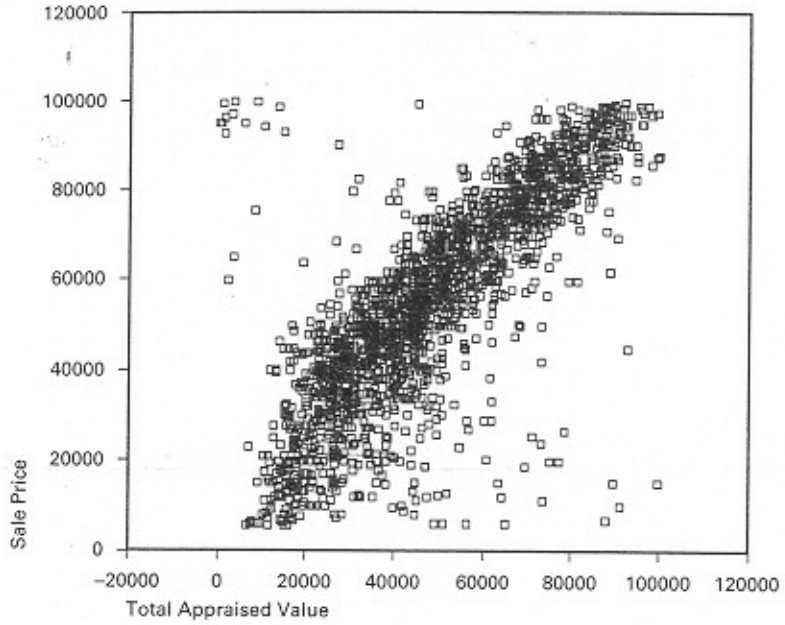


(a) Draw the regression line on the scatter gram.
(b) Find a house that fits the trend closely, indicate it on the scatter plot, and find its estimated sale price directly from the graph, then using the regression equation.
(c) For that house, give also its total appraisal value, and its actual sale price.
(d) Find the difference between the actual sale price and the estimated sale price (use the estimate from the regression equation).
(e) Repeat steps (b), (c), and (d) for a house that does not fit the trend.

# 9 INFERENTIAL STATISTICS: ESTIMATION

The purpose of this chapter is to explain the basic reasoning of inferential statistics, and then to show how confidence statements are to be made and interpreted. The calculations of the margins of error and the relationship between the confidence level and the margin of error are also shown.

After studying this chapter, the student should know:

- the meaning of inference in statistics;
- the notion of margin of error and probability of error;
- how to produce and interpret confidence statements involving means or proportions;
- how to determine the margin of error using either the table or the formulas;
- how to determine the size of the sample needed to achieve a certain precision;
- that the degree of precision increases with the risk of error.

## Inferential Statistics

We have seen in the first chapter that there are two main branches of statistics, descriptive statistics and inferential statistics (refer to Figure 1.6). Chapter 3 was devoted to descriptive statistics. We are now going to study two main techniques used in inferential statistics, **estimation** (see Figure 9.1) and **hypothesis testing**, which are two distinct ways of drawing conclusions about a whole population when only a sample is known. This chapter will be devoted to estimation, and the next one to hypothesis testing.

Recall that **the purpose of inferential statistics is to draw conclusions about a whole population on the basis of information that has been collected on a sample**. In formulating such a generalization, we have to settle two issues that are closely related.

