

Tutorial 5 (chapter 4 part 1)

Systematic Sampling

- 6.4 The number of colleges in 12 districts of a state are 8, 10, 6, 7, 7, 9, 11, 5, 6, 8, 9, and 11. List all possible samples of size 3 that can be selected from this population of 12 units using LS and CS sampling. Also, determine the average of corresponding sample means in both the cases. Are the two averages equal to the population mean? If yes, what does it indicate about the bias in the two estimators ?

Solution:

Linear systematic sampling:

Random start(r)	Serial #	y-values	Sample mean
1	(1, 5, 9)	(8,7,6)	7
2	(2, 6,10)	(10,9,8)	9
3	(3,7,11)	(6,11,9)	8.67
4	(4,8,12)	(7,5,11)	7.67

The average of the sample means is $E(\bar{y}_{sys}) = \sum_{r=1}^k \bar{y}_r = 7+9+8.67+7.67=8.08$

Population mean is $\bar{Y} = \sum_{i=1}^N Y_i = 8+10+.....+11= 8.08$

The average of the sample means is equal to the population mean, so it's unbiased estimator.

The R codes for LS is going to be in the next page.

For Circular systematic sampling:

Random start(r)	Serial #	y-values	Sample mean
1	(1, 5, 9)	(8,7,6)	7
2	(2, 6,10)	(10,9,8)	9
3	(3,7,11)	(6,11,9)	8.67
4	(4,8,12)	(7,5,11)	7.67
5	(5,8,1)	(7,6,8)	7
6	(6,9,2)	(9,8,10)	9
7	(7,10,3)	(11,9,6)	8.67
8	(8,11,4)	(5,11,7)	7.67
9	(9,12,5)	(6,8,7)	7
10	(10,1,6)	(8,10,9)	9
11	(11,2,7)	(9,6,11)	8.66
12	(12,3,8)	(11,7,5)	7.67

The average of the sample means is $E(\bar{y}_{sys}) = \sum_{r=1}^k \bar{y}_r = 7+9+8.67+7.67+....+7.67=8.08$

Population mean is $\bar{Y} = \sum_{i=1}^N Y_i = 8+10+.....+11= 8.08$

The average of the sample means is equal to the population mean, so it's unbiased estimator.

```

> ##### Q6.4 (LS)
> Y=c(8, 10, 6, 7, 7, 9, 11, 5, 6, 8, 9, 11)
> Y
[1] 8 10 6 7 7 9 11 5 6 8 9 11
>
> # compute population mean and variance:
> mean(Y)
[1] 8.083333
> var(Y)
[1] 3.901515
>
> # perform all systematic samples
> n=3 ; N=length(Y); k=N/n
> n
[1] 3
> N
[1] 12
> k
[1] 4
>
> sys_samples=matrix(0,n,k)
> sys_samples
      [,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    0    0    0    0
[3,]    0    0    0    0
> for (i in 1:k) sys_samples[,i]=Y[seq(i,N,k)]
> sys_samples
      [,1] [,2] [,3] [,4]
[1,]    8   10    6    7
[2,]    7    9   11    5
[3,]    6    8    9   11
>

```

```

> sys_mean=apply(sys_samples,2,mean)
> sys_mean
[1] 7.000000 9.000000 8.666667 7.666667
>
> #compare their mean with the population mean
> mean(sys_mean)
[1] 8.083333
> mean(Y)
[1] 8.083333
>
> #compute the variance of the sample mean
> var_sys_mean=1/k*(sum((sys_mean-mean(Y))**2))
> var_sys_mean
[1] 0.6319444
>
> # compute the variances of the systematic samples  $s_{\{i\}}^2$ 
> sys_var=apply(sys_samples,2,var)
> sys_var
[1] 1.000000 1.000000 6.333333 9.333333
> mean(sys_var)
[1] 4.416667
>
> # compute the variance of the SRS mean  $V(\bar{y})$ 
> S=sd(Y)
> varSRS=((N-n)/(N*n))*(S^2)
> varSRS
[1] 0.9753788
> varSRS/var_sys_mean*100
[1] 154.3457
>

```

- 6.5 Many trees along a canal have been uprooted by a storm. This damage persists along a 35 km stretch. The Department of Irrigation is interested in estimating total number of these damaged trees. Each one kilometer segment along the canal has been divided into 5 equal parts by stone markers. Thus, the entire 35 km long stretch is divided into 175 equal segments. Twenty five of these segments are selected using LS sampling with a sampling interval of 7 segments. The information regarding number of uprooted trees (y) obtained from this 1-in-7 systematic sample is given in the following table :

Selected segment	y	Selected segment	y	Selected segment	y
6	4	62	3	118	23
13	17	69	8	125	12
20	11	76	5	132	8
27	6	83	13	139	17
34	8	90	9	146	6
41	16	97	16	153	5
48	21	104	17	160	8
55	13	111	9	167	10
				174	15

Estimate the total number of uprooted trees, and also determine the confidence interval for it.

