4 Systematic sampling

4.1 Introduction and Description

The systematic sampling technique is operationally more convenient than the simple random sampling. It also ensures at the same time that each unit has equal probability of inclusion in the sample. In this method of sampling, the first unit is selected with the help of random numbers and the remaining units are selected automatically according to a predetermined pattern. This method is known as systematic sampling.

(a) Suppose the N units in the population are numbered 1 to N in some order.

(b) Suppose further that N is expressible as a product of two integers n and k, so that $N = nk$.

(c) To draw a sample of size n, select a random number between 1 and k. Suppose it is r. Select the first unit whose serial number is r.

(d) This first unit is called a **random start.** Select every k^{th} unit after i^{th} unit.

(e) Sample will contain $r, r + k, r + 2k, ..., r + (n - 1)\, k$ serial number units.

(f) So first unit is selected at random and other units are selected systematically. This systematic sample is called k^{th} systematic sample and k is termed as **sampling interval**.

(g) This is also known as linear systematic sampling.

Example 4.1. Let $N = 50$ and $n = 5$. So $k = 10$. Suppose first selected number between 1 and 10 is 3.

Solution 5. Then systematic sample consists of units with following serial number 3, 13, 23, 33, 43.

Example 4.2. An insurance company's claims, in dollars, for one day are 400,600,570,960, 780, 800, 460, 650,440, 530, 470, 810, 625, 510, and 700. List all possible systematic samples of size 3, that can be drawn from this set of claims using linear systematic sampling. Also, obtain corresponding sample means.

Solution 6. Here the population size $N = 15$, and the size of the sample to be selected is $n = 3$. The sampling interval k will thus be $15/3 = 5$. The random number r to be selected from 1 to k can, therefore, take any value in the closed interval $[1, 5]$. Each random start from 1 to 5 will yield corresponding systematic sample. In all, there will be $k = 5$ possible samples. These are given below in Table 9 along with their means.

Table 9: Systematic sample

Random start	Serial number of sample units	y values of sampled units	Sample mean
1	1,6,11	400,800,470	556.67
2	2,7,12	600,460,810	623.33
3	3,8,13	570,650,625	615.00
4	4,9,14	960,440,510	636.67
5	5,10,15	780,530,700	670.00

4.1.1 Advantages of systematic sampling:

(a) It is easier to draw a sample and often easier to execute it without mistakes. This is more advantageous when the drawing is done in fields and offices as there may be substantial saving in time.

(b) The cost is low and the selection of units is simple. Much less training is needed for surveyors to collect units through systematic sampling.

(c) The systematic sample is spread more evenly over the population. So no large part will fail to be represented in the sample. The sample is evenly spread and cross section is better. Systematic sampling fails in case of too many blanks.

4.2 Estimation of Population Mean, Variance and Total

4.2.1 Estimation of population mean

When $N = nk$. Let y_{ij} be observation on the unit bearing the serial number $i + (j-1)\,k$ in the population, $i = 1, 2, ..., k$, $j = 1, 2, ..., n$. Suppose the drawn random number is $i \leq k$. Sample consists of ith column (in earlier table).

Consider the sample mean given by; $\bar{y}_{sys} = \bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$ as an estimator of the population mean given by $\bar{Y} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{nk}\sum_{i=1}^{k} \bar{y}_i$. The probability of selecting i^{th} column as systematic sample $= \frac{1}{k}$. So, $E\left(\bar{y}_{sys}\right) = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i = \bar{Y}$. Therefore, $\bar{y}_{sys}$ is an unbiased estimator of $\bar{Y}$. Further, $Var\left(\bar{y}_{sy}\right) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}_i - \bar{Y}\right)^2$.

Consider $(N-1)\,S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{Y}\right)^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{n}\left[\left(y_{ij} - \bar{y}_i\right) + \left(\bar{y}_i - \bar{Y}\right)\right]^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2 + n\sum_{i=1}^{k}\left(\bar{y}_i - \bar{Y}\right)^2$

$= k\,(n-1)\,S_{wsy}^2 + n\sum_{i=1}^{k}\left(\bar{y}_i - \bar{Y}\right)^2$

where $S_{wsy}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$ is the variation among the units that lie within the same systematic sample.

