SAMPLING : HOW TO GET SOMETHING FOR A LITTLE

6.1 Introduction

´Sampling´ is the process of choosing a portion of a population to represent the whole population. It contrasts with a ´census´ when every member or unit of a defined population is included. Almost everyone is familiar with situations in which judgements are made about a whole group of items when information is available for only a few of them. If we want to find out about the quality of a sack of rice we would probably only pick out a couple of handfuls and look carefully at these; it would not be necessary to investigate the whole sackful. Similarly, when testing the grade of a shipment of copra, only a small part of the shipment is actually tested, it being assumed that the remainder will be similar.

These simple situations are examples of a common statistical technique known as sampling. We are selecting a number of units from a population, observing some characteristics for these units and then using the results from the sample to estimate values for the whole population. Obviously this will be an important, practical technique, because if we can achieve reasonable results from a sample of observations, then this will be much cheaper and more efficient than having to observe every unit in the population.

In many situations, including the examples we looked at above, the procedure of sampling is quite simple and straightforward. We know that rice selected from one part of the sack will be very similar to that taken from some other part. We will not obtain a very different result if we take two handfuls or twenty. Similarly, if a doctor wishes to take a sample of blood from a patient in order to test for the presence of some disease, he knows that wherever in the body he takes the sample, he will get the same results so he can take just one blood sample.

The important thing about all the examples we have looked at so far is that the population of items under consideration were well mixed up; they did not vary very much. Technically, we can say that these populations are relatively homogeneous, which means that the variability between units is small.

We must realise that, to obtain fisheries statistics, particularly information on artisanal and subsistence fishing which is of interest to many governments in the region, we will have to use sampling methods. It would be too expensive and too time-consuming to try to run a continuous collection of data about all the fishing effort and catch in a country. Unfortunately when we look at the kind of populations we have to deal with, we will expect to find that they are far from homogeneous, and sampling will not be nearly as easy as in our simple examples above. Some villages will no doubt catch far more fish than others, so we cannot use data from one or two villages to tell us about the total fish catch of the country. The catch will no doubt vary from day to day, and from one time of the year to another, so we cannot easily choose data for one day or one week to estimate catch accurately for a whole year; different types of boat, different fishing techniques, etc. will produce different results, and we will have to make allowance for this in designing a sample to derive our estimates.

The populations we usually have to deal with, then, are fairly heterogeneous; there is considerable variability. In this case, sampling

is more difficult. Very often we have very little prior information about the population we want to study; all we know is that it is heterogeneous.

In this topic we will look at different types of samples we can design to give us the estimates we need; a little of the mathematical theory of sampling; ways to measure the accuracy or reliability of the estimates we obtain from our sample; and how to assess the size of sample we will need in order to achieve an acceptable level of reliability in our results.

6.2 Some concepts and definitions

First, it will be useful to define some new concepts which we shall use. These follow on from our previous definitions in Topic 2. There we looked at the terms: statistical unit, observation, characteristic, and population.

Finite and infinite populations

We sometimes need to distinguish between ´finite´ and ´infinite´ populations. A finite population is one which has some limit to its size, e.g. the number of foreign longline vessels operating in a country´s waters; the total catch by artisanal fisheries in a country, and so on. An infinite population is one which has no limit (or is so large that we cannot identify a limit); for example, all the fish in the sea. It is interesting to observe that for infinite populations there is no alternative to sampling in order to make estimates of population characteristics. It is impossible to measure the average fork length of the whole population of skipjack tuna; to do that we would have to catch every skipjack in the sea. All we can do is make estimates based on a sample of fish.

Sampling unit

Elementary units, or groups of units, which are convenient for purposes of sampling, are called sampling units. For example, in a subsistence fishing survey we may find it most convenient to make a selection of villages from which to collect data. The village is the sampling unit.

We must be very careful here to distinguish between the terms ´sampling unit´ and ´statistical unit´. These may be the same, but this is not necessarily so. For instance, if we select a sample of villages, then ´village´ is the sampling unit. But if we then collect our data of catch and effort for each boat in the village, then ´boat´ is the statistical unit.

Thus, we define ´statistical unit´ to refer to the element we wish to collect information about, and the ´sampling unit´ to refer to the element or group of elements which we use as a basis for sample selection.

As we shall see later, we often select samples in two or more stages. For example, we might select a number of villages at the first stage, and then within each selected village, we could choose a sample of households from which to collect data on fishing. In this case we refer to the village as the ´primary sampling unit´ and the household as the ´secondary sampling unit´.

This multiple usage of the word 'unit' can be quite confusing. We also talk about a 'fishing unit' which we define as the smallest discrete, complete unit necessary for a fishing activity. In practice, we can expect the 'fishing unit' to equate to the 'statistical unit'. We will be trying to collect information about the 'fishing unit'.

