8. Additional Topics in Sampling

8.1. Sampling with Replacement

Sampling with replacement is where a unit can be selected more than once. For a sample size n, the n selections are independent and each unit in the population has the same probability of inclusion in the sample. Simple random sampling with replacement is characterized by the property that each possible sequence of n units - distinguishing order of selection and possibly including repeat selections - has equal probability under the design.

Let $\bar{Y}_n$ denote the sample mean of the n observations; that is

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Note that if a unit is selected more than once, its y-value is utilized more than once in the estimator. Its variance can be shown to be

$$\mathbb{V}(\bar{Y}_n) = \left(1 - \frac{1}{N}\right)\frac{\sigma^2}{n}$$

Thus the variance of the sample mean with simple random sampling without replacement is lower since

$$1 - \frac{1}{N} \geq 1 - \frac{n}{N}.$$

The estimator $\bar{Y}_n$ depends on the number of times each unit is selected, so that two surveys observing exactly the same set of distinct units, but with different repeat selections, would in general yield different estimates. This situation can be avoided using the sample mean of the distinc observations.

The number of distinct units contained in the sample, termed the effective sample size, is denoted v. Let $\bar{Y}_v$ be the sample mean of the distinc observations:

$$\bar{Y}_v = \frac{1}{v}\sum_{i=1}^{v} y_i.$$

The estimator $\bar{Y}_v$ is an unbiased estimator of the population mean. The variance of $\bar{Y}_v$ can be shown to be less than that of $\bar{Y}_n$, but still not as small as the variance of the sample mean under simple random sampling without replacement.

8.2. Unequal Probability Sampling

With some sampling procedures, different units in the population have different probabilities of being included. As an example, if a study area is divided into plots of unequal sizes, it may be desired to assign larger inclusion probabilities to larger plots.

8.2.1. Sampling with replacement

Suppose that sampling is with replacement and that on each draw the probability of selecting the ith unit of the population is p_i, for $i = 1, \ldots, N$.

An unbiased estimator of the population total τ is

$$\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i},$$

called the Hansen-Hurwitz estimator. The variance of this estimator is

$$\mathbb{V}(\hat{\tau}_p) = \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - \tau \right)^2$$

Proof: Consider a sample of size 1 and suppose the sth unit was selected. The Hansen-Hurwitz estimator can be written as $t_s = y_s/p_s$ and its expected value is

$$\mathbb{E}(t_s) = \sum_{s=1}^{N} t_s p_s = \sum_{j=1}^{N} y_j = \tau.$$

The variance of t_s is

$$\mathbb{V}(t_s) = \mathbb{E}((t_s - \mathbb{E}(t_s))^2) = \sum_{s=1}^{N} (t_s - \tau)^2 p_s = \sum_{j=1}^{N} \left(\frac{y_j}{p_j} - \tau \right)^2 p_j.$$

When sampling is with replacement, the selections are independent. Thus with n independent draws, in which unit j has selection probability p_j on each draw, the Hansen-Hurwitz estimator is the sample mean of n independent and identically distributed random variables $t_{s1}, \dots, t_{sn}$ each with mean and variance above so that one can write

$$\hat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n} t_{si}.$$

Consequently

$$\mathbb{E}(\hat{\tau}_p) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(t_{si}) = \frac{1}{n}\sum_{i=1}^{n}\tau = \tau$$

and

$$\mathbb{V}(\hat{\tau}_p) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} t_{si}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(t_{si}) = \frac{1}{n}\sum_{j=1}^{N}\left(\frac{y_j}{p_j} - \tau\right)^2 p_j \quad \square$$

An unbiased estimator of this variance is

$$\widehat{\mathbb{V}(\hat{\tau}_p)} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{\tau}_p\right)^2.$$

[Without proof]

Notice that if the selection probabilities p_i were proportional to the variables y_i, the ratio y_i/p_i would be constant and the Hansen-Hurwitz estimator would have zero variance. The variance would be low if the selection probabilities could be set approximately proportional to the y-values. Of course, the population y-values are unknown prior to sampling. If it is believed that the y-values are approximately proportional to some known variable such as the sizes of the units, the selection probabilities can be chosen proportional to the value of that known variable.

