SUMMARY

Tables are a very common way of summarizing categorical data in sociological research. It is important that they are presented clearly, using the standard conventions about their arrangement and format.

A table may show that there is association between two variables, that is, that the proportion of respondents in the categories of one variable varies according to the categories of the other variable. Association can be revealed by comparing column percentages and measured using any of several measures of association including phi (for 2×2 tables), Cramér's V (for larger tables), tau (when one variable is dependent on the other) and gamma (when the variables are measured at the ordinal level).

It is also possible to analyse the interaction in tables of three variables using a procedure called elaboration. This involves comparing the associations in each of the partial tables obtained by controlling for one of the three variables.

EXERCISES

Exercise

1. The data in Exhibit 9.28 are taken from the General Household Survey, and consist of three variables for 50 randomly selected respondents in employment. The three variables are:
   (a) Participation in an occupational pension scheme (1 = not in scheme; 2 = in scheme), `INOCCPEN`
   (b) Whether receiving less (1) or more (2) than median earnings, `POOR`
   (c) Whether male (1) or female (2), `SEX`

   Construct a table to display these data, properly labelled. Then see what conclusions you can draw from the table.
2. Exhibit 12.18 shows the social mobility table that was percentaged to obtain Exhibit 9.13. Using the counts in Exhibit 12.18, a pocket calculator and the rule for calculating expected counts (page 211), calculate the table expected if there were no association between father's social class and respondent's class. Compare the expected counts with the data in Exhibit 12.18 and explain why some cells have larger counts than would be predicted if there were no 'class inheritance'.
3. Using SPSS, crosstabulate `MEDINS` (whether covered by a private medical insurance scheme) by `SEX`. Construct a properly labelled table suitable for publication and add a paragraph in which you interpret and comment on the table.

SAMPLING AND INFERENCE 10

CONTENTS

Suppose you wanted to find the average income of working women in the UK. Interviewing every working woman would obviously be a huge and expensive task. Fortunately, results for the whole country can be obtained accurately enough for all practical purposes by finding the incomes of just a sample, and then inferring from the sample to the population as a whole. Measuring the income of the working women in the sample gives an estimate of the income of all working women, provided that some basic precautions about how the sample is obtained are satisfied.

The fact that quite small samples can give reasonable estimates about whole populations is one of the discoveries that made surveys and opinion polls possible. Before statisticians had perfected the mathematics that lie behind the process of inference in the 1920s, it was assumed that accurate measures – of poverty, for instance – could only be obtained by exhaustive surveys. Nowadays, however, market research companies regularly obtain commercially useful estimates of the opinions of the 59 million adults in the United Kingdom by interviewing samples numbering between 1000 and 2000 people.

It is quite rare for it to be sensible to conduct a **census**, that is, to obtain data from everyone in the population. By measuring just those in a sample, time and money can be saved that are better used on other aspects of the research, while still obtaining sufficient accuracy for valid conclusions to be drawn. This chapter is about assessing how good an estimate of the characteristics of a whole population can be, using measurements of a sample.

RANDOM AND REPRESENTATIVE SAMPLES

The theory behind statistical inference depends on using a random (or probability) sample to estimate the characteristics of a population. A **random sample** is one in which people are selected to be in the sample at random and every individual has a chance of being included. A random sample is not the same as a **representative sample** (one in which individuals are included in proportion to the number of those in the population like them). For example, if in the population as a whole, 27 per cent of women work full-time, 22 per cent work part-time, and the rest have no paid job, a representative sample of 100 women would be one in which 27 are full-time workers and 22 are part-timers. A random sample of 100 women would be one in which 100 women were picked out of the population entirely at random. There is no guarantee that the randomly selected women would include exactly 27 full-timers, although we might be surprised if the number were very different.

