Sampling Theory and Methods

Spring 2008

C. L. Williams

Chapter 4 Systematic Random Sampling

Outline

Systematic Sampling

In the previous chapter,

- simple random sampling of elements and discussed its importance as being "conceptually" the simplest kind of sampling -individuals are equally likely to be selected.
- problems associated with simple random sampling, including difficulties that may preclude drawing a sample by that method.

Systematic sampling is widely used in practice because it is easy to apply and can be easily taught to individuals who have little training in survey methodology.

- Systematic sampling, either by itself or in combination with some other method, may be the most widely used method of sampling.

Let's start with an example

Illustrative Example.

Let us suppose that as part of a cost-containment and quality-of-care review program, a sample of inpatient medical records is selected on an ongoing basis for a detailed audit. The total number of records in the population is not likely to be known in advance of the sampling since the records are to be sampled on an ongoing basis, and so, it would not be possible to use simple random sampling to choose the records. However, it may be possible to guess the approximate number of records that would be available for selection per time period and to select a sample of one in every k records as they become available, where k is an integer having a particular value chosen to meet the requirements of the study.

For example, suppose it is anticipated that there will be available ten new discharge records per day and that a total sample of 300 records per year is desired. Then the total number of records available per year is estimated to be $10 \times 365 = 3650$. To obtain something in the neighborhood of 300 records per year in the sample, k should be the largest integer in the quotient $3650/300$. Since the value of the quotient is 12.17, k would be equal to 12. This value of k is known as the *sampling interval*. Thus, we would take a sample of 1 from every 12 records.

One way to implement this procedure is to identify each record as it is created with a consecutive number beginning with 1. At the beginning of the study, a random number between 1 and 12 is chosen as the starting point. Then, that record and every twelfth record beyond it would be chosen. For example, if the random number is 4, then the records chosen in the sample would be 4, 16, 28, 40, 52, and so on.

Generalization

To generalize, a systematic sample is taken by first determining the desired sampling interval k, choosing a random number j between 1 and k, and select ing the elements labeled j, $j + k$, $j + 2k$, $j + 3k$ . . . Note that the *sampling fraction* for such a survey is $\frac{1}{k}$.

Slight modification

Decimal system

A slight modification of this method based on the decimal system is especially useful if the elements to be sampled are already numbered consecutively and if the actual drawing of the sample is to be done by unskilled personnel. If a sample of *1* in k units is specified (where k is, for example, a two-digit number), we may select a random two-digit number between 01 and k. If the number selected is j, then the two-digit numbers$j$, $j + k$, $j + 2k$, $j + 3k$ . . . and so on, are selected until a three-digit number is reached. All elements ending in the two- digit numbers selected are then included in the sample.

Sampling interval k=12

For example, if $k = 12$, a random two-digit number between 01 and 12 is chosen (e.g., 07). The two- digit sample numbers are then determined (e.g., 07, 19, 31, 43, 55, 67, 79, and 91), and all records ending with these two digits would be included in the sample (e.g., 07, 19, 31, 43, 55, 67, 79, 91, 107, 119, 131, 143, 155, 167, 179, 191, 207, 219, 231, etc.). This procedure is sometimes easier to use than the first procedure mentioned, especially if the individuals drawing the sample are unskilled, since they could be instructed to pull only those records ending with the specified digits.

Total Estimates

$$
\begin{aligned}
\widehat{x'_t} &= \frac{N \sum_{i=1}^{n} x_i}{n} \\
\widehat{Var(\widehat{x'_t})} &= N^2 \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right) \\
\widehat{SE(\widehat{x'_t})} &= N\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)
\end{aligned}
$$

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\widehat{Var(\overline{x})} &= \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right) \\
\widehat{SE(\overline{x})} &= \sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)
\end{aligned}
$$

$$
\begin{aligned}
p_y &= \frac{\sum_{i=1}^{n} y_i}{n} \\
\widehat{Var\left(p_y\right)} &= \left(\frac{N-n}{N}\right)\frac{p_y\left(1-p_y\right)}{n-1} \\
\widehat{SE\left(p_y\right)} &= \sqrt{\left(\frac{N-n}{N}\right)}\sqrt{\frac{p_y\left(1-p_y\right)}{n-1}}
\end{aligned}
$$

