Survey Sampling Theory

1 The population total, population mean, and population variance of a population characteristic

Let $n, N \in \mathbb{N}$, with $n \leq N$. Let $\mathcal{U} = \{1, 2, \ldots, N\}$, which represents the finite population, or universe, of N elements.

Definition 1.1 *A* population characteristic *is an $\mathbb{R}$-valued function $y : \mathcal{U} \longrightarrow \mathbb{R}$ defined on the population $\mathcal{U}$. We denote the value of y evaluated at $i \in \mathcal{U}$ by y_i. The population total, denoted by t, of y is defined:*

$$t \;:=\; \sum_{i=1}^{N} y_i \in \mathbb{R}\,.$$

The population mean*, denoted by $\overline{y}$, of y is defined by:*

$$\overline{y} \;:=\; \frac{1}{N}\sum_{i=1}^{N} y_i \in \mathbb{R}\,.$$

The population variance*, denoted by S^2, of y is defined by:*

$$S^2 \;:=\; \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2 \;=\; \frac{1}{N-1}\left\{\left(\sum_{i=1}^{N} y_i^2\right) - N\cdot\overline{y}^2\right\} \in \mathbb{R}\,.$$

In survey sampling, we seek to estimate population total t and population mean $\overline{y}$ of a population characteristic $y : \mathcal{U} \longrightarrow \mathbb{R}$ by making observations of values of y on only a (usually proper) subset of $\mathcal{U}$, and extrapolate from these observations. The subset on which observations of values of y are made is called a *sample*.

2 Simple Random Sampling (SRS)

Definition 2.1 *Let $\mathcal{U}$ be a nonempty finite set, $N := \#(\mathcal{U}) \in \mathbb{N}$, and let $n \in \{1, 2\ldots, N\}$ be given. We define the probability space $\Omega_{\mathrm{SRS}}(\mathcal{U}, n)$ as follows: Let $\Omega(\mathcal{U}, n)$ be the set of all subsets of $\mathcal{U}$ with n elements, i.e.*

$$\Omega(\mathcal{U}, n) \;:=\; \{\,\omega \subset \mathcal{U} \mid \#(\omega) = n\,\}\,.$$

Note that $\#(\Omega(\mathcal{U}, n)) = \left(\begin{array}{c} N \\ n \end{array}\right)$. Let $\mathcal{P}(\Omega(\mathcal{U}, n))$ be the power set of $\Omega(\mathcal{U}, n)$. Define $\mu : \Omega \longrightarrow \mathbb{R}$ to be the "uniform" probability measure on the (finite) σ-algebra $\mathcal{P}(\Omega(\mathcal{U}, n))$ determined by:

$$\mu(\omega) \;=\; \frac{1}{\left(\begin{array}{c} N \\ n \end{array}\right)} \;=\; \frac{n!(N-n)!}{N!}\,, \quad \text{for each } \omega \in \Omega(\mathcal{U}, n).$$

Then, $\Omega_{\mathrm{SRS}}(\mathcal{U}, n)$ is defined to be the probability space $(\,\Omega(\mathcal{U}, n)\,,\,\mathcal{P}(\Omega(\mathcal{U}, n))\,,\,\mu\,)$.

Definition 2.2 *The* simple-random-sampling sample total $\widehat{t}_{\mathrm{SRS}}$ *of the population characteristic y is, by definition, the random variable $\widehat{t}_{\mathrm{SRS}} : \Omega_{\mathrm{SRS}}(\mathcal{U}, n) \longrightarrow \mathbb{R}$ defined by*

$$\widehat{t}_{\mathrm{SRS}}(\omega) \;:=\; \frac{N}{n}\sum_{i\in\omega} y_i\,, \quad \text{for each } \omega \in \Omega.$$

The simple-random-sampling sample mean $\widehat{\overline{y}}_{\mathrm{SRS}}$ *of the population characteristic y is, by definition, the random variable $\widehat{\overline{y}}_{\mathrm{SRS}} : \Omega_{\mathrm{SRS}}(\mathcal{U}, n) \longrightarrow \mathbb{R}$ defined by*

$$\widehat{\overline{y}}_{\mathrm{SRS}}(\omega) \;:=\; \frac{1}{n}\sum_{i\in\omega} y_i\,, \quad \text{for each } \omega \in \Omega.$$

The simple-random-sampling sample variance $\widehat{s^2}_{\mathrm{SRS}}$ *of the population characteristic y is, by definition, the random variable $\widehat{s^2}_{\mathrm{SRS}} : \Omega_{\mathrm{SRS}}(\mathcal{U}, n) \longrightarrow \mathbb{R}$ defined by*

$$\widehat{s^2}_{\mathrm{SRS}}(\omega) \;:=\; \frac{1}{n-1}\sum_{i\in\omega}\left(y_i - \widehat{\overline{y}}_{\mathrm{SRS}}(\omega)\right)^2\,, \quad \text{for each } \omega \in \Omega.$$

Proposition 2.3

1. *$\widehat{\overline{y}}_{\text{SRS}}$ is an unbiased estimator of the population mean $\overline{y}$, and* $\text{Var}\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right] = \left(1-\dfrac{n}{N}\right)\dfrac{S^2}{n}$.