SAMPLING IN SPSS

In the 1995 General Household Survey data set there are 2415 women. Let us treat these women as the population from which to sample. We will first find the percentages of women who work full-time, work part-time and have no paid job in the data set as a whole. To do this using SPSS, refer to Chapter 2, **Select Cases**. The condition for case selection in this example is `sex = 2`. Make sure that you also select **Unselected Cases Are Deleted**. This is important for the random sampling step below to work correctly (see Exhibit 10.1)

The variable that measures work status is called `WKSTATE`. As in Chapter 9, we will first recode `WKSTATE` into a new variable, `TIME`, which has two categories, full- and part-time working (see Exhibit 9.2). Then, we select **Analyze|Descriptive Statistics ▸|Frequencies** ..., and then select the `TIME` variable. Exhibit 10.2, taken from the SPSS output, shows that 27.4 per cent of women work full-time and 21.9 per cent work part-time.

Second, we select a sample of 100 women at random from the data set.

In SPSS, select **Data|Select Cases** ...|**Random Sample of Cases** and then click on **Sample** ..., to see the dialog box in Exhibit 10.3. In this dialog box,

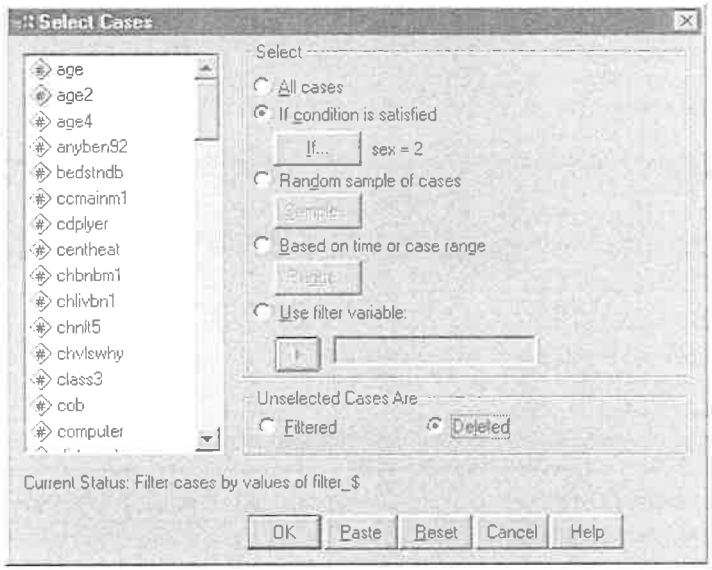

Exhibit 10.1 ***Select Cases** dialog box*

TIME Full or part-time

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Full-time	662	27.4	55.5	55.5
	2.00 Part-time	530	21.9	44.5	100.0
	Total	1192	49.4	100.0	
Missing	System	1223	50.6		
Total		2415	100.0		

Exhibit 10.2 *The variable `TIME` from the GHS 1995*

select **Exactly** 100 cases from 2415, the number of women in the sample. After pressing **Continue** in this dialog box, we return to the **Select Cases** box (Exhibit 10.1). This time, make sure that **Unselected Cases Are Filtered.**

Once again, we ask SPSS for a display of frequencies for the variable `TIME` to discover the percentages of full and part-time working women in the sample of 100. Exhibit 10.4 shows the result from this one sample of 100. In the sample, although full-time workers are fairly represented at 27 per cent, compared to the whole data set, part-time workers are over represented at the expense of those not in employment (who are coded as system missing).