The first issue has to do with the **precision** of the results. Because the generalization is some kind of (educated) guess, it is never very precise. Therefore, we will have to introduce **a margin of error** in our statement, a term that will be defined

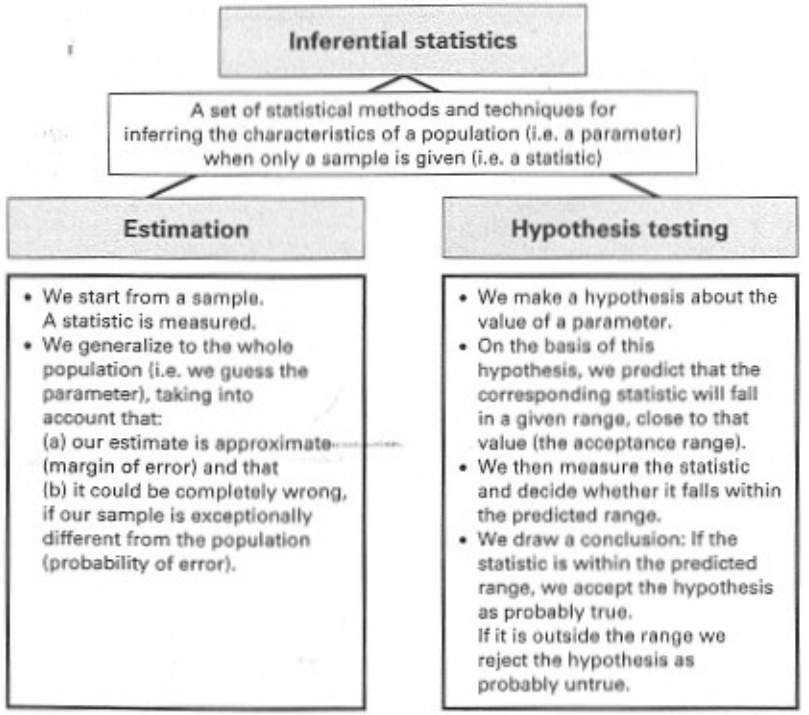


Figure 9.1 **Inferential statistics**

precisely below. For instance, if 45% of the sample of individuals who were interviewed answered Yes to some question, and if that sample of people is really representative, we estimate the percentage of people in the general population who would also answer Yes to be *around* 45%, not exactly 45%. May be somewhere between 44% and 46%, or between 43% and 47%. We will learn below how to determine this margin of error.

The second type of difficulty results from the randomness of the sample. We could be unlucky and hit a random sample that includes a large number of exceptional cases. Such a sample would not be representative, even if it had been selected at random. Only a small percentage of samples are likely to differ a lot from the general population, but the fact is that this possibility is very real. In order to take this possibility into account, we include in every inference a **probability of error**, which can be set at 10% or 5% or even 1%. Usually, the researcher sets out the risk he or she is willing to take when making a statement, and makes the inference on the basis of that level of risk. The precise way this is done will be explained below.



## The Logic of Estimation: Proportions and Percentages

Suppose you select 200 students at random in your college, and ask them whether they approve or not a decision taken by the school administration about discipline in the college. Suppose also that 76% of them declared that they approved the decision. On that basis you are trying to guess what the percentage would be for the whole student population in your college, which comprises, let us say, 2400 students. What would you say?

You could say that 76% of the population approves of the decision. However, you can never be sure that this figure is accurate. It would be safer to say that you expect the corresponding percentage for the population to be *around* 76% rather than exactly 76%. You could say you expect it to be somewhere between 75 and 77%. Or somewhere between 74 and 78%.

The statement that results from your reasoning when doing an estimation is called a **confidence statement**. It is constructed as shown in the following example.

**Example of a confidence statement:**

> The poll, conducted on 1030 individuals last week, showed that 37% of adult Canadians listen to the news on TV. These results are accurate up to ± 4%, and are reliable 95% of the time. (fictitious data)

Let us examine the various elements that are included in the statement. They have been underlined, and they are explained below.

