Contents

Chapter 1

Introduction

1.1 Frame coverage errors

In the early stage of planning a sample survey, a decision must be made on what frame to use for sampling. Sometimes, there are several frames available. In other instances, there is no suitable frame at hand, but there is a need to construct one. The significance of the choice of sampling frame ought not to be underestimated. If the sampling frame has some serious shortcomings, the quality of the final output, the survey estimates, can be questioned. More precisely, frame imperfections may bias and increase the variance of the estimators.

The purpose of the sampling frame is to offer 'observational access' to the target population. Ideally, the population linked to the frame in use is identical to the target population. If some elements of the target population lack association to the frame, this is usually referred to as undercoverage. If the frame contains some elements that are not part of the target population, the frame is correspondingly said to suffer from overcoverage. Aside from coverage errors, a sampling frame may for example have multiplicity problems or contain erroneous or incomplete auxiliary information. For a thorough treatment of various frame problems, and references, see Lessler and Kalsbeek [8].

To see the bias effect of frame coverage errors, consider the problem of estimating a population total $t = \sum_U y_k$, where U denotes the target population, y_k is the study variable value for element $k \in U$, and $\sum_U y_k$ is a shorthand for $\sum_{k \in U} y_k$. To simplify, we assume that all y_k are positive. Let

U_F denote the set of elements included in the frame. Assume that the frame has either undercoverage ($U_F \subset U$) or overcoverage ($U \subset U_F$) but no other imperfections. The total for the frame population is $t_F = \sum_{U_F} y_k$. A sample s_F of elements is drawn from U_F by some (without-replacement) sampling design. The Horvitz-Thompson estimator of t_F is $\hat{t}_{F\pi} = \sum_{s_F} y_k/\pi_k$, where π_k is the probability of including element $k \in U_F$ in the sample. If the frame has undercoverage, $t_F \leq t$ and $\hat{t}_{F\pi}$ has a negative bias as estimator of t, whereas in the case of frame overcoverage, $t_F \geq t$ and $\hat{t}_{F\pi}$ is a positively biased estimator of t.

Remark 1.1.1 *Throughout this thesis, the system of notation (and the way of tackling the inference problem in survey sampling) is taken mainly from Särndal et al. [11]. This means, among other things, that we use the term* π estimator *for the well-known Horvitz-Thompson estimator, and indicate it with a* π*. Also, the sampling designs relevant in this thesis are abbreviated as follows: SI for simple random sampling without replacement, SIR for simple random sampling with replacement, pps for probability-proportional-to-size sampling with replacement, and STSI for stratified sampling with SI sampling applied in all strata.*

1.2 Formulation of the problem

This thesis addresses a frame coverage problem urgent in a particular road traffic survey, the Survey on Vehicle Speeds (SVS), conducted annually by the Swedish National Road Administration (SNRA) since 1996. The aim of the survey is to estimate parameters such as the total vehicle mileage and, most importantly, the average vehicle speed, for the Swedish urban road network. Hypothetically, the target population, the road network, is viewed as partitioned into one-meter road sites, which represent the population elements. A three-stage sampling design is employed, where the primary sampling units are population centers and the secondary sampling units are small areas. For each selected small area, a frame of the road network is employed. The frame units are road links, and the frame contains information on the length of each link. From each small area frame road network, an *SI* sample of road sites is selected for observation.

When the small area frames were constructed, the link lengths were

determined manually from maps. Hence, the lengths may be subject to measurement errors. It is not quite obvious how this frame problem should be examined. One may say that the frame suffers from coverage errors, or that, due to faulty auxiliary information in the frame, incorrect element-inclusion probabilities are used.

The coverage error view Apart from rounding errors, a link length corresponds to a geographically ordered vector of population elements. If a road link is shorter in the frame than in reality, this corresponds to an undercoverage of target elements. Correspondingly, if a frame link is too long, the frame suffers from overcoverage.

The incorrect inclusion probabilities view The length is known for each frame unit (road link) prior to sampling; thus, length can be thought of as an auxiliary variable. If a road link is shorter or longer in the frame than in reality, this corresponds to an incorrect auxiliary variable value. For a given small area, the sum of all link lengths in the frame is supposed to be the number of road sites that make up the road network (the population). If this summed length is in error, but the sample of road sites actually is selected from the target population, the inclusion probabilities that are used for sampled road sites are incorrect.

The latter view is somewhat more general, since incorrect inclusion probabilities may arise for other reasons in other types of surveys. However, for our purposes, it does not really matter how we decide to entitle the problem.

Discrepancies between measured and actual link lengths have implications on the data collection stage of the survey. This follows since, in the presence of erroneous frame link lengths, the instructions to the field staff may no longer hold. Field staff are told to seek out a sampled road site located a certain number of meters into a specified link. In reality, the location may simply not exist, if the link is shorter than the frame says. If the link in reality is much longer than the frame says, the location will indeed exist, but on different places depending on from what direction the link is entered. In each case, the field staff adjust to real-life conditions by observing the traffic "somewhere" along the designated link.

Our aim is to investigate the bias and variance of the employed estimators under these circumstances. To simplify, all possible frame imperfections

other than erroneous road lengths are ignored. In particular, we assume that the small area frames list all the links in the areas correctly.

1.3 Our approach to the problem

Among the essential features of our approach, consider first the problem of estimating a total t for one small area. The target population U is the set of one-meter road sites (of size N) that make up the area road network. The frame population U_F of size N_F is the set of road sites represented in the frame. From U_F, an *SI* sample s_F of road sites of size n is drawn. Then, the probability π_k of including road site $k \in U_F$ in the sample is n/N_F. The resulting estimator of t is $\hat{t}_{F\pi} = N_F \bar{y}_{s_F}$, where $\bar{y}_{s_F}$ is the sample mean. Since the frame may have both undercoverage and overcoverage, the sign of possible bias of $\hat{t}_{F\pi}$ is unknown. We handle this situation by assuming that s_F in practice can be regarded as an *SI* sample s from U. Then, the only remaining difference between the unbiased estimator of t, $\hat{t}_\pi = N\bar{y}_s$, and the employed estimator, $\hat{t}_{F\pi}$, is that the latter is weighted by N_F instead of N. Due to measurement errors associated with the frame construction process, N_F may deviate from the true road length N. A model is stated in which N_F is viewed as composed of N and a random error ζ. The expectation and variance of $\hat{t}_{F\pi}$ is derived jointly with respect to the sampling design (conditional on stages one and two) and the error model.

This was the basic idea of our approach. In reality, we are interested in the statistical properties of employed estimators of totals and ratios not for a single small area but for the whole road network. Then, the full three-stage sampling design is taken under consideration. In our general model for N_F, we do not specify the relation between the true value and the error. As special cases, however, we derive results for a simple additive and multiplicative error structure. Emphasis is given to the latter, which we believe to be the most realistic. In addition to our theoretical derivations, we present results from an empirical study of the errors in the frame.

As far as we know, most studies of traffic characteristics are based on nonprobability samples. Instead of choosing road sites at random from a frame, efforts are made (by visual inspection of the road) to pick "representative" sites for observation. It is therefore not very surprising that we have not seen this frame problem treated in the traffic research literature.

The statistical literature on frame errors, on the other hand, mainly deals with errors in sampling frames used in surveys of individuals or households. The conditions of such surveys differ substantially from those in the SVS, so the methods suggested for evaluating the impact of coverage errors are not quite applicable to our problem.

Our work is, however, inspired by the approaches to measurement errors discussed in, among others, Biemer and Stokes [2] and Särndal et al. [11, Ch. 16]. In this field of research, a survey is viewed as a two-stage process such that each stage contributes with randomness to the estimators. The first stage, the sample selection, determines what part of the population to observe. The second stage is the measurement procedure, which generates an observation for each element in the sample. Unlike traditional sampling theory, the observations are not presupposed to coincide with the true values, but assumed to be subject to random errors. In order to evaluate the impact of measurement errors on the estimators, the relation between observed and true values is modeled.

1.4 Outline of the thesis

This thesis is arranged as follows. **Chapter 2** presents the main methodological features of the vehicle speed survey. The target population, the main study variables, and the related parameters are described, as well as how the variables of study are measured. Since this is a sample survey, we also describe the sampling design, the sampling frames in use and the employed estimators of the parameters of interest. **Chapter 3** is devoted to the formulation and use of our error model. We start by stating the model and discussing its plausibility. Then, we derive the expected values and variances of the survey estimators with respect jointly to the sampling design and the model. At the outset, neither a particular error structure nor identically distributed errors are assumed. Some special cases are, however, also examined, in particular the case of multiplicative errors with equal expectation and variance. **Chapter 4** treats the execution and results of the empirical study. In **Chapter 5**, some consequences of using erroneous inclusion probabilities when estimating a population total are investigated. This chapter, which does not relate directly to the vehicle speed survey, can be read independently of the others. Finally, in **Chapter 6**, we summarize

our findings and make some suggestions for further work.

Chapter 2

The survey on vehicle speeds

2.1 Background and survey objectives

Since 1993, the Swedish National Road Administration (SNRA) has had overall responsibility for traffic safety work in Sweden. The model for this work is the National Road Traffic Safety Programme for 1995-2000 [12]; on a long view, its goal is that nobody should be killed or seriously injured as a result of traffic accidents. The work is organized in focus areas, called 'road traffic safety reforms,' such as "reduction in speeding offences," "use of cycle helmets," and "use of safety equipment in cars." Operational goals are stated for each reform, and it is assumed that if a reform goal is reached, this should contribute to a reduction in traffic deaths and injuries.

In order to assess whether development is heading toward the reform goals, the SNRA conducts several surveys regularly. The largest of them, the Survey on Vehicle Speeds (SVS), aims at measuring the results of efforts within the traffic safety reforms "reduction in speeding offences" and "reduction in other driving offences." There are many possible driving offences other than exceeding speed limits, but in the "other" category the survey deals only with that of driving too close to the vehicle ahead ("too short headway").

2.2 Population

The target population is the entire Swedish road network except rural private roads. It is divided into two subpopulations of special interest – state roads and 'urban' roads (local authority roads and private roads in built-up areas) – that also serve as strata when the sampling is conducted. In this thesis, we restrict our attention to the part of the survey that concerns the urban roads, and refer throughout to the urban road network as the target population. We think of this road network as partitioned into one-meter road sites, which we call 'measurement locations,' that are the population elements.

From the target population, the following road sections are excluded.

- From major roads: 100 meters before and after each intersection with traffic lights.
- From non-major roads: 100 meters before and after each intersection.

The main reason for excluding road sections close to intersections is to avoid observational difficulties. In the SVS, observations are carried out by measurement equipment installed on the road (see Section 2.5). Certain traffic situations, such as vehicles lining up, accelerating, or decelerating, have the potential to cause measurement problems. Such situations frequently occur close to intersections.

Survey results are demanded not only for the whole target population, but also for specific subpopulations or 'domains.' One important goal of the survey is to provide results for each SNRA region. The SNRA organization includes seven regional road management directorates (Figure 2.1), which are responsible for the SNRA's regional management, including traffic safety work within their geographic areas.

