Estimate for total milk production in a district

$\hat{Y}_{ts} = NM\,\overline{Y}_{ts} = 40 \times 25 \times 10.94 = 10{,}940$ litres

Est. $V(\overline{Y}_{ts}) = N^2M^2(1/n - 1/N)\, s_b^2$

$+ NM^2(1/m - 1/M)\, s_w^{-2}$

$= 1306.9671 + 116400.00 = 117706.97$

Est. $S.E(\hat{Y}_{ts}) = \sqrt{\text{Est. } V(\hat{Y}_{ts})} = 343.08$

Confidence limits for total production:

Lower limit $= 10{,}940 - 1.96 \times 343.08 = 10{,}267.56$

Upper limit: $10{,}940 + 1.96 \times 343.08 = 11{,}612.44$

17.6 Systematic sampling

In this method, if one unit is selected at random the other units would be selected automatically. For example to estimate the area under a particular crop, one household would be selected randomly and the remaining households in the sample would be selected in a systematic manner by listing all the households in a serial order. Let N be the size of population and be the multiple of size of sample i.e. $N = nk$ where n is a size of sample and k is a positive integer. Let one unit be selected at random out of k units and the remaining $(n - 1)$ units be selected at equally spaced intervals of k units. Let 4-th unit was selected out of k units at random then the remaining units in the sample are $k + 4, 2k + 4, \ldots, (n - 1)k + 4$. This could be well understood from the following Table 17.11

TABLE 17.11

1	2	3	4	....	k
k + 1	k + 2	k + 3	k + 4	....	2k
2k + 1	2k + 2	2k + 3	2k + 4	....	3k
⋮	⋮	⋮	⋮		
(n − 1)k + 1	(n − 1)k + 2	(n − 1)k + 3	(n − 1)k + 4	....	nk

This method resembles stratified random sampling method though there is a subtle difference that one unit is selected at random from each of n strata. But the randomness was observed only for the first stratum but not for the other strata whereas this was not so in stratified random sampling. However, this method is equivalent to cluster sampling wherein one cluster is selected at random out of k clusters. This method is useful in forest research for estimating the volume of timber, total production of tamarind, lack, honey, etc. by serially numbering the trees, estimation of fish in a sea coast, etc. For further reading please refer to Sukhatme (1953).

17.7 Non-sampling errors

The standard errors of the estimates for the different sampling methods given in previous sections are known as sampling errors. The errors other than due to sampling are called non-sampling errors. The non-sampling errors might occur through (i) observational errors and (ii) incomplete samples (or non-response).

17.7.1 Observational errors: In socio-economic surveys, the observational value of the character under study may differ from investigator to investigator and from time to time even for the same investigator. In the sampling methods dealt in the previous sections, the assumption was that the value of the observation is unique but that is not so in practice. For example, estimation of a crop would differ from person to person, field to field and time to time. Further the recording of information by the investigator supplied by third person might be far away from true value. These errors form part of observational errors. For further reading please refer to Sukhatme (1953).

17.7.2 Incomplete samples: The non-sampling errors occur also through incomplete schedules furnished by field investigators or false type of information provided by respondents or investigators. If the information is not available for a complete sample the estimates based on incomplete sample are biased. Further, the cost of the survey would be increased in order to obtain the information again on the incomplete sample. For further reading please refer to Sukhatme (1953).

17.8 Tolerances in the testing of seeds

Rao and Apte (1972) gave a review on tolerances in the testing of seeds. By testing a sample of seeds in a laboratory the purity percentage would vary from scientist to scientist, sample to sample, core to core in a seed bag and finally from a lot to lot. The only way out in order to arrive at a confirmed and accurate result would be to give the confidence limits in which the true value of the purity percentage lies. The difference between upper and lower limits is known as 'Tolerance'. The tolerance value would be obtained by taking into all types of variations which might creep in at the time of drawing a sample of seeds or at the time of testing the sample in the laboratory.

The tolerance is expressed in terms of probability. The tolerances change according to the probability assumed. If the tolerance is calculated at 0.05 probability then five samples out of hundred may give results outside the expected variation.

17.8.1 Procedure of selecting a sample: The bags containing seeds, the cores within bags, the samples within cores would be selected randomly at each stage so that the drawn samples of seeds would become representative of the whole lot. If the random sampling procedure was not adopted at each stage it would be difficult to calculate the standard error of the estimate and thereafter the values of tolerances could not be estimated precisely by statistical methods. After selecting the sample by random sampling procedure it would be sent to laboratory for testing purity percentage (or germination percentage), other crop seeds, noxious, weed seeds, etc. The samples would be labelled along with tolerance values on each sample after they were tested and would be released to market.

17.8.2 Methods of computing tolerances: *1st method:* G.N. Collins proposed formulae to compute tolerance for non-chaffy and chaffy seeds based on Binomial distribution. For calculating tolerances for non-chaffy seeds (Purity and germination percentage) we have

$$T_1 = 0.6 + \frac{(0.2 \cdot X_1/X_2)}{100}$$

where 'X_1' be the percentage of the component under consideration and $X_2 = 100 - X_1$. The values of X_1 and X_2 could be obtained by testing the samples either for purity or for germination. These percentages could vary up to an extent of T_2.

