SAMPLING: PRINCIPLES FOR ESTIMATING POPULATION PARAMETERS FROM SIMPLE 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'.

For the simplest type of samples, where all units have an equal chance of selection, the expansion from sample data to population estimates is quite straightforward. In these circumstances, e.g. in simple random sampling or simple systematic sampling, the multiplier is the same for all selected units and is equal to N/n, which is the inverse of the sampling fraction - sometimes referred to as the sampling interval. All that is needed is to 'raise' sample values by this factor, to obtain population estimates, i.e. $\hat{X} = \frac{N}{n} \Sigma x$.

Population estimates and sampling error for the main types of sample

We are now in a position to look at the formulae whereby we estimate population parameters and calculate sampling errors, for the main types of sample we have been discussing.

6.7.1 Simple random sample

The notation we will use is:

N = Number of sampling units in population
n = Number of sampling units in the sample
X = Population value
$\hat{X}$ = Estimate of population value
$\hat{\bar{X}}$ = Estimate of population mean

(We could also write this, as $\hat{\mu}$, since μ was our symbol for the population mean, but in practice it is more common to write the estimate as $\hat{\bar{X}}$, pronounced X bar hat.)

$Se(\hat{X})$ = Standard error of population estimate
$Se(\hat{\bar{X}})$ = Standard error of estimate of population mean

We then have, for estimates of the population mean:

$\hat{\bar{X}} = \bar{x}$, or $\frac{\Sigma x}{n}$ — In other words we simply estimate the population mean to be equal to the sample mean.

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

and for estimates of the total population :

$\hat{X} = N\bar{x}$, or $\frac{N}{n} \Sigma x$ — That is, we estimate the value of the population total to be the value of the sample total, multiplied by the sampling interval.

$$Se(\hat{X}) = N \cdot \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

For large and infinite populations, the f.p.c. can be omitted. So we have

$$Se(\hat{\bar{X}}) = \frac{S}{\sqrt{n}} \quad \text{and} \quad Se(\hat{X}) = N \cdot \frac{S}{\sqrt{n}}$$

6.7.2 Stratified random sample

k = Number of strata

N_c = Number of sampling units in stratum c, so we have $N = \sum_{c=1}^{k} N_c$

w_c = N_c/N, i.e. the weight, or proportion, of total sampling units in stratum c

n_c = Sample size in stratum c

$\hat{\bar{X}}_c$ = Estimate of the population mean in stratum c

$\bar{x}_c$ = Sample mean in stratum c

Then $$\hat{\bar{X}} = \frac{1}{N} \sum_{c=1}^{k} N_c \, \bar{x}_c$$

or, perhaps more commonly this would be written

$$\hat{\bar{X}} = \sum_{c=1}^{k} w_c \, \bar{x}_c$$

That is to say, we have a <u>weighted</u> mean, with the means of each individual stratum weighted by the proportion of the total number of sampling units in that stratum.

The formula for the standard error is given by :

$$Se(\hat{\bar{X}}) = \frac{1}{N} \sqrt{\sum_{c=1}^{k} \frac{N_c \, (N_c - n_c)}{n_c} \, s_c^2} \qquad \text{or} \qquad \frac{1}{N} \sqrt{\sum_{c=1}^{k} \left(\frac{N_c^2 \, s_c^2}{n_c} \left(1 - \frac{n_c}{N_c}\right)\right)}$$

This may be written, in terms of the weights, w_c, as

$$Se(\hat{\bar{X}}) = \sqrt{\sum_{c=1}^{k} \left(\frac{w_c^2 \cdot s_c^2}{n_c} \left(1 - \frac{n_c}{N_c}\right)\right)}$$

In this formula the f.p.c. $(1 - n_c/N_c)$ can be omitted if the sampling fraction is small (e.g. less than 5%) for <u>every</u> stratum. In practice in stratified samples we quite often need to have fairly large sampling fractions for some strata, and the f.p.c. is an important component in the calculation of the standard error. For example, if there are a few large boats and many canoes in a local fishery, we would endeavour to stratify by type of boat, and would probably include a fairly high percentage of the 'large boat' stratum in the survey.