| | |
|---|---|
| **The population** | The population here consists of all adult Canadians. Every confidence statement must specify clearly the population to which it applies. |
| **The sample** | The sample consists of 1030 individuals taken from the population. These are the ones that have been interviewed. On the basis of their answers, the results were extended to the whole population. |
| **The variable measured** | The variable measured here is whether the television is used as a source of information for news. |
| **The measured percentage (the statistic)** | The survey has shown that 37% of the people interviewed (that is, the sample) listen to the news on TV. This percentage was *measured* as part of a survey. |

| | |
|---|---|
| **The estimated percentage (the parameter)** | On the basis of that survey it is *estimated* that 37% ± 4% of the whole population listens to the news on TV. In other words, the estimation is that the percentage of people in the whole population that listens to the news on TV is somewhere between 33% and 41%, not exactly 37%. The middle point of that interval is 37% and this is called the **point estimate**. |
| **The margin of error** | The margin of error is ±4%. This is the degree to which the point estimate is accurate. When generalizing to a whole population, some accuracy is lost. The statement above says that the percentage of people getting their news from the TV is accurate up to + or − 4%. This is why the estimated percentage is not exactly 37% but somewhere between 33% and 41%. We will see below how this margin of error is calculated. |
| **The level of confidence** | The level of confidence here is 95%. It is a measure of how certain the results are. In other words, we are saying that 95% of the time, the sample we pick is sufficiently representative of the whole population to allow us to make a generalization. Details of that calculation will be discussed below. |
| **The probability of error** | This is the risk that the sample on which the estimation has been based was misleading and more different from the general population than expected. If the level of confidence is 95%, the risk is 5%. The level of confidence and the probability of error add up to 100%. |

An important question has been left unanswered: How do we determine the margin of error and the probability of error?

The margin of error and the probability of error are closely linked. To explain this link, let us examine a familiar situation. It is a hot summer day. Two friends are arguing:

> 'I am sure it must be around 36° Centigrade, today. It is so hot!'
> 'Are you saying it is *exactly* 36°?'
> 'No. I'm saying it is probably *around* 36°. May be 35° or 37°. Something like that. I am almost sure.'



> 'Would you bet that your guess is correct and that the temperature is between 35° and 37°?'
> 'No. If you want to bet, I would say the temperature is between 34° and 38°. I am sure it must be within that range. I am ready to bet that it is within that range.'

What is going on in this discussion is that the first person is not ready to bet that the temperature is between 35° and 37°, and he figures out that there is a high risk of being wrong. However, he is more confident that the bet is correct when a wider margin of error is included. In that example, the risk of being wrong and the margin of error are not determined accurately. They are established on the basis of impressions. By contrast, in statistical inference, the level of confidence and the margin of error are determined precisely on the basis of a rigorous mathematical reasoning. However, the link between the two follows the same logic: if you want to make a guess with a high level of confidence, increase the margin of possible error. Give a wider range of possible answers: you will be more confident that the correct answer falls within that range. This relationship can be expressed in any of the following ways.

> To make estimations with a high level of confidence, we need to give a wide margin of error.
> Or: To diminish the probability of error, we need a wider margin of error.
> Or: In formulating an estimation, narrower margins of errors will necessarily imply higher probabilities of error.
> Or: Estimations that provide a wide range for the parameter can be done with a smaller risk of error than estimations that provide a narrower range.
> Or: Smaller margins of error are accompanied by greater risks of error.
> Or: Higher levels of confidence are accompanied by larger margins of error.

All these statements are logically equivalent and they express the relationship between the level of confidence and the margin of error in a confidence statement.

## Estimation of a Percentage: The Calculations

The relationship between the level of confidence and the margin of error can be proven mathematically. Such a proof is beyond the scope of this text, but we can at least examine how it is *expressed* mathematically. Let us say that a survey involves a sample of size $n$, and that the proportion found in the sample is $p$. We can prove

that in formulating a confidence statement about a proportion, the margins of error can be calculated with the formulas shown in Table 9.1.

