

## Cluster and Systematic Sampling

Although systematic sampling and cluster sampling seem on the surface to be opposites—the one spacing out the units of a sample and the other bunching them together—the two designs share the same structure. The 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 close 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 or long, narrow strips of adjacent units. A cluster sample consisting of a simple random sample of 10 primary units, each consisting of eight secondary units, is shown in Figure 12.1. A systematic sample with two randomly selected starting points is shown in Figure 12.2. The systematic sample consists of two primary units (distinguished by their shading), each with 16 secondary units.

The key point in any of the systematic or 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. In principle, one could dispense with the concept of the secondary units, regarding the primary units as the sampling units and using, as the variable of interest for any primary unit, the total of the  $y$ -values of the secondary units within it. Then all properties of estimators may be obtained based on the design by which the sample of primary units is selected. However, several common features of systematic and cluster sampling make these designs worth considering as special cases:

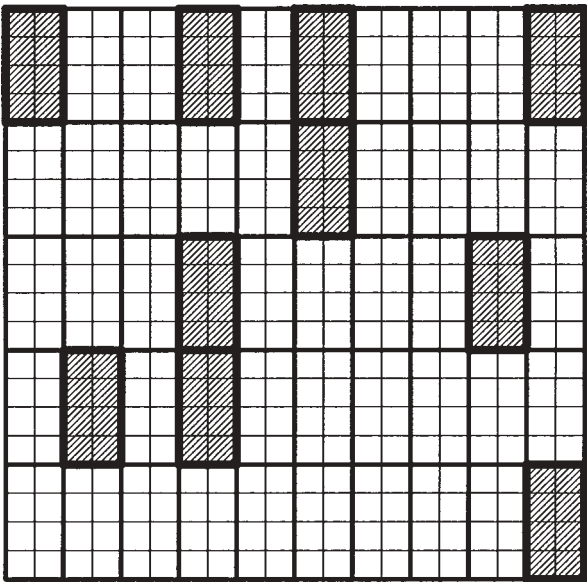


Figure 12.1. Cluster sample.

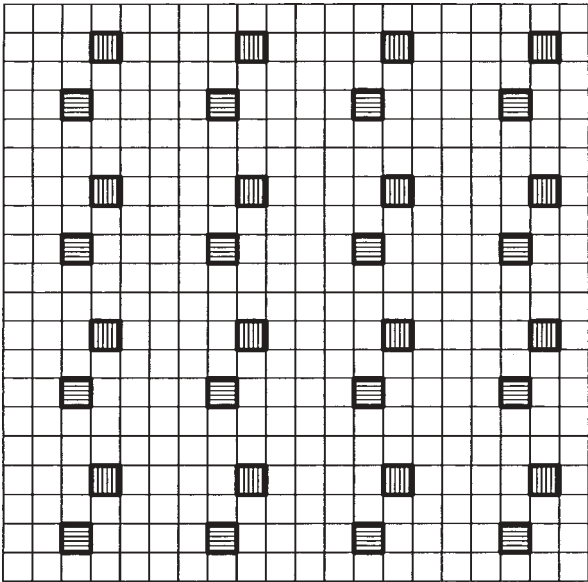


Figure 12.2. Systematic sample with two starting points.

1. In systematic sampling, it is not uncommon to have a sample size of 1, that is, a single primary unit.
2. In cluster sampling, the size of the cluster may serve as auxiliary information that may be used either in selecting clusters with unequal probabilities or in forming ratio estimators.
3. The size and shape of clusters may affect efficiency.

Let  $N$  denote 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  $i$ th primary unit. The total number of secondary 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  $\mu = \tau/M$ .