Sampling frame

In order to select a sample from a finite population we need a list of all the sampling units. We call such a list the sampling frame. The frame must be complete and up-to-date; if any unit is not included then it has no chance of being selected in the sample and this may well lead to inaccuracies in the results.

The preparation of the sampling frame can be one of the most difficult and time-consuming tasks in a sample survey. We also often find that the information that we have available to provide the frame will limit the kind of sample we can select and the results that we can obtain. Information from previous investigations is sometimes suitable, e.g. records from a population census can provide a valuable source of data for sampling frames.

Sample size

The 'sample size' is simply the number of sampling units we select to be in the sample. Obviously the sample size must be less than the number of units in the whole population.

Sampling fraction

The 'sampling fraction' refers to the proportion of the population which is included in the sample. It is usual to refer to a population of 'N' units, and the sample as consisting of 'n' units. The sampling fraction is then: n/N. For infinite populations, the concept of a sampling fraction does not exist.

Parameter

We use the word 'parameter' (or population parameter) to mean the true value of the characteristic of the population which is being estimated. Thus, for instance, in a sample survey of local fisheries in a country, the population parameters we are interested in will probably be the total catch of fish in the country, the average catch per boat day for all boats operating in the country, the proportion of all households which are engaged in subsistence fishing, and so on.

Sampling error

We use results from a sample to make estimates of population parameters. The word 'estimate' indicates that we do not know the exact value of the parameter, but that we hope to be quite close. Obviously, if we do not measure or collect data from every sampling unit within the population we cannot expect to obtain the value of a population parameter exactly. We refer to the difference between the estimate and the true value as the sampling error. In a sense this is the price we have to pay

for only observing part of the population. With a large sampling error then, the estimate obtained from our sample will be inaccurate; if the sampling error is small, the estimates will be close to the true value.

There is, however, a problem, because normally we do not know the value of the population parameter. There would be little point in undertaking a survey to collect results to estimate a value which we already know. It is, therefore, impossible to calculate the sampling error exactly. What we can do is to calculate how large we expect the sampling error to be, provided we select the sample according to certain, well-defined rules. To help understand the idea of sampling error more clearly it will be useful to look at Figure 6.1.

FIGURE 6.1 : REPRESENTATION OF SAMPLING ERROR

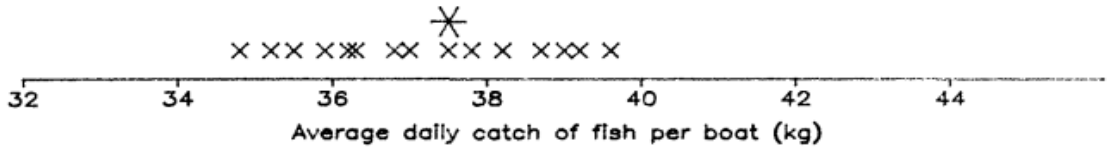


Figure 6.1 represents the results of a sample survey to measure the average daily catch of fish per boat in a country. We use a scale to represent catch and we can plot the value we obtain from a sample as a small cross. The true population value, is represented by the large asterisk. In practice, of course, this value will not be known. Let us assume that the population we are considering consists of 500 boats and that we are taking a sample of 20. With a different sample we get another estimate and we can plot this as another cross. On the diagram we have plotted the results from 15 different samples, but we could have plotted many more. In fact, there are 267×10^{33} different possible samples (that is, 267 followed by 33 zeros), and each of these may well produce different estimates. We could plot all these estimates as a frequency distribution and we would find that the shape would be almost exactly the same as a Normal distribution which we looked at in Topics 3 and 4. This will always be the case, provided the sample size is large enough (say, 20 or more), more or less regardless of the actual distribution of the population.

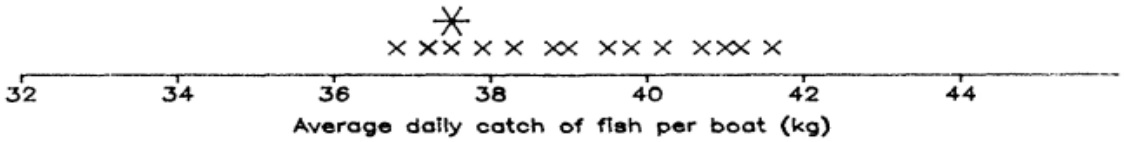
Now, as long as the sample results are clustered around the true value, or, putting it technically, that the mean of the Normal distribution is the population mean, then a measure of the accuracy of the sample estimates is provided by the degree of dispersion of the distribution. We measure dispersion by the standard deviation, but to distinguish between the standard deviation of the population and the dispersion of the sample estimates around the mean, we call the latter the <u>standard error of the estimate</u>.

