CHAPTER 12

Selecting a Sample

In this chapter you will learn about:

- The differences between sampling in qualitative and quantitative research
- Definitions of sampling terminology
- The theoretical basis for sampling
- Factors affecting the inferences drawn from a sample
- Different types of sampling including:
    - Random/probability sampling designs
    - Non-random/non-probability sampling designs
    - The 'mixed' sampling design
- The calculation of sample size
- The concept of saturation point

Keywords: *accidental sampling, cluster sampling, data saturation point, disproportionate sampling, equal and independent, estimate, information-rich, judgemental sampling, multi-stage cluster sampling, non-random sample, population mean, population parameters, quota sampling, random numbers, random sample, sample statistics, sampling, sampling design, sampling element, sampling error, sampling frame, sampling population, sampling unit, sample size, sampling strategy, saturation point, snowball sampling, study population, stratified sampling, systematic sampling.*

The differences between sampling in quantitative and qualitative research

The selection of a sample in quantitative and qualitative research is guided by two opposing

philosophies. In quantitative research you attempt to select a sample in such a way that it is unbiased and represents the population from where it is selected. In qualitative research, number considerations may influence the selection of a sample such as: the ease in accessing the potential respondents; your judgement that the person has extensive knowledge about an episode, an event or a situation of interest to you; how typical the case is of a category of individuals or simply that it is totally different from the others. You make every effort to select either a case that is similar to the rest of the group or the one which is totally different. Such considerations are not acceptable in quantitative research.

The purpose of sampling in quantitative research is to draw inferences about the group from which you have selected the sample, whereas in qualitative research it is designed either to gain in-depth knowledge about a situation/event/episode or to know as much as possible about different aspects of an individual on the assumption that the individual is typical of the group and hence will provide insight into the group.

Similarly, the determination of sample size in quantitative and qualitative research is based upon the two different philosophies. In quantitative research you are guided by a predetermined sample size that is based upon a number of other considerations in addition to the resources available. However, in qualitative research you do not have a predetermined sample size but during the data collection phase you wait to reach a point of data saturation. When you are not getting new information or it is negligible, it is assumed you have reached a data saturation point and you stop collecting additional information.

Considerable importance is placed on the sample size in quantitative research, depending upon the type of study and the possible use of the findings. Studies which are designed to formulate policies, to test associations or relationships, or to establish impact assessments place a considerable emphasis on large sample size. This is based upon the principle that a larger sample size will ensure the inclusion of people with diverse backgrounds, thus making the sample representative of the study population. The sample size in qualitative research does not play any significant role as the purpose is to study only one or a few cases in order to identify the spread of diversity and not its magnitude. In such situations the data saturation stage during data collection determines the sample size.

In quantitative research, randomisation is used to avoid bias in the selection of a sample and is selected in such a way that it represents the study population. In qualitative research no such attempt is made in selecting a sample. You purposely select ‘information-rich’ respondents who will provide you with the information you need. In quantitative research, this is considered a biased sample.

Most of the sampling strategies, including some non-probability ones, described in this chapter can be used when undertaking a quantitative study provided it meets the requirements. However, when conducting a qualitative study only the non-probability sampling designs can be used.

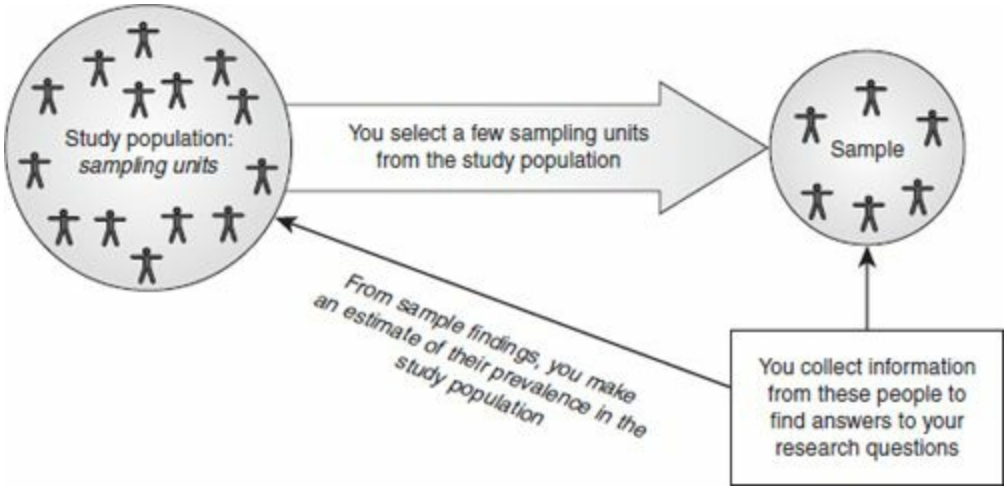

FIGURE 12.1 *The concept of sampling*

Sampling in quantitative research

The concept of sampling

Let us take a very simple example to explain the concept of sampling. Suppose you want to estimate the average age of the students in your class. There are two ways of doing this. The first method is to contact all students in the class, find out their ages, add them up and then divide this by the number of students (the procedure for calculating an average). The second method is to select a few students from the class, ask them their ages, add them up and then divide by the number of students you have asked. From this you can make an *estimate* of the average age of the class. Similarly, suppose you want to find out the average income of families living in a city. Imagine the amount of effort and resources required to go to every family in the city to find out their income! You could instead select a few families to become the basis of your enquiry and then, from what you have found out from the few families, make an estimate of the average income of families in the city. Similarly, election opinion polls can be used. These are based upon a very small group of people who are questioned about their voting preferences and, on the basis of these results, a *prediction* is made about the probable outcome of an election.

Sampling, therefore, is the process of selecting a few (a sample) from a bigger group (the **sampling population**) to become the basis for estimating or predicting the prevalence of an unknown piece of information, situation or outcome regarding the bigger group. A sample is a subgroup of the population you are interested in. See Figure 12.1.

This process of selecting a sample from the total population has advantages and disadvantages. The advantages are that it saves time as well as financial and human resources. However, the disadvantage is that you *do not find out the information* about the population's characteristics of interest to you but *only estimate* or *predict* them. Hence, the possibility of an error in your estimation exists.

Sampling, therefore, is a trade-off between certain benefits and disadvantages. While on the one hand you save time and resources, on the other hand you may compromise the level of accuracy in your findings. Through sampling you only make an estimate about the actual situation prevalent in the total population from which the sample is drawn. If you ascertain a piece of information from the total sampling population, and if your method of enquiry is correct, your findings should be reasonably

accurate. However, if you select a sample and use this as the basis from which to estimate the situation in the total population, an error is possible. Tolerance of this possibility of error is an important consideration in selecting a sample.

