## 3. Direct estimators for domain estimation

The HT type estimator does not incorporate auxiliary information. GREG estimation is assisted by a model fitted at the domain level and uses auxiliary data from the domain. Calibration incorporates auxiliary data from the domain of interest or from a higher-level aggregate. All these estimators are direct because only $y$-values are taken from the domain of interest. When domain membership is known for all population elements, domain sizes $N_d$ are also known.

### 3.1. Horvitz–Thompson estimator

The basic design-based direct estimator of the domain total $t_d$ is the HT estimator, also known as the Narain-HT and the *expansion estimator*:

$$\hat{t}_{d\mathrm{HT}} = \sum_{k\in U_d} I_k y_k/\pi_k = \sum_{k\in s_d} y_k/\pi_k = \sum_{k\in s_d} a_k y_k \tag{1}$$

(Horvitz and Thompson, 1952; Narain, 1951; notation as in Section 2.1). HT estimates of domain totals are additive: they sum up to the HT estimator $\hat{t}_{\mathrm{HT}} = \sum_{k\in s} a_k y_k$ of the population total. As $E(I_k) = \pi_k$, the HT estimator is design unbiased for $t_d$. Under mild conditions on the $\pi_k$, the corresponding mean estimator $\hat{t}_{d\mathrm{HT}}/N_d$ is also design consistent (Isaki and Fuller, 1982). The estimator $\hat{t}_{d\mathrm{HT}}$ has design variance

$$\begin{aligned}\mathrm{Var}\big(\hat{t}_{d\mathrm{HT}}\big) &= E\left(\sum_{k\in U_d} \frac{I_k-\pi_k}{\pi_k} y_k\right)^2 = \sum_{k\in U_d}\sum_{l\in U_d} E(I_k-\pi_k)(I_l-\pi_l)\,\frac{y_k}{\pi_k}\,\frac{y_l}{\pi_l} \\ &= \sum_{k\in U_d}\sum_{l\in U_d} (\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\,\frac{y_l}{\pi_l} = \sum_{k\in U_d}\sum_{l\in U_d} (a_k a_l/a_{kl}-1)y_k y_l. \end{aligned} \tag{2}$$

From $a_{kl}E(I_k I_l) = 1$, we see that an unbiased estimator for the design variance is

$$\hat{V}\big(\hat{t}_{d\mathrm{HT}}\big) = \sum_{k\in U_d}\sum_{l\in U_d} a_{kl} I_k I_l (a_k a_l/a_{kl}-1) y_k y_l = \sum_{k\in s_d}\sum_{l\in s_d} (a_k a_l - a_{kl}) y_k y_l. \tag{3}$$

An alternative Sen–Yates–Grundy formula for fixed sample size designs is (Sen, 1953; Yates, 1953):

$$\begin{aligned}\hat{V}\big(\hat{t}_{d\mathrm{HT}}\big) &= -\sum_{k\in s_d}\sum_{l<k;\,l\in s_d} a_{kl}(\pi_{kl}-\pi_k\pi_l)(a_k y_k - a_l y_l)^2 \\ &= \sum_{k\in s_d}\sum_{l<k;\,l\in s_d} (a_{kl}/a_k a_l - 1)(a_k y_k - a_l y_l)^2. \end{aligned}$$

These variance estimators are impractical because they contain second-order inclusion probabilities $\pi_{kl}$ whose computation is often laborious for practical purposes. Hájek (1964) and Berger (2004, 2005b) proposed approximations to $\pi_{kl}$. Särndal (1996) developed efficient strategies with simple variance estimators under fixed sample size probability proportional-to-size ($\pi$PS) schemes, including a combination of Poisson sampling or stratified simple random sampling without replacement (SRSWOR) with

GREG estimation. Berger and Skinner (2005) proposed a jackknife variance estimator and Kott (2006a) introduced a delete-a-group jackknife variance estimator for $\pi$PS designs. The SAS procedure SURVEYSELECT is able to compute $\pi_{kl}$ under certain unequal probability without-replacement sampling designs. Some software products can incorporate the $\pi_{kl}$ into variance estimation procedures; an example is the SUDAAN software. The SAS macro CLAN includes the Sen–Yates–Grundy formula. Such estimators are discussed in Chapter 2.

