Chapter 2
Design-Based Direct Estimation

2.1 Introduction

Survey samples provide useful information about a population and avoid the need of carrying out the more expensive and time-consuming censuses. Sampling theory covers sampling designs and inference procedures for finite populations. If the population is partitioned in domains, the estimators of parameters of the global populations can be adapted and applied to estimate domain parameters. This can be done by treating the domains as independent new populations. This approach to small area estimation yields to design-based direct estimators.

The estimation of small area parameters, like domain means, totals, or ratios of a target variable, is an inference problem in finite populations. Historically, the first estimators of population parameters defined at the domain level were adaptations of the corresponding estimators defined for the global population. Direct estimators use only the data of the target variable in the domain of interest, and their properties are studied and optimized with respect to the probability distribution of the sample design. They do not use data from other domains or time periods. Since direct estimators are simple and intuitive, researchers use them as a benchmark to establish comparisons and to measure the efficiency gain obtained by using more sophisticated small area estimators.

This manuscript dedicates an initial chapter to introduce the basic concepts and tools of sampling and inference in finite populations. Inclusion probabilities and their inverses (sampling weights) play here a relevant role. For estimating means and totals, two types of direct estimators are considered. They were introduced by Horvitz and Thompson (1952) and Hájek (1971), respectively. For estimating ratios, plug-in estimators are employed. They are defined by substituting totals by their corresponding direct estimators.

The chapter gives a short introduction to the survey sampling theory and describes some properties of direct estimators, with special emphasis on estimators of means, totals, and ratios. For each estimator, the design-based expectation and

D. Morales et al., *A Course on Small Area Estimation and Mixed Models*, Statistics for Social and Behavioral Sciences, https://doi.org/10.1007/978-3-030-63757-6_2

variance are calculated and a direct estimator of the variance is given. In many practical cases, only first order inclusion probabilities are available, and therefore it is not possible to calculate unbiased direct estimators of variances. This is why the chapter also presents design-based resampling methods, like bootstrap and Jackknife, for variance estimation. The last section contains some examples giving R codes, including functions for calculating domain-level direct estimators.

2.2 Survey Sampling Theory

A *finite population* is a collection of different units, such as people, companies, households, hospitals, and so on. The *survey sampling theory* deals with the selection of samples (subsets of the population), the observation of characteristics of sampled units, and the use of the obtained data for doing inferences about the population.

Survey sampling is interested in a fixed population from which a part is observed. In other branches of statistics, observations are realizations of random variables, and the inferences are not referred to any actual population, but to a probability law on the random variables. The following example clarifies this point.

Example 2.1 An industry is interested in determining if the units of a production line fulfill some given specifications. By assuming the general approach to statistics, we can model the data (CORRECT $=0$ and DEFECTIVE $=1$) as realizations of independent and identically distributed Bernoulli variables with parameter θ. The statistical target is the estimation of the probability θ of making a defective unit. The problem becomes a finite population survey sampling problem if we are only interested in the units produced during a given day. In the last case, we are interested in estimating the proportion

$$p=\frac{\text{number of defective units produced during the day}}{\text{number of units produced during the day}}.$$

In survey sampling, there are two main approaches. The first one assumes that the data obey the probability distribution given by the random extraction of samples from the population. This is the design-based approach. In the second case, the scores of the target variable are assumed to be the realization of a random vector with distribution given by a statistical model. This is the model-based approach. The inference procedures are built and studied depending on the assumed probability distribution.

Under the design-based approach, the vector containing the values of a variable y in all the population units $(y_1,\dots,y_N)$ is the basic parameter. A *probabilistic sampling plan (or design)* is a scheme for choosing the samples, such that each subset s of the population U has a known selection probability $p(s)$. Let us consider a population parameter T and its estimator $\widehat{T}$ based on s. The definitions of bias and variance of $\widehat{T}$ are based on $p(s)$, i.e.

$$\begin{array}{ll}\text{BIAS:} & E_\pi[\widehat{T}-T]=\sum_{s\subset U} p(s)[\widehat{T}(s)-T],\\ \text{VARIANCE:} & \mathrm{var}_\pi(\widehat{T})=\sum_{s\subset U} p(s)\big(\widehat{T}(s)-E_\pi[\widehat{T}]\big)^2.\end{array}$$

We use the notations E_π and var_π to emphasize the fact that we have expectations and variances with respect to the design-based probability distribution $p(s)$. Expectations and variances with respect to a model-based distribution are denoted by E_M and var_M.

In general, the calculation of $p(s)$ is not an easy task. Some simple cases are the simple random samplings with replacement (SRSWR) and without replacement (SRSWOR), i.e.

$$\begin{array}{ll} p(s)=\frac{1}{N^n} & \text{for a SRSWR sample } s \text{ of size } n,\\ p(s)=\frac{1}{\binom{N}{n}} & \text{for a SRSWOR sample } s \text{ of size } n.\end{array}$$

However, many calculations only require the inclusion probabilities π_i and π_{ij}, i.e.

$\pi_i = P(i \in s) = \sum_{s\in s(i)} p(s)$, where $s(i)=\{s\subset U : i\in s\}$ is the set of samples containing the unit i,

$\pi_{ij} = P(i \in s, j\in s) = \sum_{s\in s(i,j)} p(s)$, where $s(i,j)=\{s\subset U : i, j\in s\}$ is the set of samples containing the units i and j.

For example, under the SRSWOR, the inclusion probabilities are

$$\pi_i = n/N, \quad \pi_{ij}=\frac{n(n-1)}{N(N-1)} \quad \text{for } i, j\in U,\ i\neq j.$$

The following definition will be useful in some of the proofs.

Definition 2.1 The sampling design indicator functions are

$$\delta_i(s)=\begin{cases}1 & \text{if the unit } i \text{ is in the sample } s\\ 0 & \text{otherwise}\end{cases} \stackrel{d}{=} \text{Bernoulli}(\pi_i).$$

It holds that

(1) $\sum_{i=1}^N \delta_i(s)=n$, (2) $P(\delta_i(s)=1)=1-P(\delta_i(s)=0)=\pi_i$,
(3) $P(\delta_i(s)=1,\delta_j(s)=1)=\pi_{ij}$, (4) $\pi_{ii}=\pi_i$,
(5) $E_\pi[\delta_i(s)]=E_\pi[\delta_i^2(s)]=\pi_i$, (6) $E_\pi[\delta_i(s)\delta_j(s)]=\pi_{ij}$,
(7) $\mathrm{var}_\pi(\delta_i(s))=\pi_i(1-\pi_i)$, (8) $\mathrm{cov}_\pi(\delta_i(s),\delta_j(s))=\pi_{ij}-\pi_i\pi_j$.

In what follows, we simplify the notation and write δ_j instead of $\delta_j(s)$. Further, we consider only sampling without replacement, and we use the following notations:

- *Indexes:* s denotes a sample, and $d = 1, \ldots, D$, $j = 1, \ldots, N$, and $g = 1, \ldots, G$ denote domains (or small areas), units (or individuals), and groups, respectively.
- *Population and sample:* $U = \bigcup_{d=1}^{D} U_d$ for population and $s = \bigcup_{d=1}^{D} s_d$ for sample, where U_d and s_d are population and sample in domain d, respectively.
- *Sizes:* N for population and n for sample. When N and n have subindexes, they denote the corresponding size of the indexed set. For example, N_d is the population size of domain d.
- *Totals:* Y and X denote the population totals of variables y and x, respectively. If Y and X have subindexes, then they denote the corresponding totals of the indexed set.
- *Means:* $\overline{Y}$ and $\overline{X}$ denote the population means of variables y and x, respectively. If $\overline{Y}$ and $\overline{X}$ have subindexes, then they denote the corresponding means of the indexed set. For example, $\overline{Y}_d$ denotes the population mean of domain d.
- *Sampling weights:* w_j are the theoretical weights of the sampling design. They are the inverses of the inclusion probabilities, i.e. $w_j = 1/\pi_j$.

Example 2.2 For any individual j, interviewed at a labor force survey, some variables of interest are

$$y_j = \begin{cases} 1 \text{ if } j \text{ is unemployed,} \\ 0 \text{ otherwise,} \end{cases} \quad z_j = \begin{cases} 1 \text{ if } j \text{ is employed,} \\ 0 \text{ otherwise,} \end{cases} \quad t_j = \begin{cases} 1 \text{ if } j \text{ is inactive,} \\ 0 \text{ otherwise.} \end{cases}$$

Some target parameters are the totals of unemployed, employed, and inactive people and the unemployment rate, i.e.

$$Y_d = \sum_{j \in U_d} y_j, \quad Z_d = \sum_{j \in U_d} z_j, \quad T_d = \sum_{j \in U_d} t_j, \quad \text{and} \quad R_d = \frac{Y_d}{Y_d + Z_d} = \frac{\overline{Y}_d}{\overline{Y}_d + \overline{Z}_d},$$

where $\overline{Y}_d = Y_d/N_d$, $\overline{Z}_d = Z_d/N_d$, and N_d is the size of area d.

The following sections give estimators of the domain total and mean of a variable y, i.e.

$$Y_d = \sum_{j \in U_d} y_j, \quad \overline{Y}_d = \frac{1}{N_d} \sum_{j \in U_d} y_j.$$

Let us note that we assume that the units in U_d can be numbered, and in what follows, we sometimes use the notation

$$\sum_{j \in U_d} y_j = \sum_{j=1}^{N_d} y_j \,.$$

2.3 Direct Estimator of the Total and the Mean

Horvitz and Thompson (1952) proposed the following *direct* estimators of the total Y_d and the mean $\overline{Y}_d$ of domain d:

$$\hat{Y}_d^{dir1} = \sum_{j\in s_d} w_j y_j = \sum_{j\in s_d} \frac{1}{\pi_j} y_j, \quad \hat{\overline{Y}}_d^{dir1} = \frac{\hat{Y}_d^{dir1}}{N_d}, \tag{2.1}$$

where N_d is assumed to be known. Properties of these estimators are summarized in the following propositions.