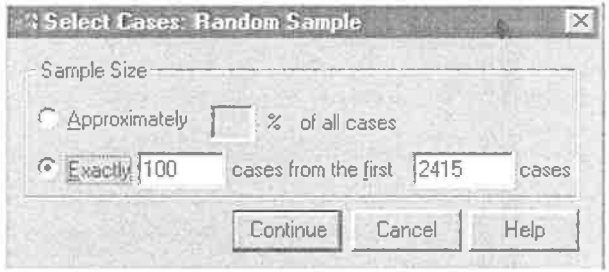

Exhibit 10.3 ***Select Cases: Random Sample*** *dialog box*

TIME Full or part-time

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Full-time	27	27.0	50.9	50.9
	2.00 Part-time	26	26.0	49.1	100.0
	Total	53	53.0	100.0	
Missing	System	47	47.0		
Total		100	100.0		

Exhibit 10.4 *The variable* `TIME` *for a sample of 100 women*

SAMPLE AND POPULATION MEANS

Although having a representative sample rather than a random sample would be the better option, because it would precisely mirror the characteristics of the population, representative samples are almost always impossible to obtain. This is because the sample needs to be representative not just on one, but on *every* characteristic that could be relevant. Usually we neither know what these relevant characteristics are, nor could we find the right sample members even if we did. Fortunately, a random sample has none of these difficulties and can be used, to make inferences about a population even though there is no guarantee that a random sample is representative. The value of a random sample is that, although one cannot be sure how representative it is, statistical theory will give us an estimate of the chances that it will be seriously unrepresentative. In other words, while we may not be exactly correct in the inferences we make from a random sample, at least we will know how likely it is that we are going to be very wrong.

Although it is unlikely that any particular random sample will be exactly

representative of the population, the characteristics of samples will tend, on average, to resemble those of the population. Consider finding the mean weekly gross earnings of working women in the UK. If we draw a random sample of working women and compute the mean weekly income of those in this sample, we might find the mean to be £184. We can use this as an estimate of the mean income of the population. However, we must remember that although the estimate is likely to be close to the actual mean earnings of the population, it will not be exactly equal to the population mean unless the random sample happens by chance to be representative. The difference between the estimate from the sample and the true value in the population is called the **sampling error**.

If we take another random sample, the mean earnings of the women in this second sample might be £187. Taking more and more samples, we would find that the means of these samples cluster around the true, population mean. In fact, if the means from a large number of samples are plotted on a graph (called a **sampling distribution**) such as Exhibit 10.5 they will always form the symmetrical bell-shaped curve known as the **normal distribution** (see Chapter 7). In Exhibit 10.5, the x or horizontal axis shows the mean of each sample, and the y or vertical axis shows the number of samples with that mean.

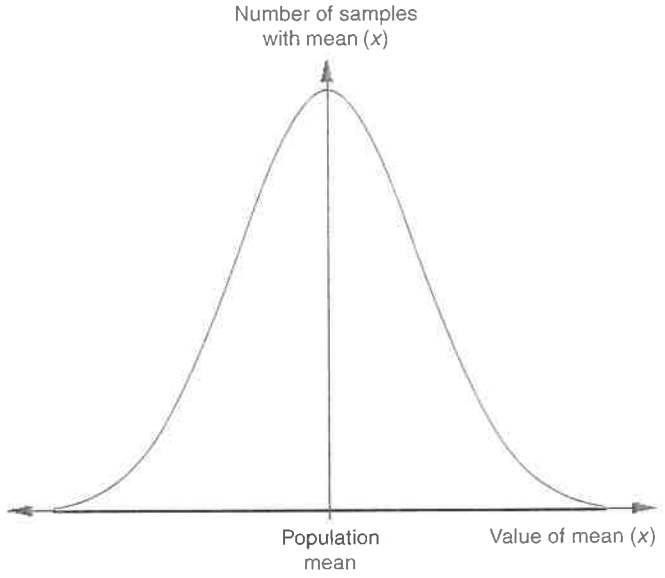

Exhibit 10.5 *The distribution of sample means tends to be normal*

OBTAINING RANDOM NUMBERS

In order to obtain a random sample, we need a numbered list of the population (for example, the electoral roll, or a list of all the pupils in a school) and a set of **random numbers**. Then those people whose numbers are in the list of random numbers are selected to be in the sample. But where do the random numbers come from? They are more difficult to obtain than you might think – the first few numbers that come into your head will certainly not be random. One way of generating random numbers is to use some chance physical event: for example, the National Lottery uses the chance that a particular numbered bouncing ball will escape from a cage. Another way of generating random numbers is to use the times of radioactive decay. But these procedures are obviously inconvenient for social surveys. A practical alternative is to use the random numbers that are printed in books of statistical tables (e.g. Fisher and Yates, 1974).