Let us look at an example where we want to estimate the average fish catch on a certain day from a population of 350 boats, and that we have resources to collect data from 50 boats. We may take a stratified sample by type of boat, and get the following results.

			Survey Results	
Type of boat	Total No. of Boats (N_c)	Sample No. of Boats (n_c)	Est. Av. Catch ($\bar{x}_c$)	Est. Standard Deviation (S_c)
Canoes	278	28	18	7
Small power boats	56	14	32	10
Large power boats	16	8	112	36
Total	350	50		

To estimate average catch we simply substitute in this formula, and we have

$$\hat{\bar{X}} = \frac{1}{350} \sum_{c=1}^{3} N_c \, x_c$$

$$= \frac{1}{350} \quad (278 \times 18 \quad + \quad 56 \times 32 \quad + \quad 16 \times 112)$$

$$= \frac{8588}{350} \qquad = \quad 24.5 \text{ kg}$$

This is a simple weighted mean. The figure of 8588 kg of course represents the total catch of all boats that day.

In similar fashion we can substitute in our formula to calculate the standard error of our estimate, as follows:

$$Se(\hat{\bar{X}}) = \frac{1}{350} \sqrt{\sum_{c=1}^{3} \left(\frac{N_c \, (N_c - n_c)}{n_c} \, S_c^2 \right)}$$

We must note that the sampling fraction is above five per cent in each stratum, so we certainly cannot ignore the f.p.c. We have

$$Se(\hat{\bar{X}}) = \frac{1}{350} \sqrt{\frac{278(250)}{28} \cdot 7^2 \quad + \quad \frac{56(42)}{14} \cdot 10^2 \quad + \quad \frac{16(8)}{8} \cdot 36^2}$$

$$= 1.14$$

So our estimate from the stratified sample is a catch of 24.5 kg per boat, with a standard error of 1.14 kg. In other words we are 95 per cent confident that the true average catch was 24.5 $\pm$ 2.3 kg, i.e. in the range 22.2 to 26.8 kg.

In is worth noting here that the predominant part of the total standard error arose from the first stratum, viz canoes. This may give us a clue that, if we were undertaking another such survey, we might reduce sampling error by increasing the sample size of this stratum. This idea will be explored in more detail in section 6.9.

6.7.3 Multi-stage sampling

Estimation of population parameters for multi-stage samples becomes more complicated because we have to use different weights, or expansion factors, at each stage of the sampling process. We have to build up our estimates at one stage before we can go on to estimating at the next stage.

The situation will be different depending on whether selection at the first stage is made with probability proportional to size or not, and whether first stage selections were made with or without replacement. We will give formulae here only for the simplest situation, but in practice it is more likely that p.p.s. would be used.

Suppose we wish to estimate the total landings of fish along a section of coastline by sampling the landings from the fishing fleet. But we know that there are a number of landing places and many vessels fishing along the coast, and we cannot visit all places. In this case we could resort to Two-Stage Sampling. First, we select at random a convenient number of landing sites from the total sites available along the coast, e.g. suppose there are 8 sites and we select 3 of these. Then:

N = Number of 1st stage units = 8
n = Number of 1st stage samples = 3

Next we select at each of these 3 sites a convenient number of boats from the total number of boats landing at these ports. Then:

M_i = Number of 2nd stage units available in 1st stage unit 'i'
m_i = Number of 2nd stage units sampled from those available in 1st stage unit 'i'

We will assume our data is as follows :

Landing site (n_i)	1	2	3
Number of boats present (M_i)	6	9	7
Number of boats sampled (m_i)	3	3	3
Landings (tonnes)	13	5	12
	9	7	8
	6	10	13
Total landings by sample vessels	28	22	33
S_i	5.3	4.7	2.6

The notation we will use is:

$\frac{n}{N}$ = 1st stage sampling fraction

$\frac{m_i}{M_i}$ = 2nd stage sampling fraction for the ith landing port

Then, if (y_{ij}) is the landing of a particular vessel, i.e. the jth vessel at the ith port, we have

$\bar{y}_i$ = Average landing per vessel at site$_i$ = $\frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$

