Taulukko 2. Suorien HT- ja GREG-estimaattoreiden keskimääräinen absoluuttinen suhteellinen virhe (Mean absolute relative error MARE %) ja keskimääräinen variaatiokerroin (mean coefficient of variation MCV %) pienissä, keskisuurissa ja suurissa domaineissa:
Suunniteltujen domainien tilanne

Auxiliary information	HT		GREG			
	1 None		2 Domain sizes and domain totals of EMP		3 Domain sizes and domain totals of EMP and EDUC	
Domain sample size class	MARE %	MCV %	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	11.9	5.8	7.7	6.4	6.8
Medium $34 \le n_d \le 45$	7.6	9.0	3.7	8.0	3.6	8.1
Major $46 \le n_d \le 277$	12.5	5.2	4.3	4.7	5.2	3.7

Esimerkki 1 Suunnitellut osajoukot, HT ja GREG: Suora estimointi
Kotitalousotos: Ositettu πPS (WOR- tyyppinen PPS)
Kokomuuttuja: Kotitalouden jäsenten lukumäärä
Ositteet: Seutukunnat (domains)
Estimaattorit: HT, kaava (21)

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k \qquad \hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2$$

Suora GREG, kaavat (30), (34) ja (36)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k = \sum_{k \in s_d} a_k g_{dk} y_k$$

$$\hat{V}_2\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$$

$$g_{dk} = I_{dk} + I_{dk}\left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}_d^{-1} \mathbf{x}_k$$

GREG-estimaattorin avustavat D-mallit:

$$Y_k = \beta_{0d} + \beta_{1d} \mathrm{EMP}_k + \varepsilon_k \quad \text{(sarake 2)}$$

$$Y_k = \beta_{0d} + \beta_{1d} \mathrm{EMP}_k + \beta_{2d} \mathrm{EDUC}_k + \varepsilon_k \quad \text{(sarake 3)}$$

Taulukko 3. Suoran HT-estimaattorin ja epäsuoran GREG-estimaattoreiden keskimääräinen absoluuttinen suhteellinen virhe (Mean absolute relative error MARE %) ja keskimääräinen variaatiokerroin (mean coefficient of variation MCV %) pienissä, keskisuurissa ja suurissa domaineissa:
Ei-suunniteltujen domainien tilanne

Auxiliary information	HT		GREG	
	1 None		2 Domain sizes and domain totals of EMP	
Domain sample size class	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	28.3	7.6	9.0
Medium $34 \le n_d \le 45$	7.6	20.3	3.8	8.1
Major $46 \le n_d \le 277$	12.5	9.6	4.1	5.0

Esimerkki 2 Ei-suunnitellut osajoukot
HT: Suora estimointi
GREG: Epäsuora estimointi
Kotitalousotos: πPS (WOR- tyyppinen PPS)
Kokomuuttuja: Kotitalouden jäsenten lukumäärä
Estimaattorit: HT, kaava (21)

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k \qquad \hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{n}{n-1} \sum_{k \in s} \left(a_k y_{dk} - \hat{t}_d / n\right)^2$$

GREG, kaavat (30), (41) ja (42)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k = \sum_{k \in s} a_k g_{dk} y_k$$

$$\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$$

$$g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$$

GREG-estimaattorin avustava P-malli:

$$Y_k = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k \quad \text{(sarake 2)}$$

the model calibration method. An assisting or "working" model is postulated in model-assisted estimation. In GREG estimation, the main goal is to obtain favorable design-based properties, such as small design bias. These design-based properties should hold even when the model is misspecified. If our model fits well, decreased design variance is expected for a GREG estimator. Thus, a model is used as an assisting tool in constructing the estimator, which is then modified to meet the desired design-based properties. For example, a GREG estimator for a domain total is often constructed by adding a bias correction term to the sum of fitted values calculated over the population domain. The bias correction term is obtained as a weighted sum of the sample residuals over the domain.

In this chapter, we do not address design-based techniques for nonresponse adjustment (see Chapter 8). Calibration approach to nonresponse treatment is discussed in Särndal and Lundström (2005). Additional topics that are not covered include informative sampling in the context of domain estimation (e.g., Pfeffermann and Sverchkov, 2007) and estimation for domains in the presence of outliers (see Chapter 11).

This chapter is organized as follows. Theoretical framework, terminology, and notation are introduced in Section 2. Section 3 discusses direct estimation for domains by the Horvitz–Thompson (HT) estimator, calibration and GREG estimators. In these cases, domains are often considered as strata in the sampling design. We extend in Section 4 our discussion to more general estimator types and domain structures that are often encountered in practice. GREG estimators for domains are discussed extensively; we also address composite estimation from a design-based perspective. In all these cases, auxiliary information is needed at an aggregated level. Extensions are discussed in Section 5, where a number of empirical examples based on simulation experiments are presented. In these cases, access to unit-level auxiliary data is assumed. Section 6 summarizes some properties of selected software products that can be used for design-based domain estimation.

