CHAPTER 17

SAMPLING METHODS

17.1 Introduction

Sampling is a part and parcel of our daily life. The housewife uses the technique of sampling in taking a decision whether the rice is cooked properly or not by inspecting a sample of grains from a cooking vessel A businessman inspects a sample of goods for ordering a large consignment. In industry, a sample would be observed to assess the quality of a product (or products). A farmer would estimate his crop prospects by observing a sample of earheads (or the plants). In the above situations, sampling is being followed to save money and time to arrive at an idea of the characteristic in the population. If there would be a considerable variation in the population, sampling adopted in the usual way might not give correct picture about the population. For example, the consumer wants to purchase rice by inspecting a handful of it from the upper portion of bag. If the quality of rice is not uniform throughout the bag the decision he takes on the basis of inspecting an upper portion of the material may bring him a monetary loss. Similarly the decision taken on only few bags out of large consignment of bags which are not having uniform quality would be of serious consequence. Hence different sampling procedures were evolved for different situations to estimate the population characteristics with minimum risk These sampling methods were developed based on probability theory. There is also a sampling method called 'purposive sampling' which do not use probability theory. The main drawback of 'purposive sampling' is that it is not possible to provide the error involved in arriving at an estimate of the population, and also the confidence intervals for the population characteristic.

17.2 Simple random sampling

In this method every unit in the population will have equal probability of being selected in the sample. Alternatively, the simple random sampling is the method of selecting 'n' sampling units out of total N units such that all the possible $\binom{N}{n}$ samples would have equal chance of being selected.

17.2.1 Sample random sampling with replacement (SRSWR): A sample is drawn such that every sampling unit drawn would be replaced back in the population. In this way the sample may contain repeated elements and any number of samples could be drawn.

17.2.2 Simple random sampling without replacement (SRSWOR): A sample is drawn such that every sampling unit drawn would not be replaced back. The sample would contain all distinct elements. If there are N units in the population and n units in the sample, there would be $\binom{N}{n}$ distinct samples by this method.

17.2.3 Selection of a random sample: List of units would be prepared by serially numbering all the sampling units from 1 to N and n random numbers would be selected from the column (or row) of a table of random numbers either by SRSWR or SRSWOR. For example, if $N=40$ and $n=5$, two columns would be selected from the table of random numbers. The maximum figure in two column table would be 99. The numbers 81 to 99 would be rejected since they have more probability than the numbers from 1 to 80. Supposing that 75 would be selected in the first draw, the actual random numbers would be the remainder after dividing 75 by 40 i. e., 35. If 80 would be selected in a particular draw the random number selected would be 40 since the remainder would be zero. In this way all the five numbers would be selected either by with or without replacement.

The method of providing estimates of population mean, standard error of mean and the confidence intervals for population mean are given as follows. Let Y_i be the i-th observational value for the character under study.

Sample

n=size of the sample

$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n} Y_i$=mean of the sample and is an unbiased estimate of the population mean, $\overline{Y}_N$.

$\hat{Y}=N.\overline{Y}_n$=estimate of the population total, $\overline{Y}$ and is an unbiased estimate.

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}Y_i^2-\frac{(\sum_{i=1}^{n}Y_i)^2}{n}\right]$$=mean square in the sample and is an unbiased estimate of S^2.

Est. $V(\overline{Y}_n)=\frac{N-n}{Nn}s^2$=estimate of the variance of the sample mean and is an unbiased estimate of $V(\overline{Y}_n)$.

Est. S.E($\overline{Y}_n$)=$\sqrt{\text{Est } V(\overline{Y}_n)}$=estimate of the standard error of sample mean and is an unbiased estimate of S.E($\overline{Y}_n$).

Population

N=size of the population

$\overline{Y}_N=1/N\sum_{i=1}^{N} Y_i$=mean of the population.

$Y=\sum_{i=1}^{N} Y_i$=Population total

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\overline{Y}_N)^2$$

$$=\frac{1}{N-1}\left[\sum_{i=1}^{N}Y_i^2-\frac{(\sum Y_i)^2}{N}\right]$$

=mean square in the population.

$V(\overline{Y}_n)=\frac{N-N}{Nn}S^2$=variance of sample mean in the population, and $\frac{N-n}{N}$ is called the finite population correction. S.E($\overline{Y}_n$)=$\sqrt{V(\overline{Y}_n)}$=standard error of sample mean in the population.

Est $V(\hat{Y})=\frac{N^2(N-n)}{Nn}s^2$=estimate of variance of the estimate of total and is an unbiased estimate of $V(\hat{Y})$.