Sampling terminology

Let us, again, consider the examples used above where our main aims are to find out the average age of the class, the average income of the families living in the city and the likely election outcome for a particular state or country. Let us assume that you adopt the sampling method – that is, you select a few students, families or electorates to achieve these aims. In this process there are a number of aspects:

- The class, families living in the city or electorates from which you select you select your sample are called the *population* or **study population**, and are usually denoted by the letter N.
- The small group of students, families or electors from whom you collect the required information to estimate the average age of the class, average income or the election outcome is called the **sample**.
- The number of students, families or electors from whom you obtain the required information is called the **sample size** and is usually denoted by the letter n.
- The way you select students, families or electors is called the **sampling design** or **sampling strategy**.
- Each student, family or elector that becomes the basis for selecting your sample is called the **sampling unit** or **sampling element**.
- A list identifying each student, family or elector in the study population is called the **sampling frame**. If all elements in a sampling population cannot be individually identified, you cannot have a sampling frame for that study population.
- Your findings based on the information obtained from your respondents (sample) are called **sample statistics.** Your sample statistics become the basis of estimating the prevalence of the above characteristics in the study population.
- Your main aim is to find answers to your research questions in the study population, not in the sample you collected information from. From sample statistics we make an estimate of the answers to our research questions in the study population. The estimates arrived at from sample statistics are called *population parameters* or the **population mean**.

Principles of sampling

The theory of sampling is guided by three principles. To effectively explain these, we will take an extremely simple example. Suppose there are four individuals A, B, C and D. Further suppose that A is 18 years of age, B is 20, C is 23 and D is 25. As you know their ages, you can *find out* (calculate) their average age by simply adding 18 + 20 + 23 + 25 = 86 and dividing by 4. This gives the average (mean) age of A, B, C and D as 21.5 years.

Now let us suppose that you want to select a sample of two individuals to make an *estimate* of the average age of the four individuals. To select an unbiased sample, we need to make sure that each

unit has an equal and independent chance of selection in the sample. **Randomisation** is a process that enables you to achieve this. In order to achieve randomisation we use the theory of probability in forming pairs which will provide us with six possible combinations of two: A and B; A and C; A and D; B and C; B and D; and C and D. Let us take each of these pairs to calculate the average age of the sample:

1. A + B = 18 + 20 = 38/2 = 19.0 years;
2. A + C = 18 + 23 = 41/2 = 20.5 years;
3. A + D = 18 + 25 = 43/2 = 21.5 years;
4. B + C = 20 + 23 = 43/2 = 21.5 years;
5. B + D = 20 + 25 = 45/2 = 22.5 years;
6. C + D = 23 + 25 = 48/2 = 24.0 years.

Notice that in most cases the average age calculated on the basis of these samples of two (sample statistics) is different. Now compare these sample statistics with the average of all four individuals – the population mean (population parameter) of 21.5 years. Out of a total of six possible sample combinations, only in the case of two is there no difference between the sample statistics and the population mean. Where there is a difference, this is attributed to the sample and is known as **sampling error**. Again, the size of the sampling error varies markedly. Let us consider the difference in the sample statistics and the population mean for each of the six samples (Table 12.1).

TABLE 12.1 *The difference between sample statistics and the population mean*

Sample	Sample average (sample statistics) (1)	Population mean (population parameter) (2)	Difference between (1) and (2)
1	19.0	21.5	−2.5
2	20.5	21.5	−1.5
3	21.5	21.5	0.0
4	21.5	21.5	0.0
5	22.5	21.5	+1.0
6	24.0	21.5	+2.5

This analysis suggests a very important principle of sampling:

Principle 1 *– in a majority of cases of sampling there will be a difference between the sample statistics and the true population mean, which is attributable to the selection of the units in the sample.*

To understand the second principle, let us continue with the above example, but instead of a sample of two individuals we take a sample of three. There are four possible combinations of three that can be drawn:

1. 1 A + B + C = 18 + 20 + 23 = 61/3 = 20.33 years;
2. 2 A + B + D = 18 + 20 + 25 = 63/3 = 21.00 years;
3. 3 A + C + D = 18 + 23 + 25 = 66/3 = 22.00 years;
4. 4 B + C + D = 20 + 23 + 25 = 68/3 = 22.67 years.

Now, let us compare the difference between the sample statistics and the population mean (Table 12.2).

TABLE 12.2 *The difference between a sample and a population average*

Sample	Sample average (1)	Population average (2)	Difference between (1) and (2)
1	20.33	21.5	−1.17
2	21.00	21.5	−0.5
3	22.00	21.5	+0.5
4	22.67	21.5	+1.17

Compare the differences calculated in Table 12.1 and Table 12.2. In Table 12.1 the difference between the sample statistics and the population mean lies between –2.5 and +2.5 years, whereas in the second it is between –1.17 and +1.17 years. The gap between the sample statistics and the population mean is reduced in Table 12.2. This reduction is attributed to the increase in the sample size. This, therefore, leads to the second principle:

Principle 2 *– the greater the sample size, the more accurate the estimate of the true population mean.*

The third principle of sampling is particularly important as a number of sampling strategies, such as stratified and cluster sampling, are based on it. To understand this principle, let us continue with the same example but use slightly different data. Suppose the ages of four individuals are markedly different: A = 18, B = 26, C = 32 and D = 40. In other words, we are visualising a population where the individuals with respect to age – the variable we are interested in – are markedly different.

Let us follow the same procedure, selecting samples of two individuals at a time and then three. If we work through the same procedures (described above) we will find that the difference in the average age in the case of samples of two ranges between –7.00 and + 7.00 years and in the case of the sample of three ranges between –3.67 and +3.67. In both cases the range of the difference is greater than previously calculated. This is attributable to the greater difference in the ages of the four individuals – the sampling population. In other words, the sampling population is more heterogeneous (varied or diverse) in regard to age.

Principle 3 *– the greater the difference in the variable under study in a population for a given sample size, the greater the difference between the sample statistics and the true population mean.*

These principles are crucial to keep in mind when you are determining the sample size needed for a particular level of accuracy, and in selecting the sampling strategy best suited to your study.

Factors affecting the inferences drawn from a sample

The above principles suggest that two factors may influence the degree of certainty about the inferences drawn from a sample:

1. **The size of the sample** – Findings based upon larger samples have more certainty than those based on smaller ones. As a rule, *the larger the sample size, the more accurate the findings*.
2. **The extent of variation in the sampling population** – The greater the variation in the study population with respect to the characteristics under study, for a given sample size, the greater the uncertainty. (In technical terms, the greater the standard deviation, the higher the standard error for a given sample size in your estimates.) If a population is homogeneous (uniform or similar) with respect to the characteristics under study, a small sample can provide a reasonably good estimate, but if it is heterogeneous (dissimilar or diversified), you need to select a larger sample to obtain the same level of accuracy. Of course, if all the elements in a population are identical, then the selection of even one will provide an absolutely accurate estimate. As a rule, *the higher the variation with respect to the characteristics under study in the study population, the greater the uncertainty for a given sample size*.

Aims in selecting a sample

When you select a sample in quantitative studies you are primarily aiming to achieve maximum precision in your estimates within a given sample size, and avoid bias in the selection of your sample.