Table: Possible Samples of One in Six Physicians Chosen (In text Table 4.4) (from Table 2.1)

Random Number Chosen	(Physicians in Sample)	Estimated Mean	(Total)	[Proportic
1	(1, 7, 13, 19, 25)	12.0	(300)	[.8]
2	(2, 8, 14, 20)	1.0	(25)	[.25]
3	(3, 9, 15, 21)	4.25	(106.25)	[.75]
4	(4, 10, 16, 22)	6.5	(162.5)	[.5]
5	(5,11,17,23)	3.5	(87.5)	[.5]
6	(6, 12, 18, 24)	1.5	(37.5)	[.5]

$$
\begin{aligned}
E(\overline{x}) &= \frac{1}{6}(12 + 1 + 4.25 + 6.5 + 3.5 + 1.5) = 4.79 \neq \overline{X} = 5.08. \\
E(x_t') &= \frac{1}{6}(300 + 25 + 106.25 + 162.5 + 87.5 + 37.5) = 119.79 \\
\neq X_T &= 127 \\
E(p_y) &= \frac{1}{6}(8 + .25 + .75 + .5 + .5 + .5) = .55 \neq P_y = .56
\end{aligned}
$$

Thus we see that in this instance systematic sampling does not lead to unbiased estimates.

The reason the estimates in the preceding example are not unbiased is that although each element has the same chance (e.g., $1/k$) of being selected, the impact made on the estimates is not the same for each element. Physician 1, for example, has less impact than physician 2 since physician 1 appears with four other physicians in the sample, whereas physician 2 appears in the sample with only three other physicians. Thus, physician 1's measurements are diluted more than those of physician 2 in obtaining the estimate. In the previous example, when N/k was an integer, each physician appeared with the same number of other physicians in the sample, and all physicians had the same impact on the resulting estimates. If the number of elements N in the population is large, the biases in the estimates from systematic samples will, in general, be quite small and will be of little concern. A slight modification of the method of choosing the initial random number, however, will result in estimates that are unbiased. This modification will be discussed later in this chapter (Section 4.5).

Variance of the Estimates

Let us now consider the variances of estimates from systematic sampling. As discussed earlier, the variance of an estimate is important because it is a measure of the reliability of the estimate. In the discussion of variances of estimates from systematic sampling, let us assume for simplicity that N/k is an integer denoted by n. Systematic sampling of one in k elements would then yield a total of k possible samples, each containing N/k elements. The possible samples are shown in Table 4.5. In examining the systematic samples shown in Table 4.5, we see that each sample is a "cluster" of n (equal to N/k) elements with the elements being k "units" apart from each other. Thus, systematic sampling is operationally equivalent to grouping the N elements into k clusters, each containing N/k elements that are k units apart on the list, and then taking a random sample of one of these clusters. To illustrate this idea, let us look at an example.

Illustrative Example

Let us suppose that a systematic sample of one in five workers from Table 3.8 is taken. For purposes of illustration the 40 workers in Table 3.8 will be considered to be a population rather than a sample from a larger population. Then $N = 40$, $k = 5$, and $N/k = n = 40/5 = 8$. The five clusters defined by the sampling design are shown in Table 4.6. These clusters represent the five possible samples of one in five workers chosen from the list in Table 3.8.

Table: Possible Samples of 1 in k Elements (N/k Is an Integer) (Table 4.5 in text)

Random Number Chosen	Elements in Sample	Value of Variable
1	1,1+k, 1+2k,. . ., 1+(n-1)k	$X_1, X_{1+k}, X_{1+2k} \ldots X_{1+(n-1)k}$
2	2,2+k2+2k,. . .,2+(n-1)k	$X_2, X_{2+k}, \ldots, X_{2+(n-1)k}$
. . .	. . .	. . .
j	j,j+k,j+2k,j+(n-1)k	$X_j, X_{j+k}, X_{j+2k}, \ldots X_{j+(n-1)k}$
k	k,2k,3k,. . . nk	$X_k, X_{k+1}, X_{k+2}, \ldots X_{k+(n-1)k}$

