Sampling theory

Sampling design and estimation methods

12

Reinder Banning, Astrea Camstra and Paul Knottnerus

Statistics Methods (201207)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/'12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/'10– 2011/'12	crop year, financial year, etc. 2009/'10 to 2011/'12 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

60165201207 X-37

Contents

1. Introduction to the subthemes

1.1 General description

This document describes various sampling designs and estimation methods used at Statistics Netherlands. The main focus is on random samples, in which units are selected from the population according to a well defined selection mechanism.

The most straightforward and familiar procedure is simple random sampling without replacement (SRSWOR), in which each possible sample (of equal size) from the population has exactly the same chance of selection. Simple random sampling is the basic selection method, and all other random sampling techniques can be viewed as an extension or adaptation of this method. These other, more complex, designs generally aim to increase the efficiency, i.e. to improve the estimates, but practical and economic considerations can also play a part in choosing a design. However, in all types of sampling, each element of the target population must have a positive and known probability of being included in the sample, which is referred to as the inclusion probability.

Having chosen a sampling design, it must be determined how to estimate the population parameters of interest, such as a population total or a poplation mean. The choice of a specific estimating method is largely determined by the presence of auxiliary information and the nature of the relationship between the target variable and the auxiliary information. Accordingly, a quotient estimator, regression estimator or poststratification estimator may be selected.

The ‘Sampling design’ subtheme is covered in Chapters 2, 3, 4 and 5. The ‘Estimation methods’ subtheme is covered in Chapters 6, 7 and 8. Furthermore, equations and formulas are numbered to facilitate referencing from elsewhere in the text. Equation numbers are marked with an asterisk (*) to indicate that a proof is provided in the appendix to the chapter concerned.

1.2 Scope and relationship with other themes

There is clearly a relationship between *estimation* and *weighting to adjust for nonresponse.* The methods and techniques discussed under that subject build upon the elementary estimation methods discussed in this document, but are oriented more to the problem of nonresponse.

There is also a relationship with the *Panels* theme, in which population elements are incorporated into the sample in the course of several consecutive observation periods. This approach is often used when the interest is in trends over time. There are also clear relationships with the theme *Experiments* and the theme *Model-based Estimation* as, for example, applied to small subpopulations.

1.3 Place in the statistical process

It will be clear from the above that the sampling design is relevant at the start of the statistical process, whereas estimation occurs more towards the end, shortly before the figures are published.

2. Simple random sampling without replacement

2.1 Short description

Simple random sampling without replacement (SRSWOR) is the most familiar sampling design. This kind of sampling is referred to as simple because it involves drawing from the entire population. Alternatively, simple random sampling can be carried out *with* replacement (SRSWR), but this type of design is not used at Statistics Netherlands. However, since the relatively simple SRSWR variance formulas can be used under certain conditions in SRSWOR, SRSWR and the associated variance formulas will be briefly discussed in Section 2.3.3.

2.1.1 Definition of SRSWOR

The rather formal definition of SRSWOR is as follows. Consider a population U of N elements, or $U = \{1,2,\cdots,N\}$. SRSWOR is a method of selecting n elements out of U such that all possible subsets of U of size n have the same probability of being drawn as a sample. Note that there are $\binom{N}{n}$ possible subsets of U of size n.

2.1.2 Sampling scheme for SRSWOR

In practice SRSWOR can involve successively selecting random numbers between 1 and N, and including each associated population element in the sample until n elements have been selected. If a number is already drawn, a new number is selected at random. A good alternative to SRSWOR is *systematic* sampling, which is defined in Section 4.3.3.

2.2 Applicability

Statistics Netherlands applies SRSWOR mainly for business statistics. The population of businesses is usually stratified and SRSWOR is performed for each stratum – see the next chapter for stratified sampling. The SRSWOR design is also frequently used as a sort of benchmark to evaluate other sampling designs by comparing the variances of the different design based estimators.

2.3 Detailed description

2.3.1 Assumptions and notation

We start with a short description of a number of important population parameters. Given a specific *target variable* we distinguish the *population total* Y, the *population mean* $\bar{Y}$, the *population variance* σ_y^2, the *adjusted population variance* S_y^2, and the *population coefficient of variation* CV_y. The adjusted population variance is often used to simplify the SRSWOR formulas. Moreover, Y_k stands for

the value of the target variable of element k, $k = 1, \cdots, N$. The stated parameters are defined as follows:

$$
\begin{aligned}
Y &= Y_1 + \cdots + Y_N = \sum_{k=1}^{N} Y_k \\
\bar{Y} &= \frac{1}{N} Y = \frac{1}{N} \sum_{k=1}^{N} Y_k \\
\sigma_y^2 &= \frac{1}{N} \sum_{k=1}^{N} (Y_k - \bar{Y})^2 \\
S_y^2 &= \frac{N}{N-1} \sigma_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \bar{Y})^2 \\
CV_y &= \frac{S_y}{\bar{Y}} \quad .
\end{aligned}
\tag{2.1}
$$

The *n* (i.e. the sample size) observations of the target variable in the sample are denoted in small letters $y_1, \cdots, y_n$. The *sample mean* $\bar{y}_s$, the *sample variance* s_y^2 and the *sample coefficient of variation* cv_y are the three most important *sample parameters*. These three sample parameters are defined as follows:

$$
\begin{aligned}
\bar{y}_s &= \frac{1}{n} \sum_{k=1}^{n} y_k \\
s_y^2 &= \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y}_s)^2 \\
cv_y &= \frac{s_y}{\bar{y}_s} \quad .
\end{aligned}
\tag{2.2}
$$

2.3.2 Estimators for the population mean, population total and adjusted population variance

For three of the five population parameters defined in (2.1) explicit estimators are formulated below. These parameters are the population total, the population mean and the adjusted population variance. Intuitively, the sample mean $\bar{y}_s$ seems a reasonable estimator of the population mean. It is denoted as follows:

$$
\hat{\bar{Y}} = \bar{y}_s = \frac{1}{n} \sum_{k=1}^{n} y_k \tag{2.3}
$$

The hat ^ on the left-hand side indicates an estimator of the corresponding parameter.

The estimator of the population mean leads to the following proposal for the estimator of the population total:

$$
\hat{Y} = N \hat{\bar{Y}} = N \bar{y}_s = \sum_{k=1}^{n} \left(\frac{N}{n} \right) y_k \tag{2.4}
$$

The parameter N/n in this expression is referred to as the *inclusion weight* of the sample because it gives the ratio of population size to sample size.

Finally, we present the sample variance as the estimator of the adjusted population variance, i.e.:

$$\hat{S}_y^2 = s_y^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}_s)^2 \quad . \tag{2.5}$$

The unbiasedness of the estimators

An estimator is unbiased if its expected value is equal to the statistic to be estimated. In order to determine the expectations of the *(direct) estimators* for the population mean and population total in SRSWOR, it is convenient to write the estimators in an alternative form. The estimator for the population mean is rewritten as:

$$\hat{\bar{Y}} = \bar{y}_s = \frac{1}{n}\sum_{k=1}^{N} a_k Y_k \tag{2.6}$$

where the binary random variable a_k is defined as:

$$a_k = \begin{cases} 1 & \text{if element } k \text{ is included in the sample} \\ 0 & \text{if element } k \text{ is } \textit{not} \text{ included in the sample} \end{cases} \tag{2.7}$$

The binary random variable a_k is also known as the *selection indicator*. The selection indicator has the following two favourable properties:

$$E(a_k) = P(a_k = 1) = \frac{n}{N} \qquad k = 1, \cdots, N \tag{2.8}$$

and:

$$E(a_k a_l) = \begin{cases} \dfrac{n}{N} & k = l \quad \wedge \quad k = 1, \cdots, N \\ \dfrac{n(n-1)}{N(N-1)} & 1 \le k \ne l \le N \quad . \end{cases} \tag{2.9}$$

Take the alternative expression (2.6) for the estimator of the population mean as starting point. Then it can be demonstrated from properties (2.8) and (2.9) of selection indicator a_k that this estimator is an *unbiased* estimator of $\bar{Y}$:

$$E\left(\hat{\bar{Y}}\right) = E\left(\frac{1}{n}\sum_{k=1}^{N} a_k Y_k\right) = \frac{1}{n}\sum_{k=1}^{N} E(a_k) Y_k$$
$$= \frac{1}{n}\sum_{k=1}^{N} \frac{n}{N} Y_k = \frac{1}{N}\sum_{k=1}^{N} Y_k = \bar{Y} \quad .$$

With this result it is also simple to verify that the estimator of the population total is unbiased, i.e.:

$$E\left(\hat{Y}\right) = E\left(N\,\hat{\bar{Y}}\right) = N\,E\left(\hat{\bar{Y}}\right) = N\,\bar{Y} = Y \quad .$$

The selection indicators are also useful in proving that s_y^2 in (2.5) is an unbiased estimator of the adjusted population variance. In other words:

$$E\left(s_y^2\right) = E\left(\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \bar{y}_s\right)^2\right) = S_y^2 \quad . \tag{2.10*}$$

The asterisk (*) alongside the number of the above equation indicates that a proof is given in the appendix to this chapter.

Variance calculations

It can be proved that the variance of the estimator of population mean is a function of the adjusted population variance:

$$\text{var}\left(\hat{\bar{Y}}\right) = \frac{1}{n}(1-f)S_y^2 \;= \frac{1}{N}\left(\frac{1-f}{f}\right)S_y^2 \qquad \left(f = \frac{n}{N}\right) \quad . \tag{2.11*}$$

The factor $(1-f)$ is referred to as the *finite population correction*, while the factor f is referred to as the *sampling fraction*. Based on this result, the variance of the estimator of the population total is:

$$\text{var}\left(\hat{Y}\right) = N^2\,\text{var}\left(\hat{\bar{Y}}\right) = N\left(\frac{1-f}{f}\right)S_y^2 \quad . \tag{2.12}$$

The value of parameter f is plotted against the value of parameter $(1-f)/f$ in Figure 2.1. If with (fixed) population size N, the sample size n increases in value, the sampling fraction f becomes 1. Reading from the graph, parameter $(1-f)/f$ then approaches 0. In other words, the variance of the estimator of population total approaches 0 as the sample becomes larger – see expression (2.12).

The variance of the estimator of the population mean, as calculated in (2.11), can be estimated using the following estimator:

$$\hat{\text{var}}\left(\hat{\bar{Y}}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)\hat{S}_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)s_y^2 \tag{2.13}$$

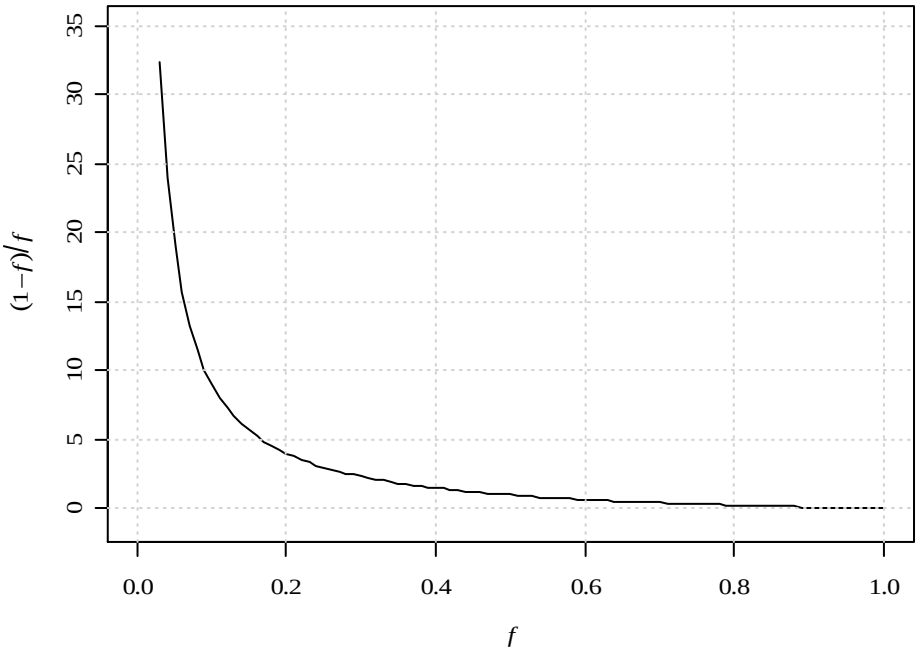
where s_y^2 represents the sample variance. The unbiasedness of this *variance estimator* follows from that of the sample variance s_y^2 as an estimator of adjusted population variance S_y^2.

An unbiased estimate of the variance of estimator $\hat{Y}$, as calculated in (2.12) can be calculated similarly using the estimator

$$\hat{\text{var}}\left(\hat{Y}\right) = N\left(\frac{1-f}{f}\right)s_y^2 \tag{2.14}$$

with s_y^2 the sample variance. The unbiasedness of this variance estimator is a direct consequence of the (proven) unbiasedness of s_y^2 as an estimator of S_y^2.

Figure 2.1. Parameter f plotted against parameter $(1-f)/f$



Coefficients of variation

Besides calculating (and estimating) an estimator's variance, it is also possible to determine the associated coefficient of variation. The coefficients of variation of the estimator of the population mean, denoted as $CV\left(\hat{\bar{Y}}\right)$, and the estimator for the population total, denoted as $CV\left(\hat{Y}\right)$, are defined as follows:

$$CV\left(\hat{\bar{Y}}\right)=\frac{\sqrt{\text{var}\left(\hat{\bar{Y}}\right)}}{\bar{Y}}$$
$$CV(\hat{Y})=\frac{\sqrt{\text{var}\left(\hat{Y}\right)}}{Y} \quad . \tag{2.15}$$

Formula (2.11) gives an expression for the variance of the estimator of the population mean under SRSWOR. This formula can be used to derive the equation below, from which the coefficient of variation of the estimator can be calculated

$$CV^2\left(\hat{\bar{Y}}\right)=\left(\frac{1}{N}\right)\left(\frac{1-f}{f}\right)CV_y^2 \quad . \tag{2.16}$$

In other words, there is a direct relationship between the coefficient of variation of the estimator of the population mean and the population coefficient of variation.

Given the relationship between the population total and the population mean, and the relationship between the estimator of the population mean and the estimator of the population total, it is possible to demonstrate that the coefficients of variation of the two estimators are related as follows:

$$CV\left(\hat{Y}\right) = CV\left(\hat{\bar{Y}}\right) \tag{2.17}$$

This equation expresses that the direct estimators of the population total and the population mean are equally accurate.

Unfortunately it is impossible to actually calculate the two coefficients of variation in (2.15), and consequently they have to be estimated. The obvious estimator for the coefficient of variation of the estimator of the population mean is:

$$\hat{CV}\left(\hat{\bar{Y}}\right) = \left(\frac{s_y}{\bar{y}_s}\right)\sqrt{\frac{1}{N}\left(\frac{1-f}{f}\right)} \tag{2.18}$$

because:

$$\hat{CV}_y = cv_y = \frac{s_y}{\bar{y}_s} \quad .$$

It is important to note that the sampling fraction is often very small in practice, i.e. $f \approx 0$. In fact, this is then equivalent to random sampling with replacement. It is also intuitively clear that replacement or non replacement no longer has an effect on the value of the estimator when the population size is much greater than the sample size. The next subsection discusses simple random sampling with replacement.

2.3.3 Simple random sampling with replacement

Suppose that we perform simple random sampling *with* replacement with equal drawing probabilities $1/N$ (SRSWR). Then $y_1, \cdots, y_n$ can be viewed as n *independent* selections from $Y_1, \cdots, Y_N$, with the following expectation and variance:

$$E\left(y_k\right) = \frac{1}{N}\sum_{l=1}^{N} Y_l = \bar{Y} \qquad\qquad k = 1, \cdots, n$$

and:

$$\text{var}\left(y_k\right) = \frac{1}{N}\sum_{l=1}^{N}\left(Y_l - \bar{Y}\right)^2 = \sigma_y^2 \qquad\qquad k = 1, \cdots, n \quad . \tag{2.19}$$

Note that these two formulas also apply to SRSWOR. In the SRSWR case, these two formulas directly lead to the expectation and the variance, respectively, of the estimator $\hat{\bar{Y}} = \bar{y}_s$:

$$E(\bar{y}_s) = \frac{1}{n}\sum_{k=1}^{n} E(y_k) = \bar{Y}$$

and:

$$\text{var}(\bar{y}_s) = \frac{1}{n}\sigma_y^2 \quad . \tag{2.20*}$$

Comparing the variance formulas in this subsection with those in the previous subsection confirms the property mentioned above that the variance formulas are approximately the same for small *f*.

Finally, note that in SRSWR $E(s_y^2) = \sigma_y^2$. The proof is largely the same as for SRSWOR, and is therefore omitted.

2.3.4 Determination of sample size

A frequently asked question in practice is how large a sample has to be to make inferences, with a certain accuracy, about Y, the population total. A commonly used criterion for accuracy is the 95% confidence interval $I_{95}(Y)$. The definition of $I_{95}(Y)$ is that there is a 95% probability of the unknown population total falling within $I_{95}(Y)$. Assuming that the statistic $\hat{Y}$ is approximately normally distributed, an approximate value of the 95% confidence interval for *Y* for sufficiently large *n* is:

$$I_{95}(Y) = \left(\hat{Y} - 1.96\sqrt{\hat{\text{var}}(\hat{Y})}, \hat{Y} + 1.96\sqrt{\hat{\text{var}}(\hat{Y})}\right) \quad .$$

Using percentage margins of uncertainty, this expression can also be formulated as:

$$\begin{aligned} I_{95}(Y) &= \hat{Y} \pm 1.96\frac{\sqrt{\hat{\text{var}}(\hat{Y})}}{\hat{Y}}100\% \\ &= \hat{Y} \pm \hat{CV}(\hat{Y}) \times 196\% \quad . \end{aligned} \tag{2.21}$$

This is explained below with reference to a simple example. Assume $\hat{Y} = 120$ and $\hat{\text{var}}(\hat{Y}) = 25$. We then obtain $I_{95}(Y) = (120 - 9.8; 120 + 9.8) = (110.2; 129.8)$. In accordance with (2.21) we can also express this interval as $120 \pm 8.2\%$.

The above equations also allow determination of the sample size that would be needed in order to remain within a given margin of uncertainty, say $r\%$. From

$$196 \times CV_y\sqrt{\frac{1}{N}\left(\frac{1-f}{f}\right)} < r$$

it follows that:

$$\frac{196^2 \times CV_y^2}{N \times r^2 + 196^2 \times CV_y^2} < f \quad .$$

If a 5% margin of uncertainty is considered acceptable, the sampling fraction *f* for a population with $N = 1{,}000$ and a $CV_y = 0.7$ must satisfy the following inequality:

$$\frac{196^2 \times 0.49}{1{,}000 \times 25 + 196^2 \times 0.49} \approx 0.43 = f_{\min} < f \quad .$$

For the given population size $N = 1{,}000$, this minimum sampling fraction corresponds to a minimum sample size of $n_{\min} = 430$.

Needless to say, in practice CV_y must first be estimated in some way. The estimate can often be based on recent past data. Sometimes it is also possible to use comparable data with the same volatility.

2.4 Example: estimating fractions

If we wish to calculate the percentage of people with a certain attribute, e.g. the number of people with a high salary, Y_k will assume the following values:

$$Y_k = \begin{cases} 1 & \text{if person } k \text{ has a high salary} \\ 0 & \text{if person } k \text{ does not have a high salary} \end{cases} \qquad k = 1, \cdots, N \quad .$$

If $P = \overline{Y}$ is the fraction of high-salary earners in the population, then (N.B. in this situation $Y_k^2 = Y_k$):

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k = P$$

$$\sigma_y^2 = \frac{1}{N}\sum_{k=1}^{N}(Y_k - P)^2 = \frac{1}{N}\sum_{k=1}^{N} Y_k^2 - P^2 = P - P^2 = PQ \qquad (Q = 1 - P)$$

$$S_y^2 = \frac{N}{N-1} PQ \quad .$$

The sampling fraction of high-salary earners is $p = \hat{P}$, in accordance with formulas (2.3) and (2.11):

$$p = \overline{y}_s$$

$$\text{var}(p) = \frac{1}{N}\left(\frac{1-f}{f}\right) S_y^2 = \left(\frac{1}{N-1}\right)\left(\frac{1-f}{f}\right) PQ \quad .$$

The latter result, the calculated variance of the estimator, can be estimated with (2.13):

$$\hat{\text{var}}(p) = \frac{1}{N}\left(\frac{1-f}{f}\right) s_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)\left(\frac{1}{n-1}\right)\sum_{k=1}^{n}(y_k - p)^2$$

$$= (1-f)\left(\frac{1}{n-1}\right) p(1-p) \quad .$$

The coefficient of variation of estimator p can therefore be estimated as follows – see (2.15) and also (2.18):

$$\hat{CV}(p) = \sqrt{(1-f)\left(\frac{1}{n-1}\right)\left(\frac{1-p}{p}\right)} \quad .$$

It follows from formula (2.21) for the 95% confidence interval *p* that for sufficiently large n, $I_{95}(p)$ can be estimated as:

$$I_{95}(P) = p \pm 196\% \times \sqrt{(1-f)\left(\frac{1}{n-1}\right)\left(\frac{1-p}{p}\right)} \quad .$$

The above formula states that the relative margin of uncertainty increases strongly with decreasing *p*. For example, if on the basis of SRSWOR parameter P is estimated at 0.001 with $n = 50{,}000$, then assuming that $f \approx 0$, the relative margin of uncertainty is 28%. This illustrates the difficulty of estimating what are known as small domains with sufficient accuracy.

2.5 Quality indicators

The quality indicators for SRSWOR are:

- the margins of uncertainty of the corresponding estimators;
- the size of nonresponse.

Nonresponse can severely affect the quality of the results if (i) the nonresponse is large and (ii) the nonresponse is selective. The bias from selective nonresponse can sometimes be corrected by using auxiliary variables that correspond with both the probability of response and the target variable.

Another important assumption in random sampling is that the *frame* from which the sample is drawn corresponds closely with the *target population* about which inferences are to be made.

2.6 Appendix

Proof of (2.10)

Before deriving a formula for the variance of estimator (2.3), it is pointed out that the variance of the estimator for the population mean does not change if we add a constant to all Y_k in the population. In other words, it may be assumed without loss of generality of the results that the population mean is zero, i.e. $\overline{Y} = 0$. Based on this assumption, we rewrite the expression for sample variance using the selection indicator as:

$$s_y^2 = \frac{1}{n-1}\sum_{k=1}^{n} y_k^2 - \frac{n}{n-1}\overline{y}_s^2 = \frac{1}{n-1}\sum_{k=1}^{N} a_k Y_k^2 - \frac{n}{n-1}\overline{y}_s^2 \quad .$$

The expectation s_y^2 can be calculated from definitions (2.2) and (2.6) as (N.B. $E(\overline{y}_s^2) = \mathrm{var}(\overline{y}_s)$ because $\overline{Y} = 0$):

$$
\begin{aligned}
E\left(s_y^2\right) &= \frac{1}{n-1}\sum_{k=1}^{N} E\left(a_k\right)Y_k^2 - \frac{n}{n-1}E\left(\bar{y}_s^2\right) \\
&= \frac{1}{n-1}\times\frac{n}{N}\sum_{k=1}^{N} Y_k^2 - \frac{n}{n-1}\operatorname{var}\left(\bar{y}_s\right) \\
&= \left(\frac{n}{n-1}\right)\sigma_y^2 - \frac{n}{n-1}\operatorname{var}\left(\hat{\bar{Y}}\right) \quad .
\end{aligned}
$$

In other words, the expectation of the sample variance is a function of the variance of the estimator of the population mean. Calculating the expectation of the sample variance therefore requires knowledge of the variance of the estimator of the population mean. In (2.11) a formula for the variance of the estimator of the population mean is derived.