$\hat{y}_i$ = Total estimated landing at site$_i$ = $M_i \cdot \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$

$\hat{y}$ = Total estimated landings for all selected sites = $\sum_{i=1}^{n} \left(M_i \cdot \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \right)$

$\hat{Y}$ = Total estimated landing for entire coast = $\frac{N}{n} \left[\sum_{i=1}^{n} \left(M_i \cdot \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \right) \right]$

Estimation of the standard error of this total estimate is a little more complicated, because we have two sources of variation in our calculation. Firstly, we have the variation in landings by the vessels at any landing site, and then we have the variation in landings between the landing sites. Our total sampling error is therefore the sum of these two factors.

It is for this reason that we said earlier that the multi-stage sampling is only achieved at the expense of some loss in accuracy of our results. We have saved in costs and ease of data collection by concentrating all our efforts on 3 out of 8 landing sites, instead of taking observations at all of them, as we would in other types of sample design. Now we have to pay the penalty for using this grouping, or 'clustering' of observations, by making allowance for the variation between landing sites.

Our formula for the variance, i.e. the square of the standard error, becomes

$$V(\hat{Y}) = \frac{N(N-n)}{n} \frac{1}{n-1} \left(\sum_{i=1}^{n} \hat{y}_i^2 - \frac{(\sum_{i=1}^{n} \hat{y}_i)^2}{n} \right) + \frac{N}{n} \sum_{j=1}^{m_i} \frac{M_i(M_i-m_i)}{m_i} \cdot S_i^2$$

where $S_i = \frac{1}{m_i-1} \left(\sum_{j=1}^{m_i} y_{ij}^2 - \frac{(\sum_{j=1}^{m_i} y_{ij})^2}{m_i} \right)$

Substituting the data from our example in this formula we have

$$V(\hat{Y}) = \left(\frac{8x5}{3} \times \frac{1}{2} \times 221\right) + \frac{8}{3} \left[\left(\frac{6x3}{3} \times 12.3\right) + \left(\frac{9x6}{3} \times 6.3\right) + \left(\frac{7x4}{3} \times 7.0\right) \right]$$

$$= 1473 + 673 = 2146$$

The first term in this expression represents the variation between landing sites. It will be noted that the contribution to variance due to this term is much greater than that which is due to difference among second-stage units within the first-stage units. This means that if we were going to carry out the exercise again, it would be preferable to increase the number of fishing sites sampled, even if it meant that we had to reduce the number of vessels sampled at each site.

We have in our example $Se(y) = \sqrt{2146} = 46.3$, therefore

95% confidence limits are $531 \pm 2\ (46.3)$
$= 531 \pm 92.6$

i.e. our estimate of the total landings, with 95 per cent confidence, is between 438 and 624 tonnes.

6.8 Ratio estimation

Ratios of population totals of two characteristics are as important as, and sometimes more important than, the population totals themselves. For example, we may obtain from a sample survey information on total catch and on total effort, and we may be more interested in estimating catch per unit effort than we are in estimating either of the totals. For surveys covering two different points of time, we may be more concerned with finding out whether total catches have gone up or down, than with measuring the level at any one point of time.

We use the term ratio estimations to refer to the method of estimating a ratio of the population by means of a ratio of the unbiased estimators of two characteristics. Thus, if $\hat{Y}$ and $\hat{X}$ are unbiased estimators of Y and X respectively, then an estimator of the population ratio $R=Y/X$ is given by the ratio estimator $\hat{Y}/\hat{X}$—R.

In situations where the actual population value for the denominator (X) is known, it might be felt that to estimate the desired population ratio (R) all that is necessary would be to estimate the numerator (Y); thus $\hat{Y}/X$—R. However, if the estimators of the numerator and the denominator are approximately proportional (that is, if the two characteristics are highly linearly related with the line passing through the origin), then an estimator based on the ratio of the estimators of the numerator and the denominator is a more efficient method.

The method has a possible application in fisheries statistics, if a country manages to conduct a complete census of all local fisheries to measure total catch in one year, and then wishes to monitor changes to the total catch in following years by a sample survey of fishing units.