The variance of estimates from systematic sampling can be understood more easily if we label the elements with double subscripts to denote the particular cluster. For example, the elements in the first cluster, elements 1, $1 + k$, $1 + 2k \ldots 1 + (n - 1)k$ would be relabeled with double subscripts as follows:

Original Label	New Label	Value of Variable
1	1,1	X_{11}
1+k	1,2	X_{12}
1+2k	1,3	X_{13}
1+(n-1)k	1,n	X_{1n}

Original Label	New Label	Value of Variable
j	j,1	X_{j1}
j+k	j,2	X_{j2}
j+2k	j,3	X_{j3}
j+(n-1)k	j,n	X_{jn}

Variance of Estimates

Total, x_t'

$$Var(x_t') = \left(\frac{N^2\sigma_x^2}{n}\right)[1+\delta_x(n-1)]$$

Mean, $\overline{x}$

$$Var(\overline{x}) = \left(\frac{\sigma_x^2}{n}\right)[1+\delta_x(n-1)]$$

Proportion, p

$$Var(p) = \left(\frac{P_y(1-P_y)}{n}\right)[1+\delta_x(n-1)]$$

$$\delta_x = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l<j}(X_{ij}-\overline{X})(X_{il}-\overline{X})}{nk(n-1)\sigma_x^2}$$

Table: 4.6 Cluster Samples Based on Data of 3.8

Cluster	Workers in Cluster	Forced Vital Capacity of Workers in Cluster
1	1, 6, 11, 16, 21, 26, 31, 36	81, 97, 71, 76, 70, 96, 84,69
2	2, 7, 12, 17, 22, 27, 32, 37	64, 82, 88, 62, 64, 62, 89, 80
3	3, 8, 13, 18, 23, 28, 33, 38	85, 99, 84, 67, 72, 67, 89, 98
4	4, 9, 14, 19, 24, 29, 34, 39	91, 96, 85, 91, 72, 95, 65, 65
5	5, 10, 15, 20, 25, 30, 35, 40	60, 91, 77, 99, 95, 87, 67, 84

Illustrative Example.

Nurse practitioner example

Suppose that a list of appointments for a nurse practitioner is available to us and that we will take a sample of one in four of the patients seen by this nurse on a given day for purposes of estimating the average time spent per patient. Suppose that on the day in which the sample was to be taken, the nurse saw a total of 12 patients in the order shown in Table 4.7. Since we specify a *one in four sample*, the four possible samples are shown in Table 4.8. The mean time $\overline{X}$ spent with the 12 patients is 29.583 mm (minutes) with variance $\sigma_x^2 = 153.08$. The variance over all possible samples of the estimated mean time spent per patient is (see Box 2.3)

Table: 4.7 Time with patients-Nurse Practitioners

Order of Visit	Time Spent with Patient (min.)	Order of Visit	Time Spent with Patient (min.)
1	15	7	49
2	34	8	40
3	35	9	25
4	36	10	46
5	11	11	33
6	17	12	14

Table: 4.8 Four Possible Samples for Data of Table 4.7 (1-in-4 sampling)

Patient	Time Spent (min.)	Patient	Time Spent (min.)	Patient	Time Spent (min.)	Patient	Time Spent (min.)
1	15	2	34	3	35	4	36
5	11	6	17	7	49	8	40
9	25	10	46	11	33	12	14
Total	51		97		117		90
Mean	17		32.33		39		30

Table: Intraclass Correlations

Sample(1,5,9)	Sample(2,6,10)
(15-29.583)(11-29.583)	(34-29.583)(17-29.583)
(15-29.583)(25-29.583)	(34-29.583)(46-29.583)
(11-29.583)(25-29.583)	(17-29.583)(46-29.583)
(422.996)	(-189.64)

