DACSEIS

IST-2000-26057

Workpackage 6

Variance Estimation for Unequal Probability Designs

Deliverable 6.1

2004

List of contributors:

Yves Berger and Chris Skinner; University of Southampton

Main responsibility:

Yves G. Berger; University of Southampton

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

Survey sampling textbooks often refer to the Sen-Yates-Grundy variance estimator for use with without-replacement unequal probability designs. This estimator is rarely implemented, because of the complexity of determining joint inclusion probabilities. In practice, the variance is usually estimated by simpler variance estimators such as the Hansen-Hurwitz variance estimator; which often leads to overestimation of the variance for large sampling fraction that are common in business surveys. We will consider alternative variance estimators that depend on the first-order inclusion probabilities only and are usually more accurate than the Hansen-Hurwitz estimator for large sampling fraction. In this report, we compare these estimators via simulation based on the German Income and Consumption Universe.

Yves Berger Southampton

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Unequal probability sampling was first suggested by Hansen and Hurwitz (1943) in the context of with-replacement sampling. Narain (1951), Horvitz and Thompson (1952) developed the corresponding theory for sampling without-replacement. Gabler (1984) shows the superiority of sampling without-replacement over sampling with-replacement. Variance estimation for sampling with-replacement is straightforward (Hansen and Hurwitz, 1943). However, for sampling without-replacement, the design unbiased Sen-Yates-Grundy variance estimator (Sen, 1953; Yates and Grundy, 1953) is hard to compute because of joint inclusion probabilities. Although exact computation of these probabilities is possible with specific sampling designs like the Chao (1982) sampling design, their calculation becomes practically impossible when the sample size is large. It is also inconceivable to provide these probabilities in released data-sets, as the set of joint inclusion probabilities is a series of $n(n-1)/2$ values; where n denotes the sample size. Moreover, standard statistical packages like SPSS, SAS, STATA do not deal with these probabilities. Specialized software like SUNDAAN needs to be used. However, even SUNDAAN does not include actual computation of these probabilities. They need to be provided by the user.

We will show that it is possible to estimate the sampling variance without computing joint inclusion probabilities. We will also show how it is possible to estimate the variance using weighted least squares (WLS) regression.

Chapter 2

Complexity of Variance Estimation

Consider a finite population $U = \{1, \ldots, i, \ldots, N\}$ containing N units. Suppose we wish to estimate a population total

$$\tau = \sum_{i \in U} y_i$$

where y_i is the value of a study variable of a unit labelled i. The π-*estimator* (Narain, 1951; Horvitz and Thompson, 1952) of Y is

$$\widehat{\tau} = \sum_{i \in s} \check{y}_i \tag{2.1}$$

where s is a sample, $\check{y}_i = y_i \pi_i^{-1}$ and where π_i is the first-order inclusion probability of unit i; that is, the probability for unit i to be sampled. The variance of the π-estimator plays an important role in variance estimation, as most estimators of interest can be linearized to involve π-estimators. The sampling variance of $\widehat{\tau}$ is given by

$$\mathrm{var}(\widehat{\tau}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \check{y}_i \check{y}_j \tag{2.2}$$

where π_{ij} is the joint inclusion probabilities of unit i and$j$; that is, the probability that both units i and$j$ are sampled, and $\pi_{ii} = \pi_i$.

Assuming n fixed, design unbiased estimator of σ_Y^2 is given by the *Sen-Yates-Grundy estimator*

$$\widehat{\mathrm{var}}(\widehat{\tau})_{YG} = \tfrac{1}{2} \sum_{i \in s} \sum_{j \in s} (\pi_{ij} - \pi_i \pi_j) \pi_{ij}^{-1} (\check{y}_i - \check{y}_j)^2 \tag{2.3}$$

The estimator (2.3) is hard to implement in practice, as the π_{ij} are often unknown except for special cases such as stratified simple random sampling (STSRS). Moreover, the double sum feature is computationally inconvenient for large samples. Furthermore, the computation of the π_{ij} requires the values of the π_i for all the units of the population, whereas it is common to know the value of π_i only for the sampled units. There are alternative methods (Smith, 2001) of variance estimation that do not involve π_{ij}, such as replication methods. In the next chapter, we show that the variance can be easily estimated without using computationally intensive methods like replication methods or any methods that would involve the actual computation of the π_{ij}.

Chapter 3

Variance estimators free of joint inclusion probabilities

We suppose that the sampling design is a single stage stratified sampling design with unequal probabilities within each stratum. Let $U_1, \ldots, U_H$ denotes the strata. We suppose that a sample s_h of size n_h is selected without replacement within each stratum U_h of size N_h. For simplicity, we assume that the sample data are free from errors due to non-response and measurement.

In the following section, we define several estimators for the variance. These estimators can be computed using the Splus library varianceht available at

`http://www.socstats.soton.ac.uk/staff/berger/unequal.html`

This library is also available for R.

3.1 The variance estimator based on simple random sampling without replacement

The variance based on the simple random sampling without replacement (SRS) sampling is

$$\widehat{\text{var}}(\widehat{\tau})_{SRS} = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\widehat{S}_h^2}{n_h}$$

where $\widehat{S}_h^2 = (n_h - 1)^{-1} \sum_{i \in s_h} (y_i - \overline{y}_h)$ and $\overline{y}_h = n_h^{-1} \sum_{i \in s_h} y_i$. This estimator ignores the unequal probabilities and is biased under unequal probability sampling.

3.2 The Hansen-Hurwitz estimator

The Hansen-Hurwitz variance estimator is defined by

$$\widehat{\mathrm{var}}(\widehat{\tau})_{HH} = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\check{y}_i - \widehat{B}_h^*)^2 \tag{3.1}$$

where $\widehat{B}_h^* = n_h^{-1} \sum_{i \in s_h} \check{y}_i$. The variance estimator $\widehat{\mathrm{var}}(\widehat{\tau})_{HH}$ is the usual stratum-by-stratum Hansen-Hurwitz variance estimator.