Using the above result and expression (2.11) gives:

$$
\begin{aligned}
E\left(s_y^2\right) &= \frac{n}{n-1}\sigma_y^2 - \frac{n}{n-1}\times\frac{1}{n}\left(1-f\right)S_y^2 \\
&= \frac{n\left(N-1\right)}{N\left(n-1\right)}S_y^2 - \frac{N-n}{N\left(n-1\right)}S_y^2 \\
&= S_y^2 \quad .
\end{aligned}
$$

Therefore s_y^2 is an unbiased estimator of S_y^2. QED.

Proof of (2.11)

We make the same simplifying assumption in proving formula (2.11) for the variance of the sample mean as before, i.e. that $\bar{Y} = Y = 0$. It follows from (2.7) – (2.9) that:

$$
\begin{aligned}
\operatorname{var}\left(\hat{\bar{Y}}\right) = E\left(\hat{\bar{Y}}\right)^2 &= E\left(\frac{1}{n}\sum_{k=1}^{N} a_k Y_k\right)^2 \\
&= \frac{1}{n^2}\left\{\sum_{k=1}^{N} E\left(a_k^2\right)Y_k^2 + \sum_{k=1}^{N}\sum_{\substack{l=1 \\ l\neq k}}^{N} E\left(a_k a_l\right)Y_k Y_l\right\} \\
&= \frac{1}{n^2}\left\{\frac{n}{N}\sum_{k=1}^{N} Y_k{}^2 + \frac{n(n-1)}{N(N-1)}\sum_{k=1}^{N} Y_k\left(0-Y_k\right)\right\} \\
&= \frac{n}{n^2}\left\{\sigma_y^2 - \frac{n-1}{N-1}\sigma_y^2\right\} \\
&= \frac{1}{n}\frac{N-n}{N-1}\sigma_y^2 \\
&= \frac{1}{N}\times\frac{N}{n}\times\frac{N-n}{N-1}\times\frac{N-1}{N}S_y^2 \\
&= \frac{1}{N}\left(\frac{1-f}{f}\right)S_y^2 \quad .
\end{aligned}
$$

Expression (2.11) for the variance of the sample mean is therefore correct. QED.

Proof of (2.20)

In SRSWR, y_k and y_l $(k \neq l)$ are independent selections, and therefore uncorrelated (this is the main difference with SRSWOR, in which y_k and y_l are negatively correlated with a correlation coefficient of $-1/(N-1)$). The variance of the sample mean can therefore be written as:

$$\text{var}(\bar{y}_s) = \text{var}\left(\frac{1}{n}\sum_{k=1}^{n} y_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} \text{var}(y_k) \quad .$$

It follows from (2.19) for the variance of selection k that:

$$\text{var}(\bar{y}_s) = \frac{1}{n^2}\sum_{k=1}^{n} \sigma_y^2 = \frac{1}{n^2} n\sigma_y^2 = \frac{1}{n}\sigma_y^2 \quad .$$

QED.

3. Stratified samples

3.1 Short description

The definition of a stratified random sample assumes division of the target population into what are known as *strata* (the singular form is *stratum*). The strata must be nonoverlapping and together they must cover the whole population. A random sample is then selected from every stratum by SRSWOR. The different SRSWOR samples are mutually independent. In other words, rather than having one large SRSWOR selection, SRSWOR is performed in several smaller steps.

3.2 Applicability

Statistics Netherlands often uses regional, demographic, or socioeconomic factors for stratification. *Stratification*, i.e. dividing the target population into strata, requires the necessary auxiliary information to be available in the sampling frame. For example, to stratify according to branch of industry, this information must be known for each company in the sampling frame. The General Business Register contains a company's standard industrial classification and size category, which are auxiliary variables that Statistics Netherlands often uses for stratification. The information for every person in the Municipal Personal Records Database (GBA) includes much personal data, such as gender, date of birth, marital status and address, which can be used to categorize the Dutch population.

Stratified random sampling can be used for a variety of reasons. First, stratification is a common way of improving the precision of estimators (i.e. to reduce the variance), in particular when estimating characteristics of the entire population. Some variables, e.g. company revenue, can have such a large population variance that very large samples are needed to make reliable inferences. If it is possible to form groups within which the target variable varies little, stratified sampling can lead to more precise outcomes than simple random sampling (with equal sample size). Precision improves because the variance within the strata is less than the variance for the population as a whole.

Second, the interest is often not only in the population as a whole, but also in specific subpopulations or in making comparisons between subpopulations. In simple random sampling, it is a matter of chance how many elements end up in the strata. Small subpopulations in particular will then be poorly represented in the sample. Stratification is a way of ensuring that all subpopulations of interest are sufficiently represented in the sample to allow reliable statements to be made.

Third, it is possible in stratification to use different data collection techniques for different strata. For instance, it may be desirable in a business survey to approach small companies by means of a brief paper questionnaire and to have large companies take part in an extensive telephone or personal interview. The selection and estimating methods may also differ for each stratum.

Fourth, for administrative reasons, sampling frames are often already divided into 'natural' parts, which may even be kept at geographically different locations. In this case separate sampling may be more economical.

3.3 Detailed description

3.3.1 Assumptions and notation

The population is divided into H *strata*. An individual *stratum* is denoted with the index $h, h = 1, \cdots, H$, and comprises N_h elements. The strata must not overlap. In other words, each element must belong to exactly one stratum. The strata jointly form the population, so that

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

where again N is the total population size. We also assume that the size N_h of every stratum is known. The value of the target variable of element k in stratum h is denoted with Y_{hk}, for $h = 1, \cdots, H$, and $k = 1, \cdots, N_h$. The sample size of stratum h is expressed as n_h; by definition $\sum_h n_h = n$. The notation for the sample observations in stratum h is y_{hk}, with $h = 1, \cdots, H$, and $k = 1, \cdots, n_h$.

For a target variable we distinguish the following *stratum parameters*: the *stratum total* Y_h, the *stratum mean* $\bar{Y}_h$, the *stratum variance* σ_{yh}^2, the *adjusted stratum variance* S_{yh}^2, and the *coefficient of stratum variation* CV_{yh}. These parameters are defined as follows:

$$\begin{aligned} Y_h &= \sum_{k=1}^{N_h} Y_{hk} \\ \bar{Y}_h &= \frac{1}{N_h} Y_h \\ \sigma_{yh}^2 &= \frac{1}{N_h} \sum_{k=1}^{N_h} \left(Y_{hk} - \bar{Y}_h\right)^2 \\ S_{yh}^2 &= \frac{1}{N_h - 1} \sum_{k=1}^{N_h} \left(Y_{hk} - \bar{Y}_h\right)^2 \\ CV_{yh} &= \frac{S_{yh}}{\bar{Y}_h} \quad . \end{aligned} \tag{3.1}$$

N.B. the above parameters refer to population parameters in stratum h.

Because a stratified sampling design involves taking a sample in each stratum, the sample parameters are discussed per stratum. The sample parameters in a stratum are: the *sample mean* $\bar{y}_h$ in stratum h, the *sample variance* s_{yh}^2 in stratum h, and the *sample coefficient of variation* cv_{yh} in stratum h. They are defined as follows:

$$\bar{y}_h = \frac{1}{n_h}\sum_{k=1}^{n_h} y_{hk}$$

$$s_{yh}^2 = \frac{1}{n_h - 1}\sum_{k=1}^{n_h}\left(y_{hk} - \bar{y}_h\right)^2 \qquad (3.2)$$

$$cv_{yh} = \frac{s_{yh}}{\bar{y}_h} \quad .$$

3.3.2 Relationships between population parameters and stratum parameters

Each element in the population belongs to exactly one stratum. This special property of strata makes it possible to derive relationships between population parameters (2.1) and stratum parameters (3.1). For instance, the population total and the population mean are related to their stratum counterparts as follows:

$$Y = \sum_{h=1}^{H}\sum_{k=1}^{N_h} Y_{hk} = \sum_{h=1}^{H} Y_h$$

$$\bar{Y} = \frac{1}{N}Y = \frac{1}{N}\sum_{h=1}^{H} Y_h = \frac{1}{N}\sum_{h=1}^{H} N_h\bar{Y}_h = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{Y}_h \quad . \qquad (3.3)$$

In other words, we can express the population total as the sum of the stratum totals, and the population mean appears to be a weighted mean of the stratum means. The weighting factor of the stratum mean corresponds with the relative size of the stratum concerned.

For the population variance σ_y^2 – see (2.1) – and the stratum variances σ_{yh}^2 the following applies:

$$\sigma_y^2 = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\sigma_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\bar{Y}_h - \bar{Y}\right)^2 \quad . \qquad (3.4^*)$$

The expression for the adjusted population variance S_y^2 can be rewritten in terms of the adjusted stratum variances, as follows:

$$S_y^2 = \sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N - 1}\right)\left(\bar{Y}_h - \bar{Y}\right)^2 \quad . \qquad (3.5^*)$$

The first term to the right of the equal sign is referred to as the *intrastrata variance* because it is composed from the individual stratum variances. The second term to the right of the equal sign is referred to as the *interstrata variance*, and gives an indication of the dispersion of the stratum means around the population mean.

Using equation (3.5), an expression is derived for the population coefficient of variation in terms of the coefficients of variation of the individual strata:

$$CV_y^2 = \sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1}\right)\left(\frac{\bar{Y}_h}{\bar{Y}}\right)^2 CV_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N - 1}\right)\left(\left(\frac{\bar{Y}_h}{\bar{Y}}\right) - 1\right)^2 \quad . \qquad (3.6^*)$$

3.3.3 Estimators of population mean and population total

SRSWOR is performed separately for each stratum in the stratified sample. That is, from the population of N_h elements in a stratum *h*, we perform SRSWOR to select a sample of size n_h. Estimators of stratum mean and then stratum total are directly available (see e.g. Chapter 2):

$$\begin{aligned} \hat{\bar{Y}}_h &= \bar{y}_h \\ \hat{Y}_h &= N_h \bar{y}_h \quad . \end{aligned} \tag{3.7}$$

These two estimators will be used in estimating the population total and population mean.

The population total Y corresponds with the sum of the stratum totals Y_h, see (3.3). It is therefore logical to estimate the population total by means of estimators of the stratum totals. An estimator of this kind that explicitly uses the stratification design is referred to as a *stratification estimator*. The *stratification estimator of the population total*, denoted as $\hat{Y}_{ST}$, is therefore by definition:

$$\hat{Y}_{ST} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \bar{y}_h \quad . \tag{3.8}$$

By analogy the *stratification estimator of the population mean*, denoted as $\hat{\bar{Y}}_{ST}$, is defined as:

$$\hat{\bar{Y}}_{ST} = \frac{1}{N} \hat{Y}_{ST} = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right) \bar{y}_h \quad . \tag{3.9}$$

To recap, both the population total and population mean can be estimated with a weighted sum of the stratum means. The respective stratum-dependent weighting factors are the absolute stratum size and the relative stratum size.

The unbiasedness of the estimators

SRSWOR is performed for each stratum to select a sample of size n_h. This type of random sample has several known properties – see Chapter 2. Accordingly, the estimators of the stratum mean and stratum total – see (3.7) – can be taken to be unbiased. Given the unbiasedness of the estimator of the stratum total, it is possible to calculate the expected value of the stratification estimator for the population total (3.8):

$$E\left(\hat{Y}_{ST}\right) = \sum_{h=1}^{H} E\left(\hat{Y}_h\right) = \sum_{h=1}^{H} Y_h = Y \quad .$$

The stratification estimator (3.8) is therefore an unbiased estimator of the population total. We use this result in calculating the expected value of the stratification estimator of the population mean:

$$E\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N} E\left(\hat{Y}_{ST}\right) = \frac{1}{N} Y = \bar{Y} \quad .$$

This demonstrates that $\hat{\bar{Y}}_{ST}$ is an unbiased estimator of the population mean.

The proofs for the unbiasedness (i.e. the formal derivations of the calculated expected values) of estimators (3.8) and (3.9) given above confirm what is clear intuitively. The fact is that if the estimators of the stratum totals are unbiased, then the weighted sum of the estimators of the stratum totals is also unbiased. Identical reasoning applies to the estimators of the stratum means and the weighted sum of the estimators of the stratum means.

Variance calculations

The variance of, for example, the estimator of the stratum total, can also be determined directly – see expression (2.12):

$$\text{var}\left(\hat{Y}_{yh}\right) = N_h \left(\frac{1 - f_h}{f_h} \right) S_{yh}^2 \quad . \tag{3.10}$$

In this formula f_h is the *sampling ratio* in stratum *h*, i.e. $f_h = n_h / N_h$. SRSWOR is performed independently in the various strata of the stratified sample. Consequently the variance of the stratification estimator of the population total equals the sum of the variances of the individual estimators of the stratum totals. That is,

$$\text{var}\left(\hat{Y}_{ST}\right) = \sum_{h=1}^{H} \text{var}\left(\hat{Y}_h\right) = N \sum_{h=1}^{H} \left(\frac{1 - f_h}{f_h} \right) \left(\frac{N_h}{N} \right) S_{yh}^2 \quad . \tag{3.11}$$

Using (3.11) and the natural relationship between the stratification estimators of the population mean and population total, the variance of the stratification estimator of the population mean can be calculated as:

$$\text{var}\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N^2} \text{var}\left(\hat{Y}_{ST}\right) = \frac{1}{N} \sum_{h=1}^{H} \left(\frac{1 - f_h}{f_h} \right) \left(\frac{N_h}{N} \right) S_{yh}^2 \quad . \tag{3.12}$$

Further inspection of formulas (3.11) and (3.12) reveals that both variances are small if and only if the individual adjusted stratum variances S_{yh}^2 are small. Small adjusted stratum variance means that the target variable varies little within the stratum, or in other words that the stratum is internally *homogeneous*. The variance of the estimator also appears to depend on the allocation scheme used, since the individual adjusted stratum variances depend on the size of the random samples.

The variance of the stratification estimator of the population total – see (3.10) – is a weighted linear combination of the adjusted stratum variances. An estimator of this variance can therefore be obtained by estimating each individual adjusted stratum variance. The sample variance in SRSWOR is known to be an unbiased estimator of the adjusted population variance. Using this result we obtain:

$$\hat{\text{var}}\left(\hat{Y}_{ST}\right) = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)\hat{S}_{yh}^2 = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)s_{yh}^2 \quad . \tag{3.13}$$

By analogy the variance of the stratification estimator of the population mean can be estimated as follows:

$$\hat{\text{var}}\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N^2}\,\hat{\text{var}}\left(\hat{Y}_{ST}\right) = \frac{1}{N}\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)s_{yh}^2 \quad . \tag{3.14}$$

Because the sample variance is an unbiased estimator of the adjusted population variance in SRSWOR – see formula (2.10) – we know that estimators (3.13) and (3.14) are unbiased.

Coefficients of variation

Analogous to definition (2.15), the coefficients of variation of the stratification estimators of the population total and population mean, are defined by:

$$\begin{aligned} CV\left(\hat{Y}_{ST}\right) &= \frac{\sqrt{\text{var}\left(\hat{Y}_{ST}\right)}}{Y} \\ CV\left(\hat{\bar{Y}}_{ST}\right) &= \frac{\sqrt{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}}{\bar{Y}} \quad . \end{aligned} \tag{3.15}$$

It can be established from formulas (3.11) and (3.12) that the above coefficients of variation are equal:

$$CV\left(\hat{\bar{Y}}_{ST}\right) = \frac{\sqrt{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}}{\bar{Y}} = \frac{\left(\frac{1}{N}\right)\sqrt{\text{var}\left(\hat{Y}_{ST}\right)}}{\left(\frac{1}{N}\right)Y} = CV\left(\hat{Y}_{ST}\right) \quad .$$

The stratification estimators of the population total and population mean are therefore equally accurate.

Expression (3.11) is used in determining $CV(\hat{Y}_{ST})$. The result can be rewritten in terms of the stratum coefficients of variation – see definition (3.1). After squaring, we obtain the following relationship between the coefficient of variation of the stratification estimator of the population mean, and the individual stratum coefficients of variation:

$$CV^2\left(\hat{Y}_{ST}\right) = N\sum_{h=1}^{N}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)\left(\frac{\bar{Y}_h}{\bar{Y}}\right)^2 CV_{yh}^2 \quad .$$

3.3.4 Sampling design evaluation

The principal idea of stratified sampling designs is that stratification yields a more precise estimator than simple random sampling. If the observations within the strata are generally more homogeneous than in the the population as a whole, the reduced variance in the strata will lead to a smaller variance of the estimator.

Sampling designs can be evaluated by comparing the variance of an estimator for a particular design with the variance of the corresponding estimator based on a SRSWOR sample of equal size (which is also known as the SRSWOR estimator). The ratio of the two variances is known as the *design effect*, or *DEFF*. The design effect for the stratification estimators of the population total and population mean, respectively, are defined (under the necessary condition that $\text{var}(\hat{Y}_{EAZT}) \neq 0$) as:

$$\begin{aligned} DEFF\left[\hat{Y}_{ST}\right] &= \frac{\text{var}\left(\hat{Y}_{ST}\right)}{\text{var}\left(\hat{Y}_{SRSWOR}\right)} \\ DEFF\left[\hat{\bar{Y}}_{ST}\right] &= \frac{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}{\text{var}\left(\hat{\bar{Y}}_{SRSWOR}\right)} \quad . \end{aligned} \tag{3.16}$$

A *DEFF* value of 1 indicates that the variances of the two estimators are equal; the precision of the stratification estimator is then the same as that of the SRSWOR estimator. *Mutatis mutandis*, a stratified random sample is more efficient than a SRSWOR sample of the same size if *DEFF is* less than 1, and less efficient if *DEFF* is greater than 1. The actual calculation of a design effect requires explicit calculations of both variances.

The design effect for the stratification estimator of the population mean can be calculated as follows:

$$DEFF\left[\hat{\bar{Y}}_{ST}\right] = \frac{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}{\text{var}\left(\hat{\bar{Y}}_{SRSWOR}\right)} = \frac{\left(\frac{1}{N^2}\right)\text{var}\left(\hat{Y}_{ST}\right)}{\left(\frac{1}{N^2}\right)\text{var}\left(\hat{Y}_{SRSWOR}\right)} = DEFF\left[\hat{Y}_{ST}\right] \quad .$$

In other words, the design effects for the stratification estimators of the population total and population mean are equal. With hindsight this is not surprising because the design effect is actually no more than the ratio of the squares of the coefficients of variation of the stratification estimator and the SRSWOR estimator.

A compact expression for the design effect of the stratification estimator of the population total can be derived if a simple relationship exists between $\text{var}(\hat{Y}_{ST})$ and $\text{var}(\hat{Y}_{SRSWOR})$. Unfortunately, there is no relationship of this kind for the general case. However, a relationship does exist under two, not particularly strict, conditions.

Assume for the remainder of the discussion in this subsection that the sampling ratio f_h is constant. This assumption simplifies equation (3.11) as follows:

$$\text{var}\left(\hat{Y}_{ST}\right) = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)S_{yh}^2 = N\left(\frac{1-f}{f}\right)\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_{yh}^2 \quad . \tag{3.17}$$

The variance of the SRSWOR estimator of the population total, based on a sample of the same size, depends on the adjusted population variance S_y^2 – see (2.12). Therefore assume in the second instance that the following conditions are fulfilled:

$$N_h \approx (N_h - 1) \quad \wedge \quad N \approx (N - 1) \quad . \tag{3.18}$$

Based on this assumption the expression for $\text{var}(\hat{Y}_{SRSWOR})$ can be rewritten as follows:

$$\text{var}\left(\hat{Y}_{SRSWOR}\right) \approx \text{var}\left(\hat{Y}_{ST}\right) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H} N_h\left(\bar{Y}_h - \bar{Y}\right)^2 \quad . \tag{3.19*}$$

To recap, the variance of the SRSWOR estimator for the population total can be rewritten as the sum of the variance of the stratification estimator of the population total and a term that is directly proportional to the interstrata variance. This result allows the design effect of the stratification estimator of the population total to be approximated as:

$$DEFF\left[\hat{Y}_{ST}\right] \approx \frac{\text{var}\left(\hat{Y}_{ST}\right)}{\text{var}\left(\hat{Y}_{ST}\right) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H} N_h\left(\bar{Y}_h - \bar{Y}\right)^2} \quad . \tag{3.20}$$

It can be derived from approximation (3.20) that the design effect of the stratification estimator is always less than or equal to 1. It also shows that a larger interstrata variance can yield more precise stratification estimators. It is therefore worthwhile in designing the stratification to maximise the spread of stratum means $\bar{Y}_h$ as much as possible.

To conclude, in a stratified sample with constant sampling ratio per stratum and sufficiently large population and strata, the stratification estimators of the population total and population mean are at least as precise as the corresponding SRSWOR estimators.

3.3.5 Allocation

The above sections referred to stratified sampling in which both the size N_h of the strata and the sample size n_h in the strata were given. Here we will not go into the questions of how to construct the strata or how many strata there should be (see, for example Cochran, 1977), but we will talk briefly about how many observations to take in each stratum. If we wish to estimate stratum means or other stratum parameters with a certain predetermined level of accuracy, we can determine the necessary stratum sample sizes using the formulas from Section 2.3.4. Collectively these samples form the total sample. Often, however, the total sample size will already be determined, and the question is how to distribute these *n* elements over the strata. This is known as the allocation problem. This section discusses three different allocation methods.

Proportional allocation

Proportional allocation is based on the original idea of a representative sample. The stratified sample is selected such that the sample reflects the population in terms of

the stratification variable(s). The size of the sample in each stratum is taken in proportion to the size of the stratum:

$$n_h = \frac{N_h}{N} \times n \quad . \tag{3.21}$$

If necessary, noninteger values of n_h are rounded to get integer sample sizes. Using the (unrounded) stratum sample size (3.21), the sampling ratio of each stratum can be calculated as:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} \quad .$$

In other words, the sampling ratio per stratum is constant. The condition underlying (3.17) is then satisfied automatically. Furthermore results (3.19) and (3.20) are true under the additional condition (3.18).

Finally, the inclusion probability of element *k* in stratum *h* can be calculated as:

$$\pi_{hk} = \frac{n_h}{N_h} = \frac{N_h}{N}\frac{n}{N_h} = \frac{n}{N} \quad .$$

All elements in the population, irrespective of stratum, have the same probability of being selected in the sample, and therefore have equal weight in the estimation. We refer in this case to a *self-weighting* sample.

In SRSWOR, every element of the population also has a probability n/N of being selected in the sample. However, a stratified sample excludes some extreme samples that would not be excluded in SRSWOR (e.g. a sample of only high or low incomes).

Optimum allocation

When the various adjusted stratum variances are equal, proportional allocation is the best allocation method for improving precision. In a sample where the elements differ in size (e.g. companies, schools, municipalities), the larger elements will generally exhibit greater variability on the target variable than the smaller units. An example would be companies' international trade; larger companies will exhibit greater differences in export and import than smaller companies. In this case we would like to have a higher proportion of larger companies in the sample.

