R&D Report
Research – Methods – Development 1998:3

Statistiska centralbyrån Statistics Sweden

On the Stratification of Highly Skewed Populations

Dan Hedlin

R&D Report
Research - Methods - Development 1998:3

On the Stratification of Highly Skewed Populations *(Dan Hedlin)*

This thesis was originally published as report No. B:41, 1998, of the Institute of Actuarial Mathematics and Mathematical Statistics and is reprinted here after kind permission by the University of Stockholm.
Statistics Sweden1998

Från trycket	Maj 1998
Ansvarig utgivare	Lars Lyberg
Producent	Statistiska centralbyrån, utvecklingsavdelningen

Förfrågningar	Dan Hedlin e-post dh1@socsci.soton.ac.uk

ISSN 0283-8680
Printed in Sweden

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1998:3. On the stratification of highly skewed population / Dan Hedlin.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1998-X101OP9803

R&D Report
Research – Methods – Development 1998:3

On the Stratification of Highly Skewed Populations

Dan Hedlin

Contents

Abstract

This paper discusses the problem of stratifying highly skewed populations, such as those encountered in many business surveys. We give conditions which must be satisfied for stratum boundaries to minimize the variance of the standard estimator of the population total. The paper appears to be the first one that deals with the combined problem of allocation and stratification in order to minimize the variance of the usual unbiased estimator, taking into account that the population is finite. The proof utilizes the Kuhn-Tucker Theorem. An iterative numerical method for practical application of the analytic results is proposed.

1 Introduction

Stratification is a widely used sample survey technique. The sampling frame is divided into strata and independent samples are drawn from the strata. There are a number of reasons for stratification. It is common in business surveys, for example, to use slightly different questionnaires for different subpopulations. Then it is natural to let each subpopulation be a stratum.

For the purpose of bringing estimator variances down there are two main types of beneficial stratifications:

(1) The survey designer forms strata as close to important study domains as possible, which will allow him to control sample sizes in strata and thereby the precision of domain statistics. We refer to these strata as ***pre-strata***.

(2) The survey designer forms ***homogenized strata***, which are obtained if important study variables vary less within strata than in the unstratified population (or in the pre-stratum). Such stratification is typically carried out as follows. Strata are formed by classifying the values of a stratification variable available in the sampling frame. Such a stratification increases the precision of the resulting statistics in cases where the stratification variable and the study variable are fairly strongly correlated. The effect of increasing precision is particularly strong when the study variables have highly skewed distributions, which is usually the case in business surveys. Then, typically, the stratum with the largest businesses is a self-representing stratum (also called "certainty stratum" or "take-all stratum") where all businesses are selected for observation.

In the sequel we will focus on objective (2).

Several problems have to be addressed when designing a stratified sample. The following list is taken from Särndal, Swensson and Wretman (1992) with some modification.

Construction of Strata
A1. Which stratification variable(s) is (are) to be used?
A2. How many strata should there be?
A3. How should strata be demarcated?

Choice of Sampling and Estimation Methods within Strata
B1. Sampling design for each stratum
B2. An estimator for each stratum
B3. The sample size for each stratum

Often the same type of design and estimator are used for all strata.

This paper focuses on questions A3 and B3 jointly, under the assumptions that the stratification variable is equal to the study variable. Section 2 gives an overview of the literature in this field and section 3 states conditions for stratum boundaries minimizing the variance. In section 4 an iterative numerical algorithm for univariate stratification of highly skewed populations is put forward. Applications are presented in section 5.

2 Overview of the Optimum Stratification Problem

There are a number of the problems associated with stratum construction in highly skewed populations. Sigman and Monsour (1995) give an overview.

The main problem to be considered in this report is: How should strata be demarcated? There is a considerable literature on optimal stratification for the usual unbiased estimator. In this section we give an overview of the most important references addressing this question in the context of homogenized strata. First we formulate some assumptions and approaches common in this literature.

This problem is usually treated as a "single-purpose" one in that just one parameter is considered. The following problem may be called ***the common optimum stratification problem***. This is the problem, with slight modifications in some cases, that most of the literature on this subject as well as this report discuss. Consider the standard estimator of the total of a study variable **y**:

$$\hat{t} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_k \,. \qquad (2.1)$$

The problem is to find the stratification that minimizes the variance of $\hat{t}$,

$$Var(\hat{t}) = \sum_{h=1}^{H} N_h^2 \frac{S_{yh}^2}{n_h}\left(1 - \frac{n_h}{N_h}\right), \qquad (2.2)$$

where N_h and n_h are the number of frame units in stratum h and the sample

size in stratum h, respectively, and S_{yh}^2 is the study variable variance in stratum h,

$$S_{yh}^2 = \frac{1}{N_h - 1}\sum_{k=1}^{n_h}(y_k - \bar{y}_h)^2 ,$$

where $\bar{y}_h$ is the study variable mean in stratum h. N_h, S_{yh}^2 and $\bar{y}_h$ are functions of the stratum boundaries. The number of strata, H, is fixed but arbitrary. A simple random sample is drawn from each stratum. The total sample size $n = n_1 + n_2 + \ldots + n_H$ is fixed.

It is well known that Neyman allocation gives the optimal sample sizes within strata, in the sense that the variance of $\hat{t}$ is minimized. If nothing else is stated the articles referred to in this section use Neyman allocation. Some authors, however, prefer other allocation schemes, thus deviating slightly from the optimum solution.

We will refer to a stratum where a frame unit is sampled with a probability less than one as a ***genuine sampling stratum*** as opposed to a ***certainty stratum*** where all frame units are included in the sample.

Either of the two following assumptions is widely used in connection with the common optimum stratification problem. Both are associated with the choice of stratification variable, problem A1. In this paper, we work under assumption A1.a only.

Assumption A1.a
The values of a single auxiliary variable are known and it is, although unrealistically, assumed that the values of study variables equal those of the stratification variable.

Assumption A1.b
The values of a single auxiliary variable are known and some stochastic relationship between the study variable and the stratification variable is assumed.

Many articles draw on the following approximation.

Approximation 1
The finite population correction is ignored when minimizing the estimator variance.

Comment: while the finite population correction is negligible in many practical applications, approximation 1 is crude if it used for a certainty stratum as this means replacing a zero variance with a strictly positive one for that stratum. Consequently, approximation 1 is questionable for highly skewed populations.

The approaches used to address the common optimum stratification

problem under assumption A1.a are organized in a tree-chart in Figure 2.1 and briefly summarized below.

Addressing the common optimum stratification problem Dalenius (1950) minimizes

$$v(\hat{t}) = \sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h}, \qquad (2.3)$$

where S_h^2 is the stratification variable variance in stratum h. Like in (2.2), both N_h and S_h^2 are functions of the stratum boundaries. The function $v(\hat{t})$ approximates (2.2) under Approximation 1 and Assumption A1.a. Let the H-1 stratum boundaries be denoted by $b_1, b_2, \ldots b_{H-1}$. They satisfy $b_1 < b_2 < \ldots < b_{H-1}$. Dalenius derives the following equations as a necessary condition for stratum boundaries minimizing (2.3):

$$\frac{S_h^2 + (b_h - \bar{x}_h)^2}{S_h} = \frac{S_{h+1}^2 + (b_h - \bar{x}_{h+1})^2}{S_{h+1}}, \quad h = 1, 2, \ldots H-1, \qquad (2.4)$$

where $\bar{x}_h$ is the mean of the stratification variable in stratum h. This condition is also discussed in Cochran (1977, section 5A.7). Schneeberger (1985) points out that a solution to (2.4) is not necessarily a local or global minimum to (2.3). The solution(s) may be one or several minima, maxima or saddle points.

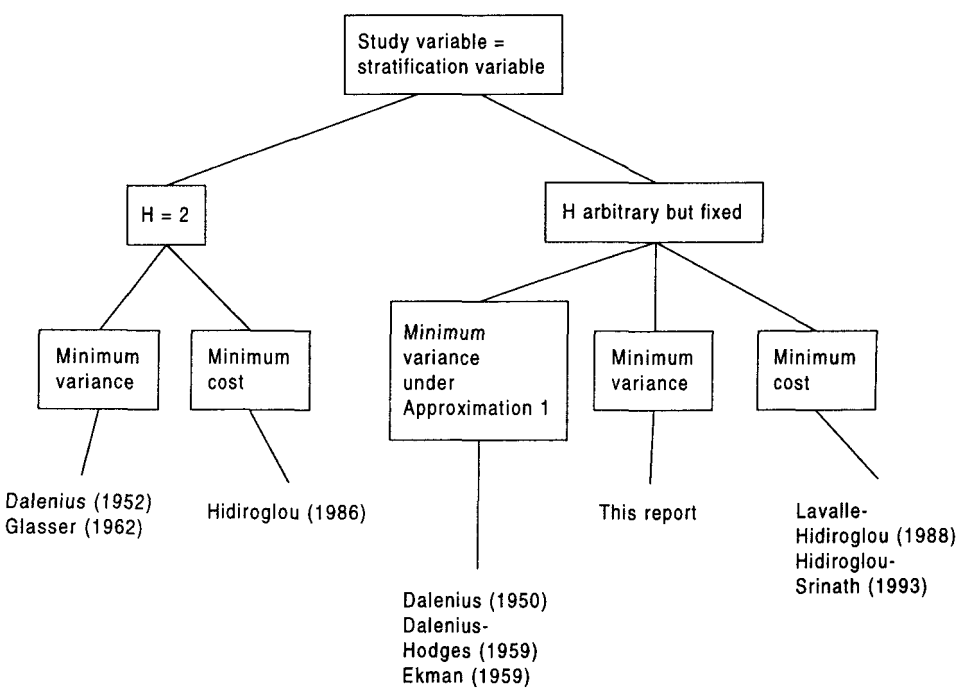


Figure 2.1 Approaches under assumption A1.a (stratification variable = study variable).

The Dalenius equations (2.4) are, however, ill adapted to practical computation. Consequently, a large number of approximate methods for constructing genuine sampling strata have been suggested. The most efficient from a precision-increasing point of view are presumably the Dalenius-Hodges rule ("the cum $\sqrt{f}$ rule") and the Ekman rule (Dalenius, Hodges (1959); Ekman (1959); Cochran (1961); Hess *et al.* (1966) and Murthy (1967)). A numerical procedure for the Ekman rule is presented in section 4 of this report. Both the Dalenius-Hodges rule and the Ekman rule give approximate solutions to the Dalenius equations (2.4). Since no stratum is allowed to be a certainty stratum these boundaries are not optimal for highly skewed populations.

Several authors have addressed the problem of finding the point where the far tail of a skewed distribution should be cut off to form a certainty stratum. All elements in the certainty stratum are included in the sample. None of the papers mentioned below, all of which consider designs that include a certainty stratum, draw on Approximation 1. Several solutions have been proposed to the special case of this problem when the population is divided into two strata only, one certainty stratum and one genuine sampling stratum. In this special case Dalenius (1952) suggests a condition for the certainty stratum. Glasser (1962) derives an exact result, as opposed to Dalenius who approximates the finite population with an infinite population. Nevertheless, the Glasser equation for stratum boundary b_1 is essentially equivalent with that of Dalenius:

$$(b_1 - \bar{x}_1)^2 = \frac{N_1 S_1^2}{n_1}, \qquad (2.5)$$

where index 1 refers to stratum 1, which is the genuine sampling stratum.

Whereas Dalenius and Glasser make estimator precision as good as possible under given total sample size, Hidiroglou (1986) minimizes sample size under prescribed estimator precision. Like Dalenius and Glasser he works under assumption A1.a. Moreover, he too limits the number of strata to two, thus also limiting the practical usefulness of the results. The approaches of Dalenius, Glasser and Hidiroglou are not easily generalized to a number of strata greater than two.

The approach of this paper differs from the ones mentioned above in that here we address the combined problem of finding the optimal allocation and optimal stratification when there are several genuine sampling strata and one certainty stratum. Condition (2.5) is a special case of the results of this report, whereas (2.4) is not. The reason for this is that approximation 1 is not invoked in this report.

Next, we briefly describe an algorithm by Lavallée and Hidiroglou (1988) and Hidiroglou and Srinath (1993), both of which provide stratum boundaries for one certainty stratum and several genuine sample strata. Both papers address the common optimum stratification problem under assumption A1.a,

however, like in Hidiroglou (1986) the sample size is minimized under a precision constraint rather than the other way around. Hidiroglou and Lavallée use a form of power allocation of the sample:

$$n_h = (n - n_H)\frac{(N_h \bar{x}_h)^p}{\sum_{h=1}^{H-1}(N_h \bar{x}_h)^p}$$

where index h indicates stratum h and n, N and $\bar{x}$ are sample sizes, frame sizes and the mean of the stratification variable, respectively (a description of power allocation is provided by Särndal, Swensson and Wretman (1992)). Strata 1, 2, ... H-1 are genuine sampling strata and stratum H is a certainty stratum with $n_H = N_H$. Hidiroglou and Srinath use a general allocation formula comprising several schemes, *e g* Neyman allocation. The stratum containing the largest units is predetermined to be a certainty stratum, the other strata to be genuine sampling strata. The iterative search algorithm finds the minimum of the objective function, which is the sample size viewed as a function of the stratum boundaries.

Sweet and Sigman (1995 b), Slanta and Krenzke (1996) and Detlefsen and Veum (1991) report on applications of the Lavallée and Hidiroglou algorithm. They found that the resulting boundaries depend on where the initial boundaries are set. Moreover, the convergence may be slow or non-existent. These findings made Detlefsen and Veum abandon the algorithm. Slanta and Krenzke studied the convergence of the algorithm applied to two populations. They propose ways of resolving difficulties with the algorithm, which they applied to one stratification variable of the Annual Capital Expenditures Survey at the US Bureau of the Census.