Proposition 2.1 *If* $\pi_j > 0$, $\forall j \in U_d$, *then*

(a) $E_\pi\left[\hat{Y}_d^{dir1}\right] = Y_d$,

(b) $\text{var}_\pi\left(\hat{Y}_d^{dir1}\right) = \sum_{i\in U_d} \sum_{j\in U_d} (\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$, *and*

(c) $\widehat{\text{var}}_\pi\left(\hat{Y}_d^{dir1}\right)] = \sum_{i\in s_d}\sum_{j\in s_d} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$ *is an unbiased estimator of* $\text{var}_\pi\left(\hat{Y}_d^{dir1}\right)$.

Proof We give two proofs of (a). The first one works directly with the probability distribution of samples s. Let $s_d(j)$ be the set of all samples such that $j \in s_d = s \cap U_d$. It holds that

$$E_\pi\left[\sum_{j\in s_d}\frac{y_j}{\pi_j}\right] = \sum_s p(s)\sum_{j\in s_d}\frac{y_j}{\pi_j} = \frac{y_1}{\pi_1}\sum_{s\in s_d(1)} p(s) + \frac{y_2}{\pi_2}\sum_{s\in s_d(2)} p(s)$$
$$+ \cdots + \frac{y_N}{\pi_N}\sum_{s\in s_d(N)} p(s) = \sum_{j\in U_d}\frac{y_j}{\pi_j}\pi_j = \sum_{j\in U_d} y_j = Y_d,$$

since $s_d(j) = \emptyset$ for $j \notin U_d$.

An alternative and more simpler proof is obtained by applying the indicator functions δ_j, i.e.

$$E_\pi\left[\hat{Y}_d^{dir1}\right] = E_\pi\left[\sum_{j\in s_d}\frac{y_j}{\pi_j}\right] = E_\pi\left[\sum_{i\in U_d}\frac{y_i}{\pi_i}\delta_i\right] = \sum_{i\in U_d}\frac{y_i}{\pi_i}E_\pi\left[\delta_i\right] = \sum_{i\in U_d} y_i = Y_d.$$

(b) By using the indicator functions, we get

$$\mathrm{var}_\pi\big(\hat{Y}_d^{dir1}\big) = \mathrm{var}_\pi\left(\sum_{j\in s_d}\frac{y_j}{\pi_j}\right) = \mathrm{var}_\pi\left(\sum_{i\in U_d}\frac{y_i}{\pi_i}\delta_i\right) = \sum_{i\in U_d}\sum_{j\in U_d}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\mathrm{cov}_\pi\left(\delta_i,\delta_j\right)$$
$$= \sum_{i\in U_d}\sum_{j\in U_d}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

(c) By using the indicator functions, we have

$$E_\pi\left[\widehat{\mathrm{var}}_\pi\big(\hat{Y}_d^{dir1}\big)\right] = \sum_{i\in U_d}\sum_{j\in U_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_j}{\pi_j}\frac{y_j}{\pi_j}E_\pi\left[\delta_i\delta_j\right]$$
$$= \sum_{i\in U_d}\sum_{j\in U_d}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} = \mathrm{var}_\pi\big(\hat{Y}_d^{dir1}\big).$$

□

Corollary 2.1 *If $\pi_j > 0,\ \forall j \in U_d$, then*

(a) $E_\pi\Big[\hat{\overline{Y}}_d^{dir1}\Big] = \overline{Y}_d$,

(b) $\mathrm{var}_\pi\Big(\hat{\overline{Y}}_d^{dir1}\Big) = \frac{1}{N_d^2}\sum_{i\in U_d}\sum_{j\in U_d}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$, *and*

(c) $\widehat{\mathrm{var}}_\pi\Big(\hat{\overline{Y}}_d^{dir1}\Big) = \frac{1}{N_d^2}\sum_{i\in s_d}\sum_{j\in s_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$ *is an unbiased estimator of* $\mathrm{var}_\pi\Big(\hat{\overline{Y}}_d^{dir1}\Big)$.

Let us consider now a simple random sampling design without replacement inside each domain (SRSWORD). This is to say, we consider a stratified random sampling design where the strata are the domains and the domain samples, $n_1,\dots,n_D$, are fixed. For $i, j \in U_d$ we have

$$\pi_i = n_d/N_d, \quad \pi_{ii} = \pi_i = n_d/N_d, \quad \pi_{ij} = \frac{n_d(n_d-1)}{N_d(N_d-1)} \ \text{ if } i \neq j.$$

Proposition 2.2 *Under a SRSWORD design, the variance of the direct estimator of the total is*

$$\mathrm{var}_\pi\big(\hat{Y}_d^{dir1}\big) = \frac{(1-f_d)N_d^2}{n_d}S_{yd}^2, \quad S_{yd}^2 = \frac{1}{N_d-1}\sum_{i\in U_d}(y_i-\overline{Y}_d)^2, \quad f_d = \frac{n_d}{N_d}.$$

Proof It holds that

$$\mathrm{var}_\pi\left(\hat{Y}_d^{dir1}\right) = \sum_{i=1}^{N_d}(\pi_{ii} - \pi_i^2)\frac{y_i^2}{\pi_i^2} + \sum_{\substack{i=1\\i\neq j}}^{N_d}\sum_{j=1}^{N_d}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

$$= \sum_{i=1}^{N_d}\frac{n_d}{N_d}\left(1-\frac{n_d}{N_d}\right)\frac{N_d^2}{n_d^2}\,y_i^2 + \sum_{\substack{i=1\\i\neq j}}^{N_d}\sum_{j=1}^{N_d}\left(\frac{n_d(n_d-1)}{N_d(N_d-1)} - \frac{n_d^2}{N_d^2}\right)\frac{N_d^2}{n_d^2}\,y_i y_j$$

$$= \sum_{i=1}^{N_d}\frac{N_d-n_d}{n_d}\,y_i^2 + \sum_{\substack{i=1\\i\neq j}}^{N_d}\sum_{j=1}^{N_d}\frac{(n_d-N_d)}{(N_d-1)n_d}\,y_i y_j$$

$$= \frac{N_d-n_d}{n_d}\left[\sum_{i=1}^{N_d}y_i^2 - \frac{1}{N_d-1}\sum_{\substack{i=1\\i\neq j}}^{N_d}\sum_{j=1}^{N_d}y_i y_j\right] = \frac{N_d-n_d}{n_d}\left[\sum_{i=1}^{N_d}y_i^2\left(1+\frac{1}{N_d-1}\right)\right.$$

$$\left. - \frac{1}{N_d-1}\left(\sum_{i=1}^{N_d}y_i\right)^2\right] = \frac{(N_d-n_d)N_d}{n_d}\left[\frac{1}{N_d-1}\sum_{i=1}^{N_d}y_i^2 - \frac{Y_d^2}{N_d(N_d-1)}\right]$$

$$= \frac{(N_d-n_d)N_d}{n_d}\,S_{yd}^2 = \frac{(1-f_d)N_d^2}{n_d}\,S_{yd}^2.$$

□

In sampling designs with $\pi_{ij} = \pi_i\pi_j,\ i \neq j$, and $\pi_{jj} = \pi_j$, it holds that

$$\mathrm{var}_\pi\left(\hat{Y}_d^{dir1}\right) = \sum_{j\in U_d}\frac{1-\pi_j}{\pi_j}\,y_j^2 = \sum_{j\in U_d}(w_j-1)y_j^2, \tag{2.2}$$

$$\widehat{\mathrm{var}}_\pi\left(\hat{Y}_d^{dir1}\right) = \sum_{j\in s_d}\frac{1-\pi_j}{\pi_j^2}\,y_j^2 = \sum_{j\in s_d}w_j(w_j-1)y_j^2. \tag{2.3}$$

For the estimator of the domain mean, we have

$$\mathrm{var}_\pi\left(\hat{\overline{Y}}_d^{dir1}\right) = \frac{1}{N_d^2}\sum_{j\in U_d}(w_j-1)y_j^2, \quad \widehat{\mathrm{var}}_\pi\left(\hat{\overline{Y}}_d^{dir1}\right) = \frac{1}{N_d^2}\sum_{j\in s_d}w_j(w_j-1)y_j^2. \tag{2.4}$$

The equalities $\pi_{ij} = \pi_i\pi_j,\ i \neq j$, hold for the Bernoulli sampling (BS) design and the SRSWR design. In sampling designs with $\pi_{ij} \approx \pi_i\pi_j$ if $i \neq j$ (i.e. under SRSWOR), the above formulas are approximations. If a SRSWORD design is employed, the approximation (2.2) is an upper bound of the variance of the estimator of the total, i.e.

$$\sum_{j\in U_d} \frac{1-\pi_j}{\pi_j}\, y_j^2 = \sum_{j=1}^{N_d} \frac{1-\frac{n_d}{N_d}}{n_d/N_d}\, y_j^2 = \sum_{j=1}^{N_d} \frac{N_d-n_d}{n_d}\, y_j^2 = \frac{(1-f_d)N_d^2}{n_d}\frac{1}{N_d}\sum_{j=1}^{N_d} y_j^2$$
$$= \frac{(1-f_d)N_d^2}{n_d}\left[\frac{N_d-1}{N_d}S_{yd}^2 + \overline{Y}_d^2\right] > \frac{(1-f_d)N_d^2}{n_d}\, S_{yd}^2 = \text{var}_\pi\left(\hat{Y}_d^{dir1}\right),$$

where the inequality holds if N_d is large enough and $\overline{Y}_d$ is not too close to zero. Särndal et al. (1992, p. 170), present the following formula for the covariance between two direct estimators:

$$\text{cov}_\pi(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1}) = \sum_{i\in U_d}\sum_{j\in U_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\, y_i z_j.$$

An unbiased estimator of the covariance is

$$\widehat{\text{cov}}_\pi(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1}) = \sum_{i\in s_d}\sum_{j\in s_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\, y_i z_j.$$

Remark 2.1 In sampling designs with $\pi_{ij} = \pi_i\pi_j,\ i \neq j$, and $\pi_{jj} = \pi_j$, we have

$$\text{cov}_\pi(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1}) = \sum_{j\in U_d} \frac{1-\pi_j}{\pi_j}\, y_j z_j,$$
$$\widehat{\text{cov}}_\pi(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1}) = \sum_{j\in s_d} \frac{1-\pi_j}{\pi_j^2}\, y_j z_j = \sum_{j\in s_d} w_j(w_j-1)\, y_j z_j,$$
$$\text{cov}_\pi(\hat{\overline{Y}}_d^{dir1}, \hat{\overline{Z}}_d^{dir1}) = \frac{1}{N_d^2}\sum_{j\in U_d} \frac{1-\pi_j}{\pi_j}\, y_j z_j,$$
$$\widehat{\text{cov}}_\pi(\hat{\overline{Y}}_d^{dir1}, \hat{\overline{Z}}_d^{dir1}) = \frac{1}{N_d^2}\sum_{j\in s_d} \frac{1-\pi_j}{\pi_j^2}\, y_j z_j = \frac{1}{N_d^2}\sum_{j\in s_d} w_j(w_j-1)\, y_j z_j.$$

Remark 2.2 For calculating $\widehat{Y}_d^{dir1}$, we need the sampling weights and the locations of sampled units. This is to say, we need the data $y_j,\ w_j,\ I_{U_d}(j),\ j \in s$, where $I_{U_d}(j)$ is the indicator function, i.e. $I_{U_d}(j) = 1$ if $j \in U_d$ and $I_{U_d}(j) = 0$ otherwise.