If the adjusted stratum variances vary strongly, *optimum* allocation of the sample leads to minimum variances of the estimators for the population total and mean. In this allocation method the sample size in stratum *h* is taken equal to:

$$n_h = \frac{N_h S_{yh}}{\sum_{h=1}^{H} N_h S_{yh}} n \quad . \tag{3.22}$$

where n_h is again rounded to a whole number. Optimum allocation does lead to equal inclusion probabilities for all elements of the population; the probability of being included in the sample is proportional to S_{yh} and therefore varies between the strata.

The number of elements selected in a stratum is greater when a stratum represents a greater proportion of the population *and* when the stratum is relatively inhomogeneous, i.e. has a relatively large S_{yh}. In applying formula (3.22) the calculated sample size may be larger than the actual size of the stratum, in which case the entire stratum is included.

Determining the optimum allocation requires knowledge of (the ratios of) the adjusted stratum variances. This information will seldom be known in practice, but the variances can sometimes be approximated based on earlier surveys. If the stratum variances may be expected not to diverge too much, so one can assume that $S_{yh} = S_y$, formula (3.22) reduces to proportional allocation.

Costs

Besides estimator precision, costs are obviously also relevant in sample surveys. Preferably, the allocation should take into account any differences in costs per observation between strata, e.g. because of different data collection techniques employed in the strata, by requiring fewer observations in relatively expensive strata. Suppose that a certain budget is available for field work, that is the maximum total costs are *C*. A simple cost function would then be

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \,, \tag{3.23}$$

where c_0 represents fixed overheads and c_h the (variable) costs of an observation in stratum *h*. We will now distribute the random sample among the strata such as to minimize the variance of the estimator given the total costs *C*. Condition (3.23) replaces the condition that the total sample size must be equal to *n*. It can be proved, given this condition, that taking the sample size in stratum *h* equal to

$$n_h = \frac{\frac{N_h S_{yh}}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_{yh}}{\sqrt{c_h}}} n \,, \tag{3.24}$$

minimizes the variance of the stratification estimator of the population total. The proof is given in Cochran (1977, Chapter 5). Thus we select more elements from a stratum if it represents a large proportion of the population, the variance within the stratum is large, and observation in the stratum is inexpensive. Optimum allocation (3.22) is actually a special case of (3.24) where the costs per stratum are equal. Optimum allocation is also known as Neyman allocation.

3.3.6 Fractions

If the target variable *Y* is an indicator variable that can only assume the values 0 and 1, the mean equals a fraction (see also Section 2.4). The fraction of elements with a given property (the fraction of 'successes') in the sample from stratum *h* is therefore $p_h = \bar{y}_h$. The fraction of successes in the population can be estimated using stratification estimator (3.9), i.e.:

$$\hat{P}_{ST} = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right) p_h \quad .$$

The variance of this stratification estimator is calculated in formula (3.12). In practical situations this variance can be estimated using formula (3.14), i.e.:

$$\hat{\text{var}}\left(\hat{P}_{ST}\right) = \sum_{h=1}^{H} (1 - f_h) \left(\frac{N_h}{N} \right)^2 \left(\frac{p_h (1 - p_h)}{(n_h - 1)} \right)$$

because the stratum variance can be calculated as:

$$s_h^2 = \left(\frac{n_h}{n_h - 1} \right) p_h (1 - p_h) \quad .$$

3.4 Quality indicators

An important quality criterion in stratified sampling is the reduction in variance compared with simple random sampling without replacement. This criterion is expressed in the design effect *DEFF*, discussed in Section 3.3.4. The greater the reduction in variance, the smaller the *DEFF*. The reduction in variance is particularly large when the stratum means for the target variable *Y* diverge greatly.

3.5 Appendix

Proof of (3.4)

It can be derived from definition (2.1) for the population variance that:

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{N_h} \left(Y_{hk} - \bar{\bar{Y}} \right)^2 \\
&= \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{N_h} \left(Y_{hk} - \bar{\bar{Y}}_h + \bar{\bar{Y}}_h - \bar{\bar{Y}} \right)^2 \\
&= \frac{1}{N} \left[\sum_{h=1}^{H} \sum_{k=1}^{N_h} \left(Y_{hk} - \bar{\bar{Y}}_h \right)^2 + \sum_{h=1}^{H} \sum_{k=1}^{N_h} \left(\bar{\bar{Y}}_h - \bar{\bar{Y}} \right)^2 \right] + \\
&\quad + \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{N_h} 2 \left(Y_{hk} - \bar{\bar{Y}}_h \right) \left(\bar{\bar{Y}}_h - \bar{\bar{Y}} \right) \quad .
\end{aligned}$$

We calculate the double product in the above term as follows:

$$\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\bar{Y}_h\right)\left(\bar{Y}_h-\bar{Y}\right)=\sum_{h=1}^{H}\left(\bar{Y}_h-\bar{Y}\right)\sum_{k=1}^{N_h}\left(Y_{hk}-\bar{Y}_h\right)$$
$$=\sum_{h=1}^{H}\left(\bar{Y}_h-\bar{Y}\right)\left[\sum_{k=1}^{N_h}Y_{hk}-N_h\bar{Y}_h\right]=0 \quad .$$

For population variance σ_y^2 we then find:

$$\sigma_y^2=\frac{1}{N}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\bar{Y}_h\right)^2+\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\bar{Y}_h-\bar{Y}\right)^2\right]$$
$$=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\sigma_{yh}^2+\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\bar{Y}_h-\bar{Y}\right)^2$$

according to definition (3.1), which proves (3.4). QED.

Proof of (3.5)

It can be derived from the definition of adjusted population variance in (2.1) that:

$$S_y^2=\frac{1}{N-1}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\bar{Y}\right)^2$$
$$=\frac{1}{N-1}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\bar{Y}_h\right)^2+\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\bar{Y}_h-\bar{Y}\right)^2\right]+$$
$$+\frac{1}{N-1}\sum_{h=1}^{H}\sum_{k=1}^{N_h}2\left(Y_{hk}-\bar{Y}_h\right)\left(\bar{Y}_h-\bar{Y}\right) \quad .$$

The double product in the above expression is known to be 0. The above expression can therefore be simplified as follows:

$$S_y^2=\frac{1}{N-1}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\bar{Y}_h\right)^2+\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\bar{Y}_h-\bar{Y}\right)^2\right]$$
$$=\sum_{h=1}^{H}\left(\frac{N_h-1}{N-1}\right)S_{yh}^2+\sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\bar{Y}_h-\bar{Y}\right)^2$$

according to definition (3.1) for S_{yh}^2, which proves equation (3.5). QED.

Proof of (3.6)

According to definition (2.1) for the population coefficient of variation:

$$CV_y^2=\frac{S_y^2}{\bar{Y}^2} \quad .$$

Using equation (3.5) for the adjusted population variance we can write:

$$CV_y^2 = \sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1}\right)\frac{S_{yh}^2}{\bar{Y}^2} + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\frac{\left(\bar{Y}_h - \bar{Y}\right)^2}{\bar{Y}^2}$$
$$= \sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1}\right)\left(\frac{\bar{Y}_h}{\bar{Y}}\right)^2 CV_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\left(\frac{\bar{Y}_h}{\bar{Y}}\right) - 1\right)^2 \quad .$$

This proves (3.6). QED.

Proof of (3.19)

Equation (3.5), which relates the adjusted population variance to the stratum variances, can be rewritten as follows based on approximation (3.18):

$$S_y^2 = \sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\bar{Y}_h - \bar{Y}\right)^2$$
$$\approx \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\bar{Y}_h - \bar{Y}\right)^2 \quad .$$

The variance of the SRSWOR estimator based on a sample of the same size produced with SRSWOR can be calculated with (2.12). Using the result above the variance can be approximated as follows:

$$\mathrm{var}\left(\hat{Y}_{SRSWOR}\right) \approx N\left(\frac{1-f}{f}\right)\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_{yh}^2 + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H}N_h\left(\bar{Y}_h - \bar{Y}\right)^2$$
$$= \mathrm{var}\left(\hat{Y}_{ST}\right) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H}N_h\left(\bar{Y}_h - \bar{Y}\right)^2 \quad .$$

This proves equation (3.19). QED.

4. Cluster sampling, Two-stage sampling and Systematic sampling

4.1 Short description

The principle of the three types of sampling designs discussed in this chapter is a division of the target population into *clusters*. The division must be such that the clusters do not overlap and together they cover the whole population.

The clusters in a cluster sample are selected by random sampling, and then each selected cluster is observed in full. Cluster sampling can therefore be interpreted as random sampling of groups. In this case too we have the choice of selecting the clusters with or without replacement.

The first stage of two-stage sampling, as in cluster sampling, involves selecting clusters at random. Subsequently the second stage of two-stage sampling involves taking a further random sample of elements from each selected cluster.

The systematic sampling design is also treated in this chapter because it can be regarded as a kind of cluster sampling.

4.2 Applicability

The previous chapter introduced stratification as a sampling method in which the population is first divided into subpopulations (the strata), and then selecting a sample from each separate stratum. Stratified sampling generally increases the precision of estimators. Although at first glance cluster sampling closely resembles stratified sampling, it has very different properties. For instance, cluster sampling usually leads to a lower precision than simple random sampling. Cluster sampling is therefore applied only when required by the practical situation, or when the loss of precision is compensated by a substantial reduction in the (data collection) costs. The loss of precision arises because, unlike stratified sampling, cluster sampling does not select all subpopulations.

Nonetheless the estimators can sometimes gain precision in a second stage by drawing samples from the clusters that were selected in the first stage. Because the clusters are no longer fully observed, one could consider selecting a few more clusters in the first stage in order to create a clearer picture of the overall population. Two-stage sampling is particularly useful with clusters that are homogeneous with respect to the target variable.

An example of two-stage sampling is a clustering of all households in the Netherlands according to regions or districts. A sample of households is selected in the second stage from each district selected in the first stage. The advantage of a regional cluster sample of this kind is the reduction in total interviewer travel expenses compared with a random sample of all Dutch households. The travel expenses are reduced because an interviewer can now interview a relatively large number of individuals who live close together in clusters.

All sampling designs discussed so far have assumed the availability of a satisfactory sampling frame for the target population. Clearly, this will not always be the case. However, sampling frames may be available for subpopulations of the target population. Suppose we want to conduct a survey of secondary school students. There is no readily available comprehensive list of secondary school students, but maybe we can lay our hands on a list of secondary schools without much difficulty. Each school will obviously know the identities of its students. Other examples of 'natural' groups in the population include municipalities, households, businesses and care homes. These groups then form the clusters.

4.3 Detailed description

This section describes three sampling designs. Firstly, we describe cluster sampling. That is, the clusters selected in the sample are completely observed. Secondly, we describe two-stage sampling where in the second stage subsamples are drawn from the clusters selected in the first stage. Finally, systematic sampling is discussed which can be seen as a special case of cluster sampling.

4.3.1 Cluster sampling

4.3.1.1 Assumptions and notation

With respect to the target variable Y we distinguish the five parameters for the population defined in (2.1).

As with the stratified sampling design, cluster sampling subdivides the population into subpopulations. A subpopulation is referred to as a *cluster* or *primary sampling unit* (PSU). An implicit assumption is that the population is fully covered by the clusters, and that the separate clusters have no common elements.

The clusters in the sampling design are initially interpreted as survey units (which is why the clusters are not allowed to overlap in the population). A specific cluster is identified with an index *d*, $d = 1, \cdots, N$; the number of elements in cluster *d* is denoted by M_d. The elements of a primary sampling unit are referred to as the *secondary sampling units* (SSU). The cluster size M_d is normally assumed known.

The total number of elements in the population is denoted by *M*. Therefore

$$\sum_{d=1}^{N} M_d = M \quad .$$

This total population size can be averaged over the number of clusters. The result is referred to as the *mean cluster size*, denoted by $\overline{M}$:

$$\overline{M} = \frac{M}{N} \quad .$$

The value of the target variable for element *k* in cluster *d* is denoted by Y_{dk} for $d = 1, \cdots, N$ and $k = 1, \cdots, M_d$. With respect to the target variable, the following

parameters are identified for each cluster: the *cluster total* Y_d, the *cluster mean* $\bar{Y}_d$, the *cluster variance* σ^2_{yd}, and the *adjusted cluster variance* S^2_{yd}. These parameters are defined by:

$$\begin{aligned} Y_d &= \sum_{k=1}^{M_d} Y_{dk} \\ \bar{Y}_d &= \frac{1}{M_d} Y_d \\ \sigma^2_{yd} &= \frac{1}{M_d} \sum_{k=1}^{M_d} \left(Y_{dk} - \bar{Y}_d\right)^2 \\ S^2_{yd} &= \frac{1}{M_d - 1} \sum_{k=1}^{M_d} \left(Y_{dk} - \bar{Y}_d\right)^2 \quad . \end{aligned} \tag{4.1}$$

To these four cluster parameters can be added the *mean cluster total* $\bar{Y}_{CT}$ and the *adjusted variance of cluster totals* S^2_{yCT}. These are defined as follows:

$$\begin{aligned} \bar{Y}_{CT} &= \frac{1}{N} Y = \frac{1}{N} \sum_{d=1}^{N} Y_d \\ S^2_{yCT} &= \frac{1}{N-1} \sum_{d=1}^{N} \left(Y_d - \bar{Y}_{CT}\right)^2 \quad . \end{aligned} \tag{4.2}$$

A simple random sample of *n clusters* is selected from the population of *N* clusters. Each of the selected clusters is observed in full, so that for each cluster there are m_d observations with $m_d = M_d$. It is therefore possible to interpret the random cluster sample as a simple random sample of *n* out of *N* primary sample units with the cluster totals as observations. The *n* observed cluster totals are denoted by lower case letters $y_1, \cdots, y_n$.

We further define the sample mean of cluster totals $\bar{y}_{CT}$, and the sample variance of cluster totals s^2_{yCT} as:

$$\begin{aligned} \bar{y}_{CT} &= \frac{1}{n} \sum_{d=1}^{n} y_d \\ s^2_{yCT} &= \frac{1}{n-1} \sum_{l=1}^{n} \left(y_d - \bar{y}_{CT}\right)^2 \quad . \end{aligned} \tag{4.3}$$

4.3.1.2 Relationships between population parameters and cluster parameters

The relationships between the population parameters and their cluster counterparts are shown below:

$$\begin{aligned} Y &= \sum_{d=1}^{N} \sum_{k=1}^{M_d} Y_{dk} = \sum_{d=1}^{N} Y_d \\ \bar{Y} &= \frac{Y}{M} = \frac{1}{M} \sum_{d=1}^{N} Y_d = \frac{1}{M} \sum_{d=1}^{N} M_d \bar{Y}_d = \sum_{d=1}^{N} \left(\frac{M_d}{M}\right) \bar{Y}_d \quad . \end{aligned}$$

The above relationships state that the population total equals the sum of all cluster totals, and the population mean can be interpreted as a weighted sum of the cluster means. Both relationships appear intuitive.

The population variance can be determined from the variances in the subpopulations, as with stratification – see Chapter 3. The following relationship can be demonstrated:

$$\sigma_y^2 = \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\overline{Y}_d - \overline{\overline{Y}}\right)^2 \qquad (4.4^*)$$

Assuming (4.4) it is possible to derive a relationship between the adjusted population variance S_y^2 and the individual adjusted cluster variances S_{yd}^2:

$$S_y^2 = N\left(\frac{\overline{M}-1}{M-1}\right)S_{intra}^2 + \overline{M}\left(\frac{N-1}{M-1}\right)S_{inter}^2 \qquad (4.5^*)$$

with:

$$S_{intra}^2 = \frac{1}{N}\sum_{d=1}^{N}\left(\frac{M_d - 1}{\overline{M}-1}\right)S_{yd}^2$$

$$S_{inter}^2 = \frac{1}{N-1}\sum_{d=1}^{N}\left(\frac{M_d}{\overline{M}}\right)\left(\overline{Y}_d - \overline{\overline{Y}}\right)^2 \quad .$$

The first term (on the right-hand side of the '=' sign) in (4.5) is directly proportional to S_{intra}^2, which is referred to as the *intracluster variance*. The intracluster variance is composed of the *N* adjusted cluster variances. The second term (on the right-hand side of the '=' sign) in (4.5) is directly proportional to S_{inter}^2, which is referred to as the *intercluster variance*. The intercluster variance depends on the difference between the cluster mean and population mean for each cluster.

4.3.1.3 Estimators of the population mean and population total

In essence the cluster sample is a simple random sample with the clusters as units and the cluster totals as observations. Assuming furthermore that the clusters are sampled without replacement, then the cluster sample is simply the result of SRSWOR applied to cluster totals. This type of sample was extensively analysed in Chapter 2. Estimators of population parameters based on this specific interpretation of cluster sampling as the sampling of cluster totals,are known as *cluster estimators*.

Based on the concept of SRSWOR of cluster totals, the estimator for the mean cluster total is straightforwardly given by

$$\hat{\overline{Y}}_{CT} = \overline{y}_{CT} = \frac{1}{n}\sum_{d=1}^{n} y_d \ .$$

Similarly, the estimator for the adjusted variance of cluster totals is

$$\hat{S}_{yCT}^2 = s_{yCT}^2 = \frac{1}{n-1}\sum_{d=1}^{n}\left(y_d - \overline{y}_{CT}\right)^2 \ .$$

There is a simple relationship between the population total Y and the mean cluster total $\bar{Y}_{CT}$, which can be derived from (2.1) and (4.2). There is therefore also a similar relationship between the population mean and the mean cluster total. Both relationships are shown below:

$$Y = N\,\bar{Y}_{CT} \qquad \wedge \qquad \bar{Y} = \frac{Y}{M} = \frac{1}{\bar{M}}\bar{Y}_{CT} \quad . \tag{4.6}$$

Based on these relationships and the estimator for the mean cluster total, we arrive at a cluster estimator for the population total Y, denoted by $\hat{Y}_{CL}$, and a cluster estimator for the population mean, denoted by $\hat{\bar{Y}}_{CL}$.

$$\begin{aligned} \hat{Y}_{CL} &= N\,\hat{\bar{Y}}_{CT} = N\,\bar{y}_{CT} = \left(\frac{1}{f}\right)\sum_{d=1}^{n} y_d \\ \hat{\bar{Y}}_{CL} &= \frac{\hat{Y}_{CL}}{M} = \frac{1}{\bar{M}}\hat{\bar{Y}}_{CT} = \frac{1}{\bar{M}}\,\bar{y}_{CT} = \left(\frac{1}{f}\right)\left(\frac{1}{M}\right)\sum_{d=1}^{n} y_d \quad . \end{aligned} \tag{4.7}$$

In these two formulas $f = n/N$ represents the fraction of clusters in the random sample. Note that the sample mean $\bar{y}_{CT}$ of cluster totals arises naturally in these estimators.

The unbiasedness of the estimators

Because the cluster sampling procedure is SRSWOR of cluster totals, the estimator of the mean cluster total is known to be unbiased. Based on the same argument it can also be concluded that the sample variance of cluster totals is an unbiased estimator of the adjusted variance of cluster totals.

By definition the cluster estimators of the population total and the population mean are scaled versions of the estimator for the mean cluster total. This fact together with the unbiasedness of the estimator of the mean cluster total and the dependencies (4.6) mean that the cluster estimators of the population total and the population mean in (4.7) are both unbiased.

Variance calculations

We have seen that the unbiasedness of the estimator of the mean cluster total is the basis of the unbiasedness of the two cluster estimators. Similarly, the variance calculation for the estimator of the mean cluster total is the basis of the variance calculations for both cluster estimators.

The variance of the estimator of the mean cluster total can be calculated as follows – see also (2.11):

$$\operatorname{var}\left(\hat{\bar{Y}}_{CT}\right) = \operatorname{var}\left(\bar{y}_{CT}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)S^2_{yCT} \quad . \tag{4.8}$$

We can infer from this result that the variance of the cluster estimator of the population total is:

$$\text{var}\left(\hat{Y}_{CL}\right) = N^2 \,\text{var}\left(\hat{\bar{Y}}_{CT}\right) = N\left(\frac{1-f}{f}\right)S^2_{yCT} \tag{4.9}$$

and that for the variance of the cluster estimator of the population mean:

$$\text{var}\left(\hat{\bar{Y}}_{CL}\right) = \text{var}\left(\frac{\hat{Y}_{CL}}{M}\right) = \frac{1}{M^2}\,\text{var}\left(\hat{Y}_{CL}\right) = \left(\frac{1}{M}\right)\left(\frac{1}{\bar{M}}\right)\left(\frac{1-f}{f}\right)S^2_{yCT} \quad . \tag{4.10}$$

Estimators for results (4.8) - (4.10) can be obtained by replacing S^2_{yCT} in the corresponding formulas by s^2_{yCT}. Because the sample variance of cluster totals is an unbiased estimator of the adjusted variance of cluster totals, the estimators of the variances (of the cluster estimators) produced in this way are unbiased.

4.3.1.4 Sampling design evaluation

To evaluate the cluster sample the design effects of the cluster estimators of the population total and population mean have to be calculated. This involves comparing the variance of a cluster estimator with the variance of the SRSWOR estimator based on a sample of the same size. Unfortunately, the size of a cluster sample is not known in advance. However, on average a cluster sample consists of $(n \times \bar{M})$ individual observations. The comparison with the SRSWOR estimator is therefore also based on a sample of $(n \times \bar{M})$ out of M elements.

The design effects of the cluster estimators for the population total and population mean, analogous to (3.16), are defined as (assume implicitly that: $\text{var}(\hat{Y}_{SRSWOR}) \neq 0$):

$$\begin{aligned} DEFF\left(\hat{Y}_{CL}\right) &= \frac{\text{var}\left(\hat{Y}_{CL}\right)}{\text{var}\left(\hat{Y}_{SRSWOR}\right)} \\ DEFF\left(\hat{\bar{Y}}_{CL}\right) &= \frac{\text{var}\left(\hat{\bar{Y}}_{CL}\right)}{\text{var}\left(\hat{\bar{Y}}_{SRSWOR}\right)} \quad . \end{aligned} \tag{4.11}$$

Definition (4.11) formally specifies the design effects of the cluster estimators of the population total and the population mean. However, the result given below shows that for the remainder of this subsection it suffices to refer only to the design effect of the cluster estimator of the population total.

$$DEFF\left[\hat{\bar{Y}}_{CL}\right] = \frac{\text{var}\left(\hat{\bar{Y}}_{CL}\right)}{\text{var}\left(\hat{\bar{Y}}_{SRSWOR}\right)} = \frac{\left(\frac{1}{M^2}\right)\text{var}\left(\hat{Y}_{CL}\right)}{\left(\frac{1}{M^2}\right)\text{var}\left(\hat{Y}_{SRSWOR}\right)} = DEFF\left[\hat{Y}_{CL}\right] \quad .$$

The variance of the SRSWOR estimator of the population total is

$$\text{var}\left(\hat{Y}_{SRSWOR}\right) = M\left(\frac{1-f}{f}\right)S_y^2$$

with $f = (n \times \overline{M})/M = n/N$. The design effect of the cluster estimator of the population total can be calculated from this result and (4.9) as follows:

$$DEFF\left[\hat{Y}_{CL}\right] = \frac{\text{var}\left(\hat{Y}_{CL}\right)}{\text{var}\left(\hat{Y}_{SRSWOR}\right)} = \frac{N\left(\frac{1-f}{f}\right)S_{yCT}^2}{M\left(\frac{1-f}{f}\right)S_y^2} = \frac{1}{\overline{M}}\frac{S_{yCT}^2}{S_y^2} \quad . \tag{4.12}$$

The design effect is therefore determined by the ratio of the adjusted variance of the cluster totals and the adjusted population variance. The adjusted population variance is fixed and is not influenced by the sampling design. The situation is different for the adjusted variance of cluster totals, which is fully determined by the design of the cluster sample.