Similarly, the tolerances for other crop seeds, weed seeds and Inert matter were calculated by the same author using the formula

$$T_2 = 0.2 + \frac{(0.2 \cdot X_1X_2)}{100}$$

In T_2, the first component of variation was found to be small, while the other component of variation remained constant.

Since chaffy seeds would not mix well, the range of variation for purity, etc. was expected to be more compared to non-chaffy seeds. The tolerance value for chaffy seeds was given by the formula

$$T_3 = T_1 \frac{(X_1 \text{ or } X_2)}{100}$$

where T_1 be the tolerance value for non-chaffy seeds, X_1 or X_2 be used in the formula whichever is less.

2nd method: S.R. Miles, A.S. Carter and L.C. Shenberger gave the following formula for tolerance of seeds.

$$T_4 = 1.414 \times t \times \sqrt{\frac{N-n}{n}\left[\frac{s_B^2}{n} + \frac{s_C^2}{n} + \frac{s_W^2}{n} + \frac{s_A^2}{n} + \frac{s_T^2}{n}\right]}$$

where t be the tabulated value of student's 't' for one tailed test at 5 per cent (or 1 per cent level), s_B^2, s_C^2, s_W^2, s_A^2 and s_T^2 are the means square among bags of seeds, cores within bags, working samples taken from the same submitted sample, Analysts, timings of testing the sample by the Analyst, respectively and n be the size of sample in each case. This formula could be used for both the chaffy and non-chaffy seeds and this was applicable to a given percentage of pure seed, other crop seeds, weed seeds or inert matter. The tolerances obtained for pure seed by this formula are somewhat small compared to tolerances obtained by previous formulae either in the case of chaffy and non-chaffy seeds. But in the case of other crop seeds, weed seeds and inert matter, the tolerances obtained by this formula are somewhat greater than previous tolerances due to small percentage of the component under consideration.

3rd method: The method for calculating tolerance for weed seeds is given here. Tests for seeds of noxious, weeds are quite different from the usual tests since the number of weed seeds per Kg (or Lb) in a lot should be determined. Usually the number of weed seeds would be quite small so the Poisson distribution is used in obtaining tolerances. These tolerances are dependent on the number of weed seeds found in each sample. They would give the maximum interval in which the number of weeds lie in each sample taking into consideration of variation due to sampling. Let T_5 be the tolerance value, we have

$$T_5 = (m + 1) + 1.96\sqrt{m}$$

where 'm' be the number of weed seeds found from a random sample of seeds and $\sqrt{m}$ be the standard deviation of the variable 'm' in a Poisson distribution. Normally, the sample selected for finding the weed seed tolerance would be 10 times of the sample taken for purity and germination percentage due to small percentage of weed seeds found in an ordinary sample.

EXERCISES

1. A sample of 20 adult women were selected from a locality containing 200 households by simple random sampling to estimate the average protein intake in a diet in that locality. The hypothetical data of average intakes of protein in a diet in a week by 20 adult women are presented here. Give the estimate of average intake in that locality and estimate of standard error and confidence limits for the population mean.

S.No.	*Average protein intake (gm)*	*S.No.*	*Average protein intake (gm)*
1	57	4	45
2	33	5	52
3	47	6	51

7	56	14	35
8	40	15	58
9	48	16	41
10	38	17	46
11	37	18	52
12	49	19	47
13	42	20	36

2. A sample survey was conducted in a locality by dividing the households into four income groups, for estimating the average height of adult males along with standard error. The sample was selected based on proportional allocation. The hypothetical data are presented. Estimate the average height of adult male along with standard error and confidence limits.

Income group	N_i	n_i	$\bar{Y}_{ni}$	s_i^2	$N_i/N \cdot s_i^2$
I	100	10	150	20.4	6.80
II	80	8	157	24.8	6.61
III	70	7	164	36.2	8.45
IV	50	5	169	45.3	7.55

3. A sample survey was conducted to estimate the total egg production in a district by cluster sampling method by randomly selecting 6 clusters of villages from the 40 clusters in a district. There are 5 poultry farms in each cluster and the egg count (in 10's) (hypothetical) in each Poultry farm is given.

Poultry farm	*Clusters*					
	1	*2*	*3*	*4*	*5*	*6*
1	42	18	18	14	13	3
2	38	37	39	17	9	13
3	26	43	42	28	31	18
4	20	25	50	16	16	24
5	19	21	35	25	19	16

Estimate the total egg production, estimate the standard error and the confidence limits for the total in a district.

4. A sample survey was conducted to estimate the total area under the high yielding varieties of paddy in a district with the stratified random sampling method with proportional allocation and the following results were obtained.

Stratum	*Stratum size* (N_i)	*Sample size* (n_i)	s_i^2	$\bar{Y}_{ni}$ (*in 100 hectares*)
1	5	15	3.5	10
2	8	25	4.2	12
3	7	20	5.4	14

Estimate the total area, standard error and confidence limits for the total area under high yielding varieties.