The approach of this report differs from that of Lavallée, Hidiroglou (1988) and Hidiroglou, Srinath (1993) in that here the estimator variance is treated as a function of the stratum boundaries and sample sizes within strata. The estimator variance is minimized under a fixed size of the total sample. The cited papers, however, solve a slightly different problem: the total sample size is seen as a function of the stratum boundaries. It is minimized under a predetermined estimator variance constraint. When the minimum size of the total sample is found, one part of the sample is allocated to the certainty stratum, and the remaining part is allocated to the genuine sampling strata according to a predetermined scheme, for example power allocation.

3 The Optimum Stratification Problem

When considering the common optimum stratification problem introduced in the previous section, we address the combined problem of A3 and B3 in section 1. The problem is now formulated in greater detail.

A sample is to be taken from the population $U = \{1,2,\ldots N\}$ with study variable $\mathbf{y} = (y_1, y_2, \ldots y_N)'$ in order to estimate the population total $t = y_1 + y_2 + \ldots y_N$. We disregard non-sampling errors, that is non-response, measurement and coverage errors.

For convenience we assume that every population unit corresponds to exactly one frame unit.

Stratified sampling with a predetermined number of strata, H, is employed. That is, the population is partitioned into H strata, denoted $A_1, A_2, \ldots A_H$. One stratification variable $\mathbf{x} = (x_1, x_2, \ldots x_N)'$ is assumed to be available with known values for every frame unit. The strata are determined by stratum boundary points $b_1, b_2, \ldots b_{H-1}$, $b_1 < b_2 < \ldots < b_{H-1}$:

$A_1 = \{u: x_u \le b_1\}$,

$A_h = \{u: b_{h-1} < x_u \le b_h\}$, $h = 2, 3, \ldots H-1$,

$A_H = \{u: b_{H-1} < x_u\}$

From each stratum a simple random sample without replacement is taken independently of samples of other strata.

The standard estimator of the total of the study variable is considered, see (2.1).

The total sample size n is predetermined whereas the sample size allocation to strata will be given by the solution to the optimization problem, that is, the sample sizes within strata, $n_1, n_2, \ldots n_H$ are treated as variables with fixed sum $n = \sum_{h=1}^{H} n_h$. A sample size in stratum h may or may not equalize the number of units in that stratum.

The variance of the standard estimator will be minimized (see (3.1) below).

Thus, the version of the ***common optimum stratification problem*** we will consider is as follows.

Find the values of $(\mathbf{n}, \mathbf{b}) = (n_1, n_2, \ldots, n_H, b_1, b_2, \ldots, b_{H-1})$ that minimizes the objective function (3.1) under the constraints (3.2) below.

$$Var(\hat{t}) = \sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right), \qquad (3.1)$$

where n_h is the sample size and S_h^2 is the study variable variance in stratum h. Under Assumption A1.a, S_h^2 is also the stratification variable variance. Since N_h and S_h^2 are functions of the stratum boundaries, (3.1) can be written:

$$w(\mathbf{n},\mathbf{b}) = \sum_{h=1}^{H} N_h^2(\mathbf{b}) \frac{S_h^2(\mathbf{b})}{n_h}\left(1 - \frac{n_h}{N_h(\mathbf{b})}\right),$$

with $w(\mathbf{n},\mathbf{b}) = Var(\hat{t})$.

In (3.2) the symbol ≡ indicates "definition":

$$\begin{cases} g_h(\mathbf{n},\mathbf{b}) \equiv n_h - N_h \le 0\ , \quad h = 1, 2 \ldots H \\ g_{H+1}(\mathbf{n},\mathbf{b}) \equiv \sum_{h=1}^{H} n_h - n \le 0 \end{cases} \qquad (3.2)$$

Note that these constraints allow any stratum to be a certainty stratum. As a useful special case the constraints will be further restricted. The constraints 1, 2, ... H in (3.3) state that strata 1, 2 ...H-1 are genuine sampling strata whereas stratum H is a certainty stratum. The constraint H+1 states that all of the available sample should be used, which in practise is no restriction of (3.2).

$$\begin{cases} g_h(\mathbf{n},\mathbf{b}) \equiv n_h - N_h < 0\ , \quad h = 1, 2 \ldots H-1 \\ g_H(\mathbf{n},\mathbf{b}) \equiv n_H - N_H = 0 \\ g_{H+1}(\mathbf{n},\mathbf{b}) \equiv \sum_{h=1}^{H} n_h - n = 0 \end{cases} \qquad (3.3)$$

3.1 A solution

We introduce a framework that will allow us to apply optimization theory for continuous functions. The framework can either be seen as a superpopulation approach or simply as an approximation approach. We start with the first one.

The finite population U is regarded as N independent realizations of a stochastic variable X with density function $f(x)$. Let x_1 and x_N be a priori known lower and upper bounds for the values of X. In practise, x_1 is often zero and x_N a value larger than any value that actually could occur. Thus, $f(x)$ is concentrated on (x_1, x_N). This interval is stratified into H intervals with variable boundaries, H being a fixed integer. Let stratum h consist of the units with x-values in the interval (b_{h-1}, b_h). Set $b_0 = x_1$ and $b_H = x_N$. We will need three properties of the strata: probability, mean and variance. Let P_h denote the probability that X falls in stratum h:

$$P_h = \int_{b_{h-1}}^{b_h} f(t)dt \qquad (3.4)$$

The mean and variance of X are denoted by μ and σ^2, respectively. The corresponding parameters of stratum h are the conditional mean and variance of X given $X \in (b_{h-1}, b_h)$:

$$\mu_h = \frac{1}{P_h} \int_{b_{h-1}}^{b_h} tf(t)dt \qquad (3.5)$$

$$\sigma_h^2 = \frac{1}{P_h} \int_{b_{h-1}}^{b_h} (t - \mu_h)^2 f(t)dt \qquad (3.6)$$

In each stratum, N_h x-values are generated from $f(x)$, where $\sum_{h=1}^{H} N_h = N$.

Let $\mathcal{E}$ denote expectation with respect to superpopulation randomness. An unbiased estimator of σ_h^2 is $S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{n_h} (y_k - \bar{y}_h)^2$, that is, $\mathcal{E}(S_h^2) = \sigma_h^2$.

From the finite population a sample is randomly selected without replacement, using the same stratum boundaries as those partitioning the superpopulation. From (3.1) we obtain

$$\mathcal{E}Var(\hat{t}) = \sum_{h=1}^{H} N_h^2 \frac{\sigma_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \qquad (3.7)$$

Note that the right-hand side of (3.7) can be seen as a function $\phi(\mathbf{n}, \mathbf{b}) = \phi(n_1, n_2, \ldots n_H, b_1, b_2, \ldots b_{H-1})$, that is, a function of stratum sample sizes and stratum boundaries. We regard $N_h = NP_h$ as a continuous function of b_{h-1} and b_h. We also treat $n_1, n_2, \ldots n_H$ as continuous variables. We have

$$\phi(\mathbf{n}, \mathbf{b}) = \sum_{h=1}^{H} N_h^2(\mathbf{b}) \frac{\sigma_h^2(\mathbf{b})}{n_h} \left(1 - \frac{n_h}{N_h(\mathbf{b})}\right) \qquad (3.8)$$

For simplicity, we will in the sequel drop the argument $\mathbf{b}$ in the functions $N_h(\mathbf{b})$, $\sigma_h(\mathbf{b})$ and other functions of the stratum boundaries.

The approximation approach is to work under the assumption that the discrete distribution of $\mathbf{x}$ can sufficiently well be approximated by a continuous distribution with density $f(x)$. The integer N_h and the finite population variance S_h^2 are assumed approximately equal to NP_h and σ_h^2, respectively. We will denote NP_h by $N_h(\mathbf{b})$ or just N_h. Thus N_h is regarded as a continuous functions of the stratum boundaries. The objective function to be minimized is again (3.8).

3.1.1 The main result

Theorem 1

Suppose strata 1, 2, ... H-1 are predetermined to be genuine sampling strata and stratum H is predetermined to be a certainty stratum. Then, if $f(x) > 0$, $x \in (x_1, x_N)$, a necessary condition for a local minimum of (3.8) with respect to stratum sample sizes and stratum boundaries under constraints (3.3) is the system of equations (3.9), (3.10) and (3.11) below.

Conditions for stratum sample sizes:

$$n_h = (n - N_H)\frac{N_h\sigma_h}{\sum_{h=1}^{H-1} N_h\sigma_h}, \quad h = 1, 2, 3 \ldots H\text{-}1. \tag{3.9}$$

Conditions for the boundaries $b_1, b_2, \ldots b_{H-2}$ of the genuine sampling strata:

$$\frac{(b_h - \mu_h)^2\left(1 - \frac{n_h}{N_h}\right) + \sigma_h^2}{\sigma_h} = \frac{(b_h - \mu_{h+1})^2\left(1 - \frac{n_{h+1}}{N_{h+1}}\right) + \sigma_{h+1}^2}{\sigma_{h+1}}, \tag{3.10}$$

$h = 1, 2, 3 \ldots H\text{-}2$.

Condition for the boundary b_{H-1} of the certainty stratum:

$$(b_{H-1} - \mu_{H-1})^2 = \frac{N_{H-1}}{n_{H-1}}\sigma_{H-1}^2 \,. \tag{3.11}$$

•

Remark 1. This report does not attempt to provide any sufficient condition for a local minimum.

Remark 2. Equation (3.9) is Neyman allocation when stratum H is a certainty stratum (see, for example, Cochran (1997, section 5.8)).

Remark 3. Equation (3.10) is a necessary condition for stratum boundaries associated with genuine sampling strata. Still, it differs from that of Dalenius, compare (2.4). The reason is that Dalenius uses Approximation 1. "Finite population correction factors" of the type $1 - \frac{n}{N}$ are often seen in survey sampling theory. Interestingly, this problem is no exception: the proper finite population result is obtained by inserting finite population corrections at appropriate places in the corresponding formula valid for an infinite population.

Remark 4. For $H = 2$, equation (3.11) is equivalent to the condition of Glasser, see (2.5).

Remark 5. When applying Theorem 1 in a practical situation, the unknown superpopulation parameters μ_h and σ_h^2 must be estimated or guessed by the corresponding parameters of the finite population. Moreover, in a practical situation the values of n_h and N_h have to be rounded to nearest integer.

3.1.2 Auxiliary results

We will use the Kuhn-Tucker Theorem which provides necessary conditions for a local optimum of a function given certain constraints. For convenience the theorem is restated in Proposition 1. Definition 1 introduces the concept of a regular point that will be needed in Proposition 1. See for example Luenberger (1973) for a more detailed account.

Definition 1
Let $(\mathbf{n},\mathbf{b})$ be a point satisfying the constraints

$d_i(\mathbf{n},\mathbf{b}) = 0,\ i = 1, 2,\ldots I,$

and

$e_j(\mathbf{n},\mathbf{b}) \le 0,\ j = 1, 2\ldots J,$

and let $\mathcal{H}$ be the set of indices j for which $e_j(\mathbf{n},\mathbf{b}) = 0$. Then $(\mathbf{n},\mathbf{b})$ is a *regular point* of the constraints, if the gradient vectors

$$\nabla d_i(\mathbf{n},\mathbf{b}) = \left(\frac{\partial d_i(\mathbf{n},\mathbf{b})}{\partial n_1}, \ldots \frac{\partial d_i(\mathbf{n},\mathbf{b})}{\partial n_H}, \frac{\partial d_i(\mathbf{n},\mathbf{b})}{\partial b_1}, \ldots \frac{\partial d_i(\mathbf{n},\mathbf{b})}{\partial b_{H-1}} \right)', \ i = 1,2,\ldots I, \text{ and}$$

$$\nabla e_j(\mathbf{n},\mathbf{b}) = \left(\frac{\partial e_j(\mathbf{n},\mathbf{b})}{\partial n_1}, \ldots \frac{\partial e_j(\mathbf{n},\mathbf{b})}{\partial n_H}, \frac{\partial e_j(\mathbf{n},\mathbf{b})}{\partial b_1}, \ldots \frac{\partial e_j(\mathbf{n},\mathbf{b})}{\partial b_{H-1}} \right)', \ j \in \mathcal{H},$$

are linearly independent.
•

Note that the constraints (3.2) are of the form $e_j(\mathbf{n},\mathbf{b}) \le 0,\ j = 1, 2\ldots H+1$ while the two last constraints of (3.3) can be written $d_i(\mathbf{n},\mathbf{b}) = 0$, $i = H, H+1$.

Proposition 1
Denote the gradient vector of $\phi(\mathbf{n},\mathbf{b})$ by

$$\nabla\phi(\mathbf{n},\mathbf{b}) = \left(\frac{\partial\phi}{\partial n_1}, \frac{\partial\phi}{\partial n_2}, \ldots, \frac{\partial\phi}{\partial n_H}, \frac{\partial\phi}{\partial b_1}, \frac{\partial\phi}{\partial b_2}, \ldots, \frac{\partial\phi}{\partial b_{H-1}} \right)'.$$

Let $(\mathbf{n}^*,\mathbf{b}^*)$ be a local minimum for the problem of mimizing $\phi(\mathbf{n},\mathbf{b})$ subject to the constraints

$d_i(\mathbf{n},\mathbf{b}) = 0,\ i = 1,2,\ldots I,$ and $e_j(\mathbf{n},\mathbf{b}) \le 0,\ j = 1, 2\ldots J,$

and suppose $(\mathbf{n}^*, \mathbf{b}^*)$ is a regular point of the constraints. Then there is a vector $\boldsymbol{\nu} \in \mathbf{R}^I$, with I real-valued components, and a vector $\boldsymbol{\lambda} \in \mathbf{R}^J$ with $\boldsymbol{\lambda} \geq \mathbf{0}$ such that

$$\nabla\phi(\mathbf{n}^*, \mathbf{b}^*) + \sum_{i=1}^{I} \nu_i \nabla d_i(\mathbf{n}^*, \mathbf{b}^*) + \sum_{j=1}^{J} \lambda_j \nabla e_j(\mathbf{n}^*, \mathbf{b}^*) = \mathbf{0} \qquad (3.12)$$

$$\lambda_j e_j(\mathbf{n}^*, \mathbf{b}^*) = 0, \; j = 1, 2 \ldots J \qquad (3.13)$$

•

We prepare the proof of Theorem 1 by calculating the partial derivatives of the functions $\phi(\mathbf{n}, \mathbf{b})$ and $g_\nu(\mathbf{n}, \mathbf{b})$ in (3.8), (3.2) and (3.3). The derivatives will be needed in Lemma 2 below.