The definition of the target population is not complete without some restriction in time. The SVS is always conducted during the summer months, but the exact period of study changes somewhat from year to year. In the last survey round, in 1999, the study period was June 14 to September 30. The period of study is thought of as a population in time with twentyfour-hour periods as population elements.

2.2. Population

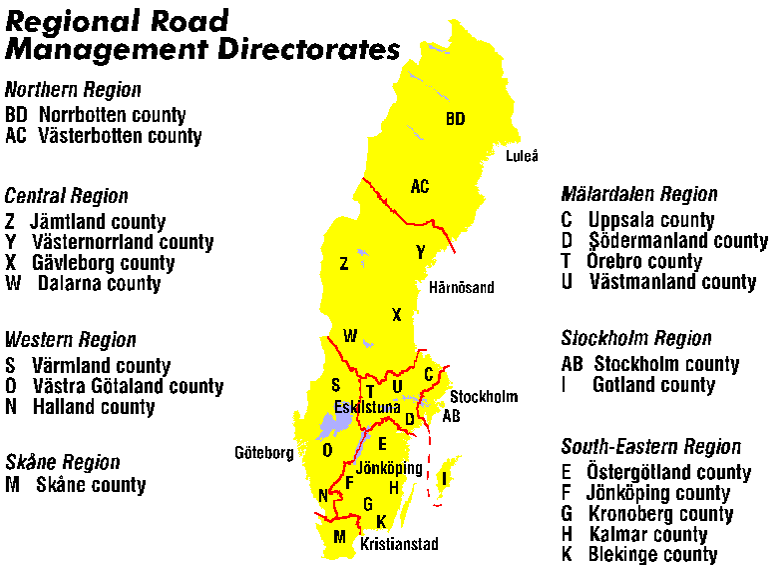

Figure 2.1: The SNRA Regional Road Management Directorates. (Source: SNRA)

2.3 Variables and parameters

The primary study variable in the SVS is the traffic flow, y. In general, for a given point on a road and a specified period of time, the traffic flow is defined as the number of passing vehicles. Since in this survey a road is viewed as made up of one-meter sections, a 'point' is interpreted as a one-meter section (a measurement location). From the total flow, other study variables of interest can be derived; some examples include:

- Flow above a certain speed limit
- Flow with less than a certain headway
- Flow by a certain type of vehicle (e.g., by cars with trailers)

The second main study variable is the travel time, z. For a given traffic flow, the travel time is the total time all vehicles take to pass the road point.

Let U denote the target population "in space" – the set of measurement locations that make up the urban road network – and U_Υ the target population "in time" – the set of twentyfour-hour periods that make up the time period of study. The population total of the study variable y, the 'total vehicle mileage' for the road network and time under study, is $\sum_{U_\Upsilon}\sum_U y_k^\upsilon$, where y_k^υ equals the traffic flow in measurement location $k \in U$ during twentyfour-hour period $\upsilon \in U_\Upsilon$. Correspondingly, the population total of z, the 'total travel time,' is given by $\sum_{U_\Upsilon}\sum_U z_k^\upsilon$. Since the total vehicle mileage is a measure of distance, and the total travel time a measure of time, their ratio is a measure of speed.

In this thesis, we ignore possible time variability in y and z. That is, we consider only the special case when $y_k^\upsilon = y_k$ and $z_k^\upsilon = z_k$ for all $\upsilon \in U_\Upsilon, k \in U$. Hence, we will hereafter drop the time index and talk simply about the parameters $t_y = \sum_U y_k$, $t_z = \sum_U z_k$ and $R = t_y/t_z$.

2.4 Sampling design

2.4.1 Use of a master frame

The SNRA's traffic safety surveys share the same sampling in the first and second stages. In each survey, the final sample consists of measurement

locations selected from a master frame of roads. Depending on the nature of the survey, the locations can be intersections with traffic lights (suitable for observing motorists who drive against red light) or, as in the SVS, one-meter sections of the road.

The method of using a master frame is discussed by, among others, Kish [7, pp. 478-480], and can briefly be described as follows. Initially, a 'master sample' of sampling units is selected. For each sampled unit, a frame is prepared. The sample for a particular survey is then selected from these frames, which serve for a longer time period.

The master sample used for the SNRA surveys was selected during 1995-96 by a two-stage sampling design. The primary sampling units (PSUs) are population centers, and the secondary sampling units (SSUs) are small areas.

2.4.2 Frames

When the master sample was selected, the frame used in the first stage was a list, supplied by Statistics Sweden (SCB), of the Swedish population centers 1990. The list contained auxiliary information on the number of inhabitants in each population center, which served as a size measure for *pps* sampling (see Remark 1.1.1). The frames used in the second stage were lists of the small areas within selected population centers. In all essentials, these small areas agree with SCB's small area market statistics (SAMS) regions. Various population statistics collected by SCB are tied to developed properties. In co-operation with the local governments, SCB has grouped similar adjacent properties. By a special technique called 'register generated borders,' geographic borders between the groups have been fixed. The resulting nationwide area division is called SAMS. There are about 9,200 SAMS regions; their main use is for statistical presentations.

For each selected small area, a list frame of road links was prepared at the SNRA from city maps. Using the intersections as breakpoints, the map road network was partitioned into links, and the link lengths were determined manually by the use of map measurers.

2.4.3 Sample selection

In this section, the sampling design of the SVS is described. In each stage, stratified sampling is used. In the first stage of sample selection, the popu-

lation centers are stratified according to *SNRA region* (see Figure 2.1) and three *size classes*:

- Large major population center of a municipality
- Other major population center of a municipality
- Other population center

In the second stage, the small areas within a selected population center are stratified according to four *development types*:

- City
- Industrial
- Residential
- Other type

In the final stage, the road sites within a selected small area are stratified according to three *road types*:

- Major roads with a speed limit of 70 kilometers per hour (km/h)
- Major roads with a speed limit of 50 km/h
- Other roads

However, to simplify, the stratification in each stage is hereafter ignored. We also ignore the fact that in stage one, the three largest PSUs (Stockholm, Göteborg, and Malmö) define a take-all stratum. The subsequent sampling stages are somewhat different in the take-all stratum than described here. All stated sample sizes refer to one stratum.

Selection of the master sample

The PSUs are the N_I population centers in Sweden, labeled $i = 1, ..., N_I$. For simplicity, we represent the ith PSU by its label i. Thus, we denote the set of PSUs as $U_I = \{1, ..., i, ..., N_I\}$. Population center $i \in U_I$ is partitioned into N_{IIi} small areas, labeled $q = 1, ..., N_{IIi}$, that represent the SSUs. Again we

represent the sampling units by their labels; hence, the set of SSUs formed by the partitioning of i is denoted $U_{IIi} = \{1, ..., q, ..., N_{IIi}\}$.

The master sample of small areas was selected in the following way. In the first stage, a *pps* sample of PSUs was drawn with probability proportional to the number of inhabitants. At every draw, p_i is the probability of selecting the ith PSU. Let i_ν denote the PSU selected in the νth draw, $\nu = 1, ..., m_I$, where m_I is the number of draws. The probability of selecting i_ν is denoted p_{i_ν}. If the ith PSU was selected in the νth draw, then $p_{i_\nu} = p_i$. The vector of selected PSUs, $(i_1, ..., i_\nu, ..., i_{m_I})$, is the resulting ordered sample os_I.

In stage two, for every i_ν that is a component of os_I, an *SI* sample s_{IIi_ν} of SSUs of size n_{IIi_ν} was selected. The resulting sample of SSUs is the master sample.

In practice, the sample sizes in each stage were $m_I = 10$ and $n_{IIi_\nu} = 1$.

Selection of the SVS final-stage sample

The road network in small area q in population center i is viewed as partitioned into N_{iq} one-meter road sections or measurement locations – the population elements. This set of locations is U_{iq}. An *SI* sample $s_{i_\nu q}$ of locations of size $n_{i_\nu q}$ is drawn for every small area $q \in s_{IIi_\nu}$. In practice, the sample sizes are $n_{i_\nu q} = 1$. The sample of locations finally obtained is denoted s.

For every location $k \in s$, one twentyfour-hour period is randomly drawn from the time population U_Υ.

2.5 Data collection and processing

A sampled location is positioned a certain number of meters into a road link. The field staff search out the location and install measurement equipment to collect data during the selected twentyfour-hour period. The equipment consists of two pneumatic tubes stretched across the road in parallel, a fixed distance apart, and connected to a traffic analyzer. When a wheel crosses a tube, this changes the air pressure in the tube. The times of such events, or *pulses*, are registered by the traffic analyzer. The analyzer further combines the pulses into vehicles and calculates their travel direction, speed, and type

(such as car or truck, with or without trailer). In the combination process, some new pulses are fabricated and some registered pulses eliminated. The variables of study are later calculated from the vehicle data produced by the traffic analyzer.

2.6 Estimation

We will now describe how t_y and $R = t_y/t_z$ are estimated from the survey data. We start by noting that the sampling design of the SVS is such that:

i. The PSUs are selected with replacement.

ii. Independent subsampling is conducted from every selection of a PSU (whether a repetition or not).

Define the population totals $t_{yiq} = \sum_{U_{iq}} y_k$, $t_{yi} = \sum_{U_{Ii}} t_{yiq}$ and t_{ziq}, t_{zi}, respectively. Further, define $E_k = y_k - Rz_k$ and the corresponding totals $t_{Eiq} = t_{yiq} - Rt_{ziq}$ and $t_{Ei} = t_{yi} - Rt_{zi}$. Estimators of the population totals are denoted by a hat.

In standard sampling theory, the probability distribution of an estimator is determined entirely by the sampling design $p(s)$ and the parameter state. In the language of Cassel et al. [4, p. 26], an estimator is said to be *p-unbiased*, or *design unbiased*, for a parameter if its expected value with respect to $p(s)$ equals the true parameter value. The variance, with respect to $p(s)$, of an estimator is called the estimator's *p-variance*. Let E_p and V_p denote the expectation and variance operators with respect to $p(s)$[1]. For some nonlinear estimators, such as the ratio of two estimated population totals, it is the practice to use the variance of a linearized statistic as an approximation to the exact variance. Let AV_p denote such an approximative variance, again with respect to $p(s)$. For details on the linearization technique, see Särndal et al. [11, Sec. 5.5].

The principal appearances of the estimators employed in the SVS are

$$\hat{t}_y = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{\hat{t}_{yi_\nu}}{p_{i_\nu}} \tag{2.1}$$

[1] Note the use of capital E as notation both for the variable $y - Rz$ and for the expectation operator.

of t_y and $\hat{R} = \hat{t}_y/\hat{t}_z$ of R. If $i \in U_I$ was selected in the νth draw, then $\hat{t}_{yi_\nu} = \hat{t}_{yi}$. The expected values and variances of $\hat{t}_y$ and $\hat{R}$ are investigated in the following lemma.

