuigenbelasting" (Center of motor vehicle taxes) and once in the license plate database of the Dutch "Rijksdienst Wegverkeer" (Vehicle Technology and Information Center). The first database was easy to use because it had been computerized, but it was also incomplete; at that time the second database was therefore used to supplement the first.

As a second example, the population registers of all Dutch municipalities together are a sampling frame for the Dutch population, except for the sailing and traveling population. Until the ninety nineties these people without fixed address were registered in the "Centraal persoonsregister" (Central population register), and the missing elements in the former frame could therefore be supplemented with the elements in the latter frame.

The second method to control for missing elements is during the survey; this is mainly useful for tracking down small numbers of missing elements. With this method, it is checked for each unit in the sample whether the next unit of the population is registered in the sampling frame. If it is not, the latter unit is added to the sample and also measured (and the next if it is also missing from the frame).

8.3 Foreign elements

It often happens that the sampling frame contains elements that do not belong to the population of interest. These are called *foreign elements*. If it is not possible to remove them before sampling, they have to be eliminated before observation or at least before computation of the population estimates.

One way to handle the problem of foreign elements that was already mentioned in Section 2.10 is to consider the sampling frame to be a hyperpopulation consisting of the sampling population plus the foreign elements. The sampling population is then a subpopulation in the sense discussed in Section 2.10. For simple random sampling the formulas in the latter section can then be applied, and for stratified random sampling those in Section 3.11. However, there we already noticed that when the size of a subpopulation is unknown the precision of sample estimates is strongly reduced, unless the subpopulation is almost as large as the

total population. This procedure also has the effect that the size of the sample from the frame will be smaller (because it will contain foreign elements which have to be removed), but this can be compensated by estimating the reduction in size as well as possible and then increase the original sample size accordingly.

Example 8-1. From a neigbourhood of 400 houses we want to draw a simple random sample of 100 houses from the municipal address register. The latter register, however, also contains the addresses of 100 buildings not destined and used for housing. This register is considered to be a hyperpopulation consisting of the sampling population (the 400 houses) *and* the foreign elements (the 100 buildings). The desired sampling fraction of $\frac{1}{4}$ is applied to the complete frame. This yields a random sample of $n = 125$, of which approximately 25 will be buildings; these are removed from the sample, and the formulas discussed in Section 2.10 are used.

Another way to handle the problem of foreign elements is to replace each foreign element encountered in the sample with the next element in the frame that *does* belong to the population of interest. This means that the sampling probabilities for the population elements are no longer equal, and usually will have to be evaluated during the field work. This is done by establishing for each element in the sample how many foreign elements were preceding it in the frame. The sampling probability for each element is then proportional to this count plus one.

Example 8-2. For the same situation as in Example 8-1, a sampling fraction of $\frac{100}{500} = \frac{1}{5}$ is applied to the complete frame, and elements are drawn randomly without replacement. This yields a random sample of size 100. Now suppose that the first part of the sampling frame has the structure shown in Table 8.1, where h is a house, f is a foreign element, and y_i is some characteristic of house or building i (e.g., its financial value).

Table 8.1: A sampling frame with foreign elements

h	h	h	f	h	f	f	h	h	h	f	h	h	h	h	...	h
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	$\cdots$	y_{500}

Also suppose that the first six elements in the random sample of size 100 happen to be the numbers 2, 12, 9, 3, 7, and 5. Then the first element in the sample, element 2, is a house, see Table 8.1, so it stays in the sample; its probability of being selected is $p_2 = \frac{1}{500}$, since it is not preceded by a foreign element in the frame. The second element in the sample is element 12, which is also a house. It therefore stays in the sample. It is preceded by a foreign element in the frame, and its probability of being selected is therefore $p_{12} = \frac{2}{500}$. The third element in the sample is 9, which is a house not preceded by a foreign element in the frame, so it stays in the sample and has a probability of being selected equal to $p_9 = \frac{1}{500}$. The fourth element in the sample is element 3, which is also a house not preceded by a

foreign element in the frame, so it stays in the sample and also has a probability of $p_3 = \frac{1}{500}$ of being selected. The fifth element in the sample is element 7. It is not a house, so it is removed from the sample and replaced by element 8 in the frame. The latter element has two foreign elements preceding it in the frame, so it has a probability of $p_8 = \frac{3}{500}$ of being selected. The sixth element in the sample is element 5. This is a house so it stays in the sample; it is preceded in the frame by a foreign element, so its probability of being selected is $p_5 = \frac{2}{500}$. An unbiased estimate of the population total is therefore