2.4 Estimator of the Ratio

In applications of statistical inference in finite populations, we often find situations where the target parameter is a ratio. Examples of ratio-type parameters are the unemployment rate or the domain mean when the population size in the denominator is unknown. This section gives some properties of estimators defined as a ratio of direct estimators of domain totals. Let us consider the domain ratio $R_d = Y_d/Z_d$, where $Y_d = \sum_{j\in U_d} y_j$ and $Z_d = \sum_{j\in U_d} z_j$, and the ratio estimator $\hat{R}_d = \hat{Y}_d^{dir1}/\hat{Z}_d^{dir1}$.

Proposition 2.3 *The standardized bias of $\hat{R}_d$ fulfills the inequality*

$$(B_\pi^{rel}[\hat{R}_d])^2 = \frac{(E_\pi[\hat{R}_d]-R_d)^2}{\mathrm{var}_\pi(\hat{R}_d)} \le \frac{\mathrm{var}_\pi(\hat{Z}_d^{dir1})}{Z_d^2}.$$

Proof It holds that

$$\begin{aligned}\mathrm{cov}_\pi(\hat{R}_d,\hat{Z}_d^{dir1}) &= E_\pi[\hat{R}_d\hat{Z}_d^{dir1}] - E_\pi[\hat{R}_d]E_\pi[\hat{Z}_d^{dir1}]\\ &= E_\pi\left[\hat{Y}_d^{dir1}\right] - E_\pi[\hat{R}_d]E_\pi[\hat{Z}_d^{dir1}]\\ &= Y_d - E_\pi[\hat{R}_d]Z_d = -Z_d\left(E_\pi[\hat{R}_d]-R_d\right).\end{aligned}$$

Therefore,

$$E_\pi[\hat{R}_d]-R_d = -\frac{\mathrm{cov}_\pi(\hat{R}_d,\hat{Z}_d^{dir1})}{Z_d}.$$

By squaring both sides of the equality and using the symbol ρ_π for correlation with respect to the design-based probability, we obtain

$$\begin{aligned}(E_\pi[\hat{R}_d]-R_d)^2 &= \frac{\left[\mathrm{cov}_\pi(\hat{R}_d,\hat{Z}_d^{dir1})\right]^2}{Z_d^2} = \frac{\rho_\pi^2(\hat{R}_d,\hat{Z}_d^{dir1})\mathrm{var}_\pi(\hat{R}_d)\mathrm{var}_\pi(\hat{Z}_d^{dir1})}{Z_d^2}\\ &\le \frac{\mathrm{var}_\pi(\hat{R}_d)\mathrm{var}_\pi(\hat{Z}_d^{dir1})}{Z_d^2},\end{aligned}$$

which proves the stated result. □

Proposition 2.3 gives the following conclusion: if

$$B_\pi^{rel}[\hat{R}_d] = \frac{B_\pi[\hat{R}_d]}{(\mathrm{var}_\pi(\hat{R}_d))^{1/2}} = \frac{E_\pi[\hat{R}_d]-R_d}{(\mathrm{var}_\pi(\hat{R}_d))^{1/2}}$$

is the standardized bias of the ratio estimator $\hat{R}_d$, then

$$(B_\pi^{rel}[\hat{R}_d])^2 \leq \frac{\mathrm{var}_\pi(\hat{Z}_d^{dir1})}{Z_d^2}.$$

Note that if the relative standard error (sampling error),

$$\frac{\sqrt{\mathrm{var}_\pi(\hat{Z}_d^{dir1})}}{Z_d},$$

of the denominator of $\hat{R}_d$ tends to zero when the sample size increases, then the relative bias of $\hat{R}_d$ also tends to zero. This is an important property for building ratio estimators.

Proposition 2.4 *If $\hat{Y}_d^{dir1}$ and $\hat{Z}_d^{dir1}$ are consistent estimators of Y_d and Z_d, respectively, then*

(a) $\hat{R}_d$ is approximately unbiased.
(b) If n_d is large enough, an approximation to the variance of $\hat{R}_d$ is

$$\mathrm{var}_\pi(\hat{R}_d) \approx \frac{1}{Z_d^2}\sum_{i\in U_d}\sum_{j\in U_d}(\pi_{ij}-\pi_i\pi_j)\frac{y_i - R_d z_i}{\pi_i}\frac{y_j - R_d z_j}{\pi_j}.$$

Proof The estimator $\hat{R}_d$ is a function of two variables, i.e.

$$\hat{R}_d = \frac{\hat{Y}_d^{dir1}}{\hat{Z}_d^{dir1}} = f(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1}).$$

As the partial derivatives of f are $\frac{\partial f}{\partial y} = \frac{1}{z}$ and $\frac{\partial f}{\partial z} = -\frac{y}{z^2}$, a first order Taylor series expansion of $f(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1})$ around (Y_d, Z_d) yields to

$$\begin{aligned}\hat{R}_d &= f(\hat{Y}_d^{dir1}, \hat{Z}_d^{dir1}) \approx f(Y_d, Z_d) + \frac{\partial f(Y_d, Z_d)}{\partial y}(\hat{Y}_d^{dir1} - Y_d)\\ &+ \frac{\partial f(Y_d, Z_d)}{\partial z}(\hat{Z}_d^{dir1} - Z_d) = R_d + \frac{1}{Z_d}(\hat{Y}_d^{dir1} - Y_d) - \frac{Y_d}{Z_d^2}(\hat{Z}_d^{dir1} - Z_d)\\ &= R_d + \frac{1}{Z_d}(\hat{Y}_d^{dir1} - R_d\hat{Z}_d^{dir1}) = R_d + \frac{1}{Z_d}\sum_{j\in s_d}\frac{y_j - R_d z_j}{\pi_j}. \qquad (2.5)\end{aligned}$$

(a) By taking expectations in (2.5), we have

$$E_\pi[\hat{R}_d] \approx R_d + \frac{1}{Z_d}(Y_d - R_d Z_d) = R_d.$$

(b) By taking variances in (2.5) and using the sampling design indicator function δ_j, we get

$$\operatorname{var}_\pi(\hat{R}_d) \approx \frac{1}{Z_d^2} \sum_{i \in U_d} \sum_{j \in U_d} \frac{y_i - R_d z_i}{\pi_i} \frac{y_j - R_d z_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j).$$

□

An estimator of the approximated variance of $\hat{R}_d$ is

$$\widehat{\operatorname{var}}_\pi(\hat{R}_d) = \frac{1}{(\hat{Z}_d^{dir1})^2} \sum_{i \in s_d} \sum_{j \in s_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i - \hat{R}_d z_i}{\pi_i} \frac{y_j - \hat{R}_d z_j}{\pi_j}. \quad (2.6)$$

The estimator $\widehat{\operatorname{var}}_\pi(\hat{R}_d)$ is approximately unbiased if $E_\pi[\hat{R}_d] \approx R_d$ and $\operatorname{var}_\pi(\hat{Z}_d^{dir1}) \approx 0$. Otherwise, it is biased.

2.5 Other Direct Estimators of the Mean and the Total

Hájek (1971) proposed the following *direct* estimators of the domain mean and total:

$$\hat{\overline{Y}}_d^{dir2} = \frac{\hat{Y}_d^{dir1}}{\hat{N}_d} = \frac{\sum_{j \in s_d} w_j y_j}{\sum_{j \in s_d} w_j}, \quad \hat{Y}_d^{dir2} = N_d \hat{\overline{Y}}_d^{dir2}. \quad (2.7)$$

These estimators have the following properties.