Lemma 2.6.1 *Under a sampling design $p(s)$ satisfying the specifications (i)-(ii), the expected value of $\hat{t}_y$ is $E_p\left(\hat{t}_y\right) = \sum_{i=1}^{N_I} E_p\left(\hat{t}_{yi}\right)$. The p-variance of $\hat{t}_y$ is*

$$V_p\left(\hat{t}_y\right) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{E_p\left(\hat{t}_{yi}\right)}{p_i} - \sum_{i=1}^{N_I} E_p\left(\hat{t}_{yi}\right)\right)^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_p\left(\hat{t}_{yi}\right)}{p_i}. \quad (2.2)$$

The estimator $\hat{R}$ has the approximate expected value $E_p\left(\hat{t}_y\right)/E_p\left(\hat{t}_z\right)$, and the approximate p-variance

$$AV_p\left(\hat{R}\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{\left[E_p\left(\hat{t}_{Ei}\right)\right]^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_p\left(\hat{t}_{Ei}\right)}{p_i}\right\}. \quad (2.3)$$

The part of Lemma 2.6.1 that refers to $\hat{t}_y$ is a slight generalization of Result 4.5.1 in [11] or Theorem 6.4 in [10], and the part that refers to $\hat{R}$ a slight generalization of results presented in [10, Sec. 6.8.2.]. Unlike the cited sources, we do not presuppose that the estimators $\hat{t}_{yi}$ and $\hat{t}_{Ei}$ are p-unbiased for t_{yi} and t_{Ei}, respectively. Later in this thesis, an expanded version of Lemma 2.6.1 is used to derive the expected values and variances of the survey estimators with respect jointly to the sampling design and an error model.

In the SVS, if the road lengths in the final-stage frames are in error, the actual sampling procedure differs from the one described in Section 2.4.3. Let U_{Fiq} denote the set of measurement locations (of size N_{Fiq}) in (i,q) *according to the frame*. For every small area $q \in s_{IIi_\nu}$, an *SI* sample $s_{Fi_\nu q}$ of locations (of size $n_{i_\nu q}$) is drawn from U_{Fiq}. In the data-collection stage, the field staff adjust to the real road network when installing the measurement equipment. Consequently, the set of locations actually observed may differ from $s_{Fi_\nu q}$. We do not, however, introduce any special notation to distinguish between these sets. The sample of locations finally obtained (as well as the sample finally observed) is denoted s_F. The relevant totals are:

- $t_{Fy} = \sum_{U_I} t_{Fyi}$

- $t_{Fyi} = \sum_{U_{IIi}} t_{Fyiq}$
- $t_{Fyiq} = \sum_{U_{Fiq}} y_k$

The totals t_{Fz}, t_{Fzi} and t_{Fziq} are defined correspondingly. Further, $R_F = t_{Fy}/t_{Fz}$, $t_{FEiq} = t_{Fyiq} - R_F t_{Fziq}$ and $t_{FEi} = t_{Fyi} - R_F t_{Fzi}$. Estimators of the population totals are denoted by a hat. In addition, π estimators (Horvitz-Thompson estimators) are denoted by a π.

The estimator of t_y in use in the SVS is

$$\hat{t}_{Fy} = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{\hat{t}_{F\pi y i_\nu}}{p_{i_\nu}} \tag{2.4}$$

where $\hat{t}_{F\pi y i_\nu} = (N_{IIi_\nu}/n_{IIi_\nu}) \sum_{s_{IIi_\nu}} \hat{t}_{F\pi y i_\nu q}$ and $\hat{t}_{F\pi y i_\nu q} = (N_{Fi_\nu q}/n_{i_\nu q}) \sum_{s_{Fi_\nu q}} y_k$. If $i \in U_I$ was selected in the νth draw, then $\hat{t}_{F\pi y i_\nu} = \hat{t}_{F\pi y i}$ and $\hat{t}_{F\pi y i_\nu q} = \hat{t}_{F\pi y i q}$. The employed estimator of R is $\hat{R}_F = \hat{t}_{Fy}/\hat{t}_{Fz}$.

Consider the special case when, for every small area q included in the master sample, the frame road length N_{Fiq} equals the true length N_{iq}, and s_{Fiq} is an *SI* sample from U_{iq}. Then, the estimators $\hat{t}_{Fy}$, $\hat{t}_{F\pi yi}$, and $\hat{t}_{F\pi yiq}$ are design-unbiased for t_y, t_{yi}, and t_{yiq}, respectively, and the frame index F is no longer needed.

Chapter 3

Model approach

3.1 The model

In order to evaluate $\hat{t}_{Fy}$ and $\hat{R}_F$ as estimators of t_y and R, respectively, we formulate the following frame error model:

(1) The sample s_{Fiq} is an *SI* sample from U_{iq}. In mathematical terms, we assume that $s_{Fiq} = s_{iq}$.

(2) The frame road length N_{Fiq} is a function of the true length N_{iq} and a random error ζ_{iq}.

(3) All N_{Fiq}'s are independent random variables with expected values μ_{iq} and variances σ^2_{iq}.

In cases of unclear instructions due to frame errors, the field staff place the measurement equipment 'somewhere' along designated links. Then, Assumption (1) holds if the road sections within the link can be considered 'randomly ordered,' or if the field staff randomly choose a road section within the designated link for measurement. The field staff's choice of a road section within the link is probably more adequately described as haphazard than as random. This follows since, when deciding upon a location, they pay regard to the road environment (e.g., by avoiding locations where cars parked by the roadside may obstruct the installation of the equipment). A 'random ordering' of the road sections is however, for the following reason, quite likely. As described in Section 2.2, only road sections located more

than one hundred meters from an intersection are included in the target population. Results from a pilot study [3] suggest that, within a link, the remaining road sections are reasonably similar with respect to the study variables. Consequently, it is not crucial which road section within link that is actually measured – the result will be about the same anyway.

Under Assumption (1), the only remaining difference between $\hat{t}_{\pi yiq}$ and the error-prone estimator $\hat{t}_{F\pi yiq}$ is that the latter is weighted by N_{Fiq} instead of N_{iq}. A random error model for N_{Fiq} is stated in Assumptions (2)-(3), but to be really useful the model needs further specification. The two most simple error structures are the additive error model,

$$N_{Fiq} = N_{iq} + \zeta_{iq} \tag{3.1}$$

and the multiplicative error model,

$$N_{Fiq} = N_{iq}\zeta_{iq} \tag{3.2}$$

We denote the expected value and variance of the random error ζ_{iq} with θ_{iq} and τ^2_{iq}, respectively. Then, under the additive error model, $\mu_{iq} = N_{iq} + \theta_{iq}$ and $\sigma^2_{iq} = \tau^2_{iq}$, whereas under the multiplicative error model, $\mu_{iq} = N_{iq}\theta_{iq}$ and $\sigma^2_{iq} = N^2_{iq}\tau^2_{iq}$. Note that, depending on the assumed error structure, θ_{iq} and τ^2_{iq} are expected to take quite different numerical values. Consider, for instance, the case when the road length measurements are rather accurate, so that μ_{iq} approximately equals N_{iq}. Under the additive error model, this occurs when θ_{iq} is close to zero; under the multiplicative error model when θ_{iq} is close to one.

3.2 Statistical properties of the error-prone estimators

3.2.1 Unspecified error structure

In this section, we express the expected value and variance of the estimators $\hat{t}_{Fy}$ and $\hat{R}_F$, taking into consideration that their probability distributions are determined jointly by the sampling design $p(s)$, the frame error model m in Section 3.1, and the parameter state. So far, we do not make any assumption on the error structure.

We call an estimator *pm-unbiased* if its expected value with respect to $p(s)$ and m equals the true parameter value. The estimator's *pm-variance* is defined correspondingly. Let E_{pm}, V_{pm}, and AV_{pm} denote the expectation, variance, and approximate variance, respectively, with respect to the sampling design and error model jointly.

In Lemma 2.6.1, the expectations and variances were taken with respect only to the sampling design. In the following lemma, a straightforward expansion of Lemma 2.6.1, the expectations and variances are taken with respect jointly to the sampling design and the error model.

Lemma 3.2.1 *Jointly under a sampling design $p(s)$ satisfying the specifications (i)-(ii) in Section 2.6, and the error model m in Section 3.1, the expected value of $\hat{t}_{Fy}$ is given by $E_{pm}\left(\hat{t}_{Fy}\right) = \sum_{i=1}^{N_I} E_{pm}\left(\hat{t}_{Fyi}\right)$. The pm-variance of $\hat{t}_{Fy}$ is given by*

$$\begin{aligned} V_{pm}\left(\hat{t}_{Fy}\right) &= \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{E_{pm}\left(\hat{t}_{Fyi}\right)}{p_i} - \sum\nolimits_{i=1}^{N_I} E_{pm}\left(\hat{t}_{Fyi}\right)\right)^2 \\ &\quad + \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{V_{pm}\left(\hat{t}_{Fyi}\right)}{p_i}. \end{aligned} \tag{3.3}$$

The estimator $\hat{R}_F$ has the approximate expected value $E_{pm}\left(\hat{t}_{Fy}\right)/E_{pm}\left(\hat{t}_{Fz}\right)$, and the approximate pm-variance

$$AV_{pm}\left(\hat{R}_F\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{\left(E_{pm}\left(\hat{t}_{FEi}\right)\right)^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{pm}\left(\hat{t}_{FEi}\right)}{p_i}\right\}. \tag{3.4}$$

We are now ready for the following theorem.

Theorem 3.2.1 *Jointly under the SVS sampling design and the error model m in Section 3.1, the expected value of $\hat{t}_{Fy}$ is given by*

$$E_{pm}\left(\hat{t}_{Fy}\right) = \sum\nolimits_{i=1}^{N_I} E_{pm}\left(\hat{t}_{F\pi yi}\right) \tag{3.5}$$

where $E_{pm}\left(\hat{t}_{F\pi yi}\right) = \sum_{U_{II_i}}\left(\mu_{iq}/N_{iq}\right)t_{yiq}$. The pm-variance of $\hat{t}_{Fy}$ is given by

$$\begin{aligned} V_{pm}\left(\hat{t}_{Fy}\right) &= \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{1}{p_i}\sum\nolimits_{U_{II_i}}\frac{\mu_{iq}}{N_{iq}}t_{yiq} - \sum\nolimits_{i=1}^{N_I}\sum\nolimits_{U_{II_i}}\frac{\mu_{iq}}{N_{iq}}t_{yiq}\right)^2 \\ &\quad + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{pm}\left(\hat{t}_{F\pi yi}\right)}{p_i} \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} V_{pm}\left(\hat{t}_{F\pi yi}\right) &= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}} t_{yiq} - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{yiq}\right)^2 \\ &\quad + \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}}\right)^2 V_p\left(\hat{t}_{\pi yiq}\right) \\ &\quad + \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left(\frac{\sigma_{iq}}{N_{iq}}\right)^2 \left[V_p\left(\hat{t}_{\pi yiq}\right) + t_{yiq}^2\right]. \end{aligned}$$

The estimator $\hat{R}_F$ is approximately pm-unbiased for $E_{pm}\left(\hat{t}_{Fy}\right)/E_{pm}\left(\hat{t}_{Fz}\right)$, with the approximate pm-variance

$$AV_{pm}\left(\hat{R}_F\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\sum\nolimits_{U_{IIi}}\left(\frac{\mu_{iq}}{N_{iq}}\right)^2 t_{Eiq}^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{pm}\left(\hat{t}_{F\pi Ei}\right)}{p_i}\right\} \tag{3.7}$$

where $V_{pm}\left(\hat{t}_{F\pi Ei}\right)$ is obtained from $V_{pm}\left(\hat{t}_{F\pi yi}\right)$ by replacing t_{yiq} with t_{Eiq} and $\hat{t}_{\pi yiq}$ with $\hat{t}_{\pi Eiq}$.

