UNIVERSITY OF LONDON

Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

4.3 Further random sampling

Systematic sampling

In systematic sampling, the sample is chosen by selecting a *random* starting point and then picking every ith element in succession from the sampling frame. The **sampling interval**, i, is determined by dividing the population size, N, by the sample size, n, and rounding to the nearest integer. When the **ordering of the elements** is related to the characteristic of interest, systematic sampling *increases the representativeness of the sample*. If the ordering of the elements produces a **cyclical pattern**, systematic sampling may actually *decrease the representativeness of the sample*.

Suppose there are 100,000 elements in the population and a sample of 1,000 is required. In this case the sampling interval is:

$$i = \frac{N}{n} = \frac{100000}{1000} = 100.$$

A random number between 1 and 100 is selected. If, for example, this number is 23, then the sample consists of elements 23, 123, 223, 323, 423, 523, and so on.

Suppose we select a random number between 1 and 5, say 2. The resulting sample of students consists of students 2, $(2 + 5 =)$ 7, $(2 + 5 \times 2 =)$ 12, $(2 + 5 \times 3 =)$ 17 and $(2 + 5 \times 4 =)$ 22. Note in this case all the students are selected from a single row.

A	B	C	D	E
1	6	11	16	21
2	**7**	**12**	**17**	**22**
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

Systematic sampling may or may not increase representativeness – it depends on whether there is any 'ordering' in the sampling frame. It is easier to implement relative to SRS.

Stratified sampling

Stratified sampling is a two-step process in which the population is partitioned (divided up) into subpopulations known as **strata**.[1] The strata should be mutually exclusive and collectively exhaustive in that every population element should be assigned to one and only one stratum and no population elements should be omitted. Next, elements are selected from each stratum by a random procedure, usually SRS. A major objective of stratified sampling is to increase the precision of statistical inference without increasing cost.

The elements *within a stratum* should be as *homogeneous* as possible (i.e. as similar as possible), but the elements *between strata* should be as *heterogeneous* as possible (i.e. as different as possible). The **stratification factors** should also be closely related to the characteristic of interest. Finally, the factors (variables) should decrease the cost of the stratification process by being easy to measure and apply.

In **proportionate stratified sampling**, the size of the sample drawn from each stratum is proportional to the relative size of that stratum in the total population. In **disproportionate (optimal) stratified sampling**, the size of the sample from each stratum is proportional to the relative size of that stratum *and* to the standard deviation of the distribution of the characteristic of interest among all the elements in that stratum.

Suppose we randomly select a number from 1 to 5 for each class (stratum) A to E. This might result, say, in the stratified sample consisting of students 4, 7, 13, 19 and 21. Note in this case one student is selected from each class.

A	B	C	D	E
1	6	11	16	**21**
2	**7**	12	17	22
3	8	**13**	18	23
4	9	14	**19**	24
5	10	15	20	25

Stratified sampling includes all important subpopulations and ensures a high level of precision. However, sometimes it might be difficult to select relevant stratification factors and the stratification process itself might not be feasible in practice if it was not known to which stratum each population element belonged.

Cluster sampling

In cluster sampling the target population is first divided into mutually exclusive and collectively exhaustive subpopulations known as **clusters**. A random sample of clusters is then selected, based on a probability sampling technique such as SRS. For each selected cluster, either *all* the elements are included in the sample (one-stage cluster sampling), or a *sample* of elements is drawn probabilistically (two-stage cluster sampling).

[1] 'Strata' is the plural of 'stratum'.

Elements *within* a cluster should be as *heterogeneous* as possible, but clusters themselves should be as *homogeneous* as possible. Ideally, each cluster should be a small-scale representation of the population. In **probability proportionate to size sampling**, the clusters are sampled with probability proportional to size. In the second stage, the probability of selecting a sampling unit in a selected cluster varies inversely with the size of the cluster.

Suppose we randomly select three clusters: B, D and E. Within each cluster, we randomly select one or two elements. The resulting sample here consists of students 7, 18, 20, 21 and 23. Note in this case there are no students selected from clusters A and C.

A	B	C	D	E
1	6	11	16	**21**
2	**7**	12	17	22
3	8	13	**18**	**23**
4	9	14	19	24
5	10	15	**20**	25

Cluster sampling is easy to implement and cost effective. However, the technique suffers from a lack of precision and it can be difficult to compute and interpret results.

Multistage sampling

In multistage sampling selection is performed at two or more successive stages. This technique is often adopted in large surveys. At the first stage, large 'compound' units are sampled (**primary units**), and several sampling stages of this type may be performed until we at last sample the basic units.

The technique is commonly used in cluster sampling so that we are at first sampling the main clusters, and then clusters within clusters etc. We can also use multistage sampling with mixed techniques, i.e. cluster sampling at Stage 1 and stratified sampling at Stage 2 etc.

An example might be a national survey of salespeople in a company. Sales areas could be identified from which a random selection is taken from these areas. Instead of interviewing every person in the chosen clusters (which would be a one-stage cluster sample), only randomly selected salespeople within the chosen clusters will be interviewed.