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = (y_2 + \frac{y_{12}}{2} + y_9 + y_3 + \frac{y_8}{3} + \frac{y_5}{2} + \dots),$$

see also Chapter 4.

8.4 Multiple registration of elements

When the sampling frame contains some elements more than once, the sampling probabilities for those elements are proportional to the number of times they are registered. We may distinguish between these three cases:

- The number of registrations per element is known or can be easily established. In this situation the formulas of Chapter 4 can be applied. The first method is particularly simple when the multiple registrations of one element are recorded together. As the frame of all practitioners of anthropology, for example, the member lists of all societies of anthropology can be used. By incorporating a question in the survey about the number of these societies of which an anthropologist is a member, the sampling probability of each member of the sample can be established.

- If one of the multiple registrations can always be distinguished from the others then the latter registrations can be considered foreign elements, to be treated using the procedures discussed in Section 8.3. When the multiple registrations are registered one after the other in the frame, for example, we may decide that the element is only added to the sample if the first registration of these registrations is selected.

- A large part of the frame consists of single registrations; multiple registrations only occur in another part of the frame. This situation arises when multiple frames are used, for example. The frame with multiple registrations can then be considered as a stratum containing foreign elements. As already discussed in Section 8.2, in the nineteen seventies passenger cars were in principle registered twice in the Netherlands: once in the database of the Center of motor vehicle taxes and once in the license plate database of the Dutch Vehicle Technology and Information Center. The latter database could be used to supplement the former. For each sampled element from the license plate database it is checked whether it is also registered in the vehicle taxes database; if so, it is put aside as a foreign element. For the stratum of the license plate database the formulas of Section 2.10 now apply.

8.5 Non-response

In almost every survey it is found to be impossible to obtain all the required information from each element sampled from the population. Reasons can be absence, illness or the unwillingness of the people in the sample to participate. There are several ways to minimize or handle this phenomenon which is known as *non-response*. Non-response introduces bias in the parameter estimates because non-respondents are usually different on the variable(s) of interest from those of the population who do participate.

A good preparation of the survey often makes it possible to restrict the amount of non-response as much as possible. The influence of absence of respondents can be reduced by choosing a suitable time for interviewing, e.g., not during holidays. The interviewing of heads of households should preferably be organized in the evenings, after an appointment has been made by mail or telephone.

Interviews should be designed such that refusal of cooperation is prevented as much as possible. In the introduction to the survey the interest of the interviewee should be aroused, the purpose of the interview must clearly be explained, a questionnaire sent by mail should come with an introduction letter, an interviewer should identify himself. The questions should be short and clear. As many guarantees of protection of privacy as possible should be provided.

Even though the reduction of non-response can often be successful, one has to take into account that there will always be some sample elements for which no data can be obtained. It is therefore usually necessary to assess the adverse effects of non-response as best as possible, or to correct for its effects.

When the cause of the non-response is unknown or if there is reason to believe that its cause is temporary, the first obvious approach to handle non-response is to repeat the attempt to obtain the required information or to make the respondent cooperate in the survey.

A second option is to draw a random sample from the non-respondents in the original sample, and then to try again to obtain the required information from this subsample. This approach is especially useful when the observation of the latter subsample is more expensive than that of the original sample. In this case, the double sampling for non-response approach of Section 7.3.4 can be applied.

The problem of non-response is clearly not solved by replacing the non-respondents in the sample with population elements that are willing to participate in the survey. At the same time, such a replacement procedure helps to prevent the reduction of the estimated minimal sample size as required to obtain a certain pre-specified precision. It is therefore advisable to always sample more elements than the estimated minimal sample size. These extra reserve elements can then be used to replace the non-respondents. However, although this helps to obtain the required precision, it does not solve the bias due to non-response.