Solution:

The population size $N = 175$, and sampling interval $k = 7$. They use linear systematic sampling for the selection. Let the random number r selected from 1 to $k(= 7)$ be 6.

The sample mean is $\bar{y}_{sys} = \sum_{i=1}^n y_i = 4 + \dots + 15 = 11.2$

The estimate of total number of uprooted trees is $y'_{sys} = N * \bar{y}_{sys} = 1960$.

The estimate of variance from equation (6.4) is $V(\bar{y}_{sys}) = 0.9207$

Then estimate of variance $V(y'_{sys}) = N^2 * 0.9207 = 281966.88$

Using the estimate for total number of uprooted trees and the estimate of its variance, we now calculate the confidence interval for population total

$$N * \bar{y}_{sys} \pm 2 * \sqrt{V(\bar{y}_{sys}) * N}$$

$$[1624.161, 2295.839]$$

```

> ##### Q6.5 (LS)
> y=c(4,17,11,6,8,16,21,13,3,8,5,13,9,16,17,9,23,12,8,17,6,5,8,10,15)
> y
[1] 4 17 11 6 8 16 21 13 3 8 5 13 9 16 17 9 23 12 8 17 6 5 8 10 15
> n=length(y)
> N=175
> sum(y)
[1] 280
> # compute sample mean
> ybar=mean(y)
> ybar
[1] 11.2
> # compute sample total
> ytotal=N*ybar
> ytotal
[1] 1960
>

```

```

> # compute population variance
> i = 1:(length(y)-1 )
> X= sum ((y[i+1]- y[i])^2)
> X
[1] 1289
>
> var_sysmean= ((N-n)/(2*N*n*(n-1)))*X
> var_sysmean
[1] 0.9207143
>
> # compute population variance
> ytotal_var=N^2*var_sysmean
> ytotal_var
[1] 28196.88
>
> sd= sqrt(var_sysmean)
> sd
[1] 0.9595386
>
> #CI for sample total
> CIL=N*(ybar-2*sd)
> CIL
[1] 1624.161
> CIU=N*(ybar+2*sd)
> CIU
[1] 2295.839
>

```

- 6.6 It is desired to estimate the average per day rent for single occupancy rooms in well known hotels of a state. In all, there are 192 such hotels in the state and these are listed in a book entitled “A Guide to Visitors”. The investigator selected a 1-in-8 sample of hotels and rang up the managers of sampled hotels. The information on rent (in rupees) so obtained is given below :

Hotel	Rent	Hotel	Rent	Hotel	Rent
1	100	9	90	17	125
2	120	10	110	18	85
3	125	11	125	19	90
4	115	12	80	20	105
5	110	13	70	21	130
6	80	14	125	22	95
7	130	15	130	23	135
8	120	16	105	24	140

Estimate the average per day rent along with the confidence limits for it.

Solution:

The population size $N = 192$, and sampling interval $k = 8$. They use linear systematic sampling for the selection.

The estimate of the average per day is the sample mean which is $\bar{y}_{sys} = \sum_{i=1}^n y_i = 100 + \dots + 140 = 110$

The estimate of variance from equation (6.5) for random population is $V(\bar{y}_{sys}) = 14.34556$

Using the estimate of the average per day and the estimate of its variance, we now calculate the confidence interval for population total

$$\bar{y}_{sys} \pm 2 * \sqrt{V(\bar{y}_{sys})}$$

$$[102.4249, 117.5751]$$

```

> ##### Q 6.6 (LS)
> y=c(100,120,125,115,110,80,130,120,90,110,125,80,70,125,130,105,125,85
+      , 90,105,130,95,135,140)
> y
[1] 100 120 125 115 110 80 130 120 90 110 125 80 70 125 130 105 125 85 90 105 130
[22] 95 135 140
> sum(y)
[1] 2640
> n=length(y)
> n
[1] 24
> N=192
> k=N/n
> k
[1] 8
> # compute sample mean
> ybar=mean(y)
> ybar
[1] 110
>
>

```

```

> # compute the variance of the sample mean
> i = 1:n
> X= sum((y[i]- ybar)^2)
> X
[1] 9050
>
> var_sysmean= ((N-n)/(N*n*(n-1)))*X
> var_sysmean
[1] 14.34556
>
> sd= sqrt(var_sysmean)
> sd
[1] 3.787554
>
> #CI for sample mean
> CIL=ybar-(2*sd)
> CIL
[1] 102.4249
> CIU=ybar+(2*sd)
> CIU
[1] 117.5751
>

```

Exercise : Assume the data that we have from 100 observations as follows and consider the simple random sampling without replacement (SRS) with $n = 20$.

set.seed(111)
Y = sample(1:40, 100, replace = TRUE)

Compute

1. the population mean and variance.
2. Perform all 1-in-5 systematic samples (LS).
3. Compute their means.
4. Compute their variances.
5. Compute all systematic sample mean.
6. Verify that systemic mean is unbiased estimator of the population mean.
7. Compute the variance of the systematic sample mean $Var(\bar{y}_{sys})$
8. Compute the variance of the SRS mean $Var(\bar{y})$ with $n = 20$.
9. Find the relative efficiency of the variance of the simple random mean, $Var(\bar{y})$, and the variance of the systematic mean, $Var(\bar{y}_{sys})$.

