Sampling design of fixed size

A sampling design with fixed sample size n verifies the followoing properties about respectively the expectation, the variance and the covariance between two units:

$$\sum_{k \in U} \pi_k = n$$

$$\sum_{k \in U, k \neq l} \pi_{kl} = \pi_l (n-1)$$

$$\sum_{k \in U} \Delta_{kl} = 0$$

$U = 1, ..., N$ is the finite population of size N. S denotes the sample. π_k is the inclusion probability of the unit k.

The Horvitz-Thompson estimator (1/2)

Let Y denote the variable of interest and$y$, one realization.The Horvitz-Thompson estimator is defined as

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}$$

If all $\pi_k > 0$, then $\hat{Y}_\pi$ is an unbiased estimator of the total Y. Indeed, we have that

$$E[\hat{Y}_\pi] = E\Big[\sum_{k \in S} \frac{y_k}{\pi_k}\Big] = E\Big[\sum_{k \in U} \frac{I_k y_k}{\pi_k}\Big] = \sum_{k \in U} \frac{E[I_k] y_k}{\pi_k} = \sum_{k \in U} y_k = Y$$

If some inclusion probabilities π_k are 0, then this estimator is biased. There exists alternatives.

The Horwitz-Thompson estimator (2/2)

The estimation of the mean is given by

$$\hat{\bar{Y}}_{\pi} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}$$

and since $N = \sum_{k \in U} 1$, we can estimate this quantity by

$$\hat{N} = \sum_{k \in S} \frac{1}{\pi_k}$$

If some inclusion probabilities π_k are 0, then this estimator is biased. There exists alternatives.

Variance of the Horvitz-Thompson estimator

The variance of the Horvitz-Thompson estimator of the total is derived as follows:

$$
\begin{aligned}
var(\hat{Y}_\pi) &= var\Big(\sum_{k\in U} \frac{I_k y_k}{\pi_k}\Big) \\
&= \sum_{k\in U} \frac{y_k^2}{\pi_k^2} var(I_k) + \sum_{k\in U} \sum_{l\in U, l\neq k} \frac{y_k}{\pi_k} cov(I_k, I_l) \\
&= \sum_{k\in U} \frac{y_k^2}{\pi_k^2} \pi_k(1-\pi_k) + \sum_{k\in U} \sum_{l\in U, l\neq k} \frac{y_k y_l}{\pi_k \pi_l}(\pi_{kl} - \pi_k \pi_l) \\
&= \sum_{k\in U} \sum_{l\in U, l\neq k} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl}
\end{aligned}
$$

Estimation of the variance of the Horvitz-Thompson estimator

In general, an unbiased estimator for the Horvitz-Thompson estimator is as follows (drawback: can take negative values):

$$v\hat{a}r(\hat{Y}_\pi) = \sum_{k \in S} \frac{y_k^2}{\pi_k^2} \pi_k (1 - \pi_k) + \sum_{k \in S} \sum_{l \in S, l \neq k} \frac{y_k y_l}{\pi_{kl} \pi_k \pi_l} (\pi_{kl} - \pi_k \pi_l)$$

Another unbiased estimator in the case of a design with fixed sample size, called the Sen-Yates-Grundy estimator (see. Sen, 1953, Yates and Grundy, 1953) can be constructed as follows:

$$v\hat{a}r(\hat{Y}_\pi) = -\frac{1}{2} \sum_{k \in S} \sum_{l \in S, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}}$$

If $\pi_k \pi_l - \pi_{kl} \geq 0$, this estimator is positive.

Confidence Intervals

Then a $(1-\alpha)*100\%$ Confidence Interval for a total is given by

$$\left[\hat{Y}_\pi - z_{1-\alpha/2}\sqrt{v\hat{a}r(\hat{Y}_\pi)}, \hat{Y}_\pi + z_{1-\alpha/2}\sqrt{v\hat{a}r(\hat{Y}_\pi)}\right]$$

For the mean, the $(1-\alpha)*100\%$ Confidence Interval is given by

$$\left[\hat{\bar{Y}}_\pi - z_{1-\alpha/2}\sqrt{v\hat{a}r(\hat{\bar{Y}}_\pi)}, \hat{\bar{Y}}_\pi + z_{1-\alpha/2}\sqrt{v\hat{a}r(\hat{\bar{Y}}_\pi)}\right]$$

Working example

Suppose you are conducting a survey to estimate the average income of households in a certain neighborhood. You use a simple random sampling design where each household has an equal probability of being selected. You collect income data from a sample of 100 households and obtain the following information:

Sample size: $n = 100$. Total number of households in the neighborhood: $N = 500$. Inclusion probabilities for each household: $\pi_k = \frac{N}{n} = \frac{500}{100} = 5$

You also find that the sample mean income is $\bar{y} = 45,000$ and the variance of the sample total income is $v\hat{a}r(\hat{Y}_\pi) = 32,000$.

What is the Horvitz-Thompson estimator for the total income of households in the neighborhood? What is a 95% confidence interval for the estimated total income using the Horvitz-Thompson estimator?

Example 2: Application 2/3

The Horvitz-Thompson estimator for the total income, given that $\pi_k = 5$ and $\bar{y}$ =45,000$, isgivenby$:

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{1}{5} \sum_{k \in S} y_k = \frac{1}{5} * 100 * 45,000 = 90,000$$

To compute a 95% confidence interval, we use the formula:

$$\left[\hat{Y}_\pi - z_{1-\alpha/2} \sqrt{v\hat{a}r(\hat{Y}_\pi)}, \hat{Y}_\pi + z_{1-\alpha/2} \sqrt{v\hat{a}r(\hat{Y}_\pi)} \right]$$

Given that $z_{1-\alpha/2} = z_{0.975} \approx 1.96$ for a 95% confidence level, we have:

$$\left[90,0001.96 * \sqrt{32,000}, 90,000 + 1.96 * \sqrt{32,000} \right] = [89,649, 90,351]$$

References

P. Ardilly, Y. Tillé, 2006, Sampling Methods: Exercices and Solutions, Springer.

course notes