

## Cluster Sampling & Systematic Sampling (Chap. 12)

Recall that cluster sampling is where we first divide the population into “clusters,” then select a simple random sample (SRS) of these clusters, and sample every unit within the selected clusters. This is a two-stage sampling plan, where we employ an SRS of clusters at the first stage and a census within these selected clusters at the second stage.

Systematic sampling is where we first select a random starting point in the population, and then sample every  $m^{\text{th}}$  unit beginning at that starting point. This also is a two-stage sampling plan, where we employ an SRS of size 1 from the list of potential starting points, and then census the sampling units at multiples of  $m$  units from the initial unit.

- Cluster sampling and systematic random sampling, as defined, are special cases of two-stage random sampling plans which will be discussed in the next chapter. The population is first partitioned into mutually exclusive groups, known as primary units, each of which contains a number of sampling units, known as secondary units. Selection occurs on the *primary* units, and then *every* secondary unit within a selected primary unit is sampled.
- This should be clear for cluster sampling. With systematic sampling, the primary units are groups of observations  $m$  units apart in the population list. For example, if we want every 5th member of a population, then there are 5 primary units corresponding to these sets of secondary units: 1,6,11,...; 2,7,12,...; 3,8,13,...; 4,9,14,...; and 5,10,15,... In systematic sampling, we generally take an SRS of one of these primary units and then examine every secondary unit in the primary unit selected.

Main Idea: The important point for these two sampling plans is that whenever a primary unit is selected, all secondary units within are sampled. In truth then, the primary units are the sampling units in a cluster or systematic sample, even though measurements or observations are actually made on the secondary units.

Thompson lists three special considerations for these two sampling plans which warrant further discussion and separate consideration (pages 129 & 131).

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

After defining the relevant notation for these sampling plans, each of the special considerations above will be addressed by way of examples.

Notation: Consider taking a cluster sample from some population with response variable  $y$ . We let:

$$\begin{aligned}
 N &= \text{the number of primary units (clusters) in the population,} \\
 n &= \text{the number of primary units (clusters) in the sample,} \\
 M_i &= \text{the \# of secondary units in the } i^{\text{th}} \text{ primary unit,} \\
 M &= \text{the total \# of secondary units in the population} = \sum_{i=1}^N M_i \\
 y_{ij} &= \text{the } y\text{-value of the } j^{\text{th}} \text{ secondary unit in the } i^{\text{th}} \text{ primary unit,} \\
 y_i &= \sum_{j=1}^{M_i} y_{ij} = \text{the total of the } y\text{'s in the } i^{\text{th}} \text{ primary unit (cluster totals),} \\
 \tau &= \sum_{i=1}^N y_i = \text{the total of the } y\text{'s in all the units,} \\
 \mu &= \frac{\tau}{M} = \text{the mean of the } y\text{'s per secondary unit,} \\
 \mu_1 &= \frac{\tau}{N} = \text{the mean of the } y\text{'s per primary unit.}
 \end{aligned}$$

Example: A sociologist wants to estimate the average per capita income in a certain small city. As no list of resident adults is available, she decides that each of the city blocks will be considered one cluster. The clusters are numbered on a city map from 1 to 415, and the experimenter decides she has enough time and money to sample  $n = 25$  clusters where every household will be interviewed within the clusters (blocks) chosen. The data on the next page give the number of residents and the total income for each of the 25 blocks sampled. [Problem taken from Scheaffer, Mendenhall, & Ott, *Elementary Survey Sampling*, page 248.] Given that  $M = 2500$  residents, use these data to estimate the average per capita income in the city.

Notation:

$$\begin{aligned}
 N &= 415 \text{ blocks, } n = 25 \text{ blocks, } M = 2500 \text{ residents,} \\
 M_i &= \text{the number of residents in the } i^{\text{th}} \text{ block,} \\
 M &= \text{the total number of residents in all 415 blocks,} \\
 y_{ij} &= \text{the income of the } j^{\text{th}} \text{ resident in the } i^{\text{th}} \text{ block,} \\
 y_i &= \text{the total income of all residents on the } i^{\text{th}} \text{ block,} \\
 \tau &= \text{the total income of all residents in the city,}
 \end{aligned}$$

- $\mu$  = the mean income per resident,
- $\mu_1$  = the mean income per block.