Solution:

1. The population mean is $\bar{Y} = 20.69$, the population variance is $V(\bar{y}_{sys}) = 120.9635$
2. The population size $N = 100$, and sampling interval $k = 5$. We use linear systematic sampling for the selection. Let the random number r selected from 1 to $k (= 5)$. Each sample has a size of ($n=20$). The R codes is attached for sampling.
3. The systematic means for all samples are
 $20.25, 22.55, 20.80, 18.10, 21.75$
4. The variances of the sample means are
 $158.7237, 93.1026, 119.8526, 115.8842, 130.6184$
5. The all systematic sample mean $\bar{y}_{sys} = \sum_{r=1}^k \bar{y}_r = 20.25 + \dots + 21.75 = 20.69$
6. From answer 5, the all systematic sample mean is equal to the population mean, therefore it's unbiased estimator.
7. The estimate of variance from equation (6.2) is $V(\bar{y}_{sys}) = 2.2994$
8. For simple random sampling without replacement with $n = 20$ and from equation 3.9,

$$Var(\bar{y}) = 4.8385$$

9. The relative efficiency of the variance of the simple random mean, $Var(\bar{y})$, and the variance of the systematic mean, $Var(\bar{y}_{sys})$ is $= 210.4263$ which it shows that the systematic sampling estimator $\bar{y}_{sys}$ for the given population, is 21.0426 times more efficient than the one based on SRS

```

> #exercise
> RNGkind(sample.kind = "Rejection")
> set.seed(111)
> Y=sample(1:40,100,replace=TRUE)
> Y
 [1] 14 20 19 25  5 35  8 26 10 36  8 14 26 25 24  7 21 15  1  9 40 25 35  7 36 28 38 25
[29]  4 31  6 30  5 37  6 36 28 18 27 29 23 32  1 16 39  4 25 27 24 16 15 27 29 24 30  1
[57] 29 25 26 18 32 19 33 25 27 28 28 11 31 12 31  6 39 31  6 18  9 12 18  6 13  5 31  6
[85] 22  6 27  7  4 24 23 35 26 13 23 37 25  6  8 36
>
> # compute population mean:
> mean(Y)
[1] 20.69
> var(Y)
[1] 120.9635

```

```

> # perform all 1-in-5 systematic samples
> k=5 ;n=20 ; N=100
> sys_samples=matrix(0,n,k)
> sys_samples
      [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
[4,]    0    0    0    0    0
[5,]    0    0    0    0    0
[6,]    0    0    0    0    0
[7,]    0    0    0    0    0
[8,]    0    0    0    0    0
[9,]    0    0    0    0    0
[10,]   0    0    0    0    0
[11,]   0    0    0    0    0
[12,]   0    0    0    0    0
[13,]   0    0    0    0    0
[14,]   0    0    0    0    0
[15,]   0    0    0    0    0
[16,]   0    0    0    0    0
[17,]   0    0    0    0    0
[18,]   0    0    0    0    0
[19,]   0    0    0    0    0
[20,]   0    0    0    0    0

```

```

> for (i in 1:k) sys_samples[,i]=Y[seq(i,N,k)]
> sys_samples
      [,1] [,2] [,3] [,4] [,5]
[1,]   14   20   19   25    5
[2,]   35    8   26   10   36
[3,]    8   14   26   25   24
[4,]    7   21   15    1    9
[5,]   40   25   35    7   36
[6,]   28   38   25    4   31
[7,]    6   30    5   37    6
[8,]   36   28   18   27   29
[9,]   23   32    1   16   39
[10,]    4   25   27   24   16
[11,]   15   27   29   24   30
[12,]    1   29   25   26   18
[13,]   32   19   33   25   27
[14,]   28   28   11   31   12
[15,]   31    6   39   31    6
[16,]   18    9   12   18    6
[17,]   13    5   31    6   22
[18,]    6   27    7    4   24
[19,]   23   35   26   13   23
[20,]   37   25    6    8   36
> sys_mean=apply(sys_samples,2,mean)
> sys_mean
[1] 20.25 22.55 20.80 18.10 21.75
> #compare their mean with the population mean
> mean(sys_mean)
[1] 20.69
> mean(Y)
[1] 20.69
> var_sys_mean=1/k*(sum((sys_mean-mean(Y))**2))
> var_sys_mean
[1] 2.2994

```

```
> # compute the variances of the systematic samples  $s_{\{i\}}^2$ 
> sys_var=apply(sys_samples,2,var)
> sys_var
[1] 158.72368  93.10263 119.85263 115.88421 130.61842
> mean(sys_var)
[1] 123.6363
> # compute the variance of the SRS mean  $V(\bar{y})$ 
> S=sd(Y)
> varSRS=((N-n)/(N*n))*(S^2)
> varSRS
[1] 4.838541
> varSRS/var_sys_mean*100
[1] 210.4263
>
```