3.3 The Hájek estimator

BERGER (2004a,b) shows that the Hájek estimator is one of the most suitable estimator for unequal probability sampling. BERGER (2004b) also shows that this estimator is as accurate as (2.3) for high entropy sampling design. The Hájek variance estimator is given by

$$\widehat{\mathrm{var}}(\widehat{\tau})_H = \sum_{i \in s} \check{c}_i \, \widehat{e}_i^2 \tag{3.2}$$

where $\check{c}_i = n_h(n_h - 1)^{-1}(1 - \pi_i) \quad (i \in U_h)$. The $\widehat{e}_i$ are the WLS residuals of the regression

$$\check{y}_i = \sum_{h=1}^{H} \widehat{B}_h \, z_{ih} + \widehat{e}_i \tag{3.3}$$

where

$$\widehat{B}_h = \left(\sum_{j \in s} \check{c}_j \, z_{jh}^2 \right)^{-1} \sum_{i \in s} \check{c}_i \, \check{y}_i \, z_{ih}$$

BERGER (1998) showed that (3.2) can be used for a class of highly randomized or high entropy sampling designs; which includes the successive (HÁJEK, 1964) and the Rao-Sampford (RAO, 1965 and SAMPFORD, 1967) sampling designs. The systematic sampling design is not a high entropy sampling design. In Section 3.5, we show briefly how the Hájek variance estimator can be extended to accommodate this sampling design.

In practice, (3.2) is simple to compute, as it does not require the π_{ij}. If we know in which stratum each unit belongs, it is easy to specify the H stratification variables z_{ih}. As $\widehat{B}_h$ is the usual WLS estimate of a regression coefficient, any standard statistical packages can be used to compute the $\widehat{B}_h$ and the set of residuals e_i. The variance (3.2) is just a weighted sum of these residuals. The merit of this method is the fact that the variance estimator is only computed through a set of residuals and only requires the values of π_i for the sampled units.

The set of $(1 - \pi_i)$ in (3.2) can be viewed as generalised finite population correction (FPC). Indeed, with STSRS $1 - \pi_i$ reduces to the usual FPC $1 - \pi_i = 1 - f_h$. A correction of degree of freedom (DF) $n_h(n_h - 1)^{-1}$ is also included in (3.2). There are other effects of

the sampling design also included in (3.2). The effect of stratification is specified by the residuals e_i, as the working regression (3.3) uses the stratification variables as independent variables. This is the major effect of the sampling design. The effect of the π_i are also included in the residuals as the dependent variables $\check{y}_i$ in (6 is the value of the study variable divided by π_i. There are remainder effects not included in (3.2); which explains the slight differences in the variance (2.2) due to the method of sampling used, given a set of π_i. The alternative expression (3.2) has the advantage of revealing the main effects of the sampling design due to: stratification, the unequal probabilities, the FPC and the correction for DF. This allows us to quantify the impact of these effects on the variance.

If we substitute c_i by $n_h(n_h-1)^{-1} \quad (i \in U_h)$ into (3.2), we get (3.1). If $\check{c}_i = (1-\pi_i)\log(1-\pi_i)\pi_i^{-1}$, (3.2) is algebraically equivalent to the Rosén (1991) estimator implemented by Statistics Sweden. If $\check{c}_i = (1-\pi_i)^{-1} \quad \{1-[\sum_{j\in s_h}\pi_j(1-\pi_j)]^2 \sum_{j\in s_h}(1-\pi_j)\} \quad (i \in U_h)$, (3.2) gives the Deville (1999) variance estimator.

3.4 The Brewer estimator

The Brewer's family of simple estimators (Brewer, 2002, Chap. 9) also merited consideration. This family uses the approximate formula for the π_{ij} derived by Hartley and Rao (1962). An estimator of this family is given by

$$\widehat{\text{var}}(\widehat{\tau})_{Brewer} = \sum_{i\in s} \check{c}_i^* \, \widehat{e}_i^{*2} \tag{3.4}$$

where $\widehat{e}_i^* = \check{y}_i - \sum_{h=1}^H \widehat{B}_h^* z_{ih}$ is the ordinary least squares (OLS) residual and $\widehat{B}_h^*$ is the OLS coefficient defined by

$$\widehat{B}_h^* = \left(\sum_{j\in s} z_{jh}^2\right)^{-1} \sum_{i\in s} \check{y}_i \; z_{ih}$$

Note that $\widehat{B}_h^* = n_h^{-1}\sum_{i\in s_h}\check{y}_i$. Brewer (2002), Chap. 9 proposed several choice for $\check{c}_i (i \in U_h)$:

(i) $\check{c}_i^* = 1-\pi_i$

(ii) $\check{c}_i^* = n_h(n_h-1)^{-1}(1-\pi_i) = \check{c}_i$

(iii) $\check{c}_i^* = \check{c}_i + (n_h-1)^{-1}(1-\pi_i)\left[n_h^{-1}\sum_{j\in U}\pi_j^2 - \pi_i\right]$

The first choice ignores the correction of DF and is not recommended when few unit are sampled per stratum. The last choice depends on $\sum_{j\in U}\pi_j^2$ which is unknown if the π_i for $i \notin s$ are not available. With the second choice, the same weights are used in (3.2) and in (3.4). In this case, the only difference is in the regression coefficients: we have the WLS regression coefficients in (3.2), and the OLS regression coefficients in (3.4). In the rest of this section, Berger (2004a) shows why WLS regression coefficients are recommended.

3.5 Estimators for systematic sampling

A systematic sample is defined by CONNOR (1966)

$$s = \{i \in U : \pi_{i-1}^c < m_j \leqslant \pi_i^c; j = 1, \ldots, n\}.$$

Where $m_j = u + j - 1$, $\pi_i^c = \sum_{j \in U; j \leqslant i} \pi_j$ and $\pi_0^{(c)} = 0$ and u is a random number generated from a uniform distribution $U(0, 1)$.