Cluster $i$	Number of Residents, $M_i$	Total Income per Cluster, $y_i$	Cluster $i$	Number of Residents, $M_i$	Total Income per Cluster, $y_i$
1	8	\$192,000	14	10	\$98,000
2	12	\$242,000	15	9	\$106,000
3	4	\$84,000	16	3	\$100,000
4	5	\$130,000	17	6	\$64,000
5	6	\$104,000	18	5	\$44,000
6	6	\$80,000	19	5	\$90,000
7	7	\$150,000	20	4	\$74,000
8	5	\$130,000	21	6	\$102,000
9	8	\$90,000	22	8	\$60,000
10	3	\$100,000	23	7	\$78,000
11	2	\$170,000	24	3	\$94,000
12	6	\$86,000	25	8	\$82,000
13	5	\$108,000			
				$\sum_{i=1}^{25} M_i = 151$	$\sum_{i=1}^{25} y_i = \$2,658,000$

There are two basic ways to approach estimation of  $\tau$  and  $\mu$ :

- Unbiased estimation: treat the sample as an SRS of clusters (blocks in this example), each with response  $y_i$  and use the SRS formulas from Chap. 2. We ignore  $M_i$ , the cluster sizes. We can estimate  $\tau$  and  $\mu_1$  (mean per cluster) in this way. If we know  $M$ , we can also estimate  $\mu$ .
- Ratio estimation: If the primary unit total  $y_i$  is correlated with the cluster size  $M_i$  (such that we expect  $y_i = 0$  when  $M_i = 0$ ), then ratio estimators of  $\tau$  and  $\mu$  may be advantageous. Ratio estimators are biased but can have substantially smaller MSE than the unbiased estimators if there's a strong relationship between the cluster sizes and the cluster totals.

#### Unbiased estimation

- Estimate  $\mu_1$ , the mean total income per block, by  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{25} [2658000] = \underline{\$106,320}$ , with standard error

$$\text{SE}(\bar{y}) = \sqrt{\left(\frac{N-n}{N}\right) \frac{s_u^2}{n}} = \sqrt{\left(\frac{415-25}{415}\right) \frac{1898226667}{25}} = \underline{\$8,447},$$

where  $s_u^2 = \frac{1}{n-1} \sum_{i=1}^{25} (y_i - \bar{y})^2$  = the sample variance of the cluster totals ( $y_i$ 's).

- Estimate  $\tau$ , the total income for the whole city by  $N\bar{y} = 415(106320) = \underline{\$44,122,800}$ , with standard error

$$\text{SE}(\hat{\tau}) = \sqrt{N(N-n) \frac{s_u^2}{n}} = \sqrt{415(415-25) \frac{1898226667}{25}} = \underline{\$3,505,584}.$$

- If we know  $M$  (= 2500 here), we can estimate  $\mu$  (average income per resident) by

$$\begin{aligned} \hat{\mu} &= \frac{\hat{\tau}}{M} = \frac{N}{M} \bar{y} = \frac{44122800}{2500} = \underline{\$17,649}, \\ \text{SE}(\hat{\mu}) &= \text{SE}\left(\frac{\hat{\tau}}{M}\right) = \frac{1}{M} \text{SE}(\hat{\tau}) = \frac{1}{2500} (3505584) = \underline{\$1,402}. \end{aligned}$$

- To form confidence intervals with the above estimators, we would multiply the SE by a  $t$  value with  $n-1$  degrees of freedom (24 df in this example).

### Ratio Estimation

With cluster sampling, we have an auxiliary variable  $M_i$  (the number of residents on each block), so we may be able to take advantage of this to improve the SRS-based estimates of  $\tau$  and  $\mu$  by using ratio estimators (Chapter 7).