Table 9.1 **Calculation of the margin of error**

| If you want to be sure of your results at a | you must allow for a margin of error of |
|---|---|
| 90% level of confidence | $\pm 1.64\sqrt{\frac{p(1-p)}{n}}$ |
| If you want to be sure of your results at a | The margin of error is |
| 95% level of confidence | $\pm 1.96\sqrt{\frac{p(1-p)}{n}}$ |
| If you want to be sure of your results at a | The margin of error is |
| 99% level of confidence | $\pm 2.58\sqrt{\frac{p(1-p)}{n}}$ |

Notice that the $p$ used in the formula is a proportion, not a percentage. You can now verify that the statements made on the previous page are correct. Examine the various formulas carefully. They all look alike except for the coefficient that precedes the square root. As the level of confidence increases, the coefficient is higher, and it produces a wider margin of error.

Now look carefully at the numbers themselves. Do they ring a bell? Have we encountered these numbers before? You may recall that we have encountered them when studying normal distributions:

- 90% of all the data in a normal distribution falls within ± 1.64 standard deviation from the mean,
- 95% of all the data falls within ± 1.96 standard deviation from the mean, and
- 99% of all the data falls within ± 2.58 standard deviations from the mean.

For the 95% level of confidence, the margins of error corresponding to various sample sizes have been computed and presented in Table 9.2. It gives the approximate margins of error for various sample sizes and various values of the percentage. It can be used instead of the formula given above.

This is how you read the table: Suppose that in a survey of 539 people, it turns out that 62% of them answered Yes to some question. In the table, the closest column to 539 is the 500 column, and the closest percentage to 62% is the 'Near 60' percentage. The corresponding margin of error is underlined in the table: it is equal to ± 5%. What this means is that your estimate for the whole population will be 62% ± 5%, which is the same as saying it is somewhere between 57% and 67%.



**NOTE:** (for those who are not scared by a mathematical reasoning)

It is not a coincidence that these same figures show up again in this section. Indeed, suppose we had a population made of two subgroups A and B, with subgroup A forming a proportion $p$ of the general population. If we formed all possible samples of size $n$ taken from that population, and counted the proportion of people from group A in each of these samples, the set of all such proportions would constitute a distribution called the **sampling distribution**. We could prove the following: the sampling distribution is a normal distribution and its standard deviation, called the **standard error**, is equal to $\sqrt{\frac{p(1-p)}{n}}$. It follows that 95% of the values of that distribution fall within ± 1.96 standard deviations of that distribution of sample proportions (that is, standard errors). But these values are the sample proportions and each one refers to one sample of size $n$. This explains the figures in Table 9.1.

Table 9.2 **Margins of error for the estimation of a percentage, at the 95% confidence level.**

| Population | Sample size | | | | | | |
|---|---|---|---|---|---|---|---|
| Percentage | 100 | 200 | 400 | 500 | 800 | 1000 | 1500 |
| Near 10 | 7 | 5 | 4 | 3 | 3 | 3 | 2 |
| Near 20 | 9 | 6 | 5 | 4 | 3 | 3 | 3 |
| Near 30 | 10 | 7 | 5 | 5 | 4 | 3 | 3 |
| Near 40 | 10 | 7 | 5 | 5 | 4 | 4 | 3 |
| Near 50 | 10 | 7 | 5 | 5 | 4 | 4 | 3 |
| Near 60 | 10 | 7 | 5 | 5 | 4 | 4 | 3 |
| Near 70 | 10 | 7 | 5 | 5 | 4 | 3 | 3 |
| Near 80 | 9 | 6 | 5 | 4 | 3 | 3 | 3 |
| Near 90 | 7 | 5 | 4 | 3 | 3 | 3 | 2 |

### *Proportions and Percentages*

The explanations given above apply equally to percentages and to proportions. The only difference is that a proportion is calculated out of 1 whereas a percentage is calculated out of 100. Thus, by multiplying a proportion by 100 we get the corresponding percentage, and by dividing a percentage by 100 we get the corresponding proportion. Some care must be given to the formulation of confidence statements in order not to confuse percentages and proportions. A given statement can be formulated either way. We could say, for instance, that an estimated percentage is 37% ± 4% or, equivalently, that the estimated proportion is 0.37 ± 0.04.