Although, it is possible to compute exactly the π_{ij} of the systematic design (CONNOR, 1966; PINCIARO, 1978; HIDIROGLOU and GRAY, 1980), most π_{ij} equal zero, implying that the Sen-Yates-Grundy variance estimator (2.3) is biased and that there is no unbiased estimator for the variance (SÄRNDAL *et al*., 1992, p. 83). Therefore, the Sen-Yates-Grundy estimator can be misleading and should not be used for systematic (SÄRNDAL *et al*., 1992, p. 47). Thus, we cannot use the classical approach, which consists of computing or approximating the π_{ij}, as it gives nonsensical estimators. This is the reason why we will not consider (2.3) in our simulation study (See Chapter 5).

As already stated, (3.2) is suitable for high entropy sampling designs. Thus, (3.2) might not be suitable for systematic sampling, as the entropy of this sampling design is low. BERGER (2003, 2004a) proposes an adjusted Hájek variance estimator for systematic sampling that includes additional independent variables in the working regression (3.3). For example, if U is composed of a single stratum ($H = 1$), we have one additional independent variable given by

$$x_i = \begin{cases} (1 - n)\pi_i^c & \text{if } 0 < \tilde{\pi}_i^c \leqslant 1 \\ \pi_i^c & \text{otherwise} \end{cases} \tag{3.5}$$

where $\tilde{\pi}_i^c = \sum_{j \in U; j \leqslant i} \pi_j - \pi_i/2$ is a smooth cumulative sum of the π_i. A series of simulations in BERGER (2003) shows that it is recommended to add (3.5) in the working regression with systematic samples. We denote by $\widehat{\text{var}}(\widehat{\tau})_{Berger}$ the resulting estimator.

An alternative approach proposed by BREWER (2002), page 159 consists on creating pseudo strata and assuming high entropy within strata. We consider that 10 units per implicit stratum. The first 10 sampled units will be in the first implicit stratum, the next 10 units will be in the second implicit stratum etc... We will estimate the variance by the stratified Hájek estimator (3.2) where the strata are the implicit strata of the systematic sampling design. We denote this stratified Hájek estimator by $\widehat{\text{var}}_{HST}(\widehat{\tau})$. If we have a stratified sampling design, pseudo strata are created within each stratum.

Chapter 4

The German Income and Consumption Survey (ICS) Universe

The ICS survey is a quota sample selected in each German federal state (`http://www.destatis.de/download/veroe/methoden.pdf`). The ICS data does not contain individual, but only household information. Based on these data, populations or universes have been created (see workpackage 3). The numbers of households in the universes (or the population sizes) are given in Table 4.1. The variables available are given in Table 4.2.

Table 4.1: Population and sample size per German Federal states. The sizes considered for BAW are smaller than the actual size in brackets.

Federal state	Population Size	Sample Size
Schleswig-Holstein (SWH)	1258500	2752
Hamburg (HAM)	881400	2002
Niedersachsen (NIE)	3434500	6803
Bremen (BRE)	344600	860
Nordrhein-Westfalen (NRW)	8031800	14614
Hessen (HES)	2707700	5496
Rheinland-Pfalz (RLP)	1757600	3719
Baden-Wurttemberg (BAW)	3761360	7220
Bayern (BAY)	5339300	10118
Saarland (SAL)	507100	1213
Berlin - West (BER)	1180111	2434
Brandenburg (BRA)	1073700	2390
Mecklenburg-Vorpommern (MVP)	760800	1750
Sachsen (SAC)	2030200	4241
Sachsen-Anhalt (SAA)	1200600	2644
Thüringen (THN)	1075600	2398

The variables available in the universes are given in Table 4.2. The target parameters are 23 totals: the total household net income, the total expenditure, the total number of

household with i individuals ($i = 1, \ldots, 8$) and with more than 9 individuals, the total number of household with a socio economic status i ($i = 1, \ldots, 4$) and total number of the different type of household.

We propose to select unequal systematic samples per federal states. Systematic sampling is widely used by statistical offices and it can be considered as an approximation of quota sampling. With a linear model, we generate a size variable correlated with the variable expenditure and with a coefficient of correlation of 0.6. The π_i are proportional to the size variable.

Table 4.2: Variables included in the ICS pseudo universes.

Federal states:	
1...Schleswig-Holstein (SWH)	9...Bayern (BAY)
2...Hamburg (HAM)	10..Saarland (SAL)
3...Niedersachsen (NIE)	11..Berlin - West (BER)
4...Bremen (BRE)	12..Brandenburg (BRA)
5...Nordrhein-Westfalen (NRW)	13..Mecklenburg-Vorpommern (MVP)
6...Hessen (HES)	14..Sachsen (SAC)
7...Rheinland-Pfalz (RLP)	15..Sachsen-Anhalt (SAA)
8...Baden-Wurttemberg (BAW)	16..Thüringen (THN)
Number of individuals:	
1 . . . 8 1 to 8 individuals	9 9 or more individuals
Socio economic status:	
0. . . Self employed	3. . . Worker
1. . . Civil servant or military	4. . . Unemployed, pensioner,
2. . . Employee	students, others
Type of household:	
0. . . Other	
1. . . Mother/father alone + 1 child	
2. . . Mother/father alone + 2 or more children	
3. . . Couple with 1 child - spouse employed	
4. . . Couple with 1 child - spouse unemployed	
5. . . Couple with 2 or more children - spouse employed	
6. . . Couple with 2 or more children - spouse unemployed	
Household net income	
Total expenditure	

Chapter 5

Monte-Carlo simulations

In each German Federal state, we have selected 10 000 systematic samples with unequal probabilities. The sample sizes used are given in Table 4.1. In this report, we only consider the change in the number of unemployed and the total incomes for Bremen (the smallest state) and Nordrhein-Westfalen (the largest state). Bremen (BRE) has the largest sampling fraction and Nordrhein-Westfalen (NRW) has the smallest sampling fraction. The simulation results for the other variables and other states will be available in workpackage 1.