- Estimate  $\mu$  by the sample ratio  $r$ :

$$\hat{\mu}_r = r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} = \frac{\text{Sample total income}}{\text{Sample total \# of residents}}$$

For this example,  $\hat{\mu}_r = 2658000/151 = \underline{\$17,603}$ . The standard error is

$$\text{SE}(\hat{\mu}_r) = \sqrt{\left(\frac{N-n}{N\mu_x^2}\right) \frac{s_r^2}{n}}, \text{ where } s_r^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rM_i)^2$$

where  $\mu_x = \bar{M} = M/N$  is the the mean cluster size for the population. If  $\bar{M}$  is unknown, then we can substitute  $\bar{m}$ = mean cluster size for the sample. For this example, where  $\bar{M} = 2500/415 = 6.24$ , we get  $\text{SE}(\hat{\mu}_r) = \underline{\$1,621}$  (see R code at end of handout for calculation).

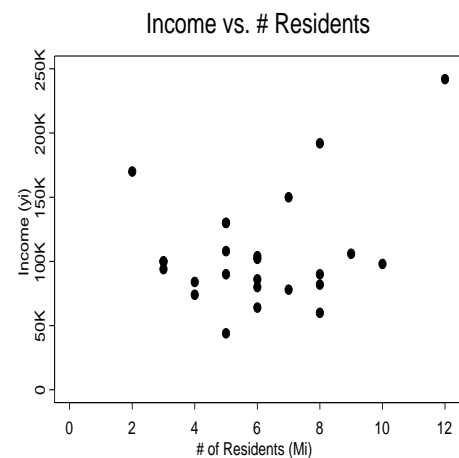
- If we know  $M$ , then the ratio estimator of the population total  $\tau$  is  $\hat{\tau}_r = Mr$  with  $SE(\hat{\tau}_r) = M SE(\hat{\mu}_r)$ . Noting that  $\mu_x = M/N$  in the expression for  $SE(\hat{\mu}_r)$ , it follows that

$$SE(\hat{\tau}_r) = \sqrt{N(N-n) \frac{s_r^2}{n}}.$$

In our example,  $\hat{\tau}_r = 2500(1763) = \underline{\$44,006,623}$  (total income for the city) with  $SE(\hat{\tau}_r) = 2500(1621) = \underline{\$4,053,522}$ .

- We can also estimate  $\mu_1$ , the mean per cluster (block), by the ratio estimator  $\hat{\mu}_{1r} = \hat{\tau}_r/N$  with  $SE(\hat{\mu}_{1r}) = SE(\hat{\tau}_r)/N$ . For our example,  $\hat{\mu}_{1r} = 44,006,000/415 = \underline{\$106,040}$  with  $SE(\hat{\mu}_{1r}) = 4053522/415 = \underline{\$9,768}$ .
- The variances of the ratio estimators are based on delta method approximations. We can also use bootstrapping to obtain standard errors, as we showed in the notes on bootstrapping.
- As with the unbiased estimators, we would form confidence intervals by multiplying the SE's by a  $t$  value with  $n - 1$  degrees of freedom.
- Ratio estimators are biased. The bias is typically negligible, and so we compare the variances of these estimators to the variances of the unbiased estimators. We can choose whichever one gives smaller variance.

Note that the ratio estimates are very similar to the unbiased estimates for the income example, but that the standard errors are higher. Why do you think this happened?



In summary, for cluster sampling, we have two basic options:

1. Work only with the primary units (clusters) and use the unbiased estimators.
2. Use ratio estimation to make use of the relationship between cluster size and cluster totals, if such a relationship exists. In many examples, we would expect such a relationship, particularly if the clusters vary greatly in size. Note that if the clusters are all the same size, then the unbiased and ratio estimators are identical.

The true (theoretical) variances of the unbiased estimators all depend ultimately on  $\sigma_u^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu_1)^2$ , the variability between cluster totals, while the true variances of the ratio estimators depend on  $\sigma_r^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - M_i \mu)^2$ , the variability between cluster totals, accounting for cluster size.

## PPS Sampling

Suppose in cluster sampling that the primary units are drawn with replacement with selection probabilities proportional to the sizes of the primary units (i.e.: larger clusters are more likely to be selected than smaller clusters). In the income example, sampling of blocks would be with probabilities proportional to the number of residents on the blocks.