2. *$\widehat{t}_{\text{SRS}}$ is an unbiased estimator of the population total t, and* $\text{Var}\left[\,\widehat{t}_{\text{SRS}}\,\right] = N^2\left(1-\dfrac{n}{N}\right)\dfrac{S^2}{n}$.

3. *$\widehat{s^2}_{\text{SRS}}$ is an unbiased estimator of the population variance S^2.*

4. $\widehat{\text{Var}}\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right] := \left(1-\dfrac{n}{N}\right)\dfrac{\widehat{s^2}_{\text{SRS}}}{n}$ *is an unbiased estimator of* $\text{Var}\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right]$.

5. $\widehat{\text{Var}}\left[\,\widehat{t}_{\text{SRS}}\,\right] := N^2\left(1-\dfrac{n}{N}\right)\dfrac{\widehat{s^2}_{\text{SRS}}}{n}$ *is an unbiased estimator of* $\text{Var}\left[\,\widehat{t}_{\text{SRS}}\,\right]$.

A quote from Lohr [3], p.37: *Hájek [2] proves a central limit theorem for simple random sampling without replacement. In practical terms, Hájek's theorem says that if certain technical conditions hold, and if n, N, and $N-n$ are all "sufficiently large," then the sampling distribution of*

$$\frac{\widehat{\overline{y}}_{\text{SRS}} - \overline{y}}{\sqrt{\left(1-\dfrac{n}{N}\right)\dfrac{S^2}{n}}}$$

is "approximately" normal (Gaussian) with mean 0 and variance 1.

Corollary 2.4 (to Hájek's theorem) *For a simple random sampling procedure, an approximate $(1-\alpha)$-confidence interval, $0<\alpha<1$, for the population mean $\overline{y}$ is given by:*

$$\widehat{\overline{y}}_{\text{SRS}} \,\pm\, z_{\alpha/2}\cdot\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$$

For sufficiently large samples, the above approximate confidence interval can itself be estimated from observations by:

$$\widehat{\overline{y}}_{\text{SRS}} \pm \text{SE}\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right] \;=\; \widehat{\overline{y}}_{\text{SRS}} \pm \sqrt{\left(1-\frac{n}{N}\right)\frac{\widehat{s^2}_{\text{SRS}}}{n}}$$

where

$$\text{SE}\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right] \;:=\; \sqrt{\widehat{\text{Var}}\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right]} \;=\; \sqrt{\left(1-\frac{n}{N}\right)\frac{\widehat{s^2}_{\text{SRS}}}{n}}$$

In order to prove Proposition 2.3, we introduce some auxiliary random variables:

Definition 2.5 *Let $n, N \in \mathbb{N}$, with $n<N$, $\mathcal{U} := \{\,1,2,\ldots,N\,\}$, and $\Omega := \{\,\omega\subset\mathcal{U}\mid \#(\omega)=n\,\}$. For each $i\in\mathcal{U} = \{\,1,2,\ldots,N\,\}$, we define the random variable $Z_i:\Omega\longrightarrow\{0,1\}$ as follows:*

$$Z_i(\omega) \;=\; \begin{cases} 1, & \text{if } i\in\omega\,, \\ 0, & \text{if } i\notin\omega \end{cases}.$$

Immediate observations:

- $\widehat{t}_{\text{SRS}} = \dfrac{N}{n}\sum_{i=1}^{N} Z_i\,y_i,$, as random variables on (Ω, P), i.e.

$$\widehat{t}_{\text{SRS}}(\omega) \;=\; \frac{N}{n}\sum_{i=1}^{N} Z_i(\omega)\,y_i, \quad \text{for each } \omega\in\Omega.$$

- $\widehat{\overline{y}}_{\text{SRS}} = \dfrac{1}{n}\sum_{i=1}^{N} Z_i\, y_i$, as random variables on (Ω, P), i.e.

$$\widehat{\overline{y}}_{\text{SRS}}(\omega) \;=\; \frac{1}{n}\sum_{i=1}^{N} Z_i(\omega)\, y_i, \quad \text{for each } \;\omega\in\Omega.$$

- $E[\,Z_i\,] = \dfrac{n}{N}$. Indeed,

$$E[\,Z_i\,] = 1\cdot P(Z_i=1) + 0\cdot P(Z_i=0) = P(Z_i=1) = \frac{\text{number of samples containing } i}{\text{number of all possible samples}} = \frac{\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{n}{N}$$