## 12.1. PRIMARY UNITS SELECTED BY SIMPLE RANDOM SAMPLING

### Unbiased Estimator

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

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y} \quad (12.1)$$

where  $\bar{y} = (1/n) \sum_{i=1}^n y_i$ , the sample mean of the primary unit totals. The variance of this estimator is

$$\text{var}(\hat{\tau}) = N(N-n) \frac{\sigma_u^2}{n} \quad (12.2)$$

where  $\sigma_u^2$  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 estimate of the variance of  $\hat{\tau}$  is

$$\widehat{\text{var}}(\hat{\tau}) = N(N-n) \frac{s_u^2}{n}$$

where  $s_u^2$  is the sample variance of the primary unit totals,

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

These results are familiar from simple random sampling. An unbiased estimator of the mean per primary unit  $\mu_1$  is  $\bar{y} = \hat{\tau}/N$ , and an unbiased estimator of the mean per secondary unit  $\mu$  is  $\hat{\mu} = \hat{\tau}/M$ . The variance of  $\bar{y}$  is  $\text{var}(\bar{y}) = (1/N^2)\text{var}(\hat{\tau})$ , and the variance of  $\hat{\mu}$  is  $\text{var}(\hat{\mu}) = (1/M^2)\text{var}(\hat{\tau})$ . The estimated variances are obtained similarly 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.

### Ratio Estimator

If primary unit total  $y_i$  is highly correlated with primary unit size  $M_i$ , a ratio estimator based on size may be efficient. The ratio estimator of the population total is

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

where the sample ratio  $r$  is

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

The population ratio is the mean per secondary unit  $\mu$ . As a ratio estimator,  $\hat{\tau}_r$  is not unbiased, but the bias tends to be small with large sample sizes, and the mean square error may be considerably less than that of the unbiased estimator when the  $y_i$  and the  $M_i$  tend to be proportionally related.

An approximate formula for the mean square error or variance of the ratio estimator is

$$\text{var}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^N (y_i - M_i \mu)^2$$

An estimator of this variance is given by

$$\widehat{\text{var}}(\hat{\tau}_r) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^n (y_i - rM_i)^2$$

or the adjusted estimator for the variance of a ratio estimator

$$\widetilde{\text{var}}(\hat{\tau}_r) = \left( \frac{nM}{N \sum_{i=1}^n M_i} \right)^2 \widehat{\text{var}}(\hat{\tau}_r)$$

Alternative variance estimators for ratio estimators are discussed in the chapter on auxiliary data and ratio estimation (Chapter 7) and are reviewed in J. N. K. Rao (1988, pp. 402–403), P. S. R. S. Rao (1988, pp. 454–456), and Royall (1988, pp. 402–403).

To estimate the population mean  $\mu_1$  per primary unit, the ratio estimator would be  $\hat{\mu}_{1r} = \hat{\tau}_r/N$ , for which the mean square error formulas above would be divided by  $N^2$ . To estimate the population mean  $\mu$  per secondary unit, the ratio estimator is  $\hat{\mu}_r = \hat{\tau}_r/M = r$ , for which one would divide the mean square error expressions above by  $M^2$ .

## 12.2. PRIMARY UNITS SELECTED WITH PROBABILITIES PROPORTIONAL TO SIZE

Suppose that the primary units are selected with replacement with draw-by-draw selection probabilities proportional to the sizes of the primary units, that is,  $p_i = M_i/M$ . One way to carry out such a design is to select  $n$  secondary units from the  $M$  in the population, using simple random sampling with replacement: A primary unit is selected every time any of its secondary units is selected.

### Hansen–Hurwitz (PPS) Estimator

An unbiased estimator of the population total under sampling with replacement with probabilities proportional to size, based on the Hansen–Hurwitz estimator, is

$$\hat{\tau}_p = \frac{M}{N} \sum_{i=1}^n \frac{y_i}{M_i}$$

with each observation utilized in the sum as many times as its primary unit is selected. The variance of this estimator is

$$\text{var}(\hat{\tau}_p) = \frac{M}{n} \sum_{i=1}^n M_i (\bar{y}_i - \mu)^2$$

where  $\bar{y}_i = y_i/M_i$ . An unbiased estimator of this variance is

$$\widehat{\text{var}}(\hat{\tau}_p) = \frac{M^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_p)^2$$

where  $\hat{\mu}_p = \hat{\tau}_p/M$ .