Recall that for PPS sampling with replacement, an unbiased estimator of the population total is given by the Hansen-Hurwitz estimator:

$$\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} = \frac{M}{n} \sum_{i=1}^n \frac{y_i}{M_i}, \quad \text{where: } p_i = \frac{M_i}{M}$$

with variance given by:

$$\begin{aligned} \text{Var}(\hat{\tau}_p) &= \frac{1}{n} \sum_{i=1}^N p_i \left( \frac{y_i}{p_i} - \tau \right)^2 = \frac{1}{n} \sum_{i=1}^N \frac{M_i}{M} \left( \frac{y_i}{M_i/M} - \tau \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^N \frac{M_i}{M} \left[ M \left( \frac{y_i}{M_i} - \mu \right) \right]^2 = \frac{M}{n} \sum_{i=1}^N M_i (\bar{y}_i - \mu)^2 \\ &= \frac{M}{n} \sum_{i=1}^N \frac{1}{M_i} (y_i - \mu M_i)^2 \end{aligned}$$

where  $\bar{y}_i = y_i/M_i$  is the average per secondary unit in cluster  $i$  (e.g., average income per resident in block  $i$ ).

- An unbiased estimator of  $\text{Var}(\hat{\tau}_p)$  given above is

$$\widehat{\text{Var}}(\hat{\tau}_p) = \frac{M^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_p)^2, \quad \text{where } \hat{\mu}_p = \hat{\tau}_p/M.$$

- The theoretical variance of the PPS estimator is roughly the same as for the ratio estimator, but the PPS estimator is unbiased. Both have low variance when the cluster total is proportional to cluster size. In the income example, this would be when block income is proportional to block size.
- A Horvitz-Thompson estimator based on the selection probabilities  $p_i = M_i/M$  can also be developed, as indicated on page 134 of the book.

## Systematic Sampling

Recall that systematic sampling is the special case of cluster sampling where each cluster is determined through some random starting point, and a single cluster is chosen.

Example: Suppose we are estimating visitation to a particular site and will count the total number of visitors every fifth day for 30 days, starting at a random day from 1 to 5. This gives the following five clusters:

- Every sampling unit is in exactly one cluster.
- In systematic sampling, experimenters generally select ONE cluster from this group. This gives a sample size of 1, which prohibits any type of inference. Why?

Note: in the above example, the population (number of days) is finite. If we sample at points in time (for example, recording the temperature every 10 minutes from 8am to 8pm, starting at a random time from 0 to 10 minutes after 8am), then the population size is infinite.

- There are two basic outlooks one takes toward this “problem”:
  1. Take more than one systematic sample (replication) and use the unbiased or ratio estimators as outlined above for cluster sampling (identical if the clusters are all the same size, as they often are in systematic sampling).
  2. Assume the variation from one systematic sample to another is not greater than (and probably less than) the variation from one SRS to another. Then use SRS formulas, assuming that they are conservative. In other words, we assume a systematic sample is likely to be more “representative” of the population (and give more accurate estimates of population parameters) than an SRS. This is by far the most common approach in practice. Systematic samples over space (e.g., evenly spaced plots along a transect) or time are generally treated as SRS’s. So are items or people selected systematically from a list.

In the following example, a systematic sample is treated as an SRS, but has an additional wrinkle.

Example: Consider sampling via line transects from an irregularly-shaped area. Interest is in estimating the proportion of the area which has some attribute (say, bare ground).

- Suppose we take a systematic random sample of 10 parallel transects, where the initial starting point is randomly chosen along a baseline, and the resulting 10 transects are evenly spaced along the baseline.
- Because the area is irregular in shape, the transects will be of different lengths.
- Although we really have a cluster sample of size 1, we will assume that the 10 transects are representative of the area and treat them as an SRS of size 10. However, we could also select multiple random starting points and do several systematic samples, perhaps with fewer transects per sample.
- So, the sampling unit here is a transect. What is the population?

How might we estimate the proportion of the area (using these 10 transects) with the desired attribute (bare ground)? Three ways are considered:

1. Simple Average of Transect Proportions: We could measure the proportion of each transect with the attribute, and average the 10 resulting proportions. Problem?
2. Ratio Estimator: Suppose we let  $y_i$  = the length of transect  $i$  with the attribute,  
 $x_i$  = the length of transect  $i$ .