The proof is given in Appendix A.1.

From Theorem 3.2.1, results can be derived for various situations of interest. An important special case is when the frame road lengths N_{Fiq} are 'unbiased' – that is, if in a (hypothetical) long run of repeated length measurements on the same small area road network, the average of the obtained values will equal the true value N_{iq}. This case is treated in the following corollary.

Corollary 3.2.1 *If the frame road lengths N_{Fiq} have expected value N_{iq}, the estimator $\hat{t}_{Fy}$ is pm-unbiased for t_y, and $\hat{R}_F$ is approximately pm-unbiased for R. The use of $\hat{t}_{Fy}$ instead of $\hat{t}_y$ as estimator of t_y increases the variance by*

$$V_{pm}\left(\hat{t}_{Fy}\right) - V_p\left(\hat{t}_y\right) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\left(\frac{\sigma_{iq}}{N_{iq}}\right)^2\left[V_p\left(\hat{t}_{\pi yiq}\right) + t_{yiq}^2\right]. \tag{3.8}$$

The variance increase due to the use of $\hat{R}_F$ instead of $\hat{R}$ as estimator of R is obtained from (3.8) by multiplying by t_z^{-2} and replacing t_{yiq} with t_{Eiq} and $\hat{t}_{\pi yiq}$ with $\hat{t}_{\pi Eiq}$.

Corollary 3.2.1 is easily derived from Theorem 3.2.1 by replacing μ_{iq} with N_{iq}.

3.2.2 Results for specified error structures

It is straightforward to adapt Theorem 3.2.1 to various error structures of interest. By replacing μ_{iq} with $N_{iq}+\theta_{iq}$ and σ_{iq}^2 with τ_{iq}^2, results are obtained for the additive error model in (3.1), whereas by replacing μ_{iq} with $N_{iq}\theta_{iq}$ and σ_{iq}^2 with $N_{iq}^2\tau_{iq}^2$, we get results for the multiplicative error model in (3.2).

In the remainder of this section, we will only look at the model we a priori believe to be the most realistic: the multiplicative error model with equal error expectations θ and variances τ^2. The multiplicative error model means that the error associated with N_{Fiq} depends on the true length N_{iq} – a view we regard as intuitively appealing. For example, it is probably harder to obtain accurate measurements for areas with extensive road networks, since such networks usually are partitioned into a large number of links. (Remember that each link length was measured separately.) Further, we have no reason to believe the error expectations and variances to differ between population centers or small areas. The same tool, a map measurer, was used everywhere, and the staff performing the measurements were given the same training. An important objective to the multiplicative model is that it states that the variances of the frame road lengths, σ_{iq}^2, are proportional to the squared true lengths. It is not obvious that this assumption holds; an equally natural assumption is that the variances are proportional to the (unsquared) lengths.

For the assumed model, the following corollary applies.

Corollary 3.2.2 *If the frame road lengths N_{Fiq} have expected value $N_{iq}\theta$, the pm-bias of $\hat{t}_{Fy}$ as estimator of t_y is given by $t_y\,(\theta - 1)$.*

Corollary 3.2.2 is easily derived from Theorem 3.2.1 by replacing μ_{iq} with $N_{iq}\theta$. Note that if θ equals one, both $\hat{t}_{Fy}$ and $\hat{R}_F$ are unbiased or approximately unbiased.

Let us proceed by investigating the variances when θ equals one.

Corollary 3.2.3 *If the frame road lengths N_{Fiq} have expected value N_{iq} and variance $N_{iq}^2\tau^2$, the use of $\hat{t}_{Fy}$ instead of $\hat{t}_y$ as estimator of t_y increases the*

variance by

$$V_{pm}\left(\hat{t}_{Fy}\right) - V_p\left(\hat{t}_y\right) = \tau^2 \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left[V_p\left(\hat{t}_{\pi yiq}\right) + t_{yiq}^2\right] \tag{3.9}$$

The approximate variance increase due to the use of $\hat{R}_F$ *instead of* $\hat{R}$ *as estimator of* R *is given by*

$$AV_{pm}\left(\hat{R}_F\right) - AV_p\left(\hat{R}\right) = \tau^2 AV_p\left(\hat{R}\right) \tag{3.10}$$

The proof of Corollary 3.2.3 is given in Appendix A.2.

3.3 Discussion

It seems reasonable to expect the measurements of road lengths to "on the average" be correct. The major sources of errors in the measurements are probably the map measurer tool producing 'shaky' results and the haste under which the measurements were performed. We have no reason to believe that these errors have a systematic influence on the measurement values. If our expectation is correct, the length error does not introduce bias in the estimators – a very encouraging result. Of course, there will still be a loss of precision due to the variability of the frame lengths.

The length of a small area road network may be viewed as a measure of the degree of difficulty of the measurement task. With this view, the multiplicative error model makes sense. Analytically, things get especially simple if these errors have the same expectations and variances; this also seems like a realistic assumption. For this case, 'unbiased' road length measurements corresponds to an error expectation equal to one. If this is fulfilled, the length error implies a relative variance increase in the estimator of average speed that is simply equal to the error variance. This variance is likely to be numerically small, since the multiplicative errors are 'relative.'

Although variance estimation is not really an issue in this thesis, we would nevertheless like to comment briefly on it here. The SVS point estimator of a total is the mean of independent and identically distributed (iid) random variables. An unbiased estimator of its variance is the sample variance, divided by the number of observations. This holds whether a length error is present or not. A nearly unbiased estimator for the variance of the

estimator of the ratio is constructed in a similar manner. Hence, the variance estimates calculated in the SVS hold also in the presence of a length error.

Chapter 4

Empirical study

4.1 Study objectives

In Chapter 3, by use of an error model, we investigated theoretically the impact of erroneous frame road lengths on the estimators $\hat{t}_{Fy}$ and $\hat{R}_F$. Hence a theoretical foundation is laid, but it needs to be complemented by knowledge about the real road length errors in the frame. Then, a choice of a realistic error structure can be made, the constant error expectations and variances assumption can be evaluated and, if proved to hold, θ and τ^2 can be estimated. To gain this knowledge, we conducted an experiment, the design and analysis of which we now present.

4.2 Design of the study

Data on the frame road length errors were collected in the following way. From the 469 small areas included in the master sample, 70 small areas were selected. A controller measured all the links in selected areas and fed the result into computer files. In the course of the work, the controller had access only to the originally used maps with the intersections numbered. Hence, for a small area, she started by making a list of all the links found on the map (using the existing numbering) and then measured them one after the other.

In the selection of small areas for the experiment, we wanted areas from different SNRA regions and from population centers of various sizes. Fur-

thermore, we wanted the areas to represent different development types. Note that SNRA region, population center size class, and development type were all used as stratification variables in the sample selection (see Section 2.4.3). An SNRA region effect was possible since, when the frame was constructed, each regional office was responsible for the work in its region, including the length measurements. Population center size and development type may correlate with the quality of available maps. To accomplish the desired dispersion of small areas, they were randomly selected within SNRA region, population center size class, and small area stratum.

For at least two reasons, the measurement values obtained in the study are probably more accurate than the frame values. Above all, when the frame was constructed, the road length measurements were made hurriedly (the entire construction work was behind the schedule). Our controller was not put under time pressure; on the contrary, she was encouraged to give priority to carefulness and to take her time. Also, when the frame was constructed, the road lengths were determined by use of a digital map measurer. This tool is convenient to use, since it can be programmed to produce length data in meters for a map with a specified scale. In our experience however, the tool is over-sensitive to the user's hand movements. The controller used a less sophisticated instrument, a common ruler, which we believe is less subject to measurement errors.

4.3 Data processing

For five of the chosen areas, the available maps were of such poor quality that the links could not be identified or measured properly. Therefore, those areas were entirely omitted from the analysis. From each remaining area, we excluded the links known to be administered by the state, as well as road links that did not occur both in the frame and in the controller's list. In practice, we applied (in turn) the following rules for excluding road links:

1. Road links, found in control, that are missing in the frame.
2. Road links that, according to the frame, are state authority roads.
3. Road links included in the frame that, according to the control, do not exist.

	Number	Per cent of original no.
Links in original data set	4123	100
Left after rule 1 applied	4013	97.3
Left after rule 2 applied	3762	91.2
Left after rule 3 applied	3618	87.8

Table 4.1: Exclusion of road links.

The resulting gradual reduction of the original data set (the set of all links occurring either in the frames or in the controller's lists) is shown in Table 4.1. The allocation of the 65 small areas over SNRA regions, population center size classes and small area development types is shown in Table 4.2. In the table, the following numbering of size classes is used: '1' for large major population center of a municipality, '2' for other major population center of a municipality, and '3' for other population center. Also, the small area development types are assigned the numbers '1' for city, '2' for industrial, '3' for residential, and '4' for other areas.

In the data processing, we encountered several frame quality problems other than erroneous road lengths. First, remember that we had to give up five chosen areas because of bad maps. Most likely, the frames in use for these areas are not, in general, very reliable. Second, we see in Table 4.1 that 110 links turned out to be missing in the frame and that 144 urban road links that were included in the frame could not be found by the controller. We take these figures as a warning signal that the frame may suffer from some serious coverage errors regarding road links. Finally, as a result of incorrect frame link lengths, some links may erroneously be excluded from or included in the target population. Among the non-major road links in our reduced data set, 42 links were shorter than 200 meters in the frame but longer than 200 meters in the control, while 34 links were longer than 200 meters in the frame but shorter than 200 meters in the control.

Like erroneous frame road lengths, all the frame imperfections discussed above may lower the quality of the survey estimates. In this thesis, we restrict our attention solely to the length problem. An expanded study would be needed in order to judge the influence and relative importance of all frame imperfections on the total error of the estimates.

Population center size class	1				2				3			
Small area development type	1	2	3	4	1	2	3	4	1	2	3	4
Central Region	1	1	1	1	0	1	1	1	1	0	1	1
Mälardalen Region	1	1	0	1	1	0	1	1	-	1	1	1
Northern Region	1	0	1	1	1	1	0	1	-	1	2	1
Skåne Region	1	1	1	0	1	0	1	1	-	-	2	1
Stockholm Region	0	2	0	0	1	1	1	1	-	1	1	1
South-Eastern Region	2	0	0	1	1	1	1	0	-	-	4	-
Western Region	1	1	1	1	1	1	0	1	-	-	1	-

Table 4.2: Number of small areas included in the analysis, by SNRA region, population center size class and small area buildings type. Non-existing strata are indicated by hyphens.