P_h, μ_h and σ_h^2 are all defined as functions of b_{h-1} and b_h on $\{b_0, b_H\}$ (see (3.4) (3.5) and (3.6)). Let in this context $N_h = N \int_{b_{h-1}}^{b_h} f(t)dt$. As $f(x)$ is assumed continuous, P_h, N_h, μ_h and σ_h^2 are continuous and differentiable. This makes (3.8) a differentiable function on the set defined by (3.2). The constraints $g_\nu(\mathbf{n}, \mathbf{b})$, $\nu = 1, 2 \ldots H+1$, are differentiable functions, too.

The partial derivatives of $g_\nu(\mathbf{n}, \mathbf{b})$, $\nu = 1, 2, \ldots H+1$, are given in Table 3.1 and Table 3.2. As $N_h = N \int_{b_{h-1}}^{b_h} f(t)dt$ the functions $g_h(\mathbf{n}, \mathbf{b}) = n_h - N_h$ are constant in all dimensions except n_h, b_{h-1} and b_h.

The partial derivatives $\frac{\partial g_\nu}{\partial n_h}$, $\nu = 1, 2, \ldots H$ and $h = 1, 2, \ldots H$ form a diagonal matrix with unity along the diagonal (Table 3.1). Furthermore, $\frac{\partial g_{H+1}}{\partial n_h} = 1, \; \forall h$.

	n_1	...	n_h	...	n_H
$g_1(\mathbf{n},\mathbf{b})$	1	...	0	...	0
⋮	⋮		⋮		⋮
$g_h(\mathbf{n},\mathbf{b})$	0	...	1	...	0
⋮	⋮		⋮		⋮
$g_H(\mathbf{n},\mathbf{b})$	0	...	0	...	1
$g_{H+1}(\mathbf{n},\mathbf{b})$	1	...	1	...	1

Table 3.1. The partial derivatives of the constraints with respect to $n_1, n_2, \ldots n_H$. The entry in the *ith* row and the *jth* column is $\frac{\partial g_i}{\partial n_j}$.

To obtain the derivatives $\frac{\partial g_v}{\partial b_h}$, $v = 1,2,\ldots H$ and $h = 1,2,\ldots H-1$, note that

$$\frac{\partial N_l}{\partial b_h} = \begin{cases} Nf(b_h), & \text{if } l = h \\ -Nf(b_h), & \text{if } l = h+1 \\ 0, & \text{else} \end{cases} \qquad (3.14)$$

and that $g_l(\mathbf{n},\mathbf{b}) = n_l - N_l$, which gives the first H rows of Table 3.2.
It is readily seen that

$$\frac{\partial g_{H+1}}{\partial b_h} = 0, \forall h,$$

which gives the last row of Table 3.2.

	b_1	...	b_{h-1}	b_h	b_{h+1}	...	b_{H-1}
$g_1(\mathbf{n},\mathbf{b})$	$-Nf(b_1)$	...	0	0	0	...	0
$g_2(\mathbf{n},\mathbf{b})$	$Nf(b_1)$	...	0	0	0	...	⋮
⋮	⋮		⋮	⋮	⋮	...	⋮
$g_{h-1}(\mathbf{n},\mathbf{b})$	0	...	$-Nf(b_{h-1})$	0	0	...	0
$g_h(\mathbf{n},\mathbf{b})$	0	...	$Nf(b_{h-1})$	$-Nf(b_h)$	0	...	0
$g_{h+1}(\mathbf{n},\mathbf{b})$	0	...	0	$Nf(b_h)$	$-Nf(b_{h+1})$	...	0
⋮	⋮		⋮	⋮	⋮	...	⋮
$g_{H-1}(\mathbf{n},\mathbf{b})$	0	...	0	0	0	...	$-Nf(b_{H-1})$
$g_H(\mathbf{n},\mathbf{b})$	0	...	0	0	0	...	$Nf(b_{H-1})$
$g_{H+1}(\mathbf{n},\mathbf{b})$	0	...	0	0	0		0

Table 3.2. The partial derivatives of the constraints with respect to $b_1, b_2, \ldots b_{H-1}$. The entry in the *ith* row and the *jth* column is $\frac{\partial g_i}{\partial b_j}$.

We rewrite the objective function in a form convenient for taking derivatives:

$$\phi(\mathbf{n},\mathbf{b}) = \sum_{h=1}^{H}\left(\frac{N_h}{n_h} - 1\right)N_h\sigma_h^2 \qquad (3.15)$$

Now, the partial derivatives of $\phi(\mathbf{n},\mathbf{b})$ with respect to the components of **n** are:

$$\frac{\partial \phi}{\partial n_h} = -\frac{N_h^2\sigma_h^2}{n_h^2}, \quad h = 1,2\ldots H \qquad (3.16)$$

To obtain partial derivatives of $\phi(\mathbf{n},\mathbf{b})$ with respect to the components of **b,** we note that σ_h^2 is constant in all dimensions except b_{h-1} and b_h. We restate an application of the chain rule called the General Leibnitz Rule (see for example Protter and Morrey (1977)).

Proposition 2. Suppose $\varphi(x,t)$ and $\dfrac{\partial \varphi}{\partial x}$ are continuous on $\{(x,t)\colon a \le x \le b,\ c \le t \le d\}$ and $h_1(x)$ are $h_2(x)$ functions on $[a,b]$ with continuous derivative and range in $[c,d]$. Let

$$\Phi(x) = \int_{h_2(x)}^{h_1(x)} \varphi(x,t)dt .$$

Then $\Phi'(x) = \varphi(x,h_1(x))h_1'(x) - \varphi(x,h_2(x))h_2'(x) + \int_{h_2(x)}^{h_1(x)} \dfrac{\partial \varphi(x,t)}{\partial x} dt .$

•

The integrand in $N_h\sigma_h^2 = \int_{b_{h-1}}^{b_h} (t-\mu_h)^2 f(t)dt$, $h = 1, 2, \ldots H$, is a function of t and μ_h, where μ_h is a function of b_{h-1} and b_h.
Hence, according to Proposition 2,

$$\frac{\partial N_h\sigma_h^2}{\partial b_h} = N(b_h-\mu_h)^2 f(b_h) + 2\int_{b_{h-1}}^{b_h} (t-\mu_h)\frac{\partial \mu_h}{\partial b_h} f(t)dt , \qquad (3.17)$$

$h = 1, 2, \ldots H\text{-}1$,

Since $\int_{b_{h-1}}^{b_h} (t-\mu_h)f(t)dt = 0$ we have

$$\frac{\partial N_h\sigma_h^2}{\partial b_h} = N(b_h-\mu_h)^2 f(b_h), \quad h = 1, 2, \ldots H\text{-}1, \qquad (3.18)$$

Analogously,

$$\frac{\partial N_h\sigma_h^2}{\partial b_{h-1}} = -N(b_{h-1}-\mu_h)^2 f(b_{h-1}), \quad h = 2, 3, \ldots H,$$

and, replacing index h with $h+1$,

$$\frac{\partial N_{h+1}\sigma_{h+1}^2}{\partial b_h} = -N(b_h-\mu_{h+1})^2 f(b_h), \quad h = 1, 2, \ldots H\text{-}1. \qquad (3.19)$$

Now, to find the derivatives of (3.15), formulae (3.14), (3.18) and (3.19) give

$$\frac{\partial \phi}{\partial b_h} = Nf(b_h)\left[(b_h-\mu_h)^2\left(\frac{N_h}{n_h}-1\right) + \frac{N_h\sigma_h^2}{n_h} - (b_h-\mu_{h+1})^2\left(\frac{N_{h+1}}{n_{h+1}}-1\right) - \frac{N_{h+1}\sigma_{h+1}^2}{n_{h+1}}\right] \qquad (3.20)$$

$h = 1, 2 \ldots H-1$.

Lemma 1

Suppose $f(x) > 0$ on (x_1, x_N). Then the gradients of the constraints are linearly independent in all feasible points, that is, all points $(\mathbf{n},\mathbf{b})$ satisfying (3.2). Thus, all points $(\mathbf{n},\mathbf{b})$ are regular points for the constraints $g_v(\mathbf{n},\mathbf{b})$ in (3.2).

•

Proof: The set of vectors $\nabla g_1(\mathbf{n},\mathbf{b}), \nabla g_2(\mathbf{n},\mathbf{b}), \ldots \nabla g_{H+1}(\mathbf{n},\mathbf{b})$ are linearly independent if there no scalars $\alpha_1, \alpha_2, \ldots \alpha_{H+1}$, except for $\alpha_1 = \alpha_2 = \ldots = \alpha_{H+1} = 0$, such that $\sum_{h=1}^{H+1} \alpha_h \nabla g_h(\mathbf{n},\mathbf{b}) = \mathbf{0}$. Thus, try to find $\alpha_1, \alpha_2, \ldots \alpha_{H+1}$ that satisfy

$$
\begin{array}{lll}
\alpha_1 & (1,0,0,\ldots,0, & -Nf(b_1),0,0,\ldots\ldots\ldots 0)' \\
+\ \alpha_2 & (0,1,0,\ldots,0, & Nf(b_1),-Nf(b_2),0,\ldots 0)' \\
+\ \vdots & \vdots & \vdots \\
+\ \alpha_H & (0,0,0,\ldots,1, & 0,0,0,\ldots\ldots\ldots, Nf(b_{H-1}))' \\
+\ \alpha_{H+1} & (1,\ 1,\ 1,\ldots,1, & 0,0,0,\ldots\ldots\ldots\ldots\ldots,0)' \\
= \mathbf{0} & &
\end{array}
$$

Then $\alpha_h = -\alpha_{H+1}$ and $Nf(b_h)(\alpha_{h+1} - \alpha_h) = 0$, $h = 1, 2, \ldots H-1$, and $Nf(b_{H-1})\alpha_H = 0$. Under the presumption that $f(x) > 0$ all $\alpha_h = 0$, $h = 1,2,\ldots H$, and we must have $\alpha_{H+1} = 0$. Thus there is no vector $\boldsymbol{\alpha}$, except for the null vector, that satisfies $\sum_{h=1}^{H} \alpha_h \nabla g_h = \mathbf{0}$.

Lemma 2

Suppose $f(x) > 0$ on (x_1, x_N). Consider a stratification and allocation that give a local minimum of (3.8) under constraints (3.2) with at least two genuine sampling strata. Then the system of equations (3.21) and (3.22) below are satisfied.

$$\frac{n_h}{n_l} = \frac{N_h \sigma_h}{N_l \sigma_l} \quad \forall\, h \text{ and } l \text{ where } n_h < N_h \text{ and } n_l < N_l \qquad (3.21)$$

$$(b_h - \mu_h)^2 \left(\frac{N_h}{n_h} - 1\right) + \frac{N_h}{n_h}\sigma_h^2 - \qquad (3.22)$$

$$(b_h - \mu_{h+1})^2 \left(\frac{N_{h+1}}{n_{h+1}} - 1\right) - \frac{N_{h+1}}{n_{h+1}}\sigma_{h+1}^2 + (\lambda_{h+1} - \lambda_h) = 0,$$

$h = 1,2\ldots H-1$,

for some non-negative real numbers λ_h and λ_{h+1}.

•

Proof: Lemma 1 justifies the use of Proposition 1. The left hand side of (3.12) is a vector. Consider the H first components, which are associated with the stratum sample sizes n_h. As seen in Table 3.1, $\frac{\partial g_v}{\partial n_h} = 1$, $v \leq H$, if and only if $v = h$ and $\frac{\partial g_{H+1}}{\partial n_h} = 1$ for all h. Then (3.16) inserted into (3.12) gives the following set of equations:

$$\lambda_h + \lambda_{H+1} = \left(\frac{N_h \sigma_h}{n_h} \right)^2, \quad h = 1,2 \ldots H. \tag{3.23}$$

By hypothesis there are at least two strata from which less than all units are sampled. Denote the indices of two such strata by s and t. The constraint associated with stratum s is $g_s(\mathbf{n},\mathbf{b}) < n_s - N_s$, and analogously for stratum t. Now (3.13) implies that $\lambda_s = 0$ and $\lambda_t = 0$. From (3.23) we conclude that

$$\lambda_{H+1} = \left(\frac{N_s \sigma_s}{n_s} \right)^2, \forall\ s \text{ where } n_s < N_s. \tag{3.24}$$

As λ_{H+1} is a constant,

$$\left(\frac{N_s \sigma_s}{n_s} \right)^2 = \left(\frac{N_t \sigma_t}{n_t} \right)^2, \tag{3.25}$$

$\forall\ s$ and t where $n_s < N_s$ and $n_t < N_t$.

Thus (3.21) is proven.

Now, turning to the condition (3.22) for one particular stratum boundary, b_h, where $h = 1, 2, \ldots H\text{-}1$, we need to know which of the multipliers $\lambda_1, \lambda_2, \ldots \lambda_{H+1}$ that vanish, if any. In Table 3.2 we see that $\frac{\partial g_v}{\partial b_h} = 0$ for all combinations of v and h, $v = 1,2 \ldots H$, except $h = v$ and $h = v + 1$ and that $\frac{\partial g_{H+1}}{\partial b_h} = 0$. That is, the multipliers are all zero except λ_h and λ_{h+1}. For a particular h, the non-vanishing values of $\frac{\partial g_v}{\partial b_h}$ are $Nf(b_h)$ and $-Nf(b_h)$, found in column b_h of Table 3.2. From (3.12) and (3.20) we obtain

$$Nf(b_h)\left[(b_h - \mu_h)^2 \left(\frac{N_h}{n_h} - 1 \right) + \frac{N_h \sigma_h^2}{n_h} - (b_h - \mu_{h+1})^2 \left(\frac{N_{h+1}}{n_{h+1}} - 1 \right) + \frac{N_{h+1} \sigma_{h+1}^2}{n_{h+1}} \right] \tag{3.26}$$

$$+ Nf(b_h)(\lambda_{h+1} - \lambda_h) = 0,$$

$$h = 1,2, \ldots H - 1$$

By hypothesis $f(b_h) \neq 0$ and (3.22) is proven.