If we start by discussing a simple random sample, the approach we have discussed so far would be to select n fishing units out of the total of N units in the population, measure the catch, y, of each sample unit, and estimate total fish catch as $\hat{Y}=\frac{n}{N}\Sigma y$.

In the ratio estimate, we would ascertain the catch by each of the n fishing units in the present survey, and also the catch which those same units obtained in the census year.

We use the notation here

X = total catch by all fishing units in the census year
y = catch by sample units in the present year
x = catch by those same units in the census year
$\hat{Y}$ = estimated total catch in the present year

Then we could say

$$\hat{Y} = \frac{\Sigma y}{\Sigma x} \cdot X$$

This is a ratio method of estimating total catch by measuring for each sample unit the ratio of catch between two different periods.

To put this in simple terms, we might come to a conclusion from our survey as follows: The total catch at last year's census was 420 tonnes. The catch by our sample boats has increased by five per cent since then, so we will estimate that the catch by all boats has increased by five per cent. Therefore, we estimate the catch this year to be 420+21 = 441 tonnes. This is the line of reasoning we follow in ratio estimation: we calculate some ratio derived from a sample, in order to estimate population parameters.

If we examine the estimator $\frac{\Sigma y}{\Sigma x}.X$, it is clear that X is not derived from a sample, so the sampling error will depend solely on the sampling error of the ratio $\frac{\Sigma y}{\Sigma x}$, with X having only the effect of a constant multiplier.

The formula can be modified to give a different weight or expansion factor to the selected units. It becomes

$$\hat{Y} = \left(\frac{\sum_{i=1}^{n} w_i \, y_i}{\sum_{i=1}^{n} w_i \, x_i} \right) \cdot X$$

where w_i is the weight, or expansion factor of the ith unit.

The actual formula for calculating standard errors for ratio estimates is complicated, and is beyond the scope of this course. We will simply observe that in some circumstances the ratio method of estimation leads to substantially lower standard errors. However, it is a biased method. Fortunately the bias tends to be negligible for moderately large samples. In many practical applications indeed, it is so small compared with the advantage gained in reducing the sampling error, that the ratio estimate is preferred to the unbiased estimate.

6.9 Determining sample size

A very important part of planning a sample survey is deciding how many units to sample, i.e. the size of the sample. The size of the sample will depend upon the resources we have available, i.e. the number of trained collectors, money, data analysis facilities, time, and on the degree of precision we need in the results. If we require only approximate estimates, a small sample may suffice, but if we require more exact estimates, a large sample may be needed.

In order to determine the size of a sample we generally need the following kinds of information:

(a) the total resources (money, manpower, etc.) available for the investigation;

(b) the cost of collecting data from one unit;

(c) the expected variability in the population;

(d) the required precision.

It is unlikely that we will ever be able to have exact information on any of these, but we can often use approximations, estimates and data from previous surveys in order to gain some idea of the size of the sample we need. In some cases, results from a pilot survey will provide estimates of costs and also give an idea of variability. Assuming, therefore, that we have some idea of this kind of information we shall look at different types of sample to to see how we can determine the sample size.

6.9.1 Simple random sample

Let us start by looking at a simple example. Suppose we wish to estimate the average per capita fish consumption per week from a simple random sample of people. We would like to have 95 per cent confidence that our estimate will be within plus or minus 0.2 kg per week. From last year's study we may have an estimate that the standard deviation is about 0.5 kg. Now we know that the range $\hat{\bar{X}}-2Se(\hat{\bar{X}})$ to $\hat{\bar{X}}+2Se(\hat{\bar{X}})$ equates to our 95 per cent confidence limits. For a simple random sample, we have approximately $Se(\hat{\bar{X}})=s/\sqrt{n}$ where s is the estimated population standard deviation. Therefore, we have

$$2Se(\hat{\bar{X}}) = 0.2$$

and $$Se(\hat{\bar{X}}) = \sqrt{(0.5^2/n)}$$

Thus $$\sqrt{(0.5^2)/n} = 0.2/2$$

which gives $n = 0.25/0.01 = 25$, i.e. we need to sample about 25 people.