Sample(3,7,11)	Sample(4,8,12)
(35-29.583)(49-29.583)	(36-29.583)(40-29.583)
(35-29.583)(33-29.583)	(36-29.583)(14-29.583)
(49-29.583)(33-29.583)	(40-29.583)(40-29.583)
(190.04)	(-195.478)

$$227.917 = (422.996 + (-189.64) + (190.04) + (-195.478))$$

$$\begin{aligned} \delta_x &= \frac{2 \times 227.917}{3 \times 4 \times (3-1) \times 153.08} \\ &= 0.124073 \end{aligned}$$

$$Var(\overline{x}) = \left(\frac{153.08}{3}\right)[1 + (0.124073)(2)]$$

Empirically,

$$\begin{aligned} Var\,(\overline{x}) &= \frac{1}{4}\left[(17 - 29.583)^2 + (32.33 - 29.583)^2\right. \\ &+ \left.(39 - 29.583)^2 + (30 - 29.583)^2\right] = 63.6 \end{aligned}$$

Table: 4.9 Data for Nurse Practitioner's Visits (Monotonically Ordered List)

Order of Visit	Time Spent with Patient (min.)	Order of Visit	Time Spent with Patient (min.)
1	49	7	33
2	46	8	25
3	40	9	17
4	36	10	15
5	35	11	14
6	34	12	11

Table: 4.10 Four Possible Samples for Data of Table 4.9 (1-in-4 sampling)

Patient	Time Spent (min.)	Patient	Time Spent (min.)	Patient	Time Spent (min.)	Patient	Time Spent (min.)
1	49	2	46	3	40	4	36
5	35	6	34	7	33	8	25
9	17	10	15	11	14	12	11
Total	101		95		87		72
Mean	33.67		31.67		29		24

Table: Intraclass Correlations

Sample(1,5,9)	Sample(2,6,10)
(49-29.583)(35-29.583)	(46-29.583)(34-29.583)
(49-29.583)(17-29.583)	(46-29.583)(15-29.583)
(35-29.583)(17-29.583)	(34-29.583)(15-29.583)
(-207.304)	(-231.308)

Sample(3,7,11)	Sample(4,8,12)
(40-29.583)(33-29.583)	(36-29.583)(25-29.583)
(40-29.583)(14-29.583)	(36-29.583)(11-29.583)
(33-29.583)(14-29.583)	(25-29.583)(11-29.583)
(-179.98)	(-63.4903)

$$-682.083 = ((-207.304) + (-231.308) + (-179.98) + (-63.4903))$$

$$\begin{aligned} \delta_x &= \frac{2 \times -682.083}{3 \times 4 \times (3-1) \times 153.08} \\ &= -0.371311 \end{aligned}$$

$$\begin{aligned} Var(\overline{x}) &= \left(\frac{153.08}{3}\right)[1 + (-0.371311)(2)] \\ &= 13.1331 \end{aligned}$$

Empirically,

$$\begin{aligned} Var\,(\overline{x}) &= \frac{1}{4}\left[(33.67 - 29.583)^2 + (31.67 - 29.583)^2\right. \\ &+ \left.(29 - 29.583)^2 + (24 - 29.583)^2\right] = 13.13 \end{aligned}$$

Adding some periodicity to the data

Table: 4.11 Data for Nurse Practitioner's Visits (Periodicity in List)

Order of Visit	Time Spent with Patient (min.)	Order of Visit	Time Spent with Patient (min.)
1(e)	11	7(m)	35
2(m)	17	8(h)	46
3(m)	36	9(e)	15
4(h)	49	10(m)	25
5(e)	14	11(m)	33
6(m)	34	12(h)	40

Table: 4.12 Four Possible Samples for Data of Table 4.11

Patient	Time Spent (min.)	Patient	Time Spent (min.)	Patient	Time Spent (min.)	Patient	Time Spent (min.)
1(e)	11	2(m)	17	3(m)	36	4(h)	49
5(e)	14	6(m)	34	7(m)	35	8(h)	46
9(e)	15	10(m)	25	11(m)	33	12(h)	40
Total	40		76		104		135
Mean	13.33		25.33		34.67		45

$$\begin{aligned} Var\left(\overline{x}\right) &= \frac{1}{4}\left[(13.33-29.583)^2+(25.33-29.583)^2\right. \\ &+ \left.(34.67-29.583)^2+(45-29.583)^2\right] = 136.41 \end{aligned}$$