The estimator  $\hat{\mu}_p$  is unbiased for the population mean per secondary unit  $\mu$  under the probability-proportional-to-size selection, while  $\hat{\mu}_{1p} = \hat{\tau}_p/N$  is unbiased for the population mean per primary unit. Variance formulas for these estimators are obtained by dividing the variance expressions for  $\hat{\tau}_p$  by  $M^2$  or  $N^2$ .

### Horvitz–Thompson Estimator

A Horvitz–Thompson estimator can also be computed for this design, using the inclusion probabilities

$$\pi_i = 1 - (1 - p_i)^n$$

and joint inclusion probabilities

$$\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n]$$

based on the selection probabilities  $p_i = M_i/M$ .

The Horvitz–Thompson estimator for the population total is

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

where  $v$  is the number of distinct primary units in the sample. Variance formulas for this estimator were given in Section 6.2.

### 12.3. THE BASIC PRINCIPLE

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 or mean square error, 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.”

With natural populations of spatially distributed plants, animals, or minerals, and with many human populations, the condition above is typically satisfied by systematic primary units, in which the secondary units are spaced apart, but not by clusters of geographically adjacent units. Cluster sampling is more often than not carried out for reasons of convenience or practicality rather than to obtain lowest variance for a given number of secondary units observed.

### 12.4. 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 or total.

A variety of procedures for estimating variance from a single systematic sample are discussed in Bellhouse (1988a), Murthy and Rao (1988), and Wolter (1984). One of the simplest is to combine pairs (or larger groups) of adjacent units into “strata” and estimate variance as if stratified random sampling had been used.

**Example 1:** The distinctions between the estimators can be illustrated with a systematic sample selected from a population in which the number  $N$  of possible systematic samples does not divide evenly into the number  $M$  of (secondary) units in the population. In a survey of bald eagle nests, a coastline 1300 km in length is divided into units of 100-km length, so that there are  $M = 13$  of these units in the population. A “one-in-three” systematic sample is selected by choosing at random one of the first three units and then including that unit and every third unit thereafter in the sample. For every unit included in the sample, all eagle nests are counted using research vessels and aircraft. Thus, a single systematic sample is selected, but the size  $M_i$  of the various possible samples differs. The number of primary units (the number of possible systematic samples) is  $N = 3$ , of which  $n = 1$  will be selected. If the first unit is chosen as a starting point,  $M_1 = 5$  units will be observed, whereas if either of the other starting points is chosen,  $M_i = 4$  units will be observed.

Suppose that unit 3 is selected at random from the first three, and the  $y$ -values observed on the survey are 5, 1, 10, and 18 nests. Then the unbiased estimate of the total number of nests on the coastline is

$$\hat{\tau} = \frac{3}{1}(5 + 1 + 10 + 18) = 102$$

However, presented with data from such a survey, many people would choose to take the average of the four units observed and multiply by the number of units in the population, obtaining the estimate

$$\hat{\tau}_r = \frac{5 + 1 + 10 + 18}{4}(13) = 110.5$$

As a ratio estimator, this estimator is not unbiased with the design used.

With a single starting point ( $n = 1$ ), the PPS estimator is identical to the ratio estimator. Hence the second method above would give an unbiased estimate if the systematic sample were selected with probability proportional to size. This could be accomplished by selecting one unit out of the 13 in the population at random and then including in the sample that unit and every third unit to the right and to the left.  $\square$

## 12.5. VARIANCE AND COST IN CLUSTER AND SYSTEMATIC SAMPLING

The effectiveness of cluster or systematic sampling depends both on the variance resulting from using primary units of a given size and shape and the cost of sampling such units. As a starting point, the variance of selecting  $n$  primary units may be compared 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  $n\bar{M}$ .