### *Point Estimates and Interval Estimates*

You may have noticed that we have formulated the estimation in two different ways, one involving a *single value together with a margin of error*, and the other one in the form of a *range*. These two formulations are called, respectively, a point estimate and an interval estimate.

The **point estimate** in the preceding example is: 'The estimated percentage is 62% (± 5%).'

The **interval estimate** is: 'The estimated percentage is between 57% and 62%.'

These two formulations are equivalent and one can convert one into the other.

### *Formulation of the Level of Confidence*

The level of confidence can be formulated as a percentage (for instance 95%) or as a ratio, as in 'These results are accurate 19 times out of 20.' The two formulations are equivalent, because if you multiply both numbers by 5 you get 'These results are accurate 95 times out of 100.'

For a level of confidence of 90%, the equivalent formulation would be: 'These results are accurate 9 times out of 10.' There is no similar simplification for a level of confidence of 99%.

## Estimation of a Mean

The estimation of a mean follows exactly the same logic as that of percentage, except that the calculation of the margin of error is done with the help of a different formula. Here is an example.

> The poll, conducted on 1030 individuals last week, showed that adult Canadians watch the television an average of 4.3 hours every day. These results are accurate up to ± 0.1 of an hour, and are reliable 95% of the time. (fictitious data)

Let us examine the various elements that are included in the statement. They have been underlined, and they are explained below.

| | |
|---|---|
| **The population** | The population here consists of all adult Canadians. |
| **The sample** | The sample consists of 1030 individuals taken from the population. These are the ones that have been interviewed. |
| **The variable measured** | The variable measured here is the daily number of hours spent watching television. |



| | |
|---|---|
| **The measured mean (the statistic)** | The survey has shown that the people interviewed (that is, the sample) watch television for 4.3 hours (that is, 4 hours and 18 minutes) every day on the average. This average was *measured*. |
| **The estimated mean (the parameter)** | On the basis of that survey it is *estimated* that the population of adult Canadians spends on the average 4.3 hours daily watching television, with a margin of error of one-tenth of an hour (6 minutes). In other words, the estimation is that the **average** daily time people spent watching television is somewhere between 4 hours and 12 minutes and 4 hours and 24 minutes. |
| **The margin of error** | The margin of error is 6 minutes. We will see below how this margin of error is calculated. |
| **The level of confidence** | The level of confidence here is 95%. |
| **The probability of error** | The probability of error here is 5%. |

The logic is exactly the same as in the case of the estimation of a percentage. Only the method for calculating the margin of error differs and we now turn to examining it.

## Estimation of a Mean: The Calculations

Let us say that a survey involves a sample of size $n$, that the mean found in the sample is $\bar{x}$, and that the standard deviation for the population is $\sigma$. We can prove that in formulating a confidence statement for the mean of the population, the margin of error can be calculated as shown in Table 9.3.

Table 9.3 **Calculation of the margin of error when estimating a mean**

| If you want to be sure of your results at a ... level of confidence | you must allow for a margin of error of |
|---|---|
| 90% | $\pm 1.64 \frac{\sigma}{\sqrt{n}}$ |
| 95% | The margin of error is $\pm 1.96 \frac{\sigma}{\sqrt{n}}$ |
| 99% | The margin of error is $\pm 2.58 \frac{\sigma}{\sqrt{n}}$ |

There are no tables for the margin of error when estimating a mean, and these calculations must be done manually or with a calculator. SPSS computes the interval estimates, as explained in Lab 13. When the sample is large, the standard deviation calculated on the sample can be used instead of the standard deviation of the population.