Thus;

$Var\left(\bar{y}_{sys}\right) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S^2_{wsy}$

$= \frac{N-1}{N}S^2 - \frac{n-1}{n}S^2_{wsy}$ where $\frac{N-1}{N}S^2$ is the variation as a whole while $\frac{n-1}{n}S^2_{wsy}$ is the pooled within variation of the k^{th} systematic sample with $N = nk$.

This expression indicates that when the within variation is large, then $Var\left(\bar{y}_i\right)$ becomes smaller. Thus higher heterogeneity makes the estimator more efficient and higher heterogeneity is well expected in systematic sample.

Example 4.3. On a particular day, 162 boats had gone to sea from the coast for fishing. It was desired to estimate the total catch of fish at the end of the day. As it was not possible to weigh the catch for all the 162 boats, it was decided to weigh fish for only 15 boats selected using circular systematic sampling. Discuss the selection procedure, and obtain the estimate of total catch of fish using data on the 15 sample boats given in the table below.

Table: Catch of fish (in quintals) for 15 selected boats

Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish
73	5.614	128	9.225	21	8.460
84	8.202	139	6.640	32	10.850
95	6.115	150	7.350	43	6.970
106	9.765	161	5.843	54	5.524
117	8.550	10	6.875	65	7.847

Solution 7. In this case, we have $N = 162$ and $n = 15$. Since $N/n = 162/15 = 10.8$ is not a whole number, the value of sampling interval k is taken as 11, an integer nearest to 10.8, and circular systematic sampling is used for selection of boats. If the selected random number r, $1 \leq \mathrm{r} \leq 162$, is 73 , then the boats bearing serial numbers $73, 84, \ldots, 65$ will be included in the sample. The serial numbers of selected boats, along with the corresponding catch of fish, are presented in table 6.4. We now proceed to estimate the total catch of fish using (6.1). This estimate is

$$\begin{aligned}\hat{\mathrm{Y}}_{\mathrm{sy}} = \mathrm{N}\overline{\mathrm{y}}_{\mathrm{sy}} &= \frac{\mathrm{N}}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{y_i} \\ &= \frac{162}{15}(5.614 + 8.202 + \ldots + 7.847) \\ &= \frac{(162)(113.83)}{15} \\ &= 1229.364\end{aligned}$$

The estimate of variance $\mathrm{V}\left(\hat{\mathrm{Y}}_{\mathrm{sy}}\right)$ is then computed by using the expression (6.4). Thus,

$$\begin{aligned} v\left(\hat{\mathrm{Y}}_{\mathrm{sy}}\right) &= \mathrm{N}^2 \mathrm{v}\left(\bar{y}_{\mathrm{sy}}\right) = \frac{\mathrm{N}(\mathrm{N}-\mathrm{n})}{2\mathrm{n}(\mathrm{n}-1)} \sum_{\mathrm{i}=1}^{\mathrm{n}-1} \left(\mathrm{y}_{\mathrm{i}+1} - \mathrm{y}_{\mathrm{i}}\right)^2 \\ &= \frac{162(162-15)}{2(15)(14)} \left[(8.202-5.614)^2 + (6.115-8.202)^2 + \ldots + (7.847-5.524)^2\right] \\ &= \frac{(162)(162-15)(67.596)}{2(15)(14)} \\ &= 3832.693 \end{aligned}$$

The confidence interval, for the total catch of fish for 162 boats, can then be calculated from

$$\begin{aligned} &\hat{Y}_{\mathrm{sy}} \pm 2\sqrt{\mathrm{v}\left(\hat{\mathrm{Y}}_{\mathrm{sy}}\right)} \\ =&1229.364 \pm 2\sqrt{3832.693} \\ =&1105.547, 1353.181 \end{aligned}$$

Thus, the estimate of total catch of fish obtained from a single sample is 1229.364 quintals. The confidence limits, obtained above, indicate that the total catch from all the 162 boats is likely to fall in the interval [1105.547, 1353.181] quintals

Example 4.4. A population is comprised of 6 households with respective sizes $2, 4, 3, 9, 1$ and 2 (the size x_k of household k is the number of people included). We select 3 households without replacement, with a probability proportional to its size.