For the unbiased estimate of the population total based on a simple random sample of  $n\bar{M}$  secondary units, write  $\hat{\tau}_{\text{srs}} = M\bar{y}$ . The variance of this design-estimator combination is

$$\begin{aligned}\text{var}(\hat{\tau}_{\text{srs}}) &= M^2 \left( \frac{(N\bar{M} - n\bar{M})}{nN\bar{M}^2} \right) \sigma^2 \\ &= N^2 \left[ \frac{\bar{M}(N - n)}{nN} \right] \sigma^2\end{aligned}$$

where  $\sigma^2$  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 repeated systematic sample, with a simple random sample of  $n$  primary units, the unbiased estimator (see Eq. (12.1)) 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  $u$ . The label  $u$  identifies the size, shape, or arrangement of primary units, which could be, for example, square clusters, rectangular clusters, or systematic samples. The variance of this design-estimator combination is

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

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

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

Thus cluster (systematic) sampling is efficient if the variance  $\sigma_u^2$  between primary units is small relative to the overall population variance  $\sigma^2$ .

To estimate this relative efficiency using data from a cluster or systematic sampling design, the usual sample variance  $s^2$  cannot be used as an estimate of  $\sigma^2$ , because the data were not obtained with simple random sampling. Instead,  $\sigma^2$  can be estimated using analysis of variance of the cluster (systematic) sample data as follows.

For simplicity, suppose that each of  $N$  primary units has an equal number  $\bar{M}$  of secondary units. The total sum of squares in the population can be partitioned as

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



where  $\bar{y}_i = \sum_{j=1}^{\bar{M}} y_{ij} / \bar{M}$ . The first term on the right contains the within-primary-unit sum of squares and the second term the between-primary-unit sum of squares.

Write  $\sigma_w^2 = \sum_{i=1}^N \sum_{j=1}^{\bar{M}} (y_{ij} - \bar{y}_i)^2 / [N(\bar{M} - 1)]$  for the within-primary-unit variance and  $\sigma_b^2 = \sum_{i=1}^N (\bar{y}_i - \mu)^2 / (N - 1)$  for the variance between primary unit means. (Note that  $\sigma_u^2 = \bar{M}^2 \sigma_b^2$  is the variance between primary unit totals.) An unbiased estimate of  $\sigma_w^2$  using the random sample of clusters is  $s_w^2 = \sum_{i=1}^n \sum_{j=1}^{\bar{M}} (y_{ij} - \bar{y}_i)^2 / [n(\bar{M} - 1)]$ , and an unbiased estimate of  $\sigma_b^2$  is  $s_b^2 = \sum_{i=1}^n (\bar{y}_i - \hat{\mu})^2 / (n - 1)$ , where  $\hat{\mu} = \hat{\tau}_u / M$ . The sum-of-squares equality may be written

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

An unbiased estimate of  $\sigma^2$  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 sampling (simple random sample of  $n$  clusters) based on the data from the cluster sample is  $\hat{\sigma}^2 / \bar{M}s_b^2 = \bar{M}\hat{\sigma}^2 / s_u^2$ . Given a cluster sample with equal-sized clusters, one can thereby compare the efficiency of a variety of smaller units.

The variance of cluster or systematic sampling can alternatively be examined in terms of the correlation within primary units. The within-primary-unit correlation coefficient is defined as