Getting unbiased estimates

Modified systematic sampling

Random Number j	j/k	Remainder	Elements in Sample
1	1/6	1	1, 7, 13, 19, 25
2	2/6	2	2,8,14,20
3	3/6	3	3,9. 15,21
4	4/6	4	4, 10, 16, 22
5	5/6	5	5, 11, 17, 23
6	6/6	0	6, 12, 18, 24
7	7/6	1	1. 7, 13, 19, 25
8	8/6	2	2,8,14,20
9	9/6	3	3,9,15,21
10	10/6	4	4, 10, 16, 22
11	11/6	5	5, 11, 17, 23
12	12/6	0	6, 12, 18, 24
13	13/6	1	1, 7, 13, 19, 25
14	14/6	2	2, 8, 14, 20
15	15/6	3	3, 9, 15,21
16	16/6	4	4, 10, 16, 22
17	17/6	5	5, 11, 17,23
18	18/6	0	6, 12, 18, 24
19	19/6	1	1, 7, 13, 19, 25
20	20/6	2	2,.8, 14, 20
21	21/6	3	3, 9, 15, 21
22	22/6	4	4, 10, 16, 22
23	23/6	5	5, 11, 17,23
24	24/6	0	6, 12, 18, 24
25	25/6	1	1, 7, 13, 19, 25

Table: 4.15 Sampling Distribution of Estimated Means from Data of Tables 4.14 and 2.1

Elements in Sample	Estimated Mean No. of Visits, $\overline{x}$	Chance of Selection, π
1, 7, 13, 19, 25	12.00	5/25
2, 8, 14, 20	1.00	4/25
3, 9, 15, 21	4.25	4/25
4, 10, 16, 22	6.50	4/25
5, 11, 17, 23	3.50	4/25
6, 12, 18, 24	1.50	4/25

Estimation of Variances

As with all sampling methods, in order to construct confidence intervals, estimates of standards errors of the estimated population parameters are needed. In this section we demonstrate how estimates of standard errors are obtained in practice under systematic sampling. The variance of an estimated mean from a systematic sample of 1 in k elements (assuming that N/k is an integer) is the average squared deviation over the k possible samples of the estimated mean from the true population mean (since the estimate is unbiased when N/k is an integer).

Variance of estimated mean

In other words, if $\overline{x}_i$ represents the mean value of the population characteristic under study for the elements i, $i + k$, $i + 2k$ $i + (n - 1)k$, then the variance of an estimated mean from a systematic sample of 1 in k elements is given by

$$Var(\overline{x}) = \frac{\sum_{i=1}^{k} \left(\overline{x}_i - \overline{X}\right)^2}{k}$$

We can compare this expression with the *variance of the estimated mean under simple random sampling*, which was defined as

$$Var(\overline{x}) = \frac{\sum_{i=1}^{M} \left(\overline{x}_i - \overline{X}\right)^2}{M}$$

where M is the total number of possible samples.

A marvelous property of simple random sampling that this variance can be expressed simply in terms of σ_x^2 the variance of the original observations, using the following formula:

$$Var(\overline{x}) = \left(\frac{N-n}{N-1}\right)\left(\frac{\sigma_x^2}{n}\right)$$

Of greater importance to the statistician is that, by estimating the variance σ_x^2 with the statistic s_x^2 computed from the observations in the sample, the variance of $\overline{x}$ can be estimated as follows:

$$Var(\overline{x}) = \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right)$$

Hence, simply by knowing the variance of the observations in a particular sample, we can estimate the variance of the sampling distribution of the estimated mean under simple random sampling. Unfortunately, the same cannot be said of systematic sampling. In order to estimate the variance given in Equation (4.6), it is necessary to have two or more of these $\overline{x}_i$ available to us. However, in our systematic sample, the estimated mean $\overline{x}$ is simply one of the $\overline{x}_i$, with the particular one depending on which random number was chosen to start the sampling. We have no information from our sample concerning the variability of estimated means over all possible samples.