2. Theoretical framework, terminology, and notation

2.1. Design-based inference at the population level

Let us consider a collection of random variables $(Y_1, Y_2, \ldots, Y_k, \ldots, Y_N)$ with unknown values $(y_1, y_2, \ldots, y_k, \ldots, y_N)$ of a variable of interest y in a *fixed* and *finite population* $U = \{1, 2, \ldots, k, \ldots, N\}$, where k refers to the label of population element. The fixed population is said to be generated from a *superpopulation*. For practical purposes, we are interested in one particular realized population U with $(y_1, y_2, \ldots, y_N)$, not in the more general properties of the model explaining how the population evolved. This is important especially in national statistical agencies, which attempt to describe the current state of the population of a country.

In the design-based approach, the values of the variable of interest are regarded as fixed but unknown quantities. The only source of randomness is the sampling design, and our conclusions should apply to hypothetical repeated sampling from the fixed population.

In estimation for the whole population, we are mainly interested in the total $t = \sum_{k\in U} y_k$ or mean $\bar{y} = \sum_{k\in U} y_k/N$ of the variable y. Notation $\sum_{k\in U}$ refers to summation over all population units $k \in U$. In practice, the values y_k of y are observed

in an n element sample $s \subset U$, which is drawn at random by a sampling design giving probability $p(s)$ to each sample s. The sampling design can be *complex* involving stratification and clustering and several sampling stages.

The design expectation of an estimator $\hat{t}$ of population total t is determined by the probabilities $p(s)$: let $\hat{t}(s)$ denote the value of estimator that depends on y observed in s. Then the expectation is $E(\hat{t}) = \sum_s p(s)\hat{t}(s)$. A *design unbiased* estimator has $E(\hat{t}) = t$. *Design variance* is defined as $\mathrm{Var}(\hat{t}) = \sum_s p(s)\,\big(\hat{t}(s) - E(\hat{t})\big)^2$. An estimator of design variance is denoted by $\hat{V}(\hat{t})$.

An estimator is *design consistent* if its design bias and variance tend to zero as the sample size increases. An estimator is *nearly design unbiased* if its bias ratio (bias divided by standard deviation) approaches zero with order $O(n^{-1/2})$ when the total sample size n tends to infinity (Estevao and Särndal, 2004). For a nearly design unbiased estimator, the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator's mean squared error (MSE) (Särndal, 2007, p. 99).

Variance estimators are derived in two steps. First, the theoretical design-based variance $\mathrm{Var}(\hat{t})$ (or its approximation if the theoretical design variance is intractable) is derived. Second, the derived quantity is estimated by a design unbiased or design-consistent estimator $\hat{V}(\hat{t})$.

When the estimator is a weighted sum of observations over sample, it is practical to derive expectation and variance using *inclusion probabilities*. An observation k is included in the sample with probability $\pi_k = P\{k \in s\}$. The inverse probabilities are called *design weights* $a_k = 1/\pi_k$. A useful tool is a sample membership indicator $I_k = I\{k \in s\}$ with value 1 if k is in the sample and 0 otherwise, $E(I_k) = \pi_k$. In variances, we have to consider inclusion of pairs of observations: the probability of including both k and $l (k \neq l)$ is $\pi_{kl} = E(I_k I_l)$ with inverse $a_{kl} = 1/\pi_{kl}$, and $a_{kl} = a_k$ when $k = l$. The covariance of I_k and I_l is $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$; this quantity is needed in constructing design variances and their estimators, especially for without-replacement type designs.

2.2. Basic features of design-based inference for domains

2.2.1. Planned and unplanned domain structures

In domain estimation, we are mainly interested in totals or averages of a variable of interest y over D nonoverlapping domains $U_d \subset U$, $d = 1, 2, \ldots, d, \ldots, D$, with possibly known domain sizes N_d. As an example, consider the population of a country divided into D domains by regional classification, with N_d households in domain U_d, and the aim is to estimate statistics on household poverty for the regional areas. A domain total is $t_d = t_{dy} = \sum_{k\in U_d} y_k$, where y_k refers to measurement for household k, and domain mean is $\bar{y}_d = t_d/N_d$, $d = 1, \ldots, D$.

Corresponding to population domains, the sample s is divided into subsamples s_d, $d = 1, \ldots, D$. Sampling design may be based on knowledge of domain membership of units in population. If the sampling design is stratified, domains being the strata, the domains are called *planned* (Singh et al., 1994) or*primary domains* (Hidiroglou and Patak, 2004); sometimes also *design domains* (Kish, 1980) or *identified domains* (Särndal, 2007). For planned domain structures, the population domains U_d can be regarded as separate subpopulations. Therefore, standard population estimators are applicable as such. The domain size N_d in every domain U_d is often assumed known and the sample size n_d