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

The sum of squares in the primary unit variance  $\sigma_u^2$  may be written

$$\begin{aligned} \sum_{i=1}^N (y_i - \mu_1)^2 &= \sum_{i=1}^N \left( \sum_{j=1}^{\bar{M}} y_{ij} - \bar{M}\mu \right)^2 \\ &= \sum_{i=1}^N \left[ \sum_{j=1}^{\bar{M}} (y_{ij} - \mu) \right]^2 \\ &= \sum_{i=1}^N \left[ \sum_{j=1}^{\bar{M}} \sum_{j'=1}^{\bar{M}} (y_{ij} - \mu)(y_{ij'} - \mu) \right] \\ &= \sum_{i=1}^N \sum_{j=1}^{\bar{M}} (y_{ij} - \mu)^2 + \sum_{i=1}^N \sum_{j=1}^{\bar{M}} \sum_{j' \neq j} (y_{ij} - \mu)(y_{ij'} - \mu) \end{aligned}$$

Substituting into Equation (12.2), the variance with cluster sampling may be written

$$\begin{aligned}\text{var}(\hat{\tau}) &= \frac{N^2(N-n)}{nN} \left( \frac{M-1}{N-1} \right) \sigma^2 [1 + (\bar{M} - 1)\rho] \\ &\approx \frac{N^2\bar{M}(N-n)}{nN} \sigma^2 [1 + (\bar{M} - 1)\rho]\end{aligned}$$

If  $\rho$  is zero, the variance with cluster sampling will be approximately the same as the variance of a simple random sample of an equal number ( $n\bar{M}$ ) of secondary units. If  $\rho$  is greater than zero, the simple random sample will give lower variance. If  $\rho$  is less than zero, the cluster sample gives lower variance.

With many natural populations, units near each other tend to be similar, so with compact clusters,  $\rho$  is greater than zero. For such populations, the value of  $\rho$ , and hence the variance of  $\hat{\tau}$ , will tend to be larger with square clusters, in which the secondary units are close together, than with long, thin clusters, in which at least some of the secondary units are far apart. With systematic sampling, the secondary units of each primary unit are spaced relatively far apart, so that  $\rho$  may well be less than zero. For this reason, systematic sampling is inherently efficient with many real populations.

The advantage of cluster sampling is that it is often less costly to sample a collection of units in a cluster than to sample an equal number of secondary units selected at random from the population. Considering the case with equal-sized clusters, let  $\bar{M}_u$  be the number of secondary units in a primary unit of type  $u$ . Let  $\sigma_u^2$  be the population variance for that type of unit. Let  $c_u$  be the cost of measuring a randomly selected unit of that type, so that the cost of a sample of  $n_u$  units is  $c_u n_u$ . Ignoring the finite-population correction factor in Equation (12.2), the variance of an estimator  $\hat{\tau}_u = N\bar{y}_u$  of the population total is approximately  $\text{var}(\hat{\tau}_u) \approx N_u^2 \sigma_u^2 / n_u$ . For a fixed cost  $C$ ,  $n_u = C/c_u$  and the variance is  $\text{var}(\hat{\tau}_u) \approx N_u^2 \sigma_u^2 c_u / C = c_u \sigma_u^2 / (M_u^2 C)$ . The choice of primary unit giving the lowest variance is the one giving the smallest value of  $c_u \sigma_u^2 / \bar{M}_u^2$ . For specified variance, the primary unit giving the lowest cost is again the one giving the smallest value of  $c_u \sigma_u^2 / \bar{M}_u^2$  (see Cochran 1977, p. 234).

In principle, the ideal size and shape of primary unit can be determined by a variance function and a cost function, each depending on the size and shape of primary unit. Such functions are not necessarily simple in real sampling situations. Examples of such functions are discussed in Cochran (1977), Hansen et al. (1953), Jessen (1978), and Kish (1965).

## 12.6. COMPUTING NOTES

Selecting a cluster sample is generally a matter of organizing the population units into primary and secondary units, selecting a sample of primary units, and calculating the primary unit totals to work with.