An unbiased estimator of the population mean μ is $\hat{\mu}_p = \frac{1}{N}\hat{\tau}_p$, having variance $\mathbb{V}(\hat{\mu}_p) = \frac{1}{N^2}\mathbb{V}(\hat{\tau}_p)$ and estimated variance $\widehat{\mathbb{V}(\hat{\mu}_p)} = \frac{1}{N^2}\widehat{\mathbb{V}(\hat{\tau}_p)}$. An approximate $(1-\alpha)100\%$ confidence interval for the population total is

$$\hat{\tau}_p \pm z\left(\frac{\alpha}{2}\right)\sqrt{\widehat{\mathbb{V}(\hat{\tau}_p)}}$$

For small sample sizes, the use of the T-distribution is recommended.

8.2.2. Any Sampling Design

With any design, with or without replacement, given probability π_i that unit i is included in the sample, $i = 1, \ldots, N$, the Horvitz-Thompson estimator of the population total τ is

$$\hat{\tau}_\pi = \sum_{i=1}^{v} \frac{y_i}{\pi_i},$$

where v is the effective sample size. The Horvitz-Thompson estimator is unbiased. Its variance is

$$\mathbb{V}(\hat{\tau}_\pi) = \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i=1}^{N} \sum_{i \neq j} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j,$$

where π_{ij} is the probability that both unit i and unit j are included in the sample. An unbiased estimator of this variance is

$$\widehat{\mathbb{V}(\hat{\tau}_\pi)} = \sum_{i=1}^{v} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^{v} \sum_{i \neq j} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \frac{y_i y_j}{\pi_{ij}}$$

if all the joint inclusion probabilities π_{ij} are greater than zero.

Proof: Define the indicator variable Z_i to be 1 if the ith unit of the population is included in the sample and zero otherwise. Thus $\mathbb{E}(Z_i) = \pi_i$ and $\mathbb{V}(Z_i) = \pi_i(1-\pi_i)$ and $\text{Cov}(Z_i, Z_j) = \pi_{ij} - \pi_i \pi_j$. The Horvitz-Thompson estimator can be written as

$$\hat{\tau}_\pi = \sum_{i=1}^{N} \frac{y_i Z_i}{\pi_i},$$

and

$$\mathbb{E}(\hat{\tau}_\pi) = \sum_{i=1}^{N} \frac{y_i \mathbb{E}(Z_i)}{\pi_i} = \sum_{i=1}^{N} y_i = \tau.$$

The variance of $\hat{\tau}_\pi$ is

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_\pi) &= \mathbb{V}\left(\sum_{i=1}^{N} \frac{y_i Z_i}{\pi_i} \right) = \sum_{i=1}^{N} \left(\frac{y_i}{\pi_i} \right)^2 \mathbb{V}(Z_i) + \sum_{i=1}^{N} \sum_{i \neq j} \text{Cov}\left(\frac{y_i Z_i}{\pi_i}, \frac{y_j Z_j}{\pi_j} \right) \\
&= \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i=1}^{N} \sum_{i \neq j} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j.
\end{aligned}$$

To see that $\widehat{\mathbb{V}(\hat{\tau}_\pi)}$ is unbiased for $\mathbb{V}(\hat{\tau}_\pi)$, define Z_{ij} to be 1 if both units i and j are included in the sample and zero otherwise. The estimator of the variance may be written as

$$\widehat{\mathbb{V}(\hat{\tau}_\pi)} = \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 Z_i + \sum_{i=1}^{N} \sum_{i \neq j} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \frac{y_i y_j Z_{ij}}{\pi_{ij}}.$$

Since $E(Z_{ij}) = \pi_{ij}$ unbiasedness follows immediately.