Proposition 2.5 *If n_d is large enough and $\pi_j > 0 \; \forall j \in U_d$, then*

(a) $E_\pi[\hat{\overline{Y}}_d^{dir2}] \approx \overline{Y}_d$ *and*

(b) $\operatorname{var}_\pi(\hat{\overline{Y}}_d^{dir2}) \approx \frac{1}{N_d^2} \sum_{i \in U_d} \sum_{j \in U_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} (y_i - \overline{Y}_d)(y_j - \overline{Y}_d).$

Proof Let $z_j = 1 \; \forall j \in U_d$, then $Z_d = N_d$ and

$$R_d = \frac{Y_d}{Z_d} = \frac{Y_d}{N_d} = \overline{Y}_d.$$

The ratio estimator of R_d is

$$\hat{R}_d = \frac{\hat{Y}_d^{dir1}}{\hat{Z}_d^{dir1}} = \frac{\sum_{j \in s_d} w_j y_j}{\sum_{j \in s_d} w_j} = \hat{\overline{Y}}_d^{dir2}.$$

Since the Hájek estimator is consistent, the proof follows immediately from Proposition 2.4. □

An estimator of the approximated variance of $\overline{Y}_d^{dir2}$ is

$$\widehat{\text{var}}_\pi\left(\hat{\overline{Y}}_d^{dir2}\right) = \frac{1}{\hat{N}_d^2}\sum_{i\in s_d}\sum_{j\in s_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\left(y_i-\hat{\overline{Y}}_d^{dir2}\right)\left(y_j-\hat{\overline{Y}}_d^{dir2}\right). \tag{2.8}$$

Corollary 2.2 *If n_d is large enough and $\pi_j > 0\ \forall j \in U_d$, then*

(a) $E_\pi[\hat{Y}_d^{dir2}] \approx Y_d$ *and*
(b) $var_\pi\left(\hat{Y}_d^{dir2}\right) \approx \sum_{i\in U_d}\sum_{j\in U_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(y_i-\overline{Y}_d)(y_j-\overline{Y}_d).$

An estimator of the approximated variance of $\hat{Y}_d^{dir2}$ is

$$\widehat{\text{var}}_\pi\left(\hat{Y}_d^{dir2}\right)] = \left(\frac{N_d}{\hat{N}_d}\right)^2\sum_{i\in s_d}\sum_{j\in s_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\left(y_i-\hat{\overline{Y}}_d^{dir2}\right)\left(y_j-\hat{\overline{Y}}_d^{dir2}\right). \tag{2.9}$$

Remark 2.3 In the case $\pi_{ij} = \pi_i\pi_j,\ i \neq j$, and $\pi_{jj} = \pi_j$, we get

$$\text{var}_\pi\left(\hat{\overline{Y}}_d^{dir2}\right) \approx \frac{1}{N_d^2}\sum_{j\in U_d}\frac{1-\pi_j}{\pi_j}(y_j-\overline{Y}_d)^2,$$

$$\text{var}_\pi\left(\hat{Y}_d^{dir2}\right) \approx \sum_{j\in U_d}\frac{1-\pi_j}{\pi_j}(y_j-\overline{Y}_d)^2,$$

$$\widehat{\text{var}}_\pi(\hat{\overline{Y}}_d^{dir2}) = \frac{1}{\hat{N}_d^2}\sum_{j\in s_d}\frac{1-\pi_j}{\pi_j^2}\left(y_j-\hat{\overline{Y}}_d^{dir2}\right)^2$$

$$= \frac{1}{\hat{N}_d^2}\sum_{j\in s_d}w_j(w_j-1)\left(y_j-\hat{\overline{Y}}_d^{dir2}\right)^2,$$

$$\widehat{\text{var}}_\pi(\hat{Y}_d^{dir2}) = \frac{N_d^2}{\hat{N}_d^2}\sum_{j\in s_d}w_j(w_j-1)\left(y_j-\hat{\overline{Y}}_d^{dir2}\right)^2.$$

Estimators of the covariance between two direct estimators of domain means and totals, respectively, are

$$\widehat{\text{cov}}_\pi(\hat{\overline{Y}}_d^{dir2}, \hat{\overline{Z}}_d^{dir2}) = \frac{1}{\hat{N}_d^2}\sum_{j\in s_d}w_j(w_j-1)\left(y_j-\hat{\overline{Y}}_d^{dir2}\right)\left(z_j-\hat{\overline{Z}}_d^{dir2}\right),$$

$$\widehat{\text{cov}}_\pi(\hat{Y}_d^{dir2}, \hat{Z}_d^{dir2}) = \frac{N_d^2}{\hat{N}_d^2}\sum_{j\in s_d}w_j(w_j-1)\left(y_j-\hat{\overline{Y}}_d^{dir2}\right)\left(z_j-\hat{\overline{Z}}_d^{dir2}\right).$$

Under the SRSWORD design, it holds that

$$\hat{\overline{Y}}_d^{dir2} = \frac{\sum_{j\in s_d} \frac{N_d}{n_d} y_j}{\sum_{j\in s_d} \frac{N_d}{n_d}} = \frac{\frac{N_d}{n_d}}{N_d} \sum_{j\in s_d} y_j = \frac{1}{n_d} \sum_{j\in s_d} y_j = \bar{y}_{ds},$$

$$\widehat{\text{var}}_\pi(\hat{\overline{Y}}_d^{dir2}) = \frac{1}{N_d^2} \sum_{j\in s_d} \frac{N_d}{n_d} \frac{N_d - n_d}{n_d} (y_j - \bar{y}_{ds})^2 = \frac{1 - f_d}{n_d} \frac{1}{n_d} \sum_{j\in s_d} (y_j - \bar{y}_{ds})^2$$
$$\approx (1 - f_d) \frac{s_{yd}^2}{n_d},$$

where

$$s_{yd}^2 = \frac{1}{n_d - 1} \sum_{j\in s_d} (y_j - \bar{y}_{ds})^2.$$

As the direct estimator is approximately unbiased, the mean squared error and its estimator are

$$MSE(\hat{\overline{Y}}_d^{dir2}) \approx \text{var}_\pi\left(\hat{\overline{Y}}_d^{dir2}\right), \quad mse(\hat{\overline{Y}}_d^{dir2}) = \widehat{\text{var}}_\pi(\hat{\overline{Y}}_d^{dir2}).$$

For more details, see Särndal et al. (1992, pp. 185, 391), or Rao (2003, p. 12).

Although it is difficult to establish general conditions under which $\hat{\overline{Y}}_d^{dir2}$ is preferred to $\hat{\overline{Y}}_d^{dir1}$, Särndal et al. (1992, pp. 183–184), show some facts in favor of the first one.

1. By comparing the variances of both estimators, we have that $\hat{\overline{Y}}_d^{dir2}$ is preferred when the values of $y_j - \overline{Y}_d$ tend to be small. An extreme case is $y_j = c \; \forall j \in U_d$. In this case, it holds that

$$\overline{Y}_d = c, \qquad \hat{\overline{Y}}_d^{dir1} = c \frac{\sum_{j\in s_d} w_j}{N_d} = c \frac{\hat{N}_d}{N_d}, \qquad \hat{\overline{Y}}_d^{dir2} = c \frac{\hat{N}_d}{\hat{N}_d} = c.$$

   As $\text{var}_\pi\left(\hat{\overline{Y}}_d^{dir2}\right) = 0$, $\hat{\overline{Y}}_d^{dir2}$ is preferred to $\hat{\overline{Y}}_d^{dir1}$ if $\text{var}_\pi(\hat{N}_d) > 0$.
2. The estimator $\hat{\overline{Y}}_d^{dir2}$ behaves better than $\hat{\overline{Y}}_d^{dir1}$ when the sample size varies. If the sample size realization, $n_d = n_d(s)$, is larger than the average sample size, then the numerator and the denominator have many summands in $\hat{\overline{Y}}_d^{dir2}$. In the opposite case, the numerator and the denominator have few summands in $\hat{\overline{Y}}_d^{dir2}$. In this way, the ratio has some kind of stability. However, $\hat{\overline{Y}}_d^{dir1}$ does not present this stability because its denominator is a known constant.

In the case of the Bernoulli sampling where each individual is included in the sample independently with probability $\pi_j = \pi$, if $y_j = c\ \forall j \in U_d$, it holds that

$$\hat{\overline{Y}}_d^{dir1} = c\,\frac{n_d(s)}{\pi N_d}, \qquad \hat{\overline{Y}}_d^{dir2} = c.$$

Therefore, the variability of $\hat{\overline{Y}}_d^{dir1}$ is only ought to the variability of n_d for different samples s. In this case, $\mathrm{var}_\pi(\hat{\overline{Y}}_d^{dir1}) > \mathrm{var}_\pi(\hat{\overline{Y}}_d^{dir2}) = 0$.

3. Another situation where $\hat{\overline{Y}}_d^{dir2}$ is preferred to $\hat{\overline{Y}}_d^{dir1}$ is when the sample contains large values y_j of the target variable associated to small inclusion probabilities π_j. In this case, the value of the numerator of both estimators tends to be quite large. This fact is compensated by $\hat{\overline{Y}}_d^{dir2}$ because its denominator also tends to be large. This compensation produces stability. However, the denominator of $\hat{\overline{Y}}_d^{dir1}$ is constant and does not compensate the extreme values of the numerator.

Särndal et al. (1992, p. 184), give the following example that illustrates the above described situation. Let us consider a domain d with $N_d = 10$ units $y_1 = \ldots = y_9 = c$ e $y_{10} = 2c$. For estimating $\overline{Y}_d = 1.1c$, we draw a random sample of size $n_d = 1$ with inclusion probabilities $\pi_1 = \ldots = \pi_9 = 0.11$ and $\pi_{10} = 0.01$. Therefore, the unit 10 has the largest value of y and the smallest value of π. It holds that

$$\hat{\overline{Y}}_d^{dir2} = \begin{cases} c & \text{if } s = \{1\}, \ldots, \{9\}, \\ 2c & \text{if } s = \{10\}, \end{cases} \qquad \hat{\overline{Y}}_d^{dir1} = \begin{cases} \frac{c}{1.1} & \text{if } s = \{1\}, \ldots, \{9\}, \\ 20c & \text{if } s = \{10\}. \end{cases}$$

Obviously, with $\hat{\overline{Y}}_d^{dir2}$, we avoid the possibility of obtaining absurd estimates of $\overline{Y}_d = 1.1c$.

2.6 Bootstrap Resampling for Variance Estimation

In this section we present a basic bootstrap procedure for estimating the variance of an estimator.

Let us consider samples s drawn at random from a population U according to a given sampling design. Let $\hat{\theta}$ be the estimator of the population parameter θ. Särndal et al. (1992, p. 442), describe the following basic bootstrap procedure:

1. From the sample s, build an artificial population U^* mimicking U. This can be done by replicating each sample register as many times as the calibrated sample weight w_j (elevation factor).