This is a very important concept in statistics. We will come back to it a little later in this topic, and will give the formulae for calculating the standard error for some principal types of samples.

Bias

In the previous section we saw how the standard error of an estimate is a measure of its precision, provided that the distribution of all the different possible estimates is centred round the true population value. If this is not the case, we say that our sampling scheme is ´biased´ and an example of this is given in Figure 6.2.

FIGURE 6.2 : EXAMPLE OF A BIASED SAMPLE



We are using the same situation as before, but with this sampling scheme we can see that the different estimates are not clustered round the true population mean but around some higher level. If we looked at the distribution of all the possible sample estimates, it would still look like a Normal distribution, but this time the mean of the Normal distribution would not be the same as the population mean. The difference between the two is the bias.

With a biased sample the accuracy of the estimate is not measured just by the standard error; it also includes the bias. Technically, we can measure the accuracy by the square root of the sum of the squares of the bias and the standard error. This will not matter very much as long as we know the value of the bias. In a very few situations we do use biased samples because they may be more accurate in the end, but the important thing is that we know what the bias will be.

In practice, however, bias may well be introduced and we do not realise it, and so we do not know how large it will be. It can arise in several different ways: from problems in preparing the sample frame, from the way the sample is selected, from the way observations are made, from non-response, from mistakes in calculations and also, in some cases, from the way we make the population estimates.

In our example of catch per boat, bias could arise because we have faulty scales which give a wrong reading; because people collecting the data decide to guess the weights, instead of measuring them accurately; because we fail to observe all the fish caught, e.g. by missing out on part of the catch at night; because we have failed to include in our frame some boats which have a different average catch from the boats we have included; or for a variety of other reasons.

Bias also introduces another drawback into samples. In general, we hope that as we increase the size of the sample we improve the estimates by reducing the sampling error. In other words we get more accurate results at the cost of having to make more observations. If the sample is biased, however, it does not matter how large the sample is made; the bias will still be present.

Generally speaking, we may say that it is very important to try to reduce bias, or eliminate it altogether from our surveys.

Non-sampling errors

We use the general term non-sampling errors to refer to all the types of errors and mistakes that can occur when we undertake a survey, other than the basic inaccuracy that is a result of the sampling process itself. Non-sampling errors can arise because of mistakes by enumerators, wrong answers given by respondents, problems with the sampling frame, poor data processing techniques, and many other reasons. They will, of course, happen in complete censuses as well as sample surveys and, in fact, can be a serious problem in these cases because of the much larger nature of the operation. In a sample enquiry, however, it is especially important to try to control these errors, because each sampling unit 'represents' many others in the population; just as we multiply our sample results to estimate population totals, so we multiply the effect of each error. Most statistical textbooks tend to concentrate on techniques for reducing the sampling error and to overlook the operational difficulties of actually carrying out the survey. This is mainly because the effect of non-sampling errors is very difficult to estimate and varies considerably from survey to survey. There is no mathematical technique we can use, as we can to calculate the sampling errors of estimates. When detailed research has been undertaken, however, it has been found that non-sampling errors in some surveys can be at least three times larger than the sampling error. What we have to do, when carrying out any investigation, is to build in as many checks and controls as possible.

6.3 Methods of selecting a sample

6.3.1 Random and non-random samples

First we should differentiate between two types of sample selection - random and non-random. A random method of selection is one which gives each of the units in the population a specified, or calculable (and non-zero) probability of being selected. This is sometimes referred to as probability sampling.

Other methods of sample selection are referred to as non-random, or non-probability. For example, suppose we wish to collect a sample to estimate the total subsistence and artisanal fish catch in a country. If, for reasons of cost and convenience, we were to restrict our sample selection to boats operating within 20 km of our urban centre, we would have a non-random sample. We cannot really expect that estimates obtained from within and near an urban centre are truly representative of the whole population.

A rather similar situation arises with 'judgement' sampling where the sample selected is one which we believe, or feel intuitively, would be representative of the whole population. We may feel confident that (say) two particular islands are 'typical' of the whole country; in other words we make a subjective judgement that we can estimate population parameters by selecting a sample from those islands only. This may be true; on the other hand it may not. There is no way of knowing, or calculating, how well such a sample does in fact represent the population.

All non-random samples suffer from one very serious drawback; there is no mathematical way to calculate the sampling error. A sample based on the

laws of chance, on the other hand, can provide a measure of how precise these estimates are. Thus, we have an objective means of evaluating the results of a survey. This is a most important characteristic, and is the biggest single reason why statisticians prefer to use random, or probability, sampling methods whenever possible. For the rest of this chapter we will concentrate our discussions entirely on random sampling.