A detailed analysis of the design effect is possible under an additional assumption. This assumption guarantees a simple relationship between S_{yCT}^2 and S_y^2. We therefore make this assumption in the remainder of this subsection.

Clusters of equal size

We continue the analysis of the design under the assumption that all clusters contain an equal number of elements. It is known because of this assumption that:

$$M_d = \overline{M} = \frac{M}{N} \quad .$$

The results for a simple random sample of cluster totals simplify substantially under this assumption. For instance, expression (4.5) for the total adjusted variance can be written as follows:

$$S_y^2 = \left(\frac{M-N}{M-1}\right)S_{intra}^2 + \left(\frac{1}{\overline{M}}\right)\left(\frac{N-1}{M-1}\right)S_{yCT}^2 \quad . \tag{4.13*}$$

This result yields the required relationship between the adjusted variance of cluster totals and the adjusted population variance. It is now possible to further develop formula (4.12). Inspecting formulas (4.5) and (4.13) it can be concluded that, under the current assumption, the intercluster variance is directly proportional to the adjusted variance of cluster totals S_{yCT}^2.

Relationship (4.13) facilitates explicit calculation of the design effect of the cluster estimator of the population total. The result of the calculation is shown below:

$$DEFF\left[\hat{Y}_{CL}\right] = -\left(\frac{M-N}{N-1}\right)\left(\frac{S_{intra}^2}{S_y^2}\right) + \left(\frac{M-1}{N-1}\right) \quad . \tag{4.14*}$$

The first term on the right-hand side of the '=' sign is directly proportional to the ratio of the intracluster variance and the population variance. In other words, the

design effect (of the cluster estimator of the population total) is a decreasing linear function of intracluster variance S^2_{intra}.

Formula (4.14) shows an indirect relationship between the way of clustering and the design effect of the estimator, because the intracluster variance is dependent on the cluster design. This means that the design effect is maximal for minimum intracluster variance and minimal for maximum intracluster variance. The range of the design effect can therefore be determined as:

$$0 \leq DEFF\left[\hat{Y}_{CL}\right] \leq \left(\frac{M-1}{N-1}\right) \quad . \tag{4.15*}$$

The lower range limit of the design effect (of the cluster estimator of the population total) is reached if and only if the intercluster variance is zero – see also Equation (4.5). Cluster sampling often appears less efficient in practice than SRSWOR. This loss of efficiency therefore has to be compensated by lower costs if cluster sampling is actually to be used.

Generally, a larger intracluster variance leads to a cluster estimator (of the population total) with greater precision. In turn, a larger intracluster variance corresponds with less internal homogeneity of the clusters. It is therefore advisable to make the clusters as diverse as possible.

An alternative analysis of the design effect of the cluster estimator of the population total uses the *intraclass correlation coefficient*. This alternative analysis is set out in the remainder of this subsection..

The *intraclass correlation coefficient* ρ_c is defined as the Pearson correlation coefficient for the $N \times \overline{M}(\overline{M}-1)$ paired observations (Y_{dk}, Y_{dl}) within the clusters, where $k \neq l$ and $d = 1, \cdots, N$. The formal definition is:

$$\rho_c = \frac{\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)}{\left(M-1\right)\left(\overline{M}-1\right)S_y^2} \tag{4.16}$$

The intraclass correlation coefficient, or *intracluster correlation coefficient*, is a measure of the homogeneity within the clusters. It indicates how alike or different the elements in a cluster are. Maximum homogeneity of the clusters corresponds with $\rho_c = 1$. Mutatis mutandis, minimum cluster homogeneity corresponds with maximum intracluster variance.

Closer examination reveals the possibility of expressing the intracluster correlation coefficient in terms of the ratio of the adjusted variance of cluster totals and the adjusted population variance, i.e.:

$$\rho_c = \left(\frac{N-1}{\left(M-1\right)\left(\overline{M}-1\right)}\right)\left(\frac{S^2_{yCT}}{S_y^2}\right) - \frac{1}{\overline{M}-1} \quad . \tag{4.17*}$$

This result shows that the intracluster correlation coefficient is a linearly increasing function in the adjusted variance of cluster totals S^2_{yCT}. In other words, ρ_c is an

increasing function in the intercluster variance S^2_{inter}, because the latter is a multiple of S^2_{yCT}. To conclude, the coefficient ρ_c is minimal for the smallest possible intercluster variance and is maximal for the greatest possible intercluster variance. The range of the intracluster correlation coefficient can thus be determined directly and equals:

$$-\left(\frac{1}{\overline{M}-1}\right) \leq \rho_c \leq 1 \quad . \tag{4.18*}$$

Finally, equation (4.16) allows the design effect of the cluster estimator of the population total to be written as a function of ρ_c. Analysis shows that for $N >> 1$:

$$\begin{aligned} DEFF\left[\hat{Y}_{CL}\right] &= \frac{M-1}{M-\overline{M}}\left\{1+\left(\overline{M}-1\right)\rho_c\right\} \\ &\approx \left\{1+\left(\overline{M}-1\right)\rho_c\right\} \quad . \end{aligned} \tag{4.19*}$$

The design effect is therefore a linear function of ρ_c; the approximation in (4.19) holds when the number of clusters *N* is large. The linearly increasing nature of the relationship means that the design effect reaches a minimum value for the smallest possible intracluster correlation coefficient. Mutatis mutandis, the maximum design effect is obtained for the greatest possible intracluster correlation coefficient. The intraclass correlation coefficients is usually positive in practice. In particular when the clusters form 'natural' groups in the population, the elements within a cluster will resemble each other more than elements selected at random from the population. This resemblance may be attributable to coming from a common background or environment. When the elements within a cluster are similar to each other, the intracluster variance S^2_{intra} will be relatively small compared with the adjusted population variance S^2_y, so that ρ_c will be positive. In addition, ρ_c will be negative only if the elements in a cluster exhibit a greater variance than a randomly selected group of elements. This will seldom be the case, but may occur with artificially formed clusters in which elements are allocated randomly to clusters.

4.3.2 Two-stage sampling

The cluster sample discussed above always involved investigating all elements of every selected cluster. It was shown that cluster sampling is not very efficient when the clusters are fairly homogeneous. If the elements within a cluster very closely resemble each other, it may actually be a waste of time and money to investigate all elements; the same information could have been obtained by examining just a couple of elements. In such cases cluster sampling will perform fairly poorly, and it may be worthwhile to take a sample from the selected clusters. This is referred to as *two-stage sampling*.

The first stage of two-stage sampling involves selecting clusters or *primary sampling units* at random. The second stage of two-stage sampling involves randomly selecting elements, which are now referred to as *secondary sampling units*, from each selected primary sampling unit.

Assuming that the clusters are homogeneous, two-stage sampling generally allows more clusters to be selected in the first sampling stage, which may increase precision. This method gives greater control of sample size where the sizes of primary sample units vary greatly. Note that, in this case too, a sampling frame is required only for the selected primary sample units.

The costs and precision of two-stage sampling fall roughly between those of cluster sampling and stratified sampling. Two-stage sampling is generally more expensive than cluster sampling with the same sample size, but less expensive than stratified sampling. On the other hand, two-stage sampling is generally more precise than cluster sampling and less precise than stratified sampling.

It is easy to extend two-stage sampling to multistage sampling (with three, four, or more stages). For example, selecting a sample of secondary school students could involve selecting a sample of schools, then selecting a number of classes within a selected school, and finally selecting a number of students within a selected class. Another example is sampling successively by region, city, postcode area, and address. The units in the final stage are the sample elements. Here, the discussion is confined to two-stage sampling, because of its simplicity, but also because it is a more common design in practice than multistage sampling.

The design of two-stage sampling involves determining the selection methods for both the primary and secondary sampling units. The two methods do not have to be the same. It is possible to select units in various ways at each stage: random, or systematic, or proportional to size (see Sections 4.3.3 and 5.1.2 for details). The units may also be stratified first.

Apart from the selection methods in the two stages, it is also necessary to decide on the distribution of the sample over the first and the second stages. Is it better to select a relatively large number of primary sampling units and few secondary sampling units, or relatively few primary sampling units and many secondary sampling units? The first case will generally yield more precise estimators, but often at greater expense.

4.3.2.1 Assumptions and notation

The current analysis assumes SRSWOR in both stages. The notation used is almost analogous to that for the cluster sampling design. The population is again divided into *N* nonoverlapping clusters, or primary sampling units. In the first stage *n* primary sampling units are selected. Then in the second stage, m_d *secondary sampling units* are selected from each primary sampling unit, where in general $m_d \neq M_d$. Sample size m_d, $d = 1, \cdots, n$ is assumed to be determined in advance. The measured value of the target variable for element *k* in selected cluster *d* is denoted by y_{dk}, $k = 1,...,m_d$. Furthermore, define for each (sub)sample the two quantities $\bar{y}_{sd}$ and s_{yd}^2 as:

$$\bar{y}_{sd} = \frac{1}{m_d}\sum_{k=1}^{m_d} y_{dk}$$

$$s_{yd}^2 = \frac{1}{m_d - 1}\sum_{k=1}^{m_d}\left(y_{dk} - \bar{y}_{sd}\right)^2 \quad .$$

(4.20)

4.3.2.2 Estimators in two-stage sampling

Two-stage sampling does not involve observing all secondary sampling units in the selected primary sampling units. The cluster parameters of the (selected) primary sampling units therefore have to be estimated. The cluster total y_d in primary *sampling* unit *d* can be estimated by:

$$\hat{y}_d = M_d \bar{y}_{sd} = \frac{M_d}{m_d}\sum_{k=1}^{m_d} y_{dk} \qquad (d = 1, \cdots, n). \tag{4.21}$$

In other words, the estimated cluster total corresponds with a weighted sum of the observed secondary sampling units (in the primary sampling unit referred to previously). The weighting factor is the reciprocal of the sampling fraction for primary sampling unit *d*.

The population mean $\bar{Y}_{CT}$ of cluster totals can be estimated by:

$$\hat{\bar{Y}}_{CT} = \bar{\hat{y}}_{CT} = \frac{1}{n}\sum_{d=1}^{n} \hat{y}_d \quad .$$

Consequently, the cluster estimators of the population total and population mean are

$$\hat{Y}_{CL2} = N\hat{\bar{Y}}_{CT} = \frac{N}{n}\sum_{d=1}^{n} \hat{y}_d$$

$$\hat{\bar{Y}}_{CL2} = \frac{1}{M}\hat{\bar{Y}}_{CT} = \frac{N}{nM}\sum_{d=1}^{n} \hat{y}_d \quad .$$

The unbiasedness of the estimators

The fact that a selected cluster (primary sampling unit) is no longer observed in full, but through an SRSWOR sample of m_d out of M_d, has no influence on the analysis of bias for the estimators. That is, all estimators defined so far in this subsection are unbiased.

Variance calculations

The variances of the estimators $\hat{Y}_{CL}$ and $\hat{\bar{Y}}_{CL}$ in conventional cluster sampling are determined only by the differences between the clusters. However, the estimated cluster totals $\hat{y}_d$ in two-stage sampling are stochastic variables. Consequently, the variation within the clusters also contributes to the variance of the estimators. The variance of the cluster estimator of the population total $\hat{Y}_{CL2}$ therefore comprises an additional term :

$$\mathrm{var}\left(\hat{Y}_{CL2}\right) = N\left(\frac{1-f_1}{f_1}\right)S^2_{yCT} + \sum_{d=1}^{N} M_d\left(\frac{1-f_{2,d}}{f_1 f_{2,d}}\right)S^2_{yd} \quad . \tag{4.22*}$$

with:

$$f_1 = \frac{n}{N} \qquad \wedge \qquad f_{2,d} = \frac{m_d}{M_d} \quad .$$

The first fraction is referred to as the *first-stage sampling fraction*; the second fraction is referred to as the *second-stage sampling fraction*.

The variance of the cluster estimator of the population mean is obtained by dividing the variance of the cluster estimator of the population total by M^2:

$$\begin{aligned}\mathrm{var}\left(\hat{\bar{Y}}_{CL2}\right) &= \frac{1}{M^2}\mathrm{var}\left(\hat{Y}_{CL2}\right) \\ &= \left(\frac{1}{M}\right)\left(\frac{1}{\bar{M}}\right)\left(\frac{1-f_1}{f_1}\right)S^2_{yCT} + \left(\frac{1}{M}\right)\sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{1-f_{2,d}}{f_1 f_{2,d}}\right)S^2_{yd} \quad .\end{aligned}$$

The effect of sampling in the second stage on the variance above is clear from a comparison of (4.22) with (4.9). The variances of the estimators comprise two terms. The first term (on the right-hand side of the '=' sign) is attributable to the differences between the totals of the primary sampling units. The second term (on the right-hand side of the '=' sign) is attributable to the variation within the primary sampling units. This term played no part in the above sections because the clusters were observed in full ($f_{2,d} \equiv 1$). In many situations the second term will be negligible relative to the first term. With a strongly varying M_d, an objection to the SRSWOR estimator in the first stage is that the cluster totals usually differ very substantially. This effect can be countered by selecting the primary sampling units in proportion to their size, or by employing the ratio estimator.

The calculated variance (4.22) can be estimated without bias by

$$\begin{aligned}\hat{\mathrm{var}}\left(\hat{\bar{Y}}_{CL2}\right) &= N\left(\frac{1-f_1}{f_1}\right)s^2_{\hat{y}CT} + \sum_{d=1}^{n} M_d\left(\frac{1-f_{2,d}}{f_1 f_{2,d}}\right)s^2_{yd} \\ s^2_{\hat{y}CT} &= \frac{1}{n-1}\sum_{d=1}^{n}\left(\hat{y}_d - \bar{\hat{y}}_{CT}\right)^2 \quad .\end{aligned} \tag{4.23*}$$

Sampling design evaluation

In two-stage sampling on average $(n \times \bar{m})$ secondary sampling units are interviewed; the expected mean of the sample sizes in the n chosen primary units is $\bar{m} = \sum_{d=1}^{N}(m_d / N)$. To evaluate this cluster sample we compare the variance of the cluster estimators with the variance of the corresponding estimators based on SRSWOR of $(n \times \bar{m})$ out of M elements. To this end, the design effect of the cluster estimator of the population total is calculated – see Definition (4.11).

The variance of the SRSWOR estimator of the population total based on SRSWOR of $(n \times \overline{m})$ elements is calculated as follows:

$$\mathrm{var}\left(\hat{Y}_{SRSWOR}\right) = M\left(\frac{1-f_1 f_2}{f_1 f_2}\right)S_y^2 \quad . \tag{4.24}$$

with:

$$f_1 = \frac{n}{N} \qquad \wedge \qquad f_2 = \frac{\overline{m}}{\overline{M}} \quad .$$

Based on expressions (4.22) and (4.24), the design effect of the cluster estimator of the population total can be calculated as follows:

$$\begin{aligned} DEFF\left[\hat{Y}_{CL2}\right] = & \left(\frac{1}{\overline{M}}\right)\left(\frac{f_2 - f_1 f_2}{1-f_1 f_2}\right)\left(\frac{S_{yCT}^2}{S_y^2}\right) + \\ & + \left(\frac{f_2}{1-f_1 f_2}\right)\sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{1-f_{2,d}}{f_{2,d}}\right)\left(\frac{S_{yd}^2}{S_y^2}\right) \quad . \end{aligned} \tag{4.25*}$$

This result can be interpreted as a more general version of expression (4.12). This conclusion is supported by the fact that under the condition of full observation of the selected primary sampling units (i.e. $f_{2,d} = f_2 = 1$), the two formulas are in agreement.

With the objective of elaborating formula (4.25) further, it is assumed that the primary sampling units have the same size: i.e. $M_d = \overline{M}$ for all *d*. Furthermore, it is assumed that the same number of secondary sampling units is selected from each primary sampling unit, i.e. $m_d = \overline{m}$ for $d = 1, \cdots, N$. It is simple to derive under these conditions that:

$$M_d = \overline{M} = \frac{M}{N} \qquad \wedge \qquad f_{2,d} = \frac{m_d}{M_d} = \frac{\overline{m}}{\overline{M}} = f_2 \quad . \tag{4.26}$$

Substituting (4.26) into (4.25) yields:

$$DEFF\left[\hat{Y}_{CL2}\right] = \left(\frac{1}{\overline{M}}\right)\left(\frac{f_2(1-f_1)}{1-f_1 f_2}\right)\left(\frac{S_{yCT}^2}{S_y^2}\right) + \left(\frac{1-f_2}{1-f_1 f_2}\right)\left(\frac{S_{intra}^2}{S_y^2}\right) \quad . \tag{4.27}$$

Formula (4.13) for the adjusted population variance is still valid, since the properties of the population do not vary with the sampling design. We can therefore write (see also (4.13)):

$$\left(\frac{S_{yCT}^2}{S_y^2}\right) = -M\left(\frac{\overline{M}-1}{N-1}\right)\left(\frac{S_{intra}^2}{S_y^2}\right) + \frac{\overline{M}(M-1)}{N-1} \quad . \tag{4.28}$$

This result yields the desired connection between the adjusted variance of cluster totals and the adjusted population variance. For two-stage sampling the design effect of the cluster estimator of the population total can also be determined after substituting (4.28) in (4.27), as follows:

$$DEFF\left[\hat{Y}_{CL2}\right]=\left(\frac{1}{1-f_1 f_2}\right)\left\{1-\left(\frac{M-1}{N-1}\right)f_2+\left(\frac{M-N}{N-1}\right)f_1 f_2\right\}\left(\frac{S^2_{intra}}{S^2_y}\right)+ \\ +\left(\frac{f_2-f_1 f_2}{1-f_1 f_2}\right)\left(\frac{M-1}{N-1}\right) \quad . \tag{4.29}$$

As expected, this expression corresponds with formula (4.14) under the condition of fully observing the selected primary sampling units (i.e. $f_2 = 1$)

4.3.2.3 Determination of sample size

As with a simple random sample, the total sample size can be determined from the precision that is required for an estimator. Once the total sample size has been determined, the question arises as to how to distribute it among the primary and secondary sampling units. In other words, how many secondary sampling units should be selected from the selected primary sample units, and how many primary sample units should be in the sample. These questions usually involve balancing costs and precision. Selecting many primary sampling units with few secondary sampling units per primary unit will generally lead to small variances. However, the potential costs saved by observing a large number of elements within one primary sampling unit are then almost totally lost. While it is less expensive to observe many secondary sampling units within a limited number of primary sampling units, it will lead to less precise estimators. Answering the question about how to distribute the sample requires some knowledge of the data collection and other costs in both stages, and of the homogeneity of the primary sample units.

With perfectly homogeneous primary sampling units, i.e. $S^2_{intra} = 0$, one might as well select just one element per primary sampling unit; observing more elements would lead only to higher costs with no gain in precision. In all other cases the distribution of the sample will depend on the costs of selecting the units in the two stages. A simple cost function is:

$$C = nc_1 + n\overline{m}c_2$$

where parameters c_1 and c_2 correspond, for example, to interview and/or travel expenses in the first and second sampling stage. Given this cost function and fixed total cost C, the variance of the cluster estimator of the population total is minimized by selecting:

$$n = \frac{C}{c_1 + c_2 m}$$

primary sampling units in the sample and:

$$m = \sqrt{\frac{\overline{M}S_{intra}^2}{\left(\overline{M}S_{inter}^2 - S_{intra}^2\right)}\frac{c_1}{c_2}} = \sqrt{\frac{c_1/c_2}{S_{inter}^2/S_{intra}^2 - \left(1/\overline{M}\right)}}$$

secondary sampling units per primary sampling unit. To determine these quantities explicitly requires some idea of the ratio S_{inter}^2/S_{intra}^2, or in other words of the homogeneity of the primary sampling units. In practice one will often proceed the other way around. This usually means trying different values of *n* and *m*, calculating the associated variance of an estimator, and assessing whether this might be acceptable.

4.3.3 Systematic sampling

Systematic sampling is frequently put forward as a good alternative to SRSWOR. In a systematic sample of *n* elements out of a population of *N* elements, first a step length $L = N/n$ is determined. We assume for convenience that the elements are numbered from 1 through *N*. A starting number *R* is then chosen at random from the interval (0, *L*). The first element selected in the sample is the one with number k_1 for which $k_1 - 1 < R \le k_1$. The second element selected in the sample has number k_2, for which $k_2 - 1 < R + L \le k_2$, and so on. The final n^{th} element in the sample has number k_n, for which $k_n - 1 < R + (n-1)L \le k_n$. In other words, the elements with numbers $k_1, \cdots, k_n$ form the sample, or the numbers in the sample are equal to the (rounded) numbers $R, R+L, R+2L, \cdots, R+(n-1)L$. It is clear that the randomly selected starting number R determines the entire sample, and that every element from the population can occur in only a single sample.

In theory, and assuming that *L* is a whole number, different samples are possible for a fixed population sequence *L*, so that the sampling probability for systematic sampling is:

$$P(s) = \frac{1}{L} = \frac{n}{N} \quad .$$

Since each element can occur in only one sample, the inclusion probability of population element *k* is equal to the sampling probability, i.e.:

$$\pi_k = \frac{1}{L} = \frac{n}{N} \quad .$$

With this form of systematic sampling the number of possible samples is less than for SRSWOR with the same sample size, as not every combination of *n* elements is possible. However (see the result above), the first-order inclusion probabilities are the same as with SRSWOR. Although the systematic sample described here in fact constitutes a cluster sample, the variance of the estimator is often approximated in practice by that of the SRSWOR estimator. The assumption is that the (fixed) sequence of the population elements does not lead to a systematic difference with the SRSWOR estimator.

Estimators, unbiasedness and variance calculations

Systematic sampling thus amounts to selecting one cluster of size *n* out of *L* clusters. The population parameters are estimated from the sample statistics of this single selected cluster. The respective cluster estimators for the population total and population mean are:

$$\hat{Y}_{sys} = N\,\bar{y} \qquad \wedge \qquad \hat{\bar{Y}}_{sys} = \bar{y} \quad .$$

The bias of the estimators follows from the properties of cluster sampling. The variance for the cluster estimator of the population mean for a sample of 1 cluster is according to formula (4.10):

$$\mathrm{var}\left(\hat{\bar{Y}}_{sys}\right) = \left(1 - \frac{1}{L}\right)\frac{S_{yCT}^2}{n^2} \tag{4.30}$$

The variance is expressed in the notation for systematic sampling. Note that *N* and *n* refer here to elements rather than clusters. The size of a cluster (systematic sampling) is now *n*; the total number of elements is *N*; the number of clusters is *L*. The variance of the cluster estimator for the population total is obtained by multiplying the variance of the cluster estimator for the population mean by N^2.