2. Extract B independent bootstrap samples from U^* by using the same sampling design as the one used for obtaining s from U. For each bootstrap sample $s_b,\ b = 1, \dots, B$, calculate the estimator $\hat{\theta}_b^*$ in the same form as $\hat{\theta}$ was calculated for s.
3. The observed distribution of $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ imitates the distribution of $\hat{\theta}$.
4. The bootstrap estimator of the *variance* of $\hat{\theta}$ is

$$\widehat{\text{var}}_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2, \qquad \text{where} \quad \hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

5. The bootstrap estimator of the *mean squared error* of $\hat{\theta}$ is

$$mse_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta})^2.$$

6. Given two population parameters θ and φ, with respective estimators $\hat{\theta}$ and $\hat{\varphi}$, the bootstrap estimators of the *covariance* and the *crossed mean squared error* of $\hat{\theta}$ and $\hat{\varphi}$ are

$$\widehat{\text{cov}}_B(\hat{\theta}, \hat{\varphi}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)(\hat{\varphi}_b^* - \hat{\varphi}^*),$$

$$mse_B(\hat{\theta}, \hat{\varphi}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta})(\hat{\varphi}_b^* - \hat{\varphi}).$$

This bootstrap method has the disadvantage of requiring the construction of an artificial population for reproducing the original sampling design. In the case of complex sampling designs with strata and clusters, like the ones implemented in some labor force surveys, rebuilding the geographic structure of the population, within the bootstrap procedure, implies the construction of artificial populations with the same or similar cluster and strata sizes as the original one. In many cases, this is simply impossible to perform.

2.7 Jackknife Resampling for Variance Estimation

The jackknife method was developed by Quenouille (1949, 1956) as a technique for bias reduction in finite populations. Tukey (1958) suggested that jackknife could also be used for variance estimation, and Durbin (1959) applied this idea in infinite populations. The jackknife method is similar to the leave-one-out cross-validation procedure, and it can also be considered as a method for data partitioning. In what follows, the basic ideas for applying the Jackknife resampling are given. For more details, see Särndal et al. (1992, pp. 437–442).

Let s be a sample of n units drawn at random by a SRSWOR design. Let $\hat{\theta}$ be an estimator of the population parameter θ. The jackknife resampling procedure gives an estimator of $\text{var}(\hat{\theta})$. The jackknife steps are

1. Partition at random the sample s in A groups of equal size $m = n/A$.
2. For each group a, $a = 1, \ldots, A$, build the subsample $s_{(a)}$ by eliminating from s the units of group a. Based on $s_{(a)}$, calculate the estimator $\hat{\theta}_{(a)}$ of θ in the same way as $\hat{\theta}$ was calculated for s.
3. The jackknife estimator of θ is $\hat{\theta}_J = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_{(a)}$.
4. The jackknife variance estimator is $\text{var}_{J1} = \frac{A-1}{A}\sum_{a=1}^{A}\left(\hat{\theta}_{(a)} - \hat{\theta}_J\right)^2.$

   In practice, var_{J1} is used as estimator of $\text{var}(\hat{\theta})$ and $\text{var}(\hat{\theta}_J)$. An alternative estimator is

$$\text{var}_{J2} = \frac{A-1}{A}\sum_{a=1}^{A}\left(\hat{\theta}_{(a)} - \hat{\theta}\right)^2.$$

   It holds that $\text{var}_{J2} \geq \text{var}_{J1}$.
5. The jackknife bias estimator is $\text{bias}_J = (A-1)(\hat{\theta}_J - \hat{\theta})$.

Remark 2.4 Särndal et al. (1992, pp. 437–442), introduce the jackknife estimator of the variance by using the pseudovalues

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\hat{\theta}_{(a)}, \quad a = 1, \ldots, A.$$

They define the jackknife estimator of θ as bias-corrected estimator, i.e.

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a = A\hat{\theta} - (A-1)\hat{\theta}_J = \hat{\theta} - (A-1)\left(\hat{\theta}_J - \hat{\theta}\right) = \hat{\theta} - \text{bias}_J\,.$$

Further, they give the variance estimator

$$\text{var}_{JK1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2,$$

which is equal to var_{J1}, because

$$\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2 = \left\{\left[A\hat{\theta} - (A-1)\hat{\theta}_{(a)}\right] - \left[A\hat{\theta} - (A-1)\hat{\theta}_J\right]\right\}^2 = (A-1)^2\left(\hat{\theta}_{(a)} - \hat{\theta}_J\right)^2.$$

For applying the jackknife method, we have to fix a number of groups A. For having a variance estimator with a good accuracy, we could take as many groups as possible, i.e. $A = n$ and $m = 1$. On the other hand, because of the computational burden, we prefer working with few groups. The extreme cases are $A = 2$ and $m = n/2$. In practice, it is quite common to take a value of A between the extreme cases $A = n$ and $A = 2$.

Remark 2.5 If $\hat{\theta}_{(a)}$, $a = 1, \dots, A$, were uncorrelated random variables with the same expectation, then var_{J1} should be unbiased for $\mathrm{var}(\hat{\theta}_J)$. However, the $\hat{\theta}_{(a)}$'s are correlated, and therefore the unbiasedness property does not hold. The properties of the jackknife estimators of a general type parameter θ under a complex sampling design have not been studied in the literature. Under a SRS and linear target parameter, the jackknife variance estimator has, in general, a good behavior.

2.7.1 Delete-One-Cluster Jackknife for Estimators of Domain Parameters

The delete-one-cluster jackknife method (see e.g. Rao and Tausi 2004) generates jackknife samples by deleting a cluster each time. There are as many jackknife samples as clusters are in the sample. Consider the jackknife sample, $s^*_{(d_*c_*)}$, obtained by excluding the cluster c_* of the domain d_* from the sample s, and denote the corresponding domain d and cluster c subsample by $s^*_{dc(d_*c_*)}$. Let D_s be the number of domains in s, m_d be the number of clusters in s_d, $C = \sum_{d=1}^{D_s} m_d$, m_{d_*} be the number of clusters in d_*, and m_{Jd_*} be the number of clusters in the jackknife subsample $s^*_{(d_*c_*)}$. The jackknife weight of individual j, cluster c, and domain d in $s^*_{(d_*c_*)}$ is

$$w_{dcj(d_*c_*)} = w_{dcj} b_{dc(d_*c_*)}, \quad b_{dc(d_*c_*)} = \begin{cases} w_{d.}/w^*_{d.} & \text{if } d = d_*, c \neq c_*, \\ 1 & \text{if } d \neq d_*, \end{cases}$$

where $w_{d.} = \sum_{c=1}^{m_d} \sum_{j \in s_{dc}} w_{dcj}$ and $w^*_{d.} = \sum_{c=1, c\neq c_*}^{m_d} \sum_{j \in s^*_{dc(d_*c_*)}} w_{dcj}$. Note that the case $d = d_*$ and $c = c_*$ does not appear in the jackknife sample $s^*_{(d_*c_*)}$. The jackknife resampling method for estimating the variance of an estimator $\hat{\theta}$ of a population parameter θ is

1. By using the procedure described above, use sample s to draw jackknife samples $s^*_{(d_*c_*)}$, $d_* = 1, \dots, D_s$, $c_* = 1, \dots, m_{d_*}$. For every jackknife sample, calculate $\hat{\theta}^*_{(d_*c_*)}$ in the same way as $\hat{\theta}$ was calculated, but using the jackknife weights $w_{dcj(d_*c_*)}$.
2. The observed distribution of $\{\hat{\theta}^*_{(d_*c_*)} : d_* = 1, \dots, D_s, c_* = 1, \dots, m_{d_*}\}$ is expected to imitate the distribution of estimator $\hat{\theta}$.

3. The jackknife estimator of θ and bias($\hat{\theta}$) is

$$\hat{\theta}_J = \frac{1}{C}\sum_{d_*=1}^{D_s}\sum_{c_*=1}^{m_{d*}} \hat{\theta}^*_{(d_* c_*)}, \quad \text{bias}_J(\hat{\theta}) = \sum_{d_*=1}^{D_s}(m_{Jd_*}-1)\sum_{c_*=1}^{m_{d*}}\left(\hat{\theta}^*_{(d_* c_*)} - \hat{\theta}_J\right). \tag{2.10}$$

4. The design-based variance of $\hat{\theta}$ can be approximated by

$$\text{var}_J(\hat{\theta}) = \sum_{d_*=1}^{D_s}\frac{m_{Jd_*}-1}{m_{Jd_*}}\sum_{c_*=1}^{m_{d*}}\left(\hat{\theta}^*_{(d_* c_*)} - \hat{\theta}_J\right)^2. \tag{2.11}$$

2.8 R Codes for Design-Based Direct Estimators

This section presents some R codes illustrating the use of the studied estimators.

2.8.1 Horvitz–Thompson Direct Estimators of the Total and the Mean

We first read the auxiliary and sample data files and rename some variables.

```
# Auxiliary data
dataux <- read.table("Nds20.txt", header=TRUE, sep = "\t", dec = ".")
# Sort dataux by sex and area:
dataux <- dataux[order(dataux$sex, dataux$area),]
# Sample data
dat <- read.table("LFS20.txt", header=TRUE, sep = "\t", dec = ".")
# number of rows (cases) in dat:
n <- nrow(dat)
# Rename some variables
y1 <- dat$UNEMPLOYED; y2 <- dat$EMPLOYED
w <- dat$WEIGHT
area <- dat$AREA; sex <- dat$SEX
```

This section describes the following activities. For domains defined as AREA crossed by SEX, do:

A1. Estimate the totals of unemployed and employed people.
A2. Estimate the variances and the coefficients of variation.
A3. Repeat A1–A2 for means.
A4. Calculate the domain unemployment rates
A5. Estimate the variance of the unemployment rate estimator.
A6. Repeat A1–A5 for domains defined by AREA.