3.1.3 Proof of the main result

Proof of Theorem 1

Lemma 2 gives an optimum under constraints (3.2). Now we are seeking an optimum under constraints (3.3). If $H = 2$, (3.9) is trivial. If $H \geq 3$ equation (3.21) in Lemma 2 can easily be restated as $n_h = n' \dfrac{N_h \sigma_h}{\sum_{A_h'} N_h \sigma_h}$ where n' is the sum of the sample sizes in the genuine sampling strata, denoted by A_h'. Equation (3.9) follows readily.

To prove (3.10) consider first (3.22) with $h = 1, 2, \ldots H\text{-}2$. Note that as constraints 1, 2 ... H-1 are predetermined to be satisfied with strict inequality, they are according to (3.13) in Proposition 1 simply dropped from (3.12). Hence, λ_h and λ_{h+1} in (3.22) both vanish. Thus, we obtain

$$(b_h - \mu_h)^2 \left(\frac{N_h}{n_h} - 1 \right) + \frac{N_h}{n_h} \sigma_h^2 - (b_h - \mu_{h+1})^2 \left(\frac{N_{h+1}}{n_{h+1}} - 1 \right) + \frac{N_{h+1}}{n_{h+1}} \sigma_{h+1}^2 = 0, \quad h = 1,2 \ldots H-2 \qquad (3.27)$$

Extract N_h/n_h and N_{h+1}/n_{h+1} from the left and right hand side, respectively, and insert (3.9) into (3.27) and (3.10) is obtained.

Consider now (3.22) with $h = H$-1. The multiplier λ_{H-1} vanishes, whereas λ_H is derived as follows. Proceeding as in the proof of Lemma 2 we have

$$\lambda_H + \lambda_{H+1} = \left(\frac{N_H \sigma_H}{n_H} \right)^2 \qquad (3.28)$$

and

$$\lambda_{H+1} = \left(\frac{N_{H-1} \sigma_{H-1}}{n_{H-1}} \right)^2 . \qquad (3.29)$$

Since $n_H = N_H$ we have

$$\lambda_H = \sigma_H^2 - \left(\frac{N_{H-1} \sigma_{H-1}}{n_{H-1}} \right)^2 . \qquad (3.30)$$

Insert (3.30) into (3.22) where $h = H$-1, $\lambda_{H-1} = 0$ and $n_H = N_H$:

$$(b_{H-1}-\mu_{H-1})^2\left(\frac{N_{H-1}}{n_{H-1}}-1\right)=\left(\frac{N_{H-1}}{n_{H-1}}\sigma_{H-1}\right)^2-\frac{N_{H-1}}{n_{H-1}}\sigma_{H-1}^2 \qquad (3.31)$$

Divide both sides by $\frac{N_{H-1}}{n_{H-1}}-1$, which by (3.3) is greater than zero, and we obtain

$$(b_{H-1}-\mu_{H-1})^2=\frac{N_{H-1}}{n_{H-1}}\sigma_{H-1}^2.$$

Thus, (3.11) is proven.

Remark 6. There is some ambiguity in the representation of λ_H in (3.30) as we could have made another choice of s in (3.29). Hence, other possibilities are $\lambda_H=\sigma_H^2-\left(\frac{N_h\sigma_h}{n_h}\right)^2$, $h=1,2\ldots H-2$. Any of these would lead to conditions for optimum equivalent to (3.11), although less appealing.

3.1.4 The special condition for certainty strata

What is the difference between (3.10) and (3.11) in Theorem 1? Let's put it in this way. Suppose you for some reason or other stratify by using a method equivalent or close to (3.10), like the cum $\sqrt{f}$ rule, using this rule for all strata. Then you allocate the sample and end up with $n_H=N_H$, what have you done? This approach corresponds to a priori letting $\lambda_{H-1}=\lambda_H=0$ in (3.22) in Lemma 2, which with $h=H$-1 becomes:

$$(b_{H-1}-\mu_{H-1})^2\left(\frac{N_{H-1}}{n_{H-1}}-1\right)+\frac{N_{H-1}}{n_{H-1}}\sigma_{H-1}^2- \qquad (3.32)$$

$$(b_{H-1}-\mu_H)^2\left(\frac{N_H}{n_H}-1\right)-\frac{N_H}{n_H}\sigma_H^2=0,$$

Compare this with an approach where strata 1, 2, ... H-1 are predetermined genuine sampling strata and stratum H may or may not be a certainty stratum. Then, by Proposition 1, $\lambda_{H-1}=0$ and $\lambda_H\geq 0$ and (3.22) for $h=H$-1 is

$$(b_{H-1}-\mu_{H-1})^2\left(\frac{N_{H-1}}{n_{H-1}}-1\right)+\frac{N_{H-1}}{n_{H-1}}\sigma_{H-1}^2- \qquad (3.33)$$

$$(b_{H-1}-\mu_H)^2\left(\frac{N_H}{n_H}-1\right)-\frac{N_H}{n_H}\sigma_H^2+\lambda_H=0,$$

The absence of λ_H in (3.32) makes either stratum H too narrow or stratum H-1 too wide.

Lavallée, Hidiroglou (1988) applied the Dalenius-Hodges rule and their own method (see section 2) to two highly skewed populations. The Dalenius-Hodges rule resulted in a much narrower certainty stratum for both populations, for all coefficients of variations requirements and for all choices of parameter p in power allocation. Their intention by using the Dalenius-Hodges rule to determine the size of a certainty stratum, despite the fact that the Dalenius-Hodges rule is derived under Approximation 1, is to "caution against its blind use in the context of highly skewed populations" (Lavallée, Hidiroglou (1988, p. 40)).

4 A Numerical Procedure for Stratification

In this section a numerical procedure for the optimum stratification problem is presented. The situation we have in mind is as follows.

There is a frame where all units have values for an auxiliary variable $\mathbf{x} = \{x_1, x_2, \ldots x_N\}$. The distribution of the values of $\mathbf{x}$ is assumed highly skewed, which calls for a certainty stratum containing the largest units. All other strata are genuine sampling strata. The strata, denoted by $A_1, A_2, \ldots A_H$, are to be determined by stratum boundary points that yield a solution to the optimum stratification problem under constraints (3.3). The solution is given by conditions (3.10) and (3.11) in Theorem 1. Once the strata are determined, the sample is allocated to strata according to condition (3.9) in Theorem 1.

However, we shall be satisfied with an approximate solution to (3.10). In doing so, we rely on the experience that the estimator variance is flat around the optimal stratum boundaries $b_1, b_2, \ldots b_{H-2}$ for genuine sampling strata. This is further discussed in section 5.

Against this background we use Approximation 1 for genuine sampling strata which simplifies (3.10) to the Dalenius equations (2.4). As already mentioned, a number of easy-to-use approximate methods have been proposed to solve (2.4). We shall be concerned with the one proposed by Ekman (1959). The degree of approximation to an exact solution of (2.4) is discussed by Ekman. References of some empirical studies are given in section 2 "Overview of the optimum stratification problem". As Ekman notes, his rule is "substantially equivalent" to the widely used Dalenius-Hodges rule (Ekman, 1959, pp 223-224).

4.1 The stratification algorithm

We aim at boundaries for genuine sampling strata given by a solution to the Dalenius equations (2.4) and at a boundary for the certainty stratum given by condition (3.11) in Theorem 1. The set of equations (2.4) requires a numerical method to be solved. Below we state the *extended Ekman rule* and propose an algorithm for it. Moreover, we propose an algorithm for the combined problem of using the extended Ekman rule for the genuine sampling strata and the condition (3.11) for the certainty stratum. This algorithm is now described.

The stratification algorithm

The algorithm will go through possible values of the size of the certainty stratum, from $N_H = 0$ to $N_H = n$, and for each value the other stratum boundaries are determined by the extended Ekman rule.

1. Let $N_H = 0$.
2. Stratify the frame with stratum H removed into H-1 strata with the extended Ekman rule. Apply the numerical procedure for stratification by the extended Ekman rule shown below.
3. Calculate the left and right hand side, respectively, of equation (3.11) in Theorem 1. Save the values in a file.
4. Transfer the K units with the largest x-values from stratum H-1 to stratum H (where K is a small positive integer, for example, $K = 1$).
5. Repeat steps 2-4 until $N_H = n$.
6. Plot the values from step 3 against N_H. You will see two curves which cross at 0, 1, 2... points. If they cross once, a solution to (3.11) is found, that is, the optimal size of the certainty stratum is found. The boundaries given by step 2 are approximately optimal sizes of the genuine sampling strata. If the curves do not cross, there is no solution with a certainty stratum. In this case, stratify the frame with the extended Ekman rule into H strata. If the curves cross at more than one point, all the points have to be evaluated. This plot will be referred to as the ***certainty stratum plot*** (an example is shown in Figure 5.4).

Clearly, this algorithm will produce all points that satisfy equation (3.11) in Theorem 1 and the extended Ekman rule.

4.2 A numerical procedure for stratification by the extended Ekman rule

When discussing the Ekman rule and its extended version (below) we assume that the size of the certainty stratum, N_H, is known. In subsection 4.2 we consider the remainder of the frame after removal of the certainty stratum. Let this part be sorted by the stratification variable. Denote the minimum value by x_1 and maximum one by x_{N-N_H}.

Let $\#E$ denote the number of elements in a set E.

The Ekman stratification rule:
Let $N_h = \#A_h$, where A_h is stratum h, $h = 1, 2, \ldots H$-1. Set $b_0 = x_1$ and $b_{H-1} = x_{N-N_H}$.

Determine the stratum boundary points $b_1, b_2, \ldots b_{H-2}$ so as to satisfy the following relation as well as possible.

$$N_1(b_1 - b_0) = N_2(b_2 - b_1) = \ldots = N_{H-1}(b_{H-1} - b_{H-2}) \qquad (4.1)$$

•

Remark. The reason for the slightly vague term "as well as possible" is that (4.1) usually lacks an exact solution when $N_1, N_2, \ldots N_{H-1}$ are confined to

integers. The extended Ekman rule, given below, admits non-integral $N_1, N_2, \ldots N_{H-1}$ and produces an exact solution under general conditions.

4.2.1 A geometric interpretation of the Ekman rule

The Ekman rule can be interpreted geometrically as in Figure 4.1, where a population divided into 3 strata is plotted. The cumulative distribution of **x** over the finite population is represented by a step function incrementing by 1 for each element in the population. Stratum 1, 2 and 3 generate rectangles, displayed in Figure 4.1, each with height N_h, $h = 1, 2, 3$, and width $(b_h - b_{h-1})$ and hence area $N_h(b_h - b_{h-1})$.

The crucial idea in the numerical algorithm for solving (4.1) is as follows. If you minimize the difference between the largest and smallest of the areas of the rectangles 1, 2 and 3 in Figure 4.1, you arrive at stratum boundaries that approximate (4.1) as well as possible. In the following we present a numerical method for finding the boundaries based on this idea.

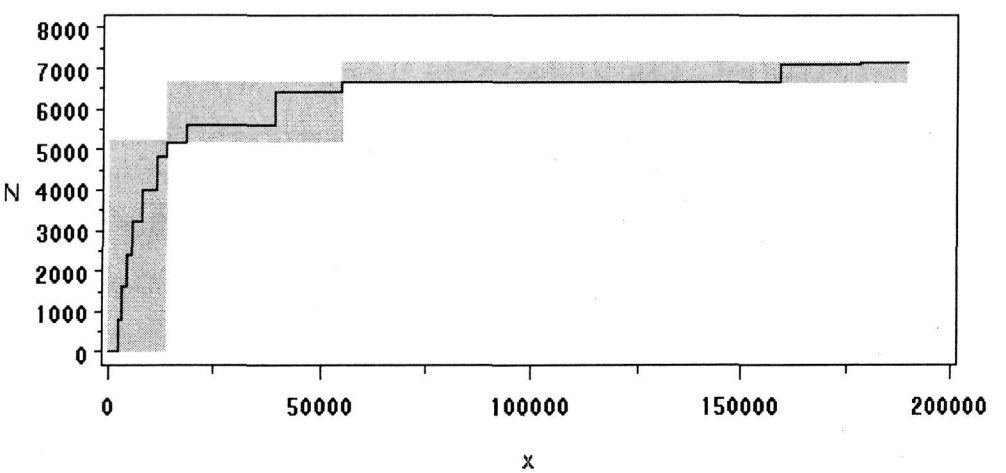


Figure 4.1. A geometric interpretation of the Ekman rule. A population where the stratification variable ranges from 0 to 190,000 is divided into 3 strata. The population is represented by a step function of cumulated frequencies.

4.2.2 Extended Ekman rule

The cumulative distribution function of **x** is

$F(x) = \#\{u : x_u \le x\}$, $b_0 \le x \le b_{H-1}$.

