Chapter Two: Sampling Methods

When you completed this chapter, you will be able to:

- ✓ know the reasons of sampling;
- ✓ use the table of random numbers;
- ✓ perform Simple Random, Systematic, Stratified, Cluster, Quota and Judgement Sampling;
- ✓ apply appropriate sampling method to different problem area; and
- ✓ identify the ways of minimising sampling bias.

Reference(s): Mason Chapter 8, Owen Chapter 15, Newbold Chapter 18
Exercise(s): Seminars 3 and 4, Mason Chapter8 Exercises 1-4

One of the first tasks for the statistician is *to determine the relevant population*. Once this is done, four basic questions must be asked about samples and the process of inference:

1. What are the ***least expensive*** methods for collecting samples that best ensure that the sample are ***representative*** of the parent population?
2. What is the best way to *describe* sample information usefully and clearly?
3. How does one go about *drawing conclusions* from samples and making inferences about the population?
4. How *reliable* are the inferences and conclusions drawn from sample information?

According to point 1, the first step of doing statistics is to obtain a ***representative sample*** from the population, the study in this area is called ***SAMPLE DESIGN***, and the study for collecting ***useful information*** from the selected sample is called ***SURVEY METHOD***.

Sample Design

Sample Design is a procedure or plan, specified before any data are collected, for obtaining a sample from a given population.

A *representative sample* contains the relevant characteristics of the population in the same proportion, as they are included in that population.

Reasons for using samples instead of the whole population

1. It is ***time consuming*** to perform complete census (examine every person or items in the population).
2. It is ***costly*** to do a complete census.
3. It is ***inefficient*** to obtain a complete count of the target population. Even funds and time are available, it is doubtful the additional accuracy of a 100 percent sample.
4. ***Destructive*** of certain tests nature prohibits the use of complete census.
5. Another reason for not using the population is where its ***full membership is unknown***. For example, the population of dog, cat and fish are undefinable, and it might not be possible to establish how many South Africans are suffering from Aids.

In the view of *time* and *cost* involved in using population, people always rely on using sampling instead of using the population. Clearly, the sample must have to be ***representative*** of the population.

There are various methods we can use to try to ensure its representativeness, and they can be classified into two main categories: ***Probability*** *Sampling* and ***Non-Probability*** *Sampling*.

Probability Sampling (Random Sampling)

Probability sampling select samples by methods that allows each possible item to have a ***known*** *and* ***equal probability*** of being included in the sample. Also the selection of one item *does not affect* the chance of any other item being selected.

1. Simple Random Sampling

Sample is selected from target population under the provision that every member of the population has a *known and equal probability* of being included in the sample. If a sample of size n is drawn from population of size N, then the probability of being selected is $\frac{n}{N}$.

If all elements of the population are listed and numbered, random sample can be selected by drawing thoroughly mixed numbered tags or tickets from a bowl, but the process of selection is tedious, especially large samples are required.

The easiest way to select a sample randomly is to use random numbers. These random numbers can be generated either by a computer programme, or by a *table of random numbers*, as its name implies, this table consists of numbers that have *no pattern* or scheme.

		Column																									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*
Row	*1*	0	2	7	1	1	0	8	1	8	2	7	5	9	9	7	7	9	8	6	6	5	8	0	9	5
	2	9	4	8	7	3	9	0	9	3	5	3	1	6	8	4	6	3	9	5	2	0	9	8	6	5
	3	5	4	9	2	1	7	8	6	8	0	0	6	6	3	5	9	8	6	8	9	1	7	3	0	6
	4	7	7	6	4	0	9	7	6	3	6	3	7	3	9	7	9	3	3	7	9	5	6	4	5	4

(Extracted from the Table of Random Numbers in Appendix E of Mason)

To select random numbers from the table, the first step is to determine the starting point. The starting point should be determined on a random basis. Someone could start from row 2 column 4, that is number 7, or from row 4 column 14, that is number 9.

The second step is to determine the number of digit we want. Suppose that samples are selected from the population of 25. Each individual should be labelled from 01, 02, ..to 25, so 2-digit number is required. Or samples are selected from the population of 340, then a 3-digit number is required.