An unbiased estimator of the population mean is $\hat{\mu}_\pi = \frac{1}{N}\hat{\tau}_\pi$ having variance $\mathbb{V}(\hat{\mu}_\pi) = \frac{1}{N^2}\mathbb{V}(\hat{\tau}_\pi)$ and estimated variance $\widehat{\mathbb{V}(\hat{\mu}_\pi)} = \frac{1}{N^2}\widehat{\mathbb{V}(\hat{\tau}_\pi)}$. An approximate $(1-\alpha)100\%$ confidence interval for the population total is

$$\hat{\tau}_\pi \pm z\left(\frac{\alpha}{2}\right)\sqrt{\widehat{\mathbb{V}(\hat{\tau}_\pi)}}$$

The Horvitz-Thompson estimator is unbiased but can have a large variance. A generalized unequal-probability estimator of the population mean is

$$\hat{\mu}_g = \frac{\sum_{i=1}^{v} y_i/\pi_i}{\sum_{i=1}^{v} 1/\pi_i}$$

Its numerator is the ordinary Horvitz-Thompson estimator, which gives an unbiased estimate of the population total τ. The denominator can be viewed as another Horvitz-Thompson estimator for the population size N. Thus $\hat{\mu}_g$ estimates $\mu = \tau/N$. But since the ratio of two unbiased estimators is not unbiased, $\hat{\mu}_g$ is not unbiased.

[No derivation of the variance of $\hat{\mu}_g$]

The generalized unequal-probability estimator for the population total is $\hat{\tau}_g = N\hat{\mu}_g$.

8.3. Cluster and Systematic Sampling

Suppose a population is partitioned into primary units, each primary unit being composed of secondary units. Whenever a primary unit is included in the sample, the y-values of every secondary unit within it are observed.

In systematic sampling, a single primary unit consists of secondary units spaced in some systematic fashion throughout the population. In cluster sampling, a primary unit consists of a cluster of secondary units, usually in closed proximity to each other. In the spatial setting, a systematic sample primary unit may be composed of a collection of plots in a grid pattern over the study area. Cluster primary units include such spatial arrangements as square collections of adjacent plots.

The key point in both systematic and clustered arrangements is that whenever any secondary unit of a primary unit is included in the sample, all the secondary units of that primary unit are included. Even though the actual measurements may be made on secondary units, it is the primary units that are selected.

Let N be the number of primary units in the population and n the number of primary units in the sample. Let M_i be the number of secondary units in the ith primary unit. The total number of units in the population is $M = \sum_{i=1}^{N} M_i$. Let y_{ij} denote the value of the variable of interest of the j-th secondary unit in the i-th primary unit. The total of the y values in the i-th primary unit will be denoted simply y_i, that is $y_i = \sum_{j=1}^{M_i} y_{ij}$. The population total is $\tau = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} y_i$. The population mean per primary unit is $\mu_1 = \tau/N$. The population mean per secondary unit is $\mu = \tau/M$.

8.3.1. Primary Units Selected by Simple Random Sampling

When primary units are selected by simple random sampling without replacement, an unbiased estimator of the population total is

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} y_i = N\bar{Y},$$

and its variance is

$$\mathbb{V}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n},$$

where σ_u^2 is the finite-population variance of the primary unit totals,

$$\sigma_u^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_1)^2.$$

An unbiased estimator of the variance of $\hat{\tau}$ is

$$\widehat{\mathbb{V}(\hat{\tau})} = N(N-n)\frac{S_u^2}{n},$$

where S_u^2 is the sample variance of the primary unit total,

$$S_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{Y})^2.$$

These results are familiar from simple random sampling.

An unbiased estimator of the mean per primary unit μ_1 is $\bar{Y} = \hat{\tau}/N$ and an unbiased estimator of the mean per secondary unit μ is $\hat{\mu} = \hat{\tau}/M$. The variance of $\bar{Y}$ is $\mathbb{V}(\bar{Y}) = (1/N^2)\mathbb{V}(\hat{\tau})$ and the variance of $\hat{\mu}$ is $\mathbb{V}(\hat{\mu}) = (1/M^2)\mathbb{V}(\hat{\tau})$. The estimated variances are obtained similary by dividing the estimated variance of $\hat{\tau}$ by N^2 for the mean per primary unit or M^2 for the mean per secondary unit.