Quite often, instead of specifying a tolerance value (such as the 0.2 kg above), we say that we want the true result to be within a certain percentage of our estimate. For example, we might want to have 95 per cent confidence that the true value will be within the range $\hat{\bar{X}} \pm 5\%\hat{\bar{X}}$. In general, we can write this as $\hat{\bar{X}} \pm p\hat{\bar{X}}$ and we have to specify p.

We know that

$$2Se(\hat{\bar{X}}) = p\hat{\bar{X}} \qquad \text{and} \qquad Se(\hat{\bar{X}}) = \sqrt{\left[\left(1-\frac{n}{N}\right)\frac{S^2}{n}\right]}$$

If we solve these two equations for n, we have

$$n = \left(\frac{2s}{p\hat{\bar{X}}}\right)^2 \cdot \frac{1}{1 + \frac{1}{N}\left(\frac{2s}{p\hat{\bar{X}}}\right)^2}$$

To see how this works, we shall use the following example: N=430; we know from previous studies that $\hat{\bar{X}}=19$ and $s^2=85.6$ and we specify p=0.10 (or 10%).

Then we have

$$n = \frac{2^2 \times 85.6}{(0.10 \times 19)^2} \cdot \frac{1}{1 + \frac{1}{430} \frac{2^2 \times 85.6}{(0.10 \times 19)^2}}$$

$$= 78$$

If we needed a precision of one per cent instead of ten per cent, we would have:

$$n = \frac{4 \times 85.6}{(0.01 \times 19)^2} \cdot \frac{1}{1 + \frac{1}{430} \frac{2^2 \times 85.6}{(0.01 \times 19)^2}}$$

$$= 411$$

that is, we would have to sample nearly the entire population. In practice, we would not take a sample, but rather a census, if the required sample size is calculated to be as close to the total population as that.

We note that the second part of the expression for n, $\frac{1}{1 + \frac{1}{n}\left(\frac{2s}{\hat{p}\hat{\bar{X}}}\right)^2}$ is the finite population correction factor. If $\frac{1}{n}\left(\frac{2s}{\hat{p}\hat{\bar{X}}}\right)^2$ is less than 0.05, then we can safely approximate n by the expression $\left(\frac{2s}{\hat{p}\hat{\bar{X}}}\right)^2$ because then the term $\frac{1}{1 + \frac{1}{n}\left(\frac{2s}{\hat{p}\hat{\bar{X}}}\right)^2}$ is very close to 1.

6.9.2 Stratified sample

In stratified random sampling we have to decide both the total sample size and how to allocate sample size in each stratum. We have four main methods for choosing overall sample size and strata sample sizes. Which one we use depends on how much prior information we have about the variability in the population and strata and also on the costs of sampling each unit. We will examine each of the four methods.

We introduce the symbols

C_c = the cost of sampling one unit in the cth stratum

d = maximum acceptable error (such as the 0.2 kg used in our example for simple random sampling)

z = a variable whose value depends on the degree of precision required, expresed in number of standard deviations from the mean. If we want confidence limits of 95 per cent, we use the value 2 for z, or if we want confidence limits of only 68 per cent, we use the value 1. Other values of z, for different confidence limits, can be found in tables of the Normal probability distribution.

Equal allocation

In equal allocation we take the same number of samples from each stratum. Therefore we have only to determine the overall sample size n, and we sample $n_c = n/k$ units from each stratum.

The formula for calculating n is:

$$n = \frac{k \sum N_c^2 s_c^2}{N^2 \frac{d^2}{z^2} + \sum N_c s_c^2}$$

We use the method of equal allocation in the following situations:

(i) When the total numbers of sample units N_c in each of the k strata are more or less equal;

(ii) When the stratum variances (s_c^2) and cost per sampling unit (C_c) do not vary much from stratum to stratum;

(iii) When there is no prior knowledge of stratum variances (s_c^2) or cost per sampling unit (C_c).