Nowadays, the usual way is to use a computer to generate random numbers. In fact, computers cannot generate random numbers because they contain no random processes. Instead a 'pseudo-random number generator' is used. This is a complex arithmetic formula that has been shown by statistical tests to yield a sequence of numbers from a given starting value (the **random number seed**) that look as though they were random. The advantage of using a pseudo-random number generator, in addition to its convenience, is that the formula always produces the same sequence for the same seed value, and a different sequence for each different seed. This means that one can carry out repeatable experiments based on random numbers just by using the same seed value.

PLOTTING SEVERAL SAMPLE MEANS USING SPSS

Using the General Household Survey data set as the population, we can draw a number of separate small samples of working women and calculate the mean income of each sample. These means can then be plotted and should result in a graph something like (but not exactly the same as) Exhibit 10.5.

An example of what is obtained is shown in Exhibit 10.6. To produce this we first selected only working women. This was done by selecting **Data|Select cases . . .|If condition is satisfied|If . . .** with the selection condition as `sex = 2 and wkstate ≤ 2`. The **Unselected Cases** were deleted.

Then 100 women were sampled from this subset, by selecting a random number of cases as in Exhibit 10.4, but choosing **Exactly** 100 cases from 1192.

Then mean weekly earnings were calculated for the sample. This was done by selecting **Analyze|Descriptive Statistics ▸|Descriptives**. The variable `EARNINGS`, usual gross weekly earnings, was used. This gave the mean earnings of one sample of 100 women. Then 19 further samples were selected and their mean earnings were found in the same way.

Finally, the means from each sample were then re-entered as the values into a new SPSS data window of a new variable and a histogram plotted (Exhibit 10.6).

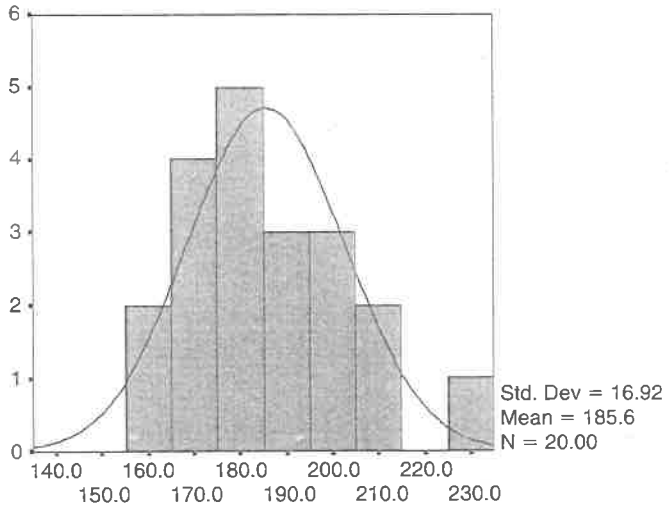

Exhibit 10.6 *Plot of the mean earnings of 20 samples of 100 working women (with a normal curve superimposed)*

The graph is not identical to the normal curve in Exhibit 10.5 because the latter is a theoretical curve obtainable only after taking the means of an infinite number of samples. Exhibit 10.6, based on only 20 samples, looks like a 'lumpy' version of the theoretical curve.

THE CENTRAL LIMIT THEOREM

Exhibit 10.6 showed the distribution of sample means obtained from a number of experiments. Each experiment contributed one sample mean. By a fundamental theorem of statistics, called the **central limit theorem**, it can be proved that the distribution of sample means always approximates to the bell-shaped normal distribution, provided that it is based on a sufficient number of samples each large enough in size.