Bias in the selection of a sample can occur if:

- sampling is done by a non-random method – that is, if the selection is consciously or unconsciously influenced by human choice;
- the sampling frame – list, index or other population records – which serves as the basis of selection, does not cover the sampling population accurately and completely;
- a section of a sampling population is impossible to find or refuses to co-operate.

Types of sampling

The various sampling strategies in quantitative research can be categorised as follows (Figure 12.2):

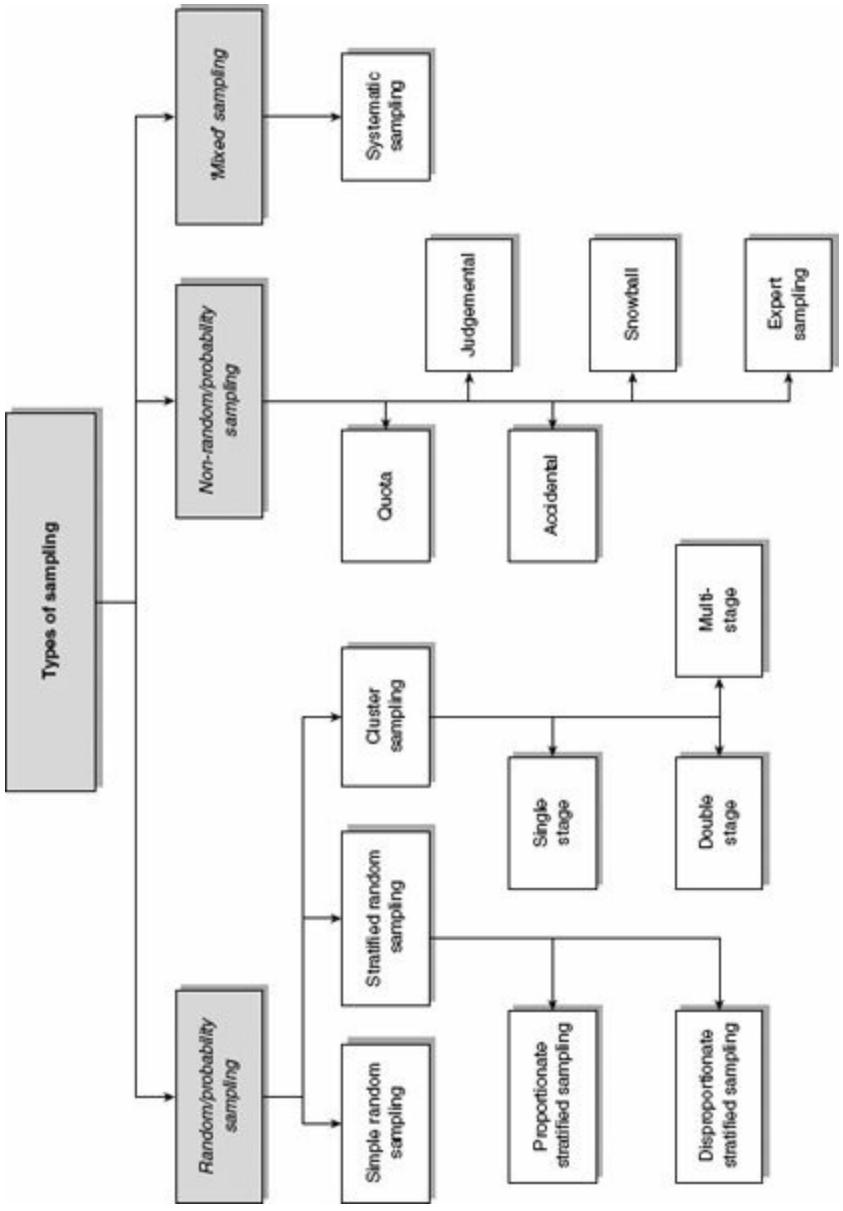

FIGURE 12.2 *Types of sampling in quantitative research*

- random/probability sampling designs;
- non-random/non-probability sampling designs selecting a predetermined sample size;
- 'mixed' sampling design.

To understand these designs, we will discuss each type individually.

Random/probability sampling designs

For a design to be called **random sampling** or **probability sampling**, it is imperative that each element in the population has an *equal* and *independent* chance of selection in the sample. Equal implies that the probability of selection of each element in the population is the same; that is, the choice of an element in the sample is not influenced by other considerations such as personal preference. The concept of independence means that the choice of one element is not dependent upon the choice of another element in the sampling; that is, the selection or rejection of one element does not affect the inclusion or exclusion of another. To explain these concepts let us return to our example

of the class.

Suppose there are 80 students in the class. Assume 20 of these refuse to participate in your study. You want the entire population of 80 students in your study but, as 20 refuse to participate, you can only use a sample of 60 students. The 20 students who refuse to participate could have strong feelings about the issues you wish to explore, but your findings will not reflect their opinions. Their exclusion from your study means that each of the 80 students does not have an equal chance of selection. Therefore, your sample does not represent the total class.

The same could apply to a community. In a community, in addition to the refusal to participate, let us assume that you are unable to identify all the residents living in the community. If a significant proportion of people cannot be included in the sampling population because they either cannot be identified or refuse to participate, then any sample drawn will not give each element in the sampling population an equal chance of being selected in the sample. Hence, the sample will not be representative of the total community.

To understand the concept of an *independent chance of selection*, let us assume that there are five students in the class who are extremely close friends. If one of them is selected but refuses to participate because the other four are not chosen, and you are therefore forced to select either the five or none, then your sample will not be considered an independent sample since the selection of one is dependent upon the selection of others. The same could happen in the community where a small group says that either all of them or none of them will participate in the study. In these situations where you are forced either to include or to exclude a part of the sampling population, the sample is not considered to be independent, and hence is not representative of the sampling population. However, if the number of refusals is fairly small, in practical terms, it should not make the sample non-representative. In practice there are always some people who do not want to participate in the study but you only need to worry if the number is significantly large.

A sample can only be considered a random/probability sample (and therefore representative of the population under study) if both these conditions are met. Otherwise, bias can be introduced into the study.

There are two main advantages of random/probability samples:

1. As they represent the total sampling population, the inferences drawn from such samples can be generalised to the total sampling population.
2. Some statistical tests based upon the theory of probability can be applied only to data collected from random samples. Some of these tests are important for establishing conclusive correlations.

Methods of drawing a random sample

Of the methods that you can adopt to select a random sample the three most common are:

1. **The fishbowl draw** – if your total population is small, an easy procedure is to number each element using separate slips of paper for each element, put all the slips into a box and then pick them out one by one without looking, until the number of slips selected equals the sample size you decided upon. This method is used in some lotteries.
2. **Computer program** – there are a number of programs that can help you to select a random

sample.

3. **A table of randomly generated numbers** – most books on research methodology and statistics include a table of randomly generated numbers in their appendices (see, e.g., Table 12.3). You can select your sample using these tables according to the procedure described in Figure 12.3.