Although we can go in any direction in the table, we always read from ***left to right*** in sequences of the required number of digit, and then from ***top to down*** without skipping until sufficient number of samples is selected.

e.g.1 Suppose that the headmaster wants to select 5 students from a class of 25 students, and he start to read from row 1 and column 1.

		Column																									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*
Row	*1*	0	2	7	1	1	0	8	1	8	2	7	5	9	9	7	7	9	8	6	6	5	8	0	9	5
	2	9	4	8	7	3	9	0	9	3	5	3	1	6	8	4	6	3	9	5	2	0	9	8	6	5
	3	5	4	9	2	1	7	8	6	8	0	0	6	6	3	5	9	8	6	8	9	1	7	3	0	6
	4	7	7	6	4	0	9	7	6	3	6	3	7	3	9	7	9	3	3	7	9	5	6	4	5	4

(Extracted from the Table of Random Numbers in Appendix E of Mason)

Since the population size is 25, 2-digit numbers, from 01 to 25, are required to label the 25 students. The first 2-digit number from the random number table is 02, so the first student selected is number 2.
The second 2-digit number is 71, which is greater than 25, so it is ignored.
The third 2-digit number is 10, so the second student selected is number 10.
The next numbers selected are 81, 82, 75, 99, 77, 98, 66, and 58, all are greater than 25, so they are ignored.
The next number is 09, so the third student selected is number 9.
The next numbers selected are 59, 48, 73, 90, 93, and 53, all are greater than 25, so they are ignored.
The next number is 16, so the fourth student selected is number 16.
The next numbers selected are 84, 63, and 95, all are greater than 25, so they are ignored.
The next number is 20, so the fifth student selected is number 20.

As a result, the five students selected by simple random sampling methods are 2, 10, 9, 16, and 20.

A computer can be used to generate and sort random numbers in order to facilitate the sample selection. For instance, the random numbers can be sorted into an ascending sequence, therefore, the sample selected will be in the same sequential order as the population.

As the method suggests, it is the only sample that is ***free from bias***.

2. Systematic Random Sampling

Elements are arranged in some ways, a starting point is randomly selected, and then other elements are selected from the population ***at an uniform interval*** that is measured in time, order or space. For example, we pick every 10th name on the student list, or pick every 50th piece coming off an assembly line.

The element of randomness is usually introduced into this kind of sampling by using ***random number*** to pick the unit with which to start.

e.g.2 Suppose we inspect every 40th piece made by a particular machine, our result would be biases if, because of a regularly recurring failure, every 20th piece were defected.

e.g.3 Suppose we were sampling paper waste produced by households, and we decided to sample 100 households every 7th day (say every Monday). Chance are that our sample would not be representative, because Monday's rubbish would very likely included the Sunday's rubbish. thus, the amount of waste would be biased, that is over-estimated by our choice of this sample procedure. The bias would be avoided if we take samples every 8th day.

Advantages of the Systematic Random Sampling

Systematic Sampling provides an improvement over Simple Random Sampling in that the sample is ***spread more evenly*** over the ***entire population***.

Shortcomings of the Systematic Random Sampling

The real dangerous of the sampling is the possible presence of ***hidden periodicity*** (sequential pattern).

3. Stratified Random Sampling

To use stratified random sampling, the population is divided into numbers of *non-overlapping homogeneous* groups, called ***Strata***. Elements with similar characteristics are grouped together, called homogeneous group. Then either ***proportional*** sample or ***non-proportional*** sample can be used.

Proportional Sample

Proportional sample requires that the number of items selected from each stratum be ***in the same proportion as in the population***. After the number of elements selected is specified, the numbers of element to be selected from each stratum are corresponding to the proportion of that stratum in the population as a whole.

Non-proportional Sample

Equal numbers of elements are selected from each stratum, and give weight to the results according to the ***stratum's proportion*** of the whole population.

With either approach, Stratified Sampling guarantees that every element in the population has a chance of being selected.

e.g.4 A Stratified Sample of size n = 60 is to be taken from a population of size N = 4000, which consists of three strata of size N_1 = 2000, N_2 = 1200 and N_3 = 800. If the allocation is to be proportional, how large a sample must be taken from each stratum?

Sol.]