$F(\cdot)$ has a piecewise continuous step graph. Let the ***extended distribution graph***, denoted by $\boldsymbol{F}$, refer to the union of the graph of $F(\cdot)$ and the vertical lines connecting steps (see Figure 4.1). $\boldsymbol{F}$ is the graph of a vector-valued function

$$\beta \mapsto \boldsymbol{F}(\beta) = (x(\beta), N(\beta))$$

where $N(\beta)$ and $x(\beta)$ are continuous versions of the discrete variables N and x. Let the parameter β have the interpretation "distance along $\boldsymbol{F}$". Let the

minimum and maximum values of β be $\beta_0 = 0$ and

$\beta_{H-1} = \left(x_{N-N_H} - x_1\right) + \left(N - N_H\right)$. The endpoints of $\boldsymbol{F}$ are $\boldsymbol{F}\left(\beta_0\right) = \left(b_0, 0\right)$ and $\boldsymbol{F}\left(\beta_{H-1}\right) = \left(b_{H-1}, N - N_H\right)$.

By an ***extended stratum boundary point*** we mean any point on the graph $\boldsymbol{F}$. We will denote the H-2 extended stratum boundary points we are interested in by $\beta_1, \beta_2, \ldots \beta_{H-2}$. Given a β_h, the corresponding proper stratum boundary b_h is the horizontal position $x\left(\beta_h\right)$ of $\boldsymbol{F}$. There is a natural order of the extended stratum boundary points and the endpoints, let them satisfy $\beta_0 < \beta_1 < \beta_2 < \ldots < \beta_{H-1}$. In the extended situation we allow formation of rectangles with lower left and upper right corner anywhere along $\boldsymbol{F}$, including the vertical parts of it. We refer to the them as ***Ekman rectangles***. The area of Ekman rectangle h is

$$E_h = \left[N\left(\beta_h\right) - N\left(\beta_{h-1}\right)\right]\left(x\left(\beta_h\right) - x\left(\beta_{h-1}\right)\right).$$

The counterpart to (4.1) becomes

$$\begin{aligned}
&\left[N\left(\beta_1\right) - N\left(\beta_0\right)\right]\left(x\left(\beta_1\right) - x\left(\beta_0\right)\right) = \\
&\left[N\left(\beta_2\right) - N\left(\beta_1\right)\right]\left(x\left(\beta_2\right) - x\left(\beta_1\right)\right) = \\
&\ldots = \\
&\left[N\left(\beta_{H-1}\right) - N\left(\beta_{H-2}\right)\right]\left(x\left(\beta_{H-1}\right) - x\left(\beta_{H-2}\right)\right).
\end{aligned} \qquad \textbf{(4.2)}$$

We will refer to (4.2) as the ***extended Ekman rule***. The geometric interpretation of a solution to (4.2) is that all Ekman rectangles have the same area. Figure 4.2 exhibits the extended Ekman rule. The difference between Figure 4.1 and Figure 4.2 is that the rectangles of Figure 4.1 have nearly the same area, whereas the areas in Figure 4.2 are exactly the same.

There are conceivable cases where (4.2) has no solution, for example, if a large proportion of the units in the frame have the same value of x, but for all practical purposes we can neglect this possibility. It is readily seen in Figure 4.2 that an exact solution $x\left(\beta_1\right), x\left(\beta_2\right), \ldots x\left(\beta_{H-2}\right)$ of (4.2) gives stratum boundaries b_1, $b_2, \ldots b_{H-2}$ that satisfy (4.1) "as well as possible". It is also readily seen that a solution to (4.2) is unique.

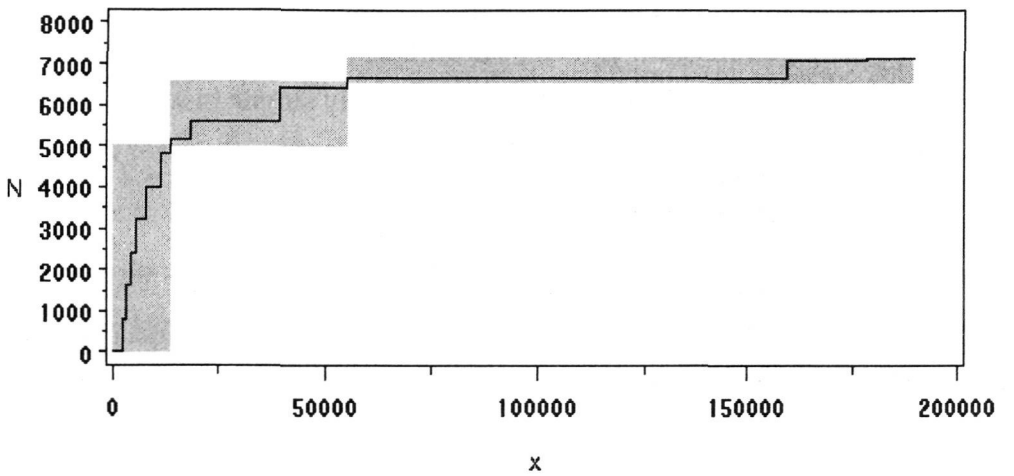


Figure 4.2. A geometrical interpretation of the *extended Ekman rule*.

4.2.3 Algorithm for solving (4.2)

First we give an outline of the algorithm, which soon will be specified. A start value β_1 is decided on. The area of the leftmost Ekman rectangle is then $E_1 = \left[N(\beta_1) - N(\beta_0)\right]\left(x(\beta_1) - x(\beta_0)\right)$. In the next step, β_h, $h = 2, 3... H\text{-}2$, are determined so as to equilize the areas of all Ekman rectangles but the rightmost one, whose area is

$E_{H\text{-}1} = \left[N(\beta_{H-1}) - N(\beta_{H-2})\right]\left(x(\beta_{H-1}) - x(\beta_{H-2})\right)$.

If $E_{H\text{-}1}$ is smaller than E_1, then β_1 is too large, if it is larger, β_1 is too small and if it equals E_1 (within some preassigned level of tolerance) a solution is found. If β_1 is too small or too large, the algorithm reiterates with a new value of β_1.

There are two main components in this procedure:

1. For given β_1, to find $\beta_2, \beta_3, \ldots \beta_{H-2}$ such that $E_2 = E_1$, $E_3 = E_1$, ... , $E_{H\text{-}1} = E_1$.
2. To pick a new value of β_1, when the current one is found too small or too large.

For both components we use the bisection method (see for example Dahlquist, Björck, 1974). The non-complicated version of this method we will need runs as follows. Let f be a continuous and monotone function on (a,b) with exactly one root ζ to the equation $f(x) = 0$ in (a,b). Divide the interval by its midpoint and check which of the two subintervals that contains ζ. The subinterval containing ζ is again divided, and so on. It is well known that this algorithm must converge to the root.

There are more efficient numerical methods for solving an equation than the bisection method. In this application, however, the rate of convergence of any iterative method and the approximation error is of minor importance since the

application is basically of discrete nature. There is no point in pursuing the algorithm until β_1 can be determined with a good number of significant decimals. Therefore, the comparatively simple bisection method is proposed.

Next, two of the steps of the algorithm that solves (4.2) are described separately.

Computation of extended stratum boundary points

Let β_1, and thus E_1, be given. In order to find the area of the second rectangle with an area E_2 that equals E_1, one wants to find the value of β_2 that solves the equation

$$E_1 - \left[N(\beta_2) - N(\beta_1)\right]\left(x(\beta_2) - x(\beta_1)\right) = 0. \qquad (4.3)$$

The function

$$Z(\beta_2) = E_1 - \left[N(\beta_2) - N(\beta_1)\right]\left(x(\beta_2) - x(\beta_1)\right)$$

is continuous and strictly decreasing on (β_1, β_{H-1}). Therefore, $Z(\beta_2)$ has at most one root in (β_1, β_{H-1}). There is exactly one root if $Z(\beta_1) > 0$ and $Z(\beta_{H-1}) \leq 0$. There is no root if $Z(\beta_1) > 0$ and $Z(\beta_{H-1}) > 0$. In this case β_2 and E_2 are set to missing.

The algorithm above is formulated for β_2, given β_1. It is repeated for the pairs (β_2, β_3), (β_3, β_4), ... $(\beta_{H-3}, \beta_{H-2})$. If β_i is missing in a pair (β_i, β_j), then β_j and E_j are set to missing.

Classification of extended stratum boundary points

A tolerance $\delta > 0$ is specified. After all extended stratum boundary points $\beta_1, \beta_2, \ldots \beta_{H-2}$ are computed, the point β_1 is classified. If the rightmost Ekman rectangle, E_{H-1}, is non-missing it is either smaller than, larger than or equal to (with tolerance δ) E_1. If it is missing, it is considered smaller than E_1. We classify β_1 into the three possible outcomes:

- β_1 is *too large* if $E_{H-1} + \delta < E_1$.
- β_1 is *too small* if $E_{H-1} > E_1 + \delta$.
- β_1 is *good* if $|E_{H-1} - E_1| \leq \delta$.

This classification divides the graph $\boldsymbol{F}$ into three parts according to the value of β_1: the first part where β_1 is too small, the second one where it is good and the last part where β_1 is too large.

An algorithm that solves (4.2)

1. Specify a pair $\left(\beta_1', \beta_1''\right)$ of a too small and a too large value of β_1, for example $\left(\beta_0, \beta_{H-1}\right)$.
2. Compute the arithmetic mean of $\left(\beta_1', \beta_1''\right)$. Denote it β_1^*.
3. Compute $\beta_2, \ldots \beta_{H-2}$ given $\beta_1 = \beta_1^*$ and classify β_1^* into good, too small or too large.
4. If β_1^* is good, a solution of (4.2) is found and the algorithm is terminated.

   Else if β_1^* is too small, go to step 1 and replace $\left(\beta_1', \beta_1''\right)$ with $\left(\beta_1^*, \beta_1''\right)$.

   Else if β_1^* is too large, go to step 1 and replace $\left(\beta_1', \beta_1''\right)$ with $\left(\beta_1', \beta_1^*\right)$.

5 Applications

In this section we give some numerical illustrations of the results in section 3 and 4. We worked under the assumption that the study variable is equal to the stratification variable. There are at least two reasons for studying practical applications under this assumption:

- Theorem 1 was derived under the assumption that the discrete distribution **x** can sufficiently well be approximated by a continuous distribution. This suggests that there may exist a stratification with lower variance than a stratification that satisfies the conditions of Theorem 1. It is therefore of interest to see how Theorem 1 works in practice (compare Remark 5 in section 3).
- It is interesting to compare the results of this report to those of other authors who work under the same assumption.

The two populations introduced next were considered.

5.1 The value added population

The annual census of Swedish manufacturing industry collects data on sales, cost of materials, energy used in the production process, etc. The value added is derived. The census together with derived variables is frequently used as a sampling frame for other surveys. We used the 1989 frame with value added as stratification variable. This frame, which in the sequel is referred to as the *value added population*, contains 7326 establishments. Its skewness is 12.4 (which could be compared with skewness 2.0 of an exponential distribution).

5.2 The log-normal population

An artificial population was created by 2000 random numbers generated from a log-normal distribution $X = e^Z$ where Z is univariate normal with mean 4 and variance 2.7 (further details in Appendix A). Again it is a highly skewed population, the skewness being 26.7.

5.3 General framework for the simulations

In the simulations we divided given populations into $H = 4$ strata. The stratum comprising units with the largest values of the stratification variable was a certainty stratum, the other strata were genuine sampling strata. A sample size was determined. The sample was allocated according to Theorem 1, that is, with stratum H as a certainty stratum, the allocation rule is

$$n_h = (n - n_H)\frac{N_h S_h}{\sum_{h=1}^{H-1} N_h S_h}, \qquad (5.1)$$

where S_h is the standard deviation of the stratification variable within stratum h. We will call this x-optimal allocation (thus adhering to the terminology of Särndal, Swensson, Wretman, 1992).

5.4 Performance measure

5.4.1 Best possible stratification

Due to the approximation mentioned in the first paragraph of section 5 there may exist a stratification with lower variance than a stratification that satisfies the conditions of Theorem 1. For each situation considered in this section we searched for the stratification with the least estimator variance (3.1), which we refer to as the ***best possible stratification***.

The values $x_1, x_2, \ldots x_N$ of the stratification variable furnish the set of all potential stratum boundaries. A boundary b_h anywhere in the interval $[x_{k-1}, x_k)$, where k-1 and k are two adjacent units in the ordered population, give the same estimator variance as the boundary $b_h = x_{k-1}$, provided the other boundaries remain unchanged. If $b_h = x_{k-1}$, unit k-1 belongs to stratum h. In the considered situations, with $H = 4$ strata, a stratification is specified by the boundaries b_1, b_2 and b_3. Alternatively, since the population size N is given, a stratification is specified by three of the stratum sizes N_1, N_2, N_3 and N_4. Clearly, as we now consider a specific situation, with specified values of $\mathbf{x} = x_1, x_2, \ldots x_N$, sample size n and number of strata H, there exists a best possible stratification (a global minimum). We denote the estimator variance by $Var(\hat{t}; \mathbf{N})$, where $\mathbf{N} = (N_1, N_2, N_3)$.

For both populations studied, $Var(\hat{t}; \mathbf{N})$ was computed for a large number of combinations of N_1, N_2 and N_3. Under variation of the three stratification parameters the estimator variance forms a response surface in a four-dimensional space. Let $\mathcal{P}_j$ be the response surface projected on the two-dimensional space $(N_j, Var(\hat{t}; \mathbf{N}))$ for j = 1, 2, 3 and 4. Figure 5.1 shows a scatter plot of $\mathcal{P}_1$. The vertical dotted lines represent estimator variances with varying N_2 and N_3 for given values of N_1. Note that a convex function is formed by the minimum values of the vertical dotted lines. This observation was used in the search method that enabled us to find the best possible stratification. We do not, however, give a full account of the search method here. Figure 5.2 displays $\mathcal{P}_j$ for j = 1, 2, 3 and 4, with the ***relative variance*** along the y-axis: the ratio of the estimator variance (3.1) obtained by a particular stratification and the estimator variance using the best possible stratification.