The procedure for selecting a sample using a **table of random numbers** is as follows:

Let us take an example to illustrate the use of Table 12.3 for random numbers. Let us assume that your sampling population consists of 256 individuals. Number each individual from 1 to 256. Randomly select the starting page, set of column (1 to 10) or row from the table and then identify three columns or rows of numbers.

Suppose you identify the ninth column of numbers and the last three digits of this column (underlined). Assume that you are selecting 10 per cent of the total population as your sample (25 elements). Let us go through the numbers underlined in the ninth set of columns. The first number is 049 which is below 256 (total population); hence, the 49th element becomes a part of your sample. The second number, 319, is more than the total elements in your population (256); hence, you cannot accept the 319th element in the sample. The same applies to the next element, 758, and indeed the next five elements, 589, 507, 483, 487 and 540. After 540 is 232, and as this number is within the sampling frame, it can be accepted as a part of the sample. Similarly, if you follow down the same three digits in the same column, you select 052, 029, 065, 246 and 161, before you come to the element 029 again. As the 29th element has already been selected, go to the next number, and so on until 25 elements have been chosen. Once you have reached the end of a column, you can either move to the next set of columns or randomly select another one in order to continue the process of selection. For example, the 25 elements shown in Table 12.4 are selected from the ninth, tenth and second columns of Table 12.3.

TABLE 12.3 *Selecting a sample using a table for random numbers*

	1	2	3	4	5	6	7	8	9	10
1	48461	14952	72619	73689	52059	37086	60050	86192	67049	64739
2	76534	38149	49692	31366	52093	15422	20498	33901	10319	43397
3	70437	25861	38504	14752	23757	29660	67844	78815	23758	86814
4	59584	03370	42806	11393	71722	93804	09095	07856	55589	46820
5	04285	58554	16085	51555	27501	73883	33427	33343	45507	50063
6	77340	10412	69189	85171	29802	44785	86368	02583	96483	76553
7	59183	62687	91778	80354	23512	97219	65921	02035	59487	91403
8	91800	04281	39979	03927	82564	28777	59049	97532	54540	79472
9	12066	24817	81099	48940	69554	55925	48379	12866	41232	21580
10	69907	91751	53512	23748	65906	91385	84983	27915	48491	91068
11	80467	04873	54053	25955	48518	13815	37707	68687	15570	08890
12	78057	67835	28302	45048	56761	97725	58438	91529	24645	18544
13	05648	39387	78191	88415	60269	94880	58812	42931	71898	61534
14	22304	39246	01350	99451	61862	78688	30339	60222	74052	25740
15	61346	50269	67005	40442	33100	16742	61640	21046	31909	72641
16	56793	37696	27965	30459	91011	51426	31006	77468	61029	57108
17	56411	48609	36698	42453	85061	43769	39948	87031	30767	13953
18	62098	12825	81744	28882	27369	88185	65846	92545	09065	22653
19	68775	06261	54265	16203	23340	84750	16317	88686	86842	00879
20	52679	19599	13687	74872	89181	01939	18447	10787	76246	80072
21	84096	87152	20719	25215	04349	54434	72344	93008	83282	31670
22	83964	55937	21417	49944	38356	98404	14850	17994	17161	98981
23	31191	75131	72386	11689	95727	05414	88727	45583	22568	77700
24	30545	68523	29850	67833	05622	89975	79042	27142	99257	32349
25	52573	91001	52315	26430	54175	30122	31796	98842	37600	26025
26	16586	81842	01076	99414	31574	94719	34656	80018	86988	79234
27	81841	88481	61191	25013	30272	23388	22463	65774	10029	58376
28	43563	66829	72838	08074	57080	15446	11034	98143	74989	26885
29	19945	84193	57581	77252	85604	45412	43556	27518	90572	00563
30	79374	23796	16919	99691	80276	32818	62953	78831	54395	30705
31	48503	26615	43980	09810	38289	66679	73799	48418	12647	40044
32	32049	65541	37937	41105	70106	89706	40829	40789	59547	00783
33	18547	71562	95493	34112	76895	46766	96395	31718	48302	45893
34	03180	96742	61486	43305	84183	99605	67803	13491	09243	29557
35	94822	24738	67749	83748	59799	25210	31093	62925	72061	69991
36	04330	60599	85828	19152	68499	27977	35611	96240	62747	89529
37	43770	81537	59527	95674	76692	86420	69930	10020	72881	12532
38	56908	77192	50623	41215	14311	42834	80651	93750	59957	31211
39	32787	07189	80539	75927	75475	73965	11796	72140	48944	74156
40	52441	78392	11733	57703	29133	71164	55355	31006	25526	55790
41	22377	54723	18227	28449	04570	18882	00023	67101	06895	08915
42	18376	73460	88841	39602	34049	20589	05701	08249	74213	25220
43	53201	28610	87957	21497	64729	64983	71551	99016	87903	63875
44	34919	78801	59710	27396	02593	05665	11964	44134	00273	76358
45	33617	92159	21971	16901	57383	34262	41744	60891	57624	06962
46	70010	40964	98780	72418	52571	18415	64362	90637	38034	04909
47	19282	68447	35665	31530	59838	49181	21914	65742	89815	39231
48	91429	73328	13266	54898	68795	40948	80808	63887	89939	47938
49	97637	78393	33021	05867	86520	45363	43066	00988	64040	09803
50	95150	07625	05255	83254	93943	52325	93230	62668	79529	66964

Source: Statistical Tables, 3e, by F. James Rohlf and Robert R. Sokal. Copyright © 1969, 1981, 1994 by W.H. Freeman and Company. Used with permission.

Step 1	Identify the total number of elements in the study population, for example 50, 100, 430, 795 or 1265. The total number of elements in a study population may run up to four or more digits (if your total sampling population is 9 or less, it is one digit; if it is 99 or less, it is two digits; ...).
Step 2	Number each element starting from 1.
Step 3	If the table for random numbers is on more than one page, choose the starting page by a random procedure. Again, select a column or row that will be your starting point with a random procedure and proceed from there in a predetermined direction.
Step 4	Corresponding to the number of digits to which the total population runs, select the same number, randomly, of columns or rows of digits from the table.
Step 5	Decide on your sample size.
Step 6	Select the required number of elements for your sample from the table. If you happen to select the same number twice, discard it and go to the next. This can happen as the table for random numbers is generated by sampling with replacement.

FIGURE 12.3 *The procedure for using a table of random numbers*

TABLE 12.4 *Selected elements using the table of random numbers*

Column in Table 12.3	Elements selected				
9	49 246 34	232 161 40	52 243	29 61	65 213
10	63 234 231	68 44	108 211	72 156	25 220
2	149	246			

Sampling with or without replacement

Random sampling can be selected using two different systems:

1. sampling without replacement;
2. sampling with replacement.