In practice, if we can assume that the list from which the systematic sample was taken represents a random ordering of the elements with respect to the variable being measured, then we can assume that the systematic sample is equivalent to a simple random sample. Therefore, the procedures developed for estimating the variances of estimates from simple random samples can be used. In other words, we estimate the population variance σ_x^2, by $\widehat{\sigma}_x^2$, as given by

$$\widehat{\sigma}_x^2 = \left(\frac{N-1}{N}\right)\left(s_x^2\right)$$

and where the $\overline{x}_i$ are the sample observations and $n = N/k$. The estimated variance of the estimated mean from a systematic sample is then given by

$$\widehat{Var}(\overline{x}) = \left(\frac{N-n}{N-1}\right)\left(\frac{\widehat{\sigma}_x^2}{n}\right)$$

Using this expression, we can then obtain confidence intervals for the population mean, $\overline{X}$, in the usual way.

Suppose we take a 1 in 5 systematic sample of physicians from the list given in Table 2.1 and that the initial random number chosen is 3. Table 4.16 lists the physicians in the sample along with their household visits. Using the data in Table 4.16, we have the following calculations (see Box 2.2 and the equations given above):

Table: 4.16 Systematic One in Five Sample Taken from Table 2.1

	Number of Visits
Physician in Sample	x_i
3	1
8	0
13	6
18	0
23	0

$$
\begin{aligned}
\overline{x} &= 1.4 \\
n &= 5 \\
N &= 25 \\
s_x^2 &= 6.8 \\
\widehat{\sigma}_x^2 &= \frac{24}{28} \times 6.8 = 6.528 \\
&= \\
\widehat{Var}(\overline{x}) &= \left(\frac{25-5}{25-1}\right)\left(\frac{6.528}{5}\right) = 1.088
\end{aligned}
$$

so that $\overline{x} \pm (1.96)\sqrt{\widehat{Var}(\overline{x})}$ or $1.4 \pm (1.96)\sqrt{1.088}$ ([-0.64,3.44]) would determine a 95% confidence inteval for the population mean.

Repeated Systematic Sampling

We discussed in previous sections how, with systematic sampling, the variances and standard errors of statistics as estimated from formulas shown in Box 4.1 can be either too small or too large if there is a relationship between the level of the particular variable and its position on the sampling frame. Also, the sample data themselves provide no insight into the nature of any ordering or periodicity in the sampling frame. It is possible, however, with a modified approach to systematic sampling to produce estimates of variances for estimated totals, means, and proportions that are unbiased no matter what kind of ordering or periodicity exists in the frame from which the sample was drawn. This modification is known as *repeated systematic sampling* and is demonstrated in the following example.

Illustrative Example.

Work Loss Data

For this example, we will use the data shown in Table 4.17. Let us suppose that we wish to take a systematic sample of approximately 18 workers from the list of 162 workers for purposes of estimating the mean number of work days lost per worker from acute illness. Since $n = 18$ and $N = 162$, a systematic sample of *1 in 9* workers would accomplish this. However, $18 = 6 \times 3$, and therefore, we can obtain a sample of 18 workers by taking 6 systematic samples, each containing 3 workers. In this case, the sampling interval $162/3 = 54$, and we would take 6 systematic samples of 1 in 54 workers. To do this, we first choose 6 random numbers between 1 and 54 (e.g., 2, 31, 46, 13, 34, 53), and then we choose systematic samples of 1 in 54 beginning with each random number. Our 6 samples are shown in Table 4.18.

Table: 4.18 Data for Six Systematic Samples Taken from Table 4.17

Random Number	Elements of Sample	Days Lost	Estimated Mean,
2	2	6	
	56	2	5.00
	110	7	
13	13	4	
	67	4	4.33
	121	5	
31	31	6	
	85	4	4.00
	139	2	
34	34	5	
	88	3	3.33
	142	2	
46	46	6	
	100	12	7.00
	154	3	
53	53	7	
	107	3	3.33
	161	0	