We suppose that each German state is a single stratum. We will compare the empirical sampling distribution of $\widehat{\text{var}}_{SRS}(\widehat{\tau})$, $\widehat{\text{var}}_{HH}(\widehat{\tau})$, $\widehat{\text{var}}_{H}(\widehat{\tau})$, $\widehat{\text{var}}_{Brewer}(\widehat{\tau})$ and $\widehat{\text{var}}_{Berger}(\widehat{\tau})$ (with a single stratum, $H = 1$). We consider a stratified Hájek estimator $\widehat{\text{var}}_{HST}(\widehat{\tau})$ with 10 household per pseudo stratum. We do not consider (2.3), as (2.3) is highly biased and not recommended for systematic sampling.

5.1 Total household income

The totals household income for Bremen (BRE) and Nordrhein-Westfalen (NRW) are 5 876 537 353 and 150 620 621 817. The sampling distribution of the standardised Horvitz-Thompson estimator $\widehat{\tau}_{st} = (\widehat{\tau} - \tau)/\tau$ is given in Figure 5.1. We see that the Horvitz-Thompson estimator is more accurate for NRW. This is what we would normally expect, as the sample size if larger for NRW.

In Figure 5.2, we present the empirical sampling distribution of $\widehat{var}(\widehat{\tau})_{\log} = \log(\widehat{\text{var}}(\widehat{\tau})/\text{var}(\widehat{\tau}))$ for $\widehat{\text{var}}_{SRS}(\widehat{\tau})$, $\widehat{\text{var}}_{HH}(\widehat{\tau})$, $\widehat{\text{var}}_{H}(\widehat{\tau})$, $\widehat{\text{var}}_{Brewer}(\widehat{\tau})$, $\widehat{\text{var}}_{Berger}(\widehat{\tau})$ and $\widehat{\text{var}}_{HST}(\widehat{\tau})$. $var(\widehat{\tau})$ is the empirical variance of $\widehat{\tau}$. We see that only $\widehat{\text{var}}_{SRS}(\widehat{\tau})$ has a different distribution, the other estimators have similar distributions. The variance estimates are more accurate with the BRE universe. The distributions of $\widehat{\text{var}}_{HH}(\widehat{\tau})$, $\widehat{\text{var}}_{H}(\widehat{\tau})$, $\widehat{\text{var}}_{Brewer}(\widehat{\tau})$, $\widehat{\text{var}}_{Berger}(\widehat{\tau})$ and $\widehat{\text{var}}_{HST}(\widehat{\tau})$ are very skewed, although unbiased.

In Table 5.1, we give the relative bias (RB) and relative mean square error (RMSE) of the estimator considered. We give also the 95% confidence interval (CI) coverage using the normal assumption. $\widehat{\text{var}}_{SRS}(\widehat{Y})$ is the less accurate estimator. The other estimators have the same accuracy. This can be explained by the small sampling fraction. Indeed,

the precision of the variance estimators defined in Section 3.2 – 3.5 might be different for large sampling fraction. BERGER (2004b) shows that $\widehat{\text{var}}_H(\widehat{\tau})$ is suitable for high entropy sampling and $\widehat{\text{var}}_{Berger}(\widehat{\tau})$ is suitable with systematic sampling.

The median of the time (in seconds) required for the computation of one variance estimate is given in Table 5.1. A 2GHz Pentium 4 CPUs with 1Gb of RAM has been used. Although $\widehat{\text{var}}_{Berger}(\widehat{\tau})$ is the most computing intensive, it can be computed in less than one second.

The variance estimators that takes the systematic sampling into account ($\widehat{\text{var}}_{Berger}(\widehat{Y})$ and $\widehat{\text{var}}_{HST}(\widehat{Y})$) has the same distribution as $\widehat{\text{var}}_H(\widehat{Y})$. Thus, for this variable, the Berger adjustment for systematic sampling has no effect. Systematic sampling is particularly efficient if there is a trend in the survey variable. In the absence of a trend, the systematic sampling is as accurate as a maximum entropy sampling design. There is probably no trend in the household incomes, implying that the distribution of $\widehat{\text{var}}_{Berger}(\widehat{Y})$ and $\widehat{\text{var}}_{HST}(\widehat{Y})$ have the same distribution as $\widehat{\text{var}}_H(\widehat{\tau})$ based on high entropy.