- $Z_i^2 = Z_i$, since $\text{range}(Z_i) = \{\,0,1\,\}$. Consequently,

$$E[\,Z_i^2\,] \;=\; E[\,Z_i\,] \;=\; \frac{n}{N}\,.$$

- $\text{Var}[\,Z_i\,] = \dfrac{n}{N}\left(1-\dfrac{n}{N}\right)$. Indeed,

$$\begin{aligned}
\text{Var}[\,Z_i\,] \;&:=\; E\left[\,(Z_i - E[\,Z_i\,])^2\,\right] \;=\; E\left[Z_i^2\right] - (E[\,Z_i\,])^2 \\
&=\; E[\,Z_i\,] - \left(\frac{n}{N}\right)^2 \;=\; \frac{n}{N} - \left(\frac{n}{N}\right)^2 \\
&=\; \frac{n}{N}\left(1-\frac{n}{N}\right).
\end{aligned}$$

- For $i \neq j$, we have $E[\,Z_i\cdot Z_j\,] = \left(\dfrac{n-1}{N-1}\right)\cdot\left(\dfrac{n}{N}\right)$. Indeed,

$$\begin{aligned}
E[\,Z_i\cdot Z_j\,] \;&=\; 1\cdot P(Z_i=1 \;\text{ and }\; Z_j=1) + 0\cdot P(Z_i=0 \;\text{ or }\; Z_j=0) \\
&=\; P(Z_i=1 \;\text{ and }\; Z_j=1) \;=\; P(Z_j=1|Z_i=1)\cdot P(Z_i=1) \\
&=\; \left(\frac{n-1}{N-1}\right)\cdot\left(\frac{n}{N}\right)
\end{aligned}$$

- For $i \neq j$, we have $\text{Cov}\,(Z_i, Z_j) = -\dfrac{1}{N-1}\left(1-\dfrac{n}{N}\right)\left(\dfrac{n}{N}\right) \leq 0$. Indeed,

$$\begin{aligned}
\text{Cov}\,(Z_i, Z_j) \;&:=\; E\,[\,(Z_i - E[\,Z_i\,])\cdot(Z_j - E[\,Z_j\,])\,] \;=\; E\,[\,Z_i\,Z_j\,] - E[\,Z_i\,]\cdot E[\,Z_j\,] \\
&=\; \left(\frac{n-1}{N-1}\right)\cdot\left(\frac{n}{N}\right) - \left(\frac{n}{N}\right)^2 \;=\; \frac{n}{N}\left(\frac{nN-N-nN+n}{N(N-1)}\right) \\
&=\; -\frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)
\end{aligned}$$

Proof of Proposition 2.3

1.

$$E\left[\,\widehat{\overline{y}}_{\text{SRS}}\,\right] \;=\; E\left[\,\frac{1}{n}\sum_{i=1}^{N} Z_i\,y_i\,\right] \;=\; \frac{1}{n}\sum_{i=1}^{N} E[\,Z_i\,]\cdot y_i \;=\; \frac{1}{n}\sum_{i=1}^{N}\left(\frac{n}{N}\right)\cdot y_i \;=\; \frac{1}{N}\sum_{i=1}^{N} y_i \;=:\; \overline{y}.$$

$$\begin{aligned}
\mathrm{Var}\left[\,\widehat{\overline{y}}_{\mathrm{SRS}}\,\right] &= \mathrm{Var}\left[\,\frac{1}{n}\sum_{i=1}^{N} Z_i\,y_i\,\right] \;=\; \frac{1}{n^2}\,\mathrm{Var}\left[\,\sum_{i=1}^{N} Z_i\,y_i\,\right] \;=\; \frac{1}{n^2}\,\mathrm{Cov}\left[\,\sum_{i=1}^{N} Z_i\,y_i\,,\,\sum_{j=1}^{N} Z_j\,y_j\,\right] \\
&= \frac{1}{n^2}\left\{\,\sum_{i=1}^{N} y_i^2\,\mathrm{Var}(Z_i) + \sum_{i=1}^{N}\sum_{i\neq j=1}^{N} y_i y_j\,\mathrm{Cov}(Z_i,Z_j)\,\right\} \\
&= \frac{1}{n^2}\left\{\,\sum_{i=1}^{N} y_i^2\,\frac{n}{N}\left(1-\frac{n}{N}\right) - \sum_{i=1}^{N}\sum_{i\neq j=1}^{N} y_i y_j\,\frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\,\right\} \\
&= \frac{1}{n^2}\,\frac{n}{N}\left(1-\frac{n}{N}\right)\left\{\,\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{i\neq j=1}^{N} y_i y_j\,\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left\{\,(N-1)\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{i\neq j=1}^{N} y_i y_j\,\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left\{\,(N-1)\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j + \sum_{i=1}^{N} y_i^2\,\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left\{\,N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)\left(\sum_{j=1}^{N} y_j\right)\,\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left\{\,\sum_{i=1}^{N} y_i^2 - N\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2\,\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left\{\,\sum_{i=1}^{N} y_i^2 - N\cdot\overline{y}^2\,\right\} \\
&= \left(1-\frac{n}{N}\right)\frac{S^2}{n}
\end{aligned}$$