There are several different types of random sample, which will be discussed a little later. In some types every unit in the population has an equal chance of being selected; in others some have more chance than others. What all random samples have in common, however, is that every unit has some chance of selection which is known or can be calculated. In a non-random sample this is not so.

6.3.2 The use of random numbers

In order to select units at random we need some kind of random process that produces results with no order or pattern, but where each unit has a known probability of selection. Some examples of such processes are:

(a) tossing a coin;
(b) throwing dice;
(c) selecting numbers out of a hat.

Any of these methods could be used to select a sample, but, in practice, they will be rather cumbersome to use, particularly if the sample size is at all large. Therefore, most people use random numbers from a computer, a calculator, or already prepared tables. An example of a table of random numbers is given in Table 6.1. This table consists of a number of digits and there is absolutely no pattern or order in the way these digits are written down. The table can be read horizontally or vertically. The gaps do not mean anything; they are simply there to make the table easier to read.

TABLE 6.1 : EXAMPLE OF A TABLE OF RANDOM NUMBERS

87 08 83 09 40	14 39 15 99 24	21 85 00 45 54	19 36 18 03 88
88 33 78 20 40	40 24 73 77 70	00 31 84 59 25	06 50 30 95 96
22 50 09 11 00	37 36 51 55 95	83 97 13 75 46	22 77 50 11 72
48 70 56 57 16	24 21 74 91 53	18 05 59 61 74	97 31 82 77 68
93 45 40 93 12	80 88 63 26 93	85 06 19 87 84	37 59 76 16 65
50 76 72 02 39	19 40 69 57 23	09 33 20 70 86	45 13 94 98 39
91 64 01 34 67	13 11 00 32 09	39 76 21 64 29	85 65 14 51 74
33 20 63 71 95	94 13 77 12 94	91 04 41 83 79	72 44 08 12 44
90 59 65 46 78	82 16 45 97 85	57 75 79 96 79	08 16 83 43 99
05 10 93 57 80	32 86 65 26 90	27 45 34 94 46	33 65 35 56 84
92 85 63 26 69	69 81 54 70 56	17 62 43 17 86	78 99 62 34 15
08 50 36 45 82	11 26 54 76 88	86 67 82 21 65	00 82 89 06 09
59 36 77 09 83	78 81 77 93 77	48 44 88 30 37	21 74 02 93 10
05 85 86 43 25	50 76 70 36 32	26 68 54 92 84	90 02 38 77 40
13 46 99 31 30	29 71 70 91 10	99 84 55 31 95	20 90 28 49 78
56 27 09 33 66	79 33 29 50 54	76 94 27 01 45	78 29 66 23 15
54 14 52 11 22	33 39 39 58 30	73 43 59 32 26	43 76 12 99 10
83 01 86 58 89	77 68 87 29 71	49 50 46 53 41	53 52 20 56 53
00 28 17 33 81	42 24 33 55 75	42 70 73 65 16	96 47 17 42 69
52 29 69 59 32	59 40 30 89 12	11 07 18 53 27	13 46 54 85 40
54 43 09 80 68	29 86 65 60 27	87 70 77 45 31	69 12 31 21 79
80 68 13 48 80	84 25 33 70 89	76 61 03 41 57	89 97 07 56 12
28 72 57 80 54	05 80 92 82 65	25 01 74 58 89	39 25 05 57 66
23 48 49 96 00	17 88 90 63 67	02 64 71 12 21	02 29 86 88 54
04 41 27 70 10	49 13 76 99 28	64 14 90 60 69	75 10 97 16 60

Let us look at an illustration of how we use this table, to select a random number between 1 and 63. We choose a random starting point on the table - say the 11th row of the 17th column of paired digits. We will observe that this starting point is the number 99. It is too large for our requirements (i.e. it is greater than 63) so we must reject it and take the next number. If we are working vertically the next number is 82 and the next 74; both are too large so they too must be rejected. The next number, 02, since it falls within our specified range, becomes our random selection. If we need another selection we must continue from immediately after our last selection. Thus, we would have to reject 90 as being too large, and our next selection would be 29.

If we had been working horizontally from our starting point, we would have rejected the first number 99 as before, and would have selected the next number, 62, which falls within our specified range. Our second selection would be the next random number, 34.

We may see from this that we really need rules about using random number tables. For our purposes we will work vertically until reaching the foot of a column, then continuing at the top of the next column, and so on. If random numbers are being extensively used, we would need more precise rules than that.

6.4 Types of random sample

There are many different types of random sample, and we will concentrate on a few of the best known and most commonly used.

6.4.1 Simple random sample

The most basic type of random sample is known as a ´simple random sample´, which can also be written as srs, for short. From a population of N units a sample of n is selected. This is done in such a way that any one of all the possible samples that could be used, is equally likely to be chosen. In effect this also means that every one of the N units has the same chance of being in the sample.