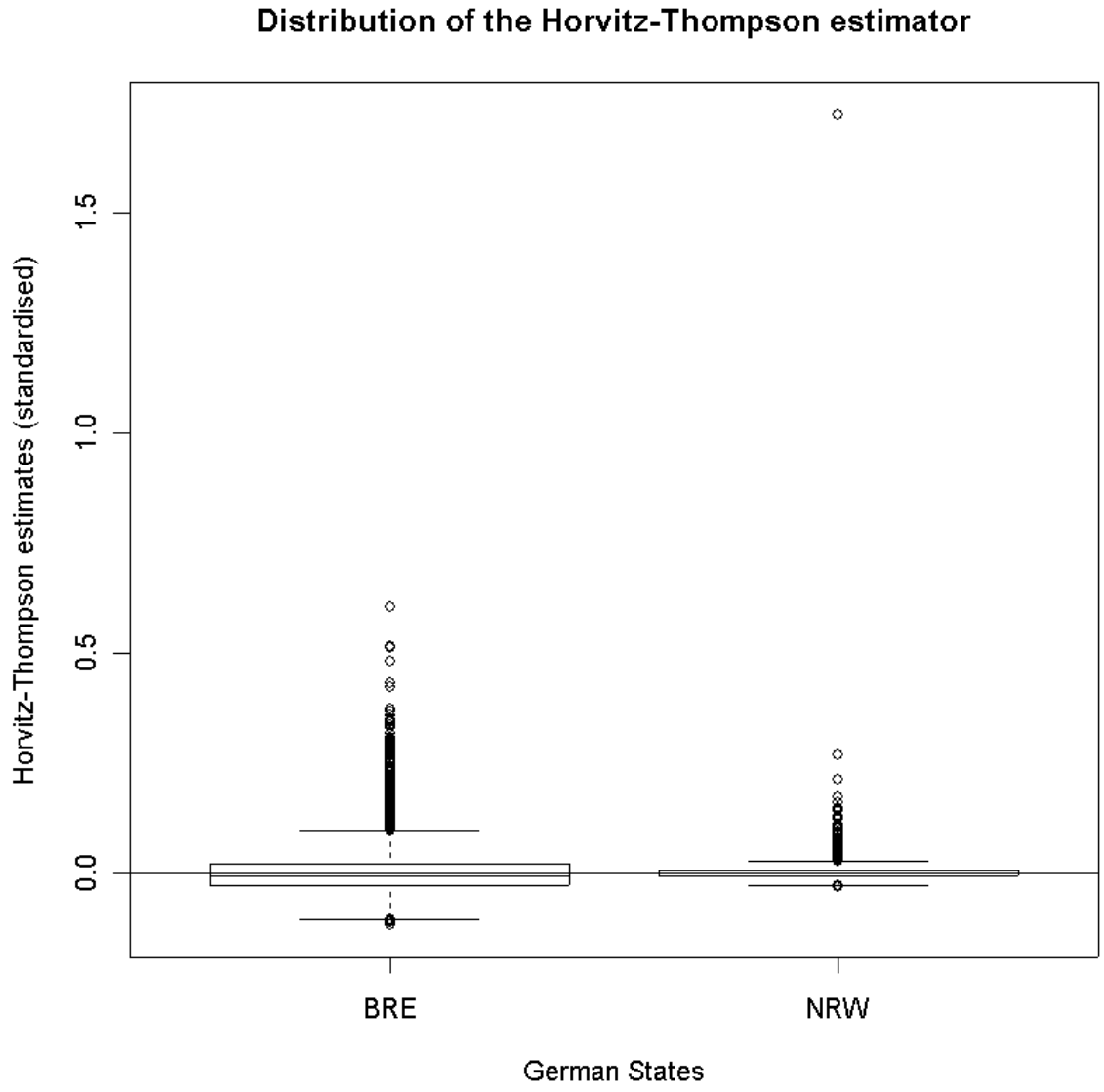

Figure 5.1: Sampling distribution of $\widehat{\tau}_{st}$ for BRE and NRW. τ = total income.

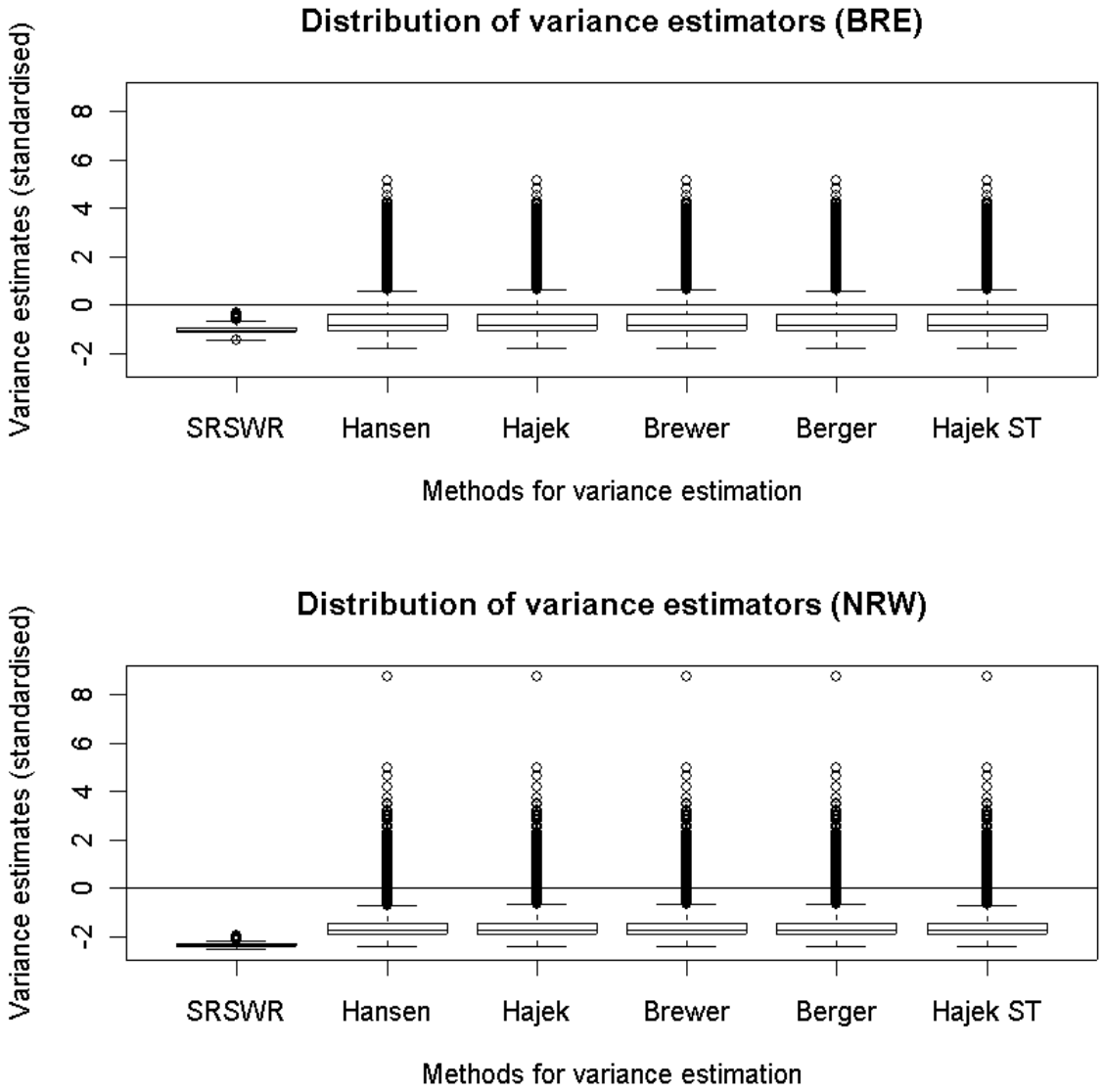

Figure 5.2: Empirical sampling distribution of $\widehat{\text{var}}(\widehat{\tau})_{\log}$. τ = total income.