If we denote $\overline{x}_i$, as the estimated mean number of work days lost due to acute illness from the ith sample and m as the number of samples taken, then our estimated mean is

$$\begin{aligned} \overline{\overline{x}} &= \frac{\sum_{i=1}^{m} \overline{x}_i}{m} \\ &= \frac{5.00 + 4.33 + 4.00 + 3.33 + 7.00 + 3.33}{6} \\ &= 4.5 \end{aligned}$$

Since the estimated mean was obtained by taking a simple random sample of $m = 6$ means, $\overline{x}_i$, from a population of $M = 54$ such means, we can use the theory of simple random sampling to obtain the estimate $Var\left(\overline{\overline{x}}\right)$ of the variance of $\overline{\overline{x}}$.

$$Var\left(\overline{\overline{x}}\right) = \left(\frac{1}{m}\right) \times \frac{\sum_{i=1}^{m}\left(\overline{x}_i - \overline{\overline{x}}\right)^2}{m-1} \times \left(\frac{M-m}{M}\right)$$

where m is the number of samples taken and M is the total number of possible systematic samples. For this example, we have

$$\frac{\sum_{i=1}^{m}\left(\overline{x}_i-\overline{\overline{x}}\right)}{m-1} = \frac{1}{5}\left[(5-4.5)^2+(4.33-4.5)^2+(4-4.5)^2\right. \\ \left. + (3.33-4.5)^2+(7-4.5)^2+(3.33-4.5)^2\right] \\ = 1.9$$

and

$$Var\left(\overline{\overline{x}}\right) = \left(\frac{1}{6}\right)\times\frac{\sum_{i=1}^{6}\left(\overline{x}_i-\overline{\overline{x}}\right)^2}{6-1}\times\left(\frac{54-6}{54}\right) \\ = \left(\frac{1}{6}\right)(1.9)\left(\frac{54-6}{54}\right) \\ = 0.2814$$

so that $\overline{\overline{x}} \pm (1.96)\sqrt{\widehat{Var}(\overline{\overline{x}})}$ or $4.5 \pm (1.96)\sqrt{0.2814}$ ([3.46,5.54]) would determine a 95% confidence interval for the population

$$\eta = \frac{Z^2_{\alpha/2} \times \frac{\sum_{i=1}^{m}\left(\overline{x}_i - \overline{\overline{x}}\right)^2/(m-1)}{\overline{\overline{x}}^2} \times \left(\frac{N}{n}\right)}{\left(\frac{N}{n} - 1\right) \times \epsilon^2 + Z^2_{\alpha/2} \times \frac{\sum_{i=1}^{m}\left(\overline{x}_i - \overline{\overline{x}}\right)^2/(m-1)}{\overline{\overline{x}}^2}}$$

Suppose we wish to sample the list of employees shown in Table 4.17 for purposes of estimating the mean number of days lost from work due to acute illnesses, and that to do so we will take repeated systematic samples of 1 in 81 employees. Suppose further we wish to be virtually certain of estimating the mean number of days lost from work to within 20% of the true value, and that we wish to determine how many systematic samples m of 1 in 81 workers are needed. Let us take a preliminary sample of six such samples for purposes of estimating η. We first choose six random numbers between 1 and 81 (e.g., 22, 48, 27, 61, 53, 10) and obtain the six samples given in Table 4.19.

We then have the following calculations:

$$\begin{aligned} \overline{\overline{x}} &= 4.67 \\ s^2_{\overline{x}_i} &= 4.367 \\ \epsilon &= 0.2 \\ N &= 162 \\ n &= 2 \\ \frac{N}{n} &= 81 \end{aligned}$$

Table: 4.19 Six Samples Taken from Table 4.17

Random Number	Workers in Sample	Estimated Mean Days Lost, $\overline{x}_i$
10	10, 91	2.0
22	22, 103	4.0
27	27, 108	8.0
48	48, 129	4.5
53	53, 134	6.0
61	61, 142	3.5

From relation (4.11), we have

$$\eta = \frac{(3)^2 \times \frac{4.367}{(4.67)^2} \times 81}{(81-1) \times (0.2)^2 + (3)^2 \times \frac{4.367}{(4.67)^2}}$$

Thus, we would need approximately 30 systematic samples of 1 in 81 workers from the list to meet the specifications of the problem.