4.4 Analysis

Assume that the road link lengths according to the control are the true lengths. Then, by summing the frame link lengths for a small area (i, q), we get an observation on N_{Fiq}, and by summing the link lengths according to the control, we get N_{iq}. Under the additive error model in Equation (3.1), the error in the frame road length N_{Fiq} is given by $\zeta_{iq} = N_{Fiq} - N_{iq}$, whereas under the multiplicative error model in Equation (3.2), the error is given by $\zeta_{iq} = N_{Fiq}/N_{iq}$. For the 65 small areas comprised by our analysis, the errors were calculated under both the additive and the multiplicative error model (see Figure 4.1). We see that in the additive case, the points scatter around an imaginary horizontal line placed at a level close to zero, whereas in the multiplicative case, the scatter is around a line at a level close to one. Hence, under both error models, data suggest that the frame road lengths, on the average, are correct. In the additive case, the variance for the scatter of points seems constant, exactly as we had hoped for. In the multiplicative case however, the point scatter shows a tendency to narrow with the true length. This is a sign that the variance of the frame road lengths rather is proportional to the true length than to the squared true length (as the model states). However, due to the shortage of observations for large values of the true length, it is hard to draw any certain conclusions.

In our study design, population center size classes are nested under the

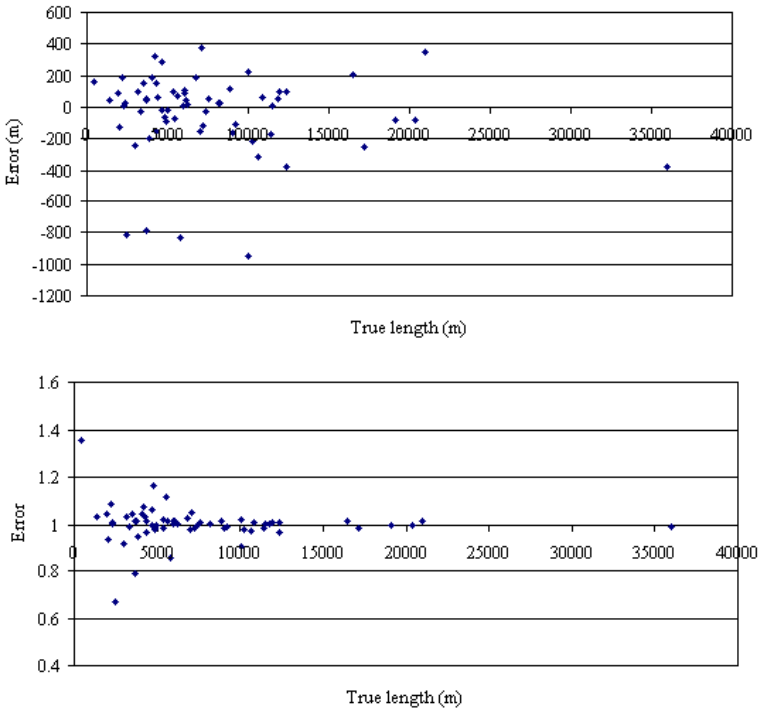

Figure 4.1: Observed errors under additive error model (top) and multiplicative error model (bottom).

SNRA regions, and small area strata are nested under the size class levels. Thus, it is a three-stage nested design (see, e.g., [9]). To account for the design, we introduce some new notation. Consider again the length error for small area (i, q), ζ_{iq}. Let $\zeta_{iq} = \zeta_{rstq}$ if population center i is included in SNRA region r and size class s, and small area q is included in small area stratum (development type) t. The analysis of variance (ANOVA) model for our design is

$$\zeta_{rstq} = \alpha + \beta_r + \gamma_{s(r)} + \delta_{t(rs)} + \epsilon_{(rst)q} \tag{4.1}$$

where α is an overall mean, β_r is the random effect of the rth region, $\gamma_{s(r)}$ is the random effect of size class s within the rth region, $\delta_{t(rs)}$ is the random effect of small area stratum t within size class s within the rth region, and $\epsilon_{(rst)q}$ is a random error.

Each of the factors – region, size class, and small area stratum – has a small number of possible levels (7, 3, and 4, respectively). Nevertheless we consider these factors as random. Regarding the regional factor, we are not interested in the administrative division in itself, but rather in potential differences in the behavior of the staff. Hence, we view the seven SNRA regions as a selection of levels from a population of behavior levels. Correspondingly, we are not interested in the divisions in size classes or small area strata, but in potential differences in the quality of the maps.

Assume that β_r, $\gamma_{s(r)}$, $\delta_{t(rs)}$ and $\epsilon_{(rst)q}$ are independent with variances σ^2_β, σ^2_γ, σ^2_δ and σ^2_ϵ, respectively. We would like to test if these variances are zero. That is, we want to know whether variability exists in the length errors that is due to SNRA region, population center size class, or small area stratum. We do not have enough data to perform such tests 'by the book,' but use instead a simplified (approximative) test procedure. To put it briefly, we look only at one effect at the time and ignore the nesting. This was done for each of the three effects and for the observed errors under both the additive and the multiplicative error model. The relevant ANOVA tables are given in Appendix B. In no case is the hypothesis of zero variance rejected. We take this as an indication that the variances σ^2_β, σ^2_γ, and σ^2_δ are all zero.

We proceed by viewing the observed length errors $\{\zeta_{iq}\}$ simply as iid random variables with mean $\theta = \alpha$ and variance $\tau^2 = \sigma^2_\zeta$. As unbiased estimators of θ and τ^2, we use the sample mean $\bar{\zeta}$ and the sample variance

Error structure	Sample statistics $\bar{\zeta}$	Sample statistics s_ζ^2	95 % c.i. for θ	95 % c.i. for τ^2
Additive	-18.262	85481.915	$[-89.340, 52.816]$	$[0, 115296.998]$
Multiplicative	1.00209	0.00629	$[0.98281, 1.02137]$	$[0, 0.00848]$

Table 4.3: Sample statistics and confidence intervals (c.i.). The intervals for τ^2 are upper bounded.

s_ζ^2, respectively. The resulting estimates are given in Table 4.3. The confidence intervals shown in the table hold under the added assumption of normally distributed errors. We see that under the additive error model, the hypothesis of $\theta = 0$ cannot be rejected. If in fact the hypothesis is true, Corollary 3.2.1 applies and the length error does not bias $\hat{t}_{Fy}$ or $\hat{R}_F$. Under the multiplicative error model, the hypothesis of $\theta = 1$ can not be rejected. If this hypothesis is true, Corollary 3.2.2 tells us that the length error does not bias the estimators. We conclude that irrespective of which error structure we look at, our data do not suggest that the length error will cause bias in the estimators.

We are also interested in the possible variance increase due to the length error. Although the additive error model with equal error expectations and variances seems to fit the data somewhat better than the multiplicative counterpart (according to Figure 4.1), we choose the multiplicative model. The reason for this is simply that if the errors are multiplicative, Corollary 3.2.3 applies and we can easily estimate the approximate variance increase due to the use of $\hat{R}_F$ instead of $\hat{R}$. Since the observed errors are numerically quite small, the choice of model is not so crucial. If our point estimate of τ^2 in Table 4.3 coincides with the true parameter value, the relative variance increase is only about 0.6 percent. Hence, at least for the ratio, the variance increase seems negligible.

Chapter 5

On erroneous weighting of survey data

5.1 Introduction

The goal of a survey is to estimate some population characteristics or parameters of interest. Typically, a parameter can be expressed as function of several population totals. A simple example of a linear function is the population mean. In the SVS, the main parameter, average speed, is a non-linear function (a ratio) of two totals. No matter what the function looks like, the problem of estimating a population total obviously has a key role in survey sampling. If the sampling is conducted without replacement, the most important unbiased estimator is the π estimator. In with-replacement designs, unbiasedness is assured by use of the pwr estimator (the name from Särndal et al. [11, p. 51]; pwr refers to "p-expanded with replacement"). We have seen that in the SVS, these two estimation principles are combined, due to the fact that in the first sampling stage, sampling is done with replacement, while in the subsequent stages, sampling is done without replacement.

The π estimator and the pwr estimator presuppose the knowledge and use of the correct inclusion or drawing probabilities, respectively, for sampled elements. The probabilities are used to weight obtained data in order to reach the population level. As pointed out in Chapter 1, the frame problem addressed in this thesis may be looked upon as a case where incorrect

inclusion probabilities are used in the estimation. This view on our problem inspired us to investigate in some generality the consequences of basing calculations of survey estimates on incorrect weights. The result of our efforts is presented in this chapter. Expressions are given for the bias of the π estimator and the pwr estimator, as well as for the estimators of their variances, if incorrect weights are used. Furthermore, some examples of use of erroneous weights are examined. In the SVS, the correct inclusion probabilites in the final sampling stage are unknown (since the true lengths of the road networks in selected small areas are unknown). We revisit this problem and fit it into the general theoretical framework (while ignoring the previous sampling stages).

Another frame problem, that of unrecognized multiplicity in the frame, is also treated. If the multiplicity of the frame can not be determined, the correct inclusion probabilities (or drawing probabilities, depending on how the sampling is done) are again unknown. An example where the correct inclusion probabilites are available, but still not used, is also given.

5.2 Sampling without replacement

In without-replacement sampling designs, unbiased estimation of a population total is achieved by use of the π estimator. Expressed in words, this estimator is simply the sample sum of the observed values, weighted by their inverse inclusion probabilities. A thorough theoretical motivation for this estimator is given in Särndal et al. [11, Ch. 2]. In our investigation of the consequences of using incorrect weights, we will follow the same path as Särndal et al, with the only difference that we replace the correct first- and second-order inclusion probabilities, π_k and π_{kl}, with some (possibly) error-prone entities π_k^* and π_{kl}^*. Hence, we talk about π^* estimation instead of π estimation.

The π^* estimator of a total t is given by

$$\hat{t}_{\pi^*} = \sum\nolimits_s \frac{y_k}{\pi_k^*} \tag{5.1}$$

where each sample value y_k is weighted by π_k^*. Equivalently, $\hat{t}_{\pi^*}$ can be written as

$$\hat{t}_{\pi^*} = \sum\nolimits_U I_k \frac{y_k}{\pi_k^*} \tag{5.2}$$

where I_k is the sample membership indicator of element k,

$$I_k = \begin{cases} 1 & \text{if } k \in S \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

($k \in U$). S is the set-valued random variable that can take on every sample s possible under the design in question. For $k, l = 1, ..., N$, the expected values and covariances are

$$E(I_k) = \pi_k; \quad C(I_k, I_l) = \pi_{kl} - \pi_k \pi_l = \Delta_{kl} \tag{5.4}$$

where $\pi_k = \sum_{s \ni k} p(s)$, $\pi_{kl} = \sum_{s \ni k\&l} p(s)$, "$s \ni k$" denotes that the sum is over those samples s that contain the given k (and "$s \ni k\&l$" that the sum is over those samples s that contain both k and l), and $p(s)$ is the probability of selecting the sample s. The expected value of $\hat{t}_{\pi^*}$ is given by

$$E\left(\hat{t}_{\pi^*}\right) = \sum\nolimits_U E(I_k) \frac{y_k}{\pi_k^*} = \sum\nolimits_U \frac{\pi_k}{\pi_k^*} y_k. \tag{5.5}$$

Thus, the bias of $\hat{t}_{\pi^*}$ as estimator of t is given by

$$B\left(\hat{t}_{\pi^*}\right) = E\left(\hat{t}_{\pi^*}\right) - t = \sum\nolimits_U \left(\frac{\pi_k}{\pi_k^*} - 1\right) y_k. \tag{5.6}$$

The variance of $\hat{t}_{\pi^*}$ is given by

$$V\left(\hat{t}_{\pi^*}\right) = \sum \sum\nolimits_U \Delta_{kl} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}. \tag{5.7}$$

An estimator of $V\left(\hat{t}_{\pi^*}\right)$ is obtained by weighting Δ_{kl} with π_{kl}^*:

$$\hat{V}\left(\hat{t}_{\pi^*}\right) = \sum \sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}^*} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}. \tag{5.8}$$

The expected value of $\hat{V}\left(\hat{t}_{\pi^*}\right)$ is given by

$$\begin{aligned} E\left[\hat{V}\left(\hat{t}_{\pi^*}\right)\right] &= \sum \sum\nolimits_U E(I_k I_l) \frac{\Delta_{kl}}{\pi_{kl}^*} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*} \\ &= \sum \sum\nolimits_U \frac{\pi_{kl}}{\pi_{kl}^*} \Delta_{kl} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}. \end{aligned} \tag{5.9}$$

Obviously, $\hat{V}\left(\hat{t}_{\pi^*}\right)$ is unbiased for $V\left(\hat{t}_{\pi^*}\right)$ if $\pi_{kl}^* = \pi_{kl}$ for all $k, l \in U$.