5.2 Total number of unemployed

The total number of unemployed for BRE and NRW are 159 831 and 2 599 080. The sampling distribution of the standardised Horvitz-Thompson estimator $\widehat{\tau}_{st}$ is given in Figure 5.3. In Table 5.2, we have the RB and the RMSE of the estimator considered. In Figure 5.4, we have the empirical sampling distribution of $\widehat{\text{var}}(\widehat{\tau})_{\log}$ for the variance estimators considered. We can draw the same conclusion as in Section 5.1. We observe a large RMSE for the NRW universe. This is probably due to the outlying sample observed in Figure 5.3. Note that Hansen and the Berger variance estimators are slightly better for the BRE universe, as far as the RB and RMSE are concerned. Thus, for this variable, the Berger adjustment for systematic sampling has an effect.

Table 5.1: Relative bias (RB), the relative mean square error (RMSE) and 95% confidence coverage of the variance estimator considered. "Time (med.)" is the median of the time for computing one estimate. τ= total income.

		SRSWR	Hansen	Hájek	Brewer	Berger	Hájek Stratified
BRE	RB	-0.60	0.00	0.00	0.00	0.00	0.00
	RMSE	0.40	17.30	17.40	17.40	17.30	17.30
	Coverage	0.85	0.91	0.91	0.91	0.91	0.91
	Time (med.)	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec
NRW	RB	-0.90	0.00	0.00	0.00	0.00	0.00
	RMSE	0.80	4115.10	4130.10	4130.10	4128.20	4128.50
	Coverage	0.79	0.93	0.93	0.93	0.93	0.93
	Time (med.)	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec

Table 5.2: Relative bias (RB), the relative mean square error (RMSE) and 95% confidence coverage of the variance estimator considered. "Time (med.)" is the median of the time for computing one estimate. τ = total number of unemployed.

		SRSWR	Hansen	Hájek	Brewer	Berger	Hájek Stratified
BRE	RB	-0.90	0.00	0.10	0.10	0.00	0.10
	RMSE	0.80	52.40	52.70	52.70	52.40	52.60
	Coverage	0.62	0.92	0.92	0.92	0.92	0.92
	Time (med.)	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec
NRW	RB	-1.00	0.00	0.00	0.00	0.00	0.00
	RMSE	1.00	8620.70	8652.20	8652.20	8648.10	8648.80
	Coverage	0.58	0.93	0.93	0.93	0.93	0.93
	Time (med.)	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec	<1 sec