2.

$$E\left[\,\widehat{t}_{\mathrm{SRS}}\,\right] \;=\; E\left[\,N\cdot\widehat{\overline{y}}_{\mathrm{SRS}}\,\right] \;=\; N\cdot E\left[\,\widehat{\overline{y}}_{\mathrm{SRS}}\,\right] \;=\; N\cdot\overline{y} \;=\; N\cdot\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) \;=\; \sum_{i=1}^{N} y_i \;=:\; t.$$

$$\mathrm{Var}\left[\,\widehat{t}_{\mathrm{SRS}}\,\right] \;=\; \mathrm{Var}\left[\,N\cdot\widehat{\overline{y}}_{\mathrm{SRS}}\,\right] \;=\; N^2\cdot\mathrm{Var}\left[\,\widehat{\overline{y}}_{\mathrm{SRS}}\,\right] \;=\; N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

3.

$$
\begin{aligned}
E\left[\,\widehat{s^2}_{\mathrm{SRS}}\,\right] &= E\left[\frac{1}{n-1}\sum_{i\in\omega}\left(y_i-\widehat{\overline{y}}_{\mathrm{SRS}}\right)^2\right] \;=\; \frac{1}{n-1}\,E\left[\sum_{i\in\omega}\left((y_i-\overline{y})-(\,\widehat{\overline{y}}_{\mathrm{SRS}}-\overline{y})\right)^2\right] \\
&= \frac{1}{n-1}\,E\left[\left(\sum_{i\in\omega}(y_i-\overline{y})^2\right)-n\left(\widehat{\overline{y}}_{\mathrm{SRS}}-\overline{y}\right)^2\right] \\
&= \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{N}Z_i(y_i-\overline{y})^2\right]-n\,\mathrm{Var}\left[\,\widehat{\overline{y}}_{\mathrm{SRS}}\,\right]\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{N}E[\,Z_i\,]\,(y_i-\overline{y})^2-n\left(1-\frac{n}{N}\right)\frac{S^2}{n}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{N}\frac{n}{N}(y_i-\overline{y})^2-\left(1-\frac{n}{N}\right)S^2\right\} \\
&= \frac{1}{n-1}\left\{\frac{n(N-1)}{N}\,\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y})^2-\left(1-\frac{n}{N}\right)S^2\right\} \\
&= \frac{1}{n-1}\left\{\frac{n(N-1)}{N}-\left(1-\frac{n}{N}\right)\right\}S^2 \\
&= \frac{1}{n-1}\left\{\frac{nN-n-N+n}{N}\right\}S^2 \;=\; S^2
\end{aligned}
$$

4. Immediate from preceding statements.
5. Immediate from preceding statements. □

3 Stratified Simple Random Sampling

Let $\mathcal{U}=\{\,1,2,\ldots,N\,\}$ be the population, as before. Let

$$\mathcal{U} \;=\; \bigsqcup_{h=1}^{H}\mathcal{U}_h$$

be a partition of $\mathcal{U}$. Such a partition is called a *stratification* of the population $\mathcal{U}$. Each of $\mathcal{U}_1,\mathcal{U}_2,\ldots,\mathcal{U}_H$ is called a *stratum*. Let $N_h := \#(\mathcal{U}_h)$, for $h=1,2,\ldots,H$. Note that $N_1+N_2+\cdots+N_H=N$.

In *stratified simple random sampling*, an SRS is taken within each stratum $\mathcal{U}_h$, $h=1,2,\ldots,H$. Let n_h, $h=1,2,\ldots,H$, be the number elements in the simple random sample taken in the stratum $\mathcal{U}_h$. In other words, a stratified simple random sample ω of the stratified population $\mathcal{U}=\bigsqcup_{h=1}^{H}\mathcal{U}_h$ has the form:

$$\omega \;=\; \bigsqcup_{h=1}^{H}\omega_h\,,\quad\text{where }\;\omega_h\in\Omega_{\mathrm{SRS}}(\mathcal{U}_h,n_h)\,,\quad\text{for each }\;h=1,2,\ldots,h\,.$$

Note that $n_1+n_2+\cdots+n_H =: n = \#(\omega)$.