Confidence limits: If S^2 is not known and the size of sample is small, the confidence limits for population mean, $\overline{Y}_N$ are given as

$$\overline{Y}_n \pm t_{(n-1)} \times \text{Est.S.E.}(\overline{Y}_n)$$

as upper and lower limits. These limits can be written as

$$\overline{Y}_n - t_{(n-1)} \times \text{Est S.E.}(\overline{Y}_n)$$

and $\overline{Y} + t_{(n-1)} \times \text{Est.S.E.}(\overline{Y}_n)$

where $t_{(n-1}$ is tabulated value of student's t-distribution with $(n-1)$ d.f.

$V(\hat{Y})=N^2\frac{(N-n)}{Nn}S^2$=variance of sample total in the population.

EXAMPLE: A sample of 50 progressive farmers were selected from a district containing 800 progressive farmers by simple random sampling method so as to estimate the total area under high yielding variety of paddy. The list of selected farmers along with corresponding areas under high yielding variety (HYV) is given in Table 17.1. Estimate the mean area under HYV, standard error and confidence limits for the mean area in the district.

TABLE 17.1

Holding	*Area (Hectares)*	*Holding*	*Area (Hectares)*	*Holding*	*Area (Hectares)*
1	3.5	18	4.2	35	2.1
2	3.2	19	6.1	36	2.4
3	2.5	20	1.1	37	1.5
4	4.0	21	1.0	38	1.1
5	3.2	22	1.7	39	0.7
6	2.0	23	2.3	40	3.1
7	2.2	24	5.2	41	3.3
8	1.5	25	4.6	42	2.8
9	2.6	26	0.8	43	2.2
10	2.8	27	1.9	44	4.3
11	3.5	28	2.5	45	3.8
12	3.0	29	2.6	46	6.2
13	1.4	30	3.1	47	5.0
14	1.2	31	6.2	48	0 7
15	1.3	32	5.4	49	0.9
16	3.6	33	3.6	50	1.2
17	3.2	34	4.5		

$\Sigma Y=141.6,\ \Sigma Y^2=517.74$

$$\overline{Y}_n=\frac{141.6}{50}=2.83,\quad s^2=\frac{1}{50-1}\left[517.74-\frac{(141.6)^2}{50}\right]=2.38$$

$$\text{Est}.V(\overline{Y}_n)=\frac{800-50}{800\times 50}\times 2.38=0.0446$$

$$\text{Est}.SE(\overline{Y}_n)=\sqrt{0.0446}=0.2112$$

Confidence limits of $\overline{Y}_N$;

Lower limit: $2.83-1.96\times 0.2112=2.42$

Upper limit: $2.83+1.96\times 0.2112=3.24$

17.3 Stratifieu random sampling

In this method the population is divided into different homogeneous groups known as strata and a simple random sampling is selected from each of the strata to estimate the population mean (or total), standard error of the estimate and confidence limits for the population mean. It could be seen from the expression of Est. $V(\overline{Y}_n)$ from Section 17.2 of simple random sampling, the precision of the estimate depends on size of sample as well as the value of s^2. If the size of sample increases and the value of s^2 decreases, the precision of the estimate $(\overline{Y}_n)$ increases. Assuming that due to limitation of time-and money it would not be possible to increase the size of sample, the only alternative would be to decrease the value of s^2 for increasing the precision of the estimate, $(\overline{Y}_n)$. Stratified random sampling method provides the scope for decreasing the value of s^2 by dividing the population into homogeneous groups such that there would be more heterogeneity between the groups (or strata) and more homogeneity within the groups (or strata). For example, considering the selection of sample of rice from a bag by a consumer, if the bag does not contain uniform quality of rice the consumer would be risking a loss. Similarly if the adoption of rice yielding varieties is not same throughout the district and differs from one Panchayat Samiti to another then the estimate of the extent of adoption by farmers in the district would be less efficient assuming that the samples size is not small. Let Y be the character under study. The estimate of population, standard error of the estimate and the confidence limits for the population mean are given as follows.

17.3.1 Proportional allocation of Sample : If the proportion of the sample size in the i-th stratum is assumed to be same as proportion of population size in the same stratum, we have $n_i/n = N_i/N = \text{constant}$. In other words, the sample size to be allocated for i-th stratum is $n_i = n.\ N_i/N$ for $i = 1, 2, \ldots, k$. This type of allocation of sample is known as proportional allocation. By substituting $n_i = n.\ N_i/N$ in the mean of the sample, variance of the estimate, estimate of the variance of the estimate in Section 17.3. We have

Sample

n = size of the sample.

n_i = size of the sample in the i-th stratum for $i=1,2,\ldots,k$, where k is the number of strata.

$$n=\sum_{i=1}^{k} n_i$$

$\overline{Y}_{ni}=\frac{1}{n_i}\sum_{i=1}^{ni} Y_i$ = mean of i-th stratum in the sample and is an unbiased estimate of $\overline{Y}_{Ni}$

$\overline{Y}_n=1/n\sum_{i=1}^{k} n_i\overline{Y}_{ni}$ = mean of the sample.