The variance increase due to the use of $\hat{t}_{\pi^*}$ instead of $\hat{t}_\pi$ is given by

$$V\left(\hat{t}_{\pi^*}\right) - V\left(\hat{t}_\pi\right) = \sum\sum\nolimits_U \Delta_{kl}\left(\frac{y_k}{\pi_k^*}\frac{y_l}{\pi_l^*} - \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right). \tag{5.10}$$

For fixed size sampling designs, the variance increase can equivalently be expressed as

$$V\left(\hat{t}_{\pi^*}\right) - V\left(\hat{t}_\pi\right) = -\frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}\left[\left(\frac{y_k}{\pi_k^*} - \frac{y_l}{\pi_l^*}\right)^2 - \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2\right]. \tag{5.11}$$

Equation (5.11) is derived by use of a variance expression due to Yates and Grundy, see [11, pp. 45-46].

We will now take a look at some examples of use of erroneous weights in without-replacement sampling designs.

5.2.1 Example 1: Frame with coverage errors

We start by looking at a weighting problem arising from use of a frame with coverage errors. The conditions resemble those in the final stage of sampling for the SVS, with one important difference. In the SVS, a multi-stage sampling design is used, and the sampling is done with replacement in the first stage. The variance estimates are then calculated by use of the estimates of the totals for the first-stage sampling units. In this example, to simplify, we assume a one-stage without-replacement sampling design. Then, the variance is estimated directly from the sample data.

Consider the set of frame units F of size M. Some target population elements may be missing in F, and the frame may also contain elements that are not part of U, but the frame has no other deficiencies. An *SI* sample s of frame units of size n is drawn from F. The target population size N is unknown. We are interested in a situations where it is reasonable to make the following assumption.

Assumption 5.2.1 *The sample s is an SI sample from U.*

For example, a survey is to be conducted by means of an area sample. The sampling frame of area units is established on the basis of poor maps, resulting in a frame with coverage errors. Furthermore, the low quality

of the maps makes it hard for the field staff to accurately locate sampled areas. When they fail, they make a haphazard choice of areas to study. This informal selection procedure is such that every area in the target population has a chance of being included in the sample, while areas not in the target population do not.

Under Assumption 5.2.1, the true, but unknown, element inclusion probabilities and delta values are

$$\pi_{kl} = \begin{cases} \frac{n}{N} & \text{for } k = l \in U \\ \frac{n(n-1)}{N(N-1)} & \text{for } k \neq l \in U \end{cases} \tag{5.12}$$

$$\Delta_{kl} = \begin{cases} \frac{n}{N}\left(1 - \frac{n}{N}\right) & \text{for } k = l \in U \\ \frac{n}{N}\frac{n-N}{N(N-1)} & \text{for } k \neq l \in U \end{cases} \tag{5.13}$$

The π estimator $\hat{t}_\pi$ of t, and its variance $V\left(\hat{t}_\pi\right)$, are obtained by inserting these values in Equation (5.1) and (5.7), respectively. We get

$$\hat{t}_\pi = N\bar{y}_s \tag{5.14}$$

where $\bar{y}_s = \sum_s y_k/n$, and

$$V\left(\hat{t}_\pi\right) = N^2\left(1 - \frac{n}{N}\right)\frac{S_U^2}{n} \tag{5.15}$$

where $S_U^2 = \sum_U \left(y_k - \bar{y}_U\right)^2 / \left(N-1\right)$, $\bar{y}_U = \sum_U y_k/N$.

The inclusion probabilities available to the investigator are

$$\pi_{kl}^* = \begin{cases} \frac{n}{M} & \text{for } k = l \in U \\ \frac{n(n-1)}{M(M-1)} & \text{for } k \neq l \in U \end{cases}$$

Using these probabilites, the estimator of t is

$$\hat{t}_{\pi^*} = M\bar{y}_s = \frac{M}{N}\hat{t}_\pi. \tag{5.16}$$

Apparently, the bias of $\hat{t}_{\pi^*}$ as estimator of t is given by

$$B\left(\hat{t}_{\pi^*}\right) = E\left(\hat{t}_{\pi^*}\right) - t = \left(\frac{M}{N} - 1\right)t. \tag{5.17}$$

The variance of $\hat{t}_{\pi^*}$ is

$$V\left(\hat{t}_{\pi^*}\right) = \left(\frac{M}{N}\right)^2 V\left(\hat{t}_\pi\right) \tag{5.18}$$

and the estimator of $V\left(\hat{t}_{\pi^*}\right)$ is

$$\begin{aligned}\hat{V}\left(\hat{t}_{\pi^*}\right) &= M^2\left(1-\frac{n}{M}\right)\frac{S_s^2}{n} \\ &= \left(1-\frac{n}{M}\right)\left(1-\frac{n}{N}\right)^{-1}\left(\frac{M}{N}\right)^2\hat{V}\left(\hat{t}_{\pi}\right). \qquad (5.19)\end{aligned}$$

The bias of $\hat{V}\left(\hat{t}_{\pi^*}\right)$ as estimator of $V\left(\hat{t}_{\pi^*}\right)$ is given by

$$\begin{aligned}B\left[\hat{V}\left(\hat{t}_{\pi^*}\right)\right] &= E\left[\hat{V}\left(\hat{t}_{\pi^*}\right)\right] - V\left(\hat{t}_{\pi^*}\right) \\ &= \left[\left(1-\frac{n}{M}\right)\left(1-\frac{n}{N}\right)^{-1}-1\right]V\left(\hat{t}_{\pi^*}\right). \qquad (5.20)\end{aligned}$$

We see that the size and direction of the bias of $\hat{t}_{\pi^*}$ and $\hat{V}\left(\hat{t}_{\pi^*}\right)$ depends on the difference in size between the frame population and the target population. One possible way of handling the problem that M but not N is known, is to view M as a function of N and a random error. This is the strategy we use for tackling the SVS frame coverage problem.

5.2.2 Example 2: Stratified disproportionate sampling

This is an example of a case when the correct design weights are available but not used. A popular sampling method is the *STSI*: stratified sampling with *SI* sampling applied in all strata. If the stratum sample sizes are determined by proportional allocation, the first-order inclusion probabilities π_k coincide with those applicable in (unstratified) *SI* sampling. Hence, the *SI* formula for point estimation of a population total can be used. This is not, however, true for disproportionate allocations in general. Hansen et al. [5, Ch. 2 Sec. 8] noted that, if disproportionate allocation is used but the investigator uses the *SI* estimator of t (and hence fails to use the proper strata inclusion probabilities), the employed estimator may be biased.

Let us look at the statistical properties of the *SI* estimator when applied on an *STSI* sample. We start by introducing some notation. Let the population of interest, U, be partitioned into H strata $U_1, .., U_h, ..., U_H$ of sizes $N_1, .., N_h, ..., N_H$, respectively. From each stratum h, an *SI* sample s_h of size n_h is selected. The resulting total sample is $s = \bigcup_{h=1}^{H} s_h$ of size $n = \sum_{h=1}^{H} n_h$. Since the strata form a partition of U, and the sampling is

independent in each stratum, the π estimator of t for a stratified design is

$$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi} \tag{5.21}$$

where $\hat{t}_{h\pi}$ is the π estimator of $t_h = \sum_{U_h} y_k$. Correspondingly, the variance of $\hat{t}_\pi$ is given by

$$V\left(\hat{t}_\pi\right) = \sum_{h=1}^{H} V\left(\hat{t}_{h\pi}\right).$$

The appearances of $\hat{t}_{h\pi}$ and $V\left(\hat{t}_{h\pi}\right)$ are given by Equation (5.14) and (5.15), respectively.

Erroneous use of the *SI* estimator of t means that the weighting is done by use of the inclusion probabilities

$$\pi^*_{kl} = \begin{cases} \frac{n}{N} & \text{for } k = l \in U \\ \frac{n(n-1)}{N(N-1)} & \text{for } k \neq l \in U \end{cases} \tag{5.22}$$

The resulting estimator of t is

$$\hat{t}_{\pi^*} = \frac{N}{n} \sum\nolimits_s y_k = \sum_{h=1}^{H} \frac{N}{n} \frac{n_h}{N_h} \hat{t}_{h\pi} = \sum_{h=1}^{H} \hat{t}_{h\pi^*}. \tag{5.23}$$

Obviously, the bias of $\hat{t}_{\pi^*}$ as estimator of t is given by

$$B\left(\hat{t}_{\pi^*}\right) = E\left(\hat{t}_{\pi^*}\right) - t = \sum_{h=1}^{H} \left(\frac{N}{n} \frac{n_h}{N_h} - 1\right) t_h \tag{5.24}$$

The variance of $\hat{t}_{\pi^*}$ is

$$V\left(\hat{t}_{\pi^*}\right) = \sum_{h=1}^{H} \left(\frac{N}{n} \frac{n_h}{N_h}\right)^2 V\left(\hat{t}_{h\pi}\right) \tag{5.25}$$

5.3 Sampling with replacement

In with-replacement sampling designs, a population total is estimated without bias by the pwr estimator. At first glance, this estimator resembles the π estimator, since the observed values are weighted (or "p-expanded")

by their inverse probabilities. The probabilities are, however, the drawing probabilities, not the inclusion probabilities. Only in a sample of size one do these coincide. Also, the pwr estimator is a sample average, not a sample sum, of the weighted observations.

The construction of the pwr estimator is explained in Särndal et al. [11, Ch. 2]. We will again follow the same line of argumentation as in the cited reference, but replace the correct drawing probabilities p_k with the (possibly) error-prone p_k^*. Hence, we will talk about the pwr* estimator instead of the pwr estimator.