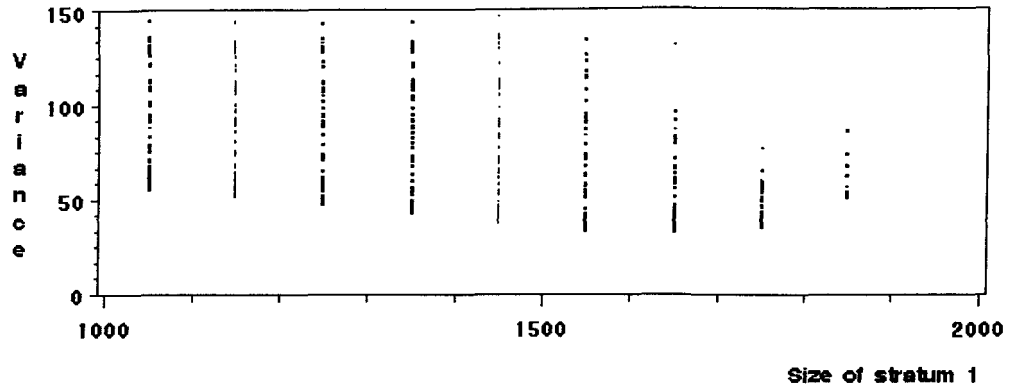
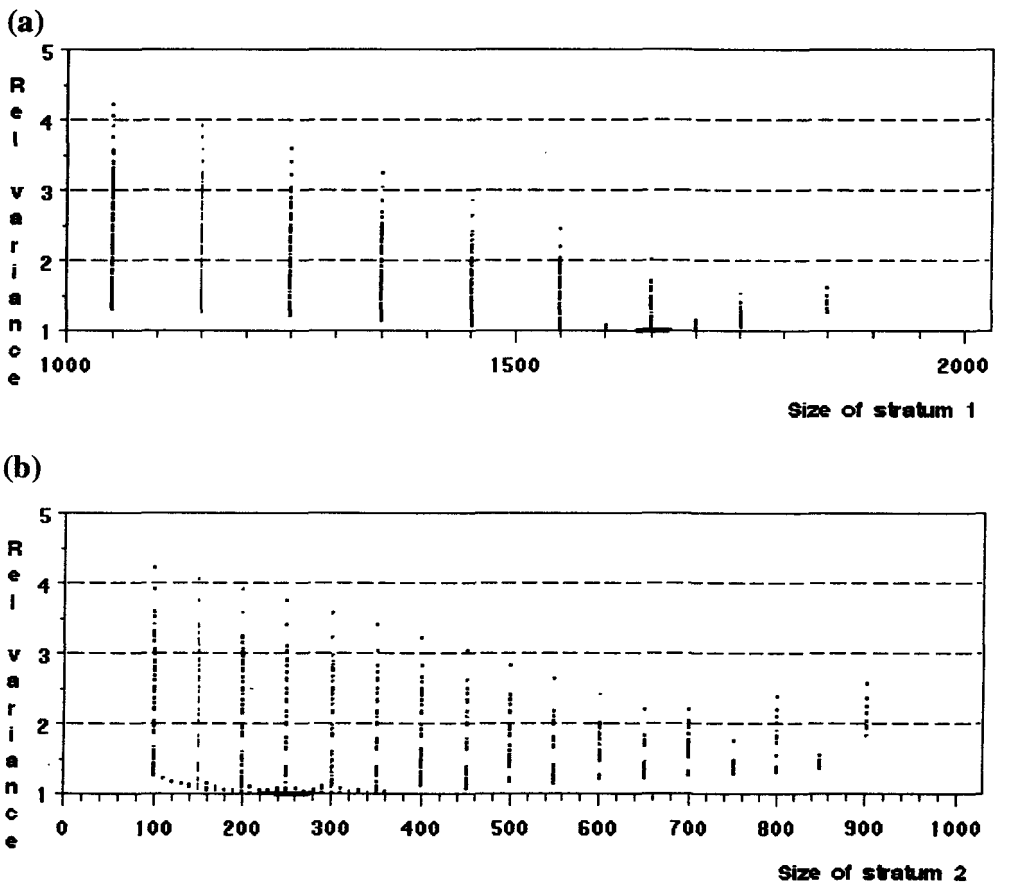
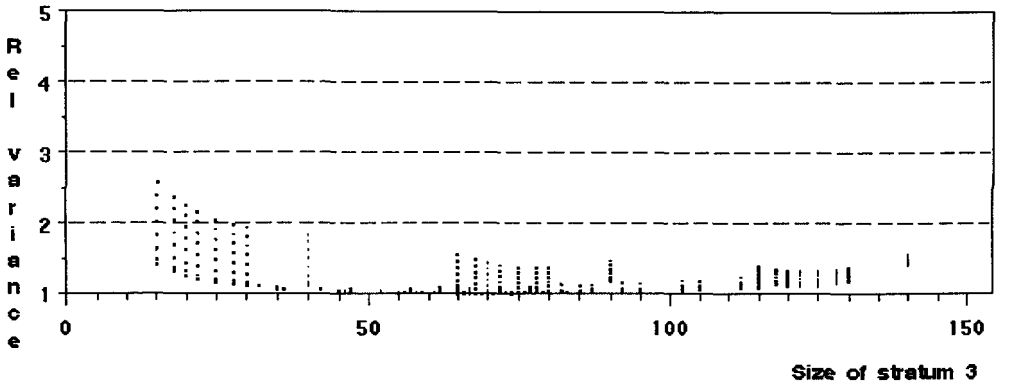


Figure 5.1. The estimator variance surface for a large number of stratifications of the log-normal population, projected on the plane given by N_1 and the variance (divided by 10^9).



(c)



(d)

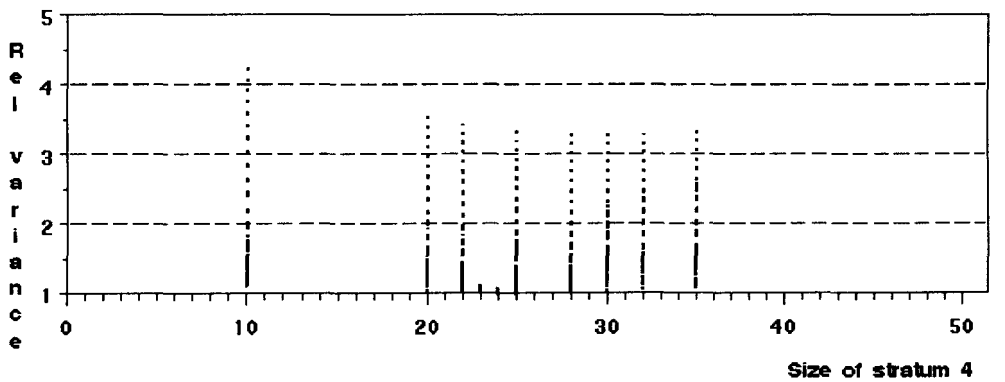


Figure 5.2. The relative variance of a large number of stratifications of the log-normal population. In scatter plot (a) different sizes of stratum 1, N_1, are plotted along the x-axis. The vertical dotted lines represent relative variances with varying N_2 and N_3 given a value of N_1. Scatter plots (b), (c) and (d) display exactly the same stratifications as (a), although with N_2, N_3 and N_4, respectively, along the x-axis.

5.4.2 Best possible stratification of the value added population

In the stratification study of the value added population the size of the total sample was set to 400, that is, an overall sampling rate of somewhat more than 5 %. The best possible size of the certainty stratum was found to be 186. Some characteristics of the best possible stratification are shown in Table 5.2. All calculations were based on values in 1000 SEK, although the values displayed in Table 5.2 are rounded to nearest million SEK. Even with stratum 4 removed, the remaining population is highly skewed, the skewness being 3.5.

The coefficient of variation (CV) is the square root of the estimator variance divided by the total. To emphasize that the CV refers to an estimate of the total of the stratification variable **x**, we denote it x-CV:

$$x\text{-CV} = \frac{\sqrt{V(\hat{t}_x)}}{t_x} \qquad (5.2)$$

The minimum x-CV of this population, constructing 4 strata of any kind and sampling 400 units, is 1.688 %.

Stratum	Minimum x-value	Maximum x-value	N_h	n_h	$1-n_h/N_h$ %	$\bar{x}_h$	S_h^2
1	0	654	1642	6	99.6	90	$0.2 \cdot 10^5$
2	664	5574	266	9	96.6	1911	$15 \cdot 10^5$
3	5801	26098	68	11	83.9	12426	$350 \cdot 10^5$
4	29444	399214	24	24	0	66420	$57573 \cdot 10^5$
Sum			2000	50			

Table 5.1. Characteristics of the best possible stratification of the log-normal population.

Stratum	Minimum x-value	Maximum x-value	N_h	n_h	$1-n_h/N_h$ %	$\bar{x}_h$	S_h^2
1	0	13.9	5225	74	98.6	5.2	$1 \cdot 10^3$
2	13.9	54.9	1433	66	95.4	27	$120 \cdot 10^3$
3	55.1	189.1	482	74	84.7	97	$314 \cdot 10^3$
4	191.6	..	186	186	0	454	$129055 \cdot 10^3$
Sum			7326	400			

Table 5.2. Characteristics of the best possible stratification of the value added population. Unit 1 million SEK.

5.4.3 Best possible stratification of the log-normal population

When stratifying the log-normal population, the sample size was set to 50 units. Some characteristics of the best possible stratification are shown in Table 5.1. The minimum x-CV, defined in (5.2), is 5.819 %.

5.5 On the equations (2.4) and (3.10)

It is interesting to see how well the best possible stratum boundaries in Table 5.1 and Table 5.2 satisfy the Dalenius equations (2.4) and the corresponding condition (3.10) in Theorem 1. We refer to the factors $1-n_h/N_h$ in condition (3.10) as finite population corrections (*fpc*). As $1-n_h/N_h < 1$ for $h = 1, 2, \ldots H-1$, the *fpcs* in (3.10) moderate the impact of $(y_h - \mu_h)^2$ and $(y_h - \mu_{h+1})^2$. If the *fpcs* increase from stratum 1 to stratum H, which is likely if the population is highly skewed, the effect of the *fpcs* is stronger on the right hand side of each equation. Consequently, (3.10) tends to produce strata less unequal in size than strata given by the Dalenius equations. This is displayed in the applications to the value added and the log-normal populations below. The relative variance, however, turned up only a trifle above 1.

5.5.1 Equations (2.4) and (3.10) applied to the value added population

The characteristics of the best possible stratification for the genuine sampling strata (strata 1, 2 and 3 given in Table 5.2) were inserted in the Dalenius equations (2.4) and in system (3.10) in Theorem 1. A value of the right and

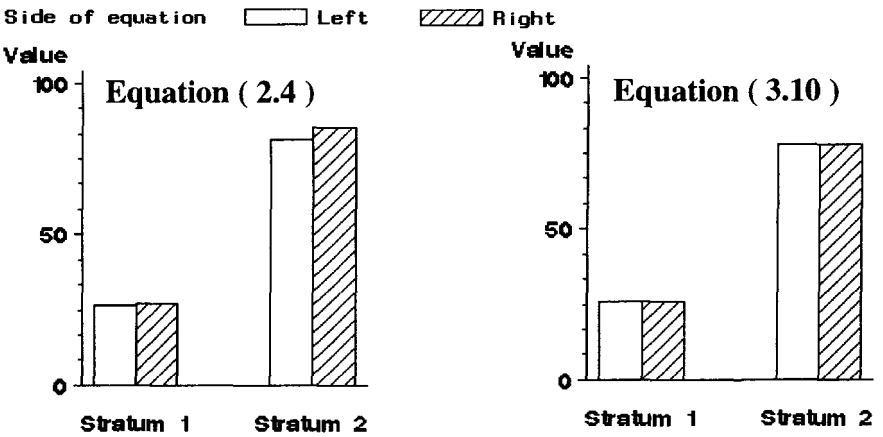


Figure 5.3. The best possible stratum boundaries for the value added population (from Table 5.2) inserted into (2.4) and into (3.10). The bars represent the value (in thousands) of the left and right hand side, respectively, of (2.4) and (3.10).

left hand side, respectively, were obtained for each of the equations with $h = 1, 2$. Figure 5.3 exhibits those values. Notice a discrepancy between the left and the right hand side for the Dalenius equation associated with stratum 2, whereas the best possible boundaries satisfy (3.10) almost exactly for both stratum 1 and 2.

It is also interesting to analyse the problem the other way around. The stratification in Table 5.3 is a solution to the Dalenius equations in the following sense. Usually, when (2.4) is applied to a finite population an exact solution does not exist. The stratum boundaries b_1 and b_2 shown in Table 5.3 minimize $D_1 + D_2$ where

$$D_h = \left| \frac{S_{h+1}^2 + (b_h - \bar{x}_{h+1})^2}{S_{h+1}} - \frac{S_h^2 + (b_h - \bar{x}_h)^2}{S_h} \right|, \quad h = 1, 2.$$

The boundaries b_1 and b_2 are the maximum x-values within strata. Stratum 4 is fixed to 186 units which is its best possible size. The relative variance turned out to be 1.004, that is, only slightly above 1.

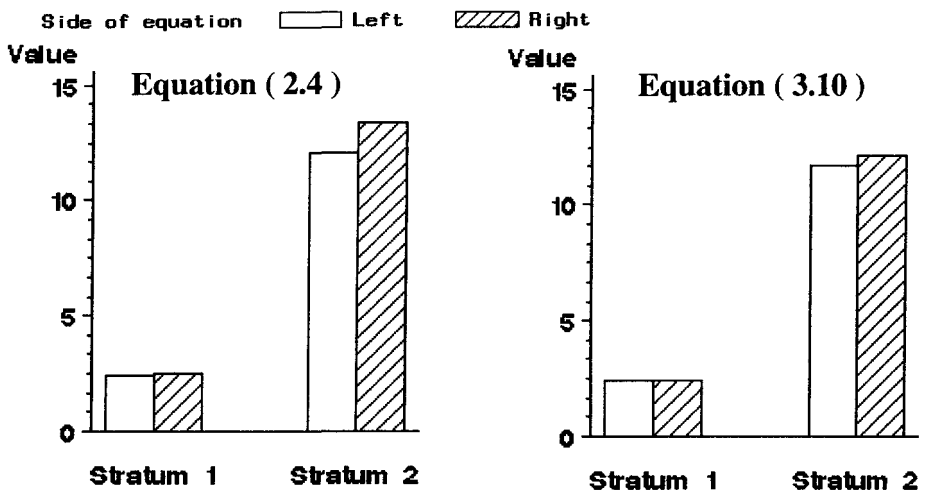


Figure 5.4. The best possible stratum boundaries of the log-normal population inserted in (2.4) and in (3.10). The bars represent the value (in thousands) of the left and right hand side, respectively, of (2.4) and (3.10).