Simple random sampling can be realised by selecting units one by one with equal probability (i) replacing units already selected before the next draw, so that in fact the same unit may be selected more than once, or (ii) without replacing the selected units before the next draw. The former is termed ´srs with replacement´ and the latter ´srs without replacement´. The latter can be shown to provide a more efficient estimate than the former.

It may be noted that srs is not widely used in practice, mainly because some information or other is usually available for all the units in the population and this information can generally be utilised in the selection schemes which are discussed below, to increase the efficiency of the sample design.

6.4.2 Systematic sample

A systematic sample is one in which the sample is selected from a list of the population according to some pre-determined systematic pattern.

Perhaps the most commonly used method is to make selection at regular intervals from the list. For example, to draw a 10 per cent sample we would select every 10th unit, and would do this by drawing a random number between 1 and 10 to choose the first unit. If this were, say, 5, then the units selected in the sample would be the 5th, 15th, 25th, 35th, and so on. With this method there is no chance for various combinations of units to be selected, e.g. it is impossible to select both the first and the second units on the list, as could occur with simple random sampling.

This fact can be turned to our advantage, and systematic sampling can provide a more efficient and more representative sample than could be obtained by simple random sampling. This is achieved by first arranging the units in the list in a suitable order. For instance, if we wished to draw a sample of villages, we might list the villages geographically. This systematic procedure guarantees a very good geographic ´spread´ of selections. It avoids the possibility, which is always present in a simple random sample, that by chance we might select, for example, a higher proportion of villages near an urban centre than is actually present in the population. So, provided the list is ordered in a satisfactory way, we can be more confident of drawing a representative sample. However, the list must not be prepared so that it contains a regularly repeating pattern, as this can lead to a most unrepresentative selection.

Another form of systematic sample, which has been shown to be very efficient, is called a ´Balanced Systematic Sample´. With this method, the population is listed in a suitable order, and selections are made at equal distances from each end of the list. For example, in a list of 100 units, if the first unit is selected by random means, then this would be balanced by also selecting the 100th unit; if the 12th unit were selected, then the 89th unit (i.e. the 12th from the other end of the list) would also be selected, and so on.

In general we can say that systematic sampling is a good method in many circumstances, since it is unbiased, is easy to understand and to operate, and gives us an efficient sample.

6.4.3 Stratified random sample

In simple random sampling the selection of the sample is left to the luck of the draw. No use is made of any knowledge that we possess about members of the population. If we have such knowledge, we should be able to improve upon simple random sampling by using the knowledge to guide us in the selection of the sample.

For example, suppose we wish to estimate the average daily landing per vessel from a fishing fleet at a particular port, by taking a random sample of the fishing boats. If all of the fishing boats are similar, then we can proceed as described before. But if the fleet consists of, say, 100 small canoes and four large motorised boats, obviously our answer will depend greatly upon whether our sample happens to include one or more of these large boats.

In circumstances like this we can often improve the accuracy of our estimates by dividing the population up into groups, or strata, and we can then take a sample from each stratum separately.

In the example we gave above, we would stratify by type of boat - with powered fishing boats in one stratum, and canoes in another. This is the principle behind stratification: we try to have each unit in a stratum as

similar to each other unit as possible (in terms of the characteristic we are measuring). Thus, no matter what unit we happen to select will be representative of other units in the stratum. However, we can make the difference between strata as great as we like, and indeed it is to our advantage to do so. We refer to this as 'low within-stratum variability' and 'high between-stratum variability'. The problem in practice is that we need a suitable sampling frame in order to make the stratification, and this may be a limiting factor.

The most common, and the most obvious, method of stratification is geographic. For a fisheries survey we would almost certainly wish to stratify between high islands and atolls, and between rural and urban areas, for instance.

As well as improving the precision of our overall population estimates, stratification is important for other reasons. We may well need, for example, estimates for different districts or provinces as well as for the whole country. By making each district a stratum we automatically get results for the district. Using stratification also allows us to change the size of the sample in each strata. If one area is very expensive to survey then the sample size can be reduced, and if another area seems to be very variable then it can be sampled more intensively.

Once the strata have been defined, a separate sample is taken from each one, using simple random or systematic sampling.

6.4.4 Multi-stage sampling

The two main types of random sample that we have looked at so far, simple random samples and stratified samples, while being very useful techniques, do have two drawbacks when used in many Pacific countries. Firstly, the cost of collecting the data for each selected sampling unit can be very high, which means that, because the overall budget is usually limited, the sample size has to be reduced. The main reason for this high 'unit cost' is the amount of time that is required to reach scattered units which can often be quite isolated. Obviously this will be much more of a problem with surveys in rural areas. Choosing a simple random, or stratified sample, could well mean that it would be necessary to visit a large number of villages. In some countries this can mean a journey of two or three days simply to obtain one observation. Since a large part of the cost of any rural survey is accounted for by salaries and transportation and since time taken to locate a unit is not productive, it is clear that we need some way of organising the sample in order to reduce the amount of travelling required.