Finally, when everything else fails, for the permanent remaining non-response group we can try to estimate the impact of their non-response on the population parameter estimates. In the Netherlands, for example, the odometer readings of each passenger car is registered in the commercial "National Car Pass" database each time it visits a service station for a roadworthiness check or a car service, as we already mentioned in Section 3.10. From the non-respondents in a sample survey to estimate the total number of kilometers driven by passenger cars in the Netherlands, therefore, a random sample could be selected, after

which their odometer readings could be obtained from this database in order to estimate and correct for the bias in the original sample.

8.6 Measurement errors

A measurement error is the difference between the observed value of an element in the sample or the population and its true value. Measurement errors arise due to the fact that the "device" used for measuring the variable of interest is imperfect. Let x_{ij} denote the observation of element i of the population ($i = 1, \ldots, N$) at time point j, x_i denote the true value of element i, and e_{ij} denote the measurement error of element i at time point j. Then we have for element i that

$$x_{ij} = y_i + e_{ij}. \tag{8.1}$$

It follows from (8.1) that, for a sample of size n,

$$\bar{x} = \bar{y} + \bar{e}. \tag{8.2}$$

The variance of the mean of the n observations in the sample is then

$$\mathrm{V}(\bar{x}) = \mathrm{V}(\bar{y} + \bar{e}) = \mathrm{V}(\bar{y}) + 2\mathrm{Cov}(\bar{y}, \bar{e}) + \mathrm{V}(\bar{e}). \tag{8.3}$$

We may distinguish the following four situations:

- The measurement errors are random, where $E(e_{ij}) = 0$: in this case the measurement device is off in a random fashion. An example is rounding error when measuring a person's length with a ruler or a person's weight with a scale. Since rounding errors are random they cancel each other out, and the resulting parameter estimate in the population is therefore unbiased.

- The measurement errors are systematic, meaning that $E(e_{ij}) \neq 0$: in this case the measurement device is systematically under- or overestimating the actual values of the sample elements. This introduces bias. The variance of the mean is not affected by this bias, but the mean squared error is: $\mathrm{MSE} = \mathrm{V}(\bar{x}) + [E(e_{ij})]^2$.

- The measurement errors are uncorrelated with the true values: in this case $\mathrm{V}(\bar{x}) = \mathrm{V}(\bar{y}) + \mathrm{V}(\bar{e})$, i.e., the parameter estimate is unbiased, but less precise.

- The measurement errors are positively correlated with the true values: in this case the variance of the observed mean will be larger than the variance of the true mean, see (8.3).

Generally, however, this is an area where we can expect naturalistic driving studies to really outperform surveys where the information is gathered through questionnaires, diaries, or interviews.

8.7 Calibration weighting and non-sampling errors

More recently, the use of auxiliary information on the population to reduce sampling error such as applied in the ratio and regression estimator discussed in Chapter 6 has been generalized to also adjust for non-response (Section 8.5), and for multiple registrations (Section 8.4) and/or missing elements (Section 8.2) in the sampling frame. These recent generalizations are known as *calibration weighting*, see for example Kott (2006) and Bethlehem, Cobben, and Schouten (2011).

In a national travel survey, for example, comparing background variables of the participants in the sample of respondents that are related to travel behavior such as age, gender, income and urbanization level with known distributions of these same variables in the total population (if available!) are used to calculate calibration weights. These calibration weights are then used to make the sample more representative of the total population, and thus to correct for selection bias resulting from non-response.

One of the simplest of these weighting techniques is called *poststratification*. Suppose we draw a simple random sample of $n = 100$ elements from a population of $N = 1000$ elements. Also suppose that the frequencies of the categories of the variables gender and age in the population and in the sample are as displayed in Table 8.2.