Suppose you want to select a sample of 20 students out of a total of 80. The first student is selected out of the total class, and so the probability of selection for the first student is 1/80. When you select the second student there are only 79 left in the class and the probability of selection for the second student is not 1/80 but 1/79. The probability of selecting the next student is 1/78. By the time you select the 20th student, the probability of his/her selection is 1/61. This type of sampling is called **sampling without replacement**. But this is contrary to our basic definition of randomisation; that is, each element has an equal and independent chance of selection. In the second system, called **sampling with replacement**, the selected element is replaced in the sampling population and if it is selected again, it is discarded and the next one is selected. If the sampling population is fairly large, the probability of selecting the same element twice is fairly remote.

Step 1	Identify by a number all elements or sampling units in the population.
Step 2	Decide on the sample size *n*.
Step 3	Select *n* using the fishbowl draw, the table of random numbers or a computer program.

FIGURE 12.4 *The procedure for selecting a simple random sample*

Specific random/probability sampling designs

There are three commonly used types of random sampling design.

1. **Simple random sampling (SRS)** – The most commonly used method of selecting a probability sample. In line with the definition of randomisation, whereby each element in the population is given an equal and independent chance of selection, a simple random sample is selected by the procedure presented in Figure 12.4.

   To illustrate, let us again take our example of the class. There are 80 students in the class, and so the first step is to identify each student by a number from 1 to 80. Suppose you decide to select a sample of 20 using the simple random sampling technique. Use the fishbowl draw, the table for random numbers or a computer program to select the 20 students. These 20 students

become the basis of your enquiry.

2. **Stratified random sampling** – As discussed, the accuracy of your estimate largely depends on the extent of variability or heterogeneity of the study population with respect to the characteristics that have a strong correlation with what you are trying to ascertain (Principle 3). It follows, therefore, that if the heterogeneity in the population can be reduced by some means for a given sample size you can achieve greater accuracy in your estimate. Stratified random sampling is based upon this logic.

   In stratified random sampling the researcher attempts to stratify the population in such a way that the population within a stratum is homogeneous with respect to the characteristic on the basis of which it is being stratified. It is important that the characteristics chosen as the basis of stratification are clearly identifiable in the study population. For example, it is much easier to stratify a population on the basis of gender than on the basis of age, income or attitude. It is also important for the characteristic that becomes the basis of stratification to be related to the main variable that you are exploring. Once the sampling population has been separated into non-overlapping groups, you select the required number of elements from each stratum, using the simple random sampling technique. There are two types of stratified sampling: **proportionate stratified sampling** and **disproportionate stratified sampling**. With proportionate stratified sampling, the number of elements from each stratum in relation to its proportion in the total population is selected, whereas in disproportionate stratified sampling, consideration is not given to the size of the stratum. The procedure for selecting a stratified sample is schematically presented in Figure 12.5.

3. **Cluster sampling** – Simple random and stratified sampling techniques are based on a researcher's ability to identify each element in a population. It is easy to do this if the total sampling population is small, but if the population is large, as in the case of a city, state or country, it becomes difficult and expensive to identify each sampling unit. In such cases the use of cluster sampling is more appropriate.

   Cluster sampling is based on the ability of the researcher to divide the sampling population into groups (based upon visible or easily identifiable characteristics), called clusters, and then to select elements within each cluster, using the SRS technique. Clusters can be formed on the basis of geographical proximity or a common characteristic that has a correlation with the main variable of the study (as in stratified sampling). Depending on the level of clustering, sometimes sampling may be done at different levels. These levels constitute the different stages (single, double or multiple) of clustering, which will be explained later.

   Imagine you want to investigate the attitude of post-secondary students in Australia towards problems in higher education in the country. Higher education institutions are in every state and territory of Australia. In addition, there are different types of institutions, for example universities, universities of technology, colleges of advanced education and colleges of technical and further education (TAFE) (Figure 12.6). Within each institution various courses are offered at both undergraduate and postgraduate levels. Each academic course could take three to four years. You can imagine the magnitude of the task. In such situations cluster sampling is extremely useful in selecting a random sample.

   The first level of cluster sampling could be at the state or territory level. Clusters could be grouped according to similar characteristics that ensure their comparability in terms of student population. If this is not easy, you may decide to select all the states and territories and then select a sample at the institutional level. For example, with a simple random technique, one

institution from each category within each state could be selected (one university, one university of technology and one TAFE college). This is based upon the assumption that institutions within a category are fairly similar with regards to student profile. Then, within an institution on a random basis, one or more academic programmes could be selected, depending on resources. Within each study programme selected, students studying in a particular year could then be selected. Further, selection of a proportion of students studying in a particular year could then be made using the SRS technique. The process of selecting a sample in this manner is called *multi-stage cluster sampling*.

Step 1 Identify all elements or sampling units in the sampling population.
Step 2 Decide upon the different strata (k) into which you want to stratify the population.
Step 3 Place each element into the appropriate stratum.
Step 4 Number every element in each stratum separately.
Step 5 Decide the total sample size (n).
Step 6 Decide whether you want to select proportionate or disproportionate stratified sampling and follow the steps below.

Disproportionate stratified sampling	Proportionate stratified sampling
Step 7 Determine the number of elements to be selected from each stratum $= \frac{\text{sample size } (n)}{\text{no. of strata } (k)}$	Step 7 Determine the proportion of each stratum in the study population (p) $= \frac{\text{elements in each stratum}}{\text{total population size}}$
Step 8 Select the required number of elements from each stratum with SRS technique	Step 8 Determine the number of elements to be selected from each stratum = (sample size) × (p) Step 9 Select the required number of elements from each stratum with SRS technique
As this method does not take the size of the stratum into consideration in the selection of the sample, it is called disproportionate stratified sampling.	As the sample selected is in proportion to the size of each stratum in the population, this method is called proportionate stratified sampling.