The second major disadvantage is that both types of sample require a complete sampling frame covering the whole population. To carry out a fishing survey in a country in which the sampling unit is to be the boat (or the fishing unit), we would need to prepare a list of every boat (or every fishing unit) in the country for our frame, and this is likely to be quite impractical. In many data collection exercises one of the most difficult problems can be the preparation of a sampling frame.

Multi-stage sampling has been developed to help overcome these two serious drawbacks, although, as we shall see later, this can only be done at the expense of a certain amount of accuracy in our estimates. The basic idea is that, instead of selecting a sample of our final sampling units, we combine these units into groups and then select a sample of these groups.

For example, in our fishing survey we could first of all list villages and select a sample of these. Since we only select a sample of villages, we have immediately cut down on the amount of travelling required to go from place to place. In addition, to select this sample all we need is a list of villages and not a complete frame of all fishing boats. We would then need to make a list of boats in our selected villages, but this is obviously a far simpler task than making a list of all the boats in the country.

We can then undertake a second stage of sampling, selecting boats within the villages already picked out. We have illustrated here two-stage sampling; the first stage was selecting villages and the second stage choosing boats. In principle, we can have any number of stages; and then we refer to multi-stage sampling. For example, we could select islands at the first stage, villages within each selected island at the second stage, boats within those villages at the third stage, and certain days of the month on which to collect data at the fourth stage.

In practice it is probable that some combination of multi-stage (or at least two-stage) and stratified sampling will prove to be the most efficient and cost-effective system we can devise.

6.4.5 Sampling with probability proportional to size

We will not attempt to examine more sophisticated designs in this course, but will mention one particular technique, namely, sampling with probability proportional to size (often referred to as 'p.p.s.' sampling), because this technique is very widely used, especially in the first stages of multi-stage sampling. With p.p.s. sampling, instead of giving each unit an equal chance of selection, we adopt procedures which give larger units a greater chance of selection than smaller units.

We will use an illustration to show how to make selection with probability proportional to size. Suppose in a multi-stage sample we wish to select one of five villages in a district, with the intention of selecting certain fishing units within the selected village at the next stage. Using the selection methods we have described so far we would simply take a random number between 1 and 5 to choose the village.

With p.p.s. sampling, we would need to know the population (or the number of households, or some other measure which is suitable for use as a measure of size) of each village. Even if we do not have a precise measure of size, there are often records available which will be adequate. For example, data from the last census would give us the comparative populations at that time, and would be good enough to be used as estimates of the current population. When we use estimated measures of size we may refer to p.p.e.s. (i.e. probability proportion to estimated size) sampling. We would make the village selection in the following manner:

Village	Population	Cumulative population	Selection range
A	226	226	1- 226
B	705	931	227- 931
C	339	1,270	932-1,270
D	104	1,374	1,271-1,374
E	295	1,669	1,375-1,669

The total population of five villages is 1,669, so we would choose a random number between 1 and 1,669. Any number between 1 and 226 would select village A, and so on, according to the figures shown in the column 'Selection range' above. With this system we are far more likely to choose the largest village, B, than the smallest village, D, because there are 705 different random numbers (from 227 to 931) which would select B, and only 104 numbers (from 1,271 to 1,374) which would select D. In fact the chances of selection are exactly proportional to the population of the villages, as shown.

We should note here that the measure of 'size' which we really wanted is presumably the catch of fish, since that is the characteristic we are trying to measure. We would really like to give probability of selection of villages in proportion to the amount of fish they catch, but that is presumably not known. However, in the case of subsistence fishing, it may be reasonable to expect that fish catch would be roughly proportional to population (at least within a certain geographic area), so we can use population figures as a useful substitute for fish catch as the basis on which to make p.p.s. selections.

The use of p.p.s. sampling, at the first or second stages of multi-stage sampling, is very common. When several units are being selected at once (e.g. if we are selecting 4 villages out of 50), it is the usual practice to make selections <u>with</u> replacement, thus giving a village the chance of being doubly-selected. This is slightly less efficient than selection without replacement - that is to say, it has slightly higher sampling errors, but it makes the estimation process and the calculation of standard error easier.

Although in p.p.s. sampling the probability of selecting sampling units at this stage is unequal, we can still make subsequent selections in such a way that the probability of selection for any final-stage sampling unit is exactly equal, if we wish.

6.5 <u>Principles of calculating standard error</u>

We will introduce this discussion by referring to the calculation for simple random samples, where the population is infinite (e.g. the total number of fish in the sea).