(a) Give, in fractional form, the inclusion probabilities of the 6 households in the sampling frame (note: we may recalculate certain probabilities).

(b) Carry out the sampling using a systematic method.

(c) Using the sample obtained in 2., give an estimation for the mean size $\bar{X}$ of households; was the result predictable?

Solution 8. (a) For all $k : \pi_k = 3\frac{x_k}{X}$, with $X = 21$ Therefore

$$\pi_k = \frac{x_k}{7}, k \in U.$$

A problem arises for unit 4 because $\pi_4 > 1$. We assign the value 1 to π_4 and for the other units we recalculate the $\pi_k, k \neq 4$, according to:

$$\pi_k = 2\frac{x_k}{X-9} = 2\frac{x_k}{12} = \frac{x_k}{6}.$$

Finally, the inclusion probabilities are presented in Table 9a. We can verify that

$$\sum_{k=1}^{6} \pi_k = 3$$

k	1	2	3	4	5	6
π_k	$1/3$	$2/3$	$1/2$	1	$1/6$	$1/3$

(b) We select a random number between 0 and 1, and we are interested in the cumulative probabilities presented in Table 10. We advance in this list using a sampling interval of 1. In each case, we obtain in fine three distinct individuals (including household 4).

Table 10: Cumulative inclusion probabilities

k	1	2	3	4	5	6
$\sum_{j \leq k}$	1/3	1	1 1/2	572	2 2/3	3

(c) We have

$\hat{\bar{X}} = \frac{1}{6} \sum_{k \epsilon S} \frac{x_k}{\pi_k} [\frac{xk_1}{1} + \frac{x_k 2}{xk_2/6} + \frac{xk_3}{x_k 3/6}]$ with $k_1 = 4$ (household 4 is definitely chosen) and k_2 and k_3 being the other two selected households

$\hat{\bar{X}} = \frac{1}{6}[9 + 6 + 6] = 3.5 = \bar{X}$

This result was obvious, as x_k and π_k are perfectly proportional, by construct (we have a null variance, thus a 'perfect' estimator for the estimation of the mean size $\bar{X}$).

4.3 Exercises

(a) Describe the circular systematic sample.

(b) Show how to estimate the mean in systematic sampling when $n \neq k$.

(c) A census was conducted in a community. In addition to obtaining the usual population information the surveys questioned the occupants of every 20^{th} household to determine how long they have occupied their present homes. The results are summarized as follows; $n = 115$, $\sum y_i^2 = 2011.15$, $\sum \bar{y}_i = 407.1$, $N = 2300$, $k = 20$. Use this results to estimate the average amount of time people have lived in their present homes and place a bound on the error of estimation.

(d) Out of 24 villages in an area, two linear systematic samples of 4 villages each were selected. The total area under wheat is given in the table below.

i. Estimate the total area under wheat.

ii. Estimate the variance of the sample mean and place an upper bound on the error of estimation.

(e) In a small municipality, we listed six businesses for which total sales (variable xk) are respectively 40, 10, 8, 1, 0.5 and 0.5 million Euros. With the aim of estimating total paid employment, select three businesses at random and without replacement, with unequal probabilities according to total sales, using systematic sampling (by justifying your process). To do this, we use the following result for a uniform random variable between [0, 1]: 0.83021. What happens if we modify the order of the list?

(f) Consider a population of 5 units. We want to select using systematic sampling with unequal probabilities a sample of two units with inclusion probabilities proportional to the following values of X_i,

$1, 1, 6, 6, 6$.

i. Calculate the first-order inclusion probabilities.

ii. Considering the two units where the value of X_i is 1, calculate their second-order inclusion probabilities for every possible permutation of the list. What is the outcome?