$\overline{Y}_{st}=1/N\sum_{i=1}^{k} N_i\overline{Y}_{ni}$ = estimate of the population mean in the stratified random sampling method and is an unbiased estimate of $\overline{Y}_N$,

$s_i^2=\frac{1}{n_i-1}\sum_{j=1}^{ni}(Y_{ij}-\overline{Y}_{ni})^2$ = mean square of i-th stratum in the sample for $i=1,2,\ldots,k$

Est. V $(\overline{Y}_{si})=\sum_{i=1}^{k}\frac{N_i^2}{N^2}\frac{(N_i-n_i)}{N_in_i}s_i^2$ = estimate of variance of the estimate.

Est.S.E$(\overline{Y}_{st})=\sqrt{\text{Est.V}(\overline{Y}_{st})}$ = estimate of the standard error of the estimate

Confidence limits for $\overline{Y}_N$: If S_i^2 is not known and the size of total sample is small, the limits are given as $\overline{Y}_{st}-t_{(n-1)}\times$ Est.S.E.$(\overline{Y}_{st})$ as upper and lower limits.

Population

N = size of the population

N_i = size of the i-th stratum in the population for $i=1,2,\ldots,k$, where k is the number of strata.

$$N=\sum_{i=1}^{k} N_i.$$

$\overline{Y}_{Ni}=\frac{1}{N_i}\sum_{i=1}^{Ni} Y_i$ = mean of i-th stratum in the population.

$\overline{Y}_N=1/N\sum_{i=1}^{k} N_i\overline{Y}_{Ni}$ = mean of the population.

$S_i^2=1/N_i-1\sum_{j=1}^{Ni}(Y_{ij}-\overline{Y}_{Ni})^2$ = mean square of i-th stratum in the population for $i=1,2,\ldots k$

$V(\overline{Y}_{st})=\sum_{i=1}^{k}\frac{N_i^2}{N^2}\frac{(N_i-n_i)}{N_in_i}S_i^2$ = variance of the estimate

S.E.$(\overline{Y}_{st})=\sqrt{V(\overline{Y}_{st})}$ = standard error of the estimate.

Sample

$\overline{Y}_{st}=\overline{Y}_n$=estimate of the population mean in the proportional allocation.

Est. V ($\overline{Y}_{st}$) Prop

$$=\frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i}{N}s_i^2$$ =estimate of the variance of the estimate in the propoitional allocation.

Est. S. E. ($\overline{Y}_{st}$) prop.

$=\sqrt{\text{Est. V }(\overline{Y}_{st})}$ =estimate of the standard error of the estimate in the proportional allocation.

Population

V ($\overline{Y}_{st}$) prop

$$=\frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i}{N}S_i^2$$ =variance of the estimate in the proportional allocation.

S. E. ($\overline{Y}_{st}$) prop.

$=\sqrt{V\ (\overline{Y}_{st})}$ prop=standard error of the estimate in the proportional allocation.

Confidence limits for $\overline{Y}_N$ can also be obtained on the similar lines as in Section 17.3 for the proportional allocation.

17.3.2 Neyman's allocation of sample : In this case the total sample would be allocated to different strata in such a way that the variance of the estimate would be minimized with the condition that the total of all the strata samples is equal to the total sample size i. e. $\sum_{i=1}^{k}n_i=n$. By minimizing the variance of the estimate in Section 17.3 with the restriction that $\sum_{i=1}^{k}n_i=n$, we have $n_i=n\times\frac{N_iS_i}{\sum_{i=1}^{k}N_iS_i}$

On substituting this expression of n_i in the variance of the estimate and estimate of the variance of the estimate in Section 17.3, the corresponding expression in optimum allocation, are obtained as follows.

Sample

Est. V ($\overline{Y}_{st}$) opt$=\frac{1}{N^2n}$ $(\sum_{i=1}^{k}N_is_i)^2-\frac{1}{N^2}(\sum_{i=1}^{k}N_is_i^2)$ =estimate of the variance of the estimate in the (optimum) Neyman's allocation.

Population

V ($\overline{Y}_{st}$) opt.$=\frac{1}{N^2}(\sum_{i=1}^{k}\frac{N_iS_i^2}{n})-$ $\frac{1}{N^2}(\sum_{i=1}^{k}N_is_i^2)$=variance of the estimate in the (optimum) Neyman's allocation.

17.3.3 Comparison of Stratified random sampling with simple random sampling: It was proved that

$$V\,(\overline{Y}_{st})_{opt} \leqslant V\,(\overline{Y}_{st})_{prop} \leqslant V\,(\overline{Y}_{n})_{ran}$$

which indicates optimum allocation is more precise in comparison to proportional allocation which in turn more precise to simple random sampling. For further reading on this topic please refer to Cochran (1953).