Table 8.2: Poststratification example

	Population		
	Male	Female	Total
Young	226	209	435
Middle	152	144	296
Elderly	133	136	269
Total	511	489	1000
	Sample		
	Male	Female	Total
Young	23	15	38
Middle	16	17	33
Elderly	13	16	29
Total	52	48	100
	Weights		
	Male	Female	
Young	0.983	1.393	
Middle	0.950	0.847	
Elderly	1.023	0.850	

We see in the example in Table 8.2 that the proportions n_h/n of elements in the six strata made up by the demographic variables age and gender in the sample are not the same as those in the population. The age and gender groups in the sample are therefore not completely representative of the age and gender groups in the population. This may be the result of sampling error, or of non-response, or of a combination of both.

However, if we happen to know the exact number of elements N_h in each of the L strata

in the population, where $h = 1, 2, \ldots, L$, and $N = N_1 + N_2 + \cdots + N_L$ is the total number of elements in the population, then we can "correct" the sample by weighting the sample elements with the following weights:

$$w_h = \frac{N_h/N}{n_h/n}, \tag{8.4}$$

for $h = 1, \ldots, L$. For females in the middle age group, for example, the correction weight equals $w_h = \frac{144/1000}{17/100} = 0.847$, see the bottom of Table 8.2. By weighting the corresponding sample elements with these six weights, the frequencies in the contingency table of the variables age and gender are as shown in Table 8.3.

Table 8.3: Contingency table of weighted age and gender in the sample

	Sample		
	Male	Female	Total
Young	22.6	20.9	43.5
Middle	15.2	14.4	29.6
Elderly	13.3	13.6	26.9
Total	51.1	48.9	100.0

Comparing the frequencies in Table 8.3 with those for the population at the top of Table 8.2, we see that the distributions of the variables age and gender in the sample now match those in the population perfectly.

Letting $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h}$ denote the mean of the variable of interest in stratum h, in post-stratification it can be shown that the correction weights (8.4) lead to the following estimate of the mean of the population

$$\hat{\bar{Y}}_{PS} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h, \tag{8.5}$$

while

$$\hat{Y}_{PS} = \sum_{h=1}^{L} N_h \bar{y}_h, \tag{8.6}$$

yields an estimate of the total of the variable of interest in the population. The correction weights (8.4) at least help to reduce the bias in the estimated population parameters insofar as auxiliary variables like age and gender are capable of capturing the mis-representation in the sample due to non-response.

For more sophisticated calibration weighting approaches we refer to Kott (2006) and Bethlehem, Cobben, and Schouten (2011).

Chapter 9

Conclusions and implications for naturalistic driving study design

In order to decide what sampling and estimation method is most appropriate in any given situation, we first of all have to consider the type of sampling frame(s) that are available for the study at hand.

When this sampling frame contains information on all *individual* population elements, a simple random sample (see Chapter 2) or a systematic sample (see Section 7.1) may be considered. If relevant auxiliary/additional information is available on a qualitative variable whose categories can be expected to have relatively homogeneous variances, precision can be increased by using a stratified random sample, see Chapter 3. The same applies to a classified quantitative auxiliary variable that is highly correlated with the variable of interest.

If the individual values of a quantitative auxiliary variable that is highly correlated with the variable of interest are known for all *sample units*, then precision can be further increased by replacing the direct estimator with a ratio or regression estimator, as discussed in Section 6.1 and Section 6.2. However, this usually requires knowledge of the sum total of the auxiliary variable in the population. Should the individual values of such a quantitative auxiliary variable also be known for all *population units*, then the selection procedure with unequal probabilities discussed in Chapter 4 can be considered as a useful improvement.

When the sampling frame only contains information on *groups* of population elements, then the two-stage sampling methods discussed in Chapter 5 can be used. Once the primary units have been selected randomly with equal probabilities, both simple random sampling and systematic sampling may be applied for the selection of the corresponding second-stage or secondary units. If the sizes of the *randomly selected* primary units are known, or the total of an auxiliary variable, the ratio estimator can be used. If the sizes of all primary units *in the population* are known, the primary units can be selected with probabilities proportional to size.

Even when the sampling frame contains information on all individual population elements, however, both stratified and multi-stage sampling methods can still be used.