For example, suppose a survey is conducted on a representative sample of 900 newborn babies in Canada and that it is found that their average weight at birth is 3.5 kg with a standard deviation of 0.5 kg. At the 95% level of confidence, the margin of error will be $\pm 1.96 \times 0.5 \div 30$ (which is the square root of 900), which gives approximately $\pm 0.033$ kg, that is, $\pm 33$ g (it is advised that you do the calculations yourself to make sure you understand the procedure). With this margin of error, we can come up with the following confidence statement:

> **The average weight of newborn babies in Canada is estimated to be 3.5 kg, with a margin of error of 33 g and a risk of error of 5%.**

Or, equivalently:

> **At a confidence level of 95%, the average weight of newborn babies in Canada is estimated to be between 3.467 kg and 3.533 kg.**

You may have noticed that the margin of error in this example is surprisingly small. This is because the sample is rather large. We are going to examine in some detail the effect of sample size on the margin of error.

As in the case of proportions, the estimate can be formulated as a *point estimate with a margin of error,* or as an *interval estimate* by subtracting and adding the margin of error to the point estimate, so as to get the whole range of values in which the estimated parameter falls.

## Effect of the Sample Size on the Margin of Error

You may have noticed that, for both percentages and means, the formula giving the margin of error includes the root of $n$ in its denominator, $n$ being the sample size. If the sample size is 400, the formula includes 20 in the denominator. If the sample size is 900, the formula includes 30 in the denominator. This means that the margin of error gets smaller and smaller as the sample size gets bigger. In fact, we can make the margin of error as small as we wish by taking a big enough sample, but that may not be practical.

> The margin of errors gets smaller and smaller as the sample size gets bigger.

For instance, suppose that the standard deviation in the population is 12 units, and that you want a 95% level of confidence. A sample of size 100 would give you the following margin of error, calculated with the formula given on the previous page:



Margin of error for $n = 100$: $\pm 1.96 \frac{\sigma}{\sqrt{n}} = \pm 1.96 \times 12 \div 10 = 2.35$ units approximately.

If you want to improve your guess and make this margin of error *half as large*, you would have to take a sample *4 times bigger*. Indeed, if the sample size is 400 units instead of 100 units, you would be dividing by the root of 400, which is 20, and you would get:

Margin of error for $n = 400$: $\pm 1.96 \frac{\sigma}{\sqrt{n}} = \pm 1.96 \times 12 \div 20 = 1.18$ units approximately.

Conclusion: **In performing an estimation, every time you quadruple your sample size, you diminish your margin of error by one half.**

Or: **In order to make the margin of error half as large as the one we have obtained, we have to take a sample which is 4 times as big as the one we have.**

A similar calculation can be done for the estimation of a proportion, because the formula for the margin of error includes root $n$ in the denominator. We can also conclude in this case that in order to cut the margin of error by half, we have to take a sample which is 4 times bigger.

## Calculation of the Sample Size Needed in a Survey

The formulas seen above are useful for planning the data collection process in a survey. One of the steps of the design of a survey consists in determining the size of the sample needed. If we plan to make inferences about the whole population, and we want the margins of error to be reasonable, we have to select a sample that is large enough. But how large is large enough? If we make it larger than necessary, the survey might be more costly and longer than needed.

Examine Table 9.1, which gives the margins of error for the estimation of a percentage. You see that if your sample includes 100 individuals, you will get margins of error as high as 10%. Notice that for every sample size the largest margin of error corresponds to a percentage of 50%, which is the percentage you may find in a sample and that you wish to generalize. Suppose that you want a margin of error no greater than 4%. What is the sample size needed? Examining the table closely, you notice that by taking a sample of 800 individuals, the margins of error when generalizing will be 4% or less.

But we can also figure out the size of the sample needed to produce a given margin of error. To do this we have to isolate the $n$ in the formula for the margin of error. For a confidence level of 95%, if $m$ is the maximum margin of error you wish to allow, the sample size must be at least:

$$\text{Size of the sample } n = \left(\frac{1.96*0.5}{m}\right)^2$$

We used 0.5 instead of $p$ in this formula because a proportion of 0.5 produces the greatest possible margin of error. If the $p$ we are generalizing is other than 0.5, the margin of error will be smaller than the maximum we have set, which is fine. Keep in mind that the numbers must be entered in this formula as *proportions* (between 0 and 1), not as percentages. Thus if you want your margin of error to be at most 4%, you enter 0.04 as the maximum margin of error accepted. What the formula gives you is the size of the sample that will give you a margin of error equal or smaller than the maximum accepted. If you take a sample greater than the $n$ you get from the formula, the margin of error will be even smaller.