We now give unbiased estimators, and their variances, of the population total and population mean of a population characteristic under stratified simple random sampling. Let $y:\mathcal{U}\longrightarrow\mathbb{R}$ be a population characteristic. Define:

$$\widehat{t}_{\mathrm{Str}} \;:=\; \sum_{h=1}^{H}N_h\cdot\widehat{\overline{y}}_{h,\mathrm{SRS}}$$

$$\widehat{\overline{y}}_{\mathrm{Str}} \;:=\; \frac{1}{N}\cdot\widehat{t}_{\mathrm{Str}} \;=\; \sum_{h=1}^{H}\frac{N_h}{N}\cdot\widehat{\overline{y}}_{h,\mathrm{SRS}}$$

Here,

$$\widehat{\overline{y}}_{h,\mathrm{SRS}} \;:\; \Omega_{\mathrm{SRS}}(\mathcal{U}_h, n_h) \longrightarrow \mathbb{R} \;:\; \omega_h \;\longmapsto\; \frac{1}{n_h}\sum_{i\in\omega_h} y_i$$

is the SRS estimator of

$$\overline{y}_h \;:=\; \overline{y|_{\mathcal{U}_h}} \;=\; \frac{1}{N_h}\sum_{i\in\mathcal{U}_h} y_i \in \mathbb{R},$$

the "stratum mean" of the "stratum characteristic" $y|_{\mathcal{U}_h} : \mathcal{U}_h \longrightarrow \mathbb{R}$, the restriction of the population characteristic $y : \mathcal{U} \longrightarrow \mathbb{R}$ to the stratum $\mathcal{U}_h$. Then,

$$E\left[\,\widehat{t}_{\mathrm{Str}}\,\right] \;=\; t \;:=\; \sum_{i=1}^{N} y_i, \quad \text{and} \quad E\left[\,\widehat{\overline{y}}_{\mathrm{Str}}\,\right] \;=\; \overline{y} \;:=\; \frac{1}{N}\sum_{i=1}^{N} y_i.$$

In other words, $\widehat{t}_{\mathrm{Str}}$ and $\widehat{\overline{y}}_{\mathrm{Str}}$ are unbiased estimators of the population total t and population mean $\overline{y}$ of the population characteristic $y : \mathcal{U} \longrightarrow \mathbb{R}$, respectively. Indeed,

$$\begin{aligned}
E\left[\,\widehat{t}_{\mathrm{Str}}\,\right] &= E\left[\sum_{h=1}^{H} N_h \cdot \widehat{\overline{y}}_{h,\mathrm{SRS}}\right] \;=\; \sum_{h=1}^{H} N_h\, E\left[\,\widehat{\overline{y}}_{h,\mathrm{SRS}}\,\right] \;=\; \sum_{h=1}^{H} N_h\,\overline{y}_h \\
&= \sum_{h=1}^{H} N_h \left(\frac{1}{N_h}\sum_{i\in\mathcal{U}_h} y_i\right) \;=\; \sum_{h=1}^{H}\left(\sum_{i\in\mathcal{U}_h} y_i\right) \;=\; \sum_{i=1}^{N} y_i \;=:\; t\,.
\end{aligned}$$

And,

$$E\left[\,\widehat{\overline{y}}_{\mathrm{Str}}\,\right] \;=\; E\left[\,\frac{1}{N}\cdot\widehat{t}_{\mathrm{Str}}\,\right] \;=\; \frac{1}{N}\,E\left[\,\widehat{t}_{\mathrm{Str}}\,\right] \;=\; \frac{1}{N}\sum_{i=1}^{N} y_i \;=:\; \overline{y}\,.$$

Furthermore,

$$\mathrm{Var}\left[\,\widehat{t}_{\mathrm{Str}}\,\right] \;=\; \mathrm{Var}\left[\sum_{h=1}^{H} N_h \cdot \widehat{\overline{y}}_{h,\mathrm{SRS}}\right] \;=\; \sum_{h=1}^{H} N_h^2 \cdot \mathrm{Var}\left[\,\widehat{\overline{y}}_{h,\mathrm{SRS}}\,\right] \;=\; \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}.$$

$$\mathrm{Var}\left[\,\widehat{\overline{y}}_{\mathrm{Str}}\,\right] \;=\; \mathrm{Var}\left[\,\frac{1}{N}\cdot\widehat{t}_{\mathrm{Str}}\,\right] \;=\; \frac{1}{N^2}\cdot\mathrm{Var}\left[\,\widehat{t}_{\mathrm{Str}}\,\right] \;=\; \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}.$$

Comparing variances of SRS and stratified simple random sampling with proportional allocation via ANOVA (analysis of variance):