In all these cases, given a pre-specified precision and a pre-specified confidence level it is possible to obtain estimates of the minimal required sample size (see Section 2.8 for simple random sampling, Section 3.6 for continuous variables and Section 3.9 for proportions in stratified random sampling, Section 5.1, Section 5.2.3, and Section 5.2.4 for two-stage sampling, Section 6.1.3 and Section 6.2.3 for the ratio and regression estimators, respectively,

and Section 7.3.4 for double-sampling for non-response). The same applies to the situation where sample size needs to be calculated for a fixed budget (see Section 3.4 for stratified random sampling, Section 5.1, Section 5.2.3, and Section 5.2.4 for two-stage sampling, and Section 7.3.4 for double-sampling for non-response).

As an illustration again consider Example 2-6, where we assumed a very large population (so the finite population correction is not required) of car drivers who on average drive $15,000$ kilometers a year. Using a confidence level of 95%, and applying formula (2.65), the minimal required sample sizes obtained in a simple random sampling scheme needed to estimate the total number of vehicle kilometers driven by cars in a year with precision levels of $\pm 10\%$, $\pm 5\%$, and $\pm 1\%$, and population standard deviations of $S = 5,000$, $S = 10,000$, and $S = 15,000$ are given in Table 9.1. As the table indicates, sample size increases both when the required precision of the estimate increases, and when the variance of the variable of interest in the population is larger.

The practical implication of the chosen precision level is that only changes between two consecutive time points or periods larger than twice this precision level will be detected with the corresponding sample. If a precision level of $\pm 5\%$ is chosen for the estimation of the total number of kilometers driven by cars in a country, for example, and the standard deviation of this variable in the population happens to be $S = 10,000$, then changes equal to or smaller than 10% in the total number of kilometers driven will go undetected with a sample size of 683 cars. When a precision level of $\pm 1\%$ is chosen, on the other hand, and the standard deviation in the population is $S = 10,000$, then only changes equal to or smaller than 2% in the total number of kilometers driven will go undetected with a sample size of 17,074 cars.

Table 9.1: Sample sizes required for the estimation of total number of motor vehicle kilometers driven by cars in a country with precision levels of $\pm 10\%$, $\pm 5\%$, and $\pm 1\%$, population standard deviations of $S = 5,000$, $S = 10,000$, and $S = 15,000$, and a confidence level of 95%.

$S = 5,000$			$S = 10,000$			$S = 15,000$		
$\pm 10\%$	$\pm 5\%$	$\pm 1\%$	$\pm 10\%$	$\pm 5\%$	$\pm 1\%$	$\pm 10\%$	$\pm 5\%$	$\pm 1\%$
43	171	4,269	171	683	17,074	385	1,537	38,416

The sample sizes in Table 9.1 are conservative in the sense that they are based on the direct estimator in simple random sampling, which have the largest standard errors and are thus the least precise. Other estimators like the ratio and regression estimators and other sampling techniques like stratified random sampling will usually require smaller sample sizes in order to obtain the same amount of precision. However, the latter approaches also all require more information about the population at hand than when the direct estimator and simple random sampling are used.

All these considerations carry over to naturalistic driving study designs. In naturalistic driving study designs, therefore, the sampling technique of choice will also first of all depend on whether a centralised national sampling frame is available or not. In the Netherlands, for example, it seems obvious that the database containing all Dutch licensed vehicles of the "RDW" (Vehicle Technology and Information Center) is the most appropriate frame from which to sample passenger cars. The database contains all registered motor vehicles

in the Netherlands, including several technical specifications of each vehicle. The latter specifications can be very useful for the stratification of the vehicle population. The Dutch Vehicle Technology and Information Center also has a database containing all driver licences issued in the Netherlands, including background variables of the drivers like age and gender. These demographic characteristics can be used for the stratification of the driver population. In the Dutch situation the available sampling frames imply that the units to be sampled and then observed should be the licensed drivers since they are the ones who give informed consent to participate in the study.