17.3.4 Selection of a Sample with probability proportional to size: Often a situation arises to draw a sample from the population with probability proportional to size. For example, a sample of farmers are to be selected with probability proportional to area under their holding. The procedure of drawing n farmers out of N with probability proportional to area under holding is as follows. Let i-th farmer holds A_i hectares and $\sum_{i=1}^{N} A_i = A$. All the farmers would be arranged serially according to the size of their holding. The third column in Table 17.2 gives cumulative totals of holdings.

TABLE 17.2

S. No. of Farmer 1	*Holding size* 2	*Cumulative totals* 3
1	A_1	A_1
2	A_2	$A_1 + A_2$
3	A_3	$A_1 + A_2 + A_3$
⋮	⋮	⋮
N	A_N	$\sum_{i=1}^{N} A_i = A$

Select a random number (say) r from the table of random numbers out of A. If this number lies between $A_1 + A_2 + \ldots A_{i-1}$ and $A_1 + A_2 \ldots + A_{i-1} + A_i$ then i–th farmer is selected. Select another random number and if it lies between 1 and A_1 then first farmer is selected. In this way all the n farmers can be selected.

17.3.5 Lahiri's method of selecting a sample with probability proportional to size: This method avoids the need of writing cumulative totals for selecting a sample when the number of units in the population is considerably large. Let N be the number of units in the population and M be the maximum of the sizes of the N units or some number greater than that maximum size (M) and let M_i be the size of the i-th unit. The procedure is as follows :

(i) Select a number at random from 1 to N, say, i

(ii) Select another number at random from 1 to M, say, R

(iii) Select i-th unit in the sample if $R \leqslant M_i$

(iv) Reject i-th unit and repeat the above process if $R > M_i$

For selecting a random sample of n units with probability proportional to size with replacement, the above procedure has to be repeated n times. If the selection is without replacement the above procedure has to be repeated till n distinct units are obtained.

EXAMPLE: A sample survey was conducted to estimate the credit needs of cultivators in a Taluka by considering panchayat samitis in that Taluka as strata. A random sample of 10 villages were selected from each of the three Panchayat Samitis of sizes 150, 100, 120 respectively. The credit needs (in thousand rupees) of the thirty villages are given in Table 17.3. Estimate the average credit needs of cultivator in the Taluka, standard error of the estimate and confidence interval for the average credit need.

TABLE 17.3 CREDIT NEED (IN THOUSANDS) SAMITHI

Village	*I*	*II*	*III*
1	15	16	7
2	12	10	8
3	8	11	5
4	6	9	10
5	10	5	12
6	4	3	9
7	3	7	11
8	16	5	10
9	7	13	6
10	5	4	2

$$n_1=n_2=n_3=10,\ N_1=150,\ N_2=100,\ N_3=120$$

$$n=\sum_{i=1}^{3} n_i=30,\ N=\sum_{i=1}^{3} N_i=370$$

TABLE 17.4

Stratum No.	N_i	n_i	$\bar{Y}_{ni}$	s_i	$\frac{N_i^2(N_i-n_i)}{N^2 N_i n_i} s_i^2$	$\frac{N_i}{N}\bar{Y}_{ni}$
1	150	10	8.6	20.49	.3135	3.53
2	100	10	8.3	18.01	.1189	2.24
3	120	10	8.0	9.33	.0896	2.99
	370				.5220	8.76

$$s_i^2=\frac{1}{n_i-1}\left[\sum_{j=1}^{n_i} Y_{ij}^2-\frac{\left(\sum_{j=1}^{n_i} Y_{ij}\right)^2}{n_i}\right]$$

$$\bar{Y}_{st}=1/N\sum N_i\,\bar{Y}_{ni}=8.76$$

$$\text{Est. } V(\bar{Y}_{st})=\sum_{i=1}^{k}\frac{N_i^2}{N^2}\,\frac{(N_i-n_i)}{N_i\,n_i}\,s_i^2=0.5220$$

$$\text{Est. S.E}(\bar{Y}_{st})=0.7225$$

Confidence limits of $\bar{Y}_N$:

The lower limit: $\bar{Y}_{st}-t_{(n-1)}\times\text{Est. S.E}(\bar{Y}_{st})$
$=8.76-2.045\times0.7225=7.28$

The upper limit$=\bar{Y}_{st}+t_{(n-1)}\times\text{S.E}(\bar{Y}_{st})$
$=8.76+2.045\times.7225=10.24.$

17.4 Cluster sampling

In this method the statistical population would be divided into groups of ultimate sampling units called clusters for the process of sampling. For example, the ultimate sampling unit might be farm holding or village or group of villages for estimating the area under high yielding varieties (HYV) in a district. The choosing of ultimate sampling unit as farm holding, etc., depends on the precision required for the estimate and the cost