Proportional allocation

The total sample size is allocated among strata in proportion to the size of each stratum. For example, if stratum 3 contains 25 per cent of the population, then 25 per cent of the overall sample will be taken in stratum 3. The formula for calculating the allocation per stratum is

$$n_c = \frac{N_c}{N} \cdot n = w_c \cdot n$$

To calculate the overall sample size, n,

$$n = \frac{N \sum N_c s_c^2}{N^2 \frac{d^2}{z^2} + \sum N_c s_c^2}$$

We use proportional allocation when the stratum total number of units, N_c, varies from stratum to stratum, and when either the stratum variances and cost per sampling unit do not vary much from stratum to stratum or when we do not have any prior knowledge about stratum variances and costs.

Neyman allocation

The Neyman method is named after the famous statistician who developed the method. We use the Neyman method when the stratum variances s_c^2 vary from stratum to stratum. The formula for allocation of sample size in stratum c is

$$n_c = \frac{N_c s_c}{\sum N_c s_c} \cdot n$$

The overall sample size, n, is calculated thus :

$$n = \frac{(\sum N_c \, s_c)^2}{N^2 \frac{d^2}{z^2} + \sum N_c \, s_c^2}$$

Optimum allocation

To use optimum allocation we need some prior knowledge of the stratum variances and cost per sampling unit in each stratum. If both stratum variance, s_c^2, and cost per sampling unit, C_c, vary from stratum to stratum, then we will obtain the greatest precision in our estimates if we use the formulae for optimal allocation. The sample size in each stratum is

$$n_c = \frac{N_c \, s_c}{\sqrt{C_c}} \cdot \frac{1}{\sum \frac{N_c \, s_c}{\sqrt{C_c}}} \cdot n$$

and the formula for overall sample size n is

$$n = \frac{(\sum N_c \, s_c \, \sqrt{C_c}) \cdot (\sum \frac{N_c \, s_c}{\sqrt{C_c}})}{N^2 \frac{d^2}{z^2} + \sum N_c \, s_c^2}$$

6.9.3 An example of sample size allocation

To see how we may apply the four methods in practice, we will calculate sample sizes using each method for the following problem.

Suppose that along a certain coast, the 100 places where fish are landed can be roughly graded into three classes according to the weight of fish landed. During a typical week, the weights landed are

Large landing places : 45, 59, 87, 41, 71, 25, 9, 69, 10, 7

Medium landing places: 17, 13, 19, 26, 1, 8, 27, 11, 12, 26
5, 8, 10, 16, 16, 4, 16, 16, 13, 29
14, 25, 29, 27, 20, 25, 2, 7, 3, 12

Small landing places : 2, 6, 7, 0, 1, 2, 1, 5, 4, 7
8, 9, 3, 2, 5, 4, 2, 0, 2, 8
5, 3, 8, 9, 8, 9, 1, 6, 5, 3
3, 4, 7, 5, 5, 3, 2, 4, 6, 1
6, 2, 5, 1, 0, 3, 8, 0, 4, 3
3, 5, 5, 0, 7, 0, 9, 7, 9, 0

Calculations on the complete census of weights landed show the following

		N_c	s_c	$\bar{X}_c$
c=1:	Large landing places	10	28.91	42.30
c=2:	Medium landing places	30	8.57	15.23
c=3:	Small landing places	60	2.81	4.20

In our formulae, we will use k=3, since landing places are divided into three strata. We want to have 95 per cent confidence that our final estimate from the stratified random sample will be within two units of the true total landed weight, therefore d=2 and z=2.

If we suppose for the purposes of optimal allocation that each unit in strata 1 and 2 (large and medium landing places) costs \$10.00 to sample, and each unit in stratum 3, the small landing places, costs \$20.00 to sample, then C_1=10, C_2=10, C_3=20. Of course, with different costs we will get different sample size allocations from those worked out in this example.

We now have all the information we need for the formulae to find n, n_1, n_2, and n_3 for each of the allocation methods. We will not show the calculations but the results are given in Table 6.2.

TABLE 6.2 : TOTAL AND STRATA SAMPLE SIZE FOR FISH LANDINGS USING FOUR DIFFERENT METHODS OF ALLOCATION

Method	n_1	n_2	n_3	n
Equal	8	8	8	24
Proportional	5	16	32	53
Neyman	9	9	6	24
Optimal	10	10	5	25

The equal and proportional methods do not take into account the differences in standard error among strata. In the present example, the differences are large and, in proportional allocation, the overall sample size is more than twice that of all other methods. The large sample size is required because no weighting is given to sampling from the strata with the largest standard errors and so the overall sample size has to be increased to obtain the required degree of precision. Equal allocation does not require an exceptionally large sample size because rather large samples are taken from the strata with smallest numbers (N_c), but greatest standard errors, i.e. strata 1 and 2.