A1. For estimating the totals of unemployed and employed people by AREA and SEX, we apply formula (2.1), i.e.

$$\hat{Y}_d^{dir1} = \sum_{j \in s_d} w_j y_j.$$

The R code is

```
dir1.ds <- aggregate(w*data.frame(y1,y2), by=list(Area=area,Sex=sex), sum)
# Assign column names
names(dir1.ds) <- c("area", "sex", "y1tot", "y2tot")
```

A2. For estimating the variance of $\hat{Y}_d^{dir1}$, we apply the formula (2.4), i.e.

$$\widehat{\mathrm{var}}_\pi\left(\hat{Y}_d^{dir1}\right) = \sum_{j \in s_d} w_j (w_j - 1) y_j^2.$$

The R code is

```
vardir1.ds <- aggregate(w*(w-1)*data.frame(y1^2,y2^2),
                        by=list(Area=area,Sex=sex), sum)
# Assign column names
names(vardir1.ds) <- c("area", "sex", "y1var", "y2var")
```

We build a table with direct estimates of totals, variances, and coefficients of variation.

```
# Add columns y1var and y2var
dir1.ds <- cbind(dir1.ds, vardir1.ds$y1var, vardir1.ds$y2var)
# CV for y1
y1cv <- 100*sqrt(vardir1.ds$y1var)/abs(dir1.ds$y1tot)
# CV for y2
y2cv <- 100*sqrt(vardir1.ds$y2var)/abs(dir1.ds$y2tot)
# Add columns y1cv and y2cv
dir1.ds <- cbind(dir1.ds, y1cv, y2cv)
# Change column names for dir1.ds
namesds <- c("area", "sex", "y1tot", "y2tot", "y1var", "y2var", "y1cv",
             "y2cv")
names(dir1.ds) <- namesds
```

A3. We calculate the estimators of the means and their variances by using the formulas (2.1) and (2.4), i.e.

$$\hat{\overline{Y}}_d^{dir1} = N_d^{-1} \hat{Y}_d^{dir1}, \quad \widehat{\mathrm{var}}_\pi\left(\hat{\overline{Y}}_d^{dir1}\right) = N_d^{-2} \widehat{\mathrm{var}}_\pi\left(\hat{Y}_d^{dir1}\right).$$

```
# Add column with population sizes
dir1.ds <- cbind(dir1.ds, dataux$N)
# Add columns with HT estimates of means
dir1.ds <- cbind(dir1.ds, dir1.ds$y1tot/dataux$N,
                 dir1.ds$y2tot/dataux$N)
# Variance estimates of HT estimator
dir1.ds <- cbind(dir1.ds, dir1.ds$y1var/dataux$N^2, dir1.ds$y2var/
                   dataux$N^2)
# Change column names for dir1.ds
names(dir1.ds) <- c(namesds, "Nds", "y1mean", "y2mean", "y1meanvar",
                    "y2meanvar")
```

A4. For estimating the unemployment rates (in %), we employ the ratio estimator

$$\hat{R}^{dir} = \frac{\hat{Y}_{1,d}^{dir1}}{\hat{Y}_{1,d}^{dir1} + \hat{Y}_{2,d}^{dir1}} 100,$$

where $\hat{Y}_{1,d}^{dir1}$ and $\hat{Y}_{2,d}^{dir1}$ are the direct estimators of the totals of unemployed and employed people, respectively. The R code is

```
# Include estimates of unemployment rates in table dir1.ds
dirrate.ds <- 100*dir1.ds$y1tot/(dir1.ds$y1tot + dir1.ds$y2tot)
dir1.ds <- cbind(dir1.ds, rate=dirrate.ds)
```

A5. For estimating the covariances $\widehat{\text{cov}}(\hat{Y}_{1,d}^{dir1}, \hat{Y}_{2,d}^{dir1})$, we apply the corresponding formula of Remark 2.1, i.e.

$$\widehat{\text{cov}}_{\pi}(\hat{Y}_{1,d}^{dir1}, \hat{Y}_{2,d}^{dir1}) = \sum_{j \in s_d} w_j (w_j - 1) y_{1,j} y_{2,j}.$$

The R code is

```
covardir1.ds <- aggregate(w*(w-1)*data.frame(y1*y2),
                          by=list(Area=area,Sex=sex), sum)
# Column names
names(covardir1.ds) <- c("area", "sex", "covar")
```

For estimating the variance of the unemployment rate estimator, we apply the formula (3.10) of Chap. 3, i.e.

$$\widehat{\text{var}}(\hat{R}_d) = \frac{\hat{Y}_{2,d}^2}{(\hat{Y}_{1,d} + \hat{Y}_{2,d})^4} \widehat{\text{var}}(\hat{Y}_{1,d}) + \frac{\hat{Y}_{1,d}^2}{(\hat{Y}_{1,d} + \hat{Y}_{2,d})^4} \widehat{\text{var}}(\hat{Y}_{2,d})$$
$$- \frac{2\hat{Y}_{1,d}\hat{Y}_{2,d}}{(\hat{Y}_{1,d} + \hat{Y}_{2,d})^4} \widehat{\text{cov}}(\hat{Y}_{1,d}, \hat{Y}_{2,d}),$$

where $\hat{Y}_{1,d} = \hat{Y}_{1,d}^{dir1}$ and $\hat{Y}_{2,d} = \hat{Y}_{2,d}^{dir1}$. The following R code calculates $\widehat{\text{var}}(\hat{R}_d)$

```
# Summands in formula of covariance estimator
s1.ds <- dir1.ds$y2tot^2*dir1.ds$y1var/(dir1.ds$y1tot+dir1.ds$y2tot)^4
s2.ds <- dir1.ds$y1tot^2*dir1.ds$y2var/(dir1.ds$y1tot+dir1.ds$y2tot)^4
s12.ds <- 2*dir1.ds$y1tot*dir1.ds$y2tot*covardir1.ds$covar/
          (dir1.ds$y1tot + dir1.ds$y2tot)^4
# Estimates of variances and coefficients of variation
dir1.ds$vrate <- 10^4*(s1.ds+s2.ds-s12.ds)
dir1.ds$cvrate <- 100*sqrt(dir1.ds$vrate)/abs(dir1.ds$rate)
```

Table 2.1 DIR1 estimates of labor status indicators for sex=1 (left) and sex=2 (right)

area	y1tot	y2tot	y1var	y2var	rate	y1tot	y2tot	y1var	y2var	rate
1	344	5422	117,992	1,548,184	5.97	452	3637	112,068	960,992	11.05
2	206	1782	42,230	433,104	10.36	222	1674	49,062	331,572	11.71
3	0	3452	0	676,846	0.00	165	1320	27,060	220,026	11.11
4	179	3388	31,862	613,772	5.02	187	2798	34,782	500,522	6.26
5	0	2549	0	421,576	0.00	137	2065	18,632	337,506	6.22
6	381	3658	72,380	695,074	9.43	200	735	39,800	108,008	21.39
7	137	2857	18,632	555,234	4.58	0	3121	0	606,322	0.00
8	188	2863	35,156	500,160	6.16	0	2625	0	452,400	0.00
9	600	6641	135,138	1,243,378	8.29	346	3124	64,512	514,402	9.97
10	156	1655	24,180	282,474	8.61	0	1313	0	233,774	0.00

The R code to save the results is

```
output1 <- data.frame(dir1.ds[,1:6], rate=round(dirrate.ds,2))
head(output1, 10)
```

A6. This activity is an exercise.

For the ten first areas, Table 2.1 presents some of the contents of the data frame dir1.ds. The columns y1tot and y2tot contain the direct estimates, $\hat{Y}_{1,d}^{dir1}$ and $\hat{Y}_{2,d}^{dir1}$, of totals of unemployed and employed people. The columns y1var, y2var, and rate give the variance estimates $\widehat{\text{var}}_\pi\big(\hat{Y}_{1,d}^{dir1}\big)$ and $\widehat{\text{var}}_\pi\big(\hat{Y}_{2d}^{dir1}\big)$ and the unemployment rates estimations $\hat{R}_d^{dir1} = \hat{Y}_{1,d}^{dir1}/\big(\hat{Y}_{1,d}^{dir1} + \hat{Y}_{2,d}^{dir1}\big)$. The left (right) part of Table 2.1 contains the results for sex=1 (sex=2). In domains with null sample size, the dir1 estimator is not calculable, and we deliver the value of 0.

2.8.2 Hájek Direct Estimator of the Mean and the Total

This section describes the following activities. For domains defined as AREA crossed by SEX, do:

B1. Estimate the proportions of unemployed and employed people.
B2. Estimate the variances and the coefficients of variation.
B3. Repeat B1–B2 for totals.
B4. Estimate the unemployment rates.
B5. Estimate the variance of the unemployment rate estimator.
B6. Repeat B1–B5 for domains defined by AREA.

B1. By applying the formula (2.1), we calculate the estimator $\hat{Y}_d^{dir1}$ of the totals of unemployed and employed people by AREA and SEX. The R code is

```
dir <- aggregate(w*data.frame(1/w,1,y1,y2), by=list(Area=area,Sex=sex), sum)
# Column names
names(dir) <- c("area", "sex", "nds", "hatNds", "y1tot", "y2tot")
```

We calculate the direct estimates of means by AREA and SEX by applying the formula (2.7), i.e.

$$\hat{\bar{Y}}_d^{dir2} = \frac{\hat{Y}_d^{dir1}}{\hat{N}_d} = \frac{\sum_{j\in s_d} w_j y_j}{\sum_{j\in s_d} w_j}.$$

The R code is

```
dir2.ds <- data.frame(area=dir$area, sex=dir$sex, nds=dir$nds,
                      hatNds=dir$hatNds)
# Estimates of means of unemployed people
dir2.ds$y1mean <- dir$y1tot/dir$hatNds
# Estimates of means of employed people
dir2.ds$y2mean <- dir$y2tot/dir$hatNds
```

B2. For estimating the variance of $\hat{\bar{Y}}_d^{dir2}$, we apply the third formula of Remark 2.3, i.e.

$$\widehat{\text{var}}_\pi\left(\hat{\bar{Y}}_d^{dir2}\right) = \frac{1}{\hat{N}_d^2}\sum_{j\in s_d} w_j(w_j-1)(y_j-\hat{\bar{Y}}_d^{dir2})^2.$$