The pwr* estimator of t is given by

$$\hat{t}_{\text{pwr}^*} = \frac{1}{m} \sum_{\nu=1}^{m} \frac{y_{k_\nu}}{p_{k_\nu}^*} \tag{5.26}$$

where m is the fixed number of draws, k_ν is the element selected in the νth draw ($\nu = 1, ..., m$), y_{k_ν} is the study variable value for k_ν and $1/p_{k_\nu}^*$ is the weight attached to k_ν. Equivalently, $\hat{t}_{\text{pwr}^*}$ can be written as

$$\hat{t}_{\text{pwr}^*} = \frac{1}{m} \sum_{\nu=1}^{m} Z_\nu^* = \bar{Z}_{os}^* \tag{5.27}$$

where Z_ν^* is the random variable such that $Z_\nu^* = y_k/p_k^*$ if $k_\nu = k$. For $\nu = 1, ..., m$, the expectation and variance of Z_ν^* are, respectively,

$$E\left(Z_\nu^*\right) = \sum\nolimits_U \frac{p_k}{p_k^*} y_k = \mu_{Z^*} \tag{5.28}$$

$$V\left(Z_\nu^*\right) = \sum\nolimits_U \left(\frac{y_k}{p_k^*} - \sum\nolimits_U \frac{p_k}{p_k^*} y_k \right)^2 p_k = \sigma_{Z^*}^2 \tag{5.29}$$

where $p_k = \Pr\left(Z_\nu^* = y_k/p_k^*\right)$. Since the Z_ν^*s are iid random variables, the expected value of $\hat{t}_{\text{pwr}^*}$ is given by

$$E\left(\hat{t}_{\text{pwr}^*}\right) = \mu_{Z^*} \tag{5.30}$$

Obviously, the bias of $\hat{t}_{\text{pwr}^*}$ as estimator of t is given by

$$B\left(\hat{t}_{\text{pwr}^*}\right) = E\left(\hat{t}_{\text{pwr}^*}\right) - t = \sum\nolimits_U \left(\frac{p_k}{p_k^*} - 1 \right) y_k \tag{5.31}$$

The variance of $\hat{t}_{\text{pwr}^*}$ is given by

$$V\left(\hat{t}_{\text{pwr}^*}\right) = \frac{\sigma^2_{Z^*}}{m} \tag{5.32}$$

and an unbiased estimator of $V\left(\hat{t}_{\text{pwr}^*}\right)$ by

$$\hat{V}\left(\hat{t}_{\text{pwr}^*}\right) = \frac{S^2_{Z^*os}}{m} \tag{5.33}$$

where $S^2_{Z^*os} = \sum_{\nu=1}^{m}\left(Z^*_\nu - \bar{Z}^*_{os}\right)^2/(m-1)$.

If the weights $p^*_k = p_k$ are used, $\mu_{Z^*} = t$ and $\sigma^2_{Z^*} = \sum_U \left[(y_k/p_k) - t\right]^2 p_k$. Then, $\hat{t}_{\text{pwr}^*}$ equals the pwr estimator $\hat{t}_{\text{pwr}}$ of t.

5.3.1 Example 3: Frame with unrecognized multiplicity

In Example 1, we looked at a weighting problem arising from use of a frame with coverage errors. Another frame error which may cause weighting problems is when the frame contains duplicate listings; that is, multiple elements in the frame population are attached to one element in the target population. As pointed out by Lessler and Kalsbeek [8, p. 73], it is not always feasible to determine multiplicity:

> For example, one may wish to survey users of a certain product; however, it may only be feasible to sample people at the time of purchase. This is a sample of uses rather than users. People will vary in terms of the number of times they use the product, and more frequent users will have greater chance of being in the sample.

As in Example 1, consider the set of frame units F of size M. The only frame imperfection now is that the number of frame units having a link with element k, $L_{\cdot k}$, exceeds one for some $k \in U$. Assume that a selection of frame units from F is conducted by simple random sampling with replacement (*SIR*). Frame units linked to the same element as a unit already obtained are not eliminated if they are drawn again. The sample selection will result in an ordered sample *os* of frame units. The notation *os* can equivalently be used for the resulting with-replacement sample of elements from U. If m independent draws are made from F in order to

obtain *os*, the probability of selecting element k_ν, p_{k_ν}, equals $L_{\cdot k_\nu}/M$ for $\nu = 1, ..., m$. Hence, *os* is a *pps* sample of elements, drawn with probability proportional to the number of duplicates in the frame.

We derive $\hat{t}_{\text{pwr}}$ and $V\left(\hat{t}_{\text{pwr}}\right)$ under current design by inserting $p_{k_\nu} = L_{\cdot k_\nu}/M$ in Equation (5.26) and (5.32), respectively. We get

$$\hat{t}_{\text{pwr}} = \frac{M}{m}\sum_{\nu=1}^{m}\frac{y_{k_\nu}}{L_{\cdot k_\nu}} \tag{5.34}$$

and

$$V\left(\hat{t}_{\text{pwr}}\right) = \frac{1}{m}\sum\nolimits_U\left(\frac{M}{L_{\cdot k}}y_k - t\right)^2\frac{L_{\cdot k}}{M} \tag{5.35}$$

If the investigator treats *os* as an *SIR* sample of elements from U, she will use the weights $1/p^*_{k_\nu} = M$ for $k \in U$, $\nu = 1, ..., m$. The resulting error-prone estimator of t is given by

$$\hat{t}_{\text{pwr}^*} = M\bar{y}_{os} \tag{5.36}$$

where $\bar{y}_{os} = \sum_{\nu=1}^{m} y_{k_\nu}/m$. From Equation (5.31), the bias of $\hat{t}_{\text{pwr}^*}$ as estimator of t is given by

$$B\left(\hat{t}_{\text{pwr}^*}\right) = E\left(\hat{t}_{\text{pwr}^*}\right) - t = \sum\nolimits_U\left(L_{\cdot k} - 1\right)y_k. \tag{5.37}$$

As estimator of the variance of $\hat{t}_{\text{pwr}^*}$, the investigator will use

$$\hat{V}\left(\hat{t}_{\text{pwr}^*}\right) = M^2\frac{S^2_{yos}}{m} \tag{5.38}$$

where $S^2_{yos} = \sum_{\nu=1}^{m}\left(y_{k_\nu} - \bar{y}_{os}\right)^2/\left(m-1\right)$. The employed variance estimator is unbiased for the variance of $\hat{t}_{\text{pwr}^*}$,

$$V\left(\hat{t}_{\text{pwr}^*}\right) = \frac{1}{m}\sum\nolimits_U\left(My_k - \sum\nolimits_U L_{\cdot k}y_k\right)^2\frac{L_{\cdot k}}{M}. \tag{5.39}$$

5.4 Summary

The use of a sampling frame suffering from either coverage errors or multiplicity may result in incorrect weighting of the observed values. Faulty weighting may also arise by, for example, not taking stratification properly into account. The use of incorrect weights will bias both the π estimator and the pwr estimator of a total. The estimator of the variance of the π estimator will also be biased, but not the variance estimator for the pwr estimator.

Chapter 6

Summary and final remarks

There is a strong belief at the SNRA that if one succeeds in reducing the average speed on the roads, this will substantially reduce the number of traffic deaths and injuries. Therefore, the annual SVS estimates of average speed receive a lot of attention, and of course they need to be reliable. The uncertainty due to the fact that only a subset of the population is surveyed is quantified by confidence intervals, but these do not give the full picture if there are nonsampling errors present. In this thesis, we investigated one possible source of added uncertainty in the survey results: the road lengths in the frames used in the final stage of sample selection. If these lengths are in error, how are the statistical properties of the estimators affected? Average speed is a complex parameter since it is the ratio of two population totals: total vehicle mileage and total travel time. At present, survey estimates of the totals are not published. Still we found it illustrative to let the estimator of a total be comprised by our study, since the problems of estimating a ratio and a total are closely related.

Our theoretical derivations, supported by an error model, resulted in expressions for the effects of the length error on the bias and variance of the estimators. In particular, we showed that if the errors are multiplicative with expectation of one and constant variance, the length error has no bias effect on the estimator of average speed, and the relative (approximate) variance increase for this estimator simply equals the error variance. We also collected some data on the real errors in the frames. The observed errors were found to be quite small, and for simplicity we choose the multiplicative model, although the additive model actually had a slightly better fit.

The multiplicative errors were found to have an expectation close to one, and their variance was estimated to less than one percent. Putting all this together, our investigation led us to make the following conclusions (which are good news to the survey management). First, neither the estimator of average speed nor the estimator of a total seems to be biased by the length error. Second, the variance increase due to the length error, for the average speed estimator, seems to be negligible.

It should be noted that our results are useful only if one trusts our model, since the entire investigation relies heavily upon it. The model includes a very strong assumption: that the actual final-stage samples are selected by simple random sampling from the true road networks. For the future, we recommend that the data-collection instructions be given an overhaul. Improved instructions would increase the chances that the model assumption really holds. Also remember that the only frame imperfection considered in this study was the length error. The empirical study exposed several other imperfections, associated with the last-stage frames, that need to be dealt with.

At first glance, the problem of erroneous road lengths seems unique to a traffic survey like the SVS. However, apart from the implications for data collection, the problem may be more generally formulated as that of using incorrect inclusion probabilities in survey sampling. Since this problem to our knowledge has not been treated in the literature, we paid it some attention in Chapter 5. The most important unbiased estimators of a total are the pwr estimator (for with-replacement sampling designs) and the π estimator (for sampling without replacement). The observations are then weighted by the inverse of their drawing or inclusion probabilities. We saw that use of other than the correct weights implies bias in both the estimators. The estimator of the variance of the π estimator will also be biased, however not the variance estimator for the pwr estimator. The latter result is important for the SVS, where the sampling in the first stage is done with replacement. It follows that despite the length error, the variances of the survey estimators will be correctly estimated.

We conclude with some suggestions for further work. Our study does not give a clear picture of the impact of the road length errors on the variances of the estimators of the totals. Thus, if publication of survey estimates of the totals is suggested, some additional work is needed. Otherwise, we suggest

that future research is directed towards other nonsampling errors of possible importance.

The error source that perhaps first comes to mind is the measurement instrument. Surely there must be errors associated with the equipment – the tubes and the traffic analyzer – used for observing the road traffic? In fact, several problems are known to occur. For example, if the tubes are not parallel, or if the distance between them differs from the intended, this will result in incorrect time registrations. If the vehicles are either accelerating or decelerating during passage, this will also result in erroneous time registrations. These types of errors have already been subjected to extensive analyses (see, e.g., [1]). However, a possible measurement error that has not yet been investigated is the effect of the person installing the equipment on the road – the analogue to the 'interviewer effect' known from interviewer surveys.

The most serious nonsampling error that remains to be investigated is probably the one of missing observations. It is common to distinguish between two types of missing data in a survey: unit nonresponse and item nonresponse. In the SVS, unit nonresponse corresponds to a complete loss of data from a measurement location, whereas item nonresponse corresponds to the loss of data for some, but not all, of the passing vehicles. Neither type of nonresponse is likely to occur at random. An important cause of nonresponse is the capacity limit of the traffic analyzer, which results in a greater loss of data if the traffic is heavy.

Another cause of nonresponse is that some locations are difficult or even impossible to observe. For example, cars parked by the roadside may obstruct the installation of the tubes. Typically, such problematic places are never observed, but replaced by neighboring locations. In other cases, measurements from sampled locations are dismissed due to low-quality data; such locations are re-measured later. These procedures represent two different kinds of 'field substitution.' Earlier work by the author [6] suggests that the latter type of substitution probably has the least impact on the survey estimates, since the variation in speed measurements due to twentyfour-hour periods was found to be smaller than the location-to-location variation.