Sampling design evaluation

Finally, the design effect of the cluster estimator for the population can be calculated using formula (4.19) as:

$$DEFF\left[\hat{\bar{Y}}_{sys}\right] = \frac{N-1}{N-n}\left\{1 + (n-1)\rho_c\right\} \quad .$$

Systematic sampling is therefore better than simple random sampling if and only if the intraclass correlation coefficient satisfies:

$$-\frac{1}{n-1} \le \rho_c \le -\frac{1}{N-1} \quad .$$

When the variance within the possible systematic samples is greater than the population variance, the cluster means will differ little from each other, and systematic sampling will be more precise than simple random sampling. Otherwise, if the variance within the systematic samples is small relative to the population variance $(0 < \rho_c)$ then all elements in the sample will give the same kind of information and the variance of the estimator for systematic sampling will be larger than for simple random sampling.

However, a significant drawback of a systematic sampling design is the lack of unbiased estimators for the different variances. Since only one cluster is selected, the sample variance of the cluster totals s_{yCT}^2 is undefined. Nonetheless, the variance can be approximated if we know something about the structure of the population. The following three situations can be identified:

1. random sequences of the sampling frame;
2. positive autocorrelation;
3. periodic fluctuations.

Note to 1 (random sequences of the sampling frame)
In many situations it can be assumed that the values of the target variables do not correspond with the sequence in the sampling frame. An example would be a frame in which people are in alphabetical order. In this case systematic sampling will correspond closely to simple random sampling $[\rho_c \approx -1/(N-1)]$ and both samples will produce the same sort of results. We can therefore use the variance formula for simple random sampling to estimate $\text{var}(\hat{\bar{Y}}_{sys})$.

Note to 2 (positive autocorrelation)
Sometimes a frame will have an increasing or decreasing sequence effect. For example, companies may appear in order of decreasing size in a list. In these cases positive autocorrelation is said to exist: successive elements resemble each other more closely than elements that are further apart. Systematic sampling would ensure more spread in the sample elements and will therefore be more precise than simple random sampling $[-1/(n-1) \le \rho_c < -1/(N-1)]$. The variance formula for simple random sampling would give an overestimate in this case.

Note to 3 (periodic fluctuations)
Problems may occur if the sampling frame exhibits a periodic variation with respect to a target variable. Systematic sampling will then be considerably less precise than simple random sampling, in particular when the step length is the same as, or a multiple of, the period. The variance of the estimator will then be extremely large, which is not apparent when using the simple random variance formula to estimate the variance.

Suppose the population elements are arranged so that the target variable shows the following pattern: $1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,\cdots\cdots$. With a step length of 5 the elements that appear in the sample will all be the same and the variance formula for simple sampling will estimate the variance as $\hat{\text{var}}(\hat{\bar{Y}}) = 0$. However, the actual value of $\text{var}(\hat{\bar{Y}}_{sys})$ for this population is $\sigma_y^2 = 2$; note that $\rho_c = 1$. In other words, a systematic sample in this case is as efficient as an SRSWOR sample of size $n = 1$.

4.4 Quality indicators

An important quality indicator for (multistage) cluster sampling is the cost reduction that results from the clustering of the units to be observed. The price to be paid for this clustering in the form of increased variance relative to simple random sampling without replacement is another important criterion.

4.5 Appendix

Proof of (4.4)

From the definition of population variance we can deduce:

$$\sigma_y^2 = \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(Y_{dk}-\bar{\bar{Y}}\right)^2$$
$$= \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(Y_{dk}-\bar{Y}_d\right)^2 + \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2$$

because the double product is zero – see also the proof of (3.4). Further elaboration of this expression gives the following result:

$$\sigma_y^2 = \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\sigma_{yd}^2 + \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2$$
$$= \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2 \quad .$$

QED.

Proof of (4.5)

It can be deduced from definitions (2.1) and (4.1) that:

$$S_y^2 = \left(\frac{M}{M-1}\right)\sigma_y^2 \qquad \wedge \qquad S_{yd}^2 = \left(\frac{M_d}{M_d-1}\right)\sigma_{yd}^2 \quad .$$

Multiplying equation (4.4) by the factor $(M/(M-1))$ gives the following relationship:

$$\left(\frac{M}{M-1}\right)\sigma_y^2 = \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{M}{M-1}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M}{M-1}\right)\left(\frac{M_d}{M}\right)\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2 \quad .$$

This equation can be rewritten as:

$$S_y^2 = \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2$$
$$= \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\frac{M_d-1}{M_d-1}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2$$
$$= \sum_{d=1}^{N}\left(\frac{M_d-1}{M-1}\right)S_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2 \quad .$$

QED.

Proof of (4.13)

The proof of (4.13) starts from formula (4.5) for adjusted population variance. The following relationships are also significant:

$$\bar{\bar{Y}} = \frac{1}{\bar{\bar{M}}} \bar{Y}_{CT} \qquad \wedge \qquad \bar{Y}_d = \frac{1}{\bar{\bar{M}}} Y_d \quad .$$

The expression for adjusted population variance can be rewritten as:

$$\begin{aligned} S_y^2 &= \sum_{d=1}^{N} \left(\frac{M_d - 1}{M - 1} \right) S_{yd}^2 + \sum_{d=1}^{N} \left(\frac{M_d}{M - 1} \right) \left(\bar{Y}_d - \bar{\bar{Y}} \right)^2 \\ &= \left(\frac{\bar{\bar{M}} - 1}{M - 1} \right) \sum_{d=1}^{N} S_{yd}^2 + \left(\frac{\bar{\bar{M}}}{M - 1} \right) \sum_{d=1}^{N} \left(\frac{1}{\bar{\bar{M}}} \right)^2 \left(Y_d - \bar{Y}_{CT} \right)^2 \\ &= \left(\frac{\bar{\bar{M}} - 1}{M - 1} \right) \left(\frac{N}{N} \right) \sum_{d=1}^{N} S_{yd}^2 + \left(\frac{1}{\bar{\bar{M}} (M - 1)} \right) \left(\frac{N - 1}{N - 1} \right) \sum_{d=1}^{N} \left(Y_d - \bar{Y}_{CT} \right)^2 \\ &= \left(\frac{M - N}{M - 1} \right) S_{intra}^2 + \left(\frac{1}{\bar{\bar{M}}} \right) \left(\frac{N - 1}{M - 1} \right) S_{yCT}^2 \quad . \end{aligned}$$

QED.

Proof of (4.14)

Formula (4.13) for the population variance can be rewritten as follows:

$$\left(\frac{1}{\bar{\bar{M}}} \right) \left(\frac{N - 1}{M - 1} \right) S_{yCT}^2 = S_y^2 - \left(\frac{M - N}{M - 1} \right) S_{intra}^2$$

from which:

$$\left(\frac{1}{\bar{\bar{M}}} \right) S_{yCT}^2 = \left(\frac{M - 1}{N - 1} \right) S_y^2 - \left(\frac{M - N}{N - 1} \right) S_{intra}^2 \quad .$$

Substituting this relationship into formula (4.12) gives:

$$\begin{aligned} DEFF\left[\hat{\bar{Y}}_{CL} \right] &= \frac{\frac{1}{\bar{\bar{M}}} S_{yCT}^2}{S_y^2} \\ &= \left(\frac{M - 1}{N - 1} \right) - \left(\frac{M - N}{N - 1} \right) \frac{S_{intra}^2}{S_y^2} \quad . \end{aligned}$$

QED.

Proof of (4.15)

The design effect of the cluster estimator of the population total is minimal when the intracluster variance is maximal. According to (4.5), the intracluster variance obtains its maximum value if and only if the intercluster variance is zero. The value of S_{intra}^2 can therefore be determined as follows:

$$S_{intra}^2 \Big|_{S_{inter}^2 = 0} = \left(\frac{M - 1}{M - N} \right) S_y^2 \quad .$$

The design effect for this value of intracluster variance is:

$$DEFF\left[\hat{Y}_{CL}\right]\Big|_{S^2_{inter}=0} = -\left(\frac{M-N}{N-1}\right)\left(\frac{M-1}{M-N}\right)+\left(\frac{M-1}{N-1}\right) = 0 \quad .$$

The design effect of the cluster estimator of the population total is maximal if the intracluster variance is minimal. The minimum value that the intracluster variance can assume is 0. The maximum value that the design effect can assume is then simply:

$$DEFF\left[\hat{Y}_{CL}\right]\Big|_{S^2_{intra}=0} = \frac{M-1}{N-1} \quad .$$

QED.

Proof of (4.17)

Algebraic manipulation shows that:

$$\begin{aligned}\sum_{k=1}^{\bar{M}}\sum_{l=1}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\left(Y_{dl}-\bar{Y}\right) &= \sum_{k=1}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\sum_{l=1}^{\bar{M}}\left(Y_{dl}-\bar{Y}\right)\\ &= \left(Y_d-\bar{Y}_{CT}\right)^2 \quad .\end{aligned}$$

It can then be deduced from this result and definition (4.2) that:

$$\begin{aligned}\sum_{d=1}^{N}\sum_{k=1}^{\bar{M}}\sum_{l=1}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\left(Y_{dl}-\bar{Y}\right) &= \sum_{d=1}^{N}\left(Y_d-\bar{Y}_{CT}\right)^2\\ &= \left(N-1\right)S^2_{yCT} \quad .\end{aligned}$$

There is an alternative approach to calculating the above threefold sum, as follows:

$$\begin{aligned}\sum_{d=1}^{N}\sum_{k=1}^{\bar{M}}\sum_{l=1}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\left(Y_{dl}-\bar{Y}\right) &= \sum_{d=1}^{N}\sum_{k=1}^{\bar{M}}\sum_{l\neq k}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\left(Y_{dl}-\bar{Y}\right)\\ &\quad+\sum_{d=1}^{N}\sum_{k=1}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)^2\\ &= \sum_{d=1}^{N}\sum_{k=1}^{\bar{M}}\sum_{l\neq k}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\left(Y_{dl}-\bar{Y}\right)+\left(M-1\right)S^2_y \quad .\end{aligned}$$

Combining these two results gives the following expression:

$$\sum_{d=1}^{N}\sum_{k=1}^{\bar{M}}\sum_{l\neq k}^{\bar{M}}\left(Y_{dk}-\bar{Y}\right)\left(Y_{dl}-\bar{Y}\right) = \left(N-1\right)S^2_{yCT}-\left(M-1\right)S^2_y \quad .$$

It can be concluded from definition (4.16) for the intracluster correlation coefficient and this result that:

$$\begin{aligned}\rho_c &= \frac{\left(N-1\right)S^2_{yCT}-\left(M-1\right)S^2_y}{\left(M-1\right)\left(\bar{M}-1\right)S^2_y}\\ &= \left(\frac{N-1}{\left(M-1\right)\left(\bar{M}-1\right)}\right)\frac{S^2_{yCT}}{S^2_y}-\left(\frac{1}{\bar{M}-1}\right) \quad .\end{aligned}$$

QED .

Proof of (4.18)

The intercluster variance is maximal if and only if the intracluster variance is 0. It follows from equation (4.13) that under this condition the adjusted variance of cluster totals is:

$$S_{yCT}^2\Big|_{S_{intra}^2=0} = \left(\frac{\overline{M}(M-1)}{N-1}\right)S_y^2 \quad .$$

Substituting this result into (4.17) gives:

$$\begin{aligned}\rho_c\Big|_{S_{intra}^2=0} &= \left(\frac{N-1}{(M-1)(\overline{M}-1)}\right)\left(\frac{\overline{M}(M-1)}{N-1}\right) - \frac{1}{\overline{M}-1} \\ &= \frac{\overline{M}}{\overline{M}-1} - \frac{1}{\overline{M}-1} = 1 \quad .\end{aligned}$$

The minimum value of the intercluster variance is 0. This means that the minimum value of the adjusted variance of cluster totals is likewise 0. From the above:

$$\rho_c\Big|_{S_{inter}^2=0} = -\frac{1}{\overline{M}-1} \quad .$$

QED.

Proof of (4.19)

It can be deduced from formula (4.17) that:

$$\frac{S_{yCT}^2}{S_y^2} = \left(\frac{(M-1)(\overline{M}-1)}{N-1}\right)\rho_c + \frac{M-1}{N-1} \quad .$$

Substituting this result in expression (4.12) for the design effect of the cluster estimator of the population total gives:

$$\begin{aligned}DEFF\left[\hat{Y}_{CL}\right] &= \frac{1}{\overline{M}}\left(\frac{(M-1)(\overline{M}-1)}{N-1}\right)\rho_c + \frac{1}{\overline{M}}\times\frac{M-1}{N-1} \\ &= \left(\frac{(M-1)(\overline{M}-1)}{M-\overline{M}}\right)\rho_c + \frac{M-1}{M-\overline{M}} \quad .\end{aligned}$$

QED.

Proof of (4.22)

The proofs of (4.22) and (4.23) are taken from Knottnerus (2011b). Denoting the variance of $\hat{Y}_d$ by σ_d^2 $(d=1,\cdots,N)$, it is seen from Chapter 2 that for SRSWOR sampling in the second stage,

$$\sigma_d^2 = \mathrm{var}(\hat{Y}_d) = M_d\left(\frac{1-f_{2,d}}{f_{2,d}}\right)S_{yd}^2 \quad .$$

Recall that an SRS sample of *n* clusters in the first stage can simply be obtained by taking the first *n* clusters after the *N* clusters are put in a random order. Denote the *n* observed clusters in the first-stage sample *s* by $d_1,...,d_{n.}$. Recall that d_k (k =1,…,n) can be regarded as a random variable with $P(d_k = d) = 1/N$ ($d = 1,...,N$). In order to avoid double subscripts, the quantities $\hat{Y}_{d_k}$ and Y_{d_k} of the clusters selected in *s* are in the remainder briefly denoted by the lower case letters $\hat{y}_k$ and y_k, respectively ($k = 1,...,n$). Define the corresponding sampling errors g_k by $g_k = \hat{y}_k - y_k$. Define σ^2_{gCT} by $\sigma^2_{gCT} = (\Sigma_{d=1}^{N}\sigma_d^2)/N$. Then using $E(g_k) = 0$, it is seen that

$$\begin{aligned}\text{var}(g_k) &= E\left(g_k^2\right) = E\left\{E\left(g_k^2 \middle| d_k\right)\right\} \\ &= \frac{1}{N}\sum_{d=1}^{N} E\left(g_k^2 \middle| d_k = d\right) = \frac{1}{N}\sum_{d=1}^{N}\sigma_d^2 = \sigma^2_{gCT} \\ \text{cov}(g_k, y_k) &= E\{E(g_k y_k | d_k)\} = \frac{1}{N}\sum_{d=1}^{N} Y_d E(g_k | d_k = d) = 0 \quad .\end{aligned}$$

In the same way it can be shown that g_k is uncorrelated with y_l and g_l ($l \neq k$). Denote the sample mean of $g_1,\cdots,g_n$ by $\bar{g}_{CT}$. Now using $\bar{\hat{y}}_{CT} = \bar{y}_{CT} + \bar{g}_{CT}$, it is seen that

$$\begin{aligned}\text{var}(N\bar{\hat{y}}_{CT}) &= N^2\,\text{var}\left(\bar{y}_{CT} + \bar{g}_{CT}\right) = N^2\,\text{var}\left(\bar{y}_{CT}\right) + N^2\,\text{var}\left(\bar{g}_{CT}\right) \\ &= N\left(\frac{1-f_1}{f_1}\right)S^2_{yCT} + \frac{N^2}{n^2}\sum_{k=1}^{n}\text{var}(g_k) \\ &= N\left(\frac{1-f_1}{f_1}\right)S^2_{yCT} + \frac{N\sigma^2_{gCT}}{f_1}.\end{aligned}$$

QED.

Proof of (4.23)

Similar to $s^2_{\hat{y}CT}$, denote the sample variances of the y_k and g_k by s^2_{yCT} and s^2_{gCT}, and their sample covariance by s_{ygCT}. The key feature in two-stage sampling is that $g_1,\cdots,g_n$ are mutually uncorrelated random variables with zero expectation and variance σ^2_{gCT}. Therefore, $E(s^2_{gCT}) = \sigma^2_{gCT}$; the proof runs along the same lines as in the well-known (hypothetical) case that $g_1,\cdots,g_n$ are independent and identically distributed random variables with zero expectation and variance σ^2_{gCT}; see also Särndal *et al.* (1992, p. 52). Furthermore, $E(s^2_{yCT}) = S^2_{yCT}$; see Chapter 2. As we have seen, g_k (k=1,…,n) is uncorrelated with $y_1,\cdots,y_n$. Hence,

$$E(s^2_{\hat{y}CT}) = E(s^2_{y+g,CT}) = E(s^2_{yCT} + s^2_{gCT} + 2s_{ygCT}) = S^2_{yCT} + \sigma^2_{gCT} \quad . \qquad (4.31)$$

In addition, define $\hat{\sigma}^2_d$ by

$$\hat{s}_d^2 = M_d\left(\frac{1- f_{2,d}}{f_{2,d}}\right)s_{yd}^2 .$$

It follows from Chapter 2 that $\hat{s}_d^2$ is an unbiased estimator of s_d^2. Furthermore, in analogy with $E(\bar{\hat{y}}_{CT}) = E(\Sigma_{k\hat{I}s}\hat{y}_k / n) = Y / N$,

$$E\left(\frac{1}{n}\sum_{k=1}^{n}\hat{s}_k^2\right) = \frac{1}{N}\sum_{d=1}^{N}s_d^2 = s_{gCT}^2 \quad . \tag{4.32}$$

Using (4.31) and (4.32), it is seen that $\hat{\text{var}}\left(\hat{Y}_{CL2}\right)$ in (4.23) is unbiased. QED.

Proof of (4.25)

The design effect of the cluster estimator of the population total is defined as the ratio between variances (4.22) and (4.24). It can therefore simply be deduced that:

$$\begin{aligned} DEFF\left[\hat{Y}_{CL2}\right] &= N\left(\frac{1- f_1}{f_1}\right)\left(\frac{1}{M}\right)\left(\frac{f_1 f_2}{1- f_1 f_2}\right)\left(\frac{S_{yCT}^2}{S_y^2}\right) \\ &\quad + \left(\frac{1}{M}\right)\left(\frac{f_1 f_2}{1- f_1 f_2}\right)\sum_{d=1}^{N}M_d\left(\frac{1- f_{2,d}}{f_1 f_{2,d}}\right)\left(\frac{S_{yd}^2}{S_y^2}\right) \\ &= \left(\frac{1}{\overline{M}}\right)\left(\frac{f_2(1- f_1)}{1- f_1 f_2}\right)\left(\frac{S_{yCT}^2}{S_y^2}\right) + \\ &\quad + \left(\frac{f_2}{1- f_1 f_2}\right)\sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{1- f_{2,d}}{f_{2,d}}\right)\left(\frac{S_{yd}^2}{S_y^2}\right) \quad . \end{aligned}$$

QED.

5. Samples with equal and unequal inclusion probabilities

5.1 Short description

5.1.1 The advantage

When drawing a sample not all the elements in a population need to be given the same inclusion probability, as has been done so far. Especially when the target variable Y_k in the population is roughly proportional to a particular auxiliary variable X_k $(k = 1, \cdots, N)$, sampling with unequal inclusion probabilities is preferable. The latter variable must then be known for all elements in the population. In practice the auxiliary variable X_k often stands for the size or the relevance of element k in the entire population.

There are many ways of sampling without replacement with unequal inclusion probabilities. This chapter focuses on systematic 'probability proportional to size' (PPS) samples, where the elements are in random order and the inclusion probabilities are proportional to a given auxiliary variable X_k – see the example in section 5.4. The great advantage of unequal inclusion probabilities is that the variances of the estimators can be reduced substantially, thereby also reducing the corresponding margins of uncertainty. Another advantage is that stratification in size classes is no longer necessary, so avoiding for a given sample size the risk of having extremely few observations in some strata.

5.1.2 The sampling scheme for systematic PPS sampling

After randomly ordering the population elements, every element k is assigned an interval with length X_k (on the real line) $k = 1, \cdots, N$. The first element is assigned the interval $(0, X_1]$ by definition, the second element the interval $(X_1, X_1 + X_2]$, and so on. As a result, the interval $(0, X]$, where X stands for the population total of the auxiliary variable, is divided into N consecutive intervals. Subsequently, $(0, X]$ is divided into n equal pieces of length L $(L = X / n)$; L is also referred to as the *step length*.

The sampling scheme is now as follows. Randomly select a number r between 0 and L. This number falls in exactly one of the intervals that comprise $(0, X]$. The population element that corresponds with this interval is the first element selected in the sample. The population element the assigned interval of which contains the number $(r + L)$ becomes the 2^{nd} element in the sample, and so on. The final element in the sample is the one where the assigned interval comprises the number $[r + (n-1)L]$. This sampling scheme can be represented in formulas as follows:

$$S_k = X_1 + \cdots + X_k \quad (k = 1, \cdots, N)$$
$$S_0 = 0$$
$$J(x) = k \quad \text{for} \quad S_{k-1} < x \leq S_k \qquad (0 < x \leq X) \qquad .$$

The systematic PPS sample then consists of the *n* elements with ranks

$$J(r), J(r+L), \cdots\cdots, J(r+(n-1)L)$$

To avoid multiple inclusion of an element in the sample, we have tacitly assumed that elements for which $L < X_k$ have already been removed from the population and included in a separate stratum, which is observed in full. This stratum is said to contain the *self-selecting* elements. Having removed the self-selecting elements, *X* must be recalculated, just as the new *L* and the new sample size *n*. If necessary, this process must be repeated until no new self-selecting elements appear.

The elements in the sampling scheme described here are first randomly ordered. However, sometimes there may be a good reason for maintaining some fixed sequence of elements, e.g. a sequence of increasing X_k. If Y_k / X_k and X_k are closely related, this kind of sequence can produce estimators with a particularly small variance. Nonetheless, identifying a good variance estimator can be a problem. Standard variance estimators do not exist, but under certain modelling assumptions reasonably reliable variance estimators can be derived.

Note that it is unnecessary to randomly order the elements when there is no substantial relation between Y_k / X_k and X_k, or, more precisely, when there is no relation between Y_k / X_k and k, $k = 1, \cdots, N$. In these situations random sequencing of the elements will have no systematic effect on the outcomes.

Finally, we mention the special case of all elements having the same inclusion probability, which is to say that all elements are assigned an interval of the same length, while the X_k differ. With elements that are randomly arranged, this approach is equivalent to SRSWOR, in which case the formulas given in Chapter 2 apply. If, conversely, the elements are in a fixed sequence, it is again much more difficult to estimate the variance of the estimator of the population total. However, it is possible to demonstrate that the variance decreases if the elements are arranged such that all possible samples resemble each other (and therefore also the population) as closely as possible in terms of the distribution of the X_k. This situation can be approached in practice by ordering the elements according to size of the X_k. A condition for reducing the variance of the estimator of the population total is that a sufficiently high correlation exists between Y_k and X_k.

Another option is to arrange the elements as much as possible according to region, so that every sample has sufficient regional spread (heterogeneity). In other words, the (fixed) sequence must prevent the samples thus selected becoming homogeneous. As with unequal inclusion probabilities, a remaining problem is tracing the variance – see Knottnerus (2003, pp 190-1 and 237-43).