However, if the sampling frame happens to be decentralised and municipal, for example, then a two-stage sampling design would be called for, as discussed in Chapter 5. A nice illustration of the latter approach to survey sampling is presented in Rofique, Humphrey, Pickering, and Tipping (2010). The methodology of this survey covers and combines many of the aspects of sampling that we discussed in the previous chapters, and we therefore discuss it here in some detail. In order to obtain estimates of personal travel of the total population within Great Britain, they used a stratified two-stage random probability sample of private households in Great Britain. The sampling frame is the 'small user' Postcode Address File (PAF), a list of all addresses (delivery points) in the country. The sample was drawn firstly by selecting the Primary Sampling Units (PSUs) in the first stage, and then by selecting addresses within PSUs in the second stage. The sample design employs postcode sectors as PSUs. There were 684 PSUs in 2010. In order to reduce the variance of estimates of year-on-year change, half the PSUs in a given year's sample are retained for the next year's sample and the other half are replaced. Hence 342 of the PSUs selected for the 2009 sample were retained for the 2010 sample, supplemented with 342 new PSUs. The PSUs carried over from the 2009 sample for inclusion in 2010 were excluded from the 2010 sample frame, so they could not appear twice in the sample. The dropped PSUs from 2009 were included in the sample frame.

While the same PSU sectors might appear in different survey years, no single addresses were allowed to be included in three consecutive years. Each year, the National Center for Social Research provided the sampling company with a list of the addresses selected for the previous three survey years. These addresses were excluded from the sampling frame before the addresses for 2010 were selected. This meant respondents to the previous year's survey in the carried over PSUs could not be contacted again.

The list of postcode sectors in Great Britain was stratified using a regional variable, car ownership and population density. This was done in order to increase the precision of the sample and to ensure that the different strata in the population are correctly represented. Random samples of PSUs were then selected within each stratum. Forty regional strata were used, and within each region, postcode sectors were listed in increasing order of the proportion of households with no car (according to the 2001 Census). Cut-off points were then drawn approximately one third and two thirds (in terms of delivery points) down the ordered list, to create three roughly equal-sized bands. Within each of the 120 bands thus created (40 times 3), sectors were listed in order of population density (people per hectare). Then 342 postcode sectors were systematically selected with probability proportional to delivery point count. Differential sampling fractions were used in Inner London, Outer London and the rest of Great Britain in order to oversample London, as response rates tend to be much lower in London compared with the rest of Great Britain. These sectors were then added to the 342 sectors carried over from the previous years survey to make the final

sample of 684 sectors for each year.

Next, within each selected sector, 22 addresses (the secondary or second-stage house-hold units) were sampled systematically, giving a sample of 15,048 addresses (684 postcodes times 22). More details of the sampling methodology used in this survey can be found in Rofique, Humphrey, Pickering, and Tipping (2010).

Besides the just mentioned considerations concerning the types of sampling frame available, based on the material presented in this report we end with the following specific conclusions and recommendations for the selection of a probabilistic sample of passenger cars in a naturalistic driving study design.

1. All sample size estimation methods have in common that they require an a priori specified degree of precision with an a priori specified confidence interval; this therefore applies to sampling methods for naturalistic driving studies as well. This degree of precision simply specifies how close we want the sample estimate (of the mean, the total, or a proportion) to be to its actual population value; this can be expressed in absolute terms (i.e., I want the sample estimate of the total number of kilometers traveled to deviate no more than 10 million from the actual total number of kilometers traveled) or in relative terms (i.e., I want the sample estimate of the total number of kilometers traveled to deviate no more than 1% of the actual total number of kilometers traveled). For sample size estimation we also have to specify how certain we want to be of obtaining the desired degree of precision with a sample.

2. All sample size estimation methods have in common that they require some knowledge of, or an estimate of, the population variance(s) of the variable(s) of interest in simple random sampling, of the population variances in the different strata in stratified random sampling, and of the variances of the primary and secondary units in two-stage sampling. When sample size is estimated for proportions or percentages the situation is easier because a conservative estimate can always be obtained by assuming the population proportion to be equal to 0.5.

3. When the objective is to measure *changes* in the population over time, as is the case in naturalistic driving studies, the required precision should be established by considering the minimal difference in parameter estimates between consecutive time points that we want to detect with certainty, see also Chapter 1.

4. When information on auxiliary variables in the population is available that are highly correlated with the variable of interest this opens up the possibility of improving the precision of the parameter estimates obtained with simple random sampling (Chapter 2) by using stratified random sampling, see Chapter 3.