A survey of sea otters in a coastal study region counts visible otters in their near-shore habitat. Because of the uneven shape of the suitable habitat in the study region, the region has been divided into primary and secondary spatial units. There are  $N = 16$  primary units, divided into a total of  $M = 33$  secondary units. A simple random sample of  $n = 4$  primary units is selected. The first of these has  $M_1 = 1$  secondary unit in it, in which  $y_1 = 3$  sea otters are observed. The second sample unit has  $M - 2 = 4$  secondary units, in which a total of  $y_2 = 24$  sea otters are observed. The third sample unit has  $M - 3 = 3$  and  $y_3 = 7$ , and the fourth sample unit has  $M - 4 = 2$  and  $y_4 = 2$ . For purposes of this example it is assumed that every sea otter in a sample unit is observed; imperfect detectability in such surveys is discussed in later chapters.

Below the population total is estimated using the expansion estimator (unbiased) and the ratio estimator.

```
# some calculations for cluster
# sampling example (sea otter survey)
# with n = 4, N = 16, M = 33
> ys <- c(3,24,7,2)
> Mi <- c(1,4,3,2)
> N <- 16
> M <- 33

# expansion estimator,
# unbiased if design was SRS:
> Tauhat <- N * mean(ys)
> Tauhat
[1] 144
> ssq <- var(ys)
> ssq
[1] 104.6667
> varhatTauhat <- N^2 * (1 - 4/N) * ssq / 4
> varhatTauhat
[1] 5024
> seTauhat <- sqrt(varhatTauhat)
> seTauhat
[1] 70.88018
> qt(.90,3)
[1] 1.637744
> Tauhat + qt(.9,3) * seTauhat
[1] 260.0836
> Tauhat - qt(.9,3) * seTauhat
[1] 27.91638
```

Had the sample of 4 primary units been selected with replacement with probability proportional to primary unit size, as measured by number of secondary units, the Hansen–Hurwitz estimator and the Horvitz–Thompson estimator would each be unbiased. Their calculation is carried out next.

```

# Hansen Hurwitz estimator,
# unbiased if design had been PPS:
> pi <- Mi/M
> pi
[1] 0.03030303 0.12121212 0.09090909 0.06060606
> Tauhatp <- mean(ys/pi)
> Tauhatp
[1] 101.75
> varhatTauhatp <- var(ys/pi) / 4
> varhatTauhatp
[1] 1217.562
> ys/pi
[1] 99 198 77 33
> var(ys/pi)
[1] 4870.25
> seTauhatp <- sqrt(varhatTauhatp)
> seTauhatp
[1] 34.89359
> qt(.9,3) * seTauhatp
[1] 57.14678
> 102+57
[1] 159
> 102-57
[1] 45

# Horvitz Thompson estimator:
> pii = 1 - (1- pi)^4
> pii
[1] 0.1158129 0.4036019 0.3169865 0.2212626
> sum(ys/pii)
[1] 116.4904
> ys/pii
[1] 25.903841 59.464533 22.082956 9.039032
>

```

A systematic sample can be selected in R as follows. Suppose we want to sample rainfall at a site every 10th day over a year. Thus,  $M = 365$  and we wish to select a 1 in  $k$  systematic sample, with  $k = 10$ . Note that a single primary unit is being selected, and primary units are of unequal size, since 365 does not divide evenly by 10.

```

M <- 365
k <- 10

start <- sample(1:k, 1)

s <- seq(start, M, k)

```

```
# The ratio estimate of mean daily rainfall during the year:
mean(y[s])

# The unbiased estimate of mean daily rainfall:
10 * sum(y[s]) / 365
```