By definition, in stratified simple random sampling with proportional allocation, the stratum sample size n_h, for each $h = 1, 2, \ldots, H$, is chosen such that $n_h/N_h = n/N$. Consequently,

$$\begin{aligned}
\mathrm{Var}\left[\,\widehat{t}_{\mathrm{PropStr}}\,\right] &= \sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \;=\; \frac{N}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} N_h\, S_h^2 \\
&= \frac{N}{n}\left(1 - \frac{n}{N}\right)\left\{\,\sum_{h=1}^{H}(N_h - 1)\, S_h^2 + \sum_{h=1}^{H} S_h^2\,\right\} \\
&= \frac{N}{n}\left(1 - \frac{n}{N}\right)\left\{\,\mathrm{SSW} + \sum_{h=1}^{H} S_h^2\,\right\},
\end{aligned}$$

where

$$\mathrm{SSW} \;:=\; \sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\left(y_i - \overline{y}_{\mathcal{U}_h}\right)^2 \;=\; \sum_{h=1}^{H}(N_h - 1)\, S_h^2.$$

is called the *inter-strata squared deviation* (or *within-strata squared deviation*), and

$$S_h^2 \;:=\; \frac{1}{N_h - 1}\sum_{i\in\mathcal{U}_h}\left(y_i - \overline{y}_{\mathcal{U}_h}\right)^2$$

is the stratum variance of the population characteristic $y:\mathcal{U}\longrightarrow\mathbb{R}$ over the stratum $\mathcal{U}_h$. The following relation between $\mathrm{Var}\big[\,\widehat{t}_{\mathrm{SRS}}\,\big]$ and $\mathrm{Var}\big[\,\widehat{t}_{\mathrm{PropStr}}\,\big]$ always holds (see [3], p.106):

$$\mathrm{Var}\big[\,\widehat{t}_{\mathrm{SRS}}\,\big] \;=\; \mathrm{Var}\big[\,\widehat{t}_{\mathrm{PropStr}}\,\big] + \left(1-\frac{n}{N}\right)\frac{N}{n}\frac{N}{N-1}\left\{\mathrm{SSB} - \sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)S_h^2\right\},$$

where

$$\mathrm{SSB} \;:=\; \sum_{h=1}^{H} N_h\left(\overline{y}_{\mathcal{U}_h} - \overline{y}_{\mathcal{U}}\right)^2 \;=\; \sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\left(\overline{y}_{\mathcal{U}_h} - \overline{y}_{\mathcal{U}}\right)^2$$

is the *inter-strata squared deviation* (or *between-strata squared deviation*). It is also an easily established fact that the sum of the inter-strata squared deviation SSB and the intra-strata squared deviation SSW is always the total population squared deviation SSTO:

$$\mathrm{SSTO} \;:=\; \sum_{i=1}^{N}\left(y_i - \overline{y}_{\mathcal{U}}\right)^2 \;=\; \sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\left(y_i - \overline{y}_{\mathcal{U}}\right)^2 .$$

Most importantly, we see from above that, for stratified simple random sampling with proportional allocation, the following implication holds:

$$\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)S_h^2 \,\leq\, \mathrm{SSB} \quad\Longrightarrow\quad \mathrm{Var}\big[\,\widehat{t}_{\mathrm{PropStr}}\,\big] \leq \mathrm{Var}\big[\,\widehat{t}_{\mathrm{SRS}}\,\big] .$$

In heuristic terms, in proportional-allocation stratification for which each stratum is relatively homogeneous and the strata are relatively dissimilar to each other (intra-strata variation being smaller than inter-strata variation), then the unbiased estimator for the population total from the proportional-allocation stratified simple random sampling is more precise than that from SRS.

4 Two-stage Cluster Sampling

The universe $\mathcal{U} = \bigsqcup_{i=1}^{N}\mathcal{C}_i$ of observation units is partitioned into N *clusters* (or *primary sampling units*, psu's) $\mathcal{C}_i$. In two-stage cluster sampling, the *secondary sampling units* (or ssu's) are the observation units. Let M_i be the number of ssu's in the ith psu; in other words, $M_i := \#(\mathcal{C}_i)$.

First Stage: Select a simple random sample (SRS) $\omega_1 = \{\,\mathcal{C}_{i_1}, \mathcal{C}_{i_2}, \ldots, \mathcal{C}_{i_n}\,\}$ of n psu's from the collection of N psu's.

Second Stage: From each psu $\mathcal{C}\in\omega_1$ selected in the First Stage, select a simple random sample (SRS) $\omega_{\mathcal{C}}$ of m_i secondary sampling units (ssu's) from the collection of M_i ssu's in $\mathcal{C}$.

The sample is then $\omega := \bigsqcup_{\mathcal{C}\in\omega_1}\omega_{\mathcal{C}}$. In other words, the sample ω consists of all the secondary sampling units (or observation units) selected (during the Second Stage) from all the primary sampling units selected in the First Stage.