All normal curves are the same basic shape, differing only in the location of their centres (given by the mean) and in their degree of spread (given by the standard deviation). This means that a normal curve can be defined by just two

parameters: the mean of the distribution and its standard deviation. The mean of the distribution indicates where the centre of the curve is located (the mean runs through the middle of the bell) and the standard deviation indicates how spread out the curve is.

If we were able to plot an infinite number of sample means, the centre of the curve would be at the population mean. For example, the mean of the distribution of the mean incomes of lots of samples of working women would be approximately equal to the mean income of the population of working women. The standard deviation of the normal curve obtained from means of samples has a special name: the **standard error of the mean**. Suppose that we had drawn a large number of samples of working women, each of size N (e.g. each of 100 women as in the previous section) and had found the average earnings of each sample. The central limit theorem tells us that the mean earnings of the samples are normally distributed and that the mean of the means (that is, the centre of the bell) will be approximately equal to the population mean.

CALCULATING THE STANDARD ERROR OF THE MEAN

The central limit theorem also tells us what the standard error of the mean, $SE(\bar{x})$ is. It is given by the formula

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{N}} \qquad (10.1)$$

where σ (sigma) is the standard deviation of the population and n is the number of people in the sample. In practice, it is unlikely that you will know what the standard deviation of the population is, but it can be approximated by using the standard deviation of the sample provided that the sample size, n, is large (more than about 100).

CONFIDENCE INTERVALS

Researchers never actually do draw lots of random samples in the way we have been describing in order to plot a sampling distribution. But knowing that we could do so and that the distribution of means will always tend towards a normal curve is very useful when it comes to making inferences from a single sample. We can use the shape of the sampling distribution to give an estimate of how accurate inferences based on that sample are likely to be. The shape of the curve tells us that sample means are most likely to fall near the middle of the bell (i.e. near the population mean) and rather unlikely to be very far from the middle. As we shall see, we can quantify this using the standard error of the mean.

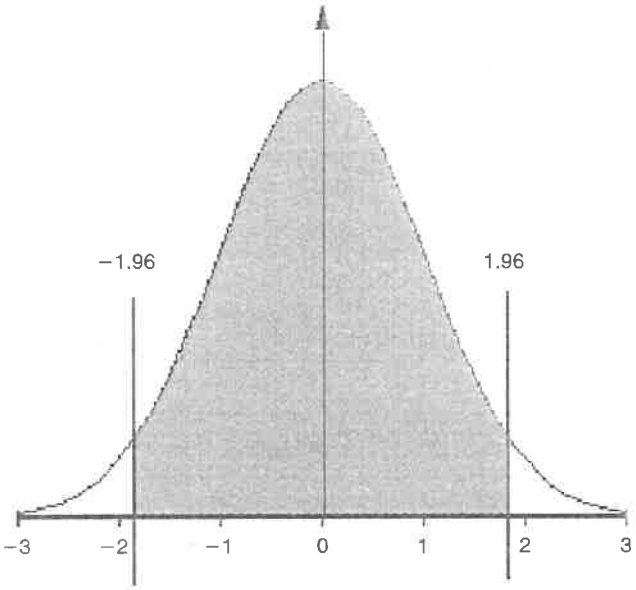

Exhibit 10.7 *The normal distribution with mean 0 and standard deviation 1, with 95 per cent of the area shaded*

For all normal curves, 95 per cent of the area of a normal curve falls between plus or minus 1.96 standard deviations from the mean (see Exhibit 10.7). The value 1.96 comes from the shape of the curve and can be found in statistical tables that describe the normal distribution. When the normal curve is the distribution of sample means, this can be interpreted as saying that 95 per cent of all the means of random samples taken from the same population will fall within the range plus or minus 1.96 standard errors from the population mean. There is thus a 95 per cent chance that the mean of any one sample will fall into the shaded area of Exhibit 10.7. Of course, that also suggests that 5 per cent of the time (100−95 per cent) the random sample will be so unrepresentative that its mean will be more than 1.96 standard errors away from the population mean.