5.2 Applicability

Statistics Netherlands applies the PPS sampling design mainly in social statistics, such as in the two-stage sampling of individuals. In the first stage, municipalities are selected with an inclusion probability proportional to the number of residents. In the second step individuals are selected by means of SRSWOR from each municipality selected in the first stage. The second-stage sampling fraction in turn is inversely proportional to the number of residents of the corresponding municipality, so that on balance all individuals have the same inclusion probability.

Following the example of other countries, Statistics Netherlands will shortly also be using the PPS sampling design for estimating the monthly producer price index (PPI) for industry. In a specific example using data from 2005, the theory for the PPI is worked out in Section 5.4.

5.3 Detailed description

5.3.1 Assumptions and notation

The starting point is a population of N elements from which a sample of size *n* is to be selected *without* replacement, unless stated otherwise. It is assumed in this chapter that the sample size *n* is fixed. As in the previous chapters, we are interested in estimating the population mean $\bar{Y}$ and the population total Y. The *n* sample observations of the target variable are denoted with $y_1,\cdots,y_n$.

The binary selection indicator a_k – see (2.7) – is a random variable that states whether or not element k from the population was ultimately included in the sample.

The probability of population element k being selected in the sample, i.e. the inclusion probability, is denoted with the symbol π_k. The formal definition of the inclusion probability is:

$$\pi_k = P(a_k = 1) \qquad k = 1,\cdots,N \tag{5.1}$$

with:

$$a_k = \begin{cases} 1 & \text{if element } k \textit{ is} \text{ included in the sample} \\ 0 & \text{if element } k \textit{ is not} \text{ included in the sample.} \end{cases}$$

The second-order inclusion probability for the two population elements k and l, denoted as π_{kl}, is the probability of population elements k and l occurring together in the sample. Expressed more formally this is:

$$\pi_{kl} = \begin{cases} P(a_k = 1 \wedge a_l = 1) & k \neq l = 1,\cdots,N \\ \pi_k & k = l = 1,\cdots,N \end{cases} . \tag{5.2}$$

5.3.2 Estimators for the population total and the population mean

Given the (unequal) inclusion probabilities π_k for all N elements of the population, the population total Y can be estimated with the Horvitz-Thompson (HT) estimator. The HT estimator, denoted as $\hat{Y}_{HT}$, is defined as:

$$\hat{Y}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} \quad . \tag{5.3}$$

A necessary condition for applying this estimator is that the inclusion probability π_k is greater than 0 for $k = 1, \cdots, N$. Definition (5.3) leads immediately to the definition of the Horvitz-Thompson estimator of the population mean, denoted with $\hat{\bar{Y}}_{HT}$, which is:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\hat{Y}_{HT} = \frac{1}{N}\sum_{k=1}^{n} \frac{y_k}{\pi_k} \quad ; \tag{5.4}$$

See Horvitz and Thompson (1952). The HT estimator is very general and can be applied to every sample *without* replacement.

The unbiasedness of the estimators

To demonstrate the unbiasedness of HT estimators (5.3) and (5.4), we change to a different notation for (5.3) and (5.4). Using the selection indicator, the formulas for the two HT estimators can be rewritten as:

$$\begin{aligned} \hat{Y}_{HT} &= \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{N} \left(\frac{a_k}{\pi_k} \right) Y_k \\ \hat{\bar{Y}}_{HT} &= \frac{1}{N}\sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{1}{N}\sum_{k=1}^{N} \left(\frac{a_k}{\pi_k} \right) Y_k \quad . \end{aligned} \tag{5.5}$$

Because $E(a_k) = \pi_k$, the estimators in (5.5) are unbiased, i.e.:

$$E\left(\hat{Y}_{HT}\right) = Y \qquad \wedge \qquad E\left(\hat{\bar{Y}}_{HT}\right) = \bar{Y} \quad . \tag{5.6}$$

Variance calculations

The calculation of the variance of the HT estimators (5.3) and (5.4), uses the following results for the (co)variances of the selection indicator a_k:

$$\begin{aligned} \operatorname{var}(a_k) &= \pi_k(1-\pi_k) \\ \operatorname{cov}(a_k, a_l) &= \pi_{kl} - \pi_k \pi_l \end{aligned} \qquad 1 \le k, l \le N \quad . \tag{5.7*}$$

This result is an extension of property (2.9) of the selection indicator. It is now possible to derive the following formula for the variance of the HT estimator:

$$\text{var}\left(\hat{Y}_{HT}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right)Y_kY_l \quad . \tag{5.8*}$$

In other words, the variance of the HT estimator of the population total is a linear combination of all possible products Y_kY_l with $k,l=1,\cdots,N$.

In order to determine variance (5.8) on the basis of observations, the following estimator is introduced, also referred to as the *HT estimator* of the variance:

$$\hat{\text{var}}_{HT}\left(\hat{Y}_{HT}\right) = \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}\pi_k\pi_l}\,y_ky_l \quad . \tag{5.9}$$

If n is fixed, there is the following alternative for this estimator (Sen (1953) and Yates and Grundy (1953)):

$$\hat{\text{var}}_{SYG}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2 \quad . \tag{5.10}$$

It is possible to demonstrate that the expected value of estimator (5.9) is equal to the variance calculated in (5.8) – see Appendix (Section 5.5). The same result applies to estimator (5.10). We have thus proved that the HT estimator (5.9) and the Sen-Yates-Grundy (SYG) estimator (5.10) are unbiased estimators of the variance of the HT estimator of the population total.

Variance estimator (5.10) usually has a somewhat smaller variance than (5.9), but both estimators may assume a negative value in practice. Consider, for example, the extreme situation where π_k is exactly proportional to Y_k, in which case the HT estimator is equal to Y with probability 1. Therefore the variance of the HT estimator in this situation is 0. However, the HT variance estimator (5.9) is not 0 with probability 1, but is nonetheless unbiased, which means that the HT variance estimator assumes a negative value for a part of all possible samples. Conversely, in this situation variance estimator (5.10) is 0 with probability 1.

In practice, the exact expressions for π_{kl} are usually difficult to trace, making it hard to use the variance formulas (5.9) and (5.10). However, reasonable approximations for the variances are available in certain situations. An example of a good approximation occurs when n is much less than N. This case can be handled as if the sample was drawn with replacement. The next subsection returns to this point.

Finally it is noted that with SRSWOR (see Chapter 2):

$$\pi_k = \frac{n}{N} \qquad \wedge \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)} \qquad\qquad k,l=1,\cdots,N \quad .$$

It is left as an exercise for the reader to verify that substituting these two formulas in (5.8)-(5.10) indeed leads to the variance formulas derived in Chapter 2 for SRSWOR. We advise to start the derivations with the rather simpler SYG formulas.

5.3.3 PPS sampling

PPS sampling without replacement

In a PPS sample without replacement of size n, the inclusion probability π_k of population element k is by definition proportional to a given auxiliary variable X_k. Since the inclusion probabilities of the population elements necessarily sum to the sample size, π_k is defined as follows:

$$\pi_k = n\frac{X_k}{X} \qquad k = 1, \cdots, N \quad .$$

In other words, the PPS sample without replacement of size n is nothing but a sample without replacement of size n with variable inclusion probabilities. The results from Section 5.3.2 therefore apply to this type of sample design.

PPS sampling with replacement

In sampling *with* replacement with sample size n, for all n drawings, population element k has the same probability of actually being selected. This probability is referred to as the drawing probability of population element k, and is denoted as p_k.

For PPS sampling with replacement, the drawing probability p_k of population element k ($k = 1, \cdots, N$) is:

$$p_k = \frac{X_k}{X} = \frac{\pi_k}{n} \qquad \wedge \qquad \sum_{k=1}^{N} p_k = 1 \quad .$$

Where π_k stands for the inclusion probability of element k for PPS sampling without replacement. In other words, $\pi_k = np_k = nX_k / X$.

The standard estimator of the population total Y in PPS sampling *with* replacement is known as the Hansen-Hurwitz (HH) estimator – see Hansen and Hurwitz (1943). The HH estimator of the population total *Y* is defined as follows:

$$\hat{Y}_{HH} = \bar{z}_s = \frac{1}{n}\sum_{k=1}^{n} z_k \qquad \text{with} \qquad z_k \equiv \frac{y_k}{p_k} = \frac{y_k}{x_k} X \quad .$$

Note that z_k ($k = 1, \cdots, n$) can be viewed as stochastic, assuming the value $Z_l = Y_l / p_l$ ($l = 1, \cdots, N$) with probability p_l – see also Section 2.3.3. The unbiasedness of the HH estimator then follows from:

$$E(z_k) = \sum_{l=1}^{N} p_l Z_l = \sum_{l=1}^{N} p_l \frac{Y_l}{p_l} = Y \qquad k = 1, \cdots, n \quad .$$

Define further:

$$\sigma_z^2 = \sum_{l=1}^{N} p_l \left(\frac{Y_l}{p_l} - Y \right)^2 \quad .$$

Because the variance of z_k is:

$$\mathrm{var}(z_k) = E(z_k - Y)^2 = \sum_{g=1}^{N} p_l \left(\frac{Y_l}{p_l} - Y \right)^2 = \sigma_z^2 \qquad k = 1, \cdots, n \quad .$$

and because the sample is drawn with replacement, $\mathrm{var}(\hat{Y}_{HH})$ is:

$$\mathrm{var}(\hat{Y}_{HH}) = \frac{1}{n^2} \sum_{k=1}^{n} \mathrm{var}(z_k) = \frac{\sigma_z^2}{n} \quad . \tag{5.11}$$

This variance can be estimated without bias by:

$$\begin{aligned} \hat{\mathrm{var}}(\hat{Y}_{HH}) &= \frac{s_z^2}{n} \\ &= \frac{1}{n(n-1)} \sum_{k=1}^{n} (z_k - \bar{z}_s)^2 \\ &= \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(\frac{y_k}{p_k} - \hat{Y}_{HH} \right)^2 \quad . \end{aligned} \tag{5.12}$$

The unbiasedness of this variance estimator follows from:

$$\begin{aligned} E(s_z^2) &= \frac{n}{n-1} E\left\{ \frac{1}{n} \sum_{k=1}^{n} (z_k - Y)^2 - (\bar{z}_s - Y)^2 \right\} \\ &= \frac{n}{n-1} \left\{ \frac{1}{n} \sum_{k=1}^{n} \sigma_z^2 - \frac{\sigma_z^2}{n} \right\} = \sigma_z^2 \quad . \end{aligned}$$

Furthermore, when n is much less than N, $\mathrm{var}(\hat{Y}_{HT})$ in a PPS sample *without* replacement can be reasonably estimated as:

$$\hat{\mathrm{var}}_{HH}(\hat{Y}_{HT}) = \frac{s_z^2}{n} = \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(\frac{y_k}{x_k / X} - \hat{Y}_{HT} \right)^2 \quad . \tag{5.13}$$

The assumption underlying this approximation is that the variance of the PPS estimator of Y hardly depends on whether or not the selected elements are replaced.

5.3.4 A simple variance estimator for systematic PPS sampling

Although variance estimator (5.13) is intended for PPS sampling *with* replacement, with slight adjustment it can also be used for PPS sampling without replacement where *n* is not appreciably less than *N*. This adjusted variance estimator is implicitly referred to by Hájek (1964) for a comparable sample design known as *rejective sampling*, and is defined as follows:

$$\hat{\mathrm{var}}_{Haj}(\hat{Y}_{HT}) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (1 - p_k) \left(\frac{y_k}{x_k / X} - \hat{Y}_{HT} \right)^2 \quad . \tag{5.14}$$

The variance formula derived by Hájek (1964, p 1520) for *rejective sampling* is

$$\mathrm{var}_{Haj}\left(\hat{Y}_{HT}\right)=\frac{1}{n^2}\sum_{k=1}^{N}\pi_k\left(1-\pi_k\right)\left(\frac{y_k}{x_k/X}-Y^*\right)^2$$

$$Y^*=\sum_{k=1}^{N}\alpha_k\frac{y_k}{x_k/X} \qquad \alpha_k=\frac{\pi_k\left(1-\pi_k\right)}{\sum_{k=1}^{N}\pi_k\left(1-\pi_k\right)} \quad .$$

Variance estimator (5.14) can be used in PPS without replacement under the condition that the elements of the population are first put in a random order before the systematic PPS sample is drawn. It can also be used when n and N are of the same order of magnitude, provided Y_k / X_k and X_k are not correlated. If Y_k / X_k and X_k are correlated and n and N are also of the same order of magnitude, it would be better to use the following variance estimator:

$$\hat{\mathrm{var}}_{Haj2}\left(\hat{Y}_{HT}\right)=\frac{1}{n(n-1)}\sum_{k=1}^{n}\left(1-\pi_k\right)\left(\frac{y_k}{x_k/X}-\hat{Y}^*\right)^2$$

$$\hat{Y}^*=\frac{\sum_{k=1}^{n}\left(1-\pi_k\right)\frac{y_k}{x_k/X}}{\sum_{k=1}^{n}\left(1-\pi_k\right)} \quad .$$

See Knottnerus (2011a) for a comparison of the variances of the HT estimator in systematic PPS sampling and the ratio estimator in combination with SRSWOR sampling. Finally, if $\pi_k = n/N$ or $x_k = X/N$, (5.14) becomes the familiar estimator (2.12) for the variance of the direct estimator $\hat{Y}_{SRSWOR} = N\bar{y}_s$.

5.4 Example

To gain an impression of the reduction in variance with systematic PPS sampling, this section describes the details of estimating a (combined) price index of 70 companies based on a sample of size $n=9$. The variance for systematic PPS sampling is compared with that for SRSWOR. The price changes are derived from the price observations of Prodcom 27 from December 2004 to December 2005. The data are shown in Table 5.1. The example is taken from Knottnerus (2011a). The two largest companies were omitted because their turnover share was greater than 1/9. They were observed in full as a separate stratum.

Consider now the following price index for N companies:

$$I=\sum_{k=1}^{N}W_kP_k \qquad (N=70)$$

P_k : the price change for company k in the stated period in per cent

W_k : the revenue share of company k in base year $\left(\sum_{k=1}^{N}W_k=1\right)$.

Table 5.1. The price changes (P_k) and turnover shares (W_k) *of 70 companies*

k	price change	turnover share	*k*	price change	turnover share
1	-18.4%	0.0608	36	34.8%	0.0427
2.10	-16.0%	0.0784	37	13.1%	0.0121
3	3.3%	0.0762	38	31.7%	0.0351
4	12.5%	0.0100	39	-24.8%	0.0074
5	0.0%	0.0029	40	55.3%	0.0009
6	8.3%	0.0006	41	40.5%	0.0066
7	-39.0%	0.0182	42	34.6%	0.0022
8	-25.1%	0.0020	43	1.7%	0.0001
9	1.1%	0.0040	44	0.0%	0.0039
10	4.4%	0.0066	45	3.9%	0.0304
11	-4.9%	0.0039	46	25.4%	0.0209
12	-8.9%	0.0070	47	25.6%	0.0062
13	-7.0%	0.0148	48	0.0%	0.0033
14	-15.0%	0.0108	49	-0.3%	0.0019
15	-10.7%	0.0087	50	66.6%	0.0346
16	-9.0%	0.1079	51	0.0%	0.0039
17	-11.3%	0.0247	52	-2.9%	0.0007
18	10.6%	0.0024	53	15.8%	0.0011
19	-23.2%	0.0001	54	0.0%	0.0026
20	-25.4%	0.0001	55	0.0%	0.0018
21	-80.7%	0.0002	56	11.6%	0.0057
22	13.4%	0.0005	57	0.0%	0.0042
23	-42.5%	0.0010	58	0.0%	0.0236
24	-34.8%	0.0014	59	-1.5%	0.0015
25	-30.0%	0.0126	60	0.0%	0.0003
26	8.0%	0.0530	61	11.7%	0.0067
27	0.0%	0.0208	62	0.0%	0.0012
28	2.1%	0.0119	63	0.8%	0.0040
29	11.3%	0.0208	64	2.0%	0.0009
30	0.7%	0.0322	65	2.3%	0.0018
31	9.5%	0.0447	66	4.7%	0.0026
32	11.5%	0.0018	67	0.9%	0.0064
33	5.8%	0.0174	68	-1.0%	0.0309
34	-6.9%	0.0197	69	-0.5%	0.0005
35	0.0%	0.0124	70	0.0%	0.0006

Note that for a price index the target variable is $Y_k = W_k P_k$. This example will be used in comparing the SRSWOR and PPS sample designs as well as the associated estimators.

The standard estimator for a price index based on SRSWOR is a pure ratio estimator defined as:

$$\hat{I}_{SRSWOR} = \frac{\sum_{k=1}^{n} w_k p_k}{\sum_{k=1}^{n} w_k} \quad . \tag{5.15}$$

Although ratio estimation is discussed below in Chapter 6, the frequently used approximation for the variance of an estimated ratio is given here:

$$\begin{aligned} \text{var}\left(\hat{I}_{SRSWOR}\right) &= \text{var}\left(\frac{\sum_{k=1}^{n} w_k p_k}{\sum_{k=1}^{n} w_k}\right) \\ &= \text{var}\left(\frac{1/n\sum_{k=1}^{n} w_k p_k}{1/n\sum_{k=1}^{n} w_k} - I\right) \\ &= \text{var}\left(\frac{1}{n}\sum_{k=1}^{n} \frac{w_k\left(p_k - I\right)}{\overline{w}_s}\right) \quad . \end{aligned}$$

Because the variance of the denominator in the above expression is relatively small in comparison with the volatility of the numerator, the denominator is usually replaced with the population mean of the W_k, i.e. $1/N$. This gives:

$$\begin{aligned} \text{var}\left(\hat{I}_{SRSWOR}\right) &= N^2 \,\text{var}\left(\frac{1}{n}\sum_{k=1}^{n} w_k\left(p_k - I\right)\right) \\ &= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left(\frac{1}{N-1}\right)\sum_{k=1}^{N} W_k^2\left(P_k - I\right)^2 \quad . \end{aligned} \tag{5.16}$$

In the second part of (5.16) formula (2.12) was used with:

$$Y_k = W_k\left(P_k - I\right) \qquad \text{Ù} \qquad \overline{Y}_p = 0 \quad .$$

Applying this formula to the 70 price changes in Table 5.1, gives:

$$\text{var}\left(\hat{I}_{SRSWOR}\right) = 101 \quad . \tag{5.17}$$

Had sampling been with replacement, then the variance without finite population correction would be:

$$\text{var}\left(\hat{I}_{SRSWR}\right) = 116 \quad .$$

In other words, the finite population correction is $(1 - f) = 0.87$.

The variance formulas in (5.16) and (5.17) will now be compared with the variance of a second estimator based on systematic PPS sampling, where the inclusion probabilities are proportional to the W_k. As seen at the start of Section 5.3.3, $p_k = W_k n.$

The PPS estimator of *I* now equals in accordance with (5.3):

$$\hat{I}_{PPS} = \sum_{k=1}^{n} \frac{y_k}{p_k} = \sum_{k=1}^{n} \frac{w_k p_k}{w_k n} = \frac{1}{n} \sum_{k=1}^{n} p_k = \bar{p}_s \quad .$$

In other words, in PPS sampling the *weighted* mean of the separate price changes in the population is estimated with the *unweighted* mean of the price changes in the sample.

We first examine the variance for PPS sampling with replacement. Formula (5.11) in this example becomes:

$$\mathrm{var}_{HH}\left(\hat{I}_{PPS}\right) = \frac{s_z^2}{n} = \frac{1}{n} \sum_{k=1}^{N} W_k \left(P_k - I\right)^2 = 44 \quad .$$

We then examine the result of the variance approximation that underlies (5.14):

$$\begin{aligned} \mathrm{var}\left(\hat{I}_{PPS}\right) &= \frac{1}{n^2} \sum_{k=1}^{N} p_k \left(1 - p_k\right)\left(P_k - I\right)^2 \\ &= \frac{1}{n} \sum_{k=1}^{N} W_k \left(1 - nW_k\right)\left(P_k - I\right)^2 = 29 \quad . \end{aligned}$$

In other words, in PPS sampling the finite population correction of 0.66 $(= 29/44)$ is again much less than that of 0.87 for SRSWOR. Table 5.2 lists the various variances.

Table 5.2. Variances in various sample designs

	with replacement	without replacement
SRS	116	101
PPS	44	29

Finally, we point out that a simulation in which $20{,}000$ systematic PPS samples of size $n = 9$ were selected from a population of 70 companies with a different random sequence on each occasion yielded a variance of 30 for $\hat{I}_{PPS}$. In other words, (5.14) in this situation is an almost unbiased variance estimator. This example also illustrates once more that the reduction in variance in systematic PPS sampling compared with SRSWOR can be substantial, up to approximately 70%.

5.5 Quality indicators

Quality indicators for PPS sampling are:

- the margins of uncertainty of the corresponding estimators;
- the size of nonresponse;
- the randomness of the sequence of the elements in the population and the degree of dependence between Y_k / X_k and X_k.

Nonresponse can severely affect the quality of the results if (*i*) the size of the nonresponse is large and (*ii*) the nonresponse is selective. Part of the bias from selective nonresponse can sometimes be corrected by using auxiliary variables that correspond with both the probability of response and the target variable.

Another important assumption in random sampling is that the *frame* from which the sample is drawn corresponds closely with the *target population* about which inferences are to be made.

The above focused in particular on PPS sampling in which the elements are, or may be considered to be, in random order. If there is no appreciable relation between Y_k / X_k and k ($k = 1, \cdots, N$), a random ordering of the population will have no systematic effect on the findings that we observed earlier in Section 5.1.2. Therefore a random order is not absolutely necessary in situations of this kind. If the elements remain in a fixed sequence, the samples must not be allowed to become homogeneous in terms of the Y_k^* ($\equiv Y_k / \pi_k$).

5.6 Appendix

For the N selection indicators, by definition:

$$\sum_{k=1}^{N} a_k = a_1 + \cdots + a_N = n \quad .$$

Taking the expectation on both sides gives:

$$\sum_{k=1}^{N} \pi_k = n \quad .$$

Likewise by definition:

$$a_k (a_1 + \cdots + a_N) = n\, a_k \quad .$$

Again, taking the expectations on either side immediately gives:

$$\sum_{l=1}^{N} \pi_{kl} = n\pi_k \quad .$$

These two results will be needed below in this section.

Proof of (5.7)

For selection indicator a_k:

$$\begin{aligned} \text{var}(a_k) &= E(a_k - E(a_k))^2 \\ &= E(a_k^2) - E^2(a_k) \\ &= E(a_k) - \pi_k^2 = \pi_k - \pi_k^2 \end{aligned}$$

and:

$$\begin{aligned}\operatorname{cov}(a_k, a_l) &= E\big((a_k - E(a_k))(a_l - E(a_l))\big) \\ &= E(a_k a_l) - E(a_k)E(a_k) \\ &= \pi_{kl} - \pi_k \pi_l \quad .\end{aligned}$$

QED.

Proof of (5.8)

It follows from (5.7) and the definition of the HT estimator of the population total – see (5.5) – that its variance is:

$$\begin{aligned}\operatorname{var}\left(\hat{Y}_{HT}\right) &= \sum_{k=1}^{N} \pi_k (1 - \pi_k) \frac{Y_k^2}{\pi_k^2} + \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l \neq k}}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{Y_k Y_l}{\pi_k \pi_l} \\ &= \sum_{k=1}^{N} \sum_{l=1}^{N} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \right) Y_k Y_l \quad .\end{aligned}$$

QED.