If primary unit total y_i is highly correlated with primary unit size M_i, the following estimator for the population total may be efficient

$$\hat{\tau}_r = rM,$$

where

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

The estimator $\hat{\tau}_r$ is usually called the *ratio estimator*. The estimator $\hat{\tau}_r$ is not unbiased. However the bias tends to be small with large sample sizes and the mean squared error may be considerably less than that of the unbiased estimator when the y_i and M_i tend to be proportionally related.

An approximated formula for the variance of the ratio estimator is

$$\mathbb{V}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)}\sum_{i=1}^{N}(y_i - M_i r)^2.$$

An estimator of this variance is given by

$$\widehat{\mathbb{V}(\hat{\tau}_r)} = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{n} (y_i - M_i r)^2,$$

this is because $r = \hat{\tau}_r / M$ is an estimator of the population mean μ per secondary unit. The ratio estimator of the population mean μ_1 per primary unit is $\hat{\tau}_r / N$.

8.3.2. The Basic Principle of Cluster and Systematic Sampling

Since every secondary unit is observed within a selected primary unit, the within-primary-unit variance does not enter into the variances of the estimators. Thus, the basic *systematic and cluster sampling principle* is that to obtain estimators of low variance, the population should be partitioned into clusters in such a way that one cluster is similar to another. Equivalently, the within-primary-unit variance should be as great as possible in order to obtain the most precise estimators of the population mean or total. The ideal primary unit contains the full diversity of the population and hence is representative.

8.3.3. Single Systematic Sample

Many surveys utilizing a systematic design select a single starting unit at random and then observe every secondary unit at the appropriate spacing from there. Thus the sample consists of a single primary unit selected at random. From a sample of size 1 it is possible to obtain an unbiased estimator of the population mean or total, but it is not possible to obtain an unbiased estimator of its variance.

Naively proceeding as if the M_1 secondary units in the single systematic primary unit were a simple random sample from the M secondary units in the population and using the variance formula from simple random sampling leads to good variance estimates only if the units of the population can reasonably be conceived as being in random order. With many natural populations, in which nearby units tend to be similar to each other, this procedure tends to overestimate the variance of the estimator of the population mean and total.

8.3.4. Variance in Cluster and Systematic Sampling

The effectiveness of cluster or systematic sampling depends on the variance resulting from using primary units of a given size and shape. We compare the variance of selecting n primary units with a simple random sample of an equivalent number of secondary units.

The average size of clusters in the population is $\bar{M} = M/N$, so the expected number of secondary units in a simple random sample of n primary units is $n\bar{M}$. Denote the unbiased estimator of the population total based on simple random sampling of $n\bar{M}$ secondary unit with $\hat{\tau}_{srs}$. Its variance is

$$\mathbb{V}(\hat{\tau}_{srs}) = M(M - n\bar{M})\frac{\sigma^2}{n\bar{M}} = \frac{\bar{M}N(\bar{M}N - n\bar{M})}{n\bar{M}}\sigma^2 = N^2\left(\frac{\bar{M}(N-n)}{nN}\right)\sigma^2,$$

where σ^2 is the finite-population variance for secondary units,

$$\sigma^2 = \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}\frac{(y_{ij} - \mu)^2}{N\bar{M} - 1}$$

and $\mu = \tau/N\bar{M}$.

For a cluster or systematic sample, with a simple random sample of n primary units, the unbiased estimator will be denoted $\hat{\tau}_u$, with the subscript u indicating that the design with which the estimator is used is a random sample of primary units of type u (for example from square clusters, rectangular clusters or systematic samples). The variance of $\hat{\tau}_u$ is

$$\mathbb{V}(\hat{\tau}_u) = N(N-n)\frac{\sigma_u^2}{n} = N^2\left(\frac{N-n}{nN}\right)\sigma_u^2,$$

where $\sigma_u^2 = \sum_{i=1}^{N}(y_i - \mu_1)^2/(N-1)$ and $\mu_1 = \tau/N$.