A similar computation can be done when you want to generalize a mean. However, you must know the standard deviation of the population, or at least an estimate of it. If you reverse the formula given for the margin of error when estimating a mean, you get the following formula, where again $m$ is the maximum margin of error allowed:

$$\text{Size of sample } n = \left(\frac{1.96*\sigma}{m}\right)^2$$

A sample of that size or larger will produce a margin of error smaller than or equal to the one we have set as the maximum margin of error allowed.

## Summary and Conclusions

In this chapter we have seen how to estimate a mean or a proportion in a population when the corresponding statistic has been measured on a sample. In other words, we have estimated a parameter (mean or proportion) from our knowledge of the corresponding statistic.

Whenever an estimation is done, there is always a *margin of error* and a *probability of error*.

The **margin of error** reflects a lack of precision: the estimate is not exactly equal to the statistic, but falls *around* the value of the statistic, because every sample is likely to differ a little from the population.

The **probability of error** measures the risk that our estimate is wrong, that is, that the real parameter falls outside of the estimated range. This happens when the sample we have picked at random, and on which we base our estimate, differs from the population *more than expected*. The sentence 'differs from the population more than expected' means that the sample is an extreme case, presenting itself rarely. In an estimation, the risk of error that we are willing to tolerate is set first (usually at 1%, or 5%, or 10%), and then the margin of error is determined accordingly. When the risk of error is set at 5%, it means that 5% of all samples are considered to be extreme, or to differ from the population more than expected. Similarly, when the risk of error is set at 1%, it means that 1% of the samples are considered to be extreme, and when the risk of error is set at 10% it means that 10% of the samples are considered to be extreme. A notion complementary to the probability of error is the **level of confidence**, which is equal to 100% – (the risk of error).



As we said before, in an estimation we first choose the probability of error we are willing to allow (or equivalently the level of confidence we wish to have) and then we calculate the margin of error. This calculation is done with the help of the formulas given in the preceding sections. When estimating a proportion we could also use a table that gives the maximum margin of error that may result with a given sample size (Table 9.2).

The conclusion of an estimation is formulated as a **confidence statement**. The sections on estimating percentages and means have illustrated and explained all the elements that should appear in a well-formulated confidence statement. Finally, the estimation can be formulated either as a point estimate accompanied by a margin of error, or as an interval estimate that incorporates the margin of error within its range, as illustrated in Figure 9.2.

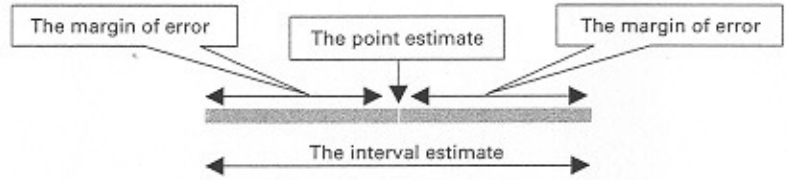


Figure 9.2

## Keywords

Confidence statement
Interval estimate
Probability of error
Point estimate
Margin of error
Confidence level

## Suggestions for Further Reading

Devore, Jay and Peck, Roxy (1997) ***Statistics, the Exploration and Analysis of Data*** (3rd edn). Belmont, Albany: Duxbury Press.

Wonnacott, Thomas H. and Wonnacott, Ronald J. (1977) ***Introductory Statistics*** (3rd edn). New York: John Wiley and Sons.

Wilcox, Rand (1996) ***Statistics for the Social Sciences***. San Diego, CA: Academic Press.