What is rather more interesting from a practical point of view is that, by the same argument, 95 per cent of the time the population mean will fall within plus or minus 1.96 standard errors of a sample mean. So, for example, if we found that the mean income of working women in a particular random sample was £184, we could then go on to infer that there is a 95 per cent chance that the mean income of the population of working women is somewhere between £184

minus 1.96 standard errors and £184 plus 1.96 standard errors. The range from the mean minus 1.96 standard errors to the mean plus 1.96 standard errors is known as the 95 per cent **confidence interval**, since we can say with 95 per cent confidence that the population mean will be within this interval. Thus, one can use the standard error of the mean to find the confidence interval around the mean of a sample and then infer from that one sample that the population mean is very likely to fall within the interval.

PUTTING IT ALL TOGETHER

We can now work through an example of calculating a confidence interval for the mean income of working women in Great Britain. The first step is to select those women in the GHS data set who have some paid employment. Then we can get SPSS to find (using the **Explore** procedure) the number of such women in this sample, N (1054 with valid information about their incomes), the mean income of these women, $\overline{X}$ (£183.94), and the standard deviation of the distribution of these incomes, s (£146.92). Using equation (10.1) and the standard deviation of the sample, s, to estimate the standard deviation of the population, σ, the standard error is:

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{N}} \approx \frac{s}{\sqrt{N}} = \frac{146.92}{\sqrt{1054}} = 4.53 \qquad (10.2)$$

We can conclude that we can be 95 per cent confident that the mean earnings of the population of working women in Great Britain is within the interval from £183.94 − (1.96 × 4.53) = £175.06 to £183.94 + (1.96 × 4.53) = £192.82. This is illustrated in Exhibit 10.8, which shows the confidence interval surrounding the sample mean.

CONFIDENCE LEVELS

In the previous example, we inferred the population mean income from evidence based on a random sample. We know that different random samples are likely to give us different estimates of the population mean income. But by using a confidence interval we can indicate that, despite the uncertainty of random samples, only in 5 per cent of samples will the population mean be outside the confidence interval, while for 95 per cent of samples the mean will be inside the confidence interval.

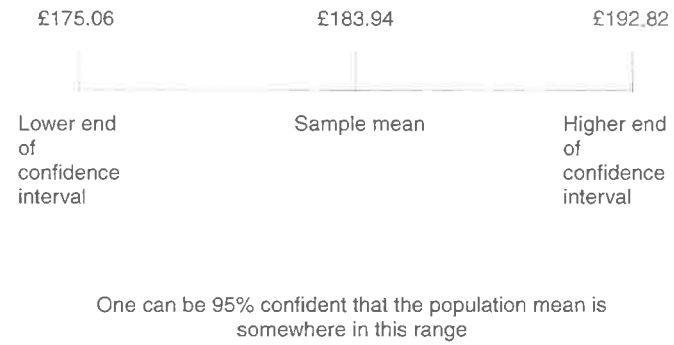

Exhibit 10.8 *A confidence interval*

Although being 95 per cent sure that the population mean is within the confidence interval is usually good enough for most issues of sociological interest, it is possible to use other probability values. For example, 99 per cent and 99.9 per cent are also used, especially in medical statistics where being highly confident of estimates is sometimes very important. The probability of including the population mean within the confidence interval is called the **confidence level** and is something that the researcher has to choose before working out a confidence interval.