FIGURE 12.5 *The procedure for selecting a stratified sample*

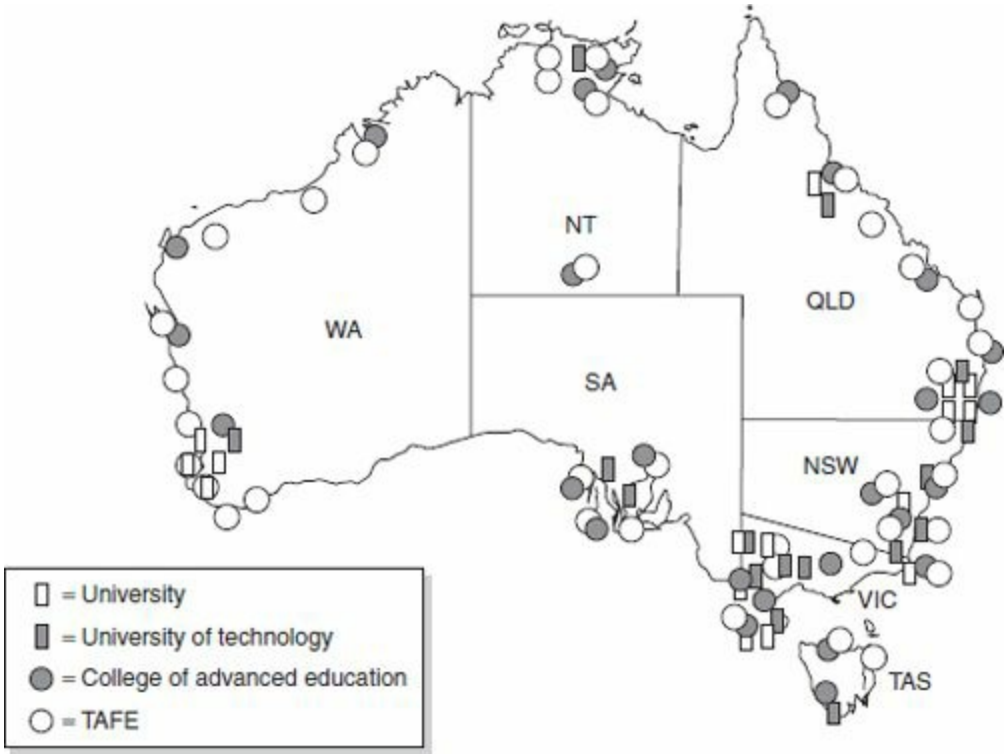

FIGURE 12.6 *The concept of cluster sampling*

Non-random/non-probability sampling designs in quantitative research

Non-probability sampling designs do not follow the theory of probability in the choice of elements from the sampling population. Non-probability sampling designs are used when the number of elements in a population is either unknown or cannot be individually identified. In such situations the selection of elements is dependent upon other considerations. There are five commonly used non-random designs, each based on a different consideration, which are commonly used in both qualitative and quantitative research. These are:

1. **quota sampling**;
2. **accidental sampling**;
3. **judgemental sampling** or **purposive sampling**;
4. **expert sampling**;
5. **snowball sampling**.

What differentiates these designs being treated as quantitative or qualitative is the predetermined sample size. In quantitative research you use these designs to select a predetermined number of cases (sample size), whereas in qualitative research you do not decide the number of respondents in advance but continue to select additional cases till you reach the data saturation point. In addition, in qualitative research, you will predominantly use judgemental and accidental sampling strategies to select your respondents. Expert sampling is very similar to judgemental sampling except that in expert sampling the sampling population comprises experts in the field of enquiry. You can also use quota and snowball sampling in qualitative research but without having a predetermined number of cases in mind (sample size).

Quota sampling

The main consideration directing quota sampling is the researcher's ease of access to the sample population. In addition to convenience, you are guided by some visible characteristic, such as gender or race, of the study population that is of interest to you. The sample is selected from a location convenient to you as a researcher, and whenever a person with this visible relevant characteristic is seen that person is asked to participate in the study. The process continues until you have been able to contact the required number of respondents (quota).

Let us suppose that you want to select a sample of 20 male students in order to find out the average age of the male students in your class. You decide to stand at the entrance to the classroom, as this is convenient, and whenever a male student enters the classroom, you ask his age. This process continues until you have asked 20 students their age. Alternatively, you might want to find out about the attitudes of Aboriginal and Torres Strait Islander students towards the facilities provided to them in your university. You might stand at a convenient location and, whenever you see such a student, collect the required information through whatever method of data collection (such as interviewing, questionnaire) you have adopted for the study.

The advantages of using this design are: it is the least expensive way of selecting a sample; you do not need any information, such as a sampling frame, the total number of elements, their location, or other information about the sampling population; and it guarantees the inclusion of the type of people you need. The disadvantages are: as the resulting sample is not a probability one, the findings cannot be generalised to the total sampling population; and the most accessible individuals might have characteristics that are unique to them and hence might not be truly representative of the total sampling population. You can make your sample more representative of your study population by selecting it from various locations where people of interest to you are likely to be available.

Accidental sampling

Accidental sampling is also based upon convenience in accessing the sampling population. Whereas quota sampling attempts to include people possessing an obvious/visible characteristic, accidental sampling makes no such attempt. You stop collecting data when you reach the required number of respondents you decided to have in your sample.

This method of sampling is common among market research and newspaper reporters. It has more or less the same advantages and disadvantages as quota sampling but, in addition, as you are not guided by any obvious characteristics, some people contacted may not have the required information.

Judgemental or purposive sampling

The primary consideration in purposive sampling is your judgement as to who can provide the best information to achieve the objectives of your study. You as a researcher only go to those people who in your opinion are likely to have the required information and be willing to share it with you.

This type of sampling is extremely useful when you want to construct a historical reality, describe a phenomenon or develop something about which only a little is known. This sampling strategy is more common in qualitative research, but when you use it in quantitative research you select a predetermined number of people who, in your judgement, are best positioned to provide you the needed information for your study.

Expert sampling

The only difference between judgemental sampling and expert sampling is that in the case of the former it is entirely your judgement as to the ability of the respondents to contribute to the study. But in the case of expert sampling, your respondents must be known experts in the field of interest to you. This is again used in both types of research but more so in qualitative research studies. When you use it in qualitative research, the number of people you talk to is dependent upon the data saturation point whereas in quantitative research you decide on the number of experts to be contacted without considering the saturation point.

You first identify persons with demonstrated or known expertise in an area of interest to you, seek their consent for participation, and then collect the information either individually or collectively in the form of a group.

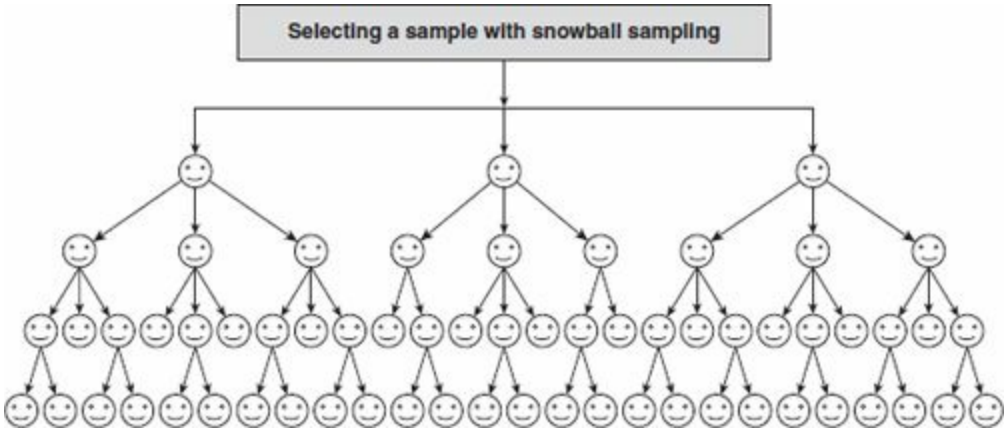

FIGURE 12.7 *Snowball sampling*

Snowball sampling

Snowball sampling is the process of selecting a sample using networks. To start with, a few individuals in a group or organisation are selected and the required information is collected from them. They are then asked to identify other people in the group or organisation, and the people selected by them become a part of the sample. Information is collected from them, and then these people are asked to identify other members of the group and, in turn, those identified become the basis of further data collection (Figure 12.7). This process is continued until the required number or a **saturation point** has been reached, in terms of the information being sought.