Stratum	b_h	N_h	n_h
1	16.0	5400	85
2	60.2	1320	67
3	189.1	420	62
4	..	186	186
Sum		7326	400

Table 5.3. Stratum boundaries for the value added population determined by (2.4). Stratum 4 was fixed to 186. Relative variance: 1.004.

5.5.2 Equations (2.4) and (3.10) applied to the log-normal population

It is interesting to see (2.4) and (3.10) applied to a population of extreme skewness, where the impact of the *fpcs* is stronger. As seen in Table 5.1, the sample from the log-normal population is not equally allocated. The Dalenius-Hodges rule makes $N_h S_h$ approximately equal for all strata, which makes x-optimally allocated sample sizes n_h (5.1) also approximately equal (Cochran 1977). This suggests that both the Dalenius equations (2.4) and the Dalenius-Hodges rule, which gives an approximate solution to (2.4), might be far from what is best possible. Figure 5.4 does exhibit discrepancies, larger for the Dalenius equations than for (3.10). The stratification in Table 5.4 is a solution to the Dalenius equations, with stratum 4 fixed to the best possible size, which is 24 units. The relative variance is 1.005.

This result and that of subsection 5.5.1 indicate that the Dalenius equations (2.4), as well as methods that give approximate solutions to (2.4), give only a minor loss of precision compared to the best possible stratification.

Stratum	b_h	N_h	n_h
1	761	1671	7
2	6144	241	9
3	26098	64	10
4	399214	24	24
Sum		2000	50

Table 5.4. Stratum boundaries for the log-normal population determined by (2.4). Stratum 4 was fixed to 24. Relative variance: 1.005.

5.5.3 The Ekman and the Dalenius-Hodges rules

The Ekman and the Dalenius-Hodges rules were applied to the value added and the log-normal population. Both rules give stratification boundaries that are approximate solutions to (2.4). Therefore, they are applicable exclusively for stratifications where you end up with genuine sampling strata only. For this reason the stratum comprising the units with the largest values was held fixed to the size found to be best possible (Table 5.1 and Table 5.2, respectively). Table 5.5 and Table 5.6 show results for the value added population. Both methods work well, the relative variance is 1.001 for the Dalenius-Hodges rule and 1.002 for the Ekman rule. When using Dalenius-Hodges rule the value added population was divided into 198 intervals and the log-normal one into 195 (a good description is provided in Särndal, Swensson, Wretman (1992, p. 463) who denote the number of intervals by J). As for the Ekman rule we used the algorithm for the extended Ekman rule described in section 4.

Table 5.8 shows that the Ekman rule works well for the log-normal population, too. The relative variance is 1.004. The Dalenius-Hodges rule yields a slightly higher relative variance: 1.026 (Table 5.7).

Stratum	b_h	N_h	n_h
1	12.6	5132	69
2	50.4	1511	69
3	189.1	497	76
4	..	186	186
Sum		7326	400

Table 5.5. Stratum boundaries given by the Dalenius-Hodges rule for the value added population. Stratum 4 was fixed to 186. Relative variance: 1.001.

Stratum	b_h	N_h	n_h
1	13.0	5086	66
2	54.9	1572	74
3	189.1	482	74
4	..	186	186
Sum		7326	400

Table 5.6. Stratum boundaries given by the extended Ekman rule for the value added population. Stratum 4 was fixed to 186. Relative variance: 1.002.

Stratum	b_h	N_h	n_h
1	772	1673	7
2	4662	220	6
3	26098	83	13
4	399214	24	24
Sum		2000	50

Table 5.7. The Dalenius-Hodges rule applied to the log-normal population. Stratum 4 was fixed to 24. Relative variance: 1.026.

Stratum	b_h	N_h	n_h
1	761	1671	7
2	6110	240	9
3	26098	65	10
4	399214	24	24
Sum		2000	50

Table 5.8. The extended Ekman rule applied to the log-normal population. Stratum 4 was fixed to 24. Relative variance: 1.004.

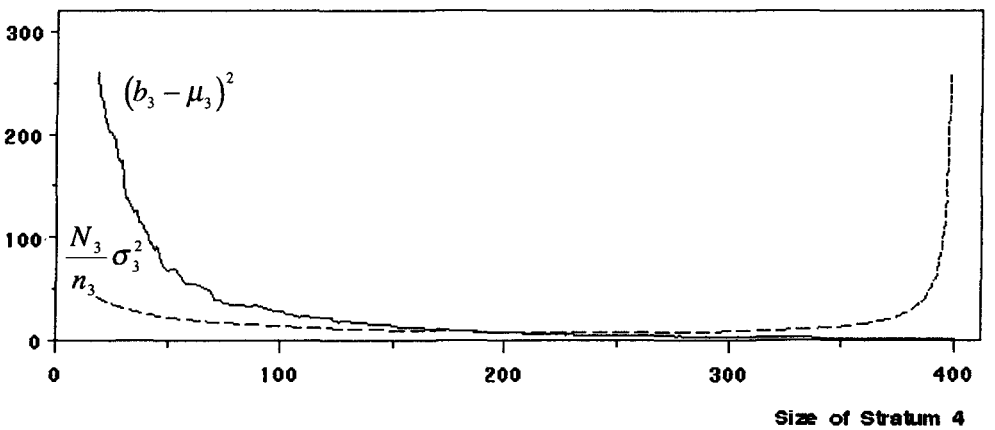


Figure 5.5. The certainty stratum plot. Each side of equation (3.11) was computed for all possible sizes of the certainty stratum. The values of the left and right hand side (divided by 10^9) are plotted against the number of units in the certainty stratum.

5.6 The stratification algorithm

5.6.1 The stratification algorithm applied to the value added population

Using the stratification algorithm (section 4) the value added population was divided into 4 strata. The extended Ekman rule was used to construct stratum 1, 2 and 3, while condition (3.11) in Theorem 1 provided the boundary of stratum 4. We used the stratification algorithm with $K = 1$. The plot described in step 6 of the stratification algorithm is shown in Figure 5.5. The curves cross at $N_4 = 186$, which coincides with the best possible size of stratum 4 (see Table 5.2). Hence, the extended Ekman rule applied to the value added population minus stratum 4 with $N_4 = 186$ yields the stratification displayed in Table 5.6. Thus the relative variance given by the stratification algorithm is 1.002.

5.6.2 The stratification algorithm applied to the log-normal population

Table 5.9 exhibits the stratification algorithm applied to the log-normal population. The relative variance is 1.012. The size of the certainty stratum differs slightly from the best possible size, which is 24.

Stratum	b_h	N_h	n_h
1	809	1680	8
2	6467	237	9
3	29444	60	10
4	399214	23	23
Sum		2000	50

Table 5.9. Stratum boundaries given by the stratification algorithm for to the log-normal population. Relative variance: 1.012.

5.7 The Lavallée and Hidiroglou algorithm

The Lavallée and Hidiroglou algorithm was applied to the value added and the log-normal population. The input and the output of the *stratification algorithm* is a sample size and an x-CV (5.2), respectively, whereas the Lavallée and Hidiroglou algorithm works the other way around. When using this algorithm, the user requests an x-CV and the algorithm responds with stratum boundaries, a minimum total sample size and a sample allocation that give the x-CV asked for (compare Lavallée, Hidiroglou, 1988). In our study, this algorithm was re-run with varying x-CV requests until it produced the same total sample size as the one that was input to the stratification algorithm. The US Bureau of the Census has kindly provided an implementation of this algorithm, modified to accommodate Neyman allocation (Sweet, Sigman, 1995 a). The stratum boundaries shown in Table 5.10 and Table 5.11 were produced by Sweet's and Sigman's program used with the option requesting x-allocation (specifications of the options used are found in Appendix B). A minor modification of the value added data set was imposed on the 67 records with null value of the stratification variable. They were replaced with random numbers taken from a uniform (0,1) distribution in order to avoid a group of values having exactly the same value of the stratification variable, which caused abnormal ending of the program. This is

discussed in Sweet and Sigman (1995 a, p 10). As seen in Table 5.10 and Table 5.11 the strata and the relative variances are similar to those obtained with the stratification algorithm (see Table 5.6 and Table 5.9).

Stratum	b_h	N_h	n_h
1	14.9	5317	80
2	59.5	1396	72
3	195.6	434	69
4	..	179	179
Sum		7326	400

Table 5.10. Stratum boundaries given by the Lavallée and Hidiroglou method for the value added population. Relative variance: 1.001.

Stratum	b_h	N_h	n_h
1	516	1595	4
2	4228	292	7
3	22605	85	11
4	399214	28	28
Sum		2000	50

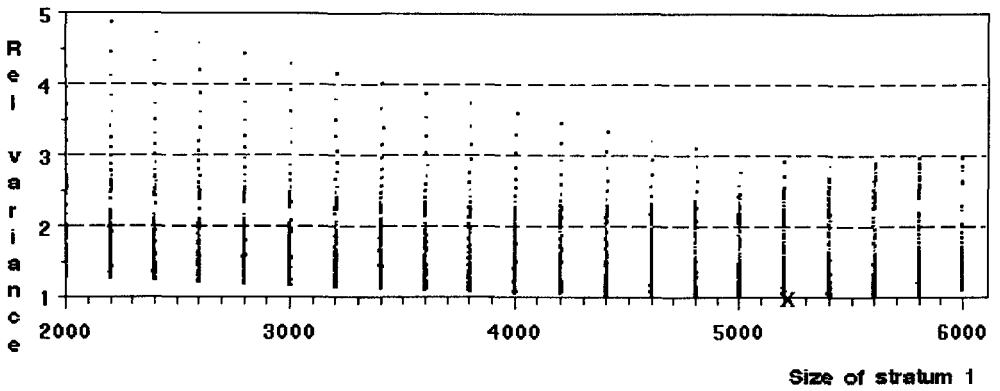
Table 5.11. Stratum boundaries given by the Lavallée and Hidiroglou method for the log-normal population. Relative variance: 1.018.

5.8 Flatness of the Objective Function

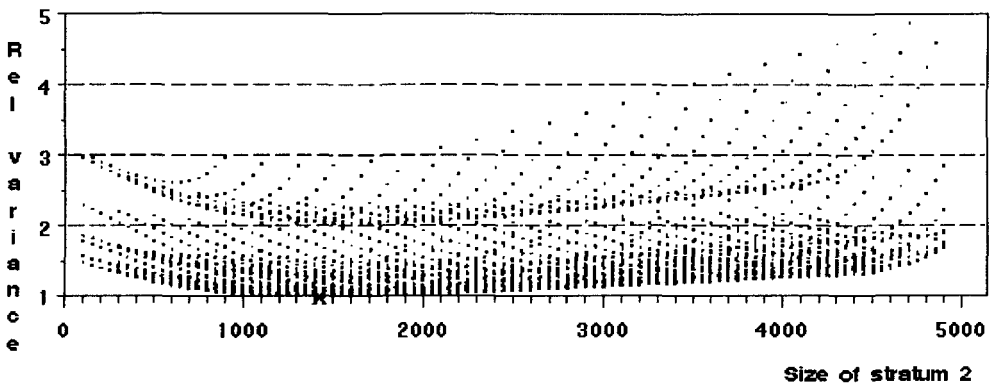
The x-CV and the relative variance were computed for 2856 stratifications of the value added population. In Figure 5.6 (a), (b), (c) and (d) the relative variance is plotted against the size of each stratum (it is the same type of figure as Figure 5.2). Recall that this population contains 7326 units and that the sample size was set to 400. The most striking feature of the plots in Figure 5.6 is the flatness of the estimator variance surface. Plot (d), for example, shows that if the size of the certainty stratum is within $(120, 230)$ it is possible to hit the minimum variance if the other strata are chosen optimally. The interval $(120, 230)$ must be considered very wide as the certainty stratum with a total sample size of 400 cannot contain more than 400 units. If the certainty stratum is chosen within this interval and the three genuine sampling strata are determined by the Ekman rule, the worst relative variance is 1.05 (achieved for $N_4 = 120$). This repeated for the interval $(140, 230)$ gives 1.02 as the worst relative variance (achieved for $N_4 = 230$). A large certainty stratum combined with a small size of stratum 3 yields relative variances that are unacceptable.

The log-normal population shows similar flatness (Figure 5.2).