In order to be confident that, for example, the population mean is within the confidence interval for 99 per cent of samples, the confidence interval has to be rather wider than it would be for 95 per cent confidence. This is because the proportion of the normal curve that is included in the interval has to be increased from 95 to 99 per cent. As was noted earlier, 95 per cent of the normal curve is included within 1.96 standard deviations of the mean. The corresponding figure for 99 per cent is 2.58. This value can be obtained from the table of the normal distribution in Appendix B. To use the table in Appendix B, first choose the confidence level you want to use, for example 99 per cent. This means that we need to find the interval that covers 99 per cent of the normal curve (see Exhibit 10.9, column A). If the shaded area of the curve includes 99 per cent of the area, 99/2 per cent or 0.495 of the area lies between the mean and the end of the interval. The table in Appendix B (a portion of which is shown in Exhibit 10.9) indicates that the ends of the intervals (z) are located at 2.58 and −2.58.

The confidence interval for the mean income of working women for a confidence level of 99 per cent is therefore from £183.94 − (2.58 × 4.53) = £172.25 to £183.94 + (2.58 × 4.53) = £194.68.

z	Area between the mean and z	Area beyond z	Area between − z and + z (i.e. twice col. A)
1.96	0.4750	0.0250	0.95
2.58	0.4950	0.0050	0.99

Exhibit 10.9 *Views of the normal curve for two selected values of z*

THE STANDARD DEVIATION OF A SAMPLE

In order to work out a confidence interval, we need to know the standard error of the mean. This is given by the central limit theorem as $\sigma/\sqrt{N}$ (see equation (10.2)). Unfortunately, we do not often know the value of the standard deviation, σ. However, it can be calculated: the standard deviation, σ, of some characteristic of a population is the square root of the average of the squared differences between the observations and the population mean:

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} \qquad (10.3)$$

(see also Chapter 5).

However, this equation for σ involves the population mean, which is also not usually known! The way out is to use the sample mean as an estimate of the population mean. However, basing the standard deviation on the sample mean can be shown to yield an underestimate of the standard deviation of the population. This can be corrected by using a formula in which the denominator is not N, but $N - 1$:

$$s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{N - 1}} \qquad (10.4)$$

In practice, it does not make much difference whether you use N or $N - 1$ as the denominator when N is greater than 100.

Procedure for confidence intervals around a mean

- A confidence interval around a mean indicates the range within which a population mean is likely to be found.

When to use a confidence interval:

- Confidence intervals are useful when estimating a population mean from statistics about a random sample.

What you need to calculate a confidence interval:

- The size of the sample (N).
- The mean of some characteristic of the sample ($\overline{x}$).
- The sample standard deviation, s, of that characteristic of the sample.
- The points of the normal distribution appropriate to the desired confidence level to be used.
- Select a confidence level (usually 95 per cent).
- Calculate the standard error of the mean from the formula

$$SE(\overline{X}) = \frac{s}{\sqrt{N}}$$

- Calculate the top and bottom points of the interval. For instance, the 95 per cent confidence interval is defined by

$$\overline{X} \pm 1.96 \times SE(\overline{X})$$

Use 2.58 instead of 1.96 for a 99 per cent confidence level.

FACTORS INFLUENCING THE WIDTH OF THE CONFIDENCE INTERVAL

If the distribution of means has a small standard error, the bell-shaped normal curve is narrow. The confidence interval within which 95 per cent of sample means fall is therefore relatively narrow and we can be relatively precise about the value of the population mean. Conversely, if the standard error is large, we will have to be relatively vague about the population mean. The size of the standard error is therefore important in determining the precision of estimates of the population mean.

The formula for standard error, equation (10.1), indicates several interesting things:

1. The standard error depends directly on the size of the standard deviation of the population, σ. The more variability there is in the population, the larger the standard error and therefore the wider the confidence interval.