Earlier in this topic, in introducing the concept of sampling error, we noted that if we took a number of separate samples, and calculated $\bar{x}$ in each case, we would finish up with a <u>distribution</u> of values of $\bar{x}$, each of which is a separate estimate of μ. If we were to repeat this process a very large number of times (which, fortunately we will not have to do in practice) we would obtain the theoretical distribution of $\bar{x}$, and we could calculate the standard deviation of this distribution.

We will give the result of this as a theorem, i.e. "If random samples of size n are taken from an infinite population, the theoretical distribution of $\bar{x}$ has a standard deviation of $\sigma/\sqrt{n}$".

We write this as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and we refer to the measure $\sigma_{\bar{x}}$ as the <u>standard error of the mean</u>. To avoid confusion in use of the symbol σ, it is more usual to write $Se(\bar{x})$ to denote the standard error of the mean. The standard error of the mean plays a very important role in statistics, because it measures the variation of the theoretical sampling distribution of $\bar{x}$. In other words, it tells us how much sample means can be expected to vary from sample to sample. We can see that, since the divisor is $\sqrt{n}$, the

standard error of the mean will decrease as we increase the sample size. So the larger we make n, the more reliable will our $\bar{x}$ be as an estimator of μ.

Of course all this is theoretical, and in practice we do not have a large number of samples, each giving a calculation of $\bar{x}$; normally we have one sample only. More importantly, we cannot ever know the value of σ (the standard deviation of the population) of an infinite population. Therefore, we need to modify our formula by replacing σ by an estimate of σ. Fortunately we can do this, because we can calculate s, the standard deviation of the sample. Provided the sample is random and unbiased and the sample size is significantly large, we can expect s to approximate quite closely to σ. However, this may not hold for small samples, for large samples then, we estimate the standard error of the mean to be equal to $s/\sqrt{n}$.

For example, s can be calculated for our yellowfin data as 1.37 kg. Now we can say that from our sample we estimate the mean weight of the total population of fish to be 4.43 kg and that we estimate the standard error of this to be $\frac{1.37}{\sqrt{63}} = 0.17$ kg.

6.5.1 The finite population correction factor

So far we have been discussing samples drawn from infinite populations. When we are sampling without replacement from finite populations we have to make an adjustment to our formula, and this becomes the following:

$$Se(\bar{x}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$$

$\sqrt{\frac{N-n}{N-1}}$ In other words we have multiplied our previous formula by the factor This is known as the 'finite population correction factor' (or f.p.c. for short). The f.p.c. serves to reduce the standard error of our estimate from the value it would have had if we had been dealing with an infinite population, or if we had used sampling with replacement. To demonstrate the effect which this factor has on our estimate of the standard error of the mean, let us consider what the value of this factor would be if we had (a) a sample of 200 observations out of a total population of 40,000 and (b) a sample of 200 observations out of a population of 400 units.

In the first case f.p.c. would be equal to:

$$\sqrt{\frac{40,000 - 200}{40,000 - 1}} = \sqrt{\frac{39,800}{39,999}} = 0.998$$

This is so close to unity that multiplying by it will have virtually no effect on the answer we obtain. In practice, where the sampling fraction is small (say less than 5%), we can ignore the f.p.c. altogether, and we can say that for very large finite populations, or with very small sampling fractions, we will treat $Se(\bar{x})$ as being the same as for our infinite population.

However, in the second case, this factor equals:

$$\sqrt{\frac{400 - 200}{400 - 1}} = \sqrt{\frac{200}{399}} = 0.701$$

Thus, multiplication of our unadjusted calculation by this factor will reduce our estimate of the standard error of the mean by almost 30 per cent, and this is so significant that it certainly cannot be ignored.

In practice it is usual to modify this correction factor slightly. We can note that

$$\sqrt{\frac{N - n}{N - 1}} = \sqrt{\frac{N - 1 + 1 - n}{N - 1}} = \sqrt{1 - \frac{n - 1}{N - 1}}$$

and this is almost equal to

$$\sqrt{1 - \frac{n}{N}}$$

or in other words the square root of one minus the sampling fraction. If we also replace σ by s in our formula (as we did before), because we normally will not know the value of σ, our formula for the estimate of the standard error of the mean can be written as:

$$Se(\bar{x}) = \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

We will find that this is the most usual presentation of the formula for practical purposes.

6.5.2 Confidence intervals

We must next ask ourselves what this really means. It is all very well to say that we estimate the mean weight of all fish in the area as 4.43 kg with a standard error of 0.17 kg, but such a statement on its own will have limited value. Users of the statistics will probably understand perfectly well what the mean of 4.43 kg denotes, but how are they to interpret the value of the standard error?