Then a ratio estimate of the proportion of the area with the attribute is:

$$r = \frac{\sum_{i=1}^{10} y_i}{\sum_{i=1}^{10} x_i}, \text{ with } \widehat{\text{Var}}(r) = \underbrace{\left(\frac{N-n}{N}\right)}_{=1 \text{ here}} \frac{1}{\mu_x^2} s_r^2, \text{ where: } s_r^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2.$$

- As a side note, we might estimate  $\mu_x$  with  $\bar{x}$  (the average length of the 10 transects), but we can actually determine  $\mu_x$  if we know the area  $A$  of the region and the length  $b$  of the baseline – then  $\mu_x = A/b$ .
- This is a perfectly valid way of estimating the proportion of the area with the attribute, although, as with all ratio estimators, the estimator is biased.

3. Unbiased Estimator: Consider only the lengths ( $y_i$ 's) of bare ground on the sampled transects and not their lengths.

- We can determine the true mean length of a transect  $\mu_x$  since we know the total area and the length along the baseline of the region. Then an unbiased estimator of the proportion of the area which is bare ground is

$$\frac{\bar{y}}{\mu_x} = \frac{\text{average length of the attribute among the transects}}{\text{true mean length of a transect}}.$$

Since  $\text{Var}(\bar{y}/\mu_x) = (1/\mu_x^2)\text{Var}(\bar{y})$ , the SE of this estimator is  $(1/\mu_x)s/\sqrt{n}$  where  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ .

- If there is large variation in transect lengths, ratio estimation will likely do better than the unbiased estimator.

Basic Principle of Cluster and Systematic Sampling: Because a census is taken on all secondary units within a primary unit, the within-primary-unit variance plays no role in the variances of population means or totals. This explains why the ideal case for cluster or systematic sampling is that when there is large variability within primary units relative to the variability between primary units, because this large within-primary-unit variability will have no effect on the variance of estimators!

Hence, we want the primary units to be “mini-populations”; that is, we want them to be representative of the population as a whole. This will minimize differences between primary units while maximizing differences within.

## R Code for Cluster Sampling Income Example

```
> N <- 415; n <- 25; fpc <- (N-n)/N; M <- 2500
> Mi <- c(8,12,4,5,6,6,7,5,8,3,2,6,5,10,9,3,6,5,5,4,6,8,7,3,8)
> y <- 1000*c(192,242,84,130,104,80,150,130,90,100,170,86,
             108,98,106,100,64,44,90,74,102,60,78,94,82)
> # Unbiased Estimators for Cluster Sampling
> # =====
> ybar1 <- mean(y)
> ybar1          # Estimated average income per block
[1] 106320
> su2 <- var(y)
> se.ybar1 <- sqrt(fpc*su2/n)
> se.ybar1      # SE of average income per block
[1] 8447.19
> tauhat <- N*ybar1
> tauhat        # Estimated total income in city
[1] 44122800
> se.tauhat <- N*se.ybar1
> se.tauhat     # SE of estimated total income
[1] 3505584
> muhat <- tauhat/M
> muhat        # Estimated average income per resident
[1] 17649.12
> se.muhat <- se.tauhat/M
> se.muhat     # SE of average income per resident
[1] 1402.234

> # Ratio Estimators with Cluster Sampling
> # =====
> r <- sum(y)/sum(Mi)
> r            # Sample ratio = estimated average income per resident
[1] 17602.65
> mux <- M/N
> sr2 <- (1/(n-1))*sum((y-r*Mi)^2)
> se.r <- sqrt(fpc*sr2/(n*mux^2))
> se.r        # SE of average income per resident
[1] 1621.409          # (assuming M = 2500 is KNOWN)
> xbar <- mean(Mi)
> se.r2 <- sqrt(fpc*sr2/(n*xbar^2))
```

```
> se.r2      # SE of average income per resident
[1] 1617.140 # (if M were UNKNOWN)
> tauhat <- M*r
> tauhat      # Estimated total income in city
[1] 44006623
> se.tauhat <- M*se.r
> se.tauhat      # SE of estimated total income
[1] 4053522
> muhat1 <- tauhat/N
> muhat1      # Estimated income per block
[1] 106040.1
> se.muhat1 <- se.tauhat/N
> se.muhat1      # SE of estimated income per block
[1] 9767.524
```