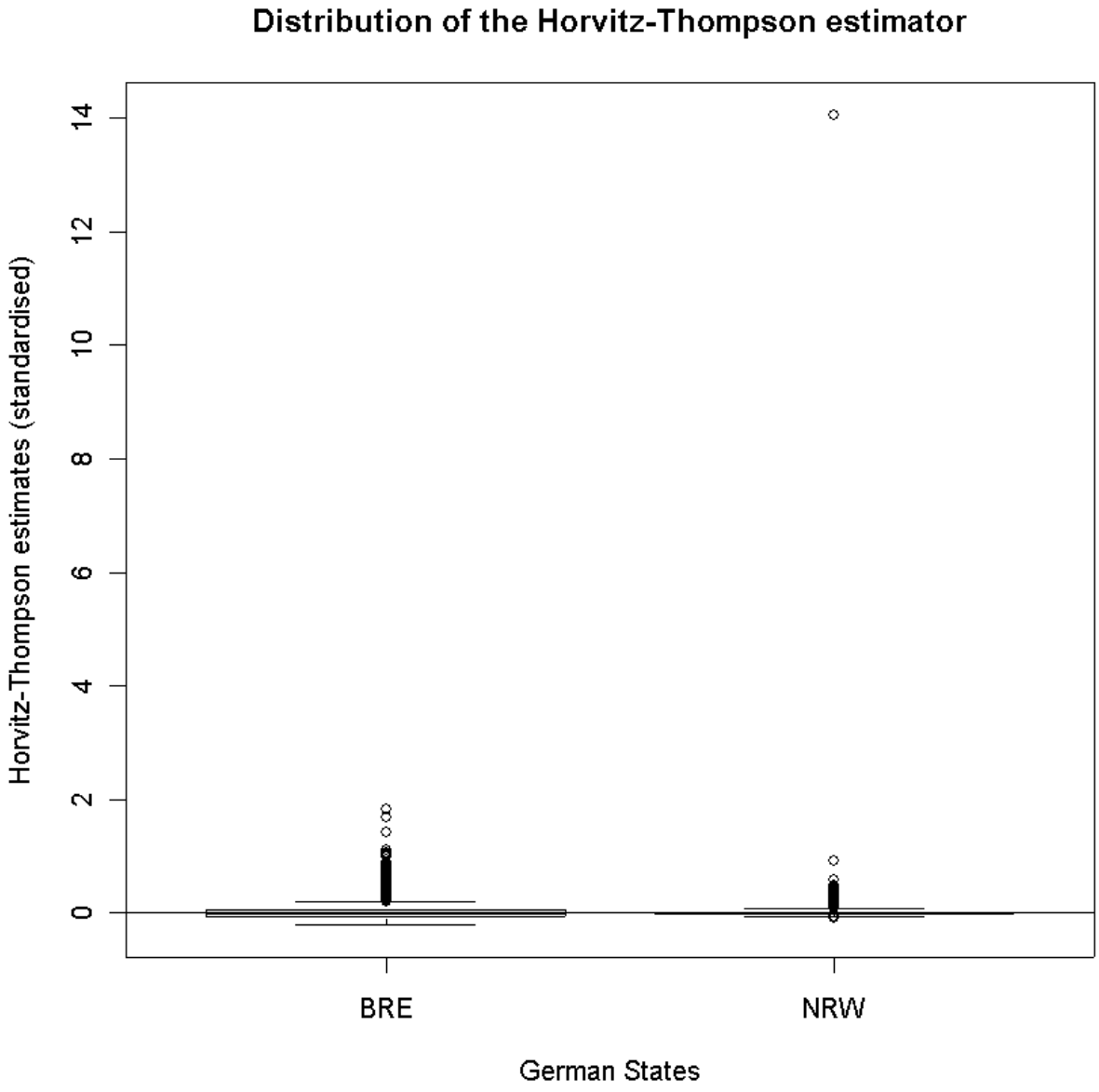

Figure 5.3: Sampling distribution of $\widehat{\tau}_{st}$ for BRE and NRW. τ = total number of unemployed.

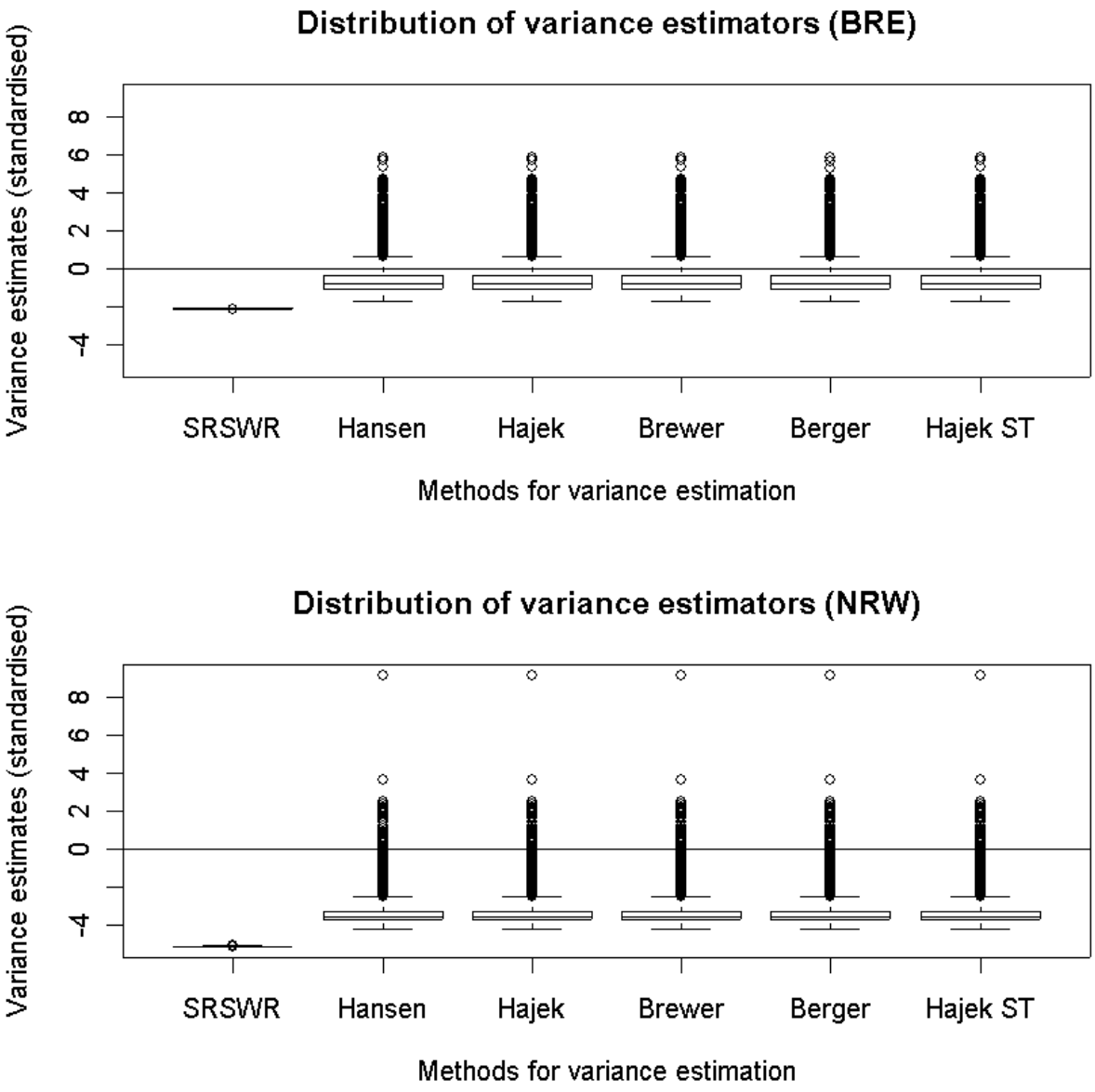

Figure 5.4: Empirical sampling distribution of $\widehat{\text{var}}(\widehat{\tau})_{\log}$. τ = total number of unemployed.

Chapter 6

Linearization and Jackknife

In this chapter, we show how the variance of a function of means can be estimated using jackknife.