This sampling technique is useful if you know little about the group or organisation you wish to study, as you need only to make contact with a few individuals, who can then direct you to the other members of the group. This method of selecting a sample is useful for studying communication patterns, decision making or diffusion of knowledge within a group. There are disadvantages to this technique, however. The choice of the entire sample rests upon the choice of individuals at the first stage. If they belong to a particular faction or have strong biases, the study may be biased. Also, it is difficult to use this technique when the sample becomes fairly large.

Systematic sampling design: a 'mixed' design

Systematic sampling has been classified as a 'mixed' sampling design because it has the characteristics of both random and non-random sampling designs.

In systematic sampling the sampling frame is first divided into a number of segments called *intervals*. Then, from the first interval, using the SRS technique, one element is selected. The selection of subsequent elements from other intervals is dependent upon the order of the element selected in the first interval. If in the first interval it is the fifth element, the fifth element of each subsequent interval will be chosen. Notice that from the first interval the choice of an element is on a random basis, but the choice of the elements from subsequent intervals is dependent upon the choice from the first, and hence cannot be classified as a random sample. The procedure used in systematic sampling is presented in Figure 12.8.

Step 1 Prepare a list of all the elements in the study population (N).

Step 2 Decide on the sample size (n).

Step 3 Determine the *width of the interval* (k) = $\frac{\text{total population}}{\text{sample size}}$

Step 4 Using the SRS, select an element from the first interval (nth order).

Step 5 Select the same order element from each subsequent interval.

FIGURE 12.8 *The procedure for selecting a systematic sample*

Although the general procedure for selecting a sample by the systematic sampling technique is described above, you can deviate from it by selecting a different element from each interval with the SRS technique. By adopting this, systematic sampling can be classified under probability sampling designs.

To select a random sample you must have a sampling frame (Figure 12.9). Sometimes this is impossible, or obtaining one may be too expensive. However, in real life there are situations where a kind of sampling frame exists, for example records of clients in an agency, enrolment lists of students in a school or university, electoral lists of people living in an area, or records of the staff employed in an organisation. All these can be used as a sampling frame to select a sample with the systematic sampling technique. This convenience of having a 'ready-made' sampling frame may be at a price: in some cases it may not truly be a random listing. Mostly these lists are in alphabetical order, based upon a number assigned to a case, or arranged in a way that is convenient to the users of the records. If the 'width of an interval' is large, say, 1 in 30 cases, and if the cases are arranged in alphabetical order, you could preclude some whose surnames start with the same letter or some adjoining letter may not be included at all.

Suppose there are 50 students in a class and you want to select 10 students using the systematic sampling technique. The first step is to determine the width of the interval (50/10 = 5). This means that from every five you need to select one element. Using the SRS technique, from the first interval (1–5 elements), select one of the elements. Suppose you selected the third element. From the rest of the intervals you would select every third element.

The calculation of sample size

Students and others often ask: 'How big a sample should I select?', 'What should be my sample size?' and 'How many cases do I need?' Basically, it depends on what you want to do with the findings and what type of relationships you want to establish. Your purpose in undertaking research is the main determinant of the level of accuracy required in the results, and this level of accuracy is an important determinant of sample size. However, in qualitative research, as the main focus is to explore or describe a situation, issue, process or phenomenon, the question of sample size is less important. You usually collect data till you think you have reached saturation point in terms of discovering new information. Once you think you are not getting much new data from your respondents, you stop collecting further information. Of course, the diversity or heterogeneity in what you are trying to find out about plays an important role in how fast you will reach saturation point. And remember: *the greater the heterogeneity or diversity in what you are trying to find out about, the greater the number of respondents you need to contact to reach saturation point.* In determining the size of your sample for quantitative studies and in particular for cause-and-effect studies, you need to consider the following:

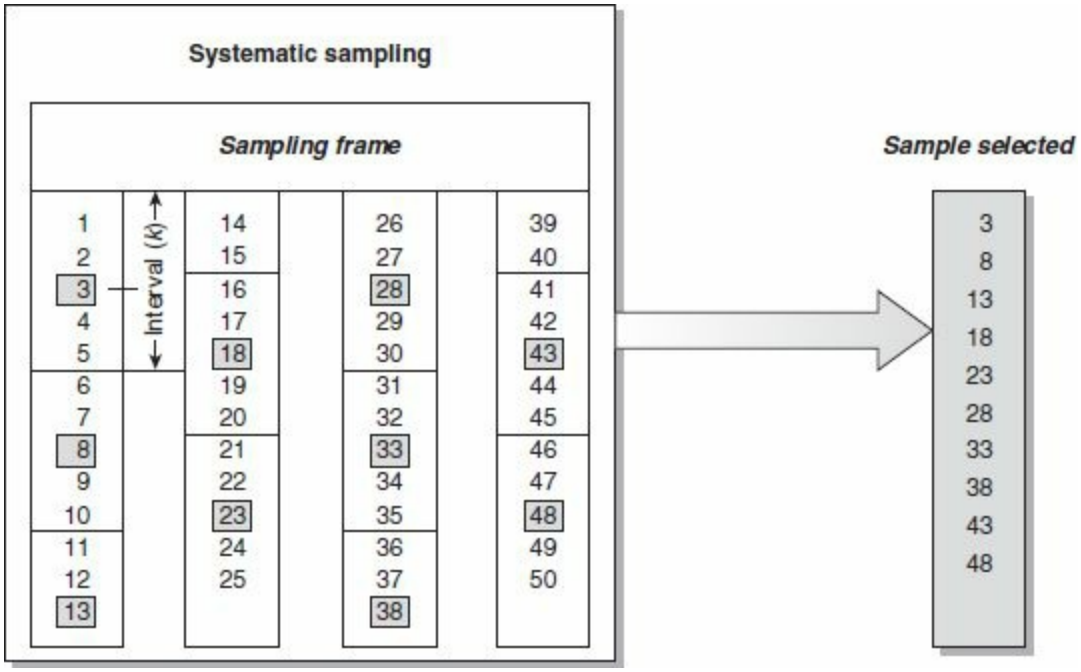

FIGURE 12.9 *Systematic sampling*

- At what *level of confidence* do you want to test your results, findings or hypotheses?
- With what *degree of accuracy* do you wish to estimate the population parameters?
- What is the estimated *level of variation* (standard deviation), with respect to the main variable you are studying, in the study population?

Answering these questions is necessary regardless of whether you intend to determine the sample size yourself or have an expert do it for you. The size of the sample is important for testing a hypothesis or establishing an association, but for other studies the general rule is: *the larger the sample size, the more accurate your estimates*. In practice, your budget determines the size of your sample. Your skills in selecting a sample, within the constraints of your budget, lie in the way you select your elements so that they effectively and adequately represent your sampling population.