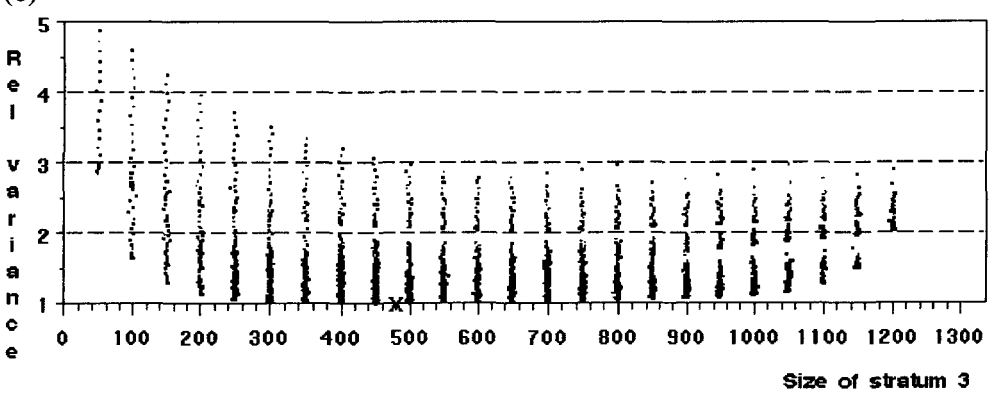
(a)


Rel variance
5
4
3
2
1
2000
3000
4000
5000
6000
Size of stratum 1

(b)


Rel variance
5
4
3
2
1
0
1000
2000
3000
4000
5000
Size of stratum 2

(c)


Rel variance
5
4
3
2
1
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
Size of stratum 3

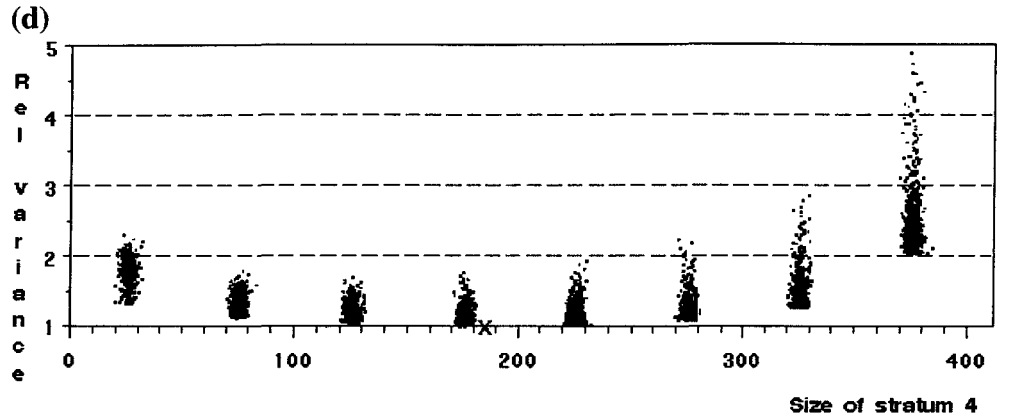


Figure 5.6. The relative variance for a large number of stratifications of the value added population. Plot (a) shows the relative variance for stratum 1 with N_1 along the x-axis. For each N_1 there are a number of choices of N_2, N_3 and N_4. The relative variance of each combination is represented by a point in the scatter plot. The points are randomly moved horizontally by addition of a small normally distributed quantity. The best possible size of stratum 1 is marked on the x-axis. Plot (b), (c) and (d) are analogous, with N_2, N_3 and N_4, respectively, instead of N_1. All combinations of N_1, N_2, N_3 and N_4 are shown in each plot.

Fairly large deviations from the best possible value of stratum boundary b_2 are not deleterious. In Figure 5.7 the certainty stratum is fixed to the best possible size and b_2 varies. Given a value of b_2 the remaining boundary b_1 was computed with the Ekman rule. If the size of stratum 3 is in the interval $(450, 515)$, the relative variance is less than 1.01. A doubled size of stratum 3, from optimal 482 to 1000, increases the relative variance to 1.10. It is the very small values of stratum 3 that give really bad stratifications.

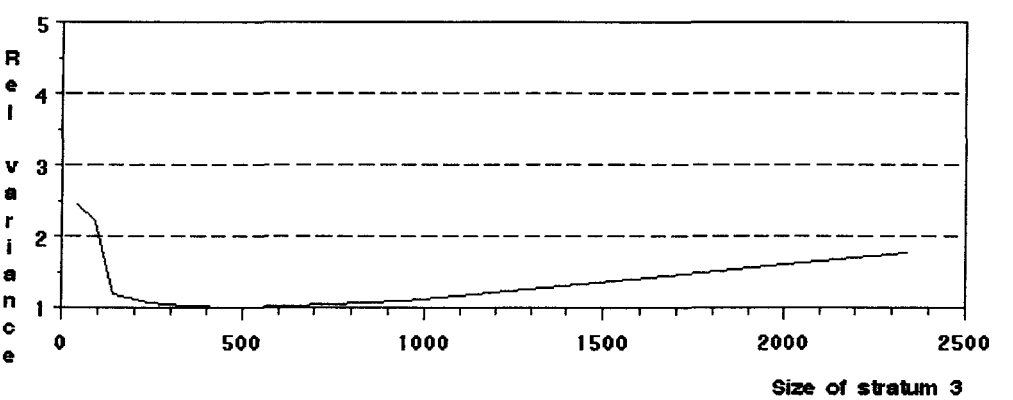


Figure 5.7. The relative variance for a large number of stratifications of the value added population. Stratum 4 is fixed to 186. The boundary between stratum 2 and 3 varies. For a given size of stratum 3, stratum 1 is computed with the extended Ekman rule.

As a concluding remark, suppose a frame is divided into H strata where the stratum containing the largest units, stratum H, is a certainty stratum and all other strata are genuine sampling strata and suppose the number of units in the frame is substantially greater than H. Then, in most practical applications, the estimator variance surface is flat around the best possible stratum boundaries. Hence, a moderate deviation from the best possible values of $b_1, b_2, \ldots b_{H-2}$ should increase the variance only negligibly. The variance around b_{H-1} is fairly flat, although not necessarily as flat as around the other stratum boundaries.

To the knowledge of the author, there is no support of the intuitive feeling that the variance is flat around the optimal boundaries of genuine sampling strata, other than empirical data. For example Dalenius and Gurney (1951, p 141) say as a final remark supporting Approximation 1 in the optimum stratification problem:

> ...as is often the case with optimum solutions in sampling, slight deviations from the absolute optimum value have almost no practical importance.

Acknowledgement: The author wishes to express his gratitude to colleagues at Statistics Sweden, in particular Professor Bengt Rosén, for encouragement and constructive criticism.

References

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed., New York: Wiley.

Cochran, W.G. (1961). Comparison of Methods for Determining Stratum Boundaries. *Bulletin de l'Institut International de Statistique*, Vol. 38:2, pp. 345-357.

Dahlquist, G., Björck, Å. (1974). *Numerical Methods*. Prentice-Hall.

Dalenius, T. (1950), The Problem of Optimum Stratification. *Skandinavisk Aktuarietidskrift*, pp. 203-213

Dalenius, T.; Gurney, M. (1951). The Problem of Optimum Stratification. II. *Skandinavisk Aktuarietidskrift*, pp. 133-148.

Dalenius, T. (1952). The Problem of Optimum Stratification in a Special Type of Design. *Skandinavisk Aktuarietidskrift*, pp. 61-70.

Dalenius, T.; Hodges, J.L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, Vol. 54, pp. 88-101.

Detlefsen, R. E.; Veum, S. C. (1991). Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods. American Statistical Association,* pp. 214-219.

Ekman, G. (1959). An Approximation Useful in Univariate Stratification. *The Annals of Mathematical Statistics*, Vol. 30:1, pp. 219-229.

Glasser, G.J. (1962). On the Complete Coverage of Large Units in a Statistical Study. *Review of the International Statistical Institute*, Vol. 30:1, pp. 28-32.

Hess, I., Sethi, V.K. and Balakrishnan, T.R. (1966). Stratification: A Practical Investigation. *Journal of the American Statistical Association*, Vol. 61, pp. 74-90.

Hidiroglou, M. A. (1986). The Construction of a Self-Representing Stratum of Large Units in Survey Design. *The American Statistician*, Vol. 40:1, pp. 27-31.

Hidiroglou, M. A., Srinath K. P. (1993). Problems Associated with Designing Subannual Business Surveys. *Journal of Business & Economic Statistics*, Vol. 11, No. 4, pp. 397-405.

Lavallée P.; Hidiroglou, M. A. (1988). On the stratification of Skewed Populations. *Survey Methodology*, Vol. 14, No. 1, pp. 33-43.

Luenberger, D. G. (1973). *Introduction to linear and nonlinear programming.* Reading, Massachusetts: Addison-Wesley.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Protter, M. H.; Morrey, C. B. (1977). *A First Course in Real Analysis.* New York: Springer-Verlag.

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

Schneeberger, H. (1985). Maxima, Minima und Sattelpunkte bei optimaler Schichtung und optimaler Aufteilung. *Allgemeines Statistisches Archiv*, Vol. 69:3, pp. 286-297.

Sigman, R.; Monsour, N. (1995). Selecting Samples from List Frames of Business. In: B. Cox, D. Binder, N. Chinappa, A. Christianson, M. Colledge, P. Kott (eds.), *Business Survey Methods*. New York: Wiley, pp. 133-152.

Slanta, J.; Krenzke, T. (1996). Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey. *Survey Methodology*, Vol. 22, No. 1, pp. 65-75.

Sweet, E. M. and Sigman R. S. (1995 a). User Guide for the Generalized SAS Univariate Stratification Program. *ESM Report Series*, ESM-9504. Bureau of the Census.

Sweet, E. M.; Sigman R. S. (1995 b). Evaluation of Model-Assisted Procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Research Methods, Vol I. American Statistical Association,* pp. 491-496.

Appendix A

The SAS-program below created the log-normal population used in the simulation studies. The program was run on SAS version 6.12 under Windows 95.

```
data lognorm;
 do i=1 to 2000;
   x=exp(4+2.7*normal(10));
   output;
 end;
 drop i;
run;
```

Appendix B

The program below is a part of an implementation of the Lavallée and Hidiroglou algorithm created by the US Bureau of the Census (Sweet, Sigman (1995 a)). The lines below show the specifications used when the stratum boundaries for the value added population were computed. The population is stored in the SAS-file s.ram with stratification variable fv. The program was run on SAS version 6.12 under Windows 95.

```
data ett;
 set s.ram;
 if fv=0 then fv=uniform(1);
run;

%global file;          * the name of the sas dataset;
%global direct;        * the directory where the sas dataset is located;
%global forder;        * indicator of first order stratification level;
%global strat;         * name of first order stratification level;
%global bintype;       * they type of intervals to make;
%global same;          * specifics when intervals are already on the dataset;
%global binvar;        * the variable to make the intervals and stratify with;
%global numbin;        * the number of intervals to make;
%global binname;       * the name of the intervals;
%global labeltyp;      * specifics for unequal intervals;
%global stvar;         * needed for stratification:  user shouldn't change;
%global numstrat;      * the number of strata to make;
%global cumtype;       * the type of stratification to use;
%global lhtype;        * stratification for certainty stratum;
%global vartype;       * the type of variance to make;
%global varvar;        * the variable to use in creating variances;
%global cv;            * the coefficient of variation to obtain;
%global cost;          * the cost per unit of sampling;
%global alpha;         * parameter estimates;
%global beta;          * parameter estimates;
%global g;             * parameter estimates;
%global delta;         * parameter estimates;
%global d0 d1 d2;      * parameter estimates;

%let file=ett;
%let direct= 'c:\dan\saswork';
%let forder=2;
%let strat=;
%let bintype=3;
%let same=1;
%let binvar=fv;
%let numbin=200;
%let binname=binnum;
%let labeltyp=1;
%let stvar=count;
%let numstrat=4;
%let cumtype=1;
%let lhtype=1;
%let vartype=1;
%let varvar=fv;
%let cv=.01688;
%let cost=1;
%let alpha=1;
%let beta=1;
%let g=1;
%let delta=1;
%let d0=1;
%let d1=1;
%let d2=1;
```

BR/BY-S

Förteckning över utkomna R&D Reports

R&D Reports är en för IT-enheten och U/STM gemensam publikationsserie, som 1988-01-01 ersatte de tidigare "gula" och "gröna" serierna. I serien ingick fram till årsskiftet 1992-93 även **Abstracts** (sammanfattning av metodrapporter från SCB).

Reports published 1996 and onwards:

1996:1 (grön) On Sampling with Probability Proportional to Size *(Bengt Rosén)*

1996:2 (grön) Bortfallsbarometern nr 11 *(Antti Ahtiainen, Stefan Berg, Margareta Eriksson,* Åsa *Greijer, Dan Hedlin, Monica Rennermalm, Anita Ullberg)*

1996:3 (grön) Regression Estimators in Theory and in Practice *(Tomas Garås)*

1996:4 (gul) Quality Aspects of a Modern Database Service *(Pat Dean, Bo Sundgren)*

1996:5 (gul) Metadata: A Quality Element in Official Statistics - the Swedish Approach *(Bo Sundgren, Pat Dean)*

1996:6 (gul) Our Legacy to Future Generation - Using Databases for Better Availability and Documentation *(Gösta Guteland, Erik Malmborg)*

1997:1 (grön) Bortfallsbarometern nr 12 *(Anti Ahtiainen, Stefan Berg, Mats Bergdahl, Fredrik Granström, Dan Hedlin, Lena Otterskog, Monica Rennermalm)*

1997:2 (grön) Quality Concept for Official Statistics - Entry in the forthcoming update of the Encyclopedia of Statistical Sciences, Wiley & Sons *(Eva Elvers, Bengt Rosén)*

1998:1 (grön) Nybyggnadsstatistik - en simuleringsstudie *(Catarina Elffors)*

1998:2 (grön) On Inclusion Probabilities for Order Sampling *(Bengt Rosén)*

1998:3 (grön) On the Stratification of Highly Skewed Populations *(Dan Hedlin)*

ISSN 0283-8680

Tidigare utgivna *R & D Reports* kan beställas genom Ingvar Andersson, SCB, U/SIB, Box 24300, 115 81 STOCKHOLM (telefon 08-783 41 47, fax 08-783 45 99, e-post ingvar.andersson@scb.se).
R & D Reports listed above as well as issues from 1988-1994 can - in case they are still in stock - be ordered from Statistics Sweden, att. Ingvar Andersson U/SIB, Box 24300, S-115 81 STOCKHOLM (telephone +46 8 783 41 47, fax +46 8 783 45 99, e-mail ingvar.andersson@scb.se).

SCB-Tryck, Örebro 1998.05