The R code for the numerator is

```
# Define all the necessary objects
difference1 <- difference2 <- numerator1 <- numerator2 <- ww1 <- list()
for(d in 1:nrow(dir2.ds)){
  # Create a logic vector with the indexes of the corresponding domains
  condition <- paste(dat$AREA,dat$SEX,sep="")==paste(dir2.ds$area,
                    dir2.ds$sex,sep="")[d]
  # Calculate the difference between data and mean of each domain
  difference1[[d]] <- y1[condition]-dir2.ds$y1mean[d]
  difference2[[d]] <- y2[condition]-dir2.ds$y2mean[d]
  ww1[[d]] <- w[condition]*(w[condition]-1)
  numerator1[[d]] <- ww1[[d]]*difference1[[d]]^2
  numerator2[[d]] <- ww1[[d]]*difference2[[d]]^2
}
```

The following R code calculates $\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{dir2})$ by AREA and SEX:

```
dir2.ds$y1meanvar <- sapply(numerator1, sum)/dir2.ds$hatNds^2
dir2.ds$y2meanvar <- sapply(numerator2, sum)/dir2.ds$hatNds^2
```

We include in `dir2.ds` the estimated coefficients of variation $\text{cv} = \text{cv}(\hat{\bar{Y}}_d^{dir2})$.

```
# cv of y1-mean (in %)
dir2.ds$y1cv <- 100*sqrt(dir2.ds$y1meanvar)/abs(dir2.ds$y1mean)
# cv of y2-mean (in %)
dir2.ds$y2cv <- 100*sqrt(dir2.ds$y2meanvar)/abs(dir2.ds$y2mean)
```

B3. We repeat steps 1 and 2 for estimating the totals of unemployed and employed people. We use the estimators (2.7) and the fourth formula of Remark 2.3, i.e.

$$\hat{Y}_d^{dir2} = N_d \hat{\overline{Y}}_d^{dir2}, \quad \widehat{\text{var}}_\pi(\hat{Y}_d^{dir2}) = \frac{N_d^2}{\hat{N}_d^2} \sum_{j \in s_d} w_j (w_j - 1)(y_j - \hat{\overline{Y}}_d^{dir2})^2.$$

This is done with the R code

```
dir2.ds$y1tot <- dir2.ds$y1mean*dataux$N
dir2.ds$y2tot <- dir2.ds$y2mean*dataux$N
dir2.ds$y1totvar <- dir2.ds$y1meanvar*dataux$N^2
dir2.ds$y2totvar <- dir2.ds$y2meanvar*dataux$N^2
```

B4. The unemployment rate and its direct estimator are

$$R_d = \frac{Y_{1,d}}{Y_{1,d} + Y_{2,d}}, \quad \hat{R}_d = \frac{\hat{Y}_{1,d}^{dir2}}{\hat{Y}_{1,d}^{dir2} + \hat{Y}_{2,d}^{dir2}}.$$

The following R code estimates the unemployment rates (in %):

```
dir2.ds$rate <- 100*dir2.ds$y1tot/(dir2.ds$y1tot + dir2.ds$y2tot)
```

B5. For estimating the covariances $\widehat{\text{cov}}(\hat{Y}_{1,d}^{dir2}, \hat{Y}_{2,d}^{dir2})$, we apply the last formula of Remark 2.3, i.e.

$$\widehat{\text{cov}}_\pi(\hat{Y}_{1,d}^{dir2}, \hat{Y}_{2,d}^{dir2}) = \frac{N_d^2}{\hat{N}_d^2} \sum_{j \in s_d} w_j (w_j - 1)(y_{1,j} - \hat{\overline{Y}}_{1,d}^{dir2})(y_{2,j} - \hat{\overline{Y}}_{2,d}^{dir2}).$$

The R code is

```
ww1s1s2 <- mapply(ww1, mapply(difference1, difference2, FUN="*"),
                  FUN="*")
sumcovardir2 <- sapply(ww1s1s2, sum)
covardir2.ds <- sumcovardir2*dataux$N^2/dir2.ds$hatNds^2
```

For estimating the variance of the unemployment rate estimator, we apply the formula (3.10) of Chap. 3, i.e.

$$\widehat{\text{var}}(\hat{R}_d) = \frac{\hat{Y}_{2,d}^2}{(\hat{Y}_{1,d} + \hat{Y}_{2,d})^4} \widehat{\text{var}}(\hat{Y}_{1,d}) + \frac{\hat{Y}_{1,d}^2}{(\hat{Y}_{1,d} + \hat{Y}_{2,d})^4} \widehat{\text{var}}(\hat{Y}_{2,d}) - \frac{2\hat{Y}_{1,d}\hat{Y}_{2,d}}{(\hat{Y}_{1,d} + \hat{Y}_{2,d})^4} \widehat{\text{cov}}(\hat{Y}_{1,d}, \hat{Y}_{2,d}),$$

where $\hat{Y}_{1,d} = \hat{Y}^{dir2}_{1,d}$ and $\hat{Y}_{2,d} = \hat{Y}^{dir2}_{2,d}$. The following R code calculates $\widehat{\mathrm{var}}(\hat{R}_d)$:

```
# Summands in formula of covariance estimator
s1.ds <- dir2.ds$y2tot^2*dir2.ds$y1totvar/(dir2.ds$y1tot+
         dir2.ds$y2tot)^4
s2.ds <- dir2.ds$y1tot^2*dir2.ds$y2totvar/(dir2.ds$y1tot+
         dir2.ds$y2tot)^4
s12.ds <- 2*dir2.ds$y1tot*dir2.ds$y2tot*covardir2.ds/
          (dir2.ds$y1tot+dir2.ds$y2tot)^4
# Estimates of variances and coefficients of variation
dir2.ds$vrate <- 10^4*(s1.ds+s2.ds-s12.ds)
dir2.ds$cvrate <- 100*sqrt(dir2.ds$vrate)/abs(dir2.ds$rate)
```

The R code to save the results is

```
output2 <- data.frame(dir2.ds[,1:2], round(dir2.ds[,11:14]),
                      rate=round(dir2.ds[,15],2))
head(output2, 10)
```

B6. This activity is an exercise.

For the ten first areas, Table 2.2 presents some of the contents of the data frame dir2.ds. The columns y1tot and y2tot contain the direct estimates, $\hat{Y}^{dir2}_{1,d}$ and $\hat{Y}^{dir2}_{2,d}$, of totals of unemployed and employed people. The columns y1var, y2var, and rate give the variance estimates $\widehat{\mathrm{var}}_\pi\big(\hat{Y}^{dir2}_{1,d}\big)$ and $\widehat{\mathrm{var}}_\pi\big(\hat{Y}^{dir2}_{2,d}\big)$ and the unemployment rates estimations $\hat{R}^{dir2}_d = \hat{Y}^{dir2}_{1,d}/\big(\hat{Y}^{dir2}_{1,d} + \hat{Y}^{dir2}_{2,d}\big)$. The left (right) part of Table 2.2 contains the results for sex=1 (sex=2). In domains with null sample size, the dir2 estimator is not calculable, and we deliver the value of 0. By comparing the results presented in Tables 2.1 and 2.2, we conclude that dir2 estimators of totals have, in general, smaller variances than dir1 estimators. However, they both give the same estimates of unemployment ratios.

Comparing the results presented in Tables 2.1 and 2.2 one can observe that the Hájek type estimator dir2 has lower variance estimates than the Horvitz–Thompson estimator dir1, particularly in the columns denoted as y2var.

Table 2.2 dir2 estimates of labor status indicators for sex=1 (left) and sex=2 (right)

area	y1tot	y2tot	y1var	y2var	rate	y1tot	y2tot	y1var	y2var	rate
1	347	5470	114,455	610,953	5.97	453	3648	107,441	568,195	11.05
2	209	1809	41,081	192,151	10.36	225	1694	47,076	194,190	11.71
3	0	3521	0	122,182	0.00	165	1317	25,787	142,520	11.11
4	182	3436	31,534	173,090	5.02	189	2828	34,115	217,891	6.26
5	0	2456	0	84,070	0.00	137	2069	18,176	163,088	6.22
6	391	3758	70,745	213,647	9.43	194	712	33,309	71,319	21.39
7	138	2885	18,584	142,130	4.58	0	3071	0	150,426	0.00
8	189	2878	33,612	115,024	6.16	0	2648	0	139,145	0.00
9	595	6587	124,176	450,588	8.29	348	3142	62,643	350,470	9.97
10	159	1687	24,034	144,069	8.61	0	1289	0	133,244	0.00

2.8.3 *Jackknife Estimator of Variances*

This section describes the following activities. For domains defined by AREA, do:

C1. Estimate the totals of unemployed and employed people.
C2. Calculate direct estimators of variances and coefficients of variation.
C3. Calculate jackknife estimators of variances and coefficients of variation.