The following results can be proved for estimators (5.9) and (5.10):

$$E\left(\hat{\operatorname{var}}_{HT}\left(\hat{Y}_{HT}\right)\right) = \operatorname{var}\left(\hat{Y}_{HT}\right) \quad \wedge \quad E\left(\hat{\operatorname{var}}_{SYG}\left(\hat{Y}_{HT}\right)\right) = \operatorname{var}\left(\hat{Y}_{HT}\right) \quad .$$

The proofs are as follows. The HT estimator $\hat{\operatorname{var}}(\hat{Y}_{HT})$ (see (5.9)) is rewritten in terms of the selection indicators as:

$$\hat{\operatorname{var}}_{HT}\left(\hat{Y}_{HT}\right) = \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} y_k y_l = \sum_{k=1}^{N} \sum_{l=1}^{N} \left(a_k a_l \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \right) Y_k Y_l \quad .$$

The unbiasedness of the variance estimator now follows from the property: $E(a_k a_l) = \pi_{kl}$ (see (2.9)). It is possible to calculate that:

$$\begin{aligned}E\left(\hat{\operatorname{var}}_{HT}\left(\hat{Y}_{HT}\right)\right) &= \sum_{k=1}^{N} \sum_{l=1}^{N} \left(E(a_k a_l) \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \right) Y_k Y_l \\ &= \sum_{k=1}^{N} \sum_{l=1}^{N} \left(\pi_{kl} \times \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \right) Y_k Y_l \\ &= \sum_{k=1}^{N} \sum_{l=1}^{N} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \right) Y_k Y_l = \operatorname{var}_{HT}\left(\hat{Y}_{HT}\right) \quad .\end{aligned}$$

A condition for unbiasedness is again that $0 < \pi_{kl} \quad 1 \leq k, l \leq N$.

Because $E(a_k a_l) = \pi_{kl}$, estimator (5.10) as such is an unbiased estimator of :

$$\operatorname{var}_{SYG}\left(\hat{Y}_{HT}\right) = -\frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right)^2$$

provided $0 < \pi_{kl}$. In other words:

$$E\left(\hat{\text{var}}_{SYG}\left(\hat{Y}_{HT}\right)\right) = -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl} - \pi_k\pi_l\right)\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2 \quad .$$

Expanding the squared term on the right-hand side of the above equation and using the two properties of the selection indicators that were derived at the start of this section, $\sum_{k=1}^{N}\pi_k = n$ and $\sum_{l=1}^{N}\pi_{kl} = n\pi_k$, again yields:

$$E\left(\hat{\text{var}}_{SYG}\left(\hat{Y}_{HT}\right)\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}\right)Y_kY_l \quad .$$

In other words, estimator (5.10) is also an unbiased estimator of (5.8).

6. Ratio estimator

6.1 Short description

Chapter 2 showed that the sample mean is an unbiased estimator of the population mean. It is also referred to as the direct estimator of $\bar{Y}$, because it uses only the observed values of the variable Y. Later chapters showed how the precision of an estimator can be improved by incorporating available information about the population into the sampling design. For instance, Chapter 3 used auxiliary information to divide the population into homogeneous groups, or strata, from which separate samples can then be drawn. Chapter 5 showed that auxiliary information about the size of the elements makes it possible to select elements with unequal probabilities. This and the following two chapters show how auxiliary information can be used in the estimation procedure. Different types of auxiliary information result in different estimators. We introduce some commonly used alternative estimators of the population mean (or total) for the simplest sampling design: simple random sampling (without replacement). The discussion is restricted to estimation methods that involve one auxiliary variable. As with stratified sampling and sampling with unequal probabilities, the use of auxiliary information leads to more precise estimators if the relation between target variables and auxiliary variables is sufficiently strong.

So far we have discussed the most important building blocks of sample surveys: simple random sampling, stratification, clustering, and sampling with unequal probabilities. Ratio and regression estimators are discussed below. Most sampling designs used at Statistics Netherlands consist of multiple building blocks. The formulas for the estimators, and in particular those for the standard errors, can then become extremely complex. In practice, weights are used to obtain point estimates of parameters, which usually simplifies matters considerably. The weights can be determined automatically using the Bascula software package, which also has various functions for calculating variances. The remaining chapters briefly discuss the relationship between estimating and weighting.

6.2 Applicability

The ratio estimator is a relatively simple and widely used alternative for the direct estimator discussed in Chapter 2. The thinking behind the ratio estimator, which uses a quantitative auxiliary variable, is that the ratio Y_k / X_k of the values of the target variable and the auxiliary variable in the population varies less than the target variable itself.

An example is turnover per employee. If this ratio varies little from one business to another, it can be used to estimate total turnover in the Netherlands. The procedure would be to estimate the ratio based on a sample and then multiply the estimated ratio by the number of employees in the Netherlands.

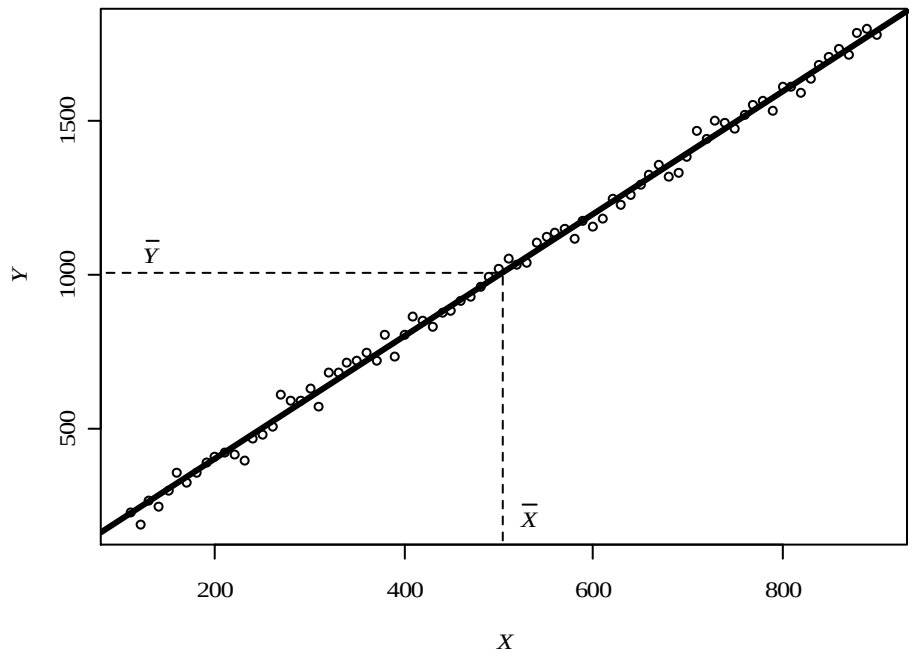
6.3 Detailed description

The limited variation of the ratio of the target variable Y and the auxiliary variable X can be expressed as:

$$\frac{Y_k}{X_k} \approx B \text{ or } Y_k \approx BX_k$$

where B is a constant. The points (X_k, Y_k) lie approximately on a straight line through the origin, which is referred to as the regression line (see Figure 6.1). If the regression line is a good description of the set of points, the individual values of Y can be predicted satisfactorily based on X. And because we have population information on X it would be logical to incorporate it into the estimator.

Figure 6.1. Linear relationship between Y and X through the origin



Summing all the elements of the population and dividing by *N* gives:

$$\bar{Y} = B\bar{X}.$$

If the population mean $\bar{X}$ is known, and the value of B is available, $\bar{Y}$ can then be estimated as $B\bar{X}$. The ideal value for B is evidently $\bar{Y}/\bar{X}$, but this cannot be calculated, because it assumes knowledge of $\bar{Y}$. However, this value can be estimated from the sample data using:

$$\hat{B} = b = \frac{\bar{y}_s}{\bar{x}_s} = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}} \quad .$$

This expression gives the ratio estimator $\hat{\bar{Y}}_{QT}$ for the population mean $\bar{Y}$:

$$\hat{\bar{Y}}_{QT} = b\bar{X} = \frac{\bar{y}_s}{\bar{x}_s}\bar{X} \quad . \tag{6.1}$$

For the ratio estimator the sample mean $\bar{y}_s$ is multiplied by a term $\bar{X}/\bar{x}_s$, which adjusts for the difference between the sample mean and the population mean of the auxiliary variable. The assumption here is that the adjustment can also be projected onto the target variable, since there is a relation between the target variable and the auxiliary variable. Figure 6.2 shows how the ratio estimator works.

Figure 6.2. The ratio estimator

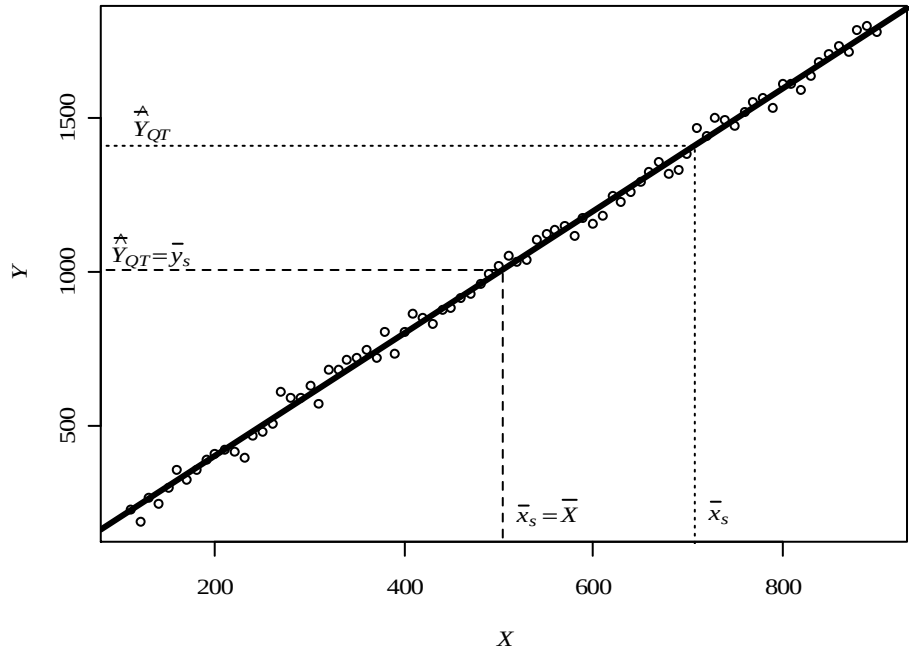


Figure 6.2 shows that a ratio estimate is obtained by adding a term $b(\bar{X}-\bar{x}_s)$ to the sample mean $\bar{y}_s$, where $b=\bar{y}_s/\bar{x}_s$, is the slope of the regression line. Therefore the ratio estimator (6.1) can also be written as:

$$\hat{\bar{Y}}_{QT} = \bar{y}_s + b(\bar{X}-\bar{x}_s) = b\bar{X} \quad .$$

Note that by setting $Y=X$ the ratio estimator equals $\bar{X}$. Regarding the auxiliary variable, the sample outcome is therefore consistent with the population, which is a very desirable property in practice. It enables us to align different statistics by using the same auxiliary variable.

The ratio estimator in (6.1) can also be written in terms of weights. That is,

$$\hat{\bar{Y}}_{QT} = b\bar{X} = \frac{\bar{y}_s}{\bar{x}_s}\bar{X} = \sum_{k=1}^{n} w_k y_k$$

$$w_k = \frac{\bar{X}}{n\bar{x}_s} \quad .$$

The weights are the same for all observations in the sample and satisfy the condition $\sum_{k=1}^{n} w_k x_k = \bar{X}$. The weight effectively represents the corresponding fraction of elements in the population with the same value y_k. The weight also gives an impression of the possible contribution (influence) of the observation to the outcome.

For convenience we assume positive target and auxiliary variables, so that B and b are also positive. As such, the auxiliary variable may also be zero as long as the sample mean $\bar{x}_s$ is positive. If, for example, Y_k is the turnover of a company k in some period, and X_k is the turnover in the previous period, then X_k will be zero if the company had yet to come into existence at the time. Similarly, $Y_k = 0$ if a company no longer exists.

The ratio estimator is not an unbiased estimator of the population mean $\bar{Y}$. Cochran (1977, p 160) proves that the bias is inversely proportional to the sample size, so that the bias is negligible for large samples. In the formula for the estimator (6.1), $\hat{B} = b$ is a stochastic variable and $\bar{X}$ is constant. The bias of $\hat{B}$ is approximately:

$$E[\hat{B}] - B \approx \frac{1-f}{n}\frac{1}{\bar{X}^2}\left[BS_x^2 - S_{xy}\right] = \frac{1}{N}\left(\frac{1-f}{f}\right)B\left[CV_x^2 - CV_{xy}\right] \quad .$$

Where CV_x is the coefficient of variation of X – see also (2.1). Furthermore:

$$CV_{xy} = \frac{S_{xy}}{\bar{X}\bar{Y}} \quad .$$

The bias is therefore approximately zero when $CV_x^2 = CV_{xy}$, in which case the regression line passes through the origin and Y_k is roughly proportional to X_k.

Unbiasedness is only one of the criteria for assessing the quality of an estimator. For an unbiased estimator, the discrepancy between the estimates and the population value, calculated for all possible samples, will average zero. A second, and usually more important, quality criterion is that the estimates fluctuate little from sample to sample, and all deviate little from the actual population value. This results in a small standard error. For biased estimators the mean squared error is often a more important criterion than the variance. The (small) bias of the ratio estimator is usually compensated by a small mean squared error. The mean squared error of $\hat{B}$ can be calculated as:

$$G\left(\hat{B}\right) = E\left(\left(\hat{B} - B\right)^2\right) = \mathrm{var}\left(\hat{B}\right) + \left(E\left(\hat{B}\right) - B\right)^2 \quad .$$

When the bias of $\hat{B}$ is small relative to the standard error of $\hat{B}$, the variance is a good approximation of the mean squared error. Then (see e.g., Muilwijk, 1992, Section 5.8.3):

$$G(\hat{B}) \leq \left[1+\frac{1}{N}\left(\frac{1-f}{f}\right)CV_x\right]\text{var}\left(\hat{B}\right)$$

so that the approximation is better for larger f and/or smaller CV_x.

A good approximation for the variance of the ratio estimator of the population mean is obtained from:

$$\text{var}\left(\hat{\bar{Y}}_{QT}\right)=\frac{1}{N(N-1)}\left(\frac{1-f}{f}\right)\sum_{k=1}^{N}(Y_k-BX_k)^2 \quad . \tag{6.2}$$

The variance is smaller as the relationship between Y and X is stronger, that is, when the values Y_k deviate less from the regression line. However, it is impossible to calculate the variance (6.2) because there are no population data of *Y*, but we can nonetheless produce an estimate based on the sample. The variance estimator of (6.2) is:

$$\hat{\text{var}}\left(\hat{\bar{Y}}_{QT}\right)=\frac{1}{N}\left(\frac{1-f}{f}\right)\left\{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-bx_k)^2\right\} \quad . \tag{6.3}$$

For large samples the variance of the ratio estimator is usually less than that of the direct estimator. It is even possible to determine the situations where this occurs. Rewrite the variance of the ratio estimator as:

$$\text{var}\left(\hat{\bar{Y}}_{QT}\right)=\frac{1}{N}\left(\frac{1-f}{f}\right)\left(S_y^2+B^2S_x^2-2BS_{xy}\right) \quad .$$

Comparing this variance with the variance for the sample mean in simple random sampling – see (2.11) – gives:

$$\begin{aligned}\text{var}\left(\hat{\bar{Y}}_{QT}\right)-\text{var}\left(\hat{\bar{Y}}\right) &\approx \frac{1}{N}\left(\frac{1-f}{f}\right)\left(S_y^2+B^2S_x^2-2BS_{xy}-S_y^2\right)\\ &=\frac{1}{N}\left(\frac{1-f}{f}\right)BS_x\left(BS_x-2\rho S_y\right) \quad ,\end{aligned}$$

where ρ is the population correlation coefficient of *X* and *Y*. The ratio estimator is therefore preferable to the direct estimator if:

$$\frac{1}{2}B\left(\frac{S_x}{S_y}\right)=\frac{1}{2}\left(\frac{CV_x}{CV_y}\right)<\rho \quad .$$

Therefore, if the coefficient of variation of X is more than twice as large as that of *Y*, i.e. $2CV_y < CV_x$, then the ratio estimator is less precise than the direct estimator (even for a correlation of $\rho=1$). If the variable X equals the variable Y, but was measured at an earlier time, then the coefficients of variation are approximately

equal and the ratio estimator is better than the direct estimator if $0.5 < \rho$. If all the values of X are the same $(S_x^2 = 0)$, the ratio estimator corresponds with the direct estimator for simple random sampling.

Again the ratio estimator of the population mean simply yields the ratio estimator of the population total after multiplying by the population size:

$$\hat{Y}_{QT} = N\bar{Y}_{QT} = bN\bar{X} = bX \quad .$$

The variance of this estimator can be calculated as:

$$\text{var}\left(\hat{Y}_{QT}\right) = \left(\frac{N}{N-1}\right)\left(\frac{1-f}{f}\right)\sum_{k=1}^{N}\left(Y_k - BX_k\right)^2 \quad .$$

The ratio estimator of the population total can also be written in terms of weights:

$$\hat{Y}_{QT} = Nb\bar{X} = \frac{\bar{y}_s}{\bar{x}_s}X = \sum_{k=1}^{n} w_k^{\#} y_k$$

$$w_k^{\#} = \frac{X}{n\bar{x}_s} \quad .$$

The weights $w_k^{\#}$ satisfy $\sum_{k=1}^{n} w_k^{\#} x_k = X$. Analogous to the ratio estimator of the mean, the weight now effectively represents the number of elements in the population with the same value y_k. For this reason the $w_k^{\#}$ are sometimes called raising weights.

The ratio estimator is usually used to improve precision. The ratio estimator also ensures the consistency of sample estimates with population numbers or totals for the auxiliary variables (see also Chapter 8) and adjusts the estimator for nonresponse. The parameter to be estimated can sometimes be a ratio (e.g. the number of fatal traffic accidents in proportion to the total number of accidents). The ratio estimator can also be used for estimating a total when the population size is unknown. This is particularly relevant when producing estimates for subpopulations.

Sometimes several auxiliary variables correlate with the target variable Y, in which case one auxiliary variable may be chosen, e.g. the one with the highest correlation. Another option would be to employ more than one auxiliary variable using the general regression estimator. Finally, we note that often nothing is known about the relationship between Y and X in the population. If this is the case, we have to rely on the sample data, provided the sample distribution with respect to Y is not skewed due to chance or nonresponse. Information from earlier surveys can also help in understanding the relation between Y and X. The ratio estimator is most appropriate when a straight line through the origin is the best description of the relationship between Y and X, and when the variance of the values Y_k about the regression line are proportional to X_k.

6.4 Example

A simple random sample of 8 towns is drawn from a population of 42 towns. Table 6.1 gives the population figures and the number of general practitioners for each town. We have to estimate the total number of general practitioners in the population. The number of general practitioners will be related to the number of residents of a town, and the numbers of residents will therfore be used as auxiliary information in the estimation proces. On average there are 23.5 general practitioners in the 8 selected towns. Without using the auxiliary variable, the total number of general practitioners is estimated at $\hat{Y} = N\bar{y} = 987$ with a standard error of 80.48. The mean number of residents in the population is 50, and in the sample 47.13. In relation to the population the mean number of residents in the sample is therefore a little too low, and this is probably also true of the mean number of general practitioners in the sample.

The estimate of the total number of general practitioners in the population using the ratio estimator is:

$$\hat{Y}_{QT} = \frac{\bar{y}_s}{\bar{x}_s} X = 0.499 \times 2{,}100 = 1{,}047 \quad .$$

The standard error of the estimator is:

$$\hat{S}_e\left(\hat{Y}_{QT}\right) = \sqrt{N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_{k=1}^{n} \left(y_k - bx_k\right)^2}$$
$$= \sqrt{42^2 \times 0.10 \times 114.68} = 54.08 \quad .$$

Table 6.1 Population and number of general practitioners in a sample of 8 towns

Town	Residents (x1000)	General practitioners	$(y_k - bx_k)^2$
1	38	20	1.10
2.10	52	23	8.59
3	75	35	5.76
4	32	20	16.34
5	60	28	3.69
6	43	25	12.65
7	57	22	41.27
8	20	15	25.27
Total	377	188	114.68

6.5 Quality indicators

Quality indicators for the ratio estimator are:

- the margins of uncertainty of the corresponding estimators;

- the size of nonresponse;
- the size of the sample.

Nonresponse can severely affect the quality of the results if (i) the nonresponse is large and (ii) the nonresponse is selective. The bias from selective nonresponse can sometimes be corrected by using auxiliary variables that correspond with both the probability of response and the target variable, like the ratio estimator described in this ch apter.

Another important assumption in employing the ratio estimator is that the size of the sample is sufficiently large to ensure that the bias of the ratio estimator and the corresponding variance estimator remain limited.

7. Regression estimator

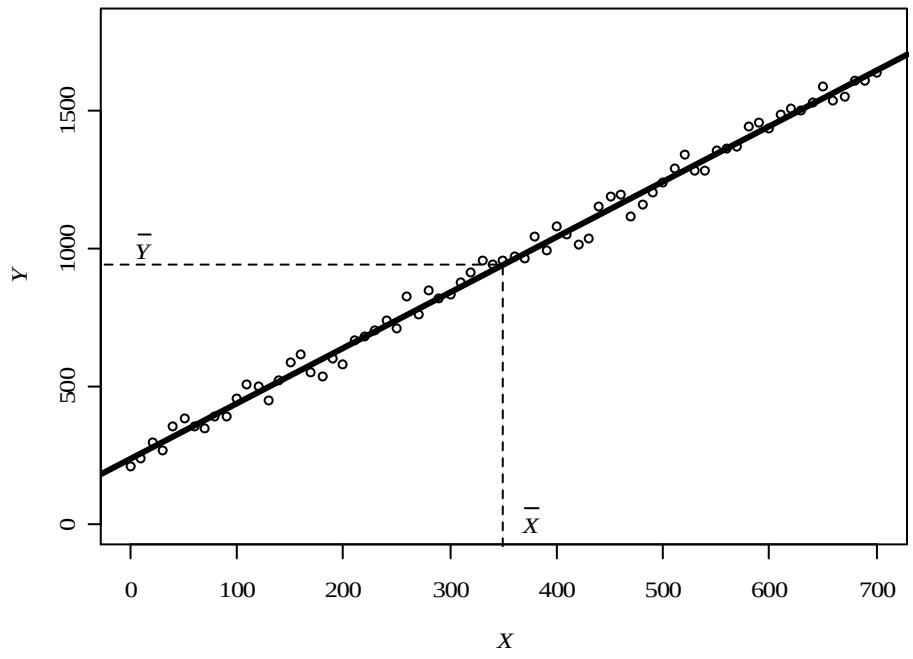
7.1 Short description

When the relationship is linear, but does not pass through the origin, it is better to use the regression estimator rather than the ratio estimator – see Figure 7.1. The relationship between Y and X is now given by:

$$Y_k \approx A + BX_k \tag{7.1}$$

where A and B are nonzero constants. These constants are referred to as the regression coefficients.

Figure 7.1. Linear relationship between Y and X



This linear relationship is the starting point for the regression estimator. Section 7.3 describes the regression estimator in more detail.