The relative efficiency of the cluster (or systematic) sample to the simple random sample of equivalent sample size, defined as the ratio of the variances, is

$$\frac{\mathbb{V}(\hat{\tau}_{srs})}{\mathbb{V}(\hat{\tau}_u)} = \frac{\bar{M}\sigma^2}{\sigma_u^2}$$

The cluster (systematic) sampling is efficient if the variance σ_u^2 between primary units is small relative to the overall population variance σ^2.

To estimate this relative efficiency using data from a cluster or systematic sampling design, the usual sample variance S^2 cannot be used as estimate of σ^2 because the data were not obtained with simple random sampling. Instead, σ^2 can be estimated using analysis of variance of the cluster (systematic) data as follows. For simplicity, assume that each of the N primary units has an equal number $\bar{M}$ of secondaty units. The total sum of squares in the population can be partitioned as

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\mu)^2 &= \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\bar{Y}_i+\bar{Y}_i-\mu)^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\bar{Y}_i)^2+\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(\bar{Y}_i-\mu)^2+2\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\bar{Y}_i)(\bar{Y}_i-\mu) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\bar{Y}_i)^2+\bar{M}\sum_{i=1}^{N}(\bar{Y}_i-\mu)^2 \qquad (8.1)
\end{aligned}$$

where $\bar{Y}_i=\sum_{j=1}^{\bar{M}}y_{ij}/\bar{M}$. This is because $\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\bar{Y}_i)(\bar{Y}_i-\mu)=0$. This is because

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\bar{Y}_i)(\bar{Y}_i-\mu) &= \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}y_{ij}\bar{Y}_i-\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}y_{ij}\mu-\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}\bar{Y}_i^2+\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}\mu\bar{Y}_i \\
&= \sum_{i=1}^{N}\bar{Y}_i\sum_{j=1}^{\bar{M}}y_{ij}-\mu\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}y_{ij}-\bar{M}\sum_{i=1}^{N}\bar{Y}_i^2+\mu\bar{M}\sum_{i=1}^{N}\bar{Y}_i \\
&= \bar{M}\sum_{i=1}^{N}\bar{Y}_i^2-\mu\bar{M}\sum_{i=1}^{N}\bar{Y}_i-\bar{M}\sum_{i=1}^{N}\bar{Y}_i^2+\mu\bar{M}\sum_{i=1}^{N}\bar{Y}_i \\
&= 0.
\end{aligned}$$

The first term in Equation (8.1) is the within-primary-unit sum of squares, whilst the second term is the between-primary-unit sum of squares. Write

$$\sigma_w^2=\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}\frac{(y_{ij}-\bar{Y}_i)^2}{N(\bar{M}-1)}$$

for the within-primary-unit variance and

$$\sigma_b^2 = \sum_{i=1}^{N} \frac{(\bar{Y}_i - \mu)^2}{N-1}$$

for the variance between primary units means. Note that $\sigma_u^2 = \bar{M}^2\sigma_b^2$.

An unbiased estimator of σ_w^2 using the random samples of clusters is

$$S_w^2 = \sum_{i=1}^{n}\sum_{j=1}^{\bar{M}} \frac{(y_{ij} - \bar{Y}_i)^2}{n(\bar{M}-1)}$$

and an unbiased estimator of σ_b^2 is

$$S_b^2 = \sum_{i=1}^{n} \frac{(\bar{Y}_i - \hat{\mu})^2}{n-1}.$$

Equation (8.1) may then be written as

$$(N\bar{M}-1)\sigma^2 = N(\bar{M}-1)\sigma_w^2 + (N-1)\bar{M}\sigma_b^2.$$

An unbiased estimator of σ^2 from the simple random cluster sample is

$$\hat{\sigma}^2 = \frac{N(\bar{M}-1)S_w^2 + (N-1)\bar{M}S_b^2}{N\bar{M}-1}.$$

The estimated relative efficiency of cluster (systematic) sampling based on the data is then

$$\frac{\bar{M}\hat{\sigma}^2}{S_u^2} = \frac{\hat{\sigma}^2}{\bar{M}S_b^2}$$