Many $\pi$PS designs allow using of Hájek approximation (Berger, 2004, 2005b; Hájek, 1964) of second-order inclusion probabilities by $\pi_{kl} \approx \pi_k \pi_l \left[1 - (1-\pi_k)(1-\pi_l) m_d^{-1}\right]$ for $k \neq l$, where $m_d = \sum_{i \in U_d} \pi_i (1-\pi_i)$. The approximation is used in a simple variance estimator $\hat{V}\left(\hat{t}_{d\mathrm{HT}}\right) = \sum_{k \in s_d} c_k e_k^2$, where $c_i = n_d (n_d - 1)^{-1} (1 - \pi_i)$ and $e_k = a_k y_k - \left(\sum_{i \in s_d} c_i\right)^{-1} \sum_{i \in s_d} c_i a_i y_i$.

For unequal probability sampling designs, the variance of the ordinary HT estimator has been approximated under a with-replacement (WR) assumption, leading to Hansen–Hurwitz (1943) type variance estimator (Lehtonen and Pahkinen, 2004, p. 228, and SAS procedure SURVEYMEANS) given by

$$\hat{V}_A(\hat{t}_{d\mathrm{HT}}) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} (n_d a_k y_k - \hat{t}_{d\mathrm{HT}})^2. \tag{4}$$

For unplanned domains, the variance estimator for HT should account for random domain sizes. An approximate variance estimator applied, for example, in SAS procedure SURVEYMEANS contains extended domain variables $y_{dk}$:

$$\hat{V}_U(\hat{t}_{d\mathrm{HT}}) = \frac{n}{n-1} \sum_{k \in s} (a_k y_{dk} - \hat{t}_d / n)^2, \tag{5}$$

where $n$ is the total sample size. Under SRSWOR, an alternative to (5) is

$$\hat{V}_{\mathrm{srswor}}(\hat{t}_{d\mathrm{HT}}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \hat{s}_{dy}^2 \left(1 + \frac{q_d}{\mathrm{c.v}_{dy}^2}\right),$$

where $p_d = n_{s_d}/n$, $q_d = 1 - p_d$, variance estimator is, $\hat{s}_{dy}^2 = \sum_{k \in s_d} (y_k - \bar{y}_d)^2 / (n_{s_d} - 1)$, and estimated coefficient of variation is $\mathrm{c.v}_{dy} = \hat{s}_{dy}/\bar{y}_d$ for $\bar{y}_d = \sum_{k \in s_d} y_k / n_{s_d}$.

The HT estimator can be regarded as a model-dependent estimator under a model $Y_k = \beta \pi_k + \pi_k \varepsilon_k$ (Zheng and Little, 2003). HT is nearly optimal estimator among weighted sums of $Y$ values when $Y$ depends on scalar $x$ as $E(Y_k) = \beta x_k$, the variance of errors is proportional to $x_k^2$, and the sampling design assigns $\pi_k$ proportional to $x_k$. On the other hand, HT is very inefficient when the intercept of the model is far from zero. Disastrous results are possible in HT estimation, as the famous example of Basu (1971) shows (e.g., citation in Little, 2004).

If the domain size $N_d$ is known, we expect better results with a "Hájek" type direct estimator $\hat{t}_{dH(N)} = N_d \hat{\bar{y}}_d$ (e.g., Hidiroglou and Patak, 2004; Särndal et al., 1992, p. 391) derived from the domain mean $\hat{\bar{y}}_d = \sum_{k \in s_d} a_k y_k / \hat{N}_d$ with $\hat{N}_d = \sum_{k \in s_d} a_k$. This is a special case of ratio estimation (Section 4.3.1). The variance of $\hat{t}_{dH(N)}$ is estimated by

$$\hat{V}(\hat{t}_{dH(N)}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl})(y_k - \hat{\bar{y}}_d)(y_l - \hat{\bar{y}}_d). \tag{6}$$