## EXERCISES

1. Assume that the following are data from cluster sampling with simple random sampling of clusters. There are 10 clusters (primary units) and a total of 100 secondary units in the population. For each of the  $n = 3$  selected clusters,  $y_i$  is the cluster total for the variable of interest and  $M_i$  is cluster size:  $y_1 = 4$ ,  $M_1 = 5$ ;  $y_2 = 12$ ,  $M_2 = 20$ ;  $y_3 = 7$ ,  $M_3 = 10$ .
  - (a) Give an unbiased estimate of the population total.
  - (b) Estimate the variance of that estimator.
2. Using the data of Exercise 1 and assuming simple random sampling, (a) give the ratio-to-size estimate of the population total and (b) estimate the variance of that estimator.
3. Using the data of Exercise 1, but assuming that the sample was obtained with selection probabilities proportional to cluster size (PPS), with replacement, (a) give an unbiased estimate of the population total and (b) estimate the variance of that estimator.
4. Use random sampling of clusters to estimate the mean or total of a population of your choice. In the process of carrying out the survey and making the estimate, think about or discuss with others the following:
  - (a) What practical problems arise in establishing a frame, such as a map or list of units, from which to select the sample?
  - (b) How is the sample selection actually carried out?
  - (c) What special problems arise in observing the units selected?
  - (d) Estimate the population mean or total.
  - (e) Estimate the variance of the estimator above.
  - (f) Give a 95% confidence interval for the population mean or total.
  - (g) How would you improve the survey procedure if you were to do it again?
5. To estimate the number of typographical errors in a 65-page manuscript, a systematic sample of pages is selected by first selecting a random number between 1 and 10 and including in the sample that numbered page and every 10th page thereafter. The random number selected was 6. The number of

typographical errors on the sample pages were 1, 0, 2, 3, 0, and 1. Assume that no errors on sample pages were missed.

- (a) Give an unbiased estimate, under the design used, of the total number of errors in the manuscript. What design was used?
  - (b) The person doing the survey estimated the total number of errors in the manuscript by  $65(1 + 0 + 2 + 3 + 0 + 1)/6 = 75.83$ . Which estimator was used? Is it unbiased with the design used?
  - (c) The variance of the estimator was estimated by  $65(65 - 6)(1.37)/6$ , where 1.37 is the sample variance of the six error counts. Is this unbiased for the actual variance of the estimator of the total number of errors? Discuss.
6. Use repeated systematic sampling to estimate the mean or total of a population of your choice. In the process of carrying out the survey and making the estimate, think about or discuss with others the following:
- (a) What practical problems arise in establishing a frame, such as a map or list of units, from which to select the sample?
  - (b) How is the sample selection actually carried out?
  - (c) What special problems arise in observing the units selected?
  - (d) Estimate the population mean or total.
  - (e) Estimate the variance of the estimator above.
  - (f) Give a 95% confidence interval for the population mean or total.
  - (g) How would you improve the survey procedure if you were to do it again?
7. With a systematic sample having a single randomly selected starting point, the expansion estimator  $\hat{\tau} = N\bar{y}$  is design-unbiased for the population total  $\tau$ , but no design-unbiased estimator of  $\text{var}(\hat{\tau})$  exists. Assume for simplicity that each possible sample (primary unit) has the same number of secondary units, that is,  $M_i = m$  for all  $i$ . With the single starting point the sample has only  $n = 1$  primary units, and  $\bar{y}$ , the sample mean of the primary unit totals, is the sum of the  $m$  secondary units in the primary unit selected.  $N$  is the number of possible starting points (the number of possible systematic samples), so the total number of secondary units in the population is  $M = Nm$ . Now consider a model-based approach, in which the population  $Y$ -values are assumed to be independent, identically distributed random variables, each with mean  $\beta$  and variance  $\gamma$ . Under the model, show that  $\hat{\tau}$  is (model) unbiased for the population total, find the mean square prediction error  $E(\hat{\tau} - \tau)^2$ , and find an unbiased estimator of  $E(\hat{\tau} - \tau)^2$ . [Hint: The notation may be simplified to avoid double subscripts by relabeling the  $M = Nm$  secondary units in the population from 1 to  $M$ , with  $Y_j$  representing the  $Y$ -value of the  $j$ th secondary unit. The population total can be partitioned into sample and nonsample parts as  $\tau = \sum_{j=1}^M Y_j = \sum_{j \in s} Y_j + \sum_{j \notin s} Y_j$ .]