We first calculate some auxiliary parameters of the sample data file LFS20.txt.

```
# Number of domains
D <- length(unique(dat$AREA))
# Domain sample sizes
nd <- tapply(rep(1,n),INDEX=list(dat$AREA),FUN=sum)
# Clusters
nCLUSTER <- unique(dat$CLUSTER)
# Number of clusters
J <- length(unique(dat$CLUSTER))
md <- vector()
# Number of clusters by domains
for (d in 1:D)
    md[d] <- length(unique(dat$CLUSTER[dat$AREA==d]))
```

C1. By applying the formula (2.1), we calculate the direct estimates, dir1, of the totals of unemployed and employed people, i.e.

```
dir.d <- aggregate(w*data.frame(y1,y2), by=list(dat$AREA), sum)
# Assign column names
names(dir.d) <- c("area", "y1tot", "y2tot")
```

C2. By applying the formula (2.3), we calculate the direct estimators of the variances, i.e.

```
vardir.d <- aggregate(w*(w-1)*data.frame(y1^2,y2^2), by=list(dat$AREA), sum)
# Assign column names
names(vardir.d) <- c("area", "y1var", "y2var")
```

The direct estimators of the coefficients of variations are

```
cvdir1 <- round(100*sqrt(vardir.d$y1var)/abs(dir.d$y1tot),2) # CV for y1
cvdir2 <- round(100*sqrt(vardir.d$y2var)/abs(dir.d$y2tot),2) # CV for y2
```

C3. For calculating the jackknife estimators of the variances, we define the auxiliary arrays

```
jackdir1 <- jackdir2 <- matrix(0, nrow=D, ncol=J)
```

We run the following jackknife loop:

```
for (j in 1:J) {
  set <- subset(dat, dat$CLUSTER!=nCLUSTER[j], na.rm=TRUE)
  # Jackknife weights
  if (length(dat$AREA[dat$CLUSTER==j])>0) {
    domjack <- unique(dat$AREA[dat$CLUSTER==j])
    jfactor <- sum(dat$WEIGHT[dat$AREA==domjack])/
      sum(set$WEIGHT[set$AREA==domjack])
    set$WEIGHT[set$AREA==domjack] <- set$WEIGHT[set$AREA==domjack]*
      jfactor
  }
  # Direct estimators
  jdir.d <- aggregate(set$WEIGHT*data.frame(set$UNEMPLOYED,
                                  set$EMPLOYED), by=list(set$AREA), sum)
  # Assign column names
  names(jdir.d) <- c("area","y1tot","y2tot")
  jackdir1[,j] <- jdir.d$y1tot
  jackdir2[,j] <- jdir.d$y2tot
}
```

We calculate the jackknife means.

```
jmeandir1 <- rowMeans(jackdir1)
jmeandir2 <- rowMeans(jackdir2)
```

We apply the formulas of Sect. 2.7.1, for calculating the jackknife variances and coefficients of variation.

```
# Number of clusters by jackknife domain
md.J <- list()
for (d in 1:D){
  md.J[[d]] <- md
  md.J[[d]][d] <- md.J[[d]][d]-1
}
factor <- Map(f="/", lapply(md.J,1,FUN="-"), md.J)
# Jackknife variances
diff.cuad.1 <- (jackdir1-jmeandir1)^2
diff.cuad.2 <- (jackdir2-jmeandir2)^2
group <- rep(1:D, md)
jvardir1 <- jvardir2 <- vector()   # declare objects for indexing
for (d in 1:D) {
  jvardir1[d] <- sum(sapply(split(diff.cuad.1[d,],group), sum)*factor[[d]])
  jvardir2[d] <- sum(sapply(split(diff.cuad.2[d,],group), sum)*factor[[d]])
}
# Jackknife coefficients of variation
jcvdir1 <- round(100*sqrt(jvardir1)/jmeandir1,2)
jcvdir2 <- round(100*sqrt(jvardir2)/jmeandir2,2)
```

The R code to save the results is

```
output3 <- data.frame(nd, y1=dir.d$y1tot, v.y1=vardir.d$y1var,
                      vJ.y1=round(jvardir1), cv.y1=cvdir1, cvJ.y1=jcvdir1,
                      y2=dir.d$y2tot, v.y2=vardir.d$y2var,
                      vJ.y2=round(jvardir2), cv.y2=cvdir2, cvJ.y2=jcvdir2)
head(output3, 10)
```

Table 2.3 presents the results for the 10 first domains (AREA). The labels y_1 and y_2 denote the dir1 direct estimates of the totals of unemployed and employed people, respectively. The direct estimates of the variances of the direct estimators of totals are denoted by $\mathrm{v}(y_1)$ and $\mathrm{v}(y_2)$. The corresponding jackknife estimates are $\mathrm{v}_J(y_1)$ and $\mathrm{v}_J(y_2)$. The direct estimates of the coefficients of variation of the direct estimators of totals are denoted by $\mathrm{c}(y_1)$ and $\mathrm{c}(y_2)$. The corresponding jackknife estimates are $\mathrm{c}_J(y_1)$ and $\mathrm{c}_J(y_2)$. The direct and jackknife estimators of variances and coefficients of variation follow the same pattern. In any case, a finer analysis cannot be done because the data used is simulated and does not come from a real survey.

Table 2.3 dir1 estimates of unemployment (left) and employment (right) totals by area

d	n_d	y_1	$\mathrm{v}(y_1)$	$\mathrm{v}_J(y_1)$	$\mathrm{c}(y_1)$	$\mathrm{c}_J(y_1)$	y_2	$\mathrm{v}(y_2)$	$\mathrm{v}_J(y_2)$	$\mathrm{c}(y_2)$	$\mathrm{c}_J(y_2)$
1	60	796	230,060	329,637	60.26	72.19	9059	2,509,176	1,365,062	17.49	12.90
2	37	428	91,292	70,084	70.59	61.84	3456	764,676	674,173	25.30	23.76
3	47	165	27,060	26,103	99.70	97.87	4772	896,872	253,103	19.85	10.54
4	55	366	66,644	46,415	70.53	58.87	6186	1,114,294	313,081	17.06	9.05
5	50	137	18,632	17,774	99.63	97.30	4614	759,082	617,055	18.88	17.03
6	43	581	112,180	307,334	57.65	95.49	4393	803,082	50,480	20.40	5.11
7	48	137	18,632	17,338	99.63	96.15	5978	1,161,556	284,300	18.03	8.92
8	48	188	35,156	33,465	99.73	97.30	5488	952,560	198,549	17.78	8.12
9	125	946	199,650	242,903	47.23	52.09	9765	1,757,780	622,368	13.58	8.08
10	41	156	24,180	22,714	99.68	96.63	2968	516,248	491,492	24.21	23.62

2.8.4 Functions for Calculating Direct Estimators

The function `dir1` calculates the Horvitz–Thompson direct estimators of the mean and the total. The R code is

```
dir1 <- function(data, w, domain, Nd) {
  if(is.vector(data)){
    last <- length(domain) + 1
    Nd.hat <- aggregate(w, by=domain, sum)[,last]
    nd <- aggregate(rep(1, length(data)), by=domain, sum)[,last]
    tot <- aggregate(w*data, by=domain, sum)
    names(tot) <- c(names(domain), "tot")
    var.tot <- aggregate(w*(w-1)*data^2, by=domain, sum)[,last]
    if(missing(Nd)){
      return(cbind(tot, var.tot, Nd.hat, nd))
    }
    else{
      mean <- tot[,last]/Nd
      var.mean <- var.tot/Nd^2
      return(cbind(tot, var.tot, mean, var.mean, Nd.hat, Nd, nd))
    }
  }
  else{
    warning("Only a numeric or integer vector must be called as data",
            call. = FALSE)
  }
}
```

The function dir2 calculates the Hájek direct estimators of the mean and the total. The R code is

```
dir2 <- function(data, w, domain, Nd) {
  if(is.vector(data)){
    last <- length(domain) + 1
    Nd.hat <- aggregate(w, by=domain, sum)[,last]
    nd <- aggregate(rep(1, length(data)), by=domain, sum)[,last]
    Sum <- aggregate(w*data, by=domain, sum)
    mean <- Sum[,last]/Nd.hat
    dom <- as.numeric(Reduce("paste0", domain))
    if(length(domain)==1){
      domain.unique <- sort(unique(dom))
    }
    else{
      domain.unique <- as.numeric(Reduce("paste0", Sum[,1:length(domain)]))
    }
    difference <- list()
    for(d in 1:length(mean)){
      condition <- dom==domain.unique[d]
      difference[[d]] <- w[condition]*(w[condition]-1)*(data[condition]-mean[d])^2
    }
    var.mean <- unlist(lapply(difference, sum))/Nd.hat^2
    if(missing(Nd)){
      return(data.frame(Sum[,-last], mean, var.mean, Nd.hat, nd))
    }
    else{
      tot <- mean*Nd
      var.tot <- var.mean*Nd^2
      return(data.frame(Sum[,-last], tot, var.tot, mean, var.mean, Nd.hat, Nd, nd))
    }
  }
  else{
    warning("Only a numeric or integer vector must be called as data",
            call. = FALSE)
  }
}
```

The following R code illustrates the use of both functions, `dir1` and `dir2`, to the data set used in this chapter. We first read the sample data files and rename some variables.

```
# Auxiliary data
```

```
dataux <- read.table("Nds20.txt", header=TRUE, sep = "\t", dec = ".")
# Sort dataux by sex and area:
dataux <- dataux[order(dataux$sex, dataux$area),]
# Sample data
dat <- read.table("LFS20.txt", header=TRUE, sep = "\t", dec = ".")
# number of rows (cases) in dat:
n <- nrow(dat)
# Rename some  variables
y1 <- dat$UNEMPLOYED
w <- dat$WEIGHT
```

Note that data and w must be a vector R object and that domains must be introduced as a list R object. The following R code calculates the direct estimator for the totals and means of unemployed people:

```
# Horvitz-Thompson direct estimator for unemployed people
direct1 <- dir1(data=y1, w=dat$WEIGHT, domain=list(area=dat$AREA,
                sex=dat$SEX), Nd=dataux$N)
head(direct1, 10)
# Hajek direct estimator for unemployed people
direct2 <- dir2(data=y1, w=dat$WEIGHT, domain=list(area=dat$AREA,
                sex=dat$SEX), Nd=dataux$N)
head(direct2, 10)
```

References

Durbin, J.: A note on the application of Quenouille's method of bias reduction to the estimation of ratios. Biometrika **46**, 477–480 (1959)

Hájek, J.: Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One". In: Godambe, V.P., Sprott, D.A. (eds.) The Foundations of Survey Sampling, pp. 236. Holt, Rinehart, and Winston, New York (1971)

Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**, 663–685 (1952)

Quenouille, M.H.: Approximation tests of correlation in time series. J. R. Stat. Soc. Ser. B. **11**, 68–84 (1949)

Quenouille, M.H.: Notes on bias in estimation. Biometrika **34**, 353–360 (1956)

Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)

Rao, J.N.K., Tausi, M.: Estimating function jackknife variance estimators under stratified multi-stage sampling. Commun. Stat. Theory Methods **33**, 2087–2095 (2004)

Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, Berlin (1992)

Tukey, J.: Bias and confidence in not quite large samples. Ann. Math. Stat. **29**, 614 (1958)