7.2 Applicability

The regression estimator and its variants are frequently used in socioeconomic statistics. The usual source of auxiliary variables is the Municipal Personal Records Database (GBA). The *Repeated Weighting* subtheme (part of the theme 'Sampling theory' of the Methods Series; Gouweleeuw and Knottnerus, 2008) also uses the regression estimator, albeit with a rather different purpose. The regression estimator is also often used in other themes

7.3 Detailed description

Assuming an approximately linear relationship between *Y* and *X*, summation in (7.1) over the elements of the population and dividing by *N* gives the following approximation:

$$\bar{Y} \approx A + B\bar{X} \quad .$$

If the constants A and B are known and the population mean of the auxiliary variable X is also known, then the population mean of the target variable can be estimated by $A + B\bar{X}$. However, in practice, we do not have the values A and B, and have to resort to estimates. The aim then is to find values for A and B that for every element k in the population make $A + BX_k$ as close as possible to Y_k. We can achieve this aim with the method of least squares, i.e. by minimizing the sum of the squares:

$$\sum_{k=1}^{N} \left(Y_k - A - BX_k\right)^2 \quad .$$

The sum of the squares is minimum when B is

$$B = \frac{\sum_{k=1}^{N} \left(X_k - \bar{X}\right)\left(Y_k - \bar{Y}\right)}{\sum_{k=1}^{N} \left(X_k - \bar{X}\right)^2} = \frac{S_{xy}}{S_x^2}$$

and when A is:

$$A = \bar{Y} - B\bar{X} \quad .$$

The values of A and B can be estimated from the outcomes of the sample by:

$$\hat{B} = b = \frac{\sum_{k=1}^{n} \left(x_k - \bar{x}_s\right)\left(y_k - \bar{y}_s\right)}{\sum_{k=1}^{n} \left(x_k - \bar{x}_s\right)^2} = \frac{s_{xy}}{s_x^2} \tag{7.2}$$

and:

$$\hat{A} = a = \bar{y}_s - b\bar{x}_s \quad .$$

The regression estimator for the population mean, denoted as $\hat{\bar{Y}}_{REG}$, is then defined by:

$$\hat{\bar{Y}}_{REG} = a + b\bar{X} = \bar{y}_s + b\left(\bar{X} - \bar{x}_s\right) \quad . \tag{7.3}$$

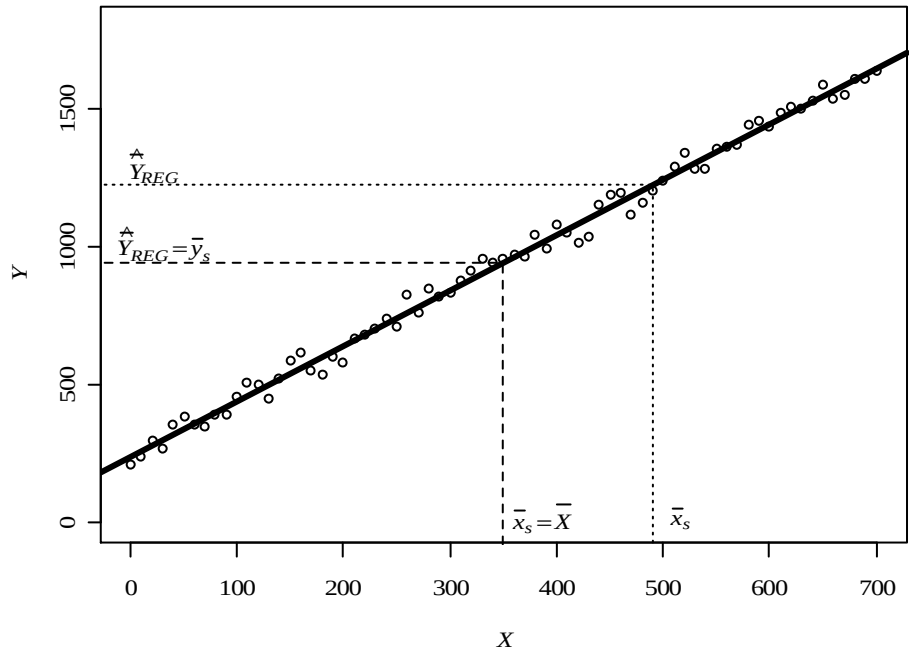
note that choosing $a = 0$, $\hat{\bar{Y}}_{REG} = b\bar{X}$ with b equal to:

$$b = \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} x_k^2} \quad .$$

However, this estimator is not the same as the ratio estimator, due to different starting points in deriving the estimators (Särndal et al. 1992, Section 6.4).

As with the ratio estimator the regression estimator is an adjustment to the sample mean $\bar{y}_s$ – see Figure 7.2. Adjustment occurs as soon as there is a difference between the population mean and the sample mean of the auxiliary variable X.

Figure 7.2. The regression estimator



By analogy with the ratio estimator, the regression estimator of the population mean can also be written in terms of weights. Substituting (7.2) into (7.3) gives:

$$\hat{\bar{Y}}_{REG} = \sum_{k=1}^{n} w_k y_k$$

$$w_k = \frac{1}{n} + \frac{(x_k - \bar{x}_s)(\bar{X} - \bar{x}_s)}{\sum_{k=1}^{n}(x_k - \bar{x}_s)^2} \quad .$$

As with the ratio estimator, the weights w_k satisfy $\sum_{k=1}^{n} w_k x_k = \bar{X}$.

The regression estimator is not an unbiased estimator of the population mean, because b is not an unbiased estimator of B. The bias of $\hat{\bar{Y}}_{REG}$ amounts to $-\text{cov}(\hat{B}, \bar{x})$, where $\text{cov}(\cdot)$ stands for covariance. As with the ratio estimator, the bias of the regression estimator is inversely proportional to sample size, so that it becomes negligible for large samples. If the regression line passes through all points (X_k, Y_k), then $\hat{B} = B$ for each possible sample and $\text{cov}(\hat{B}, \bar{x}) = 0$. The bias in this situation is zero. The variance of the regression estimator can be well approximated by:

$$\text{var}\left(\hat{\bar{Y}}_{REG}\right) = \frac{1}{N(N-1)}\left(\frac{1-f}{f}\right)\sum_{k=1}^{N}\left(Y_k - \bar{Y} - B\left(X_k - \bar{X}\right)\right)^2 \quad .$$

The variance is determined by the variation in Y that cannot be explained by X, which is referred to as the residual sum of squares. The residual for element k, denoted by e_k, is defined by:

$$e_k = Y_k - \left(A + BX_k\right) = Y_k - \bar{Y} - B\left(X_k - \bar{X}\right) \quad .$$

The more variation in Y can be explained by X, the stronger the relationship between Y and X, and the smaller the variance. A different way of achieving this result is to rewrite the variance as:

$$\text{var}\left(\hat{\bar{Y}}_{REG}\right) = \frac{1}{N}\left(1-\rho^2\right)\left(\frac{1-f}{f}\right)S_y^2 = \left(1-\rho^2\right)\text{var}\left(\bar{y}\right) \quad .$$

The closer the population correlation coefficient to $+1$ or -1, the smaller the variance will be. The variance is always a factor $(1-\rho^2)$ smaller than the variance of the direct estimator. If there is no relationship between Y and X, i.e. $\rho = 0$, then the variance of the regression estimator equals that of the direct estimator, in which case it would be pointless to use X as a auxiliary variable.

An estimator of the variance of the regression estimator based on the sample is:

$$\hat{\text{var}}\left(\hat{\bar{Y}}_{REG}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)\left\{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \bar{y}_s - b\left(x_k - \bar{x}_s\right)\right)^2\right\} \quad .$$

A regression estimator for the population total can be derived directly from regression estimator (7.3) of the population mean:

$$\hat{Y}_{REG} = N\hat{\bar{Y}}_{REG} = N\bar{y}_s + Nb\left(\bar{X} - \bar{x}_s\right) \quad .$$

This regression estimator for the population total can also be rewritten in terms of weights:

$$\hat{Y}_{REG} = \sum_{k=1}^{n} w_k^{\#} y_k$$

$$w_k^{\#} = \frac{N}{n} + \frac{N\left(x_k - \bar{x}_s\right)\left(\bar{X} - \bar{x}_s\right)}{\sum_{k=1}^{n}\left(x_k - \bar{x}_s\right)^2} \quad .$$

The weights in this regression estimator satisfy the condition $\sum_{k=1}^{n} w_k^{\#} x_k = X$. See Camstra and Nieuwenbroek (2002) and the series topic on *Repeated weighting*, for weights where there are two or more auxiliary variables.

The variance of the regression estimator of the population total can be calculated as follows:

$$\text{var}\left(\hat{Y}_{REG}\right) = N^2\,\text{var}\left(\hat{\bar{Y}}_{REG}\right) \quad .$$

The regression estimator is usable with both positive and negative correlation; Y_k and X_k can also assume zero or negative values.

7.4 Example

Examination of the sample data from Example 6.4 raises the suspicion that although it is linear, the relationship between the number of residents and number of general practitioners for each town cannot be expressed as a regression line through the origin. It would then seem to be logical to apply the regression estimator. The sample parameters are:

$$\bar{x} = 47.13 \qquad s_x^2 = 304.13$$
$$\bar{y} = 23.50 \qquad s_y^2 = 36.24$$
$$s_{xy} = 95.97 \quad .$$

The estimates of the regression coefficients are:

$$b = \frac{s_{xy}}{s_x^2} = 0.31 \qquad \wedge \qquad a = \bar{y} - b\bar{x} = 8.66 \quad .$$

The regression estimate of the total number of general practitioners is therefore:

$$\hat{Y}_{REG} = N(a + b\bar{X}) = 42 \times (8.66 + 14.84) = 987$$

which corresponds with the estimate obtained by applying the direct estimator. The standard error of the regression estimator of the total is:

$$\hat{S}_e\left(\hat{Y}_{REG}\right) = \sqrt{N\left(\frac{1-f}{f}\right)\left\{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \bar{y}_s - b\left(x_k - \bar{x}_s\right)\right)^2\right\}}$$
$$= \sqrt{42 \times \left(\frac{1-(8/42)}{(8/42)}\right) \times \left\{\frac{1}{7} \times 42.82\right\}} = 35.05 \quad .$$

The standard error is, therefore, somewhat less than the standard error of the direct estimator and also less than the standard error of the ratio estimator. For large samples the regression estimator is never worse than the ratio estimator.

7.5 Quality indicators

Quality indicators for the regression estimator closely resemble those of the ratio estimator:

- the margins of uncertainty of the corresponding estimators;
- the size of nonresponse;
- the number of auxiliary variables;
- the size of the sample.

Nonresponse can severely affect the quality of the results if (i) the nonresponse is large and (ii) the nonresponse is selective. Part of the bias from selective nonresponse can sometimes be corrected by using the regression estimator described

in this chapter with auxiliary variables that correspond with both the probability of response and the target variable.

Another important assumption in the regression estimator is that the size of the sample is sufficiently large to ensure that the bias of the regresssion estimator and the bias of its variance estimator remain limited. Unlike the ratio estimator, it is also important with the regression estimator that the number of auxiliary variables in the model is not too large in proportion to the number of observations. For reasons of simplicity this section is restricted to the case of one auxiliary variable, but in practice many more auxiliary variables can be included in the regression model, depending on the size of the sample.

8. Poststratified estimators

8.1 Short description

A poststratification estimator is obtained when the population is divided into poststrata after SRSWOR (see Chapter 2) is carried out. The sample is also divided in accordance with the poststrata. In other words, the sample size for each poststratum is stochastic in nature, which is the essential difference with ordinary (pre)stratification.

8.2 Applicability

Poststratification is used regularly when it is not known in advance to which stratum each element of the population belongs, or if no account was taken in the sampling design of the (relevant) stratification. For example, when surveying the financial situation of young adolescents, where there is no precise knowledge of whether the subjects are students, employed, or unemployed. This lack of knowledge hampers prior stratification. Once sampling has taken place, the selected elements can be divided into the different strata, which is known as poststratification. However, precautions must be taken to limit the probability of having strata without sample observations, which means not choosing excessively small strata. If empty strata occur in practice, nonetheless, it is still possible to combine strata.

8.3 Detailed description

8.3.1 The standard poststratified estimator

Unlike the ratio estimator and the regression estimator, the poststratified estimator uses only qualitative (categorical) auxiliary variables, such as gender, region and marital status. Suppose that the population is divided according to a qualitative auxiliary variable into L disjunct groups. After observing the value of the auxiliary variable, it can be established for each element in the population to which group (poststratum) it belongs. The number of elements in the sample that belong to poststratum l is denoted with n_l for $l = 1, \cdots, L$. The number of elements in the population in poststratum l is N_l (this number must be known). To distinguish the poststrata from the standard prestrata, subscript l is used for the l^{th} poststratum $(l = 1, \cdots, L)$ and subscript h for the h^{th} standard prestratum $(h = 1, \cdots, H)$. The poststratification estimator of the population mean $\bar{Y}$ is

$$\hat{\bar{Y}}_{post} = \sum_{l=1}^{L} \frac{N_l}{N} \bar{y}_l \,. \tag{8.1}$$

The poststratification estimator is unbiased under the condition that every poststratum has at least one observation. At first glance, the estimator appears to be the same as the direct estimator for a stratified sample. However, there is a conceptual difference. The sample size in a poststratum is stochastic; the number of

sample elements that will end up in the poststratum is unknown before the sample is drawn. For a (pre)stratified sample, the exact number of sample elements per stratum is determined beforehand. The stochastic nature of the sample size per poststratum is expressed in the variance of the poststratification estimator. For large samples the variance can be well approximated as:

$$\text{var}\left(\hat{\bar{Y}}_{post}\right) = \frac{1-f}{n}\sum_{l=1}^{L}\frac{N_l}{N}S_l^2 + \frac{1-f}{n^2}\sum_{l=1}^{L}\left(1-\frac{N_l}{N}\right)S_l^2 \quad . \tag{8.2}$$

The first term in the variance formula is the variance for prestratification and proportional allocation. The second term reflects the uncertainty about the sample sizes of the poststrata. This term can be interpreted as the additional contribution to the variance because the allocation that is realized differs from the proportional allocation. The *more homogeneous* the strata with respect to the target variable, i.e. the smaller the differences between the values of the target variable within a stratum, the smaller the variance of the poststratification estimator. It is therefore important to select the poststrata such that the variation of the target variable manifests itself mainly in the differences between the strata. The estimator for the variance is obtained by replacing S_l^2 in (8.2) by the sampling equivalents s_l^2.

8.3.2 The poststratified estimator for complex sampling designs

Poststratification often assumes one or more categorical auxiliary variables X that divide the population into L categories (poststrata), each with N_l elements. The population total of Y in poststratum $l = 1,\cdots,L$ can be estimated with the Horvitz-Thompson estimator:

$$\hat{Y}_{HT,l} = \sum_{k=1}^{n_l}\frac{y_{lk}}{\pi_{lk}}$$

where n_l is the number of elements in the sample that belong to poststratum l, and π_{lk} is the inclusion probability of element k in stratum l. Likewise the population size N_l of poststratum l can be estimated with:

$$\hat{N}_{HT,l} = \sum_{k=1}^{n_l}\frac{1}{\pi_{lk}}$$

so that an estimator for $\bar{Y}_l$ is:

$$\hat{\bar{Y}}_l = \frac{\hat{Y}_{HT,l}}{\hat{N}_{HT,l}} \quad .$$

The poststratified estimator of the total is then:

$$\hat{Y}_{post} = \sum_{l=1}^{L}\frac{N_l}{\hat{N}_{HT,l}}\sum_{k=1}^{n_l}\frac{y_{lk}}{\pi_{lk}} \quad .$$

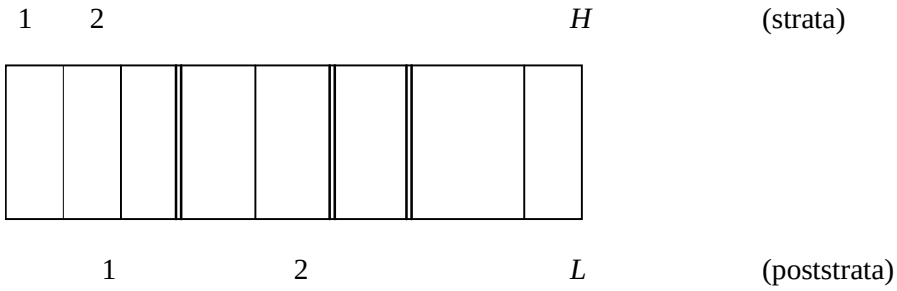
Note that L different correction weights are involved, and that in a poststratum l they equal $N_l / \hat{N}_{HT,l}$. The poststratification estimator is approximately unbiased. Although it is possible to give a good approximation formula of the variance, it sheds little light on a specific sampling design. The discussion of the poststratification estimator is therefore restricted to stratified sampling, omitting formulas.

Poststratification with a stratified sample

Suppose the population is divided into H strata and we have selected a stratified sample with n_h elements in each stratum. We distinguish four cases that may occur in practice.

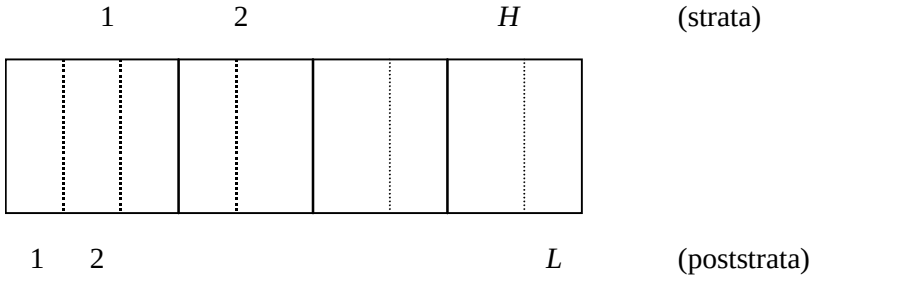
(a) The strata and the poststrata coincide (i.e. $H = L$).

(b) The poststrata are unions of strata. An example is if municipalities form the strata and country regions the poststrata.



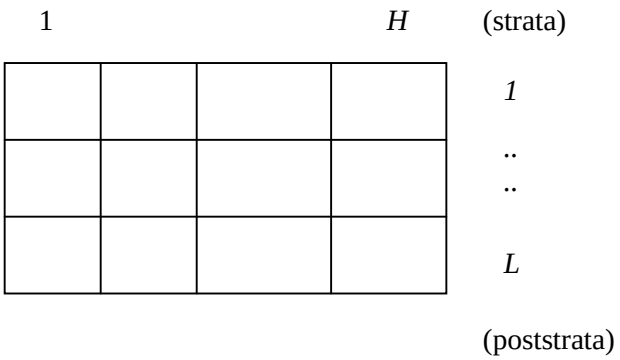
Poststrata are unions of strata

(c) The poststrata are parts of strata. An example is if regions form the strata and the municipalities the poststrata.



Poststrata are parts of strata

(d) The poststrata cross the strata. An example is if regions form the strata and the poststrata are determined by gender, or by age groups.



Poststrata cross strata

Note to (a)
Here the direct estimator (HT estimator) for stratified sampling (see Chapter 3) is appropriate. The associated variance also follows from Chapter 3.

Note to (b)
Poststratification adds nothing in this case, because the stratification has finer categories than the poststratification. $\hat{N}_l = N_l$ for all l, therefore the direct estimator for stratified sampling is again applicable.

Note to (c)
Poststratification is applied in stratum h of the simple random sample with, say L_h strata. Therefore, the variance of the poststratified estimator for simple random sampling applies within each stratum. Because the strata are sampled independently, these variances must be summed to obtain the variance of the poststratified estimator (with L poststrata).

Note to (d)
For each poststratum, the direct estimator for stratified sampling is used to estimate the total of the target variable and the population size per poststratum. Subsequently the means of the target variable in the poststrata can be estimated, which are then weighted with the actual population size of the poststrata, and summed; see Särndal et al. (1992, p. 268).

8.4 Example

A researcher is interested in the financial situation of students. He has access to university records with 5,000 students, from which he selects 500 through SRSWOR. One of the questions is about whether someone has paid work for more than 10 hours a week alongside their study. The sample consists of 300 male and 200 female students, of which 105 men and 50 women reported having a paid job of more than 10 hours a week. If the researcher were to use only the information from the sample, he would estimate that

$$\frac{5{,}000}{500} \times 155 = 1{,}550$$

students have paid work for more than 10 hours a week.

The university records also have information about student gender; in total there are $3{,}300$ male and $1{,}700$ female students registered at the university. It then follows that, relative to the student population, the male students are underrepresented, and the female students overrepresented, in the sample. Because we suspect that a student's gender is related to having a paid job of more than 10 hours a week, we would prefer a sample that is representative in terms of the numbers of male and female students. In the current situation in which sampling already has occurred, the estimator can only be adjusted for underrepresentation and overrepresentation with hindsight (*a posteriori*).

Therefore a better estimate of the number of students with a paid job is :

$$\frac{105}{300} \times 3{,}300 + \frac{50}{200} \times 1{,}700 = 1{,}580 .$$

8.5 Quality indicators

An important quality indicator for the poststratification estimator is the margin of uncertainty. As long as all the strata have sufficient elements there is little cause for concern. When, however, some strata have almost no observations, the variance of the poststratification estimator may increase. In order to avoid this, it is recommended to combine strata.

9. References

Camstra, A. and Nieuwenbroek, N.J. (2002), *Syllabus for the sampling theory course*. CBS-Academie, Voorburg.

Cochran, W.G. (1977), *Sampling Techniques*. John Wiley & Sons, New York.

Gouweleeuw, J.M. and Knottnerus, P. (2008), *Theme: Sampling Theory, Subtheme: Repeated Weighting*. Methods Series document, Statistics Netherlands, Voorburg [English translation from Dutch in 2012].

Hájek, J. (1964), Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* 35, 1491-1523.

Hansen, M.H. and Hurwitz, W.N. (1943), On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14, 333-362.

Horvitz, D.G. and Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663-685.

Knottnerus, P. (2003), *Sample Survey Theory: Some Pythagorean Perspectives*. Springer-Verlag, New York.

Knottnerus, P. (2011a), On the efficiency of randomized PPS sampling. *Survey Methodology*, 37, 95-102.

Knottnerus, P. (2011b), *Simple derivations of variance formulas in two-stage simple random sampling*, Discussion paper, Statistics Netherlands, The Hague.

Muilwijk, J., Snijders, T.A.B. and Moors, J.J.A. (1992), *Kanssteekproeven (Random sampling)*. Stenfert Kroese, Leiden.

Särndal, C.E, Swensson, B. and Wretman, J.H. (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Sen, A.R. (1953), On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* 5, 119-127.

Yates, F. and Grundy, P.M. (1953), Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* B, 15, 253-261.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Steekproeftheorie / Steekproefontwerpen en Ophoogmethoden				
1.0	26-02-2010	First Dutch version	Reinder Banning Astrea Camstra Paul Knottnerus	Kees van Berkel José Gouweleeuw Sander Scholtus Ilona Verburg
2.0	29-02-2012	Major changes in chapter 4	Reinder Banning Astrea Camstra Paul Knottnerus	Sander Scholtus
English version: Sampling theory / Sampling design and Estimation methods				
2.0E	29-02-2012	First English version	Reinder Banning Astrea Camstra Paul Knottnerus	