Assume that the parameter of interest can be expressed as a function of means, $\theta = g(\mu_1, \ldots, \mu_Q)$, where $g(\,.\,)$ is a smooth function and μ_q is the finite population mean of a survey variable labelled q. Suppose that values y_{iq}, $q = 1, \ldots, Q$, for Q survey variables are associated with the unit labelled i. We now define the point estimator, $\widehat{\theta}$, as the substitution estimator, $\widehat{\theta} = g(\widehat{\mu}_1, \ldots, \widehat{\mu}_Q)$, where

$$\widehat{\mu}_q = \sum_{i \in s} w_i \, y_{iq}$$

is the HÁJEK (1981) ratio estimator of μ_q, the weight w_i is given by

$$w_i = \frac{1}{\widehat{N} \pi_i}$$

where $\widehat{N} = \sum_{i \in s} \pi_i^{-1}$.

We propose to estimate this variance by

$$\widehat{\mathrm{var}}(\widehat{\theta})_J = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \, \varepsilon_i \, \varepsilon_j \tag{6.1}$$

where

$$\sum_i = (1 - w_i) \, (\widehat{\theta} - \widehat{\theta}_{(i)})$$

$\widehat{\theta}_{(i)} = g(\widehat{\mu}_{1(i)}, \ldots, \widehat{\mu}_{Q(i)}), \widehat{\mu}_{q(j)} = \widehat{N}_{(j)}^{-1} \sum_{i \in s_j} \pi_i^{-1} \, y_{iq}, \widehat{N}_{(j)} = \sum_{i \in s_j} \pi_i^{-1}, s_{(j)}$ consists of s with the j-th unit deleted and n is the size of s. BERGER and SKINNER (2004) show that (6.1) is a consistent estimator for the variance. BERGER and RAO (2004) extended (6.1) for imputation.

The estimator in (6.1) takes the form of the variance estimator of HORVITZ and THOMPSON (1952) for the sample sum of empirical influence values (DAVISON and HINKLEY,

1997, Ch.2), where these empirical influence values are numerically approximated by the jackknife pseudo-values ε_i. This is analogous to the linearization variance estimator (SÄRNDAL *et al*, 1992, p.175) which takes the same form but with the empirical influence values obtained by analytic differentiation. This perspective was first set out by CAMPBELL (1980) who noted how both these estimators could be constructed, but did not proceed to evaluate their properties in detail.

The factor $(1 - w_i)$ is a correction for unequal π_i, reducing the contribution of observations which have higher π_i values and thus make smaller contributions to the variance. The inclusion of this factor ensures that (6.1) reduces to the usual linearization variance estimator (SÄRNDAL *et al*, 1992, page 182) when $\widehat{\theta}$ is the Hájek estimator, $\widehat{\mu}_1$, say, in which case ε_i reduces to $w_i(y_{1i} - \widehat{\mu}_1)$. The $(1 - w_i)$ correction was suggested by CAMPBELL (1980), who noted an algebraic equivalence with the weighted jackknife method of HINKLEY (1977).

The estimators of Chapter 3 can be used to approximate (6.1). It is only necessary to substitute $\check{y}_i$ by ε_i into Chapter 3's estimators.

Chapter 7

Conclusion

Variance with unequal probability sampling without replacement can be easily estimated with the HÁJEK (1964) variance estimator. This estimator can be easily computed as it is a weighted sum of residuals. This alternative expression is computationally simpler than the Sen-Yates-Grundy variance estimator and does not require computation of joint-inclusion probabilities.

References

Berger, Y. G. (1998): Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* **74**, 149–168.

Berger, Y. G. (2003): A modified Hájek variance estimator for systematic sampling. *Statistics in Transition* **6**, 5–21.

Berger, Y. G. (2004a): A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics* **31**, 305–315.

Berger, Y. G. (2004b): Variance estimation with Chao's sampling scheme. To appear in the Journal of Statistical Planning and Inference.

Berger, Y. G. and Rao, J. N. K. (2004): Adjusted jackknife for imputation under unequal probability sampling. Manuscript.

Berger, Y. G. and Skinner, C. J. (2004): A jackknife variance estimator for unequal probability sampling. to appear in the Journal of the Royal Statistical Society, Series B.

Brewer, K. R. W. (2002): *Combined Survey Sampling Inference (Weighing Basu's Elephants)*. Arnold publishers.

Campbell, C. (1980): A different view of finite population estimation. In *Proceedings of the Section on Survey Research Methods*, pp. 319–324. Alexandria, Virginia: American Statistical Association.

Chao, M. T. (1982): A general purpose unequal probability sampling plan. *Biometrika* **69**, 653–656.

Connor, W. S. (1966): An exact formula for the probability that two specified sample units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Society* **61**, 384–390.

Davison, A. C. and Hinkley, D. V. (1997): *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Deville, J. C. (1999): Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* **25**, 193–203.

Gabler, S. (1984): On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. *Biometrika* **71**, 171–175.

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

Hájek, J. (1981): *Sampling from a Finite Population*. New York, Marcel Dekker.

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

Hartley, H. O. and Rao, J. N. K. (1962): Sampling with unequal probabilities without replacement. *Annals of Mathematical Statistics* **33**, 350–374.

Hidiroglou, M. A. and Gray, G. B. (1980): Construction of joint probability of selection for systematic pps sampling. *Applied Statistics* **29**, 107–112.

Hinkley, D. V. (1977): Jackknife in unbalanced situations. *Technometrics* **19**, 285–292.

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

Narain, R. D. (1951): On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**, 169–174.

Pinciaro, S. J. (1978): An algorithm for calculating joint inclusion probabilities under PPS systematic sampling. Technical Report pp.740, ASA Proceedings of Survey Research Methods Section.

Rao, J. N. K. (1965): On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association* **3**, 173–180.

Rosén, B. (1991): Variance for systematic pps-sampling. Technical report, Report 1991:15, Statistics Sweden.

Sampford, M. R. (1967): On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 494–513.

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

Sen, P. K. (1953): On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

Smith, T. M. F. (2001): Biometrika centenary: Sample surveys. *Biometrika* **88**, 167–194.

Yates, F. and Grundy, P. M. (1953): Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society Serie B* **1**, 253–261.