5. When several items in the population need to be estimated, then this requires sample size estimations for each of these items separately. If costs are not an issue, the largest estimated sample size should be used in order to guarantee the required precision for all items. In naturalistic driving studies where several RED and SPI items are estimated, e.g., passenger car kilometers traveled, speed, and seat belt use, sample size estimations should be made for each of these items also, and the largest estimated sample size should be used in order to guarantee the required precision for all RED

and SPI items. If the budget is fixed, it is also still possible to determine the optimal sample size in stratified random sampling and two-stage sampling.

6. Since national naturalistic driving studies are expected to extend over a number of years, the best sampling strategy for measuring change is to use a rotating sample where one half, three-quarter, or even four-fifth of the sample is retained and the remaining part of the sample is replaced after some fixed period of time.

7. The length of this fixed period of time should also take into consideration the costs and time required for the installation and de-installation in each sampled car of the chosen recording device(s).

8. In order to control for seasonal fluctuations (e.g., due to holidays) it seems that the ideal consecutive period to observe the sample units with the recording device would be one year. This could be combined with the just mentioned rotating sampling procedure, as follows. All cars in the selected sample are equipped with the recording device on time point 1, say. Half of this sample is replaced after half a year, and the replacements are then observed during one year. The other half of the sample is observed the whole first year, and then replaced with a new sample, et cetera. In this way none of the sampled cars are in the sample for more than one year, while still being rotated on a fifty percent basis.

9. The continuous nature of the measurements obtained in a naturalistic driving study implies that the ratio and/or regression estimators discussed in Chapter 7 are natural and well-suited candidates for statistically improving the precision of the population parameter estimates: the sample observations obtained for the previous time point or time period can be used to statistically increase the precision of the sample estimates in the next time point or time period. However, these estimators do require knowledge of (or estimates of) the population total or mean.

10. When estimates for sub-populations of the total passenger car population in a country are required, it is recommended to use these sub-populations as strata in a stratified random sampling design because this yields more precise estimates than when the sub-populations cut through the strata.

11. The estimation of the required sample size for a pre-specified precision should always take the problem of non-response into account, and the estimated sample size should be increased accordingly.

12. In some countries at least, it should be possible to get information on the characteristics of the non-respondents by using the double sampling for non-response approach presented in Section 7.3.4. This can be applied in two ways: either by obtaining a random sub-sample of the non-respondents and then make sure that they participate in the study after all, or by obtaining a random sub-sample of the non-respondents and then consulting a second frame also containing (estimates of) the required information.

13. Whenever possible selection bias as a result of non-response should be corrected for by poststratification based on 1) demographic information of the driver population; 2) technical characteristics of the passenger car population; and 3) odometer readings of

passenger cars as registered during roadworthiness checks. If available this last source of information is to be preferred since it is the best indicator of the actual distance traveled by passenger cars in a country.

14. Should it not be possible to install the chosen recording device in all the sampled passenger cars due to technical restrictions, then these cars should be treated the same as non-response.

Bibliography

Bethlehem, J., F. Cobben, and B. Schouten (2011). *Handbook of Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques* (Third ed.). New York: John Wiley & Sons.

Dalenius, T. and J. L. Hodges (1957). Minimum variance stratification. *Journal of the American Statistical Association 54*, 88–101.

Hays, W. L. (1970). *Statistics*. London: Holt, Rinehart and Winston.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology 2*, 133–142.

Molnar, H. R. A., G. M. E. H. Moritz, P. S. G. M. Smeets, B. A. H. Buelens, and A. Dohmen (2009). Estimating vehicle kilometres for goods vehicles and passenger cars with ododmeter readings. Technical report, Statistics Netherlands, Heerlen, The Netherlands.

Moors, J. J. A. and J. Muilwijk (1975). *Steekproeven. Een inleiding tot de praktijk (Sampling. A practical introduction)*. Amsterdam: Agon Elsevier.

Rofique, J., A. Humphrey, K. Pickering, and S. Tipping (2010). National travel survey 2010. Technical report, London, United Kingdom.