Fortunately, as we noted earlier, the distribution of calculations of $\bar{x}$ will approximate very closely to a Normal curve, and this will hold true even if the population itself was not distributed normally. It follows that the properties of the standard deviation in respect to the normal distribution, which we mentioned in Topic 4, will apply. Therefore we can say that about 68 per cent of all sample estimates will lie within one standard error either side of the mean, and over 95 per cent will lie within two standard errors.

Of course we usually have only one single sample estimate, not a large number of them, so we need to put our statement into a different form to make it more useful. In practice we say that, provided $\bar{x}$ is an unbiased estimator of the population mean, there is a probability of about 68 per cent (or that we are "68% confident") that the sample mean plus or minus one standard error will include the population mean, and over 95 per cent probability that the sample mean plus or minus two standard errors will do

so. If we assume that our 63 yellowfin comprised a sample representing the total yellowfin population of the area, then there is a 68 per cent probability that the mean weight of all yellowfin in the area is within the range (4.43-0.17) kg and (4.43+0.17) kg, i.e. between 4.26 and 4.60 kg, and that there is over 95 per cent probability that it is within the range 4.09 to 4.77 kg. These ranges are referred to as confidence intervals, and the different probabilities (i.e. 68%, 95%, etc.) are called confidence levels.

Statisticians make most use of the 95 per cent confidence levels in practice, because this is a high enough figure for us to be "fairly sure" of being correct. Thus for the fish data, we interpret our results to mean that we are 95 per cent confident that the true mean weight of the fish lies somewhere in the range 4.09 to 4.77 kg. However, we can never be really sure, and we must never assert that the true mean is within this confidence interval.

If it were decided that for some purposes a 95 per cent confidence level is inadequate, we can make a similar calculation for other levels. For example, the 99 per cent confidence level pertains to an interval of 2.6 times the standard error on either side of the sample mean. So we could say we are confident at the 99 per cent level that the true mean weight of the fish is between (4.43 - 2.6x0.17) and (4.43 + 2.6x0.17) kg, i.e. between 3.99 and 4.87 kg. A similar calculation can be made for any desired level of confidence, and we can look up special tables to find out the appropriate confidence interval for any level.

6.6 Principles for estimating population parameters from sample data

Next we must look at the techniques for making the estimates themselves - that is to say, the way in which we 'expand' the data obtained from the sample to obtain estimates for the whole population. This is in fact a very difficult subject and for surveys with a complex design, involving stratification, multi-stage selection, and probability proportional to size, the computation can become so complicated that processing will almost certainly have to be undertaken on a computer. However, the principles in making unbiased estimations still apply, even though the applications may be difficult.

The estimators commonly used for estimating population parameters are of the form

$$\hat{X} = \sum_{i=1}^{n} w_i x_i$$

where $\hat{X}$ (pronounced X hat) represents the estimate of the characteristic for the population X; x_i is the value of this characteristic for the ith selected 'final-stage' sampling unit; and w_i is the 'weight' applicable to that unit. It is this 'weight' - which is variously referred to as the multiplier, the expansion factor or the raising factor - which provides the main difficulty in the estimation phase.

Incidentally, it will be noted that we referred to x_i as the value for the 'final-stage' sampling unit; it is necessary to have this qualification because in multi-stage sampling we can have different sampling units at different stages, e.g. the village at the first stage and the fishing unit at the second stage, so we have to be careful to define just what we mean by 'sampling unit'.

so. If we assume that our 63 yellowfin comprised a sample representing the total yellowfin population of the area, then there is a 68 per cent probability that the mean weight of all yellowfin in the area is within the range (4.43-0.17) kg and (4.43+0.17) kg, i.e. between 4.26 and 4.60 kg, and that there is over 95 per cent probability that it is within the range 4.09 to 4.77 kg. These ranges are referred to as confidence intervals, and the different probabilities (i.e. 68%, 95%, etc.) are called confidence levels.

Statisticians make most use of the 95 per cent confidence levels in practice, because this is a high enough figure for us to be "fairly sure" of being correct. Thus for the fish data, we interpret our results to mean that we are 95 per cent confident that the true mean weight of the fish lies somewhere in the range 4.09 to 4.77 kg. However, we can never be really sure, and we must never assert that the true mean is within this confidence interval.

If it were decided that for some purposes a 95 per cent confidence level is inadequate, we can make a similar calculation for other levels. For example, the 99 per cent confidence level pertains to an interval of 2.6 times the standard error on either side of the sample mean. So we could say we are confident at the 99 per cent level that the true mean weight of the fish is between (4.43 - 2.6x0.17) and (4.43 + 2.6x0.17) kg, i.e. between 3.99 and 4.87 kg. A similar calculation can be made for any desired level of confidence, and we can look up special tables to find out the appropriate confidence interval for any level.