Proportional Sample

$n_1 = 60 \times 2000/4000 = 30$
$n_2 = 60 \times 1200/4000 = 18$
$n_3 = 60 \times 800/4000 = 12$

Samples of size n_1, n_2 and n_3 are selected from each stratum respectively, then the estimation of the mean of the whole population is:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3}{3}$$

Non-proportional Sample

Equal number of samples, 20, from each stratum, then the estimation of the mean of the whole population is:

$$\bar{\bar{x}} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

Advantages of the Stratified Random Sampling

Stratified random sampling are used to ***get rid of bias*** in sampling. Suppose we are going to study the income level of Hong Kong population. If Simple random sampling is used, there is a chance (even it is little) that all the collected samples are aged below 20 or over 60, and the result will be *biased*, since most of them are still studying or retired.

Stratified random sampling is useful in this case, we can divide the population into numbers of strata in terms of their age, and sample from each stratum. It makes sure that different ages of people are included in the sample.

Stratified random sampling is appropriate when the population is already divided into groups of different sizes. For example, suppose that the patients of a physician are divided into four groups according to age, as shown below, and the physician wants to find out how many hours his patients sleep, stratified random sampling would be used.

Composition of patients by age

Age Group	Percentage of Total
Birth - 19 years	30%
20 - 39 years	40%
40-59 years	20%
60 years and older	10%

If the strata are properly designed, they are more accurately reflecting characteristics of the population from which they chosen than other kind of sampling do.

4. Cluster Random Sampling

In cluster random sampling, the total area of interest (population) is divided into numbers of small, *non-overlapping blocks* (or ***clusters***), a number of these blocks (clusters) are then randomly selected for inclusion in the overall sample. We assume that these individual blocks are ***representative*** of the population as whole.

If the clusters are geographic subdivisions, this kind of sampling is called ***Area*** Sampling.

e.g.5 Suppose a market research team is attempting to determine by sampling the average number of television sets per household in a large city, they would use a city map to divide the whole city into blocks (or clusters), and then choose a certain number of blocks (clusters) for interviewing.

e.g.6 Suppose that the management of a large Chain Store organisation wants to interview a sample of its employees to determine their attitudes toward a proposed plan. If random methods are used to select, say, five stores from the list, and some or all employees of these five stores are interviewed, the resulting sample is a cluster random sample.

Advantages and Shortcomings of the Cluster Random Sampling

Although Cluster Sampling are usually ***not as reliable as*** estimates based on Simple Random Sampling of the ***same size***, they are usually ***more reliable per unit cost*** (cheaper to visit and interview employees working close together in clusters).

Comparison between Stratified Sampling and Cluster Random Sampling

Both stratified random sampling and cluster random sampling, the population is divided into well-developed groups. Stratified random sampling is that each group has *small variation within itself*, but there is *wide variation between the groups*. Cluster random sampling is in the opposite case, when there is *considerable variation within each group* but each group is essentially similar to each other (*little variation among groups*).

Non-Probability Sampling

Judgement Sampling

*Personal **judgement*** plays a significant role in their selection.

One important use of such sampling is in testing markets for new products, such test cities are usually not selected at random; instead they are carefully chosen because in someone's considered judgement, they are "typical" or "average" cities.

Quota Sampling

Interviewers are simply given ***quotas to be filled***. Once the quota is set, interviewers are granted flexibility in the choice of sample members. Interviewers naturally tend to select individuals who are most readily available.

Non-probability samplings are used primarily as a ***matter of convenience***, it may produce quite accurate estimates of population parameters, but the drawback is that since the sample is not chosen using probability methods, there is no valid way of determining of the resulting estimates.

Biased Samples

We toss a coin in a long run, there should be a tendency of coming up heads half the time and tails half the time. Suppose a person tosses a coin ten times, and it comes up heads on eight of these tosses, there will be two possibilities:

1. the coin is a biased, unfair coin (not a standard coin), or
2. the person has not tossed the coin a sufficient number of times

Biased samples will be gathered if we chose unsuitable sampling method, or collect insufficient number of samples.

Conclusions

Systematic sampling, stratified sampling, and cluster sampling attempt to approximate simple random sampling. It is because the principles of simple random sampling are the foundation for *statistical inference*, once these principles has been developed for simple random sampling, their extension to the other sampling methods is conceptually quite simple. ***All the methods are developed for their precision, economy, or physical case***.