Neyman allocation takes into account the standard errors of the strata. We see that 9 units out of 10 should be sampled from the large landing sites because of the large standard error in this strata, 9 out of 30 in the medium landing sites and only 6 out of 60 in the small landing sites where the standard error is small. The particular example of optimal allocation given here produces a similar sample allocation to that of the Neyman method, except that the relatively high cost of sampling the small landing places reduced the sample size in stratum 3 and slightly increased that of the other 2 strata. Different cost values would give different allocations.

In practice, where the calculations of sample size indicate that 8 or 9 out of 10 units should be sampled in a stratum, we would not sample but take a census of the stratum. Our overall estimated total landings would then have sampling errors due only to the strata which were sampled. Another example of a case where certain strata are completely enumerated and others are sampled only is the estimation of tuna catch. We try to get a census of large-scale foreign and domestic fishing vessel catches but we usually have to be content with sample estimates only of small-scale local catches.

In general, the recommended method of allocation is that method which uses the maximum information available. Therefore, if we have some estimates of the standard errors and per unit sampling costs in each stratum, we should use optimum allocation. If we have no idea of sampling costs, but we do have estimates of standard errors, then Neyman allocation is recommended. Equal and proportional allocation are used when costs and standard errors are not known. In the landing place example, equal allocation produced quite a reasonable allocation because of the inverse relationship between size of stratum and standard error. In other cases where the relationship is different, e.g. where standard error is proportional to stratum size or where no relationship exists, equal allocation will not be as good as proportional allocation.

6.9.4 Some conclusions about sample size

We usually undertake a sample survey because the costs of a complete census are likely to be too great. Usually money and other resources are very limited, and if we are undertaking any kind of statistical investigation we have to make sure that it is carried out as efficiently as possible. In this section we have seen how, for different types of samples, we can decide how large a sample we need for different purposes. The various formulae are just techniques for using prior information for better planning. Very often the information we have is vague, but these formulae can still provide a means for using this.

We have been looking at the problem of deciding on the best sample size by using information about just one variable. The problem is that in most of the surveys we undertake we are interested in collecting data on several characteristics. We can make the sample design optimal for one of them, but this does not guarantee that it is optimal for the rest. In practice, there is not a great deal we can do about this; it would obviously be impossible to choose a different sample for every variable. If we have sufficient information for several variables, then we can try to find a sample design that is almost optimal for all of them. This will be a compromise for each variable, but, otherwise, all we can do is base the sample on the most important variable and hope that it is not too inefficient for the others.

6.10 Concluding remarks

One final point on the whole topic of sampling needs to be made, and it is a very important one. All our calculations and formulae in this topic have been based on the assumption that the sample is random and unbiased. When we claim 95 per cent confidence that the true population mean is within two Se either side of $\bar{x}$, we have implicitly assumed that $\bar{x}$ is an unbiased estimator of X.

We discussed bias earlier in this topic, and we know that in some surveys there are substantial biases which we cannot eliminate. In the rather unlikely event that we are able to measure the bias, we can still quote confidence limits. Thus, if a bias B exists, i.e. if we know that the amount of displacement of the distribution of $\bar{x}$ away from the true position of the population mean is +B, then our 95 per cent confidence limits become $(\bar{x} - B - 2Se)$ to $(\bar{x} - B + 2Se)$.

However, in the great majority of situations we will not have any measure of bias, even if we are aware that it exists. It follows that if we know or suspect that there are substantial non-sampling errors in a

survey, it is a dangerous and misleading practice to express the results in terms of the mean and standard error, without any other qualification. The very fact that we publish a figure for standard error is likely to lead users of the statistics into believing that this is an accurate portrayal of the extent of errors in the results, and they are likely to assume that no other errors are present, unless we make it clear that this is not so.