To illustrate this procedure let us take the example of a class. Suppose you want to find out the average age of the students within an accuracy of 0.5 of a year; that is, you can tolerate an error of half a year on either side of the true average age. Let us also assume that you want to find the average age within half a year of accuracy at the 95 per cent confidence level; that is, you want to be 95 per cent confident about your findings.

The formula (from statistics) for determining the confidence limits is

$$\hat{x} = \bar{x} \pm (t_{0.05}) \frac{\sigma}{\sqrt{\eta}}$$

where

$\hat{x}$ = estimated value of the population mean

$\bar{x}$ = average age calculated from the sample

$t_{0.05}$ = value of t at 95 per cent confidence level

$\sigma/\sqrt{\eta}$ = standard error

σ = standard deviation

η = sample size

√ = square root

If we decide to tolerate an error of half a year, that means

$$\bar{x} \pm (t_{0.05}) \frac{\sigma}{\sqrt{\eta}}$$
$$= 0.5$$
$$= \bar{x} \pm 0.5$$

in other words we would like $\bar{x} \pm (t_{0.05}) \frac{\sigma}{\sqrt{\eta}}$

or (1.96*) $\sigma/\sqrt{\eta} = 0.5$ (value of $t_{0.05} = 1.96$)

$$\therefore \sqrt{\eta} = \frac{1.96 \times \sigma}{0.5}$$

*t-value from the following table

Level	0.02	0.10	0.05	0.02	0.01	0.001
t-value	1.282	1.645	1.960	2.326	2.576	3.291

There is only one unknown quantity in the above equation, that is σ.

Now the main problem is to find the value of σ without having to collect data. This is the biggest problem in estimating the sample size. Because of this it is important to know as much as possible about the study population.

The value of σ can be found by one of the following:

1. guessing;
2. consulting an expert;
3. obtaining the value of σ from previous comparable studies; or
4. carrying out a pilot study to calculate the value.

Let us assume that σ is 1 year. Then

$$\sqrt{\eta} = \frac{1.96 \times 1}{0.5} = 3.92$$
$$\therefore \eta = 15.37, \text{ say, } 16$$

Hence, to determine the average age of the class at a level of 95 per cent accuracy (assuming σ = 1 year) with half a year of error, a sample of at least 16 students is necessary.

Now assume that, instead of 95 per cent, you want to be 99 per cent confident about the estimated age, tolerating an error of half a year. Then

$$\sqrt{\eta} = \frac{2.576 \times 1}{0.5}$$
$$= 5.15$$
$$\therefore \eta = 26.54, \text{ say, } 27$$

Hence, if you want to be 99 per cent confident and are willing to tolerate an error of half a year,

you need to select a sample of 27 students. Similarly, you can calculate the sample size with varying values of σ. Remember the golden rule: *the greater is the sample size, the more accurately your findings will reflect the 'true' picture*.

Sampling in qualitative research

As the main aim in qualitative enquiries is to explore the diversity, sample size and sampling strategy do not play a significant role in the selection of a sample. If selected carefully, diversity can be extensively and accurately described on the basis of information obtained even from one individual. All non-probability sampling designs – purposive, judgemental, expert, accidental and snowball – can also be used in qualitative research with two differences:

1. In quantitative studies you collect information from a predetermined number of people but, in qualitative research, you do not have a sample size in mind. Data collection based upon a predetermined sample size and the saturation point distinguishes their use in quantitative and qualitative research.
2. In quantitative research you are guided by your desire to select a random sample, whereas in qualitative research you are guided by your judgement as to who is likely to provide you with the 'best' information.

The concept of saturation point in qualitative research

As you already know, in qualitative research data is usually collected to a point where you are not getting new information or it is negligible – the data saturation point. This stage determines the sample size.

It is important for you to keep in mind that the concept of data saturation point is highly subjective. It is you who are collecting the data and decide when you have attained the saturation point in your data collection. How soon you reach the saturation point depends upon how diverse is the situation or phenomenon that you are studying. The greater the diversity, the greater the number of people from whom you need to collect the information to reach the saturation point.

The concept of saturation point is more applicable to situations where you are collecting information on a one-to-one basis. Where the information is collected in a collective format such as focus groups, community forums or panel discussions, you strive to gather as diverse and as much information as possible. When no new information is emerging it is assumed that you have reached the saturation point.

Summary

In this chapter you have learnt about sampling, the process of selecting a few elements from a sampling population. Sampling, in a way, is a trade-off between accuracy and resources. Through sampling you *make an estimate* about the information of interest. You do not find the true population mean.

Two opposing philosophies underpin the selection of sampling units in quantitative and qualitative research. In quantitative studies

a sample is supposed to be selected in such a way that it represents the study population, which is achieved through randomisation. However, the selection of a sample in qualitative research is guided by your judgement as to who is likely to provide you with complete and diverse information. This is a non-random process.

Sample size does not occupy a significant place in qualitative research and it is determined by the data saturation point while collecting data instead of being fixed in advance.

In quantitative research, sampling is guided by three principles, one of which is that the greater the sample size, the more accurate the estimate of the true population mean, given that everything else remains the same. The inferences drawn from a sample can be affected by both the size of the sample and the extent of variation in the sampling population.

Sampling designs can be classified as random/probability sampling designs, non-random/non-probability sampling designs and 'mixed' sampling designs. For a sample to be called a random sample, each element in the study population must have an equal and independent chance of selection. Three random designs were discussed: simple random sampling, stratified random sampling and cluster sampling. The procedures for selecting a sample using these designs were detailed step by step. The use of the fishbowl technique, the table of random numbers and specifically designed computer programs are three commonly used methods of selecting a probability sample.

There are five non-probability sampling designs: quota, accidental, judgemental, expert and snowball. Each is used for a different purpose and in different situations in both quantitative and qualitative studies. In quantitative studies their application is underpinned by the sample size whereas the data saturation point determines the 'sample size' in qualitative studies.

Systematic sampling is classified under the 'mixed' category as it has the properties of both probability and non-probability sampling designs.

The last section of the chapter described determinants of, and procedures for, calculating sample size. Although it might be slightly more difficult for the beginner, this was included to make you aware of the determinants involved as questions relating to this area are so commonly asked. In qualitative research, the question of sample size is less important, as your aim is to explore, not quantify, the extent of variation for which you are guided by reaching saturation point in terms of new findings.

For You to Think About

- Refamiliarise yourself with the keywords listed at the beginning of this chapter and if you are uncertain about the meaning or application of any of them revisit these in the chapter before moving on.
- Consider the implications of selecting a sample based upon your choice as a researcher and how you could make sure that you do not introduce bias.
- In the absence of a sampling frame for employees of a large organisation, which sampling design would you use to select a sample of 219 people? Explain why you would choose this design and the process you would undertake to ensure that the sample is representative.
- From your own area of interest, identify examples of where cluster sampling could be applied.
- What determines sample size in qualitative research?
- What is the data saturation point in qualitative studies?