The Horvitz-Thompson estimator $\widehat{t}_{\mathrm{HT}}$, as defined below, is an unbiased estimator for the total of an $\mathbb{R}$-valued population characteristic $y:\mathcal{U}\longrightarrow\mathbb{R}$.

$$\widehat{t}_{\mathrm{HT}} \quad:=\quad \sum_{k\in\omega}\left(\frac{N}{n}\frac{M_{y_k}}{m_{y_k}}\right)y_k \quad=\quad \sum_{k\in\omega}\left(\frac{1}{\pi_k}\right)y_k \quad=\quad \sum_{\mathcal{C}\in\omega_1}\sum_{k\in\omega_{\mathcal{C}}}\left(\frac{N}{n}\frac{M_{y_k}}{m_{y_k}}\right)y_k,$$

where $M_{y_k} := M_i := \#(\mathcal{C}_i)$ and $m_{y_k} := m_i := \#(\omega_{\mathcal{C}_i})$ such that $\mathcal{C}_i$ is the unique psu containing the ssu $k \in \mathcal{U} = \bigsqcup_{i}^{N} \mathcal{C}_i$.

The variance of the Horvitz-Thompson estimator $\widehat{t}_{\mathrm{HT}}$ is given by:

$$\mathrm{Var}\left(\,\widehat{t}_{\mathrm{HT}}\,\right) \;=\; N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} \;+\; \sum_{i=1}^{N}\frac{N}{n}\cdot M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i},$$

where

$$S_t^2 \;:=\; \frac{1}{N-1}\sum_{i=1}^{N}\left(t_i-\frac{t}{N}\right)^2\,, \quad S_i^2 \;:=\; \frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(y_j-\frac{t_i}{M_i}\right)^2\,, \quad t \;:=\; \sum_{k\in\mathcal{U}} y_k\,, \quad \text{and} \quad t_i \;:=\; \sum_{k\in\mathcal{C}_i} y_k$$

IMPORTANT OBSERVATION: The first summand in the expression of $\mathrm{Var}\left(\,\widehat{t}_{\mathrm{HT}}\,\right)$ is due to variability in the First-Stage sampling, whereas the second summand is due to variability in the Second-Stage sampling.

5 One-stage Cluster Sampling

One-stage cluster sampling is a special form of two-stage cluster sampling in which all second-stage samples are censuses. In other words, following the notation introduced for two-stage cluster sampling, in one-stage cluster sampling, we have $\omega_{\mathcal{C}} = \mathcal{C}$, for each first-stage-selected $\mathcal{C} \in \omega_1$. This also implies $m_i = M_i$ for each $i = 1, 2, \ldots, N$.

Then, the Horvitz-Thompson estimator $\widehat{t}_{\mathrm{HT}}$ and its variance reduces to:

$$\begin{aligned}
\widehat{t}_{\mathrm{HT}} \;&:=\; \sum_{\mathcal{C}\in\omega_1}\sum_{k\in\mathcal{C}}\left(\frac{N}{n}\frac{M_{y_k}}{m_{y_k}}\right)y_k \;=\; \frac{N}{n}\cdot\sum_{\mathcal{C}\in\omega_1}\sum_{k\in\mathcal{C}} y_k \;=\; \frac{N}{n}\cdot\sum_{\mathcal{C}\in\omega_1} t_{\mathcal{C}}\,, \quad \text{where } t_{\mathcal{C}} := \sum_{k\in\mathcal{C}} y_k \\
\mathrm{Var}\left(\,\widehat{t}_{\mathrm{HT}}\,\right) \;&=\; N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} \;+\; \sum_{i=1}^{N}\frac{N}{n}\cdot M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i} \\
&=\; N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} \;+\; \sum_{i=1}^{N}\frac{N}{n}\cdot M_i^2\,(1-1)\,\frac{S_i^2}{m_i} \;=\; N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}
\end{aligned}$$

6 Stratified Simple Random Sampling as a special case of Two-stage Cluster Sampling

Stratified simple random sampling is a special case of two-stage cluster sampling in which the first-stage sampling is a census. In other words, if $\mathcal{U} = \bigsqcup_{i=1}^{N} \mathcal{C}_i$, then $\omega_1 = \{\,\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\,\}$. In particular, $n = N$.

Then, the Horvitz-Thompson estimator $\widehat{t}_{\mathrm{HT}}$ and its variance reduces to:

$$\begin{aligned}
\widehat{t}_{\mathrm{HT}} \;&:=\; \sum_{\mathcal{C}\in\omega_1}\sum_{k\in\omega_{\mathcal{C}}}\left(\frac{N}{n}\frac{M_{y_k}}{m_{y_k}}\right)y_k \;=\; \sum_{i=1}^{N} M_i\left(\frac{1}{m_i}\sum_{k\in\omega_{\mathcal{C}_i}} y_k\right) \;=\; \sum_{i=1}^{N} M_i\,\overline{y}_{\omega_{\mathcal{C}_i}} \\
\mathrm{Var}\left(\,\widehat{t}_{\mathrm{HT}}\,\right) \;&=\; N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} \;+\; \sum_{i=1}^{N}\frac{N}{n}\cdot M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i} \\
&=\; N^2\,(1-1)\,\frac{S_t^2}{n} \;+\; \sum_{i=1}^{N} 1\cdot M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i} \;=\; \sum_{i=1}^{N} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i}
\end{aligned}$$

The above formula agree exactly with those derived earlier for stratified simple random sampling (apart from obvious notational changes).