2. The standard error is inversely proportional to the square root of the sample size, n. Thus, the larger the sample, the smaller is the standard error and the narrower is the confidence interval. In other words, as you would expect, you can be more precise about the population mean if you use a larger sample to estimate it.
3. Because the standard error is inversely proportional to the square root of the sample size, there is a law of diminishing returns for increasing sample size. Multiplying the size of the sample by a factor of 4 only reduces the width of the confidence interval by a factor of 2.
4. The standard error and therefore the confidence interval do not depend on the size of the population, but only on the size of the sample. For example, using a sample of 1000 would achieve exactly the same precision in estimating the mean income of women working in the city of Chester as in estimating the mean income of women working throughout the United Kingdom.
5. The standard error does not depend on the shape of the distribution of the variable whose population mean is to be estimated, but only on its standard deviation. This is because, regardless of the shape of the distribution of the variable in the population, the sampling distribution of the means of random samples always approximates to a normal curve.

The last point, that the means of random samples are normally distributed even if the variable is not, can be illustrated with the distributions of women's earnings.

In a previous section we plotted the distribution of sample means of women's earnings from small random samples taken from the 1995 General Household Survey (Exhibit 10.6). As expected, we found that the distribution approximates to a normal curve.

While the distribution of means follows the normal distribution, the distribution of earnings has a very different shape (Exhibit 10.10): in fact, it is skewed to the right (positively skewed: see Chapter 5), so that the peak is at a relatively low income level and there is a long tail, showing that many women have low incomes and few have high incomes.

This negatively skewed shape for income is typical, not only for women in the UK, but for all groups throughout the world. Different countries, with different economic systems, vary in the degree of skewedness of their income distributions, but in none does the distribution follow the normal curve.

CONFIDENCE INTERVALS FOR A PROPORTION

What proportion of women in the UK have dishwashers? The national proportion can be estimated from a random sample using the same principles as we used to estimate the mean income of working women. As before, we find the proportion of dishwasher users in a sample, and then infer that the proportion in the population is within a confidence interval around the sample proportion. The

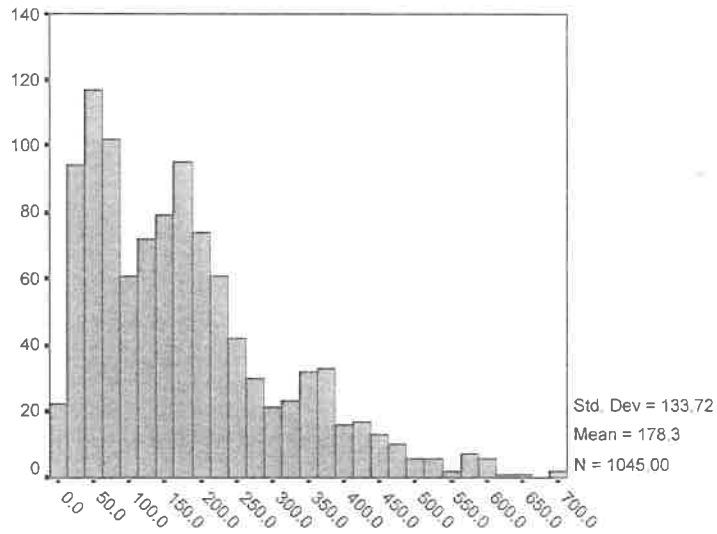

Exhibit 10.10 *Usual gross weekly earnings for working women*
Source: ONS, 1995

only difference is that the standard error is calculated using another formula, appropriate for estimating proportions.

SPSS shows that of the 2415 women in the General Household Survey, 536 own, or are in families that own, a dishwasher. The other 1879 do not have a dishwasher. The proportion who have a dishwasher is therefore the number who have a dishwasher (536) divided by the number in the sample (2415):

$$p = \frac{536}{2415} = 0.222$$

The standard error of the distribution of sample proportions is given by the formula

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} \qquad (10.5)$$

where p is the proportion who have a dishwasher (and therefore $1 - p$ is the proportion who do not have one). Remember that $p(1 - p)$ means p multiplied by $(1 - p)$; multiplication signs are conventionally omitted in formulae like these. Notice the similarity between this formula and the one for the standard