Bibliography

[1] Allogg AB, *Optimization of parameter adjustings and sensor distances for measurements with Metor 2000.* Can be ordered from Allogg AB, Box 43, SE-647 21 Mariefred, Sweden, 1996. (In Swedish).

[2] P. Biemer and S. L. Stokes, *Approaches to the modeling of measurement error*, in Measurement Errors in Surveys, P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, eds., Wiley, New York, 1991, pp. 487–516.

[3] A. Bolling and M. Wiklund, *Validation of Metor in urban areas*, VTI Meddelande 814, Statens väg- och transportforskningsinstitut, SE-581 95 Linköping, Sweden, 1997. (In Swedish).

[4] C.-M. Cassel, C.-E. Särndal, and J. H. Wretman, *Foundations of Inference in Survey Sampling*, Wiley, New York, 1977.

[5] M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*, vol. I: Methods and Applications, Wiley, New York, 1953.

[6] A. Isaksson, *Speed measurement variations in time and space*, Research report LiU-MAT-R-1999-01, Linköping University, SE-581 83 Linköping, Sweden, 1999.

[7] L. Kish, *Survey Sampling*, Wiley, New York, 1965.

[8] J. T. Lessler and W. D. Kalsbeek, *Nonsampling Error in Surveys*, Wiley, New York, 1992.

[9] D. C. Montgomery, *Design and Analysis of Experiments*, Wiley, New York, fourth ed., 1997.

BIBLIOGRAPHY

[10] D. Raj, *Sampling Theory*, McGraw-Hill, New York, 1968.

[11] C.-E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer, New York, 1992.

[12] Swedish National Road Administration, National Police Board, and Swedish Association of Local Authorities, *National road traffic safety programme for 1995-2000*. Can be ordered from the Swedish National Road Administration, SE-781 87 Borlänge, Sweden, 1994. (In Swedish).

Appendix A

Proofs

A.1 Proof of Theorem 3.2.1

Let $\mathbb{N}_F$ denote the (random) vector of all frame road lengths N_{Fiq}. By the use of conditioning, the pm-expected value and pm-variance of an estimator $\hat{\psi}$ can be written as

$$\begin{aligned} E_{pm}\left(\hat{\psi}\right) &= E_m\left[E_p\left(\hat{\psi}\,|\mathbb{N}_F\right)\right]; \\ V_{pm}\left(\hat{\psi}\right) &= E_m\left[V_p\left(\hat{\psi}\,|\mathbb{N}_F\right)\right] + V_m\left[E_p\left(\hat{\psi}\,|\mathbb{N}_F\right)\right]. \end{aligned}$$

It suffices to show that under the SVS sampling design $p(s)$ and the frame errors model m, $E_{pm}\left(\hat{t}_{Fyi}\right)$ equals the stated expression for $E_{pm}\left(\hat{t}_{F\pi yi}\right)$ and $V_{pm}\left(\hat{t}_{Fyi}\right)$ equals the stated expression for $V_{pm}\left(\hat{t}_{F\pi yi}\right)$. Let subscript $II\,|\mathbb{N}_F$ indicate conditional expected value or conditional variance with respect to the design used in stage two, given os_I and $\mathbb{N}_F$, and subscript $III\,|\mathbb{N}_F$ indicate conditional expected value or conditional variance with respect to the design used in stage three, given os_I, s_{IIi_v} and $\mathbb{N}_F$. Then,

$$\begin{aligned} E_{pm}\left(\hat{t}_{Fyi}\right) &= E_m\left[E_{II|\mathbb{N}_F}E_{III|\mathbb{N}_F}\left(\hat{t}_{Fyi}\right)\right] = E; \\ V_{pm}\left(\hat{t}_{Fyi}\right) &= V_m\left[E_{II|\mathbb{N}_F}E_{III|\mathbb{N}_F}\left(\hat{t}_{Fyi}\right)\right] + E_m\left[V_{II|\mathbb{N}_F}E_{III|\mathbb{N}_F}\left(\hat{t}_{Fyi}\right)\right] \\ &\quad + E_m\left[E_{II|\mathbb{N}_F}V_{III|\mathbb{N}_F}\left(\hat{t}_{Fyi}\right)\right] = V_1 + V_2 + V_3. \end{aligned}$$

We start with the expectation

$$E = E_m\left(\sum\nolimits_{U_{IIi}} \frac{N_{Fiq}}{N_{iq}} t_{yiq}\right) = \sum\nolimits_{U_{IIi}} E_m\left(\frac{N_{Fiq}}{N_{iq}} t_{yiq}\right) = \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{yiq}$$

which equals the stated expression for $E_{pm}\left(\hat{t}_{F\pi yi}\right)$. Now we turn to the variance. First,

$$V_1 = V_m\left(\sum_{U_{IIi}} \frac{N_{Fiq}}{N_{iq}} t_{yiq}\right) = \sum_{U_{IIi}} V_m\left(\frac{N_{Fiq}}{N_{iq}} t_{yiq}\right) = \sum_{U_{IIi}} \left(\frac{\sigma_{iq}}{N_{iq}}\right)^2 t_{yiq}^2.$$

Second,

$$\begin{aligned} V_2 &= E_m\left[N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \sum_{U_{IIi}} \left(\frac{N_{Fiq}}{N_{iq}} t_{yiq} - \frac{1}{N_{IIi}} \sum_{U_{IIi}} \frac{N_{Fiq}}{N_{iq}} t_{yiq}\right)^2\right] \\ &= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \sum_{U_{IIi}} E_m\left[\left(\frac{N_{Fiq}}{N_{iq}} t_{yiq} - \frac{1}{N_{IIi}} \sum_{U_{IIi}} \frac{N_{Fiq}}{N_{iq}} t_{yiq}\right)^2\right] \end{aligned}$$

$$\begin{aligned} &= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \left\{\frac{1}{N_{IIi}} \sum_{U_{IIi}} V_m\left(\frac{N_{Fiq}}{N_{iq}} t_{yiq}\right)\right. \\ &\quad \left. + \frac{1}{N_{IIi}-1} \sum_{U_{IIi}} \left[E_m\left(\frac{N_{Fiq}}{N_{iq}} t_{yiq}\right) - \frac{1}{N_{IIi}} \sum_{U_{IIi}} E_m\left(\frac{N_{Fiq}}{N_{iq}} t_{yiq}\right)\right]^2\right\} \\ &= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \left\{\frac{1}{N_{IIi}} \sum_{U_{IIi}} \left(\frac{\sigma_{iq}}{N_{iq}}\right)^2 t_{yiq}^2\right. \\ &\quad \left. + \frac{1}{N_{IIi}-1} \sum_{U_{IIi}} \left[\frac{\mu_{iq}}{N_{iq}} t_{yiq} - \frac{1}{N_{IIi}} \sum_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{yiq}\right]^2\right\} \end{aligned}$$

and finally,

$$\begin{aligned} V_3 &= E_m\left[\frac{N_{IIi}}{n_{IIi}} \sum_{U_{IIi}} \left(\frac{N_{Fiq}}{N_{iq}}\right)^2 V_p\left(\hat{t}_{\pi yiq}\right)\right] \\ &= \frac{N_{IIi}}{n_{IIi}} \sum_{U_{IIi}} \frac{V_p\left(\hat{t}_{\pi yiq}\right)}{N_{iq}^2} E_m\left(N_{Fiq}^2\right) \\ &= \frac{N_{IIi}}{n_{IIi}} \sum_{U_{IIi}} \frac{V_p\left(\hat{t}_{\pi yiq}\right)}{N_{iq}^2} \left(\sigma_{iq}^2 + \mu_{iq}^2\right). \end{aligned}$$

Adding V_1, V_2 and V_3 gives, after simplification, the stated expression for $V_{pm}\left(\hat{t}_{F\pi yi}\right)$.

A.2 Proof of Corollary 3.2.3

The variance increase due to the use of $\hat{t}_{Fy}$ instead of $\hat{t}_y$ as estimator of t_y is immediately obtained from Corollary 3.2.1 by replacing σ_{iq}^2 with $N_{iq}^2\tau^2$ in Equation (3.8). In the same manner, we obtain

$$AV_{pm}\left(\hat{R}_F\right) - AV_p\left(\hat{R}\right) = \tau^2\lambda$$

where

$$\lambda = \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\left[V_p\left(\hat{t}_{\pi Eiq}\right) + t_{Eiq}^2\right]$$

It remains to show that $\lambda = AV_p\left(\hat{R}\right)$. From Lemma 2.6.1, the approximate p-variance for $\hat{R}$ is given by

$$AV_p\left(\hat{R}\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{t_{Ei}^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\left[V_{EIIi} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}V_p\left(\hat{t}_{\pi Eiq}\right)\right]\right\}$$

where

$$\begin{aligned}
V_{EIIi} &= N_{IIi}^2\frac{1-f_{IIi}}{n_{IIi}}\frac{1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}}\left(t_{Eiq} - \frac{t_{Ei}}{N_{IIi}}\right)^2 \\
&= \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}t_{Eiq}^2 - \frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}}t_{Eiq}^2 - \frac{1}{n_{IIi}}\frac{N_{IIi}-n_{IIi}}{N_{IIi}-1}t_{Ei}^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
AV_p\left(\hat{R}\right) - \lambda &= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\left[\frac{t_{Ei}^2}{p_i} - \frac{1}{p_i}\left(\frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}}t_{Eiq}^2\right.\right. \\
&\qquad \left.\left. + \frac{N_{IIi}-n_{IIi}}{n_{IIi}\left(N_{IIi}-1\right)}t_{Ei}^2\right)\right] \\
&= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\left[\frac{t_{Ei}^2}{p_i}\left(1 - \frac{N_{IIi}-n_{IIi}}{n_{IIi}\left(N_{IIi}-1\right)}\right)\right. \\
&\qquad \left. - \frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}}t_{Eiq}^2\right] \\
&= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\left(t_{Ei}^2 - \sum\nolimits_{U_{IIi}}t_{Eiq}^2\right).
\end{aligned}$$

A.2. Proof of Corollary 3.2.3

Since, in practice, n_{IIi} equals one for all i, the derived expression is zero and we are ready.

Appendix B

ANOVA tables

B.1 Under additive error model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
SNRA Region	416735.9	6	69456.0	0.80	0.5761
Error	5054106.6	58	87139.8		
Total	5470842.6	64			

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Size class	89074.4	2	44537.2	0.51	0.6012
Error	5381768.1	62	86802.7		
Total	5470842.6	64			

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Small Area Stratum	101240.4	3	33746.8	0.38	0.7653
Error	5369602.2	61	88026.3		
Total	5470842.6	64			

B.2 Under multiplicative error model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
SNRA Region	0.0341	6	0.0057	0.90	0.5036
Error	0.3683	58	0.0063		
Total	0.4024	64			

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Size class	0.0238	2	0.0119	1.95	0.1507
Error	0.3786	62	0.0061		
Total	0.4024	64			

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Small Area Stratum	0.0169	3	0.0056	0.89	0.4508
Error	0.3855	61	0.0063		
Total	0.4024	64			