7 Calibration Estimators in Survey Sampling

This is a summary of [1].

Let $U = \{1, 2, \ldots, N\}$ be a finite population. Let $y : U \longrightarrow \mathbb{R}$ be an $\mathbb{R}$-valued function defined on U (commonly called a "population parameter"). We will use the common notation y_i for $y(i)$. We wish to estimate $T_y := \sum_{i\in U} y_i$ via survey sampling. Let $p : \mathcal{S} \longrightarrow (0, 1]$ be our chosen sampling design, where $\mathcal{S} \subseteq \mathcal{P}(U)$ is the set of all possible samples in the design, and $\mathcal{P}(U)$ is the power set of U. For each $k \in U$, let $\pi_k := \sum_{s\ni k} p(s)$ be the inclusion probability of k under the sampling design p. We assume $\pi_k > 0$ for each $k \in U$. Then, the Horvitz-Thompson estimator

$$\widehat{T}_y^{\mathrm{HT}}(s) \;:=\; \sum_{k\in s} \frac{y_k}{\pi_k} \;=\; \sum_{k\in s} d_k y_k \;=\; \sum_{k\in U} I_{ks}\frac{y_k}{\pi_k}, \quad \text{where } d_k := \frac{1}{\pi_k} \text{ and } I_{ks} \;:=\; \begin{cases} 1, & \text{if } k \in s \\ 0, & \text{otherwise} \end{cases}$$

is well-defined and is known to be a design-unbiased estimator of T_y; in other words,

$$E_p\left[\,\widehat{T}_y^{\mathrm{HT}}\,\right] \;=\; \sum_{s\in\mathcal{S}} p(s)\cdot\widehat{T}_y^{\mathrm{HT}}(s) \;=\; \sum_{s\in\mathcal{S}} p(s)\cdot\left(\sum_{k\in U} I_{ks}\frac{y_k}{\pi_k}\right) \;=\; \sum_{k\in U}\frac{y_k}{\pi_k}\left(\sum_{s\in\mathcal{S}} p(s) I_{ks}\right) \;=\; \sum_{k\in U}\frac{y_k}{\pi_k}\,\pi_k \;=\; T_y.$$

We will call the d_k's above the *Horvitz-Thompson weights*.

Roughly, a calibration estimator for T_y is an estimator of the form:

$$\widehat{T}_y^{\mathrm{Cal}}(s) \;:=\; \sum_{k\in s} w_k(s) y_k,$$

where the sample-dependent "calibrated" weights $w_k(s)$ are the solution of a constrained minimization problem where the objective function depends on the $w_k(s)$'s and the Horvitz-Thompson weights d_k's, while the constraints involve the $w_k(s)$'s and auxiliary information. More precisely, the calibrated weights $w_k(s)$ solve the following constrained minimization problem:

> **Conceptual framework:** Let $\mathbf{x} : U \longrightarrow \mathbb{R}^{1\times J}$ be an $\mathbb{R}^{1\times J}$-valued function defined on U. We use the common notation $\mathbf{x}_k$ for $\mathbf{x}(k)$, for each $k \in U$.
>
> **Assumptions:**
>
> - The population total of $\mathbf{x}$
>
> $$T_{\mathbf{x}} \;:=\; \sum_{k\in U} \mathbf{x}_k \in \mathbb{R}^{1\times J}$$
>
>   is known.
> - For each $s \in \mathcal{S}$, the value $(y_k, \mathbf{x}_k)$ can be observed for each $k \in s$ via the sampling procedure.
>
> **Constrained Minimization Problem:** For each $k \in U$, let $q_k > 0$ be chosen. For each $s \in \mathcal{S}$, the calibrated weights $w_k(s)$, for $k \in s$, are obtained by minimizing the following objective function:
>
> $$f_s(w_k(s)\,;\,d_k, q_k) \;:=\; \sum_{k\in s} \frac{(w_k(s) - d_k)^2}{d_k q_k}$$
>
> subject to the (vectorial) constraint on $w_k(s)$:
>
> $$\mathbf{h}(w_k(s)\,;\,\mathbf{x}_k, T_{\mathbf{x}}) \;:=\; -T_{\mathbf{x}} + \sum_{k\in s} w_k(s)\,\mathbf{x}_k \;=\; 0$$

References

[1] Deville, J.-C., and Särndal, C.-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 418 (1992), 376–382.

[2] Hájek, J. Limiting distributions in simple random sampling from a finite population. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences 5* (1960), 361–374.

[3] Lohr, S. L. *Sampling: Design